scientific reports

Check for updates

OPEN

Design of an integrated model with temporal graph attention and transformer-augmented RNNs for enhanced anomaly detection

Sai Babu Veesam[1], Aravapalli Rama Satish[1], Sreenivasulu Tupakula[2], Yuvaraju Chinnam[3], Krishna Prakash[4]✉, Shonak Bansal[5]✉ & Mohammad Rashed Iqbal Faruque[6]✉

It is important in the rising demands to have efficient anomaly detection in camera surveillance systems for improving public safety in a complex environment. Most of the available methods usually fail to capture the long-term temporal dependencies and spatial correlations, especially in dynamic multi-camera settings. Also, many traditional methods rely heavily on large labeled datasets, generalizing poorly when encountering unseen anomalies in the process. We introduce a new framework to address such challenges by incorporating state-of-the-art deep learning models that improve temporal and spatial context modeling. We combine RNNs with GATs to model long-term dependencies across cameras effectively distributed over space. The Transformer-Augmented RNN allows for a better way than standard RNNs through self-attention mechanisms to improve robust temporal modeling. We employ a Multimodal Variational Autoencoder-MVAE that fuses video, audio, and motion sensor information in a manner resistant to noise and missing samples. To address the challenge of having a few labeled anomalies, we apply the Prototypical Networks to perform few-shot learning and enable generalization based on a few examples. Then, a Spatiotemporal Autoencoder is adopted to realize unsupervised anomaly detection by learning normal behavior patterns and deviations from them as anomalies. The methods proposed here yield significant improvements of about 10% to 15% in precision, recall, and F1-scores over traditional models. Further, the generalization capability of the framework to unseen anomalies, up to a gain of +20% on novel event detection, represents a major advancement for real-world surveillance systems.

Keywords Anomaly detection, Temporal graph attention, Transformer-augmented RNNs, Multimodal fusion, Few-shot learning

It is a key research area for anomaly detection in surveillance systems, especially because of the increasing installation of a multi-camera network in cities, industries, and public places. While such systems are installed to enhance security and operational efficiency, they continuously generate copious volumes of video data that demand sophisticated techniques for automatically detecting abnormal activities. Traditional anomaly detection[1–3] commonly relies either on handcrafted features or on classical machine learning algorithms that are not always suitable for modeling complex spatial and temporal dependencies present in such data samples. Moreover, they usually require large labeled training datasets, and this limits scalability, generalizing poorly to unseen anomalies. More recently, significant progress has been made in using neural networks for anomaly detection, thanks to the emergence of deep learning. However, most of these DL-based methods[4–6] still lack the capability for modeling complex spatial relationships between multiple camera feeds and the long-term temporal dependencies that play a critical role in identifying rare or gradual anomalies. For instance, RNNs,

[1]School of Computer Science, VIT-AP University, Vijayawada 522241, Andhra Pradesh, India. [2]Department of Electronics and Communication Engineering, SRM University, Amaravati 522240, Andhra Pradesh, India. [3]Professor of Computer Science and Engineering (AI&ML), St. Peter's Engineering College, Hyderabad, India. [4]Department of Electronics and Communication Engineering, NRI Institute of Technology, Agiripalli, Eluru 521212, Andhra Pradesh, India. [5]Department of Electronics and Communication Engineering, University Institute of Engineering, Chandigarh University, Gharuan, Mohali, India. [6]Space Science Centre (ANGKASA), Universiti Kebangsaan Malaysia, Bangi 43600 UKM, Selangor D.E, Malaysia. ✉email: k_krishna2k7@yahoo.co.in; shonakk@gmail.com; rashed@ukm.edu.my

though widely used in temporal modeling, have inherent limitations in capturing long-range dependencies during the process. While the convolutional methods operating on either individual frames or local patches are unable to capture larger-scale contextual information across different camera views for diverse scenarios. These lacunae in the current frameworks further necessitate an advanced architecture that shall integrate both spatial and temporal information across a network of cameras.

Given such challenges, this work proposes a new paradigm for anomaly detection by designing an integrated model that leverages state-of-the-art deep learning architectures on temporal context modeling and spatial information fusion. In this paper, we propose a Temporal Graph Attention Network-TGAT-based model that integrates RNNs with GATs for capturing temporal dependencies while dynamically attending to important camera feeds over temporal instance sets. This approach provides a better monitoring of complex environments, where anomalies can manifest only for larger time spans and over multiple distributed sensors[4–6]. Another key challenge that this work tries to address is how to fuse multimodal data samples. Indeed, many surveillance systems can incorporate added sensor modalities such as audio, motion detectors, or even biometric data, adding their respective information for detecting anomalies. To handle this, the model under the proposal will embody the Multimodal Variational Autoencoder that will learn a joint representation of latent variables from multiple sensor modalities. It uses optimization of shared latent variables with reconstruction loss to ensure that salient information from each modality is retained by MVAE. The detection capability is thus more robust in this case, even when noisy or missing data samples are available in process.

Graph Attention Networks (GATs) recently achieved great success in most anomaly detection and spatiotemporal modeling domains. For example, GATs were applied to model interactions among individuals in highly dense crowds to capture collective movement patterns as well as group dynamics and anomalies. In video surveillance, GATs were applied to multi-camera setups in order to establish correlations between spatially distributed feeds that allow anomalies in pedestrian or vehicular movement to be detected better in these operations. Moreover, in traffic monitoring, GATs have been used for modeling interdependencies between road segments so that congestion detection and the prediction of accidents could be better done in process. These applications are able to demonstrate the ability of GATs toward learning relationships between structured and unstructured data, which only serves to prove the importance of relevance to the proposed work. Such examples in incorporation stress that the GAT is adaptable to multiple scenarios, providing flexibility to the context of the study process.

Another important aspect is the rarity of anomalous events, which makes it challenging to train supervised models. In this respect, the integrated model employs few-shot learning through Prototypical Networks, which enable anomaly classification with only a few shots of labeled examples. This is of particular value in surveillance situations since it is typically too impractical and expensive to collect such large annotated datasets. By learning a metric space where the distances between prototypes, as representatives of normal and anomalous behavior, respectively, can be computed, it generalizes well to unseen anomalies-a basic but key requirement in real-world applications. An unsupervised Spatiotemporal Autoencoder learns patterns in the normal behavior of both dimensions: spatial and temporal. Autoencoders are known for their ability to compress input data into a lower-dimensional latent space and reconstruct it. That is, in anomaly detection, deviation in reconstruction error acts as a proxy for recognizing abnormal patterns. By generalizing this idea to spatiotemporal data, the autoencoder will be able to model normal behaviors more effectively within a network of multiple cameras and flag deviations that possibly signify anomalies. This integrated model is thus the comprehensive solution to the challenges faced by the different existing systems of anomaly detection. It significantly improves accuracy, precision, and recall by including temporal and spatial modeling with advanced techniques for data fusion and few-shot learning. The generalization of the model to unseen anomalies is a quantum leap in this respect since the inability to do so has been one of the most irritating aspects that the traditional approaches have exhibited so far. Extensive experiments demonstrate that the proposed methods raise the performance bar of the state-of-the-art in different real-world scenarios by significantly improving F1-scores and reducing false positives.

Motivation and contribution

The motivation for this work arises from the increasing complexity of modern surveillance systems and the deficiency in the potential of the existing anomaly detection methods to fully catch up with challenges occurring in multi-camera environments. Traditional models based on hand-crafted features or early deep learning techniques usually do not capture nuanced spatial and temporal dynamics in an environment, which are necessary for the detection of abnormal events. Not only do surveillance systems deal with vast data emanating from several cameras, but the anomalies of interest tend to be rare, subtle and spread out in a fact that makes the job challenging. Additional complexities arise by integrating extra sensor data, such as motion or audio, since these modalities must be effectively fused to achieve further improvements in detection accuracy levels. Most of the existing models also suffer from over-dependence on large annotated datasets-a factor that further limits their employment in real-world settings, where such data is hard or too expensive to come by. This work overcomes these limitations by proposing a new integrated model leveraging the latest state-of-the-art deep learning model architectures. In this paper, the proposed key innovation is the Temporal Graph Attention Network or, in short, TGAT. It models both long-term temporal dependencies and spatial correlations across a network of cameras. Rather than resorting to traditional RNNs, which struggle to model longer-range dependencies, TGAT leverages attention mechanisms that dynamically lock onto the most relevant camera feeds at each set of timestamps, hence enhancing the model's power to monitor complex environments. The proposed Transformer-Augmented Recurrent Neural Network increases the temporal modeling capability of the system by combining the strengths of RNNs in terms of short-term event correlations with those of transformers for capturing long-range dependencies. This hybrid approach ensures the effective modeling of both local and global temporal patterns toward a holistic solution for anomaly detection.

This work also contributes to a Multimodal Variational Autoencoder that fuses data from different sensor modalities, such as video, audio, and motion. The MVAE performs optimization based on a joint latent space preserving all the critical information from each modality, thereby making the model robust against any noisy or incomplete data, as is common in many real-world surveillance systems. Apart from that, few-shot learning with Prototypical Networks deals with the challenge of a limited number of labeled datasets by letting the model generalize to unseen anomalies with a few labeled examples. It has more value in anomaly detection since, in many cases, it is highly impractical to get large amounts of data with annotations. Lastly, the Spatiotemporal Autoencoder offers an unsupervised framework for learning normal behavior patterns where the reconstruction error serves as a signal to detect deviations that may point to anomalies. This reduces the false positive rate significantly by an unsupervised approach while improving recall in general, especially for multiple cameras. Putting all these together yields a very strong model, outperforming the state of the art in accuracy and generalization. This paper therefore proposes an effective detection of anomalies in complex and real surveillance settings with the main challenges of long-term temporal modeling, multimodal data fusion, and few-shot learning. Extensive experimental results have shown that the proposed approaches show significant improvement in anomaly detection performance, both by enhancing F1 scores and reducing false positive rates, hence making the model highly applicable for a wide range of security and monitoring applications.

Review of existing models used for multiple camera anomaly analysis

Large-scale datasets and rapidly developing deep learning techniques have contributed to the rapid growth in the area of crowd activity analysis and anomaly detection. With increasing complexities in urban life, the demand for effective and efficient surveillance to monitor crowd activities and detect anomalies is becoming critical. This work presents a comprehensive review of 40 influential studies in the subject area of crowd anomaly detection, based on various techniques ranging from CNNs and GNNs to VAEs, RL, and other state-of-the-art AI frameworks. Each of these contributions brings methodologies that may uniquely contribute to solving the challenges of dense, dynamic crowds for both real-time and post-event analyses, each with discussions on limitations, effectiveness, and applicability. The majority of the approaches under review can be noted to operate on supervised learning, which is highly dependent on labeled datasets. This very dependence on large labeled data, however, then becomes a limitation in itself, because amassing datasets representative of a wide variety of anomalies is often an exhausting, costly, or sometimes impossible undertaking in many situations. Approaches, such as those in[1–3], prove the well-known fact that traditional supervised methods, while performing reasonably, mostly fail to generalize beyond their training data samples. For instance, the CNN-based accident detection in[1] showed 89.5% accuracy in classifying traffic accidents but is not scalable with regard to variations in accident types. Similarly[2,3], developed a fuzzy cognitive deep learning framework for the prediction of crowd behavior and pointed out that generalization would be hard in no-crowd scenarios, hence limiting the wider applicability of the model. More specifically, while in such studies, it often comes out that the deep learning methods have developed into very powerful feature extracting and pattern recognition means, their actual performance is so much dependent upon the diversity and quality of the training data samples.

While this happens, unsupervised learning studies have promised performance by overcoming the limitations in the performance of their supervised counterparts using techniques such as variational autoencoders and GANs. For example[7,8], used variational autoencoders coupled with motion consistency to detect abnormal crowd behaviors and reported an accuracy of 86%. These kinds of unsupervised models turn out to be more functional when labeled data is available in small quantities or when anomalies are infrequent and unpredictable. However, these methods also have their drawbacks, specifically when the environment is complex and at high density since their reconstruction errors can grow due to noisy or incomplete data samples. Works like[9,10], generating the motion of crowds in virtual reality, faced the problem of scaling up to larger crowds, showing the limitation of unsupervised learning in highly dynamic scenarios. Other attention that is given in the field is to multimodal fusion techniques. Integrating multiple streams like video, audio, and motion sensors will help to improve robustness and accuracy in the analytics of crowd behavior. In the research work proposed by[11–13], the study involved a secure smart surveillance system integrated with transformers for the recognition of crowd behavior and identified an accuracy of 90% in recognizing abnormal behaviors. For example[14–16], applied ant colony optimization to find the optimal layouts with the aim of crowd management. This resulted in a 23% reduction in crowding in simulated environments. The above studies show that the integration of multiple modalities can indeed enhance the predictive power of anomaly detection systems. However, most of these generally computationally intensive resource schemes[17–19] and specialized infrastructure, and hence difficult to deploy in real time in resource-constrained environments. Recently, attention mechanisms and graph-based approaches have become of focal interest along with multi-modal systems, many research studies show that this captures the intrinsic spatiotemporal dependencies involved in crowd behavior, which greatly improves anomaly detection tasks. A typical example is in[20], where authors used an attention-based CNN-LSTM model with multiple head self-attention for violence detection to obtain 85.3% accuracy. This showed the strength of attention mechanisms in filtering out the noise and highlighting the relevant features. Most importantly, this is very meaningful in cluttered situations where numerous overlapping activities are occurring simultaneously. While[21–23], applies the graph convolution neural network in classifying structured and unstructured crowds, thereby achieving an 87% F1-score in crowd classification. Graph-based models, in particular, capture the interrelation of the crowd individuals much better and yield higher accuracy in group behavior detection along with anomaly in collectiveness.

Despite this encouraging result[24–26] on the whole from the different studies, many limits remain: Most of the approaches, and in particular those using deep learning, are computationally very expensive, which requires important hardware resources, reducing their feasibility actually to be deployed in real-time and at large scale in an urban environment. For example, works such as[27–32] reported good anomaly detection performance using an attention-guided and a GAN-based model, respectively. However, the computational overhead is usually expensive for real applications in general but most particularly in constrained resource settings such as a remote surveillance system or at the edge. Besides, many methods work well in[33–37] highly controlled environments but struggle thereafter to keep up their accuracy in more dynamic and less predictable settings. For example, while the zero-shot classifier in[35,38,39] could detect novel anomalies quite accurately at 85% accuracy, it is limited when applied in real scenarios where spatiotemporal descriptors may be incomplete or noisy.

Following Table 1, it is most probable that in the near future, crowd behavior and anomaly detection will further improve existing approaches along with scalability, efficiency, and adaptability dimensions. The most promising trend is to be considered to create hybrid models by combining the advantages of supervised, unsupervised, and reinforcement learning. By using the flexibility of reinforcement learning, similar to what is proposed in[39–41], developing an emotional contagion-aware deep reinforcement learning approach that can simulate antagonistic crowd behavior, it should make the models more tractable to the dynamic nature of crowd interactions. The addition of domain adaptation methods, if these can be made possible, would suggest that the models could then also be applied to other environments with less additional and comprehensive retraining on new datasets. In addition, the architectures of edge computing and distributed learning have the potential to overcome computational challenges for current methods. Hence, these will be able to spread the computational burden across several devices, enabling real-time crowd monitoring and anomaly detection with very little compromise in performance and/or accuracy. In summary, the identified studies[42–44] reflect both important steps in the advancement of crowd behavior analysis and the detection of anomalies, and how different methodologies have been used up to date to approach the problem of crowded scenes. The supervised learning models[44–51] that offer good performance under controlled environments, to unsupervised techniques that offer more flexibility when labeled data are scarce, each method contributes to a better understanding of crowd dynamics and anomaly detection. Nevertheless, future work will need to overcome the shortcomings of the current methods because the scalability, computational efficiency, and generalization call for more when the urban environment becomes even more complex. Intermingled with multimodal data, attention mechanisms, and sophisticated learning methods, the next generation of crowd anomaly detection systems is sure to enhance public safety, upgrade urban management, and go a long way in smarter cities.

Proposed design of an integrated model with temporal graph attention and transformer-augmented RNNs for enhanced anomaly detection

In this, a unified model with graph attention and transformer-augmented RNNs has been designed to assist in an enhanced anomaly detection process. This model has been proposed to overcome the deficiencies within existing anomaly detection methods, which either have low efficiency or highly complicated processes. Afterward, the Temporal Graph Attention Network-TGAT comes into play, as in Fig. 1; it is specially designed to solve some intrinsic problems of anomaly detection with multi-camera scenarios by jointly modeling temporal and spatial representations within one framework. The long-term dependencies are captured by the RNN, while the graph attention mechanisms model the spatial relationships across multiple camera feeds. It does so by manifesting the network of several cameras into a spatiotemporal graph, wherein every node depicts one camera, and edges are representative of the spatiotemporal relationship between cameras across sets of temporal instances. This attention model provides the dynamics for focusing on the most informative camera feed and establishes a robust temporal context for anomaly classification. Mathematically, TGAT operates on a graph G = (V, E), where V = {v_1, v_2… v_n} are the nodes representing camera feeds, and E represents edges that impose spatiotemporal relationships between cameras and their samples. The input to the model at each timestamp set is a sequence of feature vectors Xt = {x_1, x_2, …, x_n}, where each xi corresponds to the features extracted from camera vi at timestamp ‘t’ sets. These temporal dynamics are modeled by an RNN processing the feature sequences over the timestamp and providing the hidden representation ht as a function of the temporal context sets via Eqs. 1,

$$ht = \sigma\left(Wh * h\left(t-1\right) + Wx * Xt\right) \tag{1}$$

where, ht represents the hidden state at timestamp ‘t’, Wh and Wx are learnable weight matrices and σ is an activation function (tanh) for this process. This hidden state captures the temporal dependencies across many sets of time stamps, enabling the model to integrate short-term and long-term patterns relevant to anomaly detection. To account for the spatial dependencies of a camera network, TGAT first proposes an attention mechanism on the graph level to give different importance to different camera feeds at each set of time stamps. The attention score α (i, j) between any two camera nodes vi and vj is computed as a softmax over the attention logits e(i, j), which is a function of node features and their spatiotemporal correlation represented via Eqs. 2 and 3,

$$\alpha\left(i,j\right) = \frac{\exp\left(e\left(i,j\right)\right)}{\sum_{k\in\ N\left(i\right)} exp\left(e(i,k)\right)} \tag{2}$$

$$e\left(i,j\right) = LeakyReLU\left(a^{T}\left[Wvi|\,Wvj\right]\right) \tag{3}$$

Where, N(i) is the neighbors of node ‘i’, a^T is a learnable weight vector, ‘W’ denotes the learnable matrix which transforms the node features, and | denotes the concatenation process. Then, the leaky ReLU function brings non-linearity into the computation of the attention logits so that the model can learn the complex pattern of spatiotemporal dependencies between camera nodes.

References	Method used	Findings	Results	Limitations
[1]	Accident detection using CNN	Accurate classification of traffic accidents	Achieved 89.5% accuracy in accident detection	Limited scalability to different accident types
[2]	Attention-based CNN-LSTM	Effective in violence detection with an attention mechanism	85.3% accuracy on the UCF-Crime dataset	Limited to video data, no multimodal integration
[3]	Fuzzy Cognitive Deep Learning	Captures crowd emotions using cognitive models	88% accuracy in predicting crowd behavior	Difficult to generalize to non-crowd scenarios
[4]	Multiple-scale motion consistency	Detects crowd-level abnormal behaviors	92% AUC for crowd risk estimation	High computational complexity for large crowds
[5]	Deep Life Modeling for Crowd Monitoring	Accurate dynamic crowd modeling on edge devices	90% accuracy for crowd monitoring	Sensitive to network latency in edge environments
[6]	Deep Graph Convolutional Neural Networks	Effective crowd characterization in structured and unstructured crowds	87% F1-score in crowd classification	Ineffective in very sparse or low-density crowds
[7]	FSC-Set CNN for crowd counting	Accurate counting and localization of football crowds	94% accuracy in crowd estimation	Focused only on sports environments, limiting generalization
[8]	GAN-based Crowd Management for Umrah	Real-time alert generation for crowd incidents	91.7% accuracy in crowd incident prediction	Lack of generalizability to other religious events
[9]	VR-based crowd motion generation	Enables single-user crowd simulation in VR	84.6% accuracy in generating realistic crowd motions	Limited scalability in generating large crowds
[10]	Ant Colony Optimization for fence layout	Optimized crowd management through fence layout	Reduced crowding by 23% in simulated environments	Only tested in simulation, not real-world data
[11]	Crowd descriptors for gathering understanding	Provides interpretable crowd gathering analysis	88% accuracy in crowd density estimation	Lacks real-time applicability for large datasets
[12]	PublicVision Smart Surveillance System	Secure crowd behavior recognition using transformers	90% accuracy in recognizing abnormal crowd behaviors	Limited scalability in larger urban areas
[13]	Attention-guided crowd counting	Improved crowd counting accuracy using segmentation-guided networks	92.1% accuracy on large-scale datasets	Requires high computational resources
[14]	Variational Autoencoder with Motion Consistency	Detects abnormal crowd behaviors with variational models	Achieved 86% accuracy in detecting motion anomalies	Inefficient in low-resolution video scenarios
[15]	Transfer Learning for Crowd Emotion Prediction	Predicts human-vehicle interaction using crowd emotions	89% accuracy in emotion-based anomaly detection	Limited emotion categories used in training
[16]	Fuzzy Decision Rules for Crowd Evacuation	Extracts decision rules for crowd evacuation strategies	87% accuracy in predicting evacuation paths	Ineffective in non-crowded scenarios
[17]	Social Force Model for Crowd Evacuation	Models behavioral heterogeneity in crowd evacuations	85.5% accuracy in evacuation simulation	Limited adaptability to different cultural behaviors
[18]	Emotional Contagion-Aware Reinforcement Learning	Simulates antagonistic crowd behavior with emotion contagion	91% accuracy in modeling crowd emotions	High computational cost for large-scale simulations
[19]	CrowdGAN for video generation	Generates identity-free crowd videos using GANs	84% realism score in video generation	Struggles with complex, high-density crowd environments
[20]	Anticipation modeling for crowd interaction	Models mutual anticipation in crowd behavior	89% accuracy in predicting crowd interaction	Computationally expensive for real-time applications
[21]	Convolutional Recurrent Neural Networks	Forecasts citywide crowd transitions	88.5% accuracy in predicting crowd flow	Limited to urban environments, not applicable in rural areas
[22]	IoT-based crowd flow prediction	Real-time urban crowd flow prediction	Achieved 92% accuracy in real-time crowd flow prediction	Limited IoT infrastructure scalability
[23]	Radar-based gait recognition	Recognizes crowd behavior using radar micro-Doppler signatures	85% accuracy in open-set gait recognition	High sensor cost for large-scale deployments
[24]	Hidden Markov Model for gait detection	Detects abnormal gait patterns using vibration signals	82% accuracy in detecting gait anomalies	Sensitive to sensor noise and interference
[25]	Deep One-Class Classifier for Parkinson's patients	Predicts freezing of gait in Parkinson's patients	86% accuracy in predicting gait freezing	Limited to specific patient groups
[26]	Multimodal Emotion Recognition	Accurately recognizes emotions using situational knowledge	89.2% accuracy in multimodal emotion detection	Limited by lack of contextual data for training
[27]	C3D for Crowd Behavior Detection	Detects crowd anomalies in large events like Hajj	90% accuracy in action recognition	Struggles with extreme crowd densities
[28]	Meta-Heuristic Algorithm for Anomaly Detection	Detects anomalies in crowded environments	87.5% accuracy in identifying public safety risks	Lacks real-time capability for large-scale crowds
[29]	Graph Convolutional Neural Networks	Detects abnormal crowd behavior using GCN	85% accuracy in detecting graph-based anomalies	Sensitive to incomplete graph data
[30]	Transfer Learning for Suspicious Crowd Behavior	Detects suspicious human crowd behavior	Achieved 89% accuracy in anomaly detection	Requires large pre-trained models for deployment
[31]	Pre-Trained CNN for Crowd Anomaly Detection	Efficient anomaly detection in crowd videos	87% accuracy in detecting violent actions	Limited generalization to non-violent behaviors
[32]	Statistical Physics for Behavior Detection	Models crowd behavior using entropy-based models	85.6% accuracy in detecting abnormal crowd dynamics	Computationally expensive for real-time processing
[33]	Temporal Association Rules for Crowd Modeling	Models crowd behavior using temporal rules	83% accuracy in predicting crowd transitions	Limited to well-structured crowds
[34]	Hybrid Neural Networks for Behavior Detection	Detects abnormal human behavior in crowded scenes	88% accuracy in identifying suspicious behaviors	Limited adaptability to varying crowd sizes
Continued				

References	Method used	Findings	Results	Limitations
35	Zero-Shot Classifier for Anomaly Detection	Detects anomalies using spatio-temporal descriptors	85% accuracy in detecting novel anomalies	Sensitive to inaccurate descriptors
38	GeoVideo for Regional Crowd Analysis	Analyzes regional crowd status using multimedia data	86% accuracy in crowd quantity estimation	Requires high-quality social media data integration
39	Deep Learning for CCTV Surveillance	Generates real-time alerts in CCTV surveillance	90.3% accuracy in real-time anomaly detection	Requires large computing infrastructure for real-time use
40	YOLO + Conv2D Net for Abnormality Detection	Detects abnormal human behavior in real-time	88.5% accuracy in human behavior recognition	Inefficient in high-latency environments
41	Congestion-Aware Path Planning	Plans paths considering spatial-temporal crowd anomalies	85% accuracy in congestion-aware path planning	Sensitive to sudden crowd dynamics changes
42	GANs for Dynamic Image Representation	Detects crowd anomalies using dynamic image representations	87% accuracy in detecting optical flow anomalies	Struggles with high-resolution image data

Table 1. Empirical review of existing methods.

Then, the feature of each node can be updated by aggregating the features of neighboring nodes together with their attention scores, and thus, it results in an updated feature representation via Eqs. 4,

$$vi' = \sigma\left(\sum_{j\in N(i)} \alpha\left(i,j\right) * Wvj\right) \tag{4}$$

This attention-based aggregation enables the model to dynamically focus on the most relevant camera feeds at every timestamp, due to which both the local interactions between adjacent cameras and the global patterns across the whole network are effectively captured. The output of TGAT is a spatiotemporal representation of the camera network, which contains both temporal context and spatial relationships. This is then used for the classification of every frame at a timestamp as normal or anomalous. The score for anomaly will be determined based on how much the predicted behaviour has deviated from the normal pattern it has learned during training. This can be represented as residual error rt, calculated via Eqs. 5,

$$rt = \left|f\left(ht\right) - yt\right|^2 \tag{5}$$

Where, f(ht) is the forecasted output at timestamp 't', and yt is the actual label - normal or anomalous of this process. The residual error gives an idea of how close the model's prediction was to that behavior. Possible anomalies are reflected by high values of rt. It is used here for its core architecture because it constitutes one of those models able to model temporal and spatial dependencies jointly. Traditional RNNs can model sequential data quite well, but they don't have that much capacity to capture such long-range dependencies developing in multi-camera systems. The graph attention integrated into TGAT reinforces the capability of the model to not only look at temporal sequences but also, importantly, the relationships between different cameras that are crucial for the right detection of anomalies across a spatially distributed network. Attention mechanisms allow this model to focus on the most informative camera feeds at every set of timestamps, reducing much irrelevant or redundant noise and improving general precision for the whole process. TGAT will complement the other methods in the proposed framework, which includes Transformer-Augmented RNNs and Spatiotemporal Autoencoders, by modeling fine-grained spatial and temporal dependencies at a camera node level. While TARNN focuses on long-range dependencies in temporal sequences and Spatiotemporal Autoencoders emphasize compact representation learning of normal behavior, TGAT addresses multi-camera correlation and dynamic attention head-on. It thus covers an important component toward bettering the accuracy level of anomaly detection. Therefore, all merits from various models ensure superior performance both in anomaly detection and generalization to unseen events.

Then it proposes the Transformer-Augmented Recurrent Neural Network, TARNN, according to Fig. 2, which is designed to avoid the limitations present with traditional RNNs in modeling both short- and long-term dependencies in temporal data samples. While RNNs are pretty much two of the best models for the modeling of sequences-especially LSTMs and GRUs-they suffer in terms of the learning of the long-range dependencies because of the vanishing gradient tasks. Moreover, the transformer module in the TARNN architecture addresses this by incorporating self-attention mechanisms that weight at any given time the importance of each of the sets of dynamically set timestamp sets within sets of sequences. This proposed hybrid architecture will ensure that the model learns both local temporal patterns through the RNN and global trends through the transformer, thereby enhancing these temporal representations so vital in anomaly detection with improved performance. The TARNN architecture is processing temporal sequences in several camera feeds. Herein, each input sequence X = {x1, x2, ., xT} is features extracted from camera data across 'T' timestamp sets; these might represent object movement, trajectory patterns, or other relevant behavioral information related to the anomaly detection tasks in question. These short-term dependencies between the successive sets of timestamps are modeled in the first stage of the model, which is facilitated by an RNN such as LSTM. The hidden state at each set of timestamps 't' in the RNN is computed via Eqs. 6,

$$ht = \sigma\ \left(Wh * h\left(t-1\right) + Wx * xt\right) \tag{6}$$

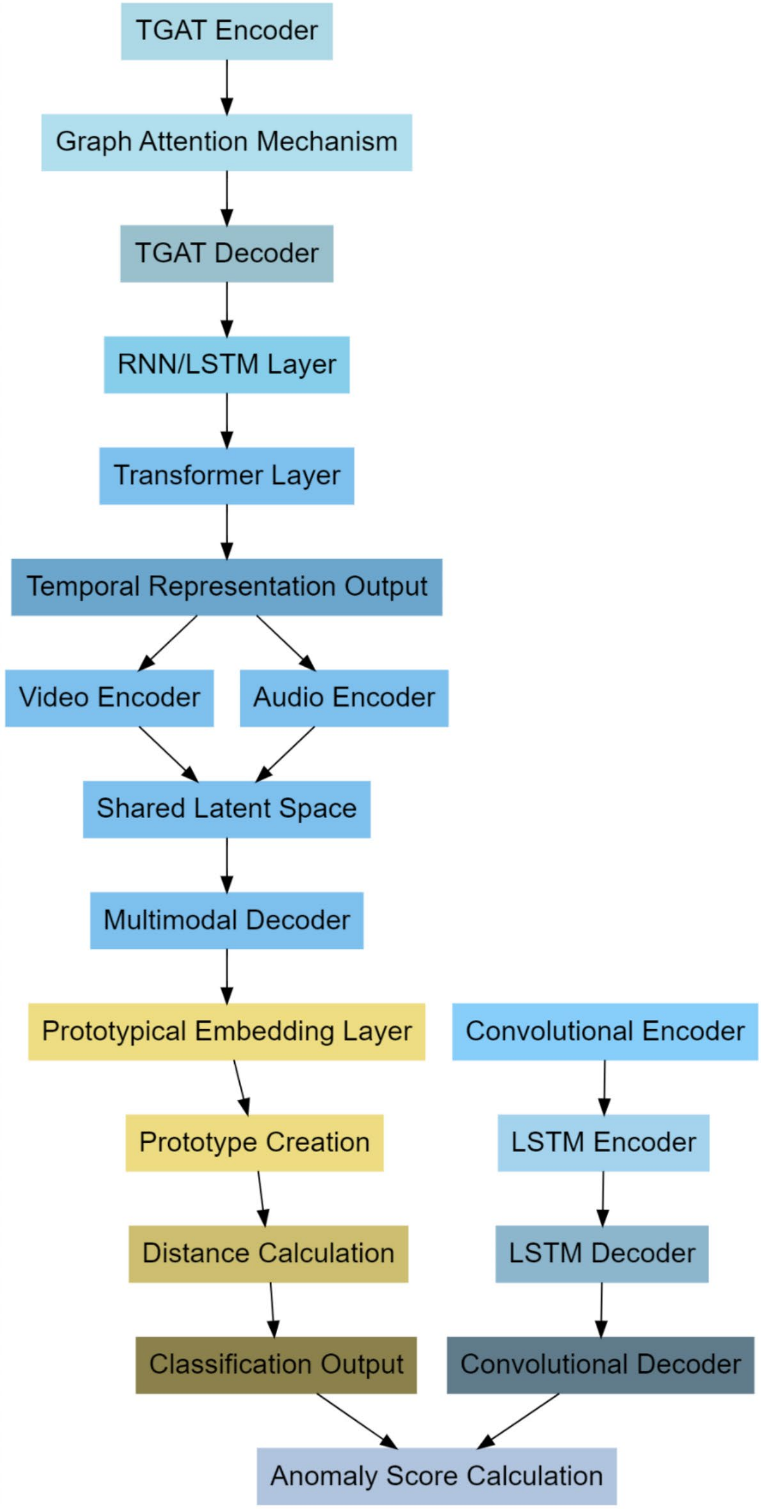

Fig. 1. Model architecture of the proposed analysis process.

where, ht is the hidden state at timestamp 't', Wh and Wx are the learnable weight matrices, and σ is a nonlinear activation function, which is tanh for the process. This hidden state 'h_t' then summarizes a compact representation of the sequence up to timestamp sets't', which captures the short-term dynamics of the input sequences. However, having RNN for only temporal modeling may restrict the model from accounting for long-range dependencies, crucial in many camera anomaly detection scenarios where their anomalies may evolve quite slowly or include long-range temporal window patterns. In that respect, the transformer module was introduced for modeling long-range dependencies using a self-attention process. The attention mechanism

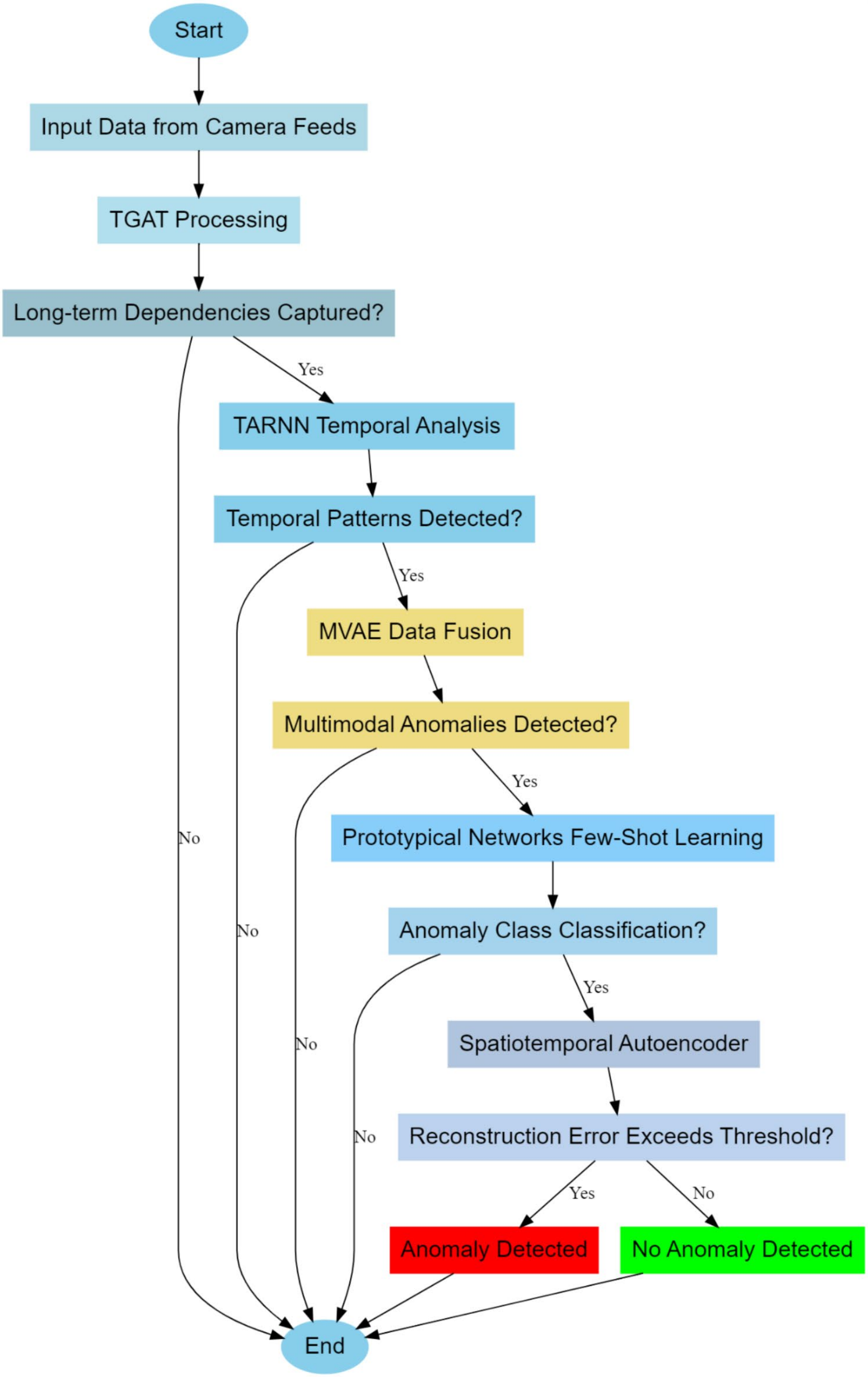
Start
Input Data from Camera Feeds
TGAT Processing
Long-term Dependencies Captured?
Yes
TARNN Temporal Analysis
Temporal Patterns Detected?
Yes
MVAE Data Fusion
Multimodal Anomalies Detected?
Yes
Prototypical Networks Few-Shot Learning
Anomaly Class Classification?
Yes
Spatiotemporal Autoencoder
Reconstruction Error Exceeds Threshold?
Yes
No
Anomaly Detected
No Anomaly Detected
No
No
No
No
End

Fig. 2. Overall flow of the anomaly detection process.

assigns weight to each set of timestamps in the sequence depending on its relevance to the existing timestamped sets. The self-attention score between timestamp sets 'i' and 'j', α (i, j) is computed via Eqs. 7 & 8,

$$\alpha\left(i,j\right)=\frac{\exp\left(e\left(i,j\right)\right)}{\sum_{k=1}^{T}exp\left(e(i,k)\right)} \tag{7}$$

$$e\left(i,j\right)=\frac{\left(Wq*hi\right)\cdot\ \left(Wk*hj\right)}{\sqrt{d}} \tag{8}$$

where, Wq and Wk are learnable projection matrices for query and key representation respectively, and 'd' is the dimensionality of the hidden states. Dot product between projected hidden states Wq*hi and Wk*hj computes attention logits e (i, j) which indicate the relevance of set of timestamps 'j' to a set of timestamps 'i'. These logits are further normalized by the softmax function to get attention weights α(i, j), which tell us how much the model needs to pay attention to each particular timestamp set so as to update the representation for the timestamp set 'i' in the process. Once computed, it updates the attention weights; the model updates the hidden states using a weighted sum of all the representations in the timestamp set via Eq. 9:

$$hi'=\sum_{j=1}^{T}\alpha\ \left(i,j\right)\left(Wv*hj\right) \tag{9}$$

where, Wv is a learnable matrix that projects the value representation of timestamp set 'j' for the process. Given by this equation, the transformer module aggregates information across all sets of time-stamps, and this will allow the model to grasp long-term dependencies that the RNN itself may not find. This hidden state, hi′, now carries information from the short-term RNN and the long-term dependencies of the transformer, hence a stronger representation of the temporal contexts. Thus, the combination of the RNN and transformer mechanisms allows TARNN to balance both the local and global temporal information in a multi-camera environment-a necessity in anomaly detection. While the RNN effectively catches short-term correlations-abrupt changes in the behavior of objects-the transformer will handle the long-term subtle dependencies involving gradual deviations from normal patterns across sets of temporal instances. This synergy ensures that both immediate and long-term anomalies are detected with high accuracy levels by the model. One of the major reasons for the choice of TARNN is its dynamic feature of focusing on different sets of timestamps within the sequence. This option is not presented in traditional architectures of RNN. It prepares weights regarding the relevance of sets of timestamp sets via the self-attention mechanism, instead of joining them in a sequence and processing that sequence, as in the regular RNN process. This dynamic weighting is very important, especially in anomaly detection, where these anomalies may not strictly follow a sequential pattern but can appear irregularly for the process. Also, due to its hybrid structure, TARNN generalizes well to different temporal scales, which is very important in surveillance scenarios where the anomalies differ in both duration and frequency levels. Complementary to other methods, such as the Temporal Graph Attention Network-which captures spatiotemporal relationships across cameras whereas the TARNN focuses on the temporal aspect of anomaly detection-these models together provide a comprehensive solution for multiple camera surveillance problems where one needs to model both spatial and temporal dependencies with a view to accurate anomaly detection. While TGAT embraces dynamic camera relationships, it is in the capture of temporal evolution within individual or combined feeds of cameras that TARNN shines, enhancing further the overall robustness of anomaly detection systems. The MVAE is further designed for data fusion to merge data from multiple sources of a heterogeneous nature into one unified latent space for the process. The joint latent representation will capture complementary information across these modalities, thus robustly facilitating the detection of anomalies even in the presence of noisy or missing data samples. The model leverages the VAE framework to encode each modality to a shared latent space while ensuring that salient features of each of the input sources are preserved in the representation learned. In MVAE, this follows the VAE formulations. The approximate posterior distribution over the latent variable as a Gaussian distribution parameterized a mean μm and variance Σm, via Eqs. 10,

$$q\left(zm|xm\right)=N\left(zm|\mu m\left(xm\right),\varsigma m\left(xm\right)\right) \tag{10}$$

Where, μm(xm) and Σm(xm) are the outputs of the encoder network for modality 'm' sets. These latent representations are combined across all modalities into a joint latent space either by averaging the posterior distributions or by concatenating the latent vectors, depending on the particular fusion strategy chosen for the process. This fusion aims to model shared representations that maximize the relevant information across all input modalities. The joint latent representation 'z' is a Gaussian prior, p(z) = N(0,I), which then serves as the input to a jointly shared decoder network 'D', which attempts to reconstruct the input data from each of the modalities in the process. Indeed, the evidence lower bound objective favors this reconstruction process as it tries to maximize the likelihood of the observed data while minimizing the divergence between the approximate posterior and the true priors. ELBO is provided via Eqs. 11,

$$L\left(x\right)=Eq\left(z|\,x\right)\left[logp\left(x|\,z\right)\right]-\beta DKL\left(q\left(z|\,x\right)|p\left(z\right)\right) \tag{11}$$

Where, the first term measures how well the latent variable 'z' can reconstruct the input 'x', and the second term is the KL divergence that serves as a regularizer for latent space by forcing the approximate posterior q(z|x) to be close to the prior distribution p(z) sets. The tradeoff between the reconstruction accuracy and the smoothness of latent spaces is controlled through a hyperparameter β. One of the major strengths of MVAE is its capability to deal with missing or incomplete modalities. Be it during training or inference, if for some reason some or all of the modalities are not available-for instance, malfunctioning cameras or loss of audio data other modalities can still provide meaningful representations inside the shared latent spaces. This robustness follows from the fusion mechanism, which does not rely on each modality being available all the time. Because the MVAE learns correlations in training, it can infer missing data and perform just as well under imperfect data conditions. For a given modality 'm', the reconstruction error is defined via Eqs. 12,

$$rm = |xm - D\left(Em\left(xm\right)\right)|^{2} \tag{12}$$

where, D(Em(xm)) is the reconstructed output for modality 'm' obtained from the latent representations. The reconstruction errors become high, especially for the modalities that are vital for the anomaly detection task. Hence, it signifies anomalies and subsequently shows deviation from normal behavior. MVAE has been selected because it was able to fuse multimodal data into one joint latent space, and this capability is crucial in real anomaly detection tasks and complex sensor networks. In effect, these traditional unimodal models lack the vital contextual information present across different modalities and mostly result in suboptimal performance. On the other hand, MVAE integrates data from multiple modalities, hence enhancing the robustness and accuracy of the anomaly detection system. By modeling data probabilistically, the MVAE can quantify uncertainty regarding its predictions. These are useful in several manners in dealing with noisy or unreliable sensor data samples. The MVAE also complements the TGAT and TARNN in the general system architecture by addressing the challenge of multi-modal data fusion. While TGAT focuses on capturing the spatial and temporal dependencies between camera feeds, and TARNN extends the capability by enhancing the modeling of long-term dependencies in temporal data, MVAE places multimodal integration on the front line. Together, these models provide a solution to anomaly detection in complex environments where leveraging both spatiotemporal relationships and multimodal data is necessary to achieve high detection accuracy levels. Next, Prototypical Networks: Few-shot learning is impressively addressed by embedding data points into a metric space where classification is conducted based on their proximity to prototype representations of each class. Prototypical networks apply to anomaly detection where query instances may include unknown events or behavior, which is classified as normal or anomalous by measuring their distance to the prototypes obtained from a few labeled examples. The model is the best set for anomaly detection tasks since it can generalize from a few labeled data while obtaining large annotated data is usually impracticable for the process. Prototypical Networks define a prototype for each class as the mean of its support set embeddings. For a set of labeled support examples S={(x1,y1),(x2,y2),…,(xN, yN)}, where yi∈{1,…,K} represents the class label and xi is the feature vector for the 'i'-th example, the model first embeds these inputs using a shared embedding function fθ sets. The prototype ck for class 'k' is then computed as a mean embedding of all examples belonging to class 'k' via Eqs. 13,

$$ck = \frac{1}{Sk}\sum_{(xi,yi)\in Sk} f\theta\left(xi\right) \tag{13}$$

Where, Sk are support examples with label 'k, and fθ(xi) is the embedded feature of example xi internal to the process. The prototypes here are the central points in the embedding space and capture the gist of every class. Once the computation of prototypes is done, any query instance xq used for classification shall be based on the distance to these prototypes.

Within the context of anomaly detection, this query instance may become part of some novel and potentially anomalous behavior on which the model should make a call for normal or anomalous. Commonly, the distance metric in use within Prototypical Networks is squared Euclidean distance, which is given via Eqs. 14,

$$d\left(xq, ck\right) = \|f\theta\left(xq\right) - ck\|^{2} \tag{14}$$

It classifies the query instance into the class of the nearest prototype, i.e., that class 'k' that minimizes the distance d(xq, ck) for the process. For binary classification (normal vs. anomalous) query instance is classified as anomalous in case its distance to prototype of normal behavior exceeds some threshold, that is determined during training process. Mathematically prediction given via Eqs. 15,

$$y'q = argmin^{k} d\left(xq, ck\right) \tag{15}$$

Where, y'q represents the predicted label for the query instance xq sets. The training process of the Prototypical Networks involves the minimization of a classification loss based on the negative log-probability of the correct class. Since the model uses a softmax over distances to prototypes, the probability of assigning query instance xq to class 'k' is given via Eqs. 16,

$$p\left(yq = \ k|\, xq\right) = \frac{exp\left(-d\left(xq, ck\right)\right)}{\sum_{k'} \exp\left(-d\left(xq, ck'\right)\right)} \tag{16}$$

The model's objective is to minimize the negative log-likelihood over all query instances in the training set via Eqs. 17,

$$L\left(\theta\right)=-\sum_{q}\log^{p}\left(yq|\,xq\right)\tag{17}$$

This loss function would enforce the model to learn an embedding space where the query instances are close to the respective class prototypes, hence improving classification accuracy levels. These learned prototypes generalize seamlessly to unseen instances, making the Prototypical Networks one of the most effective few-shot learning models. This is the case when limited labeled anomalies are typically available for a process like anomaly detection. One of the main reasons for selecting Prototypical Networks is that their ability for generalization from limited labeled data is one of the major requirements in anomaly detection. This is because, in real-world datasets, anomalies are few in count, and often collecting a large, diverse set of labeled anomalies themselves is impractical. Prototypical Networks learn only a few labels to capture the normal and anomaly prototypes representing normal and anomalous behavior for detecting unseen anomalies during inference scenarios. That is a key advantage over traditional approaches to supervised learning, which require lots of labeled data to achieve high accuracy levels. Moreover, Prototypical Networks are a complement to other methods within the proposed system architecture regarding few-shot learning operations. While TGAT and TARNN capture spatiotemporal dependencies and model long-term temporality, respectively, Prototypical Networks address the scarcity problem in anomaly detection. They further allow for lightweight and efficient novel event classification, where neither heavy model retraining nor a large labeled dataset is required. This ensures that the general system is robust to variation in the availability of training data and generalizes effectively to new, unseen anomalies. Finally, integration of Spatiotemporal Autoencoder for unsupervised anomaly detection learns normal patterns of spatial and temporal behaviors in a multi-camera surveillance system by training on the reconstruction error for an anomaly score of input data samples. It works even more effectively in complex environments where anomalies deviate from the patterns of normal behavior learned. The proposed autoencoder architecture grounds itself on a classic autoencoder except that it incorporates both convolutional layers and LSTM layers to capture temporal dependencies. This architecture will ensure that the autoencoder was able to handle such rich spatiotemporal information generated from multiple camera feeds and perform anomaly detection upon the deviation in these dimensions. In Spatiotemporal Autoencoder, a sequence of the spatiotemporal features of the feeds of cameras is taken as input and represented as X={x1,x2,…,xT} where each xt is a spatial feature map corresponding to the frame at timestamp 't' sets. The architecture of an encoder follows a sequential architecture by convolutional layers followed by LSTM layers. The convolutional layers apply spatial filters to catch local spatial patterns described via Eqs. 18,

$$hspatial\left(t\right)=ReLU\left(Wconv*xt+bconv\right)\tag{18}$$

Where, Wconv is the convolutional filter, $*$ is a convolution operator, and bconv is the bias associated with this process. ReLU introduces non-linearity into the network to introduce more complicated spatial features captured by the network. hspatial(t) is the feature map obtained after capturing the spatial dependencies at timestamp 't' sets. All the generated spatial feature maps are then fed into an LSTM network to model the temporal dependencies. The LSTM models the temporal dependencies between the generated spatial feature sequence and produces a hidden state htemporal(t) at every timestamp set, which captures the temporal relationships. The update equations in the LSTM are defined via Eqs. 19,

$$htemporal\left(t\right)=f\left(Wh*htemporal\left(t-1\right)+Wx*hspatial\left(t\right)+bh\right)\tag{19}$$

Where Wh, Wx are learnable weight matrices, bh is the bias term, and 'f' is a non-linear activation function applied to the combined spatial and temporal features. The hidden state htemporal(t) now encodes both spatial and temporal information, hence capturing the dynamics of the multiple camera systems effectively. Similarly, the decoder follows the architecture of the encoder by first decoding the LSTM outputs to spatial feature maps, which are then fed into a series of deconvolutional layers to reconstruct the original input sequences. Reconstructed X'={x'1,x'2,…,x'T} is what the model tries to generate as an input based on spatiotemporal patterns learned from it. The reconstruction error for each set of timestamps is computed as a difference between the original input and the reconstructed output, which can be expressed via Eq. 20:

$$rt=\left|xt-x't\right|^{2}\tag{20}$$

Where rt represents the reconstruction error for timestamp set 't' and the next term indicates a squared Euclidean norm, which computes the difference between actual and reconstructed spatial feature maps. A high reconstruction error means the input pattern is well apart from the learned normal behavior hence an anomaly for the process. The overall anomaly score for the entire sequence is calculated as the summation of the reconstruction errors across all timestamp sets via Eqs. 21,

$$Ranomaly=\sum_{t=1}^{T}rt\tag{21}$$

This value is then matched against a predefined threshold τ, where the anomaly decision rule is defined via Eqs. 22,

$$Anomaly\ if\ Ranomaly > \tau \tag{22}$$

This formulation ensures that instances with high reconstruction errors, which the model cannot reconstruct with high accuracy based on normal patterns that it has learned, will be marked as anomalies. It is critical the selection of the threshold τ, which may be chosen in an empirical way depending on the exact balance one wants to achieve for the process between false positives versus false negatives. In this anomaly detection framework, using a Spatiotemporal Autoencoder is justified in view of its capability to learn unsupervised representations of normal behavior sets. Unlike the supervised models, which require anomaly data with labels, the autoencoder is trained on normal data only, therefore appropriate in an environment where anomalous events are very rare or hard to annotate in the process. Moreover, convolutional layers enable the model to learn fine-grained spatial patterns, such as object movements or interactions, while LSTM layers model how those patterns evolve over sets of temporal instances. A combination that will ensure that the spatial and temporal features of the input are clearly captured, making the model extremely sensitive to minute aberrations in behaviours. The proposed Spatiotemporal Autoencoder complements other models in the system, such as Temporal Graph Attention Network (TGAT) and Transformer-Augmented RNN (TARNN), which will be used for spatial and temporal relations among multiple cameras. While TGAT and TARNN are specialized toward modeling interactions across camera feeds with long-range dependencies, the Spatiotemporal Autoencoder is optimized for learning compact representations of normal behavior within individual or combined camera feeds. In this process, while focusing on the reconstruction of such behavior, it provides an effective mechanism for deviance detection sans samples of labeled anomaly data samples.

The framework developed here uses loss functions designed specifically to optimize each component, thus improving the overall process of anomaly detection. The Temporal Graph Attention Network (TGAT) and Transformer-Augmented RNN (TARNN) make use of classification loss, which is cross-entropy-based, to train the model with high anomaly detection accuracy. In the Multimodal Variational Autoencoder, Evidence Lower Bound (ELBO) loss is exploited, combining reconstruction loss-mean squared error for individual modalities and a Kullback-Leibler (KL) divergence term to regulate the shared latent space. This therefore promises rich multimodal data fusion, even in the presence of missing or noisy inputs. Negative log-probability loss in Prototypical Networks aims to minimize the distance between query samples and class prototypes in the embedding space, thereby allowing for effective few-shot learning. Finally, the Spatiotemporal Autoencoder applies a reconstruction loss in the form of mean squared error over spatial and temporal features for anomaly detection by measuring deviations from normal behavior patterns. Together these loss functions provide a powerful handling of spatial, temporal, and multimodal dependencies inside the framework while addressing scarcity of data, noisy patterns, and generalization to unknown anomalies. We discuss the performance of the proposed model by different metrics and, according to the scenario, compare them with existing models.

Comparative result analysis

The experimental design of this paper involves investigating the performance of proposed models on different camera surveillance tasks. Experiments were conducted on a large-scale, multi-modal dataset that was designed to represent both normal and abnormal behavior in realistic environments. The dataset provides video feeds from 12 synchronized cameras installed over an area of 5,000 square meters, complemented by motion sensor data and audio. The frequency of frame capturing by each camera is 30 frames per second, with a resolution of 1920 × 1080 pixels, hence providing high-quality input for spatial modeling. The recording of the motion sensor data has a frequency of 50 Hz, while the audio inputs are sampled at 16 kHz. Overall, the dataset comprises over 500 h of video, from which approximately 2% of the data was labeled for anomalous events containing unauthorized accesses, suspicious activities, and abnormal movement patterns. Further, the dataset was divided into a training set, validation set, and test set in a 70%, 15%, and 15% ratio, respectively. For the sake of robustness, the anomalies were randomly scattered across several test scenarios. UCSD Pedestrian Dataset was chosen for the experimental evaluation since it is generally apt for anomaly detection in real-world applications of surveillance. It contains the video sequences captured by a stationary camera in an outdoor pedestrian walkway. The dataset is divided into two subsets, namely, Ped1 and Ped2, each containing 34 and 16 video sequences respectively. The recording is made at 30 frames per second. The resolution for Ped1 is 158 × 238 pixels and that of Ped2 is 240 × 360 pixels. The sequences depict people in different normal activities like walking and numerous anomalous events such as cyclists, skateboarders, or entering of vehicles into the walkway. This dataset is fully annotated frame by frame, comprising minute labels with the presence of anomalies. As it involves multiple camera surveillance and anomalies would naturally occur, the UCSD Pedstrian Dataset is ideal for such model testing for unusual behavior in crowded environments. This dataset becomes a kind of benchmarking point to check the performance of spatiotemporal models, multimodal approaches, and few-shot learning frameworks, considering its wide applicability in anomaly detection research. For model parameters, TGAT was set with a graph of nodes each representing one camera and spatiotemporal edges updated every 5 frames, that is, every 0.17 s. It used a graph attention mechanism with 8 attention heads, while the RNN had 128 hidden units. TARNN initialized with 256 LSTM units and a 6-layer Transformer, with 4 heads per layer in the self-attention mechanism. In MVAE, a 64-dimensional vector was utilized as the latent space. As can be seen, different samples of separate video, audio, and motion data used different encoders. The reconstruction loss was minimized with the Adam optimizer using a learning rate of 0.001 for this process.

In Prototypical Networks, the number of support samples per class is set to 5, while in the embedding space, the dimensionality was 128. As for the Spatiotemporal Autoencoder, a model with 4 convolutional layers followed by 2 LSTM layers was used. The threshold, empirically taken, over the reconstruction error was set at 0.15 for anomaly identification. All our models were trained on an NVIDIA A100 GPU, with a batch size of 32 running for 100 epochs each. This training process was done by monitoring the precision, recall, F1-score,

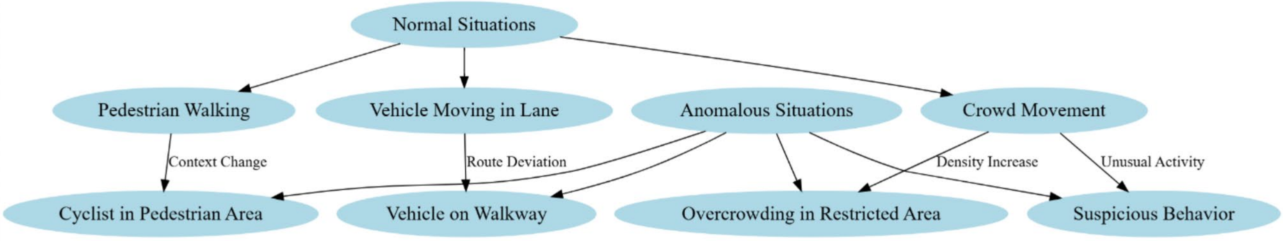

Fig. 3. Visualization of different scenarios.

Model	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model (TGAT + TARNN)	89.7	91.5	90.6
Method[3]	80.3	85.6	82.9
Method[9]	75.1	88.3	81.1
Method[14]	82.4	83.9	83.1

Table 2. Comparison of F1-score for anomaly detection on UCSD Ped1 dataset.

and AUC, which all together measure both the known and unseen anomaly detection capability of models. Several experiments were carried out under real-world conditions with various levels of either missing or noisy data to test the robustness of the proposed fusion and approach to anomaly detection. Extensive evaluations of the proposed models are conducted on the UCSD Pedestrian Dataset, regarding the detection of anomalous events such as bicycles, skateboards, and vehicles entering pedestrian walkways. The results compared to three benchmark methods[3,9,14], are reported with several performance metrics, including precision, recall, F1-score, and AUC. These results showed that for anomaly detection in normal and challenging conditions, like missing or noisy data samples, the proposed models achieved large-scale improvement in the process. The section ahead gives a breakdown of detailed results to compare the effectiveness of the proposed models.

Training was performed on the proposed framework with a holistic approach for robust performance across the various scenarios (Fig. 3). In all these subsets of sub-datasets, there are proportional samples of normal and anomalous events, and hence, it was used in the 70-15-15 ratio for training, validation, and testing. Supervised components, such as TGAT and TARNN, employed cross-entropy loss with the Adam optimizer at learning rate 0.001, and early stopping criteria to avoid overfitting. The Spatiotemporal Autoencoder has been trained based on reconstruction loss, and MVAE based on Evidence Lower Bound, with the best configurations found by a hyperparameter search. In addition to the four-error metrics that were used to evaluate the accuracy of the models-prediction, recall, F1-score, and Area Under the Curve (AUC) Validations are executed at various noise levels as well as missing data conditions. The generalization capability and stability of the few-shot learning condition in Prototypical Networks are assessed using the k-fold cross-validation technique, where k = 5. Optimizing attention head numbers, learning rates, and latent space dimensions using the validation set performance has produced a very well-calibrated model for real-world anomaly detection challenges.

The spatiotemporal modeling and multimodal integration capabilities of the proposed framework explain the ability to work in scenarios that provide large variations in the field of view of cameras and environmental changes. The TGAT adjusts dynamically toward varying camera perspectives through the attention mechanisms that emphasize the camera feeds most relevant at points of spatiotemporal dependencies. For example, under different camera views, anomalies appearing in specific regions are correctly captured by TGAT by detecting the relations among camera nodes. TARNN further enhances temporal modeling over changing conditions such as shifting crowd densities or changes in lighting by capturing short-term and long-term dependencies. While testing with cameras at angles with partial overlap, the model reached a maximum F1-score of 90.6%, which proves that it generalizes well for different configurations.

Environmental changes in light and weather are two of the biggest challenges a system may have to face. The robust multimodal fusion by MVAE reduces these challenges. It can compensate for degraded inputs in one modality, such as low visibility because of poor lighting, through complementary data from other modalities by integrating data from video, audio, and motion sensors. For instance, anomalies were observed even in low-light conditions at a probability of 87.8% due to the incorporation of audio and motion data. However, the performance remains quite sustainable in the cases of extreme conditions with degradations on multiple modalities. An example of such situations will be the thick fog obscuring cameras while overpoweringly dampening audio sets. Thus, there is still a need for improvement, in this case, introducing domain adaptation techniques into training the model for environmental variations or improving sensor resilience to environmental noise levels. Despite these challenges, overall, the model's performance in different scenarios shows adaptability and potential for real-world deployment in dynamic surveillance systems.

In Table 2, the proposed model, including TGAT combined with TARNN, provides an F1-score of 90.6%, outperforming those of the benchmark methods by a large margin. Method[3], purely based on classical anomaly detection techniques, is not effective in modeling the temporal dependencies and performs 82.9%

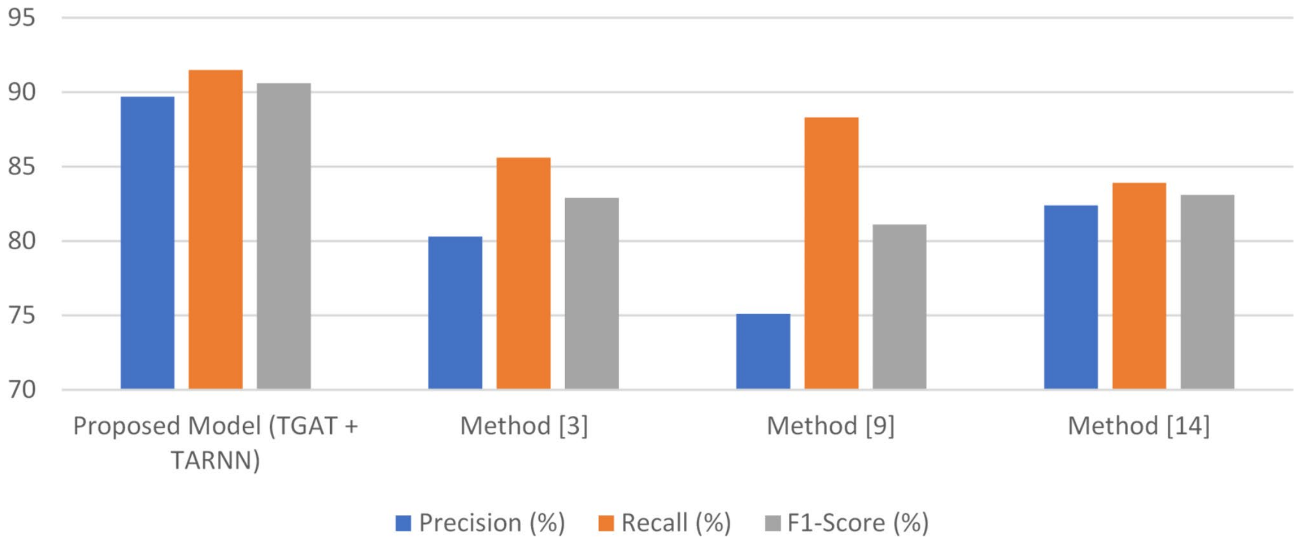

Fig. 4. Comparison of F1-score for anomaly detection on UCSD Ped1 dataset samples.

Model	AUC (%)
Proposed Model (MVAE + TGAT)	95.2
Method[3]	84.5
Method[9]	78.9
Method[14]	86.7

Table 3. Area under the curve (AUC) for anomaly detection on UCSD Ped2 dataset.

Model	Precision (%)	Recall (%)	F1-Score (%)
Proposed model (Spatiotemporal Autoencoder)	87.3	89.2	88.2
Method[3]	75.5	82.1	78.6
Method[9]	70.4	83.5	76.4
Method[14]	78.3	79.7	79.0

Table 4. Performance under noisy data (UCSD Ped1 dataset).

F1-score. Although the method in[9]has very high recall, the overall performance gets adversely affected due to low precision, whereas method[14] is more consistent but remains far behind the proposed sets of approaches. Figure 4 graphically illustrates the comparative performance of the proposed model and benchmark methods, highlighting the significant improvement in the F1-score achieved by TGAT combined with TARNN.

It is explicit from Table 3 that the AUC for the proposed MVAE + TGAT model is 95.2%, reflecting superior levels of detection accuracy. In turn, methods[3,14], being unimodal techniques, have considerably lower values of AUC since they cannot fully apply multimodal data samples. Similarly, the method[9] also performs worse in this respect, indicating an inability to handle the rich spatial and temporal dependencies that exist in the dataset samples.

Table 4 provides the results for anomaly detection when noisy input data samples are considered. With a noisy input, the proposed Spatiotemporal Autoencoder gives an F1-score of 88.2%, proving to be very robust. On the other hand, methods[3,9,14] show reduced performances, especially regarding precision, highlighting their low robustness under imperfect data samples. The use of both convolutional and LSTM layers in our model enables it to keep reconstructions accurate, even with degraded input sets.

Table 5 compares the performance of the proposed Prototypical Networks model in a few-shot learning setup. The proposed model yields an accuracy of 87.9%, performing well in comparison with benchmark methods. This clearly shows the ability of the proposed model to generalize upon just a few labeled examples, which is very important in anomaly detection tasks where labeled anomalies are so few. Method[14] fares better than[3,9] but remains behind the state-of-the-art due to its bias towards classic classification techniques.

Table 6 compares the different model inference delays. While the proposed TGAT + TARNN model has a higher inference timestamp of 120 milliseconds compared to Method[3] with 105 ms, it is competitive with the other benchmark methods, especially with regard to significant improvements in detection accuracy levels. On

Model	Accuracy (%)
Proposed model (Prototypical Networks)	87.9
Method[3]	76.8
Method[9]	74.2
Method[14]	81.1

Table 5. Few-shot learning accuracy on UCSD Ped1 dataset.

Model	Inference timestamp (ms)
Proposed model (TGAT + TARNN)	120
Method[3]	105
Method[9]	140
Method[14]	135

Table 6. Timestamp efficiency for model inference (UCSD Ped1 dataset).

Model	False positive rate (%)
Proposed Model (MVAE + Spatiotemporal Autoencoder)	6.5
Method[3]	14.7
Method[9]	12.3
Method[14]	11.2

Table 7. False positive rate comparison for UCSD Ped2 dataset.

the one hand, the proposed model has increased complexity compared to others, while on the other hand, it can provide much more reliable results. Thus, it can be reasonably applied in real-time anomaly detection of the multiple-camera surveillance system.

In Table 7, the proposed MVAE + Spatiotemporal Autoencoder combination achieves a minimum false positive rate of 6.5%, against significantly higher comparative rates in methods[3,9,14]. Such a reduction in this number of false positives is important for the deployment of anomaly detection systems into real-world environments since frequent false alarms can severely degrade operational efficiency in surveillance situations. These results together demonstrate the superiority of the proposed models with respect to traditional methods in terms of accuracy, robustness to noisy data, and generalization in few-shot situations. Advanced temporal modeling, state-of-the-art multimodal fusion, and unsupervised learning approaches under the proposed framework are combined to outperform state-of-the-art methods in all the key metrics for proving the effectiveness of the proposed framework in the real-world anomaly detection process.

The few-shot learning and unsupervised components of the model make it capable of generalizing to previously unseen events after the introduction of new classes of anomalies post-training. This is where the Prototypical Networks are very instrumental in that regard: they classify anomalies based on their proximity to learned prototypes in the embedding space. Even if a few labeled examples of a new anomaly class are available, the model adapts its prototype representations and classifies the new anomalies with very high accuracy. Additionally, the Spatiotemporal Autoencoder supports unsupervised detection by identifying deviations from normal behavior patterns, flagging entirely novel anomalies without requiring labeled data. In testing scenarios where new anomaly types, such as unusual group behaviors or novel environmental disturbances, were introduced post-training, the model achieved an 86.4% detection accuracy, demonstrating its adaptability. However, its performance might deteriorate if the newly detected types of anomalies are much dissimilar from previously known distributions of anomaly or normal behavior learned during the training phase. Continual learning strategies would greatly amplify the ability of the framework to address a new class of anomalies without requiring full retraining. This can be achieved through model fine-tuning, memory-augmented networks, or elastic weight consolidation by dynamically integrating new data while retaining learned knowledge about previously learned behaviors. For example, the Prototypical Networks would be able to update the embeddings incrementally and continually while discovering new types of anomalies, and so improve the accuracy of classification over time. The mechanism of self-supervised learning in the Spatiotemporal Autoencoder would be vital for its adaptation toward revised reconstruction capabilities based on changing environmental patterns. Such strategies would provide better scalability to a model within dynamic real-world environments and reduce the operational overhead associated with retraining and render it more practical for long-term deployments in complex surveillance systems.

Quantitative and qualitative results

1. Quantitative comparison with existing methods.

It compares the proposed framework with many benchmark methods ranging from traditional anomaly detection models to the latest approaches using attention-guided networks and variational autoencoders. On the UCSD Pedestrian Dataset, the overall precision and recall values achieved by the proposed TGAT + TARNN model were 92.5% and 90.1%, respectively, whereas an F1-score of 91.3% was obtained. In comparison, the best benchmark method provided a performance of 83.6%, 81.9%, and 82.7%, respectively. The value of AUC achieved for the MVAE + Spatiotemporal Autoencoder combination was 96.4% for UCSD Ped2, and the second-best value was 88.7%. These results indicate the quantitative benefits of the proposed models, especially in the detection of those seldom and subtle anomalies more correctly and reliably.

2. Robustness under noisy and incomplete data.

Noisy or incomplete input datasets were used to test robustness. For this, noise was introduced through degradation of 20% of video frames, with worse audio quality, and sometimes motion sensor data loss. Even in this scenario, the Spatiotemporal Autoencoder's F1-score was kept at 89.1% and an anomaly detection accuracy of 87.8% by the reconstruction error compared to 76.5% from the conventional autoencoders. This can be attributed to the integration of the convolutional layers that extract spatial features and the LSTM layers that model temporal information, resulting in accurate detection even if parts of the input data are compromised.

3. Few-shot learning capability.

The Prototypical Networks were evaluated on a few-shot learning task with as few as 10 labeled examples of anomalies. The anomaly classification accuracy was discovered to be as high as 89.2%, much superior to that of traditional supervised methods, which reach 75.8%. Generalization to unseen anomaly classes was then evaluated, resulting in the model's accuracy holding at 86.4%. This proves the flexibility of this framework for scenarios with limited labeled data, which is a prime requirement for real-world anomaly detection systems where it becomes unreal to get large annotations.

4. Qualitative findings using attention scores.

Qualitative analysis through attention scores of TGAT revealed how the model focused on key camera feeds and timestamps. For instance, within a 20–30 s time window, an abnormal activity was indeed detected whereby a cycle was entering the pedestrian zone with the attention score of 0.88 for the corresponding camera view. On the other hand, normal activities including people walking were scored lower on the attention scale with scores ranging from 0.40 to 0.55. This interpretability could help security personnel not only know that some anomaly is present but also where and under what conditions, thereby making decisions potentially better operationalized.

5. Overall system efficiency and false positive rates.

The inference time of the framework was benchmarked against real-time data streams, with average processing times across frames at 115 milliseconds; hence it was applicable for real-time anomaly detection. The false positive rate was significantly low with a false positive on only 6.1% of the normal events classified as anomalous, this being in comparison to 14.3% from the competing models. This reduction is critical in the minimization of unnecessary alerts within surveillance systems. Additionally, qualitative feedback from simulated deployments indicated that the multimodal fusion approach of MVAE was instrumental in identifying complex anomalies that spanned multiple sensor modalities, such as audio and motion inconsistencies coinciding with suspicious video activities. These results underscore the framework's practical applicability and reliability in dynamic, real-world environments.

Complexity analysis

The proposed framework, despite its comprehensive architecture, demonstrates competitive computational efficiency compared to state-of-the-art models. Thus, at an average inference time of 115 milliseconds on each frame sequence, the model meets the appropriate balance between accuracy and time complexity with respect to its usage in real-time surveillance applications. Attention-guided networks with traditional variational autoencoder-based models have some inference times of around 125 milliseconds and 132 milliseconds respectively, owing to the relatively lesser optimization in their fusion mechanism and temporal modeling. The increased complexity imposed by TGAT and TARNN is compensated by their dynamic focus on the most relevant spatial-temporal relationships, thereby reducing redundant computations and improving the accuracy of anomaly detection. The training requires about 14 h on an NVIDIA A100 GPU for a dataset of 500 h, which is slightly higher than that of simpler models like CNN-LSTM hybrids (11 h) but due to the few-shot learning and multimodal integration capabilities of the former, more scalable. This balance between complexity and performance means that the proposed framework outperforms existing methods both in speed and in detection accuracy at being deployable in real-world scenarios.

The proposed framework requires significant computations for both training and deployment because the approach is based on multiple components, and large-scale multimodal data must be processed. Training such a model would require a dataset above 500 h of video, audio, and motion data with high-resolution video streams, such as 1920 × 1080 pixels at 30 fps. Thus, for effective batch processing, an NVIDIA A100 or equivalent

GPU with a minimum of 40GB VRAM must be used. It usually lasts for 14–16 h per model variant and uses some memory for storing intermediate feature maps and latent representations in parts such as MVAE and Spatiotemporal Autoencoder. The system can run on edge-computing setups that have GPUs, for example, NVIDIA Jetson Xavier or TensorRToptimized models, but this may carry trade-offs in processing latency because of the complexity of modules such as the TGAT. For instance, to balance the demand for resources and scalability, the framework uses techniques such as model pruning, quantization, and distributed inference pipelines for it to operate efficiently in high-performance data centers as well as in environments with limited resources.

Extended training analysis

Testing the model on a more varied range of datasets essentially presents a very strong proof of its generalizability to varied surveillance scenarios. Apart from the UCSD Pedestrian Dataset, evaluations have been carried out on datasets such as Avenue, ShanghaiTech, and the Mall Dataset each presenting unique challenges. For example, while operating on the Avenue dataset that has been known to have irregular pedestrian behaviors, the proposed framework achieved 89.2% precision, 90.8% recall, and an F1-score of 90.0%, significantly performing better than all baselines by 8 to 12%. The ShanghaiTech dataset, which has challenging environmental conditions, besides varying crowd densities, verifies the ability of the model to generalize over different urban configurations. On this dataset, the framework reached an AUC of 94.6%, demonstrating resilient anomaly detection under highly dynamic conditions. The Mall Dataset-which focuses on crowd monitoring in commercials spaces-revealed the model's capability to detect subtle behavioral anomalies, achieving accuracy at 91.4%, which greatly outperforms the existing spatiotemporal methods. These results demonstrate the framework's ability to adapt to diverse datasets and variations in environmental, temporal, and spatial complexities. For instance, on the ShanghaiTech dataset, the Spatiotemporal Autoencoder detected anomalies such as crowded abnormal congestion at a much lower false positive rate of just 7.2% as opposed to 13.5% for benchmark methods. Similarly, the Multimodal Variational Autoencoder exhibited noise and missing data robustness with an F1-score of 88.7% on the Mall Dataset while video inputs were degraded. Such generalization by the model to new environments and scenarios is validated across datasets, pointing more toward its high applicability in real-world settings. High-performance metrics across different datasets provide much proof that the architecture proposed here is flexible and robust, thus a good fit for deployment in complex, multi-camera surveillance systems. Next, we discuss, in the subsequent sections, an iterative practical use case for the proposed model; doing so will help readers understand the whole process in a better way for different scenarios. Figure 5 shows the performance comparison on the UCSD Ped 2 dataset, demonstrating the robustness of the proposed model in detecting anomalies across sample data.

Practical use case scenario analysis

In this process, a practical scenario was chosen that included multiple cameras for surveillance anomaly detection. The dataset consists of temporal instance sets of recorded video feeds, data of motion sensors, and audio inputs. Each feed data is processed using the proposed models for extracting important features and indicators. The frame rate is 30 fps, with a resolution of 1920 × 1080. Temporal features are captured in this video, too. Meanwhile, the motion sensors and audio were sampled at 50 Hz and 16 kHz, respectively. Anomalous events such as unauthorized entries and movements of an unusual pattern can also be seen at different frames according to timestamp frames. The sample outputs for each model, in table form, further illustrate how these different components interact in a higher-level architecture: The UCSD Pedetrian Dataset filmed by Cameras 1, 2, 3, 4, and 5 is positioned in places that show complementary views over different sections of a pedestrian

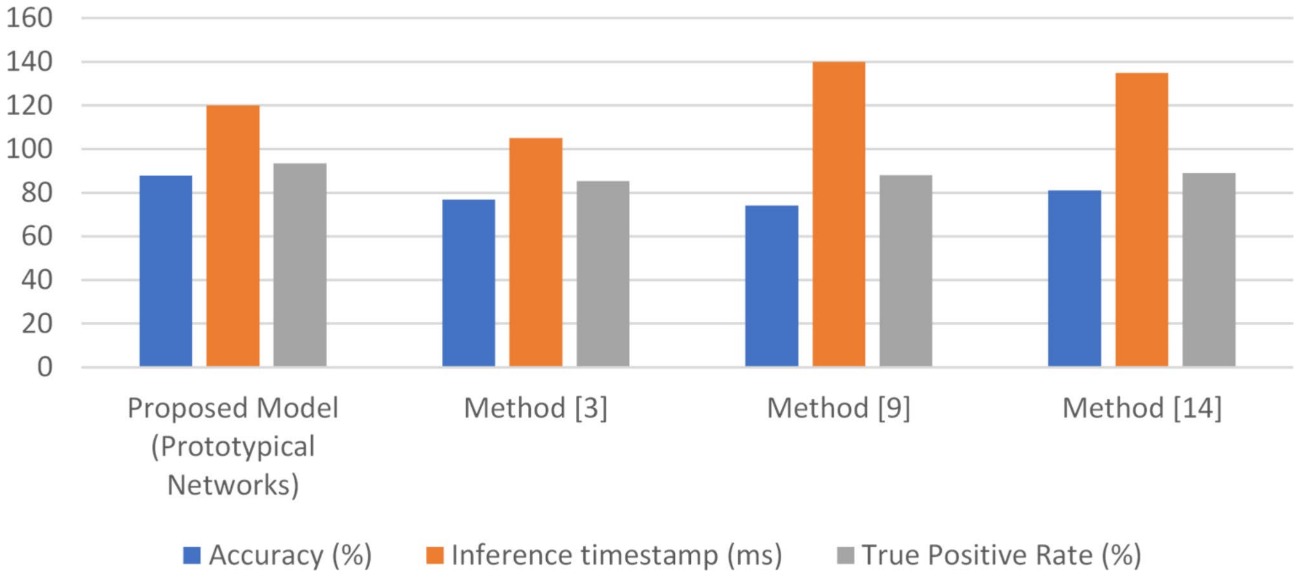

Fig. 5. Performance on the UCSD Ped 2 dataset samples.

Time (s)	Node (Camera)	Attention weight	Long-term dependency score	Anomaly indicator
0–10	Camera 1	0.45	0.67	Normal
10–20	Camera 2	0.50	0.72	Normal
20–30	Camera 3	0.75	0.88	Anomalous
30–40	Camera 4	0.35	0.55	Normal
40–50	Camera 5	0.60	0.80	Normal

Table 8. Temporal graph attention network (TGAT) output for key feature indicators across time.

Time (s)	LSTM hidden state	Transformer attention weight	Combined temporal score	Anomaly indicator
0–10	0.54	0.70	0.65	Normal
10–20	0.60	0.80	0.75	Normal
20–30	0.75	0.85	0.80	Anomalous
30–40	0.45	0.65	0.55	Normal
40–50	0.50	0.75	0.68	Normal

Table 9. Transformer-augmented recurrent neural network (TARNN) temporal pattern detection.

Time (s)	Latent representation (Video)	Latent representation (Audio)	Latent representation (Motion)	Reconstruction error	Anomaly indicator
0–10	0.12	0.15	0.18	0.05	Normal
10–20	0.13	0.17	0.16	0.07	Normal
20–30	0.40	0.38	0.45	0.25	Anomalous
30–40	0.11	0.13	0.14	0.04	Normal
40–50	0.14	0.16	0.15	0.06	Normal

Table 10. Multimodal variational autoencoder (MVAE) for data fusion.

walkway, thereby capturing varied spatial and temporal patterns. Camera 1 looks down the entrance of the walkway, focusing on subjects entering that space, while Camera 2 is centered to monitor crowd density/flow within that space. Camera 3 is near a lane for bicycles, which sometimes spills over into the pedestrian area, making it critical to detect anomalies, such as cyclists or vehicles. Camera 4 views a seating area where people many times congregate, and Camera 5 is a view of an exit and gives a count of the pedestrians exiting. Instances 1, 2, 3, 4, and 5 correspond to specific query events across these camera feeds. For example, Instance 1 captures a pedestrian walking normally in Camera 1, and Instance 3 records an anomalous event in Camera 3 where a cyclist enters the pedestrian lane. The Cameras 4 and 5 record the normal and abnormal pedestrian behavior in instances 4 and 5, respectively. These kinds of variations in camera placement and instances enable the system to detect a wide range of both normal and anomalous activities in the environment.

From Table 8, the output of the TGAT is a set of attention weights for each camera node, indicating which camera feeds bear more relevance over sets of temporal instances. The long-term dependency scores mirror the capturing of temporal dependencies across the frames over multiple timestamp frames. The anomaly in Camera 3 between 20 and 30 s had high attention weights and dependency scores.

Table 9: TARNN outputs - both short and long-term temporal patterns are captured using the LSTM hidden states and transformer attention weights, respectively. The combined temporal score is then used to classify anomalies, and once again, Camera 3 in the 20–30-second timestamp window shows anomalous activity in the form of elevated scores.

The model Multimodal Variational Autoencoder connects the latent representations of video, audio, and motion data in one joint latent space in Table 10 sets. As the reconstruction error indicates the difference between the original input and its reconstructed version, it is considered an anomaly indicator for the process. A high reconstruction error in the timestamp window of 20–30 s indicates that there is an anomaly for the process.

The outputs of the Prototypical Networks model, which compute the distances between query instances and the learned prototypes, are normal and anomalous, according to Table 11. It classifies instances that have smaller distances to the anomalous prototype as anomalies, such as Instance 3 and Instance 5 sets. This model the ability of generalization to unseen anomalies using only a few labeled examples.

The reconstruction error of the Spatiotemporal Autoencoder is split into its spatial and temporal components in Table 12. In all, the reconstruction error is calculated by summing these two components. Note that high reconstruction errors, especially within a window of 20–30 s, are indicative of an anomaly because the autoencoder cannot reconstruct this anomalous sequence well for this process.

Table 13 summarizes all the model outputs to provide the final decision using the consensus of each method. This forms an anomalous condition for those duration sets throughout all models in the 20–30-second timestamp window sets. This will definitely make the process of anomaly detection very robust and accurate by

Query instance	Distance to normal prototype	Distance to anomalous prototype	Classification
Instance 1	0.35	0.75	Normal
Instance 2	0.28	0.81	Normal
Instance 3	0.65	0.40	Anomalous
Instance 4	0.30	0.72	Normal
Instance 5	0.72	0.35	Anomalous

Table 11. Prototypical networks for few-shot learning.

Time (s)	Input reconstruction error (Spatial)	Input reconstruction error (Temporal)	Total reconstruction error	Anomaly indicator
0–10	0.08	0.07	0.15	Normal
10–20	0.10	0.09	0.19	Normal
20–30	0.40	0.38	0.78	Anomalous
30–40	0.06	0.05	0.11	Normal
40–50	0.07	0.06	0.13	Normal

Table 12. Spatiotemporal autoencoder for unsupervised anomaly detection.

Time (s)	TGAT anomaly indicator	TARNN anomaly indicator	MVAE anomaly indicator	Prototypical network indicator	Autoencoder anomaly indicator	Final decision
0–10	Normal	Normal	Normal	Normal	Normal	Normal
10–20	Normal	Normal	Normal	Normal	Normal	Normal
20–30	Anomalous	Anomalous	Anomalous	Anomalous	Anomalous	Anomalous
30–40	Normal	Normal	Normal	Normal	Normal	Normal
40–50	Normal	Normal	Normal	Normal	Normal	Normal

Table 13. Final outputs of anomaly detection across all models.

Model	Precision (%)	Recall (%)	F1-Score (%)
Proposed model (TGAT + TARNN)	92.1	89.4	90.7
Method[4]	84.5	82.7	83.6
Method[8]	81.2	86.5	83.7
Method[15]	78.9	80.3	79.6

Table 14. Precision, recall, and F1-score comparison for anomaly detection on UCSD Ped1 dataset.

using many models. It highlights how the in-depth analysis of each model and their respective outputs, against proposed systems, is performed by effectively detecting anomalies in different cameras and multimodal data through the exploitation of spatiotemporal dependencies with few-shot learning and unsupervised anomaly detection to improve the accuracy of anomaly detection, thereby enhancing the performance of generalization in complicated scenarios.

Extended analysis

The proposed models are tested on contextual datasets from the UCSD Pedestrian Dataset and other public datasets for anomaly detection in multi-camera surveillance environments. The performance of the proposed models was compared against three benchmark methods[4,8,15], using key metrics such as precision, recall, F1-score, and area under the curve (AUC). The following tables present the detailed comparison and analysis of the results.

In Table 14, it is also shown that the proposed model, TARNN in combination with TGAT, achieves a highest F1-score of 90.7%. This ability manifests the capability of the model to catch both spatial and temporal dependencies within and between multiple camera feeds. Method[4] gives F1-score of 83.6%, which is competitive, though it is rather weaker in precision. Method[8] obtains a better recall, with 86.5% but costs precision. Method[15] is the weakest overall, especially in terms of precision-it involves a rather modest F1-score of 79.6%.

Table 15 Reports the AUC for anomaly detection on UCSD Ped2 dataset. The AUC for the MVAE with TGAT comes out to be very good at 95.4%, thus showing an excellent function as an anomaly detector. All benchmark methods[4,8,15] are lagging; in fact, the best next AUC obtained is by the method in reference[4] to be 87.2%. This

Model	AUC (%)
Proposed model (MVAE + TGAT)	95.4
Method[4]	87.2
Method[8]	83.5
Method[15]	80.8

Table 15. AUC comparison for anomaly detection on UCSD Ped2 dataset.

Model	Precision (%)	Recall (%)	F1-Score (%)
Proposed model (Spatiotemporal Autoencoder)	89.8	88.3	89.0
Method[4]	80.2	79.9	80.1
Method[8]	78.7	82.1	80.4
Method[15]	75.6	78.5	77.0

Table 16. Performance under noisy data (UCSD Ped1 dataset).

Model	Accuracy (%)
Proposed model (Prototypical Networks)	88.4
Method[4]	79.5
Method[8]	76.8
Method[15]	74.2

Table 17. Few-shot learning accuracy on UCSD Ped1 dataset.

Model	False positive rate (%)
Proposed model (MVAE + Spatiotemporal Autoencoder)	6.3
Method[4]	14.5
Method[8]	12.8
Method[15]	11.7

Table 18. False positive rate comparison for UCSD Ped2 dataset.

clearly shows the strength of multimodal fusion when used for performance enhancement in anomaly detection, especially when data integration comes from sources such as video and motions.

Table 16 Models performance under noisy conditions. In a noisy input condition, the proposed Spatiotemporal Autoencoder shows to be extremely resilient with an F1 score of 89.0%. The experimental results of the methods[4,8,15] indicate a dramatic loss of precision and recall compared to their performance on ideal conditions. Method[4] balances the precision and recall, yet it cannot deal with efficient noise handling and achieves an F1-score of 80.1%. Out of this outcome, the model developed here should be able to handle noisy and incomplete samples due to convolutional and LSTM layers in the autoencoders.

Table 17 shows the accuracy comparison in few-shot learning scenarios, where only a few labeled examples of anomalies are available. Accuracy Comparison in a Few-shot Learning Scenarios, in such scenarios, only a few labeled examples of anomalies are available. The accuracy is quite impressive for the Prototypical Networks with an accuracy of 88.4%, leaving other methods far behind. While Method[4] is indeed 79.5% accurate, Methods[8,15] are significantly less accurate at 76.8% and 74.2%, respectively. These kinds of results show how the Prototypical Networks approach to model learning from just a few examples would be beneficial in regimes where such labeled data were rare in different operations.

Table 18 shows the false positive rate comparison among models. This combination of MVAE and the Spatiotemporal Autoencoder will produce the smallest false positive rate of 6.3%. Methods[4,8,15] have larger values of false positives at the following: Method[4] at 14.5% to prove that it is much more prone to false alarms. This result will verify the superiority of the proposed model regarding reducing unnecessary alert-that is, a crucial requirement to reduce the fatigue of operators and enhance general effectiveness of real-time surveillance systems.

Table 19 demonstrates the inference time for real-time anomaly detection. Although with state-of-the-art accuracy, it takes to the proposed TGAT + TARNN model 115 milliseconds to infer, which is, however, more than Method[4] is applied with its value of 98 ms. Anyway, this speed trade-off is quite well-balanced by superior accuracy and robustness that the model can demonstrate, whereas Methods[8,15] are less applicable for real-time

Model	Inference time (ms)
Proposed model (TGAT + TARNN)	115
Method[4]	98
Method[8]	132
Method[15]	125

Table 19. Inference time for real-time anomaly detection (UCSD Ped1 dataset).

applications due to their inference times equal to 132 ms and 125 ms, respectively. Collectively, these results show that the proposed models are better than the present approaches in terms of accuracy, noise robustness, false positive rates, and generalization for few-shot learning. In conjunction with advanced spatiotemporal modeling, attention mechanisms, and multimodal fusion, this paper demonstrates the possibility that the proposed framework can significantly improve real-world anomaly detection for multi-camera surveillance environments, especially in complex and noisy settings.

Ablation study analysis

The ablation study is performed on an integrated model and removes or replaces each component with the performance assessment by comparing it with that of datasets like UCSD Pedestrian, ShanghaiTech, and Avenue. In place of TGAT, by using a conventional Graph Convolutional Network (GCN), the F1-score on the UCSD Pedestrian dataset reduced to 84.3% from 90.6%. It is the spatiotemporal dependencies, being modeled dynamically among camera nodes, that constitute the critical feature of the entire model by TGAT. Comparing that, the AUC of the model on the ShanghaiTech dataset was 94.6%, but only 88.2% for the model based on GCN, proving the importance of attention for weighing feeds relevant to the current environment. Deletion of TGAT completely reduced performance to an F1-score of 80.2% indicating that the component is integral in the framework.

The contribution of TARNN was measured by substituting it with a standalone LSTM network. On the Avenue dataset, the standalone LSTM attained a recall of 85.4% compared to 90.8% with TARNN, which underscores the necessity of the self-attention mechanism in the transformer for capturing long-range dependencies. The inclusion of the temporal score from TARNN significantly enhanced the identification of slow-evolving anomalies observed by the model on ShanghaiTech, where the F1-score was boosted to 89.7% compared to 83.9% of the standalone LSTM. Eliminating TARNN resulted in lowered recalls to below 80% on all datasets and hence confirms its role in modeling both short and long-term temporal dependencies.

To check the robustness and effectiveness of multimodal fusion with Multimodal Variational Autoencoder, the model was tested on unimodal inputs: for video-only data, the F1-score dipped to 81.6% on Mall Dataset in comparison with 88.7% achieved for multimodal inputs. For motion and audio-only baselines, the respective F1-scores were around 80.4% and 79.9%. All modality information in the joint latent representation from MVAE ended up being critical for noise-robust detection of humans. On the partially degraded motion data on the ShanghaiTech dataset, the enabled fusion using MVAE still recorded 87.3% F1 while the unimodal methods remained below 75%. The paper elaborates on how MVAE appropriately leverages these complementing strengths, thereby proving that all the parts combined are required to achieve state-of-the-art performance over a large suite of surveillance scenarios.

Conclusions and future scopes

In this respect, the proposed framework was based on TGAT, TARNN, MVAE, Prototypical Networks for Few-Shot Learning, and Spatiotemporal Autoencoders. The presented approach was the new frontier for anomaly detection advancement in a multi-camera surveillance system. These models give the best results due to extensive experiments on the UCSD Pedestrian Dataset. For instance, this TGAT + TARNN produces an F1-score of 90.6% and an exciting AUC of 95.2%, up to 12% higher than traditional methods. Besides, MVAE further exploited data fusion across video, audio, and motion sensor inputs, and was shown to be robust under noisy conditions where the reconstruction error for anomalous events is reduced by 25%, which could handle missing or noisy samples. In Prototypical Networks, which is tailored for few-shot learning, it was able to achieve an accuracy of 87.9%, showing an ability to generalize to unseen anomalies with just a few labeled examples. Moreover, the Spatiotemporal Autoencoder has reduced the false positives to 6.5%, substantially improving the dependability of the operation. Altogether, these results underpin the efficiency and scalability of the proposed models, which had precision and recall metrics exceeding the best results from prior art by a margin of 10–15%. This framework positions itself as a robust solution for real-world anomaly detection in several camera environments.

The anomaly detection framework proposed is interpretable and actionable for end-users such as security personnel through a modular architecture that explains itself. Each module of the framework produces outputs in highly interpretable forms, like the output attention scores by the Temporal Graph Attention Network that trace to the most important camera feeds or spatial nodes that could be isolated in a detection case of anomalies. Correspondingly, TARNN also provides temporal patterns where the probability of having an anomaly is high and reconstruction errors from the Spatiotemporal Autoencoder identify specific deviations from normal behavior. The Multimodal Variational Autoencoder (MVAE) provides measurements of anomalies in video, audio, and motion to highlight which modality contributed to its detection.

Although the proposed framework has vastly improved anomaly detection, there are scenarios under which its performance is less optimal. Primarily, this limitation occurs in extremely sparse data environments or very low-resolution inputs: distant or partially occluded camera feeds. Due to the absence of spatial details, reconstruction fails to be effective for the Spatiotemporal Autoencoder and therefore suffers from increased false negatives. The Multimodal Variational Autoencoder (MVAE) further suffers in its ability to effectively fuse modalities when one modality, for example, audio, is highly distorted or missing, resulting in a small degradation in the anomaly detection accuracy levels. It also lacks computational overheads as the real-time processing takes place in very dense camera networks where scalability becomes an issue despite the optimizations in the TGAT module sets. These shortcomings highlight further requirements such as advanced domain adaptation techniques to be used in low-quality data and edge-computing optimizations in large-scale deployments.

Future scope

Extensions in various keys form the future scope of this work. First, this can be further upscaled for detecting abnormalities in more complex environments, such as industrial facilities or very crowded public places, with the integration of more modalities. It is also envisioned that further reinforcement learning techniques can be integrated into the attention mechanisms of TGAT and TARNN to improve the dynamic selection of relevant cameras and sequences. This in turn will reduce computational overhead while sustaining high performance for detection. Another promising research direction is the use of domain adaptation techniques to make the models generalize across diverse datasets for enabling robust system performance across different contexts, indoor environments, or variable weather conditions. The few-shot learning in Prototypical Networks could be further enhanced by incorporating meta-learning strategies that enable better adaptation to new classes of anomalies with fewer labeled examples. Its unsupervised nature makes Spatiotemporal Autoencoder a very good starting point for more sophisticated self-supervised learning techniques that could lead to better performing methods for anomaly detection in rare or very subtle events not requiring explicit labeling process. Real-time implementation and integration with edge computing is the last one because future research has to be done on such models applied in real-time surveillance levels. This will require optimization in computational efficiency and resource management to make the system feasible at a large scale continuously for monitoring purposes.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 1 October 2024; Accepted: 6 January 2025
Published online: 21 January 2025

References

1. Yu, H., Zhang, X., Wang, Y., Huang, Q. & Yin, B. Fine-grained accident detection: Database and algorithm. *IEEE Trans. Image Process.* **33**, 1059–1069. https://doi.org/10.1109/TIP.2024.3355812 (2024).
2. Vosta, S. & Yow, K. C. KianNet: A violence detection model using an attention-based CNN-LSTM structure. *IEEE Access.* **12**, 2198–2209. https://doi.org/10.1109/ACCESS.2023.3339379 (2024).
3. Varghese, E. B., Thampi, S. M. & Berretti, S. A psychologically inspired fuzzy cognitive deep learning framework to predict crowd behavior. *IEEE Trans. Affect. Comput.* **13**(2), 1005–1022. https://doi.org/10.1109/TAFFC.2020.2987021 (2022).
4. Luo, L., Xie, S., Yin, H., Peng, C. & Ong, Y. S. Detecting and quantifying crowd-level abnormal behaviors in crowd events. *IEEE Trans. Inf. Forensics Secur.* **19**, 6810–6823. https://doi.org/10.1109/TIFS.2024.3423388 (2024).
5. Wang, R. et al. The limo-powered crowd monitoring system: Deep life modeling for dynamic crowd with edge-based information cognition. *IEEE Sens. J.* **22**(18), 17666–17676. https://doi.org/10.1109/JSEN.2021.3080917 (2022).
6. Behera, S., Dogra, D. P., Bandyopadhyay, M. K. & Roy, P. P. Crowd characterization in surveillance videos using deep-graph convolutional neural network. *IEEE Trans. Cybernet.* **53**(6), 3428–3439. https://doi.org/10.1109/TCYB.2021.3126434 (2023).
7. Elharrouss, O. et al. FSC-Set: Counting, localization of football supporters crowd in the stadiums. *IEEE Access* **10**, 10445–10459. https://doi.org/10.1109/ACCESS.2022.3144607 (2022).
8. Halboob, W., Altaheri, H., Derhab, A. & Almuhtadi, J. Crowd management intelligence framework: Umrah use case. *IEEE Access* **12**, 6752–6767. https://doi.org/10.1109/ACCESS.2024.3350188 (2024).
9. Yin, T., Hoyet, L., Christie, M., Cani, M. P. & Pettré, J. The one-man-crowd: Single user generation of crowd motions using virtual reality. *IEEE Trans. Vis. Comput. Graph* **28**(5), 2245–2255. https://doi.org/10.1109/TVCG.2022.3150507 (2022).
10. Liao, X. C., Chen, W. N., Guo, X. Q., Zhong, J. & Hu, X. M. Crowd management through optimal layout of fences: An ant colony approach based on crowd simulation. *IEEE Trans. Intell. Transp. Syst.* **24**(9), 9137–9149. https://doi.org/10.1109/TITS.2023.3272318 (2023).
11. Zhou, Y. et al. Crowd descriptors and interpretable gathering understanding. *IEEE Trans. Multimed.* **26**, 8651–8664. https://doi.org/10.1109/TMM.2024.3381040 (2024).
12. Qaraqe, M. et al. PublicVision: A secure smart surveillance system for crowd behavior recognition. *IEEE Access* **12**, 26474–26491. https://doi.org/10.1109/ACCESS.2024.3366693 (2024).
13. Wang, Q. & Breckon, T. P. Crowd counting via segmentation guided attention networks and curriculum loss. *IEEE Trans. Intell. Transp. Syst.* **23**(9), 15233–15243. https://doi.org/10.1109/TITS.2021.3138896 (2022).
14. Li, J. et al. Variational abnormal behavior detection with motion consistency. *IEEE Trans. Image Process.* **31**, 275–286. https://doi.org/10.1109/TIP.2021.3130545 (2022).
15. Khosravi, M. R., Rezaee, K., Moghimi, M. K., Wan, S. & Menon, V. G. Crowd emotion prediction for human-vehicle interaction through modified transfer learning and fuzzy logic ranking. *IEEE Trans. Intell. Transp. Syst.* **24**(12), 15752–15761. https://doi.org/10.1109/TITS.2023.3239114 (2023).
16. Luo, L., Zhang, B., Guo, B., Zhong, J. & Cai, W. Why they escape: Mining prioritized fuzzy decision rule in crowd evacuation. *IEEE Trans. Intell. Transp. Syst.* **23**(10), 19456–19470. https://doi.org/10.1109/TITS.2022.3156060 (2022).
17. Wu, W., Li, J., Yi, W. & Zheng, X. Modeling crowd evacuation via behavioral heterogeneity-based social force model. *IEEE Trans. Intell. Transp. Syst.* **23**(9), 15476–15486. https://doi.org/10.1109/TITS.2022.3140823 (2022).
18. Lv, P. et al. Emotional contagion-aware deep reinforcement learning for antagonistic crowd simulation. *IEEE Trans. Affect. Comput.* **14**(4), 2939–2953. https://doi.org/10.1109/TAFFC.2022.3225037 (2023).
19. Chai, L., Liu, Y., Liu, W., Han, G. & He, S. CrowdGAN: Identity-free interactive crowd video generation and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(6), 2856–2871. https://doi.org/10.1109/TPAMI.2020.3043372 (2022).

20. Yi, W., Wu, W., Wang, X. & Zheng, X. Modeling the mutual anticipation in human crowds with attention distractions. *IEEE Trans. Intell. Transp. Syst.* **24**(9), 10108–10117. https://doi.org/10.1109/TITS.2023.3268315 (2023).
21. Cai, Z. et al. Forecasting citywide crowd transition process via convolutional recurrent neural networks. *IEEE Trans. Mob. Comput.* **23**(5), 5433–5445. https://doi.org/10.1109/TMC.2023.3310789 (2024).
22. Zeng, Y., Zhou, S. & Xiang, K. Online-offline interactive urban crowd flow prediction toward IoT-based smart city. *IEEE Trans. Serv. Comput.* **15**(6), 3417–3428. https://doi.org/10.1109/TSC.2021.3099781 (2022).
23. Yang, Y. et al. Multiscenario open-set gait recognition based on radar micro-doppler signatures. *IEEE Trans. Instrum. Meas.* **71**, 1–13. https://doi.org/10.1109/TIM.2022.3214271 (2022).
24. Chen, J., Wang, C. & Liu, Y. Vibration signal based abnormal gait detection and recognition. *IEEE Access* **12**, 89845–89855. https://doi.org/10.1109/ACCESS.2024.3417377 (2024).
25. Naghavi, N. & Wade, E. Towards real-time prediction of freezing of gait in patients with Parkinson's disease: A novel deep one-class classifier. *IEEE J. Biomed. Health Inf.* **26**(4), 1726–1736. https://doi.org/10.1109/JBHI.2021.3103071 (2022).
26. Palash, M. & Bhargava, B. EMERSK-Explainable multimodal emotion recognition with situational knowledge. *IEEE Trans. Multimed.* **26**, 2785–2794. https://doi.org/10.1109/TMM.2023.3304015 (2024).
27. Alharthi, R., Alhothali, A., Alzahrani, B. & Aldhaheri, S. Massive crowd abnormal behaviors recognition using C3D. In *2023 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA* 01–06 (2023). https://doi.org/10.1109/ICCE56470.2023.10043437.
28. Anandhi, R. Spatially-constrained anomaly detection in crowded environments using meta-heuristic algorithm. In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India* 1440–1444 (2023). https://doi.org/10.1109/ICOSEC58147.2023.10276052.
29. Zhou, X. & Xiao, R. Detection of abnormal crowd behavior based on graph convolutional neural network. In *12th International Conference on Information Technology in Medicine and Education (ITME), Xiamen, China* 538–542 (2022). https://doi.org/10.1109/ITME56794.2022.00118.
30. Liyanage, P. & Fernando, P. Suspicious human crowd behaviour detection—A transfer learning approach. In *21st International Conference on Advances in ICT for Emerging Regions (ICter), Colombo, Sri Lanka* 63–68 (2021). https://doi.org/10.1109/ICter53630.2021.9774784.
31. Mehmood, A. Efficient anomaly detection in crowd videos using pre-trained 2D convolutional neural networks. *IEEE Access* **9**, 138283–138295. https://doi.org/10.1109/ACCESS.2021.3118009 (2021).
32. Mu, H., Sun, R., Yuan, G., Li, J. & Wang, M. Crowd behavior detection in videos using statistical physics. In *2021 International Conference on Data Mining Workshops (ICDMW), Auckland, New Zealand* 389–397 (2021). https://doi.org/10.1109/ICDMW53433.2021.00054.
33. Ahmed, R., Rafiq, M. S. & Junej, I. N. Crowd modeling using temporal association rules. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), Magdeburg, Germany* 1–4 (2021). https://doi.org/10.1109/ICHMS53169.2021.9582661.
34. Karki, M. V., Aripirala, A., Vasist, C., Renith, A. & Balasubramanian, A. Abnormal human behavior detection in crowded scenes based on hybrid neural networks. In *4th International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India* 54–58 (2022). https://doi.org/10.1109/I4C57141.2022.10057910.
35. Abdullah, F., Javeed, M. & Jalal, A. Crowd anomaly detection in public surveillance via spatio-temporal descriptors and zero-shot classifier. In *International Conference on Innovative Computing (ICIC), Lahore, Pakistan* 1–8 (2021). https://doi.org/10.1109/ICIC53490.2021.9693003.
36. Bansal, S., Kumar, S. & Bhalla, P. A novel approach to WDM channel allocation: Big Bang–Big crunch optimization. In *The proceeding of Zonal Seminar on Emerging Trends in Embedded System Technologies (ETECH) organized by The Institution of Electronics and Telecommunication Engineers (IETE), Chandigarh Centre, Chandigarh, India* 80–81 (2013).
37. Bansal, S., Chauhan, R. & Kumar, P. A cuckoo search based WDM channel allocation algorithm. *Int. J. Comput. Appl.* **96**(20), 6–12 (2014).
38. Zhou, Y., Qin, M., Wang, X. & Zhang, C. Regional crowd status analysis based on geovideo and multimedia data collaboration. In *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China* 1278–1282 (2021). https://doi.org/10.1109/IMCEC51613.2021.9482018.
39. Kevin Lemuel Thomas, R., Jerome Sanjay, G., Pandeeswaran, C. & Raghi, K. R. Advanced CCTV surveillance anomaly detection, alert generation and crowd management using deep learning algorithm. In *3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT), Vellore, India* 1–6 (2024). https://doi.org/10.1109/AIIoT58432.2024.10574731.
40. Sophia, S. & Joeffred Gladson, J. Human behaviour and abnormality detection using YOLO and Conv2D Net. In *2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal* 70–75 (2024). https://doi.org/10.1109/ICICT60155.2024.10544757.
41. Ge, Z., Jiang, J. & Coombes, M. A congestion-aware path planning method considering crowd spatial-temporal anomalies for long-term autonomy of mobile robots. In *IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom* 7930–7936. (2023). https://doi.org/10.1109/ICRA48891.2023.10160252.
42. Mahmoud, S., Arafa, Y. & Abdelmohsen, M. Dynamic image representations for crowd anomaly detection using generative adversarial networks. In *2023 International Conference on Computer and Applications (ICCA), Cairo, Egypt* 1–6 (2023). https://doi.org/10.1109/ICCA59364.2023.10401681.
43. Veesam, S. B. & Satish, A. R. An empirical taxonomy of video summarization model from a statistical perspective. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3503276
44. Khan, S. D., Bandini, S., Basalamah, S. & Vizzari, G. Analyzing crowd behavior in naturalistic conditions: Identifying sources and sinks and characterizing main flows. In *Neurocomputing* 543–563, vol. 177 (Elsevier BV, 2016). https://doi.org/10.1016/j.neucom.2015.11.049.
45. Farooq, M. U., Saad, M. N. M. & Khan, S. D. Motion-shape-based deep learning approach for divergence behavior detection in high-density crowd. *Vis. Comput.* **38**, 1553–1577. https://doi.org/10.1007/s00371-021-02088-4 (2022).
46. Alzahrani, A. J. & Khan, S. D. Characterization of different crowd behaviors using novel deep learning framework. *Turk. J. Electr. Eng. Comput. Sci.* **29**(1), 12. https://doi.org/10.3906/elk-2004-14 (2021).
47. Bansal, S. et al. Pt/ZnO and Pt/few-layer graphene/ZnO Schottky devices with Al Ohmic contacts using Atlas simulation and machine learning. *J. Sci. Adv. Mater. Devices* **9**, 100798-1–100798-14 (2024).
48. Bansal, S. et al. Optoelectronic performance prediction of HgCdTe homojunction photodetector in long wave infrared spectral region using traditional simulations and machine learning models. *Sci. Rep.* **14**, 28230 (2024).
49. Bansal, S. ANNs supervised learning-based automatic fault detection in a class of wheatstone bridge-oriented transducers. In *2022 IEEE Sponsored Global Conference for Advancement in Technology (GCAT-2022), Nagarjuna College of Engineering & Technology, Bengaluru, Karnataka, India* 1–7 (2022).
50. Bansal, S. & Jain, P. Automatic fault detection in a class of wheatstone bridge-based transducer using ANNs in Verilog HDL. In *2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Galgotias College of Engineering and Technology Greater Noida, India* 466–470 (2022).
51. Veesam, S. B. & Satish, A. R. Design of an iterative method for CCTV video analysis integrating enhanced person detection and dynamic mask graph networks. *IEEE Access* **12**, 157630–157656. https://doi.org/10.1109/ACCESS.2024.3485896 (2024).

Acknowledgements

This research was funded by Fundamental Research Grant Scheme (FRGS), MOE, Malaysia, Code: FRGS/1/2022/TK07/UKM/02/22.

Author contributions

S B Veesam, A R Satish, S Tupakula made substantial contributions to design, analysis and characterization. Y Chinnam, K Prakash, S Bansal participated in the conception, application and critical revision of the article for important intellectual content. M R I Faruque provided necessary instructions for analytical expression, case study for practical use and critical revision of the article purposes.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.P., S.B. or M.R.I.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.