scientific reports

Check for updates

OPEN

Contextual information based anomaly detection for multi-scene aerial videos

Girisha S[1], Ujjwal Verma[2], Manohara M M. Pai[3]✉ & Radhika M. Pai[1]

Aerial video surveillance using Unmanned Aerial Vehicles (UAV) is gaining much interest worldwide due to its extensive applications in monitoring wildlife, urban planning, disaster management, anomaly detection, campus security, etc. These videos are processed and analyzed for strange/odd/anomalous patterns, which are essential requirements of surveillance. But manual analysis of these videos is tedious, subjective, and laborious. Hence, developing computer-aided systems for analyzing UAV-based surveillance videos is crucial. Despite this interest, in the literature, most of the video surveillance applications are developed focusing only on CCTV-based surveillance videos which are static. Thus, these methods cannot be extended for scenarios where the background/context information is dynamic (multi-scene). Further, the lack of standard UAV-based anomaly detection datasets has restricted the development of novel algorithms. In this regard, the present work proposes a novel multi-scene aerial video anomaly detection dataset with frame-level annotations. In addition, a novel Computer Aided Decision (CAD) support system is proposed to analyze and detect anomalous patterns from UAV-based surveillance videos. The proposed system holistically utilizes contextual, temporal, and appearance features for the accurate detection of anomalies. A novel feature descriptor is designed to effectively capture contextual information necessary for analyzing multi-scene videos. Additionally, temporal and appearance features are extracted to handle the complexities of dynamic videos, enabling the system to recognize motion patterns and visual inconsistencies over time. Furthermore, a new inference strategy is proposed that utilizes a few anomalous samples along with normal samples to identify better decision boundaries. The proposed method is extensively evaluated on the proposed UAV anomaly detection dataset and performs competitively with respect to state-of-the-art methods with an AUC of 0.712.

Keywords Unmanned Aerial Vehicles (UAV), Video Anomaly Detection, Surveillance, Auto-Encoders

Anomalous events are defined as an occurrence of odd incidents. The detection of these events is crucial for providing security. In general, anomalous events can be broadly grouped into three categories namely: spatial anomalies, temporal anomalies and appearance anomalies[1,2]. Spatial anomalies are related to the location of the object with respect to the scene. Temporal and appearance anomalies are based on the trajectories and appearance of the object respectively. The definition of anomaly is mostly subjective and depends on the context of the scene[2]. An event that is an anomaly in one scene may be normal in another scene. For instance, a truck on road is normal, while on sidewalk is anomaly. Hence, defining anomaly is challenging. But, accurate and early detection of these events can reduce the risk to human life and thus is an essential aspect of security[3,4].

The current technology available for providing security, analyzes the videos acquired from monitoring systems such as CCTVs installed at fixed locations. These surveillance videos are manually analysed by security personal for detecting anomalies which is time-consuming and tedious process. Therefore, the past few years have seen several works on designing algorithms for detecting anomalous activities in videos[5–7]. However, a majority of existing methods analyse videos from static surveillance cameras installed at fixed locations such as pedestrian walkaway (CUHK Avenue[8], Ped1, Ped2[9]) Subway entrance/exit[10]. In the recent past, a few studies have focused on analysing videos from multiple scenes[4]. However, the videos analysed in these works contain either minor or no camera motion. There is hardly any study on detecting anomalous activity in a video acquired by a moving

[1]Department of Data Science and Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India. [2]Department of Electronics and Communication Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India576104. [3]Department of Information and Communication technology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal 576104, India. ✉email: mmm.pai@manipal.edu

camera. For example, popular datasets such as ShanghaiTech and UCF crime contains videos collected from different locations with a static camera (little to no camera motion). Thus, there are very few standard datasets available for the development of multi-scene anomaly detection algorithms.

Last few years have seen an increased interest in using Unmanned Aerial Vehicles (UAVs) for surveillance[11]. Compared to a static CCTVs with a fixed coverage area, videos acquired from UAV based surveillance systems can cover larger area with varying perspectives. Besides, UAVs have the advantage of mobility and can be rapidly deployed. Despite the better flexibility and mobility offered by UAV based surveillance system, there are limited works on anomaly detection in videos acquired using UAVs[12–14]. However, the videos for these studies are acquired at a single location (car parking) and the variations in the background scene information are limited. The present study focuses on detecting anomalous activities by analysing videos from UAVs. These videos are acquired at multiple locations (multi-scene) and consist of significant camera motion (Figure 1) with varying viewpoints. No published reports are available on this kind of multi-scene anomalous activity detection from UAV aerial videos with significant variations in background scene information.

In general, an anomaly detection algorithm comprises of two steps namely training and detection[1,515]. The training phase consists of developing a model for normal activities using the features extracted from the training videos containing only normal activities. In the detection phase, an anomaly score is assigned to the test video by feeding the same types of extracted features to the developed model. However, these models are only trained on normal patterns and hence the decision boundary may not accurately discriminate anomalous patterns. There exist other methods which address the problem of anomaly detection as an outlier detection where all the outliers are considered anomalous[16–18]. However, these methods fail to detect local anomalies which can be defined as those patterns that are closer to normal patterns but are anomalous in nature. For example, consider a situation of a stationary vehicle present in a parking zone vs a stationary vehicle present on road. Here, the appearance and temporal characteristics of both vehicles are similar. However, the vehicle present on road is an example of a local anomaly since its characteristics are similar to normal patterns. Recently, few studies focuses on features extracted from object of interest (pedestrian). This object centric approach allows to accurately locate the anomalous object in the frame[5].

The majority of the anomalous activity detection algorithms are developed for single-scene scenarios where the background is constant. Hence, these models cannot be adopted to detect anomalies in multi-scene scenarios such as UAVs where the background changes and hence the context of the scene. In these situations, realizing the context of the scene is important as it facilitates accurate detection of anomalies. For instance, a pedestrian on the road is an anomalous event while a person walking on sidewalk is a normal event. Therefore, the background scene information (road/sidewalk) would be useful in detecting an anomalous event. This work proposes to integrate the background scene information along with appearance and motion information for anomalous activity detection. To the best of our knowledge, there is no existing work on incorporating background scene information for anomalous activity detection.

The proposed approach is specifically designed for UAV-based surveillance, integrating contextual, temporal, and appearance-based features to handle dynamic multi-scene environments. Unlike traditional static CCTV-based methods, our object-centric algorithm effectively adapts to aerial video scenarios where backgrounds and perspectives continuously change. These enhancements ensure robustness in UAV applications, addressing challenges that static-camera-based methods cannot overcome. In the present study, an object-centric approach is presented which extracts contextual (background scene), appearance and motion features with a focus on object of interest (vehicles and human). A four class semantic segmentation algorithm is utilized to extract the background scene information, while autoencoder based reconstruction error is utilized for extracting

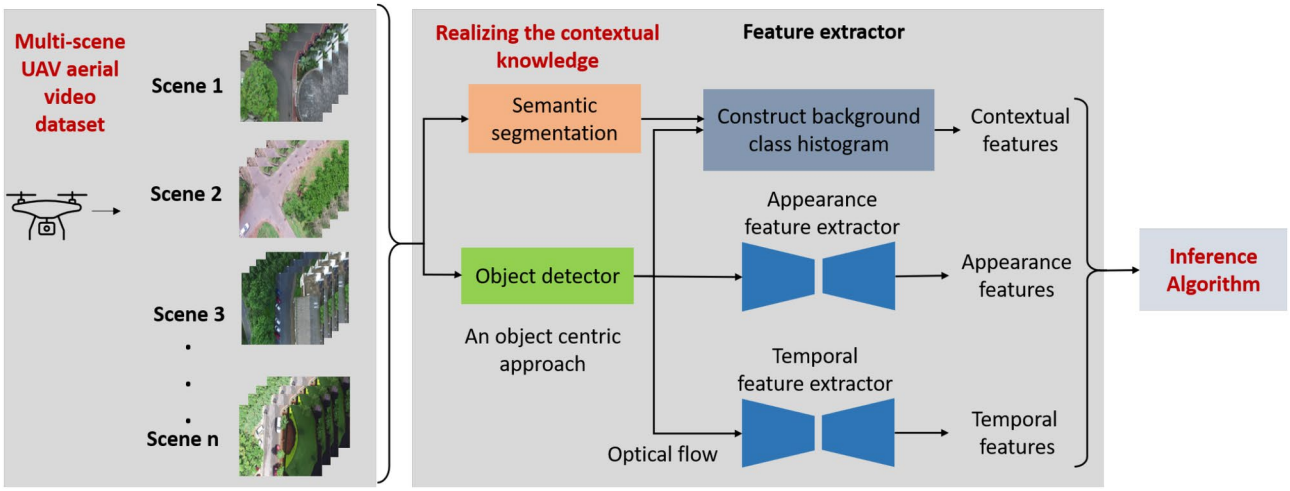


Fig. 1. Overview of proposed model: Multi-scene UAV aerial video dataset creation. Object-centric contextual, appearance and temporal features are first extracted from each frame of the aerial video. The inference algorithm is formulated as a multi-class classification algorithm on these extracted features to identify anomalous events.

appearance and motion features. Subsequently, a new inference algorithm is designed which constitutes a set of one-class SVM classifiers that are trained on the extracted features to detect anomalies. In addition, the new inference algorithm is capable of handling local anomalies that contribute to the accuracy of the model. We argue that incorporating features extracted from few anomalous videos can improve the performance of the inference model. The contributions of the research are as follows:

1. A new multi-scene UAV anomaly detection dataset (MUAAD: Manipal UAV Anomalous Activity Dataset) is presented. This provides a common platform for researchers to compare the developed models for UAV-based anomaly detection.
2. A new object-centric method for anomaly detection is proposed which captures contextual, temporal and appearance-based features. This allows the model to process multi-scene scenarios and detect spatial, temporal and appearance based anomalies.
3. A new inference algorithm is introduced to detect anomalous patterns. This algorithm is trained such that it is capable of detecting local anomalies efficiently which in-turn improves the accuracy of detection.

The structure of the paper is as follows: Section “Related works” presents the recent literature regarding the anomaly detection in videos. The details of the proposed dataset is presented in Section "UAV anomaly detection dataset generation". Section “Methodology” describes the methodology for detecting anomalies from UAV aerial videos. Section "Results and discussion" presents the various results of the proposed system. Finally, conclusion of the paper is given in section “Conclusion”.

Related works

Various approaches have been put forward to solve the issue of anomaly detection in surveillance videos. Also, several datasets are available for the development and evaluation of anomaly detection models. A detailed discussion of these can be found in[1–3]. In this section, initially, a summary of various datasets available for anomaly detection followed by a brief discussion on various methods proposed for video anomaly detection has been presented.

Video anomaly detection datasets

The development of anomaly detection system for videos is dependent on the availability of annotated datasets. Furthermore, they provide a common platform for the researcher to evaluate the developed algorithm. In literature, several popular benchmark datasets are available for anomaly detection[8–10,19]. The Subway dataset proposed in[10], provides videos collected at a subway entrance and exit. The UMN[19] dataset has 11 outdoor video scenes with staged anomalies. In[8], the authors proposed CUHK Avenue dataset which has 37 videos collected from a single camera which contains 47 anomalous events. UCSD ped1 and ped2[9] are another popular dataset used for anomaly detection. Ped 1 and Ped 2 dataset contains video acquired from two stationary camera overlooking pedestrian walkaways. The videos of UCSD ped1 and ped2 are collected at a low resolution. All these datasets are developed for single-scene video anomaly detection. A few multi-scene video anomaly detection datasets developed are available in literature such as ShanghaiTech[4] and UCF crime[20]. ShanghaiTech dataset contains videos collected from 13 different scenes with 130 abnormal events. However, ShanghaiTech and UCF crime datasets contains videos collected from static CCTV cameras with a limited coverage area and fixed background with little or no camera motion. In the present study, a new anomaly detection dataset containing the videos acquired from the camera on-board a moving Unmanned Aerial Vehicle (UAV) is proposed. Unlike existing datasets, the proposed dataset contains background scene information in the form of pixel level mask for four classes (greenery, road, construction and water bodies). The contextual scene information can be inferred from these masks which would aid in identifying anomalous activities in multi-scene videos with significant variation in the background information.

Video anomaly detection

Video anomaly detection algorithms can be broadly grouped into three categories namely: distance-based methods, probabilistic methods and reconstruction based methods. The distance-based methods construct a model of normal patterns[5,21–24]. The distance of test patterns from the model is utilized to estimate the anomaly score. In these methods, handcrafted features[25,26], as well as deep features[6,24,27], are widely used to represent the appearance and temporal domain of the object. Subsequently, one class SVMs are popularly used in these methods to identify the decision boundaries of normal patterns required for estimating distances[5–7]. Also, distance metrics such as Mahalanobis is used to measure the distance from normal patterns[27,28]. The probabilistic methods measure anomaly scores in probability space[18,29–33]. Gaussian Mixture Models and Markov Random Fields are popularly used traditional methods[30,32]. In the reconstruction based method, the given input data is converted to a high-level representation or a latent space[5,34–36]. Further, these inputs are reconstructed from this high-level representation. Since these models are trained only on normal patterns, any anomalies would produce higher reconstruction errors which are used for determination of anomalies. To this end, auto-encoders are extensively used to transform input to latent space. Recently, GAN based methods are gaining popularity for anomaly detection. However, these methods do not consider the contextual knowledge required for detecting multi-scene video anomalies.

Recent advancements uses transformer-based future frame prediction networks and spatio-temporal cross-transformers to enhance anomaly detection performance[37]. However, real-time detection remains a challenge due to computational constraints, environmental changes, and occlusions, necessitating further exploration into multi-camera integration, edge computing, and hybrid deep learning architectures[38]. The introduction of large-scale datasets like CADG, which captures crowd anomalies from both drones and ground cameras, has

helped bridge domain gaps, yet stable anomaly detection from moving drones continues to be an issue[39]. While deep learning techniques such as CNNs, transformers, and YoLo have significantly advanced video surveillance, there is still a need for more robust models capable of handling large-scale, real-time surveillance data[40]. Comprehensive surveys on video anomaly detection highlight existing research gaps, analyzing the strengths and limitations of various methods while identifying future trends such as multi-modal anomaly detection, process automation, and improved dataset curation[40]. Additionally, a focused review on moving camera video anomaly detection (MC-VAD) categorizes methods across diverse domains, emphasizing dataset constraints and the need for novel approaches to improve anomaly detection in UAV surveillance applications[41]. These studies collectively underscore the importance of continued research in UAV-based anomaly detection, particularly in developing real-time, scalable, and computationally efficient models for intelligent surveillance systems.

Aerial videos often suffer from degradations due to sensor noise, compression artifacts, motion blur, and adverse environmental conditions, which can significantly impact anomaly detection performance. Ensuring high-quality input data is essential for reducing false positives and improving the robustness of detection algorithms[42,43]. Recent studies have developed tailored methodologies to assess screen content, highlighting differences in perceptual characteristics, benchmarking techniques, and future research directions in quality of experience (QoE) modelling[44]. Beyond traditional quality assessment, emerging deep learning-based methods enhance image quality prediction by incorporating rich subjective rating distributions[45,46]. Aerial video anomaly detection has traditionally focused on visual signals alone, overlooking the potential influence of audio cues in real-world scenarios. To bridge this gap, multimodal approaches integrating both visual and audio information can enhance anomaly detection performance. Inspired by studies on audiovisual saliency, where humans are naturally drawn to sound sources, a fusion of spatial-temporal visual features with audio cues can improve the detection of anomalies in aerial videos[47,48].

The anomalous activities occur rarely and developing a robust model for detection of anomalous activities is a challenging task. Hence, popular methods proposed in literature are semi-supervised where the model is trained on normal patterns. However, these models fails to detect local anomalies. In this context, this work formulates the inference algorithm such that it utilizes a few samples of anomalous events to identify better decision boundaries. Note that a recent work[15] formulated the few shot anomaly detection task as identifying the anomalous activity with few sequential frames of a video. However, our proposed approach utilizes the features from few anomalous videos to learn a more robust model. To the best of our knowledge, very limited work exists that considers minor supervision to anomalous samples to improve the performance of anomaly detection in UAV videos.

The scarcity and ambiguous definition of anomalies make it challenging to develop anomaly detection datasets. Also, lack of standard UAV based anomaly detection datasets, make it difficult to develop Computer Aided Decision (CAD) support systems. Hence, it urges to develop a new dataset for detecting anomalous activities in UAV aerial videos. Besides, the existing methods for anomalous activity detection are designed for single-scene videos with a fixed or constant background. Therefore, these approach would have limited success when there is a significant change in the background scene information, such as the video from the camera on-board a moving UAV. The UAV videos contain a different perspective view (topdown) of the scene as compared to the videos acquired from CCTVs. In this regard, the present work aims at the development of CAD systems to detect anomalies from UAV aerial videos. Furthermore, this study presents a novel feature extractor that holistically extracts contextual, temporal and appearance features required for multi-scene video anomaly detection. Also, the present work proposes a new inference algorithm to efficiently determine the decision boundaries needed for the accurate detection of anomalies.

UAV anomaly detection dataset generation

The present study proposes a new UAV aerial video dataset for anomaly detection. The videos are acquired from the camera on-board a UAV (DJI Phantom 3 drone) at the campus of Manipal Institute of Technology, Manipal, India. We refer to this dataset as Manipal UAV Anomalous Activity Dataset (MUAAD). These videos are acquired at 29 fps and are of resolution 1024×720 pixels. The videos are collected during different time of the day at which maximum vehicular traffic is expected (morning 9am, afternoon 12 noon and evening 4pm). The minimum and maximum duration of the videos are 7 seconds and 1 minute 30 seconds respectively. The videos are collected at an altitude of 25–30 metres. In total, the dataset contains 60 videos collected from nine different locations within the campus.

In the present study, seven general anomalous patterns are considered for annotations, viz. : vehicles parked in a no-parking zone, over-speeding vehicles, frantic trajectories of vehicles, vehicles found on non-pavement areas, pedestrians walking on the road, pedestrians gathering on-road and frantic movements of pedestrians. A few examples of these patterns are highlighted and shown in Figure 2. Similar to the standard practice followed in[4], a frame-level annotation (normal/anomalous) for each frame in the video is provided for a total of 68,687 frames. All the anomalous frames are labeled as 1 and normal frames are labeled as 0. The dataset contains a total of 92 anomalous events distributed in the 60 videos. A summary of the proposed dataset is given in Table 1 and 2. The training and the testing set consists of 8 and 52 videos respectively. The dataset is available at *https://github.com/uverma/MUAAD*. Our dataset explicitly accounts for occlusion challenges by including sequences where anomalous events are partially or fully obscured before reappearing. Despite the dataset containing only 92 anomalous events across 68,687 frames, these events span multiple frames, reflecting real-world rarity. The MUAAD dataset encompasses a diverse range of anomalous events relevant to UAV-based surveillance while maintaining a focus on generic anomaly detection.

The existing anomaly detection datasets contains videos acquired from static CCTV camera at fixed locations[8–10,19]. The videos present in these dataset contain very little or no camera motion. Beside, there is no variation in the background scene information in these videos. As discussed in[1], there is a need to develop

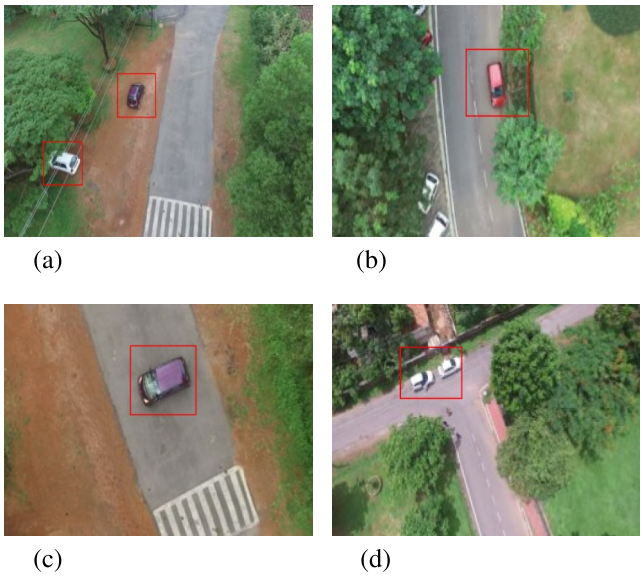


Fig. 2. Few frames from the dataset which contain anomalous event. The anomalous object is shown as a red bounding box. An example of contextual anomaly is shown in (**a**), while an example of vehicle parked on road is shown in (**b**). Figure (**c**) shows an example of vehicle with random trajectories, and (d) shows an example of vehicle parked in no-parking zone.

Dataset	Total No. Of Videos	Annotated frames	Anomalous events	Multi-scene	Camera motion	Modality
CHUK Avenue[8]	37	30,652	47	No	Limited camera shake	CCTV
ShanghaiTech[49]	437	317,398	130	\checkmark	-	CCTV
UCSD Ped1[50]	70	14,000	40	-	-	CCTV
UCSD Ped2[51]	28	4,560	12	-	-	CCTV
MUAAD (Ours)	60	68,687	92	\checkmark	\checkmark	UAV

Table 1. Comparison of MUAAD with other datasets.

Training videos	8
Testing videos	52
Number of scenes	9
Objects considered	Vehicle, Humans
Number of annotated frames	68,687
Types of anomalous events	7
Occurrence of anomalous events	92
Altitude	25–30 mts
Resolution	1280×720p
Frame rate	29.97 frames/second
Minimum duration	7 seconds
Maximum duration	1 min 30 seconds
Camera motion	Present
Illumination variation	Present

Table 2. Details of the proposed dataset.

anomalous activity detection algorithms in multi-scene videos with significant camera motion. The present study attempts to address this research gap by proposing a new multi-scene moving camera dataset. This dataset (MUAAD) contains videos acquired from the camera on-board a moving UAV and collected at 9 distinct *locations* (multi-scene). A few frames from this dataset is shown in Figure 3. It can be observed that there is significant variation in the background information along with the camera motion. Moreover, the videos acquired from

Fig. 3. Shows a few sample variations in scene of the proposed UAV anomaly detection dataset.

UAV contains a different perspective (topdown) view of the object (vehicles/human) as compared to the front view captured by the CCTVs.

Methodology

Proposed model

This study proposes an object-centric multi-scene video anomaly detection algorithm for UAV surveillance videos. Figure 1 shows the overview of the proposed UAV based anomaly detection system. In this work, an object-centric approach is adopted, since anomalies are related to the objects (human, vehicles) in the scene.

The workflow of the proposed system is as follows: The input to the proposed model is a UAV aerial video frame. Initially, an object detector is used to detect all the objects in the given video frame. Further, temporal and appearance features are extracted for every instance of the object of interest (vehicles/humans) detected in the frame. Also, the given input frame is semantically segmented to capture the contextual information around the detected objects. Finally, the contextual features along with the temporal and appearance features are given as input to the inference algorithm. The inference algorithm assigns an anomaly score for every instance of the object of interest (human/vehicle) present in the scene. The object level score is assigned as the frame level anomaly score if a single object of interest is present in the frame. In case of multiple objects present in the frame, the frame-level anomaly score is thus estimated as the maximum of the anomaly score assigned to these objects.

Object detection

Normally, anomalous events are associated with the objects in the scene[5,15]. Therefore, recent work have focused on object of interest in the scene to identify the presence/absence of anomalous events. Moreover, the detection of an anomalous event linked to an object in the scene allows us to identify the location of the anomalous event in the scene[5].

In this work, two classes of objects namely vehicles and humans are considered for anomaly detection. In this study, YoloV3[52] is used to detect these objects in the given scene. YoloV3 is trained from scratch on MUAAD dataset to detect humans and vehicles in each frame of the video.

Feature extraction

An anomalous event is generally categorized into three groups namely contextual, temporal and appearance anomaly. Hence, for an accurate prediction of anomalous patterns, it is important to extract contextual, temporal and appearance features. Moreover, in a multi-scene scenario such as videos acquired from a moving UAV, the definition of anomaly depends on the context of the scene. Despite this fact, methods proposed in the literature ignore contextual features. This limits the application of developed methods to multi-scene scenarios. Hence, in the present study, a novel feature extractor is proposed which considers contextual, temporal and appearance features. It is observed that so far these features have not been considered holistically for detecting anomalies in video sequences. However, in the present analysis, for each of the detected object of interest q (vehicle, humans), the contextual (f_q^c), temporal (f_q^t) and appearance (f_q^a) features are extracted. The final feature descriptor F_q for the object of interest q is the combination of f_q^c, f_q^t and f_q^a and is of dimension 22.

$$F_q = \{f_q^c, f_q^t, f_q^a\} \tag{1}$$

The process for extraction of these features for a particular object of interest q is explained below. Note that for the sake for simpler notation, f^c, f^t and f^a refers to the contextual, temporal and appearance features respectively for a particular object of interest q unless otherwise specified.

Temporal features (f^t): The proposed temporal feature extractor relies on the auto-encoder based reconstruction error computed from the optical flow. Initially, the dense optical flow[53] of the bounding box corresponding to the object of interest is estimated from the two consecutive image frames. These dense motion vectors are represented as RGB color image patches. Further, these RGB image patches are given as input to the auto-encoders. The auto-encoders converts the input image to latent space and back to the image domain using an unsupervised learning method. A few sample images of dense optical flow and the corresponding reconstructed flow is shown in Figure 5. Here, the auto-encoders are trained on normal motion patterns.

Hence, these auto-encoders fails to reconstruct those motion patterns which deviates from normal patterns thus resulting in a higher reconstruction error. The reconstruction error is computed as the absolute difference between the original input RGB image and the reconstructed image (output of auto-encoder) for each individual color channel (E_r^t, E_g^t, E_b^t). The proposed auto-encoder for temporal feature extraction has 4 encoding layers and 3 decoding layers. The overview of the proposed auto-encoder is shown in Figure 6. Each layer has a convolution layer followed by a batch normalization layer and an activation layer. In the present study, a filter size of 3x3 is utilized. ReLU activation function is used. In the end, softmax layer is applied to reconstruct the input image. Further, first-order statistical features such as mean (S_1^t), variance (S_2^t), kurtosis (S_3^t), energy (S_4^t), skewness (S_5^t) and entropy (S_6^t)[54] of the reconstructed image patch along with the reconstruction error (E_r^t, E_g^t, E_b^t) produced by the model are considered as the temporal feature vector for the object and is given by:

$$f^t = \{E_r^t, E_g^t, E_b^t, S_1^t, S_2^t, S_3^t, S_4^t, S_5^t, S_6^t\} \tag{2}$$

Appearance features (f^a): An appearance feature extractor is defined similar to the temporal feature extractor. In particular, the appearance feature extractor consists of an auto-encoder whose input is the image patch corresponding to object of interest. The model is trained to reconstruct the normal objects. Hence, any new/anomalous objects will produce a higher reconstruction error. A few samples of detected objects and their reconstructed image is shown in Figure 4. The reconstruction error in the RGB (E_r^a, E_g^a, E_b^a) color plane along with first-order statistical features mean (S_1^a), variance (S_2^a), kurtosis (S_3^a), energy (S_4^a), skewness (S_5^a) and entropy (S_6^a) is considered as appearance feature vector which is given as follows:

$$f^a = \{E_r^a, E_g^a, E_b^a, S_1^a, S_2^a, S_3^a, S_4^a, S_5^a, S_6^a\} \tag{3}$$

Contextual features (f^c): Semantic segmentation algorithms assigns pixel level labels to each individual pixels in the image and are widely used to realize the context of the scene[55,56]. A similar approach is considered to extract contextual information required for anomaly detection. UVid-Net proposed in[56], is used to semantically segment the given UAV aerial video frame. UVid-Net is the state-of-the-art algorithm for semantic segmentation of UAV aerial videos. A brief overview of UVid-Net is provided below: UVid-Net is an encoder-decoder based architecture which incorporates temporal information for semantic segmentation for aerial videos. The two consecutive keyframes as the input to the encoder ensures that the segmentation output contains temporally consistent labels without the need for any additional sequential module. Besides, a modified decoder module consisting of multiplication operation instead of concatenation produces a much finer segmentation result. More details about this architecture can be found in[56].

In this work, UVid-Net is trained to segment the given frame into four classes namely greenery, construction, roads and water bodies. These four broad semantic classes help in modeling the background scene information which can be utilized to learn context of the scene.

Generally, the bounding box produced by the object detectors are small and tight around the object. Also, context of the object is determined by its surrounding pixels. Hence, in this work, a small region R_q(4 pixel width) around and within the bounding box of the detected object is considered for extracting contextual information. Further, the total number of occurrence of the four class labels (greenery, construction, roads and water bodies) is computed in the identified small region R_q. This histogram of class labels defines the context around the object and is considered as the contextual feature. For instance, in case of object situated on the road, pixels belonging to the road class will have a maximum count in the histogram (Figure 7). Let us represent the four bins of the histogram as H_1, H_2, H_3, H_4 corresponding to the total number of the occurrence of the pixels belonging to greenery, road, construction and water classes respectively in the small region R_q. The contextual feature vector is then defined as:

Fig. 4. Appearance Feature Extraction: A few sample input images (top row) and its corresponding reconstructed images (bottom row). Note that a large reconstruction error is observed in few images representing abnormal event.

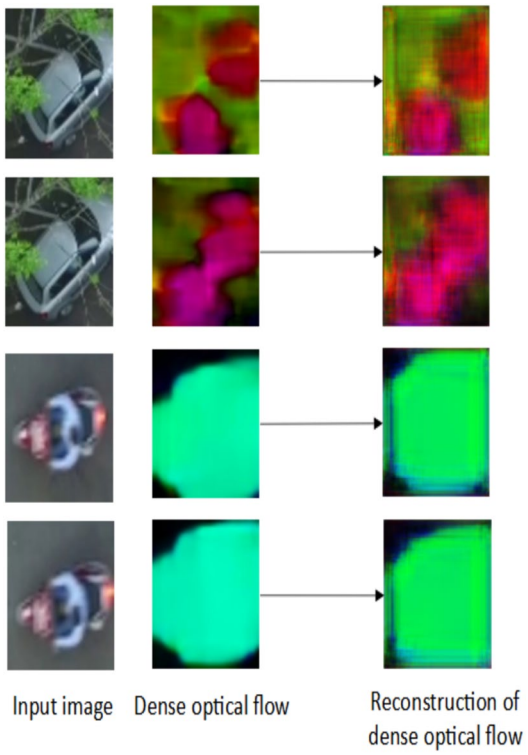


Fig. 5. Temporal Feature Extraction: Reconstruction of dense optical flow for temporal feature extraction. Note that a large reconstruction error is observed for objects associated with anomalous event.

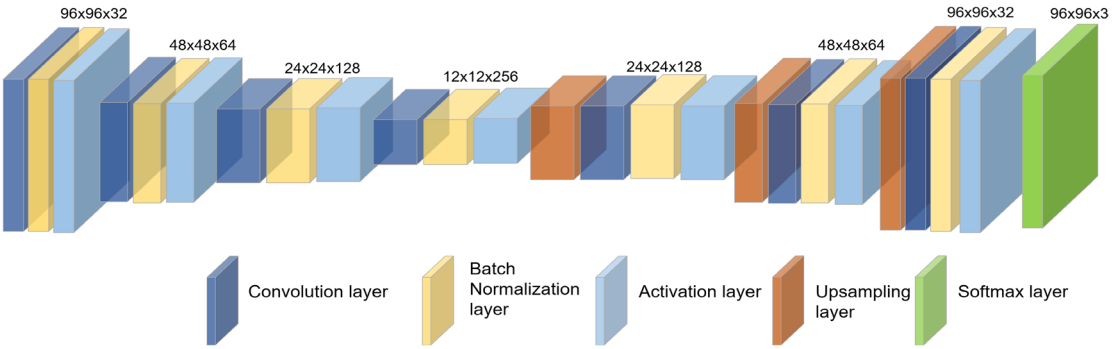


Fig. 6. Architecture of the auto-encoder.

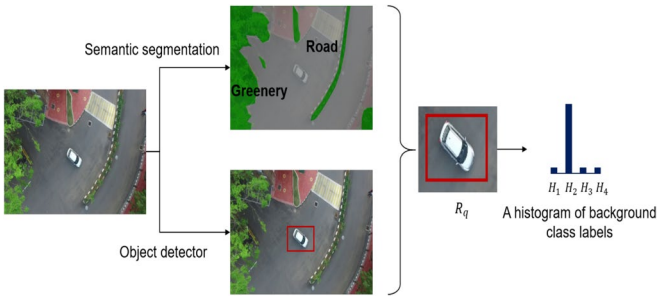


Fig. 7. Contextual Feature Extraction: The distribution of pixels belonging to four background classes is utilized for representing contextual information around the object of interest.

$$f^c = \{H_1, H_2, H_3, H_4\} \tag{4}$$

Model training

This section describes the training procedure adopted for training the object detector, auto-encoders and UVid-Net for object detection and feature extraction. It may be noted that object detector, temporal, appearance and contextual feature extractor modules are trained independently of each other as described below. In the present study, two generic classes of objects (vehicles, humans) are considered for detecting anomalies.

Object detector: The YoloV3[52] model is trained on UAV aerial images to detect vehicles and humans. To this end, 300 UAV aerial images are selected from MUAAD dataset and annotations are provided for objects vehicles and humans. These images and annotations are further utilized for training the YoloV3 model. The K-Means algorithm is used to determine the sizes of 9 anchor boxes. The model is trained for 250 epochs with Adam optimizer and learning rate set to 0.01.

Feature extractor: From the training data, objects representing normal patterns are manually cropped to create a training set for appearance auto-encoder. In addition, dense optical flow for these cropped images is estimated to create a training set for temporal auto-encoders. The two auto-encoders are trained separately using Binary cross-entropy loss function and Adam optimizer. Also, data augmentation is employed to increase the training dataset size and prevent the model from overfitting. Various image transformations such as flipping, rotation, translation and shearing are utilized to augment data. Subsequently, these augmented data is utilized for training the auto-encoders. For contextual feature extractor, the UVid-Net model is trained on ManipalUAVid dataset[56] to semantically segment the given aerial image into four classes. The categorical-cross-entropy loss is used with Adam optimizer to train the UVid-Net.

In this study, YoloV3[52] and UVid-Net[56] were employed for object detection and semantic segmentation, respectively, without structural modifications. However, both models were fine-tuned on our proposed UAV anomaly detection dataset, which presents unique challenges such as dynamic backgrounds and varying perspectives. Careful optimization of hyperparameters, including learning rates and augmentation strategies, further enhanced their generalization capability in UAV-based scenarios.

Inference

In the present study, a new inference algorithm is proposed for detecting anomalies. In the literature, it is observed that the anomaly detection models are solely trained on normal patterns due to the rarity of abnormal events. However, these algorithms may fail to detect local anomalies since the decision boundaries are determined based on normal patterns exclusively. The local anomaly is referred to as the anomalous events which closely resembles a normal event. Hence, in the present study, a learning strategy is proposed to address the issue of local anomalies. In this strategy, the inference algorithm is trained on a larger set of normal patterns and a limited set of anomalous patterns. The inclusion of a smaller set of anomalous patterns allows the model to identify better decision boundaries and aids in improved accuracy.

The proposed method is inspired by the inference algorithm presented in[5]. In[5], the authors proposed to train *K* SVM classifiers on *K* clusters of normal patterns to create a model for *K* normal events. Subsequently, the score from binary classifiers in one-vs-rest scheme is utilized to determine the anomalous events. The use of one-vs-rest scheme creates artificial dummy abnormal events (one normal vs others dummy abnormal events). However, this approach considers only normal patterns to identify decision boundaries. In this work, we propose incorporating a limited set of anomalous event in multi-class classification for anomalous activity detection to learn better discriminative decision boundaries. This work first train two disjoint sets of SVM classifiers, one trained only on normal events and other trained only on abnormal events. During inference, the maximum classification scores from these two sets of classifiers for the test sample are compared to identify the anomalous events.

The 22 dimension feature vector extracted by the proposed feature extractor is used to train the inference algorithm. Let the given training set *T* consists of *M* normal samples and *N* anomalous samples, where $N << M$. Further, the K-Means algorithm is employed to cluster the *M* normal samples and *N* anomalous samples into K^1 and K^2 groups respectively. Subsequently, we initialize an SVM classifier for each of the $K^1 + K^2$ clusters which results in $K = K^1 + K^2$ SVM classifiers. The SVM classifier initialized for normal patterns is trained to classify normal patterns while SVM classifiers trained on anomalous patterns are trained to classify anomalous patterns. While training i^{th} SVM classifier, the samples belonging to other clusters $K - 1$ clusters are considered anomalous patterns. Hence, each SVM classifier is trained as a binary classifier. Since these classifiers are trained on the training set *T* which consists of both normal and anomalous patterns, the decision boundaries identified will be tuned to separate local anomalies. Note that the sample here refers to the feature vector corresponding to a particular object.

During the test time, for every detected object in the given frame, we extract the feature vector *F*. This feature vector is classified by *K* SVM classifiers. Further, we calculate the scores as follows:

$$\alpha = max(m_1, m_2, ...m_{k^1}) \tag{5}$$

$$\beta = max(n_1, n_2, ...n_{k^2}) \tag{6}$$

here, α represents the maximum score of K^1 SVM classifiers of normal patterns, while β represents the maximum score of K^2 SVM classifiers of anomalous patterns. A value of α significantly higher than β indicates that the object represents an normal event, while $\alpha < \beta$ indicates an abnormal event corresponding to the object. Therefore, the samples satisfying the following criterion is regarded as normal objects:

$$\alpha > \beta \quad AND \quad \alpha > \mu \tag{7}$$

where μ is a parameter determined experimentally and *AND* is a logical AND operator. Similarly, the samples satisfying the following condition is considered as abnormal objects:

$$\alpha < \beta \quad AND \quad \beta > \nu \tag{8}$$

where ν is a parameter determined experimentally and *AND* is a logical AND operator. Moreover, if any samples does not satisfy any of the above two conditions (Eqn 7 and 8), it is regarded as anomalous object. The above conditions ensure that the object which has been classified by one of the K^1 classifiers with a high confidence (high α) is regarded as an object associated with normal event. Similarly, the object for an abnormal event would be classified by one of the K^2 classifiers with high confidence (high β). Moreover, it is not possible to model all the possible anomalous activities. Therefore, an object classified as anomalous with less confidence (low β) is also considered as anomalous event.

The final frame level anomaly score is computed as the maximum of all the score obtained by the object present in the frame.

In general, a new test sample may belong to one of the three groups namely *normal pattern*, *anomalous pattern* and *unknown pattern*. Unknown patterns are those patterns that are new to the model and may be anomalous. Hence, these patterns should also raise an alarm. Moreover, the sensitivity of anomaly detectors to normal and anomalous patterns plays an important role in providing security. In this regard, this paper presents a simple logic to control the sensitivity of the model to normal and anomalous patterns. Here, the threshold value μ and η are used to regulate the model sensitivity to normal and anomalous patterns respectively. A given test sample is considered as a normal pattern if $\alpha > \beta$ and $\alpha > \mu$. Further, the given test sample is considered anomalous if $\alpha < \beta$ and $\beta > \eta$. Else, the test sample is considered as an unknown pattern. By increasing the value of μ and η, the model's responsiveness to the normal and abnormal patterns reduces. Hence, these parameters can be used to control the effectiveness of the anomaly detector depending on the application. The selection of these parameters are discussed in section 5.2.

Results and discussion

The proposed algorithm is evaluated on the MUAAD dataset containing the videos acquired from a moving UAV at multiple locations. As discussed earlier, the existing datasets (Ped1, Ped2, ShanghaiTech) focuses on single scene videos acquired from static cameras with little or no camera motion. Since, the proposed algorithm is designed for a multi-scene scenario, it will not be prudent to evaluate the performance of the proposed algorithm on existing single scene datasets. Moreover, the existing datasets does not provide pixel level labels required for extracting contextual information.

This section first presents the performance metrics (Section 5.1), and the approach utilized to select the parameters of the proposed approach (Section 5.2). Subsequently, the performance of the proposed contextual feature extractor is studied in Section 5.3, while the performance of the proposed inference algorithm is discussed in Section 5.4. Finally, the proposed approach is compared with the existing methods in Section 5.5.

Performance metric

Several recent works of literature used Area under the ROC Curve (AUC) metric to evaluate the performance of anomaly detector[5,15,57,58]. Following their procedure, in the present study, the AUC metric is used for evaluating the performance of the proposed method. In this method, the frame-level anomaly score is compared against the ground truth frame level annotation, to compute the AUC. Specifically, the frame level AUC is computed from the frame level total positive rate and false positive rate. As discussed earlier, a frame is considered anomalous if it contains at least one object with abnormal activity.

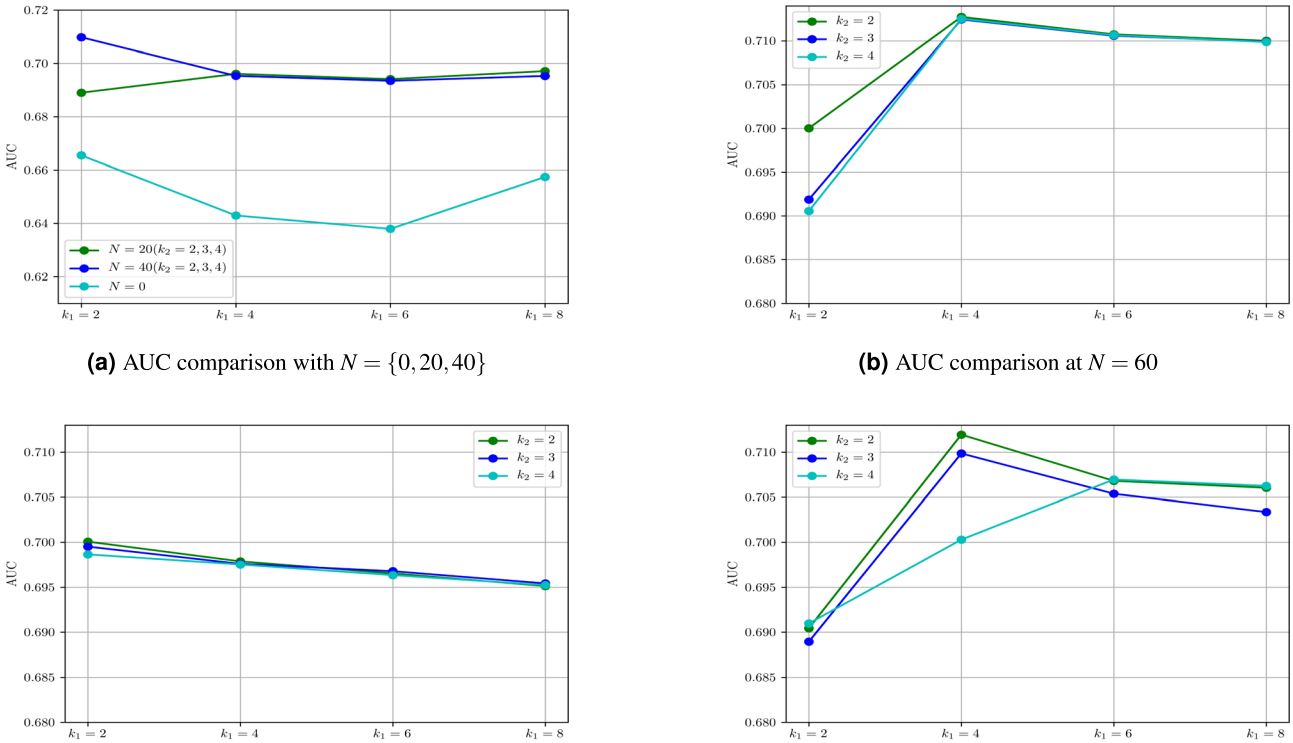
Parameter selection

This section discusses the selection of parameters used in the proposed model such as the number of normal K^1 and abnormal K^2 clusters and the thresholds μ and η. It may be noted that a normal/anomalous sample for the inference model refers to the feature vector associated with normal/anomalous object of interest.

Selection of K^1 and K^2: The parameters K^1 and K^2 represent the number of clusters in normal and anomalous training sample and has an influence on the performance of the model. The values for these two parameters are determined experimentally. The AUC metric is studied for different pair of values of K^1 and K^2. The K^1 value is selected from the set $\{2, 4, 6, 8\}$ while K^2 value is selected from the set $\{2, 3, 4\}$and a grid search like method is employed to find the optimal value. It may be noted that a lower values of the number of anomalous cluster K^2 was considered due to limited number of anomalous samples present in the dataset. The maximum AUC of 0.712 was observed for $K^1 = 4$ and $K^2 = 3$(Figure 8b).

Given the limited number of anomalous samples (N) available in the dataset, we also studied the effect on K^1 and K^2 due to variation in the number of anomalous samples N. Figures 8a, 8b, 8cand 8dshows the AUC obtained by varying N, K^1 and K^2. It can be observed that a maximum AUC of 0.712 is obtained for $N = 60$, $K^1 = 4$ and $K^2 = 3$. This AUC is slightly higher than $N = 100$, $K^1 = 4$ and $K^2 = 2$. Also, it is observed that the AUC obtained for $K^2 = \{2, 3, 4\}$ are closer by at $N = 60$ and $k^1 = 4$. Therefore, in the present study, K^1 and K^2 value is determined to be 4 and 3 respectively.

Selection of μ and η: The parameters μ and η(Equations 7, 8) determines the sensitivity of the proposed inference model. A given pattern is considered normal of the the maximum score of normal SVM classifiers (α) is greater than the maximum score of abnormal SVM classifiers (β) and the threshold μ. Consequently, regulating the μ value allows us to control the sensitiveness of the model to normal patterns. For example, a



(a) AUC comparison with $N = \{0, 20, 40\}$

(b) AUC comparison at $N = 60$

(c) AUC comparison at $N = 80$

(d) AUC comparison at $N = 100$

Fig. 8. AUC comparison of proposed method with different values of N, K_1, and K_2.

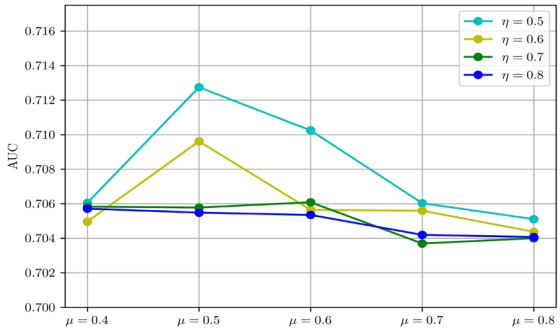


Fig. 9. AUC comparison of proposed method with different values of μ and η.

higher value of μ, determines the given pattern as normal if the confidence score (α) is greater than μ. This allows the model to identify normal patterns with more confidence. On the contrary, reducing the μ value allows the model to determine normal patterns with a lower degree of confidence. Similarly, threshold η controls the model's sensitivity to anomalous patterns. Hence, μ and η values should be selected carefully.

In the present study, μ value is selected from set $\{0.4, 0.5, 0.6, 0.7, 0.8\}$. Concerning anomaly detection, it is always desirable to have a model with high sensitivity to anomalous patterns. Hence, in the present study, η is selected from a lower set of threshold values $\{0.5, 0.6, 0.7, 0.8\}$. The grid search method is used to determine the optimal value of μ and η. This experiment is conducted by setting the values of N, K^1 and K^2 to 60, 4 and 3. The AUC of the proposed model with different values of μ and η is estimated and is plotted in Figure 9. It was found that the AUC of the model reaches 0.712 when μ is set of 0.5 and η set to 0.5. Since the model is trained on a smaller set of anomalous patterns, reducing the sensitivity of the model to anomalous patterns (increasing the η) reduces the AUC of the model. Hence, for a higher value of η, the AUC of the model reduces. In the present study, μ and η values are determined to be 0.5 and 0.5 respectively.

Evaluation of contextual knowledge

The performance of anomaly detection algorithms in a multi-scene scenario depends on contextual knowledge about the scene. The proposed work uses appearance features to realize the appearance change of the object

Proposed method	AUC
With contextual knowledge	**0.712**
Without contextual knowledge	0.6753

Table 3. AUC comparison of proposed method with and without contextual knowledge.

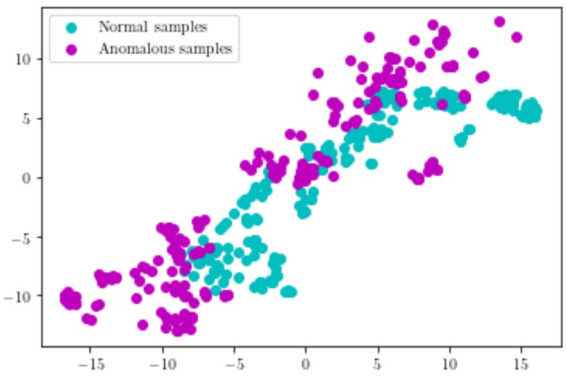


Fig. 10. Plot of normal and anomalous feature vectors without contextual knowledge.

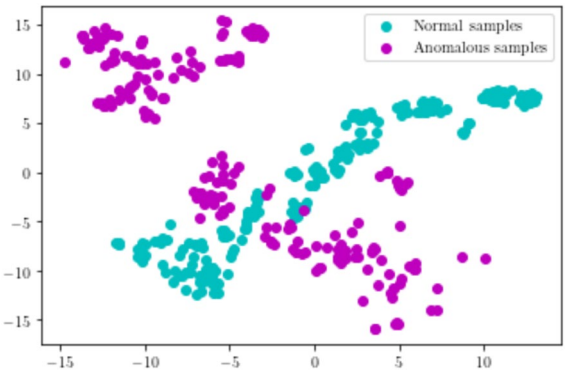


Fig. 11. Plot of normal and anomalous feature vectors with contextual knowledge.

while contextual feature builds a summary of the background scene information around the object. To this end, in the present study, a given UAV aerial scene is semantically segmented to realize the layout of the scene. Subsequently, a histogram of class labels is calculated around the object. Finally, this histogram is considered as a feature vector along with the appearance and temporal features to describe the object. This feature vector is further given to the inference algorithm to decide on the anomaly. In the present study, an experiment is executed to evaluate the influence of contextual knowledge on the performance of anomaly detection. In this experiment, the performance of the proposed model is compared *with* contextual information and *without* contextual information. Initially, the performance of the model is evaluated using appearance and temporal features. Subsequently, the performance of the model is evaluated using contextual, appearance and temporal features. In this experiment, the proposed inference algorithm is utilized to infer the anomalies. The AUC of the proposed model with and without contextual information is shown in Table 3. It was found that there was a significant improvement in AUC (0.712 vs 0.675) of the proposed model with contextual knowledge. This result is significant as it demonstrates the importance of contextual knowledge in a multi-scene scenario.

To qualitatively evaluate the importance of contextual knowledge, we visualized the feature vector with and without contextual knowledge. A total of random 400 sample patterns are considered among which 200 belongs to normal patterns and 200 belongs to anomalous patterns. Let the feature vector with contextual knowledge be represented as F_1 and without contextual knowledge be represented as F_2. The dimension of the considered feature vectors F_1 and F_2 are 22 and 18 respectively. Dimensionality reduction is essential to visualize these feature vectors. To this end, Principal Component Analysis (PCA) is employed to reduce the feature dimensions of F_1 and F_2 to two Subsequently, we plotted the features in 2D space. This is shown in Figure 10 and Figure 11. From these figures, it is observed that the feature vector F_2 discriminates normal and anomalous patterns efficiently. Hence, the inclusion of contextual knowledge aids one-class SVMs to determine better decision boundaries. This further substantiates the improvement in AUC of the model with contextual information. This

experiment demonstrates that importance of contextual knowledge in the detection of anomalous patterns from the multi-scene scenario.

Further, we compared the anomaly score at the frame level of the proposed model with and without contextual knowledge. Figure 12, shows the frame-level AUC comparison of the proposed model on few multi-scene aerial video samples. These video samples constitute contextual anomalies such as stationary vehicles on the greenery, road. It is observed that the proposed model with contextual knowledge produces a higher anomaly score for anomalous frames. However, the proposed model without contextual knowledge produces a lower anomaly score. This result is anticipated since appearance and temporal features alone are not adequate enough to detect contextual anomalies. However, the proposed feature extractor which encapsulates contextual, appearance and temporal features holistically, can effectively detect these anomalies. Besides, the inclusion of contextual information helps in better identification of anomalous frames as compared to the model without any contextual information. It can be observed in Figure 12 that a higher number of anomalous frame are correctly classified as anomalous by incorporating contextual information. This experiment demonstrates the importance of contextual knowledge in multi-scene anomaly detection.

Evaluation of inference algorithm

A majority of existing woks utilizes only normal pattern to model the normal events. Hence, the decision boundaries are tuned to identify normal patterns and may fail to detect local anomalies. For example, in the case of UAV aerial videos, a car parked in the parking area and on the road may have similar appearance characteristics. However, a car parked on road is an example of local anomalies since its attributes are closer to normal patterns. Hence, an inference algorithm trained only on normal patterns may not be sufficient to distinguish these local anomalies. In the present study, we argue that training the model on few samples of known anomalous patterns can significantly improve the performance of the model. In this context, we visualized the decision boundaries of K^1 and K^2 SVM classifiers. To this end, we considered 300 normal samples and 100 anomalous samples. Figure 14ashows the scatter plot of considered samples. Few examples of local anomalies are highlighted in RED. An inference algorithm trained only on normal patterns may not distinguish these samples (14d, 14c). However, we can observe that the decision boundaries identified by K^1 SVM classifiers are effectively separating local anomalies since they are trained using both normal and few anomalous samples (14b). Hence, the proposed model can classify local anomalies more accurately. This result validates the assumption of considering few anomalous samples in identifying the decision boundaries.

Evaluating the influence of number of anomalous events

The rarity of anomalous samples poses a significant challenge for the development of anomaly detection algorithms. The proposed work utilizes few samples of anomalous patterns along with normal patterns to determine the decision boundaries. The inclusion of few anomalous samples allows the SVM classifiers to determine accurate decision boundaries thereby improving anomaly detection. In this context, an experiment is

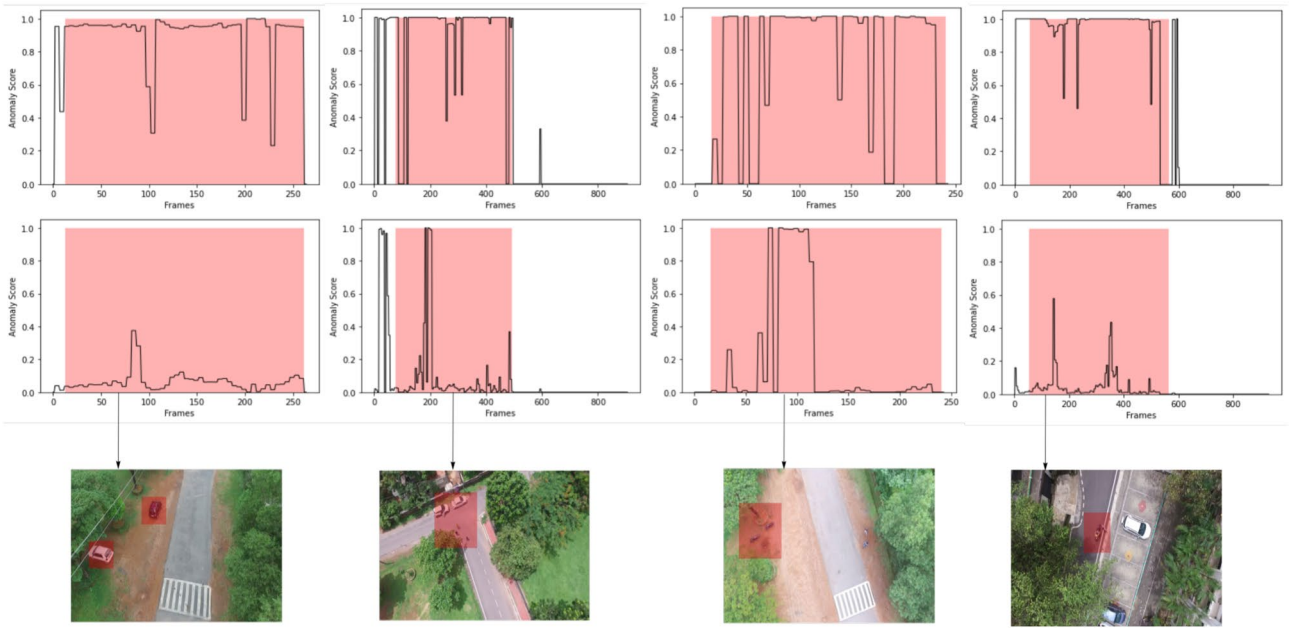


Fig. 12. Anomaly score comparison of proposed model with and without contextual knowledge. The red region in the plot indicates the anomalous frames. The first row shows the frame-level anomaly score of the proposed model with contextual knowledge. Second row shows the frame-level anomaly score of the proposed model without contextual knowledge. The bottom row shows a sample frame from the video with the anomalous object.

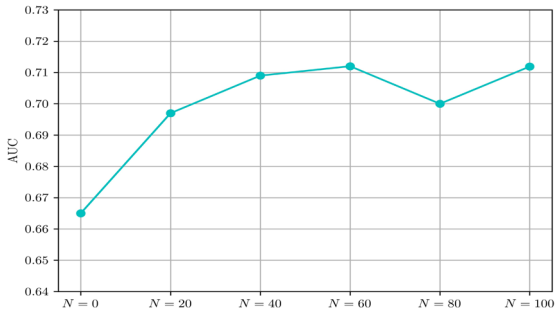
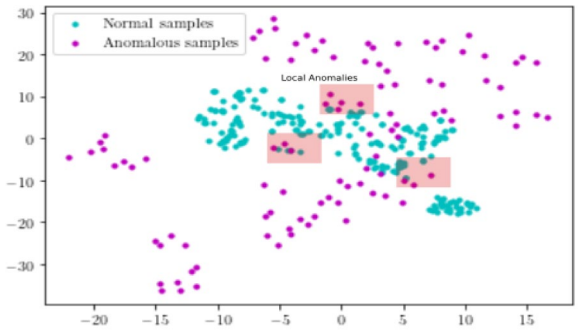
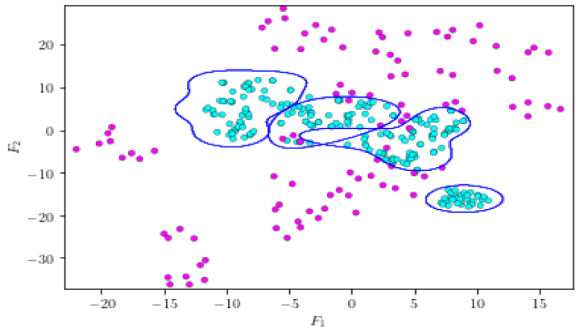


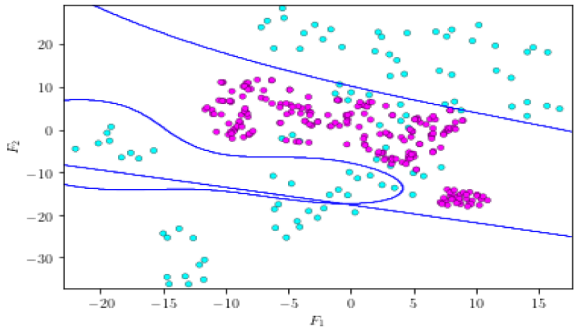
Fig. 13. AUC comparison of proposed model with different values of N. The best K_1 and K_2 values are considered for the comparison of the performance.



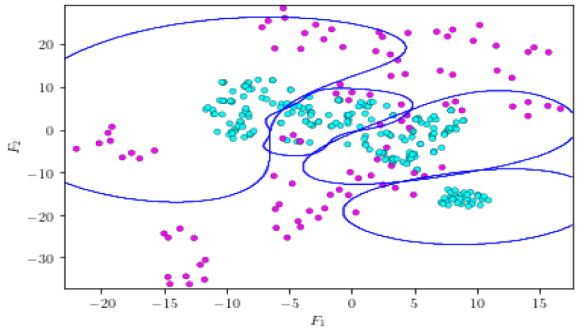
(a) Scatter plot of normal and anomalous samples



(b) SVM trained on normal + few anomalous samples



(c) SVM trained on anomalous samples



(d) SVM trained on normal samples only

Fig. 14. Visualization of decision boundaries of SVM classifiers.

performed to learn the influence of anomalous samples on anomaly detection. In this experiment, the proposed inference algorithm is evaluated with different values of number of anomalous events (N). The N value is selected from the set $\{0, 20, 40, 60, 80, 100\}$. The M value (number of normal training samples) is set to 300. In every iteration of the experiment, N anomalous training samples are selected randomly. The grid search approach is utilized to determine the parameters of SVM classifiers. Grid search is a hyperparameter tuning method that implements exhaustive searching to determine the hyperparameters. Figure 13, shows the AUC of the proposed method with different values of N. At $N = 0$, the model is trained only on normal samples. Hence, the decision boundaries may fail to accurately classify anomalous samples. Consequently, it is seen that the proposed method achieves an AUC of 0.665 at $N = 0$. However, as the N value increases, the AUC of the algorithm increases and reaches an AUC of 0.712 at $N = 60$. An important observation here is that at $N = 20$, the model achieved an AUC of 0.697 which is notably greater than the AUC (0.665) obtained at $N = 0$. This is a significant result since it demonstrates the importance of a *few anomalous samples* in identifying the decision boundaries for accurate anomaly detection.

As discussed earlier, the proposed inference algorithm is inspired by the work of[5] which formulated abnormal event detection as multi-class classification problem. Therefore, we also evaluate the proposed inference algorithm by comparing it with the inference algorithm proposed in[5]. In this experiment, we utilized the proposed feature extractor to extract contextual, temporal and appearance features. Subsequently,

Inference method	AUC
Ionescu et al.[5] (Appearance features)	0.587
Ionescu et al.[5] (Temporal features)	0.6028
Ionescu et al.[5] (Contextual features)	0.566
Ionescu et al.[5] (Proposed feature extractor)	0.648
Ours (Appearance features)	0.5973
Ours (Temporal features)	0.6332
Ours (Contextual features)	0.5513
Ours (Proposed feature extractor)	**0.712**

Table 4. AUC comparison of proposed model with different inference algorithm feature extractor.

Algorithm	AUC
Ionescu et al.[5]	0.568
Doshi et al.[15]	0.641
Li T et al.[60]	0.359
Astrid et al.[59]	0.487
Ours	**0.712**

Table 5. AUC comparison of proposed and other methods.

the performance of proposed and existing[5] inference algorithms are studied for inferring the anomaly. Note that the existing inference algorithm is only trained on normal patterns with the number of clusters set to 5. The proposed inference algorithm is trained on normal and known anomalous patterns with K^1 set to 4 and K^2 set to 3. Table 4 shows the AUC results of the two inference algorithm on the proposed UAV anomaly detection dataset with different feature extractors. It is found that the temporal features produced higher AUC as compared to contextual and appearance features on the existing and proposed inference algorithms. This result is justified by the fact that the MUAAD dataset has higher temporal anomalies due to object motion. Further, it is seen that contextual features alone are not sufficient to detect anomalies in UAV aerial videos. However, by considering the appearance, temporal and contextual features holistically, the performance of the existing and the proposed inference algorithm improved significantly. Specifically, the proposed and the baseline inference algorithm achieved an AUC of 0.712 and 0.648 respectively. It was found that the proposed inference algorithm produced a significant improvement in AUC (7%) as compared to[5]. This finding confirms that the decision boundaries identified by the proposed inference algorithm are tuned to separate both normal and anomalous patterns. Furthermore, the proposed feature extractor has improved the performance of both proposed and baseline inference algorithm. This improvement is consistent and hence validates the proposed feature extractor for UAV anomaly detection.

Comparative study

In the present study, two semi-supervised anomaly detection algorithms are considered as baseline models. In the first method[5], the authors proposed to use auto-encoders to extract features. Subsequently, SVM based inference algorithm is utilized to infer the anomalies. In[15], the authors proposed a continual learning approach using the KNN algorithm to detect anomalies in videos. In addition, the proposed method is also compared with the approaches proposed in[59] and[60]. Both these methods are based on auto-encoders and are non-object centric. All these methods are designed for CCTV videos. Here it is important to note that, the feature extractor of the proposed model is improved upon the feature extractor proposed in[5] and[15]. Furthermore, the inference algorithm is designed based on the idea proposed in[5].

The AUC results of the proposed and compared methods on UAV anomaly detection dataset is given in Table 5. It was found that the methods proposed in[5] and[15] achieved an AUC of 0.568 and 0.641 respectively on UAV anomaly detection dataset. The methods proposed in[59] and[60] achieved an AUC of 0.487 and 0.359 respectively on MUAAD dataset. These methods[59,60] are designed for CCTV videos where the background is constant with little or no camera motion. However, UAV aerial videos contains significant camera motion. Hence, the temporal features extracted by these methods may produce false positives in the presence of camera motion. In addition, the methods proposed in[59] and[60] are non-object centric. Hence, a higher reconstruction error will be observed in a multi-scene scenario such as UAV aerial video. These high reconstruction error would be due to variation in background information and need not represent an actual anomalous event. This contributes in higher false positives and lower AUC. This result is significant as it highlights the effectiveness of object centric methods in a multi-scene scenario. The proposed **object-centric method** uses auto-encoders and optical flow to extract temporal features. Since the auto-encoder is trained on normal motion patterns, it produces in higher reconstruction error for anomalous patterns. Also, the proposed method calculates temporal features concerning the object bounding box which further reduces the influence of camera motion on temporal

features. It can be observe that the proposed method achieved an AUC of 0.712 which is significantly greater than baseline models.

Furthermore, the UAV anomaly detection dataset has videos taken from various locations (multi-scene scenarios). Also, each video constitutes camera motion. Hence, the context of the scene differs within a video. In these situations, realizing the context of the scene aids in improving the performance of anomaly detection. Despite this interest, the baseline models ignore contextual features. However, the proposed method captures the contextual knowledge required for context-aware anomaly detection. This result is significant since the definition of anomaly is dependent on the context of the scene which is generally ignored in the literature. The proposed method highlights the importance of contextual information required for the accurate detection of anomalies in multi-scene scenarios such as UAV.

Conclusion

This work is intended to develop a video anomaly detection algorithm for UAV aerial videos. A new multi-scene UAV anomaly detection dataset is proposed to address the lack of standard datasets for anomaly detection in UAV surveillance videos. The proposed dataset has scene variations and camera motion that provides a standard challenging platform for researchers to develop and evaluate their algorithms. Frame-level annotations are provided for 52 UAV aerial videos. Further, two baseline models are implemented and evaluated for validating the dataset.

This study also introduces a new UAV video anomaly detection algorithm that holistically uses contextual, temporal and appearance features to detect the anomalies. Furthermore, a novel inference algorithm is presented that uses a few-shot learning strategy to infer the anomalies. The proposed algorithm achieved an AUC of 0.712 which is significantly greater than the compared baseline models. The extensive evaluation of the proposed model revealed that contextual knowledge plays a significant role in multi-scene video anomaly detection. The inclusion of contextual knowledge leverages the performance of video anomaly detection algorithms. Furthermore, this study demonstrated that the incorporation of few known anomalous samples in the training process can prominently improve the performance of anomaly detection. This study highlights the importance of contextual knowledge and learning strategy for video anomaly detection. In future work, we aim to enhance the real-time feasibility of our method by integrating shot boundary detection algorithms and optimizing computational efficiency to achieve near real-time anomaly detection in UAV surveillance scenarios.

Data availability

The datasets generated and analysed during the current study are not publicly available due to privacy reasons but are available from the corresponding author on reasonable request.

Received: 25 November 2024; Accepted: 16 June 2025
Published online: 16 July 2025

References

1. Ramachandra, B., Jones, M. & Vatsavai, R.R. A survey of single-scene video anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
2. Chalapathy, R. & Chawla, S. Deep learning for anomaly detection: A survey.(2019) arXiv preprint arXiv:1901.03407.
3. Santhosh, K. K., Dogra, D. P. & Roy, P. P. Anomaly detection in road traffic using visual surveillance: A survey. *ACM Comput. Surv.***53**, 1–26 (2020).
4. Liu, W., Luo, W., Lian, D. & Gao, S. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6536–6545 (2018).
5. Ionescu, R.T., Khan, F.S., Georgescu, M.-I. & Shao, L. 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 (CVPR)* (2019).
6. Xu, D., Ricci, E., Yan, Y., Song, J. & Sebe, N. Learning deep representations of appearance and motion for anomalous event detection. arXiv preprint arXiv:1510.01553 (2015).
7. Ma, K., Doescher, M. & Bodden, C. Anomaly detection in crowded scenes using dense trajectories. *Univ. of Wisconsin-Madison* (2015).
8. Lu, C., Shi, J. & Jia, J. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, 2720–2727 (2013).
9. Mahadevan, V., Li, W., Bhalodia, V. & Vasconcelos, N. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1975–1981 (IEEE, 2010).
10. Adam, A., Rivlin, E., Shimshoni, I. & Reinitz, D. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence* **30**, 555–560 (2008).
11. Teng, S., Zhang, S., Huang, Q. & Sebe, N. Viewpoint and scale consistency reinforcement for uav vehicle re-identification. *Int. J. Comput. Vis.* **129**, 719–735 (2021).
12. Chriki, A., Touati, H., Snoussi, H. & Kamoun, F. Deep learning and handcrafted features for one-class anomaly detection in uav video. *Multimed. Tools Appl.* **80**, 2599–2620 (2021).
13. Bozcan, I. & Kayacan, E. Uav-adnet: Unsupervised anomaly detection using deep neural networks for aerial surveillance. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1158–1164, (2020) https://doi.org/10.1109/IROS45743.2020.9341790.
14. Chriki, A., Touati, H., Snoussi, H. & Kamoun, F. Uav-based surveillance system: an anomaly detection approach. In *2020 IEEE Symposium on computers and communications (ISCC)*, 1–6 (IEEE, 2020).
15. Doshi, K. & Yilmaz, Y. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 934–935 (2020).
16. Xu, D., Yan, Y., Ricci, E. & Sebe, N. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Comput. Vis. Image Underst.* **156**, 117–127 (2017).
17. Zhao, B., Fei-Fei, L. & Xing, EP. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, 3313–3320 (IEEE, 2011).
18. Cheng, K.-W., Chen, Y.-T. & Fang, W.-H. 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*, 2909–2917 (2015).

19. Raghavendra, R., Bue, A. & Cristani, M. Unusual crowd activity dataset of university of minnesota (2006).
20. Sultani, W., Chen, C. & Shah, M. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6479–6488 (2018).
21. Ramachandra, B., Jones, M. & Vatsavai, R. Learning a distance function with a siamese network to localize anomalies in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2020).
22. Sabokrou, M., Khalooei, M., Fathy, M. & Adeli, E. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
23. Ionescu, R. T., Smeureanu, S., Popescu, M. & Alexe, B. Detecting abnormal events in video using narrowed normality clusters. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1951–1960 (IEEE, 2019).
24. Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E. & Sebe, N. 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)*, 1689–1698 (IEEE, 2018).
25. Dalal, N., Triggs, B. spsampsps Schmid, C. Human detection using oriented histograms of flow and appearance. In *European conference on computer vision*, 428–441 (Springer, 2006).
26. Dalal, N. & Triggs, B. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, 886–893 (Ieee, 2005).
27. Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z. & Klette, R. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Comput. Vis. Image Underst.* **172**, 88–97 (2018).
28. Sabokrou, M., Fayyaz, M., Fathy, M. & Klette, R. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing* **26**, 1992–2004 (2017).
29. Antić, B. & Ommer, B. Spatio-temporal video parsing for abnormality detection. arXiv preprint arXiv:1502.06235 (2015).
30. Kim, J. & Grauman, K. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE conference on computer vision and pattern recognition*, 2921–2928 (IEEE, 2009).
31. Benezeth, Y., Jodoin, P.-M., Saligrama, V. & Rosenberger, C. Abnormal events detection based on spatio-temporal co-occurences. In *2009 IEEE conference on computer vision and pattern recognition*, 2458–2465 (IEEE, 2009).
32. Kratz, L. & Nishino, K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE conference on computer vision and pattern recognition*, 1446–1453 (IEEE, 2009).
33. Feng, Y., Yuan, Y. & Lu, X. Learning deep event models for crowd anomaly detection. *Neurocomputing***219**, 548–556 (2017).
34. Vu, H., Nguyen, T. D., Le, T., Luo, W. & Phung, D. Robust anomaly detection in videos using multilevel representations. *In Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 5216–5223 (2019).
35. Chong, YS. & Tay, YH. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*, 189–196 (Springer, 2017).
36. Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K. & Davis, L. S. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 733–742 (2016).
37. Tran, T. M., Bui, D. C., Nguyen, T. V. & Nguyen, K. Transformer-based spatio-temporal unsupervised traffic anomaly detection in aerial videos. *IEEE Transactions on Circuits and Syst. for Video Technology* **34**, 8292–8309. https://doi.org/10.1109/TCSVT.2024.3376399 (2024).
38. Pathirannahalage, I., Jayasooriya, V., Samarabandu, J. & Subasinghe, A. A comprehensive analysis of real-time video anomaly detection methods for human and vehicular movement. *Multimed. Tools Appl.* 1–46 (2024).
39. Ito, Y., Sasaki, T. & Kondo, S. Crowd anomaly detection from drone and ground. *IEEE Access* (2025).
40. Duja, K. U. & Khan, I. A. *& Alsuhaibani, M* (A review on deep learning benchmarks. IEEE Access, 2024).
41. Jiao, R., Wan, Y., Poiesi, F. & Wang, Y. Survey on video anomaly detection in dynamic scenes with moving cameras. *Artif. Intell. Rev.* **56**, 3515–3570 (2023).
42. Zhai, G. & Min, X. Perceptual image quality assessment: a survey. *Sci. China Inf. Sci.* **63**, 1–52 (2020).
43. Min, X., Duan, H., Sun, W., Zhu, Y. & Zhai, G. Perceptual video quality assessment: A survey. *Sci. China Inf. Sci.***67**, 211301 (2024).
44. Min, X. et al. Screen content quality assessment: Overview, benchmark, and beyond. *ACM Comput. Surv.***54**, 1–36 (2021).
45. Min, X., Zhai, G., Gu, K., Liu, Y. & Yang, X. Blind image quality estimation via distortion aggravation. *IEEE Transactions on Broadcasting* **64**, 508–517 (2018).
46. Min, X. et al. Blind quality assessment based on pseudo-reference image. *IEEE Transactions on Multimedia* **20**, 2049–2062 (2017).
47. Min, X., Zhai, G., Zhou, J., Farias, M. C. & Bovik, A. C. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Trans. Image Process.***29**, 6054–6068 (2020).
48. Min, X., Zhai, G., Gu, K. & Yang, X. Fixation prediction through multimodal analysis. *ACM Trans. Multimedia Comput. Commun. Appl.***13**, 1–23 (2016).
49. Liu, W., W.Luo, D.L. & Gao, S. Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).
50. Wang, S. & Miao, Z. Anomaly detection in crowd scene. In *IEEE 10th International Conference on Signal Processing Proceedings*, 1220–1223 (IEEE, 2010).
51. Li, W., Mahadevan, V. & Vasconcelos, N. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence* **36**, 18–32 (2013).
52. Redmon, J. & Farhadi, A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018).
53. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, 363–370 (Springer, 2003).
54. Sainju, S., Bui, F. M. & Wahid, K. A. Automated bleeding detection in capsule endoscopy videos using statistical features and region growing. *Journal of medical systems* **38**, 1–11 (2014).
55. Cordts, M. *et al.* The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3213–3223 (2016).
56. Girisha, S., Verma, U., Pai, M. M. & Pai, R. M. Uvid-net: Enhanced semantic segmentation of uav aerial videos by embedding temporal information. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **14**, 4115–4127 (2021).
57. Cong, Y., Yuan, J. & Liu, J. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, 3449–3456 (IEEE, 2011).
58. Liu, Y., Li, C.-L. & Póczos, B. Classifier two sample test for video anomaly detections. In *BMVC*, 71 (2018).
59. Astrid, M., Zaheer, M.Z. & Lee, S.-I. Synthetic temporal anomaly guided end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 207–214 (2021).
60. Li, T., Chen, X., Zhu, F., Zhang, Z. & Yan, H. Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection. *Neurocomputing* **439**, 256–270 (2021).

Acknowledgements

This research was supported by Manipal Academy of Higher Education for the dataset generation inside the campus using drones (MUAAD, ManipalUAVid).

Author contributions

MP - Supervision, Problem formulation, Result Analysis, Review of the Manuscript UV - Result Analysis, Re-

view of the Manuscript RMP - Result Analysis, Review of the Manuscript GS - Implementation, Manuscript Drafting.

Funding
Open access funding provided by Manipal Academy of Higher Education, Manipal

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.M.M.P.

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.