Towards Interpretable Video Anomaly Detection

Keval Doshi
University of South FLorida
4202 E Fowler Ave, Tampa, FL 33620
kevaldoshi@usf.edu

Yasin Yilmaz
University of South FLorida
4202 E Fowler Ave, Tampa, FL 33620
yasiny@usf.edu

Abstract

Most video anomaly detection approaches are based on data-intensive end-to-end trained neural networks, which extract spatiotemporal features from videos. The extracted feature representations in such approaches are not interpretable, which prevents the automatic identification of anomaly cause. To this end, we propose a novel framework which can explain the detected anomalous event in a surveillance video. In addition to monitoring objects independently, we also monitor the interactions between them to detect anomalous events and explain their root causes. Specifically, we demonstrate that the scene graphs obtained by monitoring the object interactions provide an interpretation for the context of the anomaly while performing competitively with respect to the recent state-of-the-art approaches. Moreover, the proposed interpretable method enables cross-domain adaptability (i.e., transfer learning in another surveillance scene), which is not feasible for most existing end-to-end methods due to the lack of sufficient labeled training data for every surveillance scene. The quick and reliable detection performance of the proposed method is evaluated both theoretically (through an asymptotic optimality proof) and empirically on the popular benchmark datasets.

1. Introduction

With an ever-increasing number of closed-circuit television (CCTV) cameras and the subsequent amount of video data generated continuously in real-time, it has now become inefficient and nearly impossible for human operators to manually analyze the collected data. Particularly, the ability to detect events in real-time is critical for prevention of potential catastrophes. Hence, video anomaly detection has been attracting an increasing amount of research interest. Most of the recent approaches depend on spatiotemporal features extracted in a black-box fashion, which offer limited visual interpretability and cannot explicitly explain the context of the anomaly.

A crucial task which is neglected by almost all existing algorithms is interpretable decision making, where a model is able to accurately interpret the cause of the anomaly. One of the critical applications of video anomaly detection is to take appropriate action whenever an anomaly occurs. However, the suitable response to an anomalous event is often dependant on its severity, which cannot be accurately assessed without model interpretability. For example, a car accident might require immediate attention, whereas a person simply jaywalking does not. Recent video anomaly detection approaches extract appearance and motion features of objects detected in the video and raise an alarm when there is a shift in the learned nominal data patterns [17, 12, 8, 9]. However, while objects and their relative motion are the core building blocks of a video, it is generally the relationships between objects that define its holistic interpretation. For example, a video consisting of a person and a bike could involve the person *riding*, *standing next to*, or even *carrying* the bike. Different from the existing approaches, we here focus on *object interactions*, in addition to monitoring objects independently.

Another key observation is that often activities are semantically related to each other, and yet semantic information has not yet been leveraged by any existing video anomaly detection approach. For example, a "person on a bike" and a "person on a skateboard" are semantically similar, and hence even without sufficient training data for one class, we should be able to infer it from semantically related classes. Similarly, existing approaches are unable to perform cross-domain adaptability, where a model trained on a surveillance scene is able to achieve competitive performance on a completely new scene with few or no additional training. While a similar task was discussed in [24], the proposed approach still required some training data from the new scene to fine-tune its model using meta-learning. This approach might not always be feasible since it requires a human operator to manually collect a representative set of nominal frames which also includes new activities pertaining to the surveillance scene. On the other hand, humans are able to interpret the cause of the anomaly, and adapt

the learnt knowledge to different scenes. We believe that understanding the diversity of *relationships between different objects* is essential for interpretability and cross-domain adaptability.

In recent years, scene graphs have attracted an increasing amount of research in image processing due to their interpretability and the ability to generalize to different tasks. Specifically, scene graphs combine computer vision and natural language processing to generate a visual graphical representation of an image, where nodes represent objects and edges represent the relationships between them. In this paper, we aim to address the *interpretability and cross-domain adaptability* challenges by monitoring each object individually as well its interactions with other objects in the scene. Our contributions in this paper can be summarized as follows:

- We propose a novel interpretable approach for video anomaly detection using scene graphs.
- We propose a novel semantic embedding based approach for video anomaly detection using deep metric learning, which in turn significantly reduces the memory and computational requirements.
- We extensively evaluate our proposed approach using publicly available datasets and show that it performs competitively in addition to being interpretable and cross-domain adaptable.

To the best of our knowledge, the proposed approach is the first interpretable and semantic embedding based method in the video anomaly detection literature.

2. Related Work

Anomaly detection in videos has been extensively studied for several years. While early approaches focused on using handcrafted motion features such as histogram of oriented gradients (HOGs) [2, 3, 20], Hidden Markov Models (HMMs) [18, 16], sparse coding [47, 27], and appearance features [4, 20], recent approaches have been completely dominated by deep learning algorithms. Recent algorithms can be broadly classified into reconstruction based approaches [13, 15, 26, 29, 30], which try to classify frames based on the reconstruction error, and prediction based approaches [22, 19, 7, 9], which attempt to predict a future frame, primarily by using generative adversarial networks (GANs) [14]. More recently, skeletal trajectory based approaches [28, 33] have been proposed since a large proportion of anomalies in the benchmark datasets involve anomalous human poses. In such algorithms, an RNN architecture is typically used to learn nominal poses, and estimation error is used during testing to detect the level of abnormality. Apart from these approaches, [32] proposed a Siamese network to learn spatiotemporal patches and detect an anomaly using the dissimilarity between patches. While these methods perform competitively on the popular benchmark datasets, they are completely dependant on complex neural networks and mostly end-to-end trained. This limits their interpretability and makes them notoriously difficult to train on new data, which is crucial in complex sequential applications such as video anomaly detection. Furthermore, there is no clear procedure for these methods to adapt to different nominal baselines.

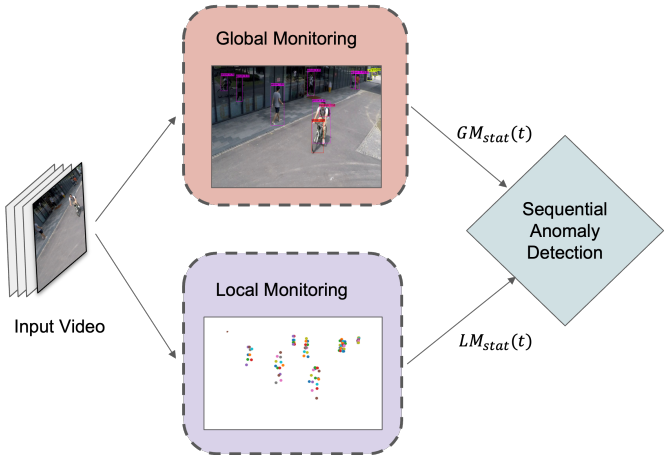

Figure 1. Proposed video anomaly detection framework. We propose a two branch pipeline, in which the global monitoring branch observes the interactions between different objects in an interpretable manner, and the local monitoring branch observes the individual skeletal pose of each person. Finally, their statistics are sequentially monitored to quickly and reliably detect anomalous events.

Recently, several visual relationship understanding approaches have been proposed for image and language related tasks. A common approach employed by several methods [23, 48, 41, 5, 39, 37] is to use an off-the-shelf detector to detect objects of interest and then treat the second step as a classification task [23, 48, 41, 5, 39, 38, 21, 36, 40, 42, 44, 45, 46], which takes in features of object pairs and outputs a label for their relationship. For video anomaly detection, to the best of our knowledge, we are the first to propose a visual relationship understanding based approach.

3. Proposed Technique

3.1. Overall Structure

In the existing video anomaly detection literature, the singular goal is to detect frames which are unexpected with respect to the training data. Most detectors train a reconstruction or prediction based deep learning model on a batch of video frames, typically in an end-to-end fashion to learn nominal appearance or motion features. However, we ar-

gue that for general video surveillance, such a setup is not optimal since learned visual embeddings are exceedingly dependant on conditions such as illumination, view point variation, occlusion, etc. Furthermore, due to the black-box nature of end-to-end trained neural networks, these models are not interpretable. Also, the standard framework implicitly assumes that sufficient training data is available for each activity from the target scene where the detector will be deployed [24]. Such an assumption requires a human to manually annotate hours of videos from each scene to generate an anomaly-free training dataset, which is far from ideal. Since humans perceive a visual environment in terms of activities, we believe that it is more natural and efficient to learn video activities *semantically* rather than storing entire frames in buffer or learning high-dimensional visual embeddings.

Motivated by these shortcomings, we propose a novel *dual-monitoring* approach for detecting video anomalies. The proposed approach consists of two branches of monitoring, namely global and local object monitoring, followed by sequential anomaly detection (Fig. 1). The global object monitoring branch exclusively observes the interactions between different objects in the scene and generates a scene graph, whereas the local object monitoring branch monitors each person in the video independently. The sequential anomaly detection block, in the end, monitors the statistics of the two object monitoring blocks to detect anomalous events in a quick and reliable manner. In the following sections, we discuss our proposed framework in detail.

3.2. Global Object Monitoring

We aim to capture and monitor the interactions between pairs of objects in a surveillance video (Fig. 2). Similar to [46, 43, 23], we propose a visual-semantic approach to detect interactions between object pairs through a scene graph. A scene graph consists of objects as graph vertices and predicates (i.e., words that relate two objects, such as verb) as graph edges. The interaction detection network is trained in a fully supervised fashion using the VRD dataset [23], which consists of a set of images with object and relationship annotations. Each detected interaction is then monitored for possible anomalies by comparing it with the detected interactions in the nominal training dataset for video surveillance (e.g., the ShanghaiTech dataset [26]). The output $GM_{stat}(t)$ is a scalar denoting the semantic distance of the interactions in frame t with respect to the nominal interactions.

Interaction Detection: As shown in Fig. 2, the proposed method first detects the bounding boxes in each frame and then processes them in pairs. A convolutional neural network (CNN) is used to extract the individual appearance features from each bounding box separately (Individual CNN). In parallel, another CNN processes the union of bounding boxes to extract the joint appearance features (Joint CNN). The joint and individual appearance features are concatenated and passed through a multilayer perceptron (MLP) to obtain a score vector $\boldsymbol{v}_1$ for each predicate class present in the VRD dataset, as well as the no-predicate class. Similarly, the individual appearance features are processed by two MLPs to obtain two $(K+1)$-dimensional score vectors $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$, where K denotes the number of predicate classes. Another score vector $\boldsymbol{v}_4$ is provided by the Semantic Module which takes the object labels (e.g., person and bike) from the Individual CNN and outputs the score $\log(q_i/q_\emptyset)$ for each predicate class i using the empirical probabilities $\{q_i, \ldots, q_K, q_\emptyset\}$ of the predicate classes for the considered subject-object pair from the annotated training data. These prior probabilities provide a baseline knowledge for predicate prediction. Heuristically, the possible combinations between two objects is usually limited. For example, the relationship between a person-bike pair would usually be "riding" or "on" and would never be "eat" or "wear". Thus this allows us to generate an empirical distribution $q_i = P(pred\ i|sub, obj)$. The final step in interaction detection is summing the four score vectors $\boldsymbol{v} = \boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3 + \boldsymbol{v}_4$ and applying the softmax function to compute the predicate class probabilities

$$p_i = \frac{e^{v(i)}}{\sum_{i=1}^{K+1} e^{v(i)}},\ i = 1, \ldots, K+1,$$

where $v(i)$ is the ith element of vector $\boldsymbol{v}$.

For training the predicate detection/classification network, we augment the well-known cross-entropy loss for quicker convergence with the contrastive loss [46], which aims to maximize, for each subject and object, the margin between the lowest affinity for related objects and the highest affinity for unrelated objects:

$$\begin{aligned} m_1^s(i) &= \min_{j \in \mathcal{X}_i^+} \Phi(s_i, o_j^+) - \max_{k \in \mathcal{X}_i^-} \Phi(s_i, o_k^-), \\ m_1^o(j) &= \min_{i \in \mathcal{X}_j^+} \Phi(s_i^+, o_j) - \max_{k \in \mathcal{X}_j^-} \Phi(s_k^-, o_j). \end{aligned} \tag{1}$$

For subject s_i, $\mathcal{X}_i^+$ and $\mathcal{X}_i^-$ represent the sets of related and unrelated objects, respectively. Similarly, $\mathcal{X}_j^+$ and $\mathcal{X}_j^-$ are defined for object o_j as the sets of related and unrelated subjects. Those sets are predetermined from the training set. $\Phi(s, o) = 1 - p_\emptyset$ is the affinity between subject s and object o, which represents the probability of interaction between the subject-object pair. In training with the VRD dataset for each image with M bounding boxes, considering all $N = M(M-1)$ subject-object pairs the following loss is computed and backpropagated through the MLPs that pro-

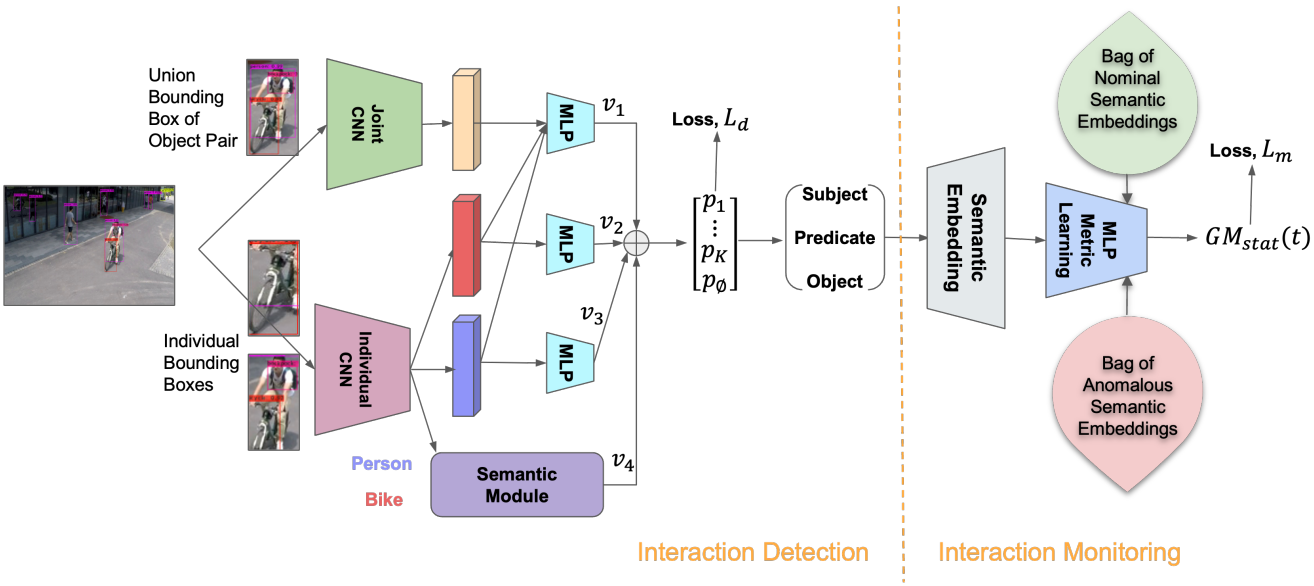

Figure 2. The global object monitoring branch of the proposed approach detects and monitors interactions between objects. In each video frame t, it detects the interaction triplets (subject, predicate, object) for the possible object pairs. Predicate can be empty for unrelated objects. Then, each interaction is monitored by computing its semantic distance to a set of nominal embeddings learned from the training data. A set of artificial anomalous interaction embeddings are generated to support the metric learning network. Finally, the largest of semantic distances in frame t is used as the global monitoring statistic $GM_{stat}(t)$ for sequential anomaly detection.

duce $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$, as well as the Joint CNN[1]:

$$L_d = \sum_{i=1}^{N} \frac{1}{N} \big[\max\{0, \alpha_1 - m_1^s(i)\} + \max\{0, \alpha_1 - m_1^o(i)\} \big] - \lambda \log p_{i^*}, \tag{2}$$

where α_1 is the margin threshold, margins below which incur a loss, λ helps combine the cross-entropy loss with the contrastive loss, and $i^* \in \{1, \ldots, K+1\}$ denotes the actual predicate class. Minimizing this loss function, the network is trained to choose the correct predicate class while maximizing the affinity between the related subject-object pairs and minimizing the affinity between the unrelated pairs.

Interaction Monitoring: For each detected interaction i in each frame t, the (subject, predicate, object) triplet with the highest probability predicate class is monitored for possible anomaly evidence. First, the word triplet is mapped to numerical vectors using a semantic embedding network, such as the Word2Vec model. Empirically, 300-dimensional embedding is found to be useful. Then, the average of the three embeddings, a_t^i, is input to an MLP for metric learning. The goal with metric learning is to extract an anomaly absence/presence information from the interaction embedding. To that end, we use another contrastive loss which minimizes the Euclidean distance between the nominal embeddings and maximizes the distance between nominal and anomalous embeddings. Denoting the metric learning embedding with $g(\cdot)$, the objective is to minimize $\|g(a_t^i) - g(p)\|$ and maximize $\|g(a_t^i) - g(n)\|$ where a_t^i, p, n respectively represent the semantic embedding vector of the anchor, positive, and negative instances. While the positive instance (i.e., interaction triplet) p is randomly selected from the nominal training set, the negative instance is randomly sampled from an artificially generated set of anomalous interaction triplets, e.g., person hits person. Note that generating anomalous (subject, predicate, object) triplets is a straightforward task and does not require actual anomalous videos. During training, only nominal instances are used as anchor so that the metric learning MLP learns to map nominal semantic embeddings close to each other and away from the anomalous instances through the loss function

$$L_m = \max\{0, \alpha_2 + \|g(a) - g(p)\| - \|g(a) - g(n)\|\} + \mu \|g(a)\|^2, \tag{3}$$

where α_2 serves as the desired margin between the distance to nominal instances and the distance to anomalous instances. μ helps combine the contrastive loss with the L2-regularizer, which ensures small embeddings for nominal (positive) instances during training.

During testing, the expectation is that the embedding $g(a_t^i)$ for the anchor instance will be small similar to $g(p)$ when the detected interaction is nominal and statistically larger than $g(p)$ when the interaction is anomalous. We use a scalar embedding $g(a_t^i) \in \mathbb{R}$ and use the largest embed-

[1]Individual CNN is separately trained using only individual objects, not pairs, as explained in Section 3.6.

ding among all detected interactions in a frame as the global monitoring statistic for anomaly detection in that frame,

$$GM_{stat}(t) = \max_i g(a_t^i).$$

3.3. Local Object Monitoring

To study the social behavior in a video, it is important to study the human motion closely. For inanimate objects like cars, trucks, bikes, etc., monitoring the optical flow is sufficient to judge whether they portray some sort of anomalous behavior. However, with regard to humans, we also need to monitor their poses to determine whether an action is anomalous or not. Hence, using a pre-trained multi-person pose estimator such as AlphaPose [10] is proposed to extract skeletal trajectories. In each frame at time t, pose i is represented by a set of joint locations in image coordinates $\theta_t^i = \left(x_t^i, y_t^i\right)_{i=1,\ldots,J}$, where J is the number of pose joints.

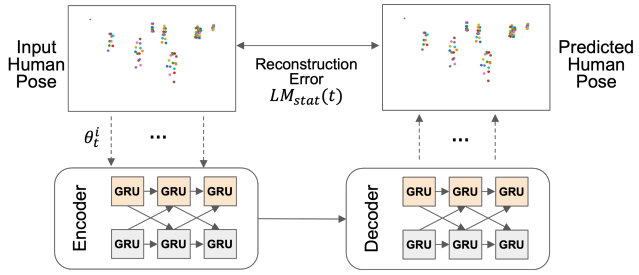

Figure 3. The local monitoring branch of the proposed approach. We input the skeletal pose for each person into a message-passing encoder-decoder network and compute the reconstruction error.

However, due to fluctuating conditions such as view point variation and occlusion, the extracted skeleton motions need to be normalized first. This can be broken into two categories, global body movement and local body posture. Specifically, global body movement describes the pose configuration with respect to the environment, whereas the local posture describes the deformity in the pose, irrespective of its position in the environment. Since the depth is missing in a 2-dimensional image coordinate, the shape of the bounding box is used to normalize the global pose component as:

$$\theta_t^{i,g} = (x_t^{i,g}, y_t^{i,g}, w_t^i, h_t^i)$$
$$x_t^{i,g} = \frac{\max_i\{x_t^i\} + \min_i\{x_t^i\}}{2}, \; y_t^{i,g} = \frac{\max_i\{y_t^i\} + \min_i\{y_t^i\}}{2},$$
$$w_t^i = \max_i\{x_t^i\} - \min_i\{x_t^i\}, \; h_t^i = \max_i\{y_t^i\} - \min_i\{y_t^i\},$$

which is then used to normalize the local component as

$$\theta_t^{i,l} = (x_t^{i,l}, y_t^{i,l}), \quad x_t^{i,l} = \frac{x_t^i - x_t^{i,g}}{w_t^i}, \quad y_t^{i,l} = \frac{y_t^i - y_t^{i,g}}{h_t^i}.$$

Inspired by the recently proposed Gated Recurrent Unit (GRU) based model called Message-Passing Encoder-Decoder Recurrent Neural Network [28], we propose the local monitoring branch shown in Fig. 3, where both the encoder and decoder networks are divided into two GRU branches. While one row of GRU processes global features, the other processes local features.

$$LM_{stat}(t) = \max_i \|\hat{\theta}_t^i - \theta_t^i\|,$$

will be of similar magnitude compared to the reconstruction errors in training, whereas for anomalous frames, it will be statistically greater than the nominal reconstruction errors.

3.4. Sequential Anomaly Detection

The global and local monitoring statistics from each frame t are combined $\boldsymbol{z}_t = [GM_{stat}(t), LM_{stat}(t)]$ and sequentially monitored for possible anomalies (Fig. 1). Both the largest semantic distance of detected object-object interactions, $GM_{stat}(t)$, and the largest reconstruction error of human poses, $LM_{stat}(t)$, are expected to take values similar to those of nominal training dataset and grow when an anomalous event starts. However, in general it is not straightforward to determine a separating boundary between the nominal and anomalous $\boldsymbol{z}_t$ values due to several factors, such as the inherent high variance in nominal and anomalous video events and the imperfect nature of feature extraction pipelines. Furthermore, instantaneous anomalous statistics for a frame in z_t may easily cause frequent false alarms. As a result of the temporal nature of anomalous real-world events unfolding in successive video frames, we consider the following sequential change detection problem:

$$\boldsymbol{z}_t \sim f_0, \; t < \tau; \; \boldsymbol{z}_t \sim f_1, \; t \geq \tau, \tag{4}$$

where f_0 and f_1 denotes the nominal and anomalous probability distributions, and τ represents the anomaly onset time.

In order to statistically monitor $\boldsymbol{z}_t$ in a sequential manner, we measure its Euclidean distance to the nominal training set $\mathcal{Z}$. In particular, the distance d_t to the kth nearest neighbor (kNN) in $\mathcal{Z}$ is computed and compared to a nominal baseline d_α to quantify any anomaly evidence in frame t. The nominal baseline d_α is obtained in training as the $(1-\alpha)$ percentile of the kNN distances of training instances with respect to each other, where α is a statistical significance level such as 0.05. That is, the kNN distance of each training $\boldsymbol{z}$ vector with respect to the other vectors within $\mathcal{Z}$ is computed, and the distance which is greater than the $(1-\alpha)\%$ of all distances is selected as d_α.

During testing, for each vector $\boldsymbol{z}_t$ at time t, the proposed algorithm computes the instantaneous frame-level anomaly

evidence δ_t as

$$\delta_t = \log d_t^2 - \log d_\alpha^2. \tag{5}$$

This specific form of δ_t enables the asymptotic optimality result presented in Theorem 1. Finally, we update a sequential decision statistic s_t as

$$s_t = \max\{s_{t-1} + \delta_t, 0\}, s_0 = 0. \tag{6}$$

We decide that there is an anomaly when the decision statistic s_t exceeds the threshold h,

$$T = \min\{t : s_t \geq h\}.$$

Theorem 1 *When the nominal distribution $f_0(\boldsymbol{z}_t)$ is finite and continuous, and the anomalous distribution $f_1(\boldsymbol{z}_t)$ is a uniform distribution, as the training set grows, the decision statistic δ_t converges in probability to the log-likelihood ratio,*

$$\delta_t \xrightarrow{p} \log \frac{f_1(\boldsymbol{z}_t)}{f_0(\boldsymbol{z}_t)} \quad as \quad |\mathcal{Z}| \to \infty, \tag{7}$$

i.e., the proposed data-driven detector converges to CUSUM, which is minimax optimum in minimizing expected detection delay while satisfying a false alarm constraint [1]:

$$\min_T \max_\tau \max_{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_\tau} \mathsf{E}_\tau[(T-\tau)^+ | \boldsymbol{z}_1, \ldots, \boldsymbol{z}_\tau] \ \ s.t. \ \ \mathsf{E}_\infty[T] \geq \beta. \tag{8}$$

E_τ denotes the expectation given the anomaly occurs at time τ, $(.)^+ = \max\{., 0\}$, E_∞ indicates the expectation given that the anomaly never occurs, i.e., $\mathsf{E}_\infty[T]$ is the expected false alarm period.

The proof is provided in the supplementary file.

3.5. Anomaly Interpretation

We perform an in-depth analysis to determine which branch of the proposed approach detected the anomaly. We begin by examining the decision statistic s_t and determining which branch is causing the increase. Hence, after an alarm is raised at time T, we

- first determine the time instance $\hat{\tau}$ when the test statistic s_t started to increase since the last time it was zero, which can be seen as an estimate for the anomaly onset time,
- then compute the average statistic

$$\bar{d}_n = \frac{1}{T - \hat{\tau} + 1} \sum_{t=\hat{\tau}}^{T} d_t^n, \tag{9}$$

  for each branch $n \in \{\text{global}, \text{local}\}$, where $(d_t^{\text{global}})^2 + (d_t^{\text{local}})^2 = d_t^2$,
- finally compare $\bar{d}_{\text{global}}$ and $\bar{d}_{\text{local}}$ with a threshold η to determine the anomaly source.

For example, if we detect the source of the anomaly as the global object monitoring branch, then it implies that the anomaly was caused due to a previously unseen interaction between two objects, and the interaction that causes $GM_{stat}(t)$ to be large during $[\hat{\tau}, T]$ is provided as the anomaly cause. If the source of the anomaly is the local object monitoring branch, then it implies that the anomaly was caused due to an anomalous human action. During training, we extract fixed length segments from a person's trajectory and separate them into local and global features, which are then input to the proposed model.

3.6. Implementation Details

We use YOLO-v4 object detection model pretrained on the MS-COCO dataset to generate bounding boxes in the global monitoring branch. The object detection model is not fine tuned on the VRD dataset for simplicity. We use VGG-16 as the Individual CNN model for extracting appearance features. It is pretrained on the MS-COCO dataset and then fine tuned on the VRD dataset in an end-to-end fashion. The Joint CNN model is also a VGG-16 model, which is initialized using the weights from the Individual CNN model. During training the Joint CNN model with the VRD dataset, we follow [46] and independently sample related and unrelated pairs. Specifically, we sample 128 subjects/objects, and then for each of them choose the related and unrelated objects/subjects according to Eq. (1). The model is trained using a learning rate of 0.001 and Adam optimizer. During inference, we take each object pair and rank the relationship proposal by multiplying the predicted subject, object, predicate probabilities. The MLPs in the interaction detection model have two fully-connected layers. Word2Vec method is used for semantic embedding in the interaction monitoring part. The MLP for metric learning has three fully-connected layers. We extract pose information from the videos using AlphaPose [10] and track them across the video using PoseFlow. The nearest neighbor ($k = 1$) is used in the sequential anomaly detection algorithm, along with the significance level $\alpha = 0.05$. The overview of the proposed framework is summarized in Algorithm 1.

4. Experiments

In this section, we evaluate the performance of the proposed approach in terms of interpretability, online anomaly detection, anomalous frame localization, and cross-domain adaptability on two benchmark datasets. We consider two publicly available benchmark datasets, namely the CUHK Avenue dataset and the ShanghaiTech campus dataset. We do not consider the UCF-Crime [34] dataset in our evaluations since it is intended for a different formulation of the

Algorithm 1: Structure of the overall algorithm

Input: Video Frames $F_1, F_2, \ldots$
Training phase:
Train the interaction detection module in the global object monitoring branch on VRD dataset using Eq. (2).
Extract semantic embeddings for triplets detected in the training video frames to form bag of nominal semantic embeddings.
Generate a set of artificial anomalous interactions and extract its corresponding semantic embeddings.
Train the interaction monitoring MLP using Eq. (3).
Extract human skeletal pose using AlphaPose and train the local monitoring branch.
Test phase:
while $s_t < h$ **do**
    $t \leftarrow t + 1$
    Get video frame F_t at time instance t.
    Compute $GM_{stat}(t)$ and $LM_{stat}(t)$ to form z_t.
    Calculate the anomaly evidence δ_t and decision statistic s_t according to Eq. (5) and Eq. (6).
end while
Declare an anomaly at time t and identify the source by comparing Eq. (9) to η.

video anomaly detection problem.

4.1. Results

Interpretability: To show the the interpretability performance of our proposed approach, we first manually annotated the root cause for each video. However, there are multiple possible predicates to explain each anomalous activity. For example, in the case of a person riding a bike, "person on bike" and "person use bike" are both possible explanations. Since it is not feasible to manually annotate all such possible combinations, we use the Recall@k metric to evaluate the interpretability performance. Since we are the first to propose an interpretable model for video anomaly detection, we are unable to compare our performance with any other method. The quantitative performance of the approach is shown in Table 1.

Interpretability	
Method	**Recall@5**
Ours	0.373

Table 1. Interpretability performance on the ShanghaiTech dataset. We compute recall for the top-5 predictions of the proposed detector by comparing it with the annotated ground truth.

In Fig. 4, we show the qualitative performance of our model. As demonstrated in the figure, the proposed approach is able to interpret and explain the true cause of the anomalies. In the middle column, we visualize the feature maps from the Joint CNN, where it can be seen that the classifier is able to learn the interaction between two objects. On the other hand, existing approaches only learn appearance or motion features, which cannot explain the anomalous object interactions. More qualitative analysis can be found in the supplementary file.

Cross-Domain Adaptability: Our goal here is to compare the cross-domain scene adaptation capability of the proposed algorithm and see how well it can generalize to new scenarios. In this case, we only train our model on the training videos from a single camera (camera 1) in the ShanghaiTech dataset and evaluate its performance on the test videos from the rest of the cameras, and also on the Avenue dataset. Cross-domain scene adaptation is mostly unexplored and to the best of our knowledge only [24] discusses a similar few-shot adaptation concept. However, the proposed approach discussed in [24] requires several anomaly-free video frames for adapting their model to the new scene, which might not always be feasible. Particularly, a GAN-based framework is used in [24] similar to [22], and MAML algorithm [11] is used for meta-learning. As shown in Tables 2 and 3, considering zero-shot adaptability the proposed approach is able to outperform the state-of-the-art methods in terms of frame-level AUC. In both of the considered datasets, behaviors that are considered anomalous are similar, which satisfies our inherent assumption. Since no zero-shot adaptability result was presented in the literature, we compared our performance with the methods whose code is available and suitable for zero-shot adaptability. The few-shot adaptable method in [24] requires videos to adapt to new scenes, hence is not suitable for zero-shot adaptation.

Anomalous Frame Localization: To show the anomaly localization capability of our algorithm, we also compare our algorithm to a wide range of state-of-the-art methods, as shown in Table 4, using the commonly used frame-level AUC criterion. The pixel-level criterion, which focuses on the spatial localization of anomalies, can be made equivalent to the frame-level criterion through simple post-processing techniques [31]. Hence, for anomaly localization, we consider the frame-level AUC criterion. While [17] recently showed significant gains over the other algorithms, their methodology of computing the average AUC over an entire dataset gave them an unfair advantage. Specifically, as opposed to determining the AUC on the concatenated videos, first the AUC for each video segment was computed and then those AUC values were averaged. As shown in Table 4, our proposed algorithm outperforms the existing algorithms on the CUHK Avenue dataset, and performs competitively on the ShanghaiTech dataset. The multi-timescale framework [33] is the only one that outperforms ours on the ShanghaiTech dataset since the anomalies are mostly

Method	Cam-1	Cam-2	Cam-3	Cam-4	Cam-5	Cam-6	Cam-7	Cam-8	Cam-9	Cam-10	Cam-11	Cam-12
Stacked RNN [26]	0.6412	0.6083	0.6116	0.6231	0.6834	0.6951	0.6482	0.6294	0.6867	0.6789	0.6924	0.6485
Future Frame Prediction [22]	0.6780	0.6178	0.6632	0.6588	0.6984	0.7351	0.6814	0.6186	0.6743	0.6789	0.6548	0.6509
Ours	0.7429	0.6924	0.7411	0.6914	0.7441	0.7985	0.7763	0.6258	0.7125	0.658	0.7531	0.7038

Table 2. Performance of the proposed detector in terms of frame-level AUC for cross-domain adaptability on different cameras from the ShanghaiTech Dataset.

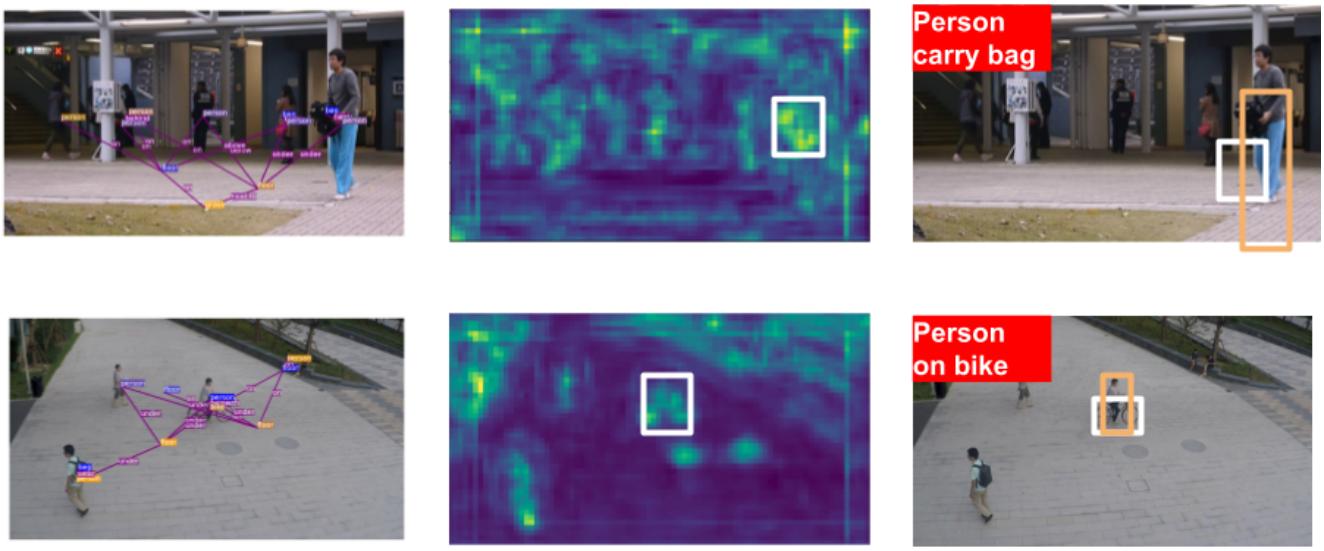

Figure 4. An example from the CUHK Avenue and ShanghaiTech datasets showing the interpretability of the proposed approach. The middle column is a visualization of the Joint CNN features by averaging over the channel dimension, where the white box represents the location where the anomaly occurs. We show that as opposed to appearance or motion based approaches, the proposed approach learns the real cause of the anomaly and outputs an interpretable result.

Frame-level AUC		
Approach	**ShanghaiTech**	**Avenue**
Stacked RNN [26]	0.643	0.724
Future Frame Prediction [22]	0.652	0.749
Skeletal Trajectory [28]	0.683	0.702
Ours	0.689	0.79

Table 3. Overall performance of each model in terms of frame-level AUC for cross-domain adaptability when trained on camera 1 from the ShanghaiTech dataset and tested on the entire ShanghaiTech and Avenue datasets.

Anomaly Localization (AUC)		
Method	**CUHK Avenue**	**ShanghaiTech**
Del et al. [6]	78.3	-
Conv-AE [15]	80.0	60.9
ConvLSTM-AE[25]	77.0	-
Growing Neural Gas [35]	-	-
Stacked RNN[26]	81.7	68.0
Future Frame [22]	85.1	72.8
Skeletal Trajectory [28]	-	73.4
Multi-timescale Prediction [33]	82.85	**76.03**
Ours	**85.78**	71.18

Table 4. Anomalous frame localization comparison in terms of frame-level AUC on two datasets.

caused by previously unseen human poses and [33] extensively monitors them using a past-future trajectory prediction based framework.

5. Conclusion

For video anomaly detection, we present an interpretable framework, which is also capable of zero-shot cross-domain adaptability. We propose a dual branch pipeline, which monitors the local and global activities in a video. While the global branch observes interactions between different objects, the local branch observes human behaviors. A sequential detector statistically monitors both branches for possible anomalies and determines the root cause of the anomaly when detected, providing interpretability. The asymptotic optimality of the proposed sequential detector is derived in terms of minimizing the average detection delay in the minimax sense while controlling the false alarm rate. Through extensive testing on the benchmark datasets, we show that the proposed approach is able to interpret the cause of detected anomalies and significantly outperforms the state-of-the-art methods in terms of cross-domain adaptability.

References

[1] Michèle Basseville and Igor V Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs, 1993.

[2] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1932–1939. IEEE, 2009.

[3] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):673–682, 2016.

[4] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011.

[5] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017.

[6] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *European Conference on Computer Vision*, pages 334–349. Springer, 2016.

[7] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020.

[8] Keval Doshi and Yasin Yilmaz. Continual learning for anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 254–255, 2020.

[9] Keval Doshi and Yasin Yilmaz. Online anomaly detection in surveillance videos with asymptotic bound on false alarm rate. *Pattern Recognition*, 114:107865, 2021.

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

[12] Mariana-Iuliana Georgescu, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. A scene-agnostic framework with adversarial training for abnormal event detection in video. *arXiv preprint arXiv:2008.12328*, 2020.

[13] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1705–1714, 2019.

[14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[15] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016.

[16] Timothy Hospedales, Shaogang Gong, and Tao Xiang. A markov clustering topic model for mining behaviour in video. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1165–1172. IEEE, 2009.

[17] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019.

[18] Louis Kratz and Ko Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1446–1453. IEEE, 2009.

[19] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. *IEEE Transactions on Image Processing*, 29:2395–2408, 2019.

[20] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013.

[21] Yikang Li, Wanli Ouyang, and Xiaogang Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. In *CVPR*, 2017.

[22] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018.

[23] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer, 2016.

[24] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. *arXiv preprint arXiv:2007.07843*, 2020.

[25] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444. IEEE, 2017.

[26] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017.

[27] Xuan Mo, Vishal Monga, Raja Bala, and Zhigang Fan. Adaptive sparse representations for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(4):631–645, 2013.

[28] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019.

[29] Trong-Nguyen Nguyen and Jean Meunier. Anomaly detection in video sequence with appearance-motion correspondence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1273–1283, 2019.

[30] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14372–14381, 2020.

[31] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020.

[32] Bharathkumar Ramachandra, Michael Jones, and Ranga Vatsavai. 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*, pages 2598–2607, 2020.

[33] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2626–2634, 2020.

[34] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.

[35] Qianru Sun, Hong Liu, and Tatsuya Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187–201, 2017.

[36] Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

[37] Xu Yang, Hanwang Zhang, and Jianfei Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *ECCV*, 2018.

[38] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *ECCV*, 2018.

[39] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, 2017.

[40] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018.

[41] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017.

[42] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *CVPR*, 2017.

[43] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9185–9194, 2019.

[44] Ji Zhang, Kevin Shih, Andrew Tao, Bryan Catanzaro, and Ahmed Elgammal. An interpretable model for scene graph generation. *arXiv preprint arXiv:1811.09543*, 2018.

[45] Ji Zhang, Kevin Shih, Andrew Tao, Bryan Catanzaro, and Ahmed Elgammal. Introduction to the 1st place winning model of openimages relationship detection challenge. *arXiv preprint arXiv:1811.00662*, 2018.

[46] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019.

[47] Bin Zhao, Li Fei-Fei, and Eric P Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR 2011*, pages 3313–3320. IEEE, 2011.

[48] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, 2017.