frontiers | Frontiers in Sustainable Cities

TYPE Original Research
PUBLISHED 07 June 2023
DOI 10.3389/frsc.2023.1197434

OPEN ACCESS

EDITED BY
Nan Gao,
Tsinghua University, China

REVIEWED BY
Shiyi Liu,
Arizona State University, United States
Shusen Jing,
University of California, Davis, United States

*CORRESPONDENCE
Chen Jin
✉ jinchen@fastmail.com

RECEIVED 31 March 2023
ACCEPTED 10 May 2023
PUBLISHED 07 June 2023

CITATION
Liu G, Shu L, Yang Y and Jin C (2023) Unsupervised video anomaly detection in UAVs: a new approach based on learning and inference. *Front. Sustain. Cities* 5:1197434. doi: 10.3389/frsc.2023.1197434

Unsupervised video anomaly detection in UAVs: a new approach based on learning and inference

Gang Liu[1], Lisheng Shu[1], Yuhui Yang[2] and Chen Jin[1,3]*

[1]Research Department of Aeronautics, Zhejiang Scientific Research Institute of Transport, Hangzhou, China, [2]School of Telecommunications Engineering, Xidian University, Xi'an, China, [3]College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

In this paper, an innovative approach to detecting anomalous occurrences in video data without supervision is introduced, leveraging contextual data derived from visual characteristics and effectively addressing the semantic discrepancy that exists between visual information and the interpretation of atypical incidents. Our work incorporates Unmanned Aerial Vehicles (UAVs) to capture video data from a different perspective and to provide a unique set of visual features. Specifically, we put forward a technique for discerning context through scene comprehension, which entails the construction of a spatio-temporal contextual graph to represent various aspects of visual information. These aspects encompass the manifestation of objects, their interrelations within the spatio-temporal domain, and the categorization of the scenes captured by UAVs. To encode context information, we utilize Transformer with message passing for updating the graph's nodes and edges. Furthermore, we have designed a graph-oriented deep Variational Autoencoder (VAE) approach for unsupervised categorization of scenes, enabling the extraction of the spatio-temporal context graph across diverse settings. In conclusion, by utilizing contextual data, we ascertain anomaly scores at the frame-level to identify atypical occurrences. We assessed the efficacy of the suggested approach by employing it on a trio of intricate data collections, specifically, the UCF-Crime, Avenue, and ShanghaiTech datasets, which provided substantial evidence of the method's successful performance.

KEYWORDS

drone video anomaly detection, spatio-temporal graph, unsupervised learning, Unmanned Aerial Vehicles, Variational Autoencoder

1. Introduction

The detection of abnormal events in videos, including those captured by Unmanned Aerial Vehicles (UAVs), poses a formidable obstacle as a result of the extensive spectrum of occurrences, coupled with the restricted accessibility of learning resources, and the contextualized definition of abnormal events (Li et al., 2013; Ionescu et al., 2019; Song et al., 2019). UAVs, also known as drones, have seen a significant rise in popularity and usage in recent years due to their cost-effectiveness, versatility, and ability to access hard-to-reach areas. These advanced aerial systems have been increasingly utilized for various applications, such as surveillance in military and civilian contexts, search and rescue operations in disaster-stricken areas, environmental monitoring to track changes in ecosystems, infrastructure inspection, and even in the entertainment industry for aerial photography and filming.

As a result, detecting abnormal events in UAV-captured videos has become increasingly important for ensuring safety and security. Abnormal events in this context can refer to a wide range of occurrences, from intrusions and suspicious activities in surveillance scenarios to detecting signs of natural disasters or accidents in search and rescue operations. The challenge lies in the fact that these events are often context-dependent and can vary greatly in appearance, making it difficult for traditional computer vision algorithms to detect and classify them effectively.

In order to confront this particular issue related to Unmanned Aerial Vehicles (UAVs) or drones, a multitude of pre-existing approaches have been put forth to address challenges such as object detection, tracking, and anomaly recognition. These techniques endeavor to acquire customary spatial and temporal configurations pertaining to appearance and movement of UAVs in various environments. Consequently, they facilitate the identification of irregular occurrences, such as unauthorized UAVs entering restricted areas or deviating from their designated flight paths, by differentiating them from the established normative patterns (Feng et al., 2016; Xu et al., 2017a).

Typically, visual features are extracted from either an entire image (Hasan et al., 2016; Chong and Tay, 2017) or a specific zone of focal concern (Ionescu et al., 2019) to acquire a comprehensive understanding of the inherent spatial and temporal configurations in nature. For instance, these techniques might analyze the shape, size, color, and texture of the UAVs in the imagery data. Additionally, the extracted features can be used to classify the UAVs into different categories, such as fixed-wing, rotary-wing, or hybrid designs, as well as to determine their speed, altitude, and flight patterns.

Investigations within the realm of psychological science have established that individuals possess the capability to accurately identify entities and environments through the employment of contextual visual data (Bar, 2004; Tang et al., 2020). Furthermore, context information has been shown to be beneficial for various computer vision tasks (Ionescu et al., 2017; Hasan et al., 2019; Sun et al., 2019; Kang et al., 2022), including those involving Unmanned Aerial Vehicles (UAVs) or drones. UAVs, which have rapidly advanced in recent years, offer a versatile and efficient solution for remote sensing, surveillance, and data collection in various domains, such as agriculture, disaster response, and law enforcement.

Hence, it is crucial to extract extensive contextual data that surpasses the scope of image-based and object-based characteristics for the precise detection of atypical occurrences within a video, as the visual context serves a pivotal function in ascertaining the existence of such anomalies. In the case of UAV-based surveillance systems, context information can enhance the system's ability to recognize and respond to unusual events or potential threats, which in turn leads to improved safety and security. Taking pedestrian behavior as an example, as depicted in Figure 1, strolling along a pedestrian pathway is perceived as a customary occurrence, whereas ambulating on a busy highway constitutes an atypical incident. Ignoring context information related to sidewalks and highways could lead to false detection in UAV-based surveillance systems.

In the realm of computer vision, context-based abnormal event detection has consistently garnered considerable scholarly attention, including UAV-captured videos. As the use of UAVs, commonly known as drones, has become more widespread in various industries such as agriculture, surveillance, disaster management, and aerial photography, there is an increased demand for reliable and accurate automated analysis of the captured data. Previous efforts in this area manually predefined the collections of context based on human experiences, such as relationship context and scene context. However, the correctness and completeness of these collections are difficult to guarantee, and it is impossible to consider all possible context information. To address these issues, we propose an automatic methodology of inferential reasoning within contextual framework that mines elevated contextual insights derived from the fundamental visual characteristics of information sources. In our approach, a solitary frame is employed to generate a spatial context graph, which facilitates the understanding of object characteristics and their spatial associations. Subsequently, a structural recurrent neural network incorporates these graphs to establish a spatio-temporal context graph. To reason about context, an iterative process involving the mean-field technique is utilized to modify the conditions of nodes and edges within the spatio-temporal graph, ultimately enabling the extraction of semantic context from visual attributes.

UAVs, also known as drones, have gained significant traction in various applications, such as monitoring system, and search and rescue operations. They are equipped with sophisticated sensors and cameras to capture high-resolution aerial imagery and video footage. Analyzing this data for the detection of abnormal events is crucial to ensuring the safety and efficiency of UAV operations. To account for the rarity of abnormal events, we present an unsupervised approach for inferring spatio-temporal context graphs through the implementation of a scene segmentation technique. By employing a deep Variational Autoencoder model grounded in graph theory, we effectively partition scenes into distinct clusters. Subsequently, event categorization into normal and anomalous occurrences is performed based on their respective cluster affiliations. The proposed methodology is rigorously evaluated using the UCF-Crime, Avenue, and ShanghaiTech datasets, including UAV-captured videos. Our approach substantially surpasses the performance of leading unsupervised techniques while yielding a noticeable enhancement in comparison to established supervised methodologies.

Although we mentioned earlier that visual contextual information is crucial for comprehensive object and scene recognition, it is especially beneficial for various computer vision tasks, including those involving UAVs. However, manually defining context sets is limited in the case of diverse, changing, and unpredictable context-related events (Pang et al., 2020). Therefore, our method automatically learns contextual information from data instead of manually pre-defining contextual content. By performing a contextual reasoning approach, we establish a connection between the visual context and the interpretation of deviant occurrences, thus overcoming the semantic disparity that exists between the two. The methodology we propose is versatile, and relevant to an extensive variety of applications in

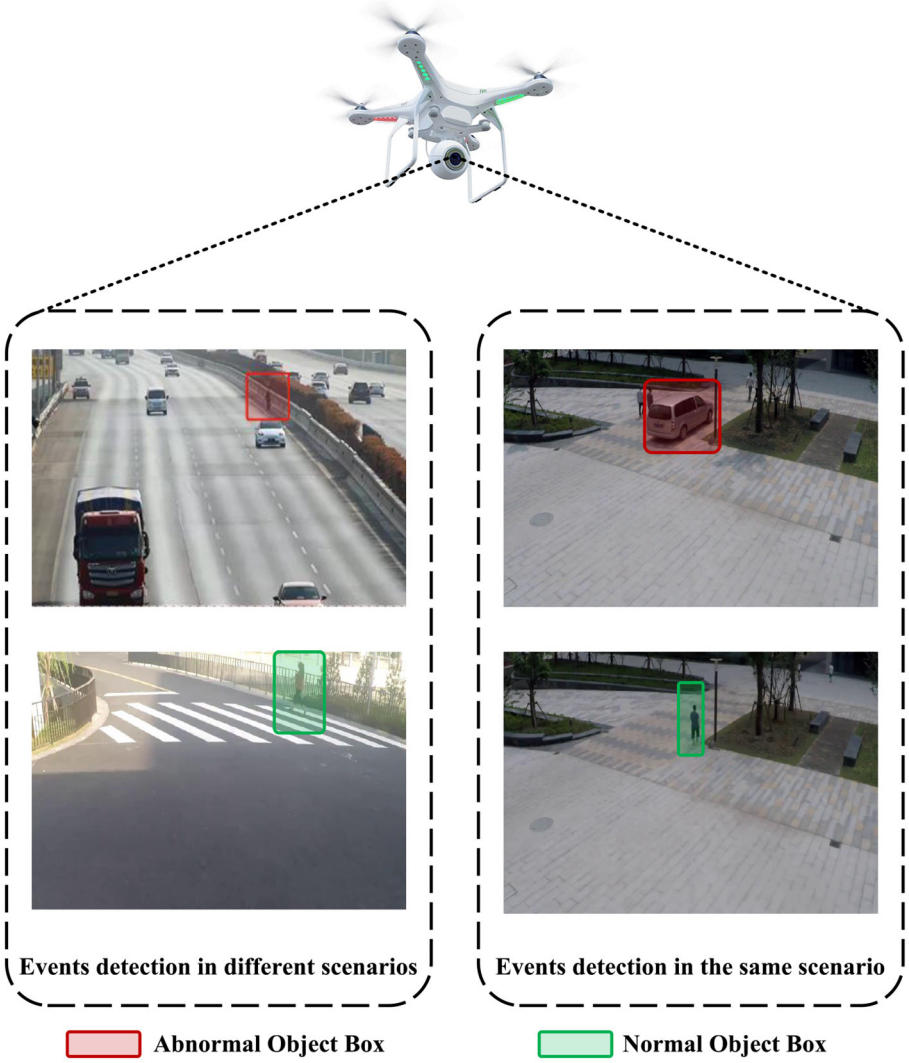

FIGURE 1
Unmanned Aerial Vehicle (UAV) perspective for abnormal event detection. The **(Left)** subplot illustrates the same entity(the pedestrian) in different scenes (highway and sidewalk) generating an abnormal event and a common event, respectively. The **(Right)** subplot depicts the abnormal and normal events caused by different objects (vehicles and the pedestrian) in the same environment(the park on a university campus).

which contextual factors are crucial, such as monitoring through Unmanned Aerial Vehicle systems.

To summarize, our work offers the subsequent contributions.

- An innovative method that utilizes scene-aware context reasoning: Our study presents a novel methodology for identifying anomalous incidents in videos, including those captured by UAVs, by leveraging scene-aware context reasoning, which helps bridge the disparity in meaning between the visual environment and anomalous occurrences. This is rare in methods of the field.
- Development of a contextual graph for spatio-temporal data: We construct a spatio-temporal context graph that encodes and reasons about context information, thereby enhancing the precision of identifying anomalous events in various scenarios, including UAV-captured videos.
- Introduction of deep Variational Autoencoder architecture that incorporates graph-based techniques for unsupervised visual environment clustering: Our method incorporates a graph-based deep Variational Autoencoders for unsupervised scenario clustering, which enables the identification of different scene types and the accurate detection of aberrations that are obscure and contextual in nature.
- Enhanced discrimination between normal and those deemed abnormal events: By leveraging scene clustering, our approach can better discriminate between normal and abnormal events in various scenes, leading to more accurate detections, particularly in UAV-based surveillance systems.
- Significant improvement in unsupervised abnormal event detection accuracy: Our proposed method demonstrates a substantial increase in the accuracy of unsupervised abnormal event detection when evaluated against cutting-edge approaches, including those applied to UAV-captured videos.

2. Related work

Over the last several years, a considerable number of scholars (Mehran et al., 2009; Li et al., 2013; Luo et al., 2017; Ribeiro et al., 2018; Feng et al., 2021; Georgescu et al., 2021) have conducted research on video anomaly recognition. Generally, the studies can be partitioned into three stages according to the specific technology used, which are the traditional machine learning (Mahadevan et al., 2010; Antić and Ommer, 2011; Li et al., 2013; Lu et al., 2013; Cheng et al., 2015; Hasan et al., 2016), the hybrid stage combining machine learning and deep learning (Hinami et al., 2017; Luo et al., 2017; Smeureanu et al., 2017; Ravanbakhsh et al., 2018; Sabokrou et al., 2018a), and the deep learning stage (Sabokrou et al., 2015, 2017, 2018b; Xu et al., 2015; Chong and Tay, 2017; Ionescu et al., 2019).

The initial video anomaly detection studies mainly used manual features to construct feature spaces. These studies used traditional machine learning methods, such as basic methods to determine whether events obey the normal state distribution (Saligrama et al., 2010), Gaussian mixture models (Kratz and Nishino, 2009), and Markov models to infer anomalous features (Tipping and Bishop, 1999; Leyva et al., 2017), and sparse learning methods (Luo et al., 2017) that are more effective than the former two (Lu et al., 2013). However, these traditional machine learning methods have a certain degree of dependence on the selection of features and are often adapted to particular scenarios.

Subsequently, with the emergence of deep learning methods, artificial features are replaced by deep features, which can effectively monitor and analyze video semantic concepts. With the properties of automatically learning and extracting video features according to the environment, deep learning methods have led to a series of studies on unsupervised learning methods (Xu et al., 2015; Hasan et al., 2016; Luo et al., 2017; Sabokrou et al., 2017; Liu et al., 2018; Wang et al., 2018; Ye et al., 2019). These works reconstruct the video so that the model gets a stronger response to anomalous frames during testing. Although such studies overcome the problem of feature dependency, they are only applicable to video types with few anomalous patterns and short time series, having the limitation of low generalization ability. In contrast to these approaches, Zhao et al. (2017) started to consider the use of video local spatio-temporal information for video reconstruction by 3D convolutional autoencoder, yet their study still has weak generalization ability for anomalous event diversity. Besides, due to redundancy on consecutive frames (Zhou et al., 2018), using 3D convolutional kernels (Tran et al., 2015) to extract features in dense RGB frame sequences can sometimes be computationally more expensive.

In addition, along with the objective social needs such as diversification and complexity of video anomaly detection, researchers have also been prompted to focus on the full and reasonable utilization of video multidimensional information. Sultani et al. (2018) adopted multiple instance learning methods to achieve better abnormality detection. This work is one of the early studies to provide new solutions for weakly supervised video abnormality detection. On the contrary, Zhu and Newsam (2019) focused on the influence of dynamic information on "anomalies" and introduced the attention mechanism to highlight the response of anomalous video feature segments. But the potential relationship between instances is still not exploited. Nevertheless, the overall unreasonable assumption of independent identical distribution between instances remains yet. Furthermore Zhang et al. (2019) analyzed the contrast between affirmative and opposing states of instances in multiple instance learning, proposed the concept of quantification of intra-packet loss, and used temporal convolutional network (TCN) for temporal information correlation. Although this work mostly achieved anomalous differentiation between normal and abnormal, it was not sensitive to other more neutral video segments within and the differentiation was not significant. In order to clean away label noise, enhancing sensitivity to neutral segments, Zhong et al. (2019) adopted a hybrid approach of weakly supervised and fully supervised to handle the detection task. They used graph convolution to transfer information to denoise the normal clips in the abnormal video, obtaining pseudo-labels to train 3D convolutional network (C3D) (Tran et al., 2015) for abnormality recognition. With high training complexity in the denoising process, it is very likely that abnormal "miscleaning" will occur. The outcome of this phenomenon is a reduction in the precision of detecting and locating abnormal events.

Among the approaches above, reconstruction and prediction is the current mainstream video anomaly event detection method. Although it settled the problems of strong feature dependency and insufficient utilization of video multidimensional information in previous, the present investigation into the detection of anomalous video activity is faced with two major challenges. The first challenge is the considerable complexity inherent in the training process of the employed methods. The second challenge pertains to the burden of differentiating noise from redundant information in the features. Notably, some interesting studies using relational inference modules to enhance the efficacy of video anomalous event detection (Choi et al., 2012; Park et al., 2012; Leach et al., 2014) have been proposed recently. Most of them have detected deviating anomalous events by defining the rule set of context for inference with the help of temporal relationships between video frames. While such methods alleviate the ambiguity of anomalous event definitions in redundant information, they in turn increase the scene-dependence of anomalous event definitions. Inspired by them, we adopt the unsupervised idea of scene-aware inference, and after encoding the contextual information into a spatio-temporal relationship graph, we attempt attention mechanisms to perform the next inference step and anomaly detection by updating the state in the graph. Differently, our method can achieve automatic mining of high-level features directly related to anomalous events based on the underlying visual information. That means it can be applied to detect anomalous occurrences in diverse settings while bridging the huge gap between the underlying data and the concepts of anomalous events.

3. Methodologies

During our study, we have evolved a spatio-temporal graph-based method for context-aware anomaly event inference in videos. As illustrated in Figure 2, we modeled graphical representation of spatial relationships using every individual frame of the video

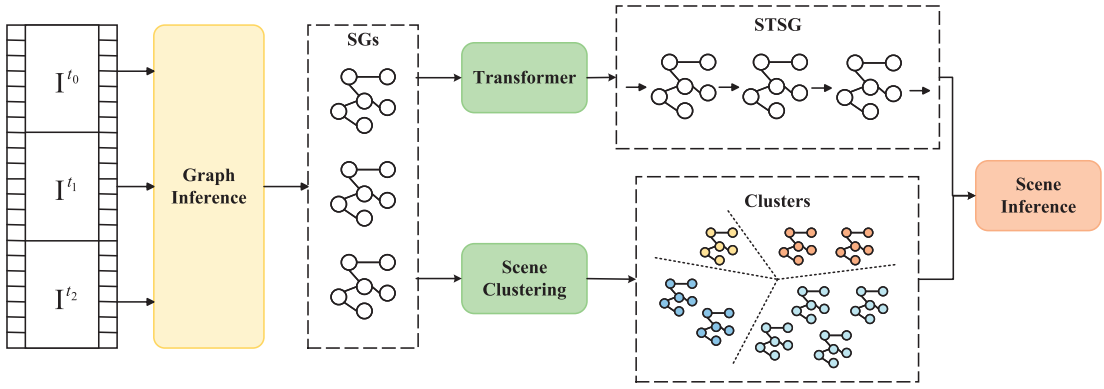

FIGURE 2
Overview of our proposed framework. The framework is proposed that involves the use of a spatial scene graph (SSG) to analyze each frame *I* based on object bounding boxes. The SSG nodes are denoted by *i*, *j*, and *k*, with timestamps t_0, t_1, and t_2. A spatio-temporal scene graph (STSG) is then constructed by incorporating the temporal modeling of the SSG using a Transformer. Moreover, scene clustering is concurrently performed on the SSG to differentiate between different scene types. Finally, a reasoning model is established on top of the STSG to detect abnormal events across different scenes.

as an object. In this spatial graph model, we encapsulated the manifestation of entities and their spatial interconnections within each frame. The spatial graph was then input into a transformer to learn dynamic features of each object in the temporal dimension, ultimately developing the spatio-temporal graph model. Inspired by Shao et al. (2016), we employed unsupervised clustering to classify scenes and deduce the spatio-temporal scene graph model. Finally, we detected anomalous events utilizing the spatio-temporal scene graph model and the context features acquired through scene clustering.

3.1. Spatio-temporal scene graph

We represent video feature information as a Spatio-Temporal Scene Graph (STSG), where nodes encode object appearance, spatial graph model edges depict object relationships, and temporal edges model object dynamic features. This encoding scheme allows us to infer more semantic information about objects (nodes), object spatio-temporal relationships (edges), and event occurrence scenarios (entire graph model) compared to existing research. Consequently, we can detect single-point anomalies, relationship anomalies, and group anomalies. We achieve the goal of detecting various scenario anomalies in videos by constructing and inferring Spatio-Temporal Scene Graphs, implementing the inference through iterative updates of the node and edge states in the Spatio-Temporal Graph Model.

3.1.1. Formulation

Assuming an input of a visual medium in the form of a video consisting of a total of N frames, denoted as $V = [I^1, I^2, \cdots, I^N]$, where I is a certain frame of the video. The Region Proposal Network (RPN) (Ren et al., 2015) is employed to generate object bounding boxes for individual frames. The top-K bounding boxes in the n-th frame are selected as B^n, which includes the entire frame as an additional bounding box. To construct a spatial scene graph (SSG) for each frame, we use the image containing k enclosure boxes. In the SSG, each node v_i represents an object, while the edge $e_{i,j}$ represents the association among the objects. To enable the inference function through iterative graph updates, we assign a "normal" or "abnormal" label to each node and edge of the SSG.

In the context of the n-th frame, the designation of the i-th object is denoted by the label y_i^n, while the label $y_{i,j}^n$ is assigned to the relationship between the i-th and j-th objects. We utilize binary categorization, whereby the label '0' denotes normal objects or inter-object relationships, while '1' indicates one-field objects or inter-object relations. To establish a comprehensive definition of anomalous labels, we define the set encompassing all such labels: $y^n = \{y_i^n, y_{i,j}^n | i, j = 1, 2, \cdots, K; i \neq j\}$. And then SSG model can be formalized as $arg\ \max_{y^n} P(\mathbf{y}^n | I^n, B^n)$, where

$$P\left(\mathbf{y}^n \mid I^n, B^n\right) = \prod_{i,j \in K} \prod_{i \neq j} P\left(y_i^n, y_{i,j}^n \mid I^n, B^n\right). \tag{1}$$

Next, we use a Transformer to incorporate temporal information of multiple SSGs to generate the STSG. In the n-th frame, the node denoted as v_i is linked solely to its corresponding node v_i in the $n+1$-th frame by means of the temporal edge denoted as $e_{i,i}$, which has a corresponding relation label denoted as $y_{i,i}^n$. Inspired by (Sun et al., 2020), the conclusive probability distribution is determined with

$$P(\mathbf{y} \mid V, B) = \prod_{n \in N} \prod_{i,j} \prod_{i \neq j} P\left(y_i^n, y_{i,j}^n, y_{i,i}^n \mid V, B\right), \tag{2}$$

where $\mathbf{y} = \{y_i^n, y_{i,j}^n | n = 1, 2, \cdots, N\}$ denotes collection of total exception tags of the input.

3.1.2. Inferencing on graphs

As previously mentioned, we perform contextual semantic inference in videos by generating states of nodes and edges in the

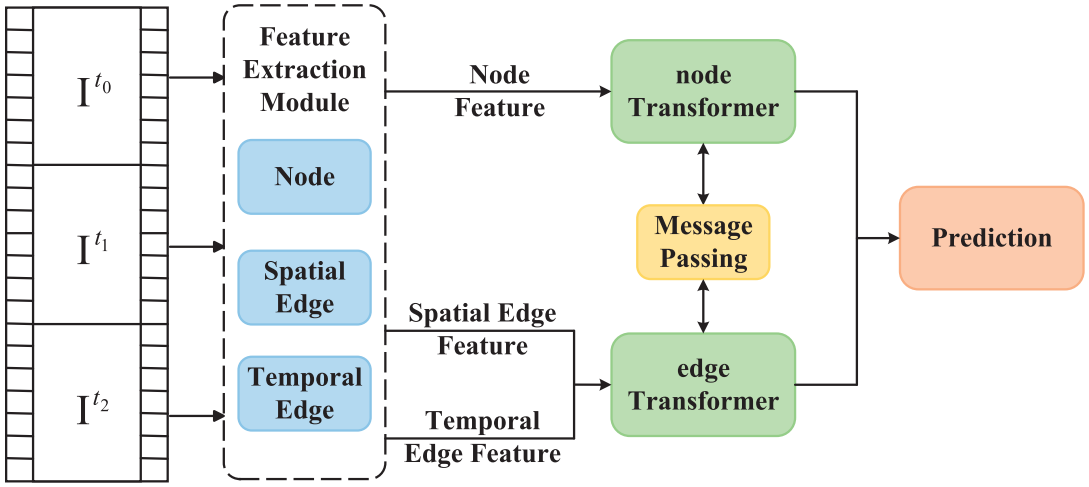

FIGURE 3
Schematic representation of the STSG inference process. The delineation of object boundaries is accomplished through the utilization of a region proposal network (RPN), which processes raw frames. Subsequently, node spatial edge and temporal edge attributes are derived through specialized feature extraction modules. These features serve as the initial input for nodeTransformer and edgeTransformer in reasoning. The message-passing technique is employed to renew Transformers. After computation, nodeTransformer and edgeTransformer yield the predicted anomaly labels y.

model of graph. In this work, we adopt the mean-field graph model inference method (Xu et al., 2017b; Qi et al., 2019). The probability function $P(\mathbf{y} \mid \cdot)$ will be approximated as $Q(\mathbf{y} \mid \cdot)$. In particular, we utilized the states $h_i^n, h_{i,j}^n$ to denote the present state of node i and the edge between node i, j in frame n, individually. $P(y_i^n \mid \cdot)$ of node i depends on the state $h_i^n, h_{i,j}^n$ of all nodes and edges.

The approximation $Q(y_i^n \mid \cdot)$ depends only on the current state, i.e., $Q(y_i^n \mid \cdot) = Q(y_i^n \mid h_i^n)$. The probability distribution approximation for the edge is also implemented using this idea.

Drawing inspiration from Vaswani et al. (2017); Xu et al. (2022), we leverage Transformers to calculate the state of Q. Depicting in Figure 3, STSG model we proposed employs the nodeTransformer and edgeTransformer to model the hidden states of nodes and spatio-temporal edges, respectively. Both the nodeTransformer and edgeTransformer serve to update node and edge states, allowing for the inference of context semantics from visual features. The nodeTransformer computation is expressed as follows:

$$\begin{aligned}
\mathbf{Q} &= \mathbf{W}_Q^{(1)}\mathbf{E}_{\text{node}}^\top + \mathbf{W}_Q^{(2)}\mathbf{H}_{\text{node}}^t, \\
\mathbf{K} &= \mathbf{W}_K^{(1)}\mathbf{E}_{\text{node}}^\top + \mathbf{W}_K^{(2)}\mathbf{H}_{\text{node}}^t, \\
\mathbf{V} &= \mathbf{W}_V^{(1)}\mathbf{E}_{\text{node}}^\top + \mathbf{W}_V^{(2)}\mathbf{H}_{\text{node}}^t, \\
\mathbf{M} &= \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}, \\
\mathbf{H}_{\text{node}}^{t+1} &= \text{LN}(\mathbf{E}_{\text{node}} + \text{MLP}(\mathbf{M}, \mathbf{H}_{\text{node}}^t)),
\end{aligned} \tag{3}$$

where $\mathbf{E}_{\text{node}} \in \mathbb{R}^{n\times d}$ is the node embedding matrix, $\mathbf{H}_{\text{node}}^t \in \mathbb{R}^{n\times d_h}$ is the hidden state matrix of nodeTransformer at time t, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{n\times d_k}$ are the query, key, and value matrices, respectively, $\mathbf{M} \in \mathbb{R}^{n\times d_v}$ is the output of the self-attention mechanism, LN denotes layer normalization, and MLP represents a multi-layer perceptron that applies non-linear transformations to the concatenated input features. Here, $\mathbf{W}_Q^{(1)}, \mathbf{W}_K^{(1)}, \mathbf{W}_V^{(1)} \in \mathbb{R}^{d\times d_k}$ and $\mathbf{W}_Q^{(2)}, \mathbf{W}_K^{(2)}, \mathbf{W}_V^{(2)} \in \mathbb{R}^{d_h\times d_k}$ are learnable parameters, d_k and d_v denote the dimensions of the key and value vectors, respectively, and MLP consists of two fully connected layers with ReLU activation and a skip connection.

Similarly, the computation of edgeTransformer is formulated as

$$\begin{aligned}
\mathbf{Q} &= \mathbf{W}_Q^{(3)}\mathbf{E}_{\text{edge}}^\top + \mathbf{W}_Q^{(4)}\mathbf{H}_{\text{edge}}^t, \\
\mathbf{K} &= \mathbf{W}_K^{(3)}\mathbf{E}_{\text{edge}}^\top + \mathbf{W}_K^{(4)}\mathbf{H}_{\text{edge}}^t, \\
\mathbf{V} &= \mathbf{W}_V^{(3)}\mathbf{E}_{\text{edge}}^\top + \mathbf{W}_V^{(4)}\mathbf{H}_{\text{edge}}^t, \\
\mathbf{M} &= \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}, \\
\mathbf{H}_{\text{edge}}^{t+1} &= \text{LN}(\mathbf{E}_{\text{edge}} + \text{MLP}(\mathbf{M}, \mathbf{H}_{\text{edge}}^t)),
\end{aligned} \tag{4}$$

where $\mathbf{E}_{\text{edge}} \in \mathbb{R}^{m\times d_e}$ is the edge embedding matrix, $\mathbf{H}_{\text{edge}}^t \in \mathbb{R}^{m\times d_h}$ is the hidden state matrix of edgeTransformer at time t, and $\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}$ are defined similarly as in the node self-attention mechanism. Here, $\mathbf{W}_Q^{(3)}, \mathbf{W}_K^{(3)}, \mathbf{W}_V^{(3)} \in \mathbb{R}^{d_e\times d_k}$ and $\mathbf{W}_Q^{(4)}, \mathbf{W}_K^{(4)}, \mathbf{W}_V^{(4)} \in \mathbb{R}^{d_h\times d_k}$ are learnable parameters, and MLP consists of two fully connected layers with ReLU activation and a skip connection. In this formulation, the edge embeddings are fed into the edge self-attention mechanism, and the resulting representation is used to update the hidden states of edges through a multi-layer perceptron.

To enhance the inference process's efficiency, we utilize message-passing techniques during computation. The message-passing matrix $\mathbf{H}_{\text{msg}}^{t+1}$ at time $t+1$ is computed using a simple linear transformation:

$$\mathbf{H}_{\text{msg}}^{t+1} = \text{ReLU}(\mathbf{W}_{\text{msg}}^{(1)}\mathbf{H}_{\text{node}}^{t+1} + \mathbf{W}_{\text{msg}}^{(2)}\mathbf{H}_{\text{edge}}^{t+1}), \tag{5}$$

where $\mathbf{H}_{\text{msg}}^{t+1} \in \mathbb{R}^{n\times d_h}$ is the message passing matrix at time $t+1$, $\mathbf{W}_{\text{msg}}^{(1)}, \mathbf{W}_{\text{msg}}^{(2)} \in \mathbb{R}^{d_h\times d_h}$ are learnable parameters, and ReLU denotes the rectified linear unit activation function.

In the aforementioned formulation, node and edge representations are integrated using a linear transformation. This can be perceived as a simplified version of message passing, where messages are computed as linear combinations of the hidden states of nodes and edges. Such a formulation can be more efficient than conventional message passing, particularly for large graphs with dense connections. The resulting model can encapsulate both spatial and temporal dependencies within the graph and learn to deduce context semantics from visual features.

3.2. Scenario clustering

The identification of scene types is essential for comprehending abnormal events since it is typical for a standard event in one scene to be considered abnormal in another. To tackle this issue, we propose an unsupervised scene clustering approach to discern scene types and deduce the STSG. Given that humans can effortlessly differentiate The categorization of scenarios based on a solitary image, we cluster the SSG of static frames to distinguish between various scenes. By categorizing events into different scenes, each group can possess its own standard events, which can be utilized to deduce the contextual framework and identify anomalous occurrences.

To cluster the scenarios, we introduce a graphic tech-based Variational Autoencoder (VAE). Specifically, we consider a graph represented by an adjacency matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$, where each node is associated with a feature vector $\mathrm{X} \in \mathbb{R}^{K\times D}$. Our objective is to learn a node representation $\mathrm{Z}^{(l)}$ in the l-th layer of a Graph Convolutional Network (GCN), as depicted in Figure 4. The representation of a node in the l-th layer is provided by

$$\mathrm{Z}_i^{(l)} = \sigma\left(\sum_{j=1}^{K} \mathbf{A}_{i,j}\mathrm{Z}_j^{(l-1)}\mathbf{W}_c^{(l)}\right), \tag{6}$$

where $\sigma(\cdot)$ represents an activation function, and $\mathbf{W}_c^{(l)}$ denotes the trainable weight matrix of the l-th layer. In the first layer, we initialize $\mathrm{Z}^{(0)} = \mathrm{X}$. To generate the scene graph, we assume that each node is connected to all other nodes and thus set all elements of $\mathbf{A}$ to 1.

To execute clustering in the latent space, the Soft K-Means algorithm can be employed with the subsequent equations:

$$\begin{aligned} z_i &= q(z|x_i), \\ \hat{\gamma}_{i,m} &= \frac{\exp(-\beta||z_i-\mu_m||^2)}{\sum_{j=1}^{M}\exp(-\beta||z_i-\mu_j||^2)}, \\ \hat{\phi}_m &= \frac{1}{N}\sum_{i=1}^{N}\hat{\gamma}_{i,m}, \\ \hat{\mu}_m &= \frac{\sum_{i=1}^{N}\hat{\gamma}_{i,m}z_i}{\sum_{i=1}^{N}\hat{\gamma}_{i,m}}, \end{aligned} \tag{7}$$

where z_i represents the latent variable of data point i, μ_m denotes the mean of the m-th cluster, β is the temperature parameter controlling the assignment's softness, and $\hat{\gamma}i, m$ is the soft assignment of data point i to the m-th cluster.

The soft mixture-component membership prediction $\hat{\gamma}$ can be obtained using the softmax function as follows:

$$\hat{\gamma} = \mathrm{softmax}(\alpha\hat{\gamma}'), \tag{8}$$

where $\hat{\gamma}'$ is the soft assignment matrix derived from the Soft K-Means algorithm and α is the temperature parameter.

To obtain the estimated parameters of each cluster, we can employ the following equation

$$\hat{\Sigma}_m = \frac{\sum_{i=1}^{N}\hat{\gamma}_{i,m}(z_i-\hat{\mu}_m)(z_i-\hat{\mu}_m)^T}{\sum_{i=1}^{N}\hat{\gamma}_{i,m}}, \tag{9}$$

where $\hat{\Sigma}_m$ is the covariance matrix of the m-th cluster.

The energy function can be expressed as follows:

$$E = \sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\gamma}_{i,j}\left(\beta||z_i-\mu_j||^2 + \log\hat{\phi}_j - \log\hat{\gamma}_{i,j}\right), \tag{10}$$

where β is the temperature parameter and $\hat{\phi}_j$ is the fraction of data points assigned to the j-th cluster.

The loss function for clustering using the VAE model can be formulated as follows:

$$\mathcal{L}_{\mathrm{clu}} = \frac{1}{N}\sum_{i=1}^{N}E + \lambda_1\sum_{m=1}^{M}\sum_{j=1}^{d}\frac{1}{\hat{\Sigma}_{m,j,j}}, \tag{11}$$

where λ_1 represents a trade-off factor, while d represents latent-space'dimensionality.

During clustering, the number of clustering centers m denotes the preset number of scene types, and we set m to 10 to cover common scenes (e.g., campuses, highways, subways, etc). The training batch of scene clustering is set to 1,024, and we use the RMSprop optimizer with a 0.0001 learning rate to train the clustering model.

The VAE model learns a lower-dimensional representation of the graph utilized for clustering. The Soft K-Means algorithm is applied to cluster the nodes in the latent space, and the estimated parameters of each cluster are obtained using equations akin to those employed in the Gaussian Mixture Model (GMM). The loss function for clustering using the VAE model is adapted to encompass the reconstruction loss and the Kullback-Leibler (KL) divergence term, which serve to train the VAE model.

3.3. Model optimization

Utilizing the methodology of segmenting visual environments, the contextual scenarios in videos were partitioned into distinct groups, and the objects and relationships in these scenes were accurately labeled. This labeled data was then utilized in the STSG inference, wherein the network's nodes were trained to predict the abnormality of objects, while the edges were trained to predict the abnormality of the relationships between the objects. The architecture underwent comprehensive education through the utilization of the backpropagation method, employing a holistic approach from inception to completion, and all the learnable model parameters optimized simultaneously. To accomplish this,

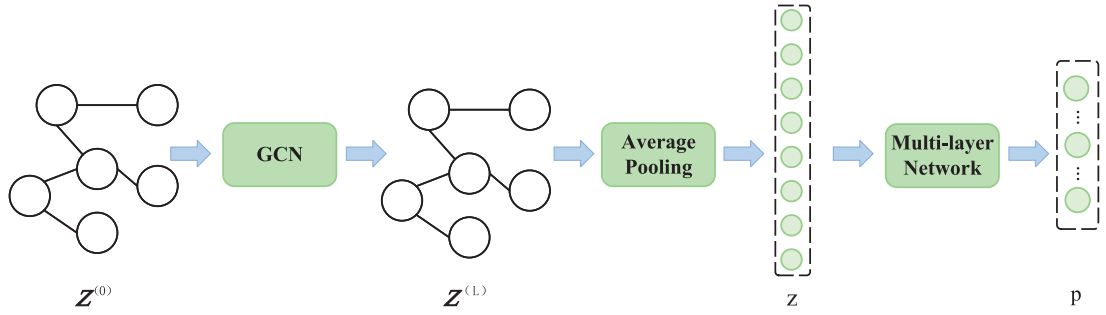

FIGURE 4
Depiction of the graph-based deep Variational Autoencoders (VAEs).

we utilized a cross-entropy loss function with regularization, the purpose of which was to augment the likelihood delineated before. For every cluster, the probability of normalcy or deviation was ascertained for each vertex v_i and corresponding edge $e_{i,j}$ within the n-th frame, individually.

$$\begin{aligned} P\left(y_i^n \mid V, B\right) &= \text{softmax}\left(MLN\left(h_i^n\right)\right), \\ P\left(y_{i,j}^n \mid V, B\right) &= \text{softmax}\left(MLN\left(h_{i,j}^n\right)\right), \end{aligned} \tag{12}$$

where the term MLN refers to a multi-layer neural network that is comprised of two fully-connected layers. In order to simplify the notation, we denote the anomaly probability of y_i^n and $y_{i,j}^n$ given V and B as p_i^n and $p_{i,j}^n$, respectively. The loss function of graph inference is given by

$$\begin{aligned} \mathcal{L}^m = & \frac{1}{NK} \sum_{n=1}^{N} \sum_{i=1}^{K} \mathcal{L}_{\text{cls}}\left(y_i^n, p_i^n\right) \\ & + \frac{1}{NK^2} \sum_{n=1}^{N} \sum_{i=1}^{K} \sum_{j=1}^{K} \mathcal{L}_{\text{cls}}\left(y_i^n, p_{i,j}^n\right) \\ & + \lambda_2 \left\|\mathbf{W}_m\right\|_1 . \end{aligned} \tag{13}$$

The loss function utilized in the training of the MLN classifier represents the logistic loss function, denoted as $\mathcal{L}_{\text{cls}}(\cdot,\cdot)$. Variables associated with the Multilayer Neural Network(MLN) are denoted as $\mathbf{W}_m$. In this investigation, the regularization parameter, denoted as λ_2, is assigned a value of 0.0001. The exponent m in $\mathcal{L}^m$ indicates that the classification model has undergone training with the m-th cluster. Within each cluster, all entities y_i^n and associations $y_{i,j}^n$ receive a designation as typical, while data points from other groups are randomly selected and labeled as simulated irregularities to facilitate the training of the model.

3.4. Irregularity score

In order to detect abnormal events in videos, independent classifiers are trained for objects and relationships within each group. These classifiers generate classification scores, which are then utilized to calculate the final anomaly scores. In the context of the m-th group, assessment scores for categorization, represented as $P^m\left(y_i^n \mid V, B\right)$ and $P^m\left(y_{i,j}^n \mid V, B\right)$, are derived from the examination of test data through the evaluation of Equation 12. The anomaly score is then determined as the minimal categorization metric observed across all scenario clusters.

Deviation metrics pertaining to entities and their interconnections are denoted as s_i^n and $s_{i,j}^n$, respectively, and are computed for each of the M groups. The former score reflects the detection of individual anomalies, while the latter reflects the detection of group anomalies. To achieve granular detection at the frame level, the highest score encompassing all entities and associations within a single frame is identified as the representative anomaly score for that particular frame. To ensure that the irregularity score varies smoothly across frames, we apply a Gaussian filter to enforce temporal smoothness of the final frame-level anomaly scores.

4. Experiments

4.1. Datasets

The effectiveness of our proposed method is evaluated on three benchmark datasets, namely UCF-Crime (Sultani et al., 2018), Avenue (Lu et al., 2013), and ShanghaiTech (Luo et al., 2017). The UCF-Crime dataset is an extensive compilation of authentic surveillance footage encompassing 13 distinct categories of anomalous occurrences across various environments. This collection is comprised of 1,610 videos for training purposes and 290 videos designated for evaluation, all of which were utilized in our experiments. Avenue, on the other hand, contains 16 training and 21 testing videos with a total of 35,240 frames, each lasting about 2 minutes. The dataset includes four types of abnormal events: running, walking in opposite direction, throwing objects, and loitering. Lastly, the ShanghaiTech dataset includes 13 scenes with complex light conditions and various viewpoints, and consists of over 270,000 training frames and 130 abnormal events. We utilized all of these datasets to comprehensively evaluate the performance of our proposed method.

4.2. Evaluation metric

We assess the performance of our proposed method at the frame level by computing anomaly scores for each frame. The performance of the method is evaluated using the Receiver Operating Characteristic (ROC) curve (Fawcett, 2006), which involves progressively adjusting the benchmark for irregularity values. The relevant evaluation metrics employed encompass the Area Under the Curve (AUC ↑) and the Equal Error Rate (EER ↓). Moreover, the false alarm rate serves as an assessment indicator for the likelihood of incorrect categorization. Enhanced performance of the anomaly detection technique is signified by an elevated AUC merit (Lobo et al., 2008), a diminished EER merit, and other merits.

TABLE 1 A comparative evaluation of abnormal event detection outcomes between unsupervised and supervised techniques.

Training	Method	AUC ↑	False alarm ↓
Unsupervised	Hasan et al. (2016)	49.8%	26.9%
	Ionescu et al. (2019)	62.1%	9.3%
	Lu et al. (2013)	67.4%	3.8%
	Ours	**76.3%**	2.2%
Supervised	SVM baseline	50%	–
	Sultani et al. (2018)	69.2%	**2.1%**

↑ Indicates that advanced values correspond to superior performance, and ↓ signifies that lower scores are indicative of better results.
The bold values indicate the best value among all experimental results.

4.3. Comparisons

4.3.1. Analysis on the UCF-crime dataset

The methodology we put forth undergoes assessment and juxtaposition with numerous prevalent unsupervised and supervised techniques, employing the UCF-Crime dataset for this comparative analysis. The performance of our method is reported in Table 1 in terms of the AUC and false alarm rate, respectively. To ensure a fair comparison, we reconstructed the research conducted by Ionescu et al. (2019), substituting their employed detection mechanism with the Region Proposal Network (RPN) detector to enhance the methodology. The performances of other compared methods are taken from Sultani et al. (2018). The results show that our method outperforms cutting-edge unsupervised technique, with an improvement of 8.9 and 1.6% on the Area Under the Curve (AUC) and false alarm rate evaluations, respectively. Moreover, our approach exhibits similarity to the most advanced supervised technique (Sultani et al., 2018) currently available in the field, achieving comparable AUC scores and false alarm rates without the need for video-level annotations. This demonstrates the effectiveness of our method in detecting unknown abnormal events in real-world applications. The ROC curves of our method are plotted in Figure 5, which encompasses the contours of unsupervised methodologies and surpasses the study of Hasan et al. (2016); Ionescu et al. (2019); Lu et al. (2013) at diverse benchmarks. True positive of the proposed method marginally exceeds the research of Sultani et al. (2018) when a middle threshold is selected, indicating the effectiveness of our method.

4.3.2. Analysis on the avenue dataset

On the Avenue dataset, our method outperforms all existing methods in terms of both the AUC and EER evaluations, as shown in Table 2. The cutting-edge research of Ye et al. (2019) achieved AUC values of 85.9%, while our approach gained an advancement of 4.0%, demonstrating the effectiveness and robustness of our method.

4.3.3. Analysis on the ShanghaiTech dataset

Furthermore, we present the findings of our experimental evaluation conducted on the demanding ShanghaiTech dataset, which contains complex scenes and various actions. According to the information presented in Table 3, the proposed approach overtakes the leading-edge strategies on this dataset, demonstrating its effectiveness in detecting abnormal events in challenging settings.

4.4. Ablation study

Table 4 showcases a comparative analysis of distinct constituents' contributions within the proposed technique designed for detecting unsupervised abnormal events. The term "w/o spatial relationships" signifies the exclusion of associations in space dimension, wherein the STSGs are converted into object-oriented series over numerous frames, which are then simulated by Transformers. The term "w/o temporal relationships" implies its performance on the SSG inference disregarding any temporal connections, while "w/o relationships" employs a twin set of fully-connected layers to simulate individual objects in a standalone manner. We conducted identical scene clustering for the aforementioned three scenarios. The term "w/o scene clustering" denotes the exclusion of scenario clustering and solely relying on a one-class discriminator to differentiate between regular and aberrant occurrences. Referring to Table 4, it is observed that discarding spatial dependencies, temporal dependencies, or spatio-temporal dependencies decreases the AUC execution by 6.9% − 13.6%, indicating the relevance of information associations for distinguishing irregular occurrences. Scenario clustering significantly improves performance, and the exhibited performance in distinguishing diverse environments to detect anomalous incidents affirms the efficacy of this approach. Furthermore, the enhancement reinforces the effectiveness of the unsupervised scene clustering technique utilized during the training phase.

4.5. Incident reckon

In order to detect and determine anomalous incidents based on irregularity values, we employ a method that involves choosing the local maxima among the chronological progression of irregularity

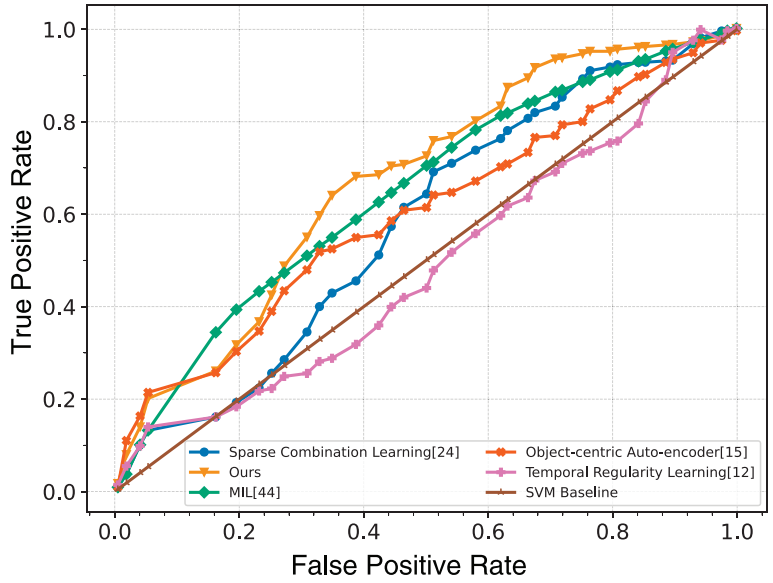

FIGURE 5
Comparison of ROC curves for various unsupervised and supervised approaches on the UCF-Crime dataset.

values within a given video. To identify meaningful local maxima, we utilize the persistence1D algorithm, and then define a fixed time interval for the region. To group nearby expanded local maximum regions, we adopt the approach outlined in Luo et al. (2017), which results in the final abnormal temporal regions where anomalous incidents can be accurately determined.

On the Avenue dataset, the outcomes of the proposed approach are presented in Table 5, which exhibits the quantity of identified anomalous incidents and false alarms. The strategy we developed can reliably identify anomalous incidents in comparison to the approaches taken in Wang et al. (2018) and Luo et al. (2017). The false alarm rate of the proposed approach exceeds the work in Medel and Savakis (2016), primarily due to their use of minute benchmarks of irregularity values to identify anomalous incidents. However, the approach we proposed identifies 47 true anomalous incidents, compared to the 39 anomalous incidents identified by the strategy in Wang et al. (2018). These outcomes manifest the superior validity of the methodology we developed in verifying the time span of anomalous incidents, rendering it a more feasible choice for implementation in real-world scenarios.

TABLE 2 Comparative evaluation of abnormal event detection performance using AUC and EER metrics.

Method	AUC ↑	EER ↓
Chong and Tay (2017)	78.2%	21.3%
Hasan et al. (2016)	69.4%	26.1%
Ionescu et al. (2019)	81.0%	–
Luo et al. (2017)	82.1%	–
Liu et al. (2018)	83.5%	–
Wang et al. (2018)	84.7%	22.9%
Morais et al. (2019)	85.6%	–
Ye et al. (2019)	85.9%	–
Ours	**89.9%**	**20.4%**

The bold values indicate the best value among all experimental results.

TABLE 3 Comparative evaluation of abnormal event detection performance using frame-level AUC and EER metrics.

Method	AUC ↑	EER ↓
Chong and Tay (2017)	61.2%	–
Luo et al. (2017)	67.9%	–
Wang et al. (2018)	71.7%	–
Liu et al. (2018)	72.1%	-
Ionescu et al. (2019)	72.9%	–
Ours	**75.2%**	**25.1%**

The bold values indicate the best value among all experimental results.

5. Conclusion

Throughout this article, we propose a novel approach for unsupervised anomalous incidents identification in videos, particularly those acquired via Unmanned Aerial Vehicles (UAVs), which involves the utilization of a contextually responsive reasoning strategy. As UAVs are increasingly utilized in various applications such as surveillance, search and rescue, and environmental monitoring, anomalous incident detection in UAV-captured videos is pivotal for ensuring safety and security. Contextual inference overtly entails the extraction of high-level environmental knowledge from low-level vision-oriented characteristics. Our approach generates a spatiotemporal scenario graph to facilitate the explicit establishment of the

TABLE 4 Performance evaluation of individual components of the proposed approach in terms of AUC and false alarm.

Method	AUC ↑	False alarm ↓
W/o temporal relationships	69.4%	2.9%
W/o spatial relationships	62.7%	5.2%
W/o relationships	62.8%	12.5%
W/o scene clustering	64.9%	7.1%
Ours	**76.3%**	**1.9%**

The bold values indicate the best value among all experimental results.

TABLE 5 Outlier event identification outcomes in terms of count of identified occurrences and false alarms.

Method	True positives ↑	False alarm ↓
Wang et al. (2018)	39	3
Luo et al. (2017)	44	5
Morais et al. (2019)	42	13
Ours	**47**	**3**

The bold values indicate the best value among all experimental results.

vision-oriented environment, by embedding objects' visual morphology and their spatiotemporal associations in graphic representations. This approach is particularly beneficial in UAV-captured videos, where the aerial perspective offers unique contextual information. Furthermore, we evolve a graph-based deep Variational Autoencoder model for scenario clustering that can capably ascertain scenario categories and deduce the spatio-temporal scenario graph within unsupervised. This enables our method to accurately detect aberrant occurrences with contextual dependencies and ambiguous sources in various environments, including those captured by UAVs. Our experiments on three datasets, including UAV-captured videos, exhibit the superiority of our approach over current unsupervised methodologies, while simultaneously highlighting its comparability with contemporary supervised techniques that represent the cutting-edge of the field. Subsequent research will endeavor to investigate more detailed contextual feature in order to expand the methodology from detecting anomalies at the frame dimension to the more precise pixel level, further enhancing the effectiveness of atypical occurrences identification in UAV-captured videos.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

GL proposed the idea and wrote this article with YY and LS. YY conducted the experiments. CJ organized the entire research, provided funding supports, handles the manuscript, and correspondence during the publication process. All authors contributed to manuscript writing, revision, read, and approved the submitted version.

Funding

This work was supported by the Zhejiang 'JIANBING' R&D Project (No. 2022C01055) and R&D Project of Department of Transport of Zhejiang Province (No. 2021010).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Antić, B., and Ommer, B. (2011). "Video parsing for abnormality detection." in 2011 *International Conference on Computer Vision* (Barcelona: IEEE), 2415–2422. doi: 10.1109/ICCV.2011.6126525

Bar, M. (2004). Visual objects in context. *Nat. Rev. Neurosci.* 5, 617–629. doi: 10.1038/nrn1476

Cheng, K.-W., Chen, Y.-T., and Fang, W.-H. (2015). "Video anomaly detection and localization using hierarchical feature representation and gaussian process regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA: IEEE), 2909–2917. doi: 10.1109/CVPR.2015.7298909

Choi, M. J., Torralba, A., and Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recogn. Lett.* 33, 853–862. doi: 10.1016/j.patrec.2011.12.004

Chong, Y. S., and Tay, Y. H. (2017). "Abnormal event detection in videos using spatiotemporal autoencoder," in *Advances in Neural Networks-ISNN 2017: 14th International Symposium, ISNN 2017, Sapporo, Hakodate, and Muroran, Hokkaido, Japan, June 21-26, 2017, Proceedings, Part II 14* (Cham: Springer), 189–196. doi: 10.1007/978-3-319-59081-3_23

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.* 27, 861–874. doi: 10.1016/j.patrec.2005.10.010

Feng, J.-C., Hong, F.-T., and Zheng, W.-S. (2021). "Mist: multiple instance self-training framework for video anomaly detection," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 14009–14018. doi: 10.1109/CVPR46437.2021.01379

Feng, Y., Yuan, Y., and Lu, X. (2016). "Deep representation for abnormal event detection in crowded scenes," in *Acm on Multimedia Conference*, 591–595. doi: 10.1145/2964284.2967290

Georgescu, M. I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., and Shah, M. (2021). "Anomaly detection in video via self-supervised and multi-task learning," in *Computer Vision and Pattern Recognition* (Nashville, TN: IEEE). doi: 10.1109/CVPR46437.2021.01255

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., and Davis, L. S. (2016). "Learning temporal regularity in video sequences, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, USA: IEEE). doi: 10.1109/CVPR.2016.86

Hasan, M., Paul, S., Mourikis, A. I., and Roy-Chowdhury, A. K. (2019). "Context-aware query selection for active learning in event recognition," in *IEEE Transactions on Pattern Analysis & Machine Intelligence* (IEEE), 1.

Hinami, R., Mei, T., and Satoh, S. (2017). "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 3619–3627. doi: 10.1109/ICCV.2017.391

Ionescu, R. T., Khan, F. S., Georgescu, M.-I., and Shao, L. (2019). "Object-centric auto-encoders and dummy anomalies for abnormal event detection in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 7842–7851. doi: 10.1109/CVPR.2019.00803

Ionescu, R. T., Smeureanu, S., Alexe, B., and Popescu, M. (2017). "Unmasking the abnormal events in video, in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice, Italy: IEEE). doi: 10.1109/ICCV.2017.315

Kang, Y., Rahaman, M. S., Ren, Y., Sanderson, M., White, R. W., and Salim, F. D. (2022). App usage on-the-move: context-and commute-aware next app prediction. *Pervasive Mobile Comput.* 87, 101704. doi: 10.1016/j.pmcj.2022.101704

Kratz, L., and Nishino, K. (2009). "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 1446–1453. doi: 10.1109/CVPR.2009.5206771

Leach, M. J., Sparks, E. P., and Robertson, N. M. (2014). Contextual anomaly detection in crowded surveillance scenes. *Pattern Recogn. Lett.* 44, 71–79. doi: 10.1016/j.patrec.2013.11.018

Leyva, R., Sanchez, V., and Li, C. T. (2017). "Video anomaly detection with compact feature sets for online performance," in *IEEE Transactions on Image Processing* (IEEE), 3463–3478. doi: 10.1109/TIP.2017.2695105

Li, W., Mahadevan, V., and Vasconcelos, N. (2013). Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intellig.* 36, 18–32. doi: 10.1109/TPAMI.2013.111

Liu, W., Luo, W., Lian, D., and Gao, S. (2018). "Future frame prediction for anomaly detection-a new baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 6536–6545. doi: 10.1109/CVPR.2018.00684

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). Auc: a misleading measure of the performance of predictive distribution models. *Glob. Ecol. Biogeogr.* 17, 145–151. doi: 10.1111/j.1466-8238.2007.00358.x

Lu, C., Shi, J., and Jia, J. (2013). "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE International Conference on Computer Vision* (Sydney, NSW: IEEE), 2720–2727. doi: 10.1109/ICCV.2013.338

Luo, W., Liu, W., and Gao, S. (2017). "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 341–349. doi: 10.1109/ICCV.2017.45

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Francisco, CA: IEEE), 1975–1981. doi: 10.1109/CVPR.2010.5539872

Medel, J. R., and Savakis, A. (2016). Anomaly detection in video using predictive convolutional long short-term memory networks. *arXiv [Preprint]*. arXiv:1612.00390.

Mehran, R., Oyama, A., and Shah, M. (2009). "Abnormal crowd behavior detection using social force model," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 935–942. doi: 10.1109/CVPR.2009.5206641

Morais, R., Le, V., Tran, T., Saha, B., Mansour, M., and Venkatesh, S. (2019). "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 11996–12004. doi: 10.1109/CVPR.2019.01227

Pang, G., Yan, C., Shen, C., Hengel, A., v. d., and Bai, X. (2020). "Self-trained deep ordinal regression for end-to-end video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 12173–12182. doi: 10.1109/CVPR42600.2020.01219

Park, S., Kim, W., and Lee, K. M. (2012). "Abnormal object detection by canonical scene-based contextual model," in *Computer Vision-ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12* (Berlin; Heidelberg: Springer), 651–664. doi: 10.1007/978-3-642-33712-3_47

Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., and Van Gool, L. (2019). Stagnet: an attentive semantic rnn for group activity and individual action recognition. *IEEE Trans. Circuits Syst. Video Technol.* 30, 549–565. doi: 10.1109/TCSVT.2019.2894161

Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., and Sebe, N. (2018). "Plug-and-play cnn for crowd motion analysis: an application in abnormal event detection," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE), 1689–1698. doi: 10.1109/WACV.2018.00188

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Cambridge, MA: MIT Press), 91–99.

Ribeiro, M., Lazzaretti, A. E., and Lopes, H. S. (2018). A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recogn. Lett.* 105:13–22. doi: 10.1016/j.patrec.2017.07.016

Sabokrou, M., Fathy, M., Hoseini, M., and Klette, R. (2015). "Real-time anomaly detection and localization in crowded scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE), 56–62. doi: 10.1109/CVPRW.2015.7301284

Sabokrou, M., Fayyaz, M., Fathy, M., and Klette, R. (2017). Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Trans. Image Process.* 26, 1992–2004. doi: 10.1109/TIP.2017.2670780

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., and Klette, R. (2018a). Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. *Comput. Vis. Image Understanding* 172, 88–97. doi: 10.1016/j.cviu.2018.02.006

Sabokrou, M., Khalooei, M., Fathy, M., and Adeli, E. (2018b). "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3379–3388. doi: 10.1109/CVPR.2018.00356

Saligrama, V., Konrad, J., and Jodoin, P.-M. (2010). Video anomaly identification. *IEEE Signal Process. Magaz.* 27, 18–33. doi: 10.1109/MSP.2010.937393

Shao, W., Salim, F. D., Song, A., and Bouguettaya, A. (2016). Clustering big spatiotemporal-interval data. *IEEE Trans. Big Data* 2, 190–203. doi: 10.1109/TBDATA.2016.2599923

Smeureanu, S., Ionescu, R. T., Popescu, M., and Alexe, B. (2017). "Deep appearance features for abnormal behavior detection in video," in *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19* (Catania: Springer), 779–789. doi: 10.1007/978-3-319-68548-9_70

Song, H., Sun, C., Wu, X., Chen, M., and Jia, Y. (2019). "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," in *IEEE Transactions on Multimedia* (IEEE), 1.

Sultani, W., Chen, C., and Shah, M. (2018). "Real-world anomaly detection in surveillance videos," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 6479–6488. doi: 10.1109/CVPR.2018.00678

Sun, C., Jia, Y., Hu, Y., and Wu, Y. (2020). "Scene-aware context reasoning for unsupervised abnormal event detection in videos," in *Proceedings of the 28th ACM International Conference on Multimedia* (New York, NY: ACM), 184–192. doi: 10.1145/3394171.3413887

Sun, C., Song, H., Wu, X., and Jia, Y. (2019). "Learning weighted video segments for temporal action localization," in *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8-11, 2019, Proceedings, Part I 2* (Xian: Springer), 181–192. doi: 10.1007/978-3-030-31654-9_16

Tang, K., Zhang, H., Wu, B., Luo, W., and Liu, W. (2020). "Learning to compose dynamic tree structures for visual contexts," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA: IEEE). doi: 10.1109/CVPR.2019.00678

Tipping, M. E., and Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural Comput.* 11, 443–482. doi: 10.1162/089976699300016728

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision* (Cambridge, MA: IEEE), 4489–4497. doi: 10.1109/ICCV.2015.510

Vaswani, A., Shazeer, N., Parmer, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY: Curran Associates), 6000–6010.

Wang, L., Zhou, F., Li, Z., Zuo, W., and Tan, H. (2018). "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *2018 25th IEEE International Conference on Image Processing (ICIP)* (IEEE), 2276–2280. doi: 10.1109/ICIP.2018.8451070

Xu, D., Ricci, E., Yan, Y., Song, J., and Sebe, N. (2015). Learning deep representations of appearance and motion for anomalous event detection. *arXiv [Preprint]*. arXiv:1510.01553. doi: 10.5244/C.29.8

Xu, D., Yan, Y., Ricci, E., and Sebe, N. (2017a). Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Understanding* 156, 117–127. doi: 10.1016/j.cviu.2016.10.010

Xu, D., Zhu, Y., Choy, C. B., and Fei-Fei, L. (2017b). "Scene graph generation by iterative message passing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI: IEEE), 5410–5419. doi: 10.1109/CVPR.2017.330

Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., and Ma, J. (2022). Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers. *arXiv [Preprint]*. arXiv:2207.02202.

Ye, M., Peng, X., Gan, W., Wu, W., and Qiao, Y. (2019). "Anopcn: Video anomaly detection via deep predictive coding network," in *Proceedings of the 27th ACM International Conference on Multimedia* (New York, NY: ACM), 1805–1813. doi: 10.1145/3343031.3350899

Zhang, J., Qing, L., and Miao, J. (2019). "Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection," in *2019 IEEE International Conference on Image Processing (ICIP)* (Taipei: IEEE), 4030–4034. doi: 10.1109/ICIP.2019.8803657

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., and Hua, X.-S. (2017). "Spatio-temporal autoencoder for video anomaly detection," in *Proceedings of the 25th ACM international conference on Multimedia* (New York, NY: ACM), 1933–1941. doi: 10.1145/3123266.3123451

Zhong, J.-X., Li, N., Kong, W., Liu, S., Li, T. H., and Li, G. (2019). "Graph convolutional label noise cleaner: train a plug-and-play action classifier for anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 1237–1246. doi: 10.1109/CVPR.2019.00133

Zhou, B., Andonian, A., Oliva, A., and Torralba, A. (2018). "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich: Springer International Publishing), 803–818. doi: 10.1007/978-3-030-01246-5_49

Zhu, Y., and Newsam, S. (2019). Motion-aware feature for improved video anomaly detection. *arXiv [Preprint]*. arXiv:1907.10211.