Ano-Graph: Learning Normal Scene Contextual Graphs to Detect Video Anomalies

Masoud Pourreza[*], Mohammadreza Salehi[*], Mohammad Sabokrou
Institute For Research In Fundamental Sciences(IPM)

Abstract

Video anomaly detection has proved to be a challenging task owing to its unsupervised training procedure and high spatio-temporal complexity existing in real-world scenarios. In the absence of anomalous training samples, state-of-the-art methods try to extract features that fully grasp normal behaviors in both space and time domains using different approaches such as autoencoders, or generative adversarial networks. However, these approaches completely ignore or, by using the ability of deep networks in the hierarchical modeling, poorly model the spatio-temporal interactions that exist between objects. To address this issue, we propose a novel yet efficient method named Ano-Graph for learning and modeling the interaction of normal objects. Towards this end, a Spatio-Temporal Graph (STG) is made by considering each node as an object's feature extracted from a real-time off-the-shelf object detector, and edges are made based on their interactions. After that, a self-supervised learning method is employed on the STG in such a way that encapsulates interactions in a semantic space. Our method is data-efficient, significantly more robust against common real-world variations such as illumination, and passes SOTA by a large margin on the challenging datasets ADOC and Street Scene while stays competitive on Avenue, ShanghaiTech, and UCSD.[1]

1. Introduction

Video Anomaly Detection (VAD) is the task of detecting those events which have rarely (or not) been observed in available training samples. Due to the ambiguity of the abnormal concept and inaccessibility of such data in training duration, learning the shared characteristics of abnormal events is not straightforward. In this circumstance, the anomaly detector is usually supposed to be trained in the absence of abnormal classes[60]. For VAD, such a detector needs to be able to detect not only both spatially and
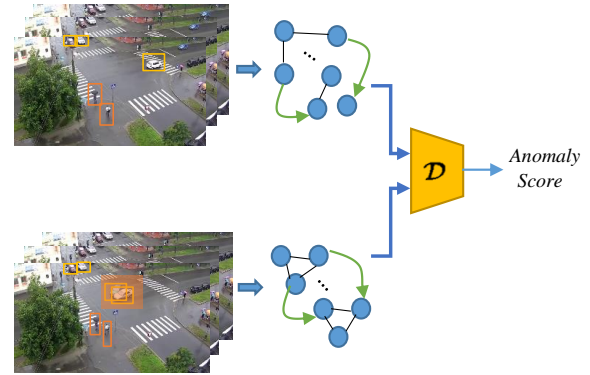

Figure 1. Our method can understand normal Spatio-temporal interactions of objects and distinguish anomalous behaviors based on them. Features of objects are considered as nodes of the graph. Spatial edges are undirected black ones, and directed green ones show temporal interactions. At the test time, a discriminator $\mathcal{D}$ is trained that discern normal graphs from abnormal ones using an anomaly score.

temporally anomalous event [55, 69, 63, 90], but also can understand the irregular video object interactions. Owing to the mentioned challenges and growing related critical applications such as detecting criminal events [70], road-traffic accidents [36] and, vehicle collisions for self-driving cars [84], VAD has recently gained significant attention [13, 14, 15, 19, 25, 26, 34, 42, 48, 50, 51, 88, 80, 17].

To model spatio-temporal features, traditionally, probabilistic approaches such as [4, 7, 28, 46, 81] were used, which were not accurate and robust to noise. However, with the advancement of deep learning, deep neural network approaches have become dominant. As an effective model that learns the semantic representation space, Generative Adversarial Networks (GANs) have been used broadly for the anomaly detection tasks [68, 62, 3, 32, 65]. Nevertheless, they suffer from serious problems such as unstable training process, irreproducibility of results, and mode

---

collapse [11]. On the other hand, Autoencoders (AEs) have a handy training process and their results are reproduced easily. Therefore, recently, AE-based approaches [1, 64, 9, 51, 19, 61, 21, 91] have been used dominantly aiming to encode and decode every consecutive fixed T frames of normal training samples, supposedly failing to reconstruct anomalous ones. Nevertheless, As [66, 67] show, AE performances are susceptible to background information or complex scenes. Therefore, variants of them based on object detection [25, 87], Self-Supervised Learning (SSL) methods [39, 48, 72, 89], or latent space clustering have been proposed [26]. Although such methods have achieved the state-of-the-art (SOTA) results, they were not able to recognize the abnormal interaction which occurred in videos.

Lately, [49] has shown the conspicuous effect of explicitly considering the interactions of objects through time and space in modeling video contexts for video captioning tasks. Despite this fact, some previous VAD approaches completely ignore the interactions of objects through space or time or build them using external supervision [71, 92]. Others tend to learn such relationships implicitly based on raw pixels using the hierarchical Deep Neural Network (DNN) structures in a very local period for each T frame. This results in downsides such as the need for fine-tuning the parameter T for each training set, and also the inability in representing and modeling long-term temporal dependencies.

On the other hand, in a supervised setting, [73, 78] have used Spatio-Temporal Graphs (STG) [86] to explicitly model high-level entities' interactions and by alleviating mentioned downsides achieved satisfying results. Also, [16, 23, 78, 79] have achieved great success on the supervised classification tasks using STG models. Nevertheless, taking advantage of the STG for anomaly detection as an unsupervised task is not a straightforward procedure.

To overcome the mentioned difficulties for modeling the interactions of objects while doing the training process in a completely unsupervised setting, we propose *Ano-Graph*. It tries to make a specific kind of STGs that captures spatio-temporal interactions while is trained using a SSL method. In this way, we not only can make a semantic normal representation space in both spatial and temporal domains without the need of fine-tuning any parameter such as T but also in a completely unsupervised training manner. Besides, SSL methods on graphs have recently shown significant performance on unsupervised graph representation learning problems [85, 93, 22, 76].

Our method consists of several steps. As Fig. 1 shows, at first, a real-time object detector such as Faster-RCNN [57] is exploited to detect objects that exist in each frame. Then each object is considered as a node of the graph and some spatial edges are built based on the Intersection Over Union (IOU) of objects' bounding boxes. This helps with better modeling of objects' interactions thorough the space. Then, for modeling temporal relations, only consecutive frames' interactions are exploited. This not only reduces the training complexity significantly but also by using Graph Convolution Networks (GCNs) [30] local information could be passed and aggregated from nodes to model complex global interactions. Hence, some edges are built based on the cosine similarity of objects in consecutive timestamps t and $t + 1$. Finally, inspired by [76], as a well-known self-supervised learning method on graphs, we learn node representations of which containing both local and global information of all spatio-temporal interactions of objects.

To the best of our knowledge, we are the first to explicitly model spatio-temporal interactions of objects for anomaly detection without exploiting any external supervision. Also, we show the effectiveness of graphical modeling in combination with SSL methods through extensive experiments. In summary, our main contributions are: (1) modeling normal interactions of objects using a STG and introducing a new perspective for the anomaly detection domain, (2) using graph SSL methods to eliminate the cost of extra labeling tasks and adopt our framework with unsupervised training procedure, (3) providing a real-time and precise anomalous object localization thanks to its object-level training procedure and, (4) conducting a huge number of diverse experiments, and outperforming SOTA methods on many datasets yet staying competitive on the rest.

2. Related Works

Traditionally, video anomaly detection was done using hand-engineered features for motion and appearance [2, 4, 10, 35, 41, 44, 46]. However, almost all of the recent works are based on deep neural networks and try to extract normal representation space from the training dataset using different approaches such as AE-based methods [1, 86, 19, 64], SSL methods [89, 17, 39], using pre-trained neural networks [50, 88, 63], extracting human body's skeleton graphs [47, 45], GAN [20] based approaches [56, 32] or combinations of them. Here we explain the closest approaches.

Latent space auto-regression for novelty detection (LSA) [1], for example, is an AE-based framework that tries to train an auto-regressive model on the latent space of the AE. Then both scores of the auto-regressive model and the reconstruction error are used to make the final abnormality score. [64, 88] attempt to train an AE in a GAN-based [20] framework. This facilitates taking the advantage of using the discriminator's output instead of reconstruction loss which is susceptible to noise, for detecting anomalies. [19, 51] try to use memory-based AEs to learn different normal patterns for the normal representation space. This consequently helps to increase the reconstruction error of anomalous test time samples and
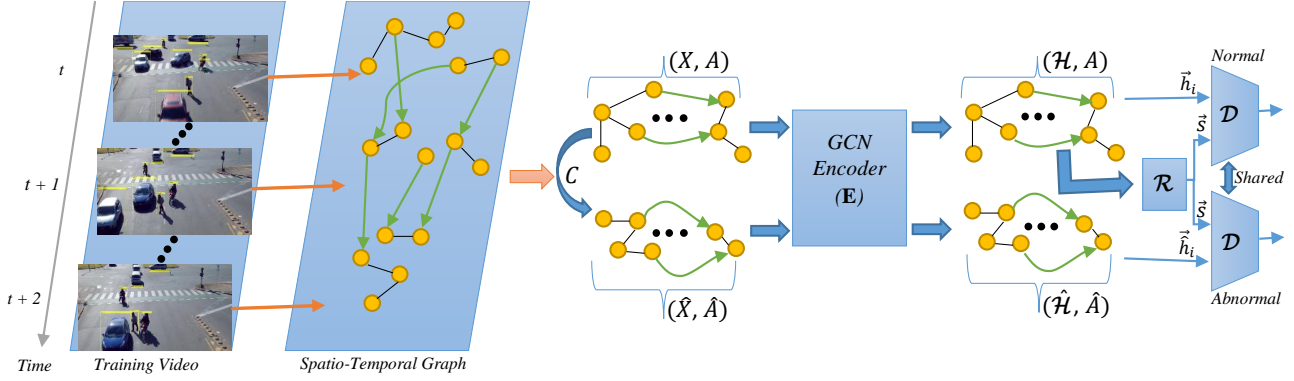
Figure 2. Visualized summary of our proposed framework. At first, features of objects are extracted using an off-the-shelf object detection model. Then, using IOU and cosine similarity, spatio-temporal interactions are modeled that is shown by STG G. Having made the spatio-temporal structure, the whole graph is passed to an encoder $\mathbf{E}$ to obtain the embeddings of nodes $\vec{h_i}$, and graph's summary $\vec{s}$. Then the discriminator $\mathcal{D}$ is trained to distinguish match or mismatch exist between the global level embedding and local ones for normal and abnormal graphs using $\vec{s}$ and $\vec{h_i}$. Abnormal graphs $(\hat{X}, \hat{A})$ are made using the stochastic corruption function C. At the test time, the videos are concatenated to each other, and a whole graph is extracted out of them. Then the graph is passed to $\mathbf{E}$ and test time embeddings are extracted. At last, the discriminator finds abnormal embeddings of nodes using their discrepancy with the summary vector $\vec{s}$ that is obtained from the training process.

discern them from normal ones.

As AEs do not generalize well on unseen normal test time samples in complex scenes [67], [39, 72] try to use U-Net [59] to predict future frames using previously seen frames. This helps the framework to bypass semantically irrelevant information to the output, which results in increasing normal sample generalization and reduces the amount of False-Positive-Rate (FPR). Similarly, [89, 48] use U-Net in combination with inpainting and appearance motion correspondence to learn better normal representation spaces.

Recently [25] has proposed a framework that focuses on the objects of interest in the training process. To do so, it firstly uses the off-the-shelf SSD object detector [37] to recognize objects that exist in each frame, then encodes motion and appearance information of the corresponding objects in each frame using the latent space of a CAE (Convolutional AE). Finally, k binary SVMs are trained on the latent space based on pseudo-labels obtained using a clustering algorithm to be used in assigning abnormality scores at the test time.

From a different point of view, [47, 45] attempt to extract each person skeleton's graph or pose and train a framework to learn normal skeleton trajectories from the training dataset. This helps them to identify human-related irregular events from video sequences more accurately. Also, [92] attempts to use GCN to approach the problem in a semi-supervised noisy training label setting using a pretrained action recognition network. The same procedure is followed in [71] that tries to build scene-aware contextual graphs. It models temporal interactions using a RNN and generates pseudo-labels based on a clustering approach to

classify the graphs as normal or anomaly. Nonetheless, in order to achieve meaningful centers, a pre-trained network on Visual-Genome [31] dataset is exploited and some prior knowledge on the normal distribution modes is needed.

3. Proposed Method

As mentioned, the spatio-temporal interactions of objects are very informative for video understanding. For instance, to prevent car crashes in the application of self-driving cars, noticing both spatial and temporal anomalous behaviors of objects is important for making real-time reactions. Graphs are a well-form data structure for representing and modeling different kinds of interactions [73, 78, 49, 86, 16]. Therefore, we propose to represent the interactions of normal video objects using a STG G. Then, inspired by [76], a discriminator ($\mathcal{D}$) is trained to detect a correspondence between the normal graph's summary information ($\vec{s}$) obtained from ($\mathcal{R}$), and the embedding of each node ($\vec{h_i}$) obtained using the Encoder ($\mathbf{E}$). $\vec{s}$ is supposed to convey global level information of the graph while $\vec{h_i}$ includes the average of neighbourhood level information for each node. Here, $\mathcal{R}$ is a simple averaging function and $\mathbf{E}$ is a graph convolutional network.

In general, $\mathcal{D}$ is supposed to learn to distinguish between corrupted graph's local embeddings and normal ones. This happens using the mismatch between the summary vector obtained using the only normal graph, and abnormal corrupted graph's local embeddings. At the test time, $\mathcal{D}$'s output could be used to detect abnormalities since they do not obey learned normal semantic regularities and structures. An overview of our method has been shown in Fig. 2.

Feature Representation: In order to extract the features of nodes, Faster-RCNN [57] as a baseline for object detection is used. We run the mentioned object detector on each frame to extract 2048 dimensional features of objects $X = \{x_1^1, x_2^1, ..., x_t^j, ..., x_T^{N_T}\}$ as will be discussed in Sec. 4.1. x_t^j is the jth from N_t objects in the frame t, and T is the size of the video (*i.e.* the number of frames). In this way, X is a high-level representation of the video which can be easily formed into a Spatio-Temporal Graph (STG). Each element of X is considered as a node of the STG and is shown by $\overrightarrow{x_i}$ in some parts of further explanations.

STG Generation: spatio-temporal graph *i.e.* G^{st} of a video with T frames is made by using the spatial G_t^{space} and the temporal G_t^{time} graphs for all timestamps $t \in T$. Spatial relations are modeled based on the IOU of each frame's objects. For the temporal part, the relationships of objects in consecutive frames are only considered. This not only significantly reduces the training complexity but also helps the network to reduce the effect of object detection noise. Additionally, by using Graph Convolutional Networks (GCN) [30], temporal information could be passed through the network and model long-term temporal interactions of objects effectively.

Spatial Graph : Similar to [49, 78, 79], the normalized IOU between different objects in each frame, as a good criterion to show the amount of their co-occurrence, is used. This helps the model to learn which objects have more spatial dependency on each other. Eq. 1 shows how a weighted edge is made based on the value of relative intersections of objects at the time t that is shown by $\sigma_{t_{ij}}$.

$$G_{t_{ij}}^{space} = \frac{e^{\sigma_{t_{ij}}}}{\sum_{j=1}^{N_t} e^{\sigma_{t_{ij}}}}, \qquad (1)$$

G_t^{space} is a $R^{N_t \times N_t}$ undirected graph that each of its elements (i, j) shows the normalized spatial connectivity of ith and j^{th} objects at time step t.

Temporal Graph: For finding temporal relations, the relative cosine similarity between features of objects in consecutive frames is exploited. This helps the model to learn spatial relations through a temporal perspective, which results in learning long-term semantic interactions. Eq. 2 shows how temporal graph G_t^{time} is made.

$$G_{t_{ij}}^{time} = \frac{e^{cos(x_t^i, x_{t+1}^j)}}{\sum_{j=1}^{N_{t+1}} e^{cos(x_t^i, x_{t+1}^j)}}, \qquad (2)$$

G_t^{time} is a $R^{N_t \times N_{t+1}}$ directed graph that each of its elements (i, j) shows the normalized temporal interaction of ith and j^{th} objects at time steps t and $t + 1$.

Spatio-Temporal Graph: After making each of the previous graphs, inspired by [49], the final

spatio-temporal graph G^{st} is formed as follows:

$$G^{st} = \begin{bmatrix} G_1^{space} & G_1^{time} & 0 & . & . & . & 0 \\ 0 & G_2^{space} & G_2^{time} & . & . & . & 0 \\ 0 & 0 & G_3^{space} & . & . & . & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & . & . & . & . & . & G_T^{space} \end{bmatrix}$$

where G^{st} is $R^{N \times N}$ and each of its elements are built as mentioned above. Also, $N = \sum_{t=1}^{T} N_t$, and zero elements are zero matrices whose shapes are adjusted based on their neighbours. Note that for making G^{st} at both test and training times all frames of all video samples are concatenated to each other respectively. This is more similar to real-world scenarios when the training video might contain different contexts. Thanks to GCN, these contexts are learned during the training process. From now on, For simplicity we use A instead of G^{st}.

Normal Interactions Learning: Given a set of node features $X = \{\overrightarrow{x_1}, \overrightarrow{x_2}, ..., \overrightarrow{x_N}\}$ where N is the number of nodes in the graph, $\overrightarrow{x_i} \in R^k$ represents the feature vector of node i, and $A \in R^{N \times N}$ as an adjacency matrix that shows relational information between nodes (G^{st}), we attempt to find embeddings $\mathcal{H} = \{\overrightarrow{h_1}, \overrightarrow{h_2}, ..., \overrightarrow{h_N}\}$ that capture both local and global information existing in the graph for each node. $\mathcal{H}$ is obtained by passing the graph representation to an encoder $\mathbf{E}(\mathcal{H}, A) : R^{N \times k} \times R^{N \times N} \longrightarrow R^{N \times K'}$ that is a graph convolutional encoder, and K' is the size of its output embedding. $\mathbf{E}(\mathcal{H}, A)$ is supposed to generate node representations by repeated aggregation over local node neighborhoods, and its key goal is finding node representations $\overrightarrow{h_i}$ that summarize a patch of the graph centered around node i rather than just the node itself.

To do so, inspired by [76], we try to find a global graph-level summary vector $\overrightarrow{s}$ which is obtained by leveraging a *readout function* $\mathcal{R} : R^{N \times K} \longrightarrow R^K$ that tries to aggregate patch-level representations $\overrightarrow{h_i}$ through Eq. 3.

$$\mathcal{R}(\mathcal{H}) = \sigma\left(\frac{1}{N} \sum_{i=1}^{N} \overrightarrow{h}_i\right) \qquad (3)$$

Then a discriminator $\mathcal{D} : R^K \times R^K \longrightarrow R = \sigma(\overrightarrow{h_i}^T W \overrightarrow{s})$ is used to assign a probability score based on the existence of the patch level summary $\overrightarrow{h_i}$ in the global graph level summary $\overrightarrow{s}$. Here, W is a learnable scoring matrix and σ is the logistic sigmoid probability function.

Using discriminator in the aforementioned setting could lead to trivial solutions due to the absence of any negative samples. To address this problem, a simple contrastive learning approach is exploited. Using an explicit stochastic corruption function $C(X, A) = (\hat{X}, \hat{A}) : R^{N \times K} \times R^{N \times N} \longrightarrow R^{M \times K} \times R^{M \times M}$ negative sample graphs that could be seen as generated abnormal samples $(\hat{X}, \hat{A})$ are

made. We observed that setting the C as a simple row-wise permutation function on only the matrix A achieves equal results compared to more complex ones. Finally, the total training loss is defined as Eq. 4.

$$\mathcal{L} = \frac{1}{N+M}\Big(\sum_{i=1}^{N}\mathbf{E}_{(X,A)}\big[\log\mathcal{D}(\overrightarrow{h}_i,\overrightarrow{s})\big] \\ + \sum_{j=1}^{M}\mathbf{E}_{(\hat{X},\hat{A})}\big[\log(1-\mathcal{D}(\overrightarrow{\hat{h}}_j,\overrightarrow{s}))\big]\Big) \quad (4)$$

This procedure maximizes the mutual information between $\overrightarrow{h_i}$ and $\overrightarrow{s}$, which means semantic normal representation embeddings are obtained in both local in global resolutions. Meanwhile, it trains a discriminator to distinct information that does not correspond to the structure of normal training samples, which could be used at the test time. For theoretical proof please refer to Appendix.

3.1. Anomaly Score

As mentioned earlier, owing to specific role of the discriminator in the training process, it learns how to discern structural irregularities with respect to normal training samples. Therefore, at the test time, the G^{st} of test time videos is obtained using the same training time procedure. Then it is passed through the encoder $\mathbf{E}$, and test time features of objects $\overrightarrow{h_i^{test}}$ are extracted. Finally, using the training summary $\overrightarrow{s}$ the anomaly score is computed at the object level as shown in Eq. 5.

$$\text{Anomaly Score} = 1 - \mathcal{D}\big(\overrightarrow{h_i^{test}},\overrightarrow{s}\big) \quad (5)$$

The more the anomaly score, the higher the probability of being an anomalous object for the test input.

4. Experiments

In this section, the proposed method *i.e.* Ano-Graph is evaluated on standard VAD benchmarks. The performance results are analyzed in details and are compared with SOTA techniques.

4.1. Setups

As mentioned above, to make the training time graph all video samples are concatenated to each other. However, for the test time one, all respective frames from $t-i$ to $t+i$ for an arbitrary length i at a specific time t are concatenated to each other. We set i with respect to each dataset's frame per second value in such a way that includes the next and previous 1 second period at a specific time t. Even so, our method is not sensitive to a specific value of i and it could be set to higher or lower values without any need of retraining. Note that all the process is done in real-time since at each time step $t+1$ only G_{t+i+1}^{space} and G_{t+i+1}^{time} are added to the current G^{st} while G_{t-i}^{space} and G_{t-i}^{time} could be removed.

In order to make the process even faster, a one-layer Graph Convolutional Network (GCN) is used as our encoder $\mathbf{E}$. To extract object features, we have used Detectron2 [82]. First apply a Faster R-CNN (with ResNet-101-C4 backbone) [18] pre-trained on COCO Dataset [38] to generate object bounding boxes for each frame. We set the confidence score threshold for a detection to 0.65 for all training datasets. Given the output bounding boxes we apply RoI pooling [18] to extract features of the corresponding regions. Specifically, we first project the bounding boxes onto the feature map from the last convolutional layer of ResNet-101, then apply RoI pooling to crop and rescale the object features within the projected bounding boxes into the same spatial dimension. This generates a 14×14×2048 feature for each object, which is then average-pooled to 1×1×2048. Moreover, There is no constraint on the amount of objects in one frame.

Also, adam SGD optimizer [29] with an initial learning rate of 0.001 in combination with early stopping strategy on the observed training loss, with a patience of 200 epochs are used. Batch size is set to 64 for 10000 epochs and all experiments are conducted on GeForce GTX 1080 Ti.

4.2. Datasets

We evaluate our method on popular datasets such as UCSD-Ped2 [44], Avenue [41], ShanghaiTech [43], and also the challenging recently introduced datasets ADOC [52] and Street Scene [53] as well. Each dataset has pre-defined training and test sets, anomalous events being included only at test time.

UCSD-Ped2: UCSD-Ped2 contains 16 training and 12 test videos. The videos illustrate various crowded scenes, and anomalies include bicycles, vehicles, skateboarders and wheelchairs crossing pedestrian areas. The resolution of each frame is 240 × 360 pixels.

Avenue: This dataset is consist of 16 training videos with normal activity and 21 test videos. Anomalous events consists of people running, throwing objects or walking in the wrong direction. The resolution of each video is 360 × 640 pixels.

ShanghaiTech: ShanghaiTech is one of the largest video anomaly detection datasets that consists of 330 training videos and 107 test videos. The training videos merely consists of normal events. However, the test videos contain both normal and abnormal ones including robbing, jumping, fighting and bikers in pedestrian areas. The resolution of each video is 480 × 856 pixels.

ADOC: A Day on Campus (ADOC), with 25 event types, spanning over 721 instances and occurring over a period of 24 hours with different lighting situations and extreme imbalanced event frequencies, is one of the most challenging anomaly detection datasets. The resolution of each video is 1080p and the frame-rate of 3 frames per second.

Street Scene: This dataset consists of 46 training video sequences and 35 testing video sequences with resolution of 1280 x 720 pixels which taken from a static USB camera looking down on a scene of a two-lane street with bike lanes and pedestrian sidewalks. Videos were collected from the camera at various times during two consecutive summers. All of the videos were taken during the daytime. It includes 205 abnormal events.

4.3. Evaluation

Similar to previous methods [39, 64, 26, 25, 51, 19, 1] the frame-level *area under the curve* (AUC) is exploited for evaluation the performance of our method on **Avenue**, **ShanghaiTech**, **UCSD-Ped2**, **ADOC**, and **Street Scene** datasets. We consider a frame as anomaly, if it contains at least one detected abnormal pixel. For **ADOC** we not only use frame-level AUC but also for the test time threshold in which the difference of True-Positive-Rate (TPR) and False-Positive-Rate (FPR) is maximized, the values of True Negative (TN), False Positive (FN), True Positive(TP), False Positive (FP), accuracy, and nlilo-Accuracy are computed as suggested in the respective paper. Also, Track-Based Detection Criterion (TBDR) and Region-Based Detection Criterion (RBDR) are reported on **Street Scene** dataset. In all experiments positive class is considered as anomaly.

4.4. Results

Table 1, Table 2, and Table 3 show the performance of our method *i.e.* Ano-Graph in comparison with previous SOTA methods for VAD. The reported results are borrowed from original papers or standard benchmarks (whenever available) on Avenue, ShangahiTech, UCSD-Ped2, ADOC, and Street Scene. Since no results are available on the recently proposed Street Scene dataset, we have reported the performance of the official source code of some of the most popular SOTA methods on this dataset. For the sake of consistency the same set of SOTA methods are reported for both ADOC and Street Scene.

4.4.1 Results on Avenue, ShanghaiTech & UCSD-Ped2:

First, we evaluate our method on these conventional VAD datasets as mentioned in Sec. 4.3. The number of anomalous events for Avenue, ShanghaiTech, and UCSD-Ped2 are 47, 130, and 20 respectively, and all of them are less than 3.6 hours [52]. We compare our method with an exhaustive set of SOTA approaches, including generative, SSL and AE-based methods, in Table 1. As it can be seen, the results of our method is comparable or even better compared with the other considered methods. Note that some approaches such as Chang et al. [6] and Rodrigues et al. [58] while

Table 1. AUC in % for frame-level anomaly detection. As shown, our method shows SOTA results on Avenue, Shanghai, and USCD-Ped2. AVG shows the average AUC value on the 3 datasets.

Year	Method	Avenue	ShanghaiTech	USCD-Ped2	AVG
2016	Del Giorno et al. [12]	78.3	-	-	-
2016	Hasan et al. [21]	70.2	60.9	90.0	73.7
2016	Zhang et al. [90]	-	-	91.0	-
2017	Hinami et al. [24]	-	-	92.2	-
2017	Ionescu et al. [74]	80.6	-	82.2	-
2017	Luo et al. [43]	81.7	68.0	92.2	-
2017	Ravanbakhsh et al. [56]	-	-	93.5	-
2017	Smeureanu et al. [69]	84.6	-	-	-
2017	Xu et al. [83]	-	-	90.8	-
2017	Chong et al. [8]	80.3	-	87.4	-
2018	Lee et al. [33]	87.2	-	96.5	-
2018	Liu et al. [40]	85.1	72.8	95.4	84.43
2018	Liu et al. [39]	84.4	-	87.5	-
2018	Ravanbakhsh et al. [55]	-	-	88.4	-
2018	Sultani et al. [70]	-	**76.5**	-	-
2019	Gong et al. [19]	83.3	71.2	94.1	82.86
2019	Ionescu et al. [26]	**88.9**	-	-	-
2019	Nguyen et al. [48]	86.9	-	96.2	-
2019	Vu et al. [77]	71.5	-	99.2	-
2019	Wu et al. [80]	86.6	-	-	-
2020	Dong et al. [13]	84.9	73.7	95.6	84.73
2020	Doshi et al. [14]	86.4	71.6	97.8	85.26
2020	Ji et al. [27]	78.3	-	**98.1**	-
2020	Park et al. [51]	88.5	70.5	97.0	**85.33**
2020	Ramachandra et al. [53]	72.0	-	88.3	-
2020	Ramachandra et al. [54]	87.2	-	94	-
2020	Tang et al. [72]	85.1	73.0	96.3	84.4
2020	Rodrigues et al. [58]	82.85	76.03		-
2020	Chang et al. [6]	86.0	73.3	96.5	85.26
2021	Ano-Graph (Ours)	**86.2**	**74.42**	**96.68**	**85.57**

achieve great performance need either a prior knowledge about normal distribution modes to fine-tune their clustering hyper-parameters or time scales. On the contrary, Ano-Graph does not need any assumption on the selection of parameters such as T. It provides significantly more *flexibility* compare to a large number of previous methods in modeling long-term dependencies. Also, other methods can not easily get adapted to settings in which T is set to a large number. For instance, suppose GAN-based methods, They become seriously unstable when the parameter T is increased. Also, AE-based methods' performance is susceptible to the number of objects and scene complexities. This happens because of their reconstruction based approach [67]. We, however, try to use an object detector in combination with a STG and SSL method to model these complexities in a more powerful way.

4.4.2 Results on ADOC & Street Scene:

To further evaluate our method, we use the two most challenging recently introduced datasets. Surprisingly, some of the very recent SOTAs that work pretty well on the conventional datasets achieve near-random performance on these, while our method shows much more reliable performance

Table 2. Comparison of our method with some of the best SOTAs using different standard criteria. As shown, our method reaches to a new SOTA by a large margin of 16% on the 3 experiments average in AUC, Accuracy, and nlilo-Accuracy, which shows the applicability and robustness of our method on this challenging dataset.The top two methods are in bold.

Experiment	Year	Method	True Negative(↑)	False Negative(↓)	True Positive(↑)	False Positive(↓)	AUC(↑)	Accuracy(↑)	nlilo-Accuracy(↑)
Exp 1	2017	Chong et al. [8]	9831	**16429**	**37726**	714	**84.6**	**73.5**	**69.8**
	2018	Sabokrou et al. D(R(x)) [64]	**10432**	53072	1083	113	25.7	17.8	29.9
	2018	Sabokrou et al. D(x) [64]	10427	52970	1185	**118**	24.7	17.9	30.7
	2018	Liu et al. [39]	7477	29166	24989	3068	60.5	50.1	58.3
	2019	Gong et al. [19]	8381	35283	18872	2164	57.0	42.1	59.8
	2020	Park et al. [51]	9082	43607	10548	1463	52.3	30.3	58.8
	2021	Ano-Graph (Ours)	**12348**	13772	37989	591	84.22	77.8	72.6
Exp 2	2017	Chong et al. [8]	22226	514	20258	27402	**64.1**	60.3	**81.9**
	2018	Sabokrou et al. D(R(x)) [64]	1187	**191**	**20581**	48441	38.3	30.9	62.7
	2018	Sabokrou et al. D(x) [64]	1352	202	20570	48276	37.7	31.1	65.6
	2018	Liu et al. [39]	**49625**	20766	6	**3**	34.6	**70.4**	58.3
	2019	Gong et al. [19]	3080	608	20164	46548	40.7	33.0	59.6
	2020	Park et al. [51]	12041	3239	17533	37587	52.2	42.0	65.1
	2021	Ano-Graph (Ours)	**27109**	**573**	24706	18012	70.81	73.6	84.1
Exp 3	2017	Chong et al. [8]	53374	63426	11501	6799	44.5	48.0	30.9
	2018	Sabokrou et al. D(R(x)) [64]	45071	**14845**	**60082**	15102	**79.5**	**77.8**	82.4
	2018	Sabokrou et al. D(x) [64]	44415	19042	55885	15758	74.0	74.2	79.5
	2018	Liu et al. [39]	47733	22835	52092	12440	72.7	73.8	82.4
	2019	Gong et al. [19]	**60173**	74911	16	**0**	25.2	44.5	19.2
	2020	Park et al. [51]	49589	23294	51633	10584	70.0	74.9	**88.1**
	2021	Ano-Graph (Ours)	**51937**	**18902**	58035	6226	87.05	81.4	89.5

and reaches a new SOTA on both ones.

ADOC: As Table 2 shows, Ano-Graph passes the best-reported method in frame-level AUC by at least **16%** margin on the average of 3 experiments. Experiment. 1 contains only day images, Experiment. 2 contains only night images, and Experiment. 3 contains both day and night images for both train and test times. Our results show the significant applicability of our method in different real-world scenarios, which means learning semantic embeddings for objects' interactions independent of irrelevant variations. As [52] suggests, extra evaluation metrics are also reported on ADOC which shows the consistent superiority of our method's accuracy and nlilo-accuracy with respect to others. Note that some methods have lower FP compare to us, nonetheless, this happens because of the extreme imbalance that exists towards normal samples in this dataset. The results apparently show that these methods not only produce low FP but also low TP, which means they get overfitted on reporting every event as a normal one. Also, The nlilo-Accuracy of such methods shows their mentioned deficiency.

Street Scene: The performance of the proposed method in comparison to the other SOTA methods is listed in Table 3. Following [53], We also report the performance of our method in terms of AUC, TBDR, and RBDR. As can be seen, Ano-Graph achieves better results rather than other SOTA methods by a considerable margin. The AUC, TBDR, and RBDR are improved by ours to 8.19%, 9.91%, and 16.7%, respectively. This outperforming is because of focusing on object-level anomalous sample localization and spatio-temporal modeling of the normal scene structures.

Table 3. TBDR, RBDR, and frame-level AUC in % on Street Scene dataset. As shown, our method has significantly better performance in all criteria and reaches to a new SOTA.The symbol(*) means that we use their official implementation.

Year	Method	AUC	TBDR	RBDR
2017	Chong et al. [8]*	64.42	58.9	49.02
2018	Sabokrou et al. D(R(x)) [64]*	43.77	47.01	39.14
2018	Sabokrou et al. D(x) [64]*	41.57	46.15	36.72
2018	Sabokrou et al. [65]*	47.68	56.72	42.12
2018	Liu et al. [39]*	61.1	32.91	27.62
2019	Gong et al. [19]*	55.72	49.12	30.12
2020	Park et al. [51]*	51.07	38.18	29.7
2021	Ano-Graph (Ours)	**72.61**	**68.81**	**65.72**

4.5. Running Time

As our method consists of two different parts, we report their execution times separately. For the object detection part, as mentioned above, Faster-RCNN has been used that is significantly slower than SOTA object detection methods such as YOLOv4 [5]. The execution time of Faster-RCNN on GeForce GTX 1080 Ti is about 8 frames per second (FPS). However, it could get faster by substituting YOLOv4 to above 20FPS [5]. The rest of the method has significantly low execution time. It takes the amortized value of about 0.00198 second for each object to be detected as anomaly or normal. This means STG generation and anomalous interaction detection is extremely fast and can be done in real-time even for almost all crowded scenes.

5. Ablation Studies

Independent effects of spatial and temporal interactions: We conduct two experiments on different datasets to

show the effectiveness of joint modeling of spatio-temporal interactions. As the Table 4 shows, considering both kinds of interactions is necessary for well-encapsulating the content of videos and gets better test time AUC by a large margin of 7% to 10% with respect to each dataset.

Table 4. As it is shown in the table, both spatial and temporal interactions are necessary for achieving the best results.

Interaction	Avenue	UCSD-Ped2
Spatial	70.42	84.21
Temporal	76.62	89.41
Spatio-Temporal	**86.24**	**96.68**

Effect of making local test time graphs: In this experiment, we try to assess the performance of our method by making different local test time graphs using the length parameter i. We change i from 1 to 20 and report the test time accuracy on Avenue dataset without any retraining. Avenue's videos have 25 frames per second. As it is shown in the Table 5, large numbers of i achieve almost better results, which approves the usability of our method for modeling long-term dependencies. Besides this help users to dynamically change the test time period on demand.

Table 5. As it is shown in the table, the more global a graph is made, the better AUC values are achieved. Parameter i shows the number of previous or next frames are used for this experiment.

Time Scale	Avenue
$i = 1$	83.32
$i = 3$	84.12
$i = 5$	84.69
$i = 10$	85.37
$i = 15$	85.52

Visualization of node embeddings: We performed an analysis on the embeddings learned by the *Ano-Graph* algorithm to better understand its properties. The Avenue is chosen as the training dataset due to its small number of nodes which significantly aiding clarity. Fig. 3, depicts $\overrightarrow{h_i}$ values of test time samples using t-SNE [75] with its default settings. Color is set in correspondence with the input's anomaly score. As it is shown, our method not only produces significantly higher scores for the abnormal interactions but also they are geometrically dense and separated. This supports our primary assumption about achieving semantic space for normal interactions.

Data Efficiency: We compare the data efficiency of our method with one of the SOTA on UCSD dataset. We reduce the number of training frames from 10% to 50% by random discarding and report the results of Ano-Graph in comparison with MNAD [51]. As Fig. 4 shows, despite the better performance of MNAD when 100% of frames are used, its performance decreases significantly and stays 7%

below of Ano-Graph when 50% of the frames are discarded. This shows our method's data efficiency that makes it more applicable for real-world scenarios.
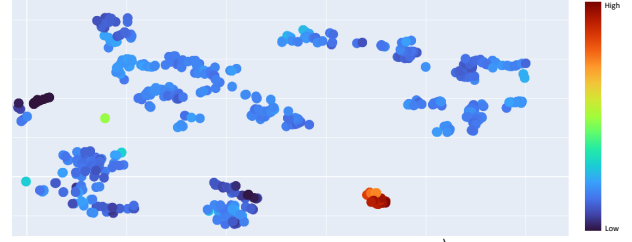

Figure 3. Visualization of node embeddings ($\overrightarrow{h_i}$) for test time samples when the graph is trained on the Avenue dataset. The color of each sample sets with respect to the discriminator's anomaly score. As it is shown anomalous interactions have higher anomaly scores and have been separated geometrically from normal ones, which supports our assumption of obtaining semantic representation space for different normal contexts and abnormal ones.
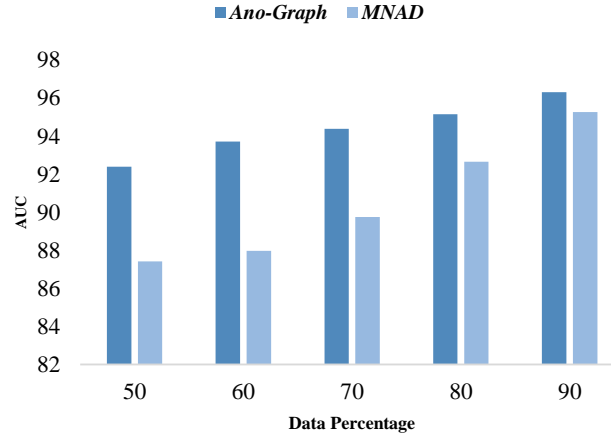

Figure 4. The performance of our method in comparison with MNAD on UCSD when 10% to 50% of the training frames are randomly discarded. The performance of MNAD decreases roughly **2.5** times faster than Ano-Graph.

6. Discussion and Conclusion

In this work, we have proposed a novel anomaly detection approach based on a SSL method and a STG, presenting comprehensive results on five datasets Avenue, ShanghaiTech, UCSD-Ped2, ADOC, and Street Scene. Experiments and ablation studies show that we not only pass SOTA by a *large* margin but also we are more *flexible*, *data efficient*, and *robust* compared to others. Also, we believe Ano-Graph could make a new perspective towards **Early Anomaly Prediction**. For instance, it seems obvious that 2 cars are going to crash by considering their direction and speed before it really happens. Ano-Graph could be easily adapted to all kinds of these scenarios as well.

References

[1] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2019. 2, 6

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008. 2

[3] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*, pages 622–637. Springer, 2018. 1

[4] Borislav Antić and Björn Ommer. Video parsing for abnormality detection. In *2011 International Conference on Computer Vision*, pages 2415–2422. IEEE, 2011. 1, 2

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 7

[6] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *European Conference on Computer Vision*, pages 329–345. Springer, 2020. 6

[7] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. 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*, pages 2909–2917, 2015. 1

[8] Yong Shean Chong and Yong Haur Tay. Abnormal event detection in videos using spatiotemporal autoencoder. In *International symposium on neural networks*, pages 189–196. Springer, 2017. 6, 7

[9] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011. 2

[10] Yang Cong, Junsong Yuan, and Ji Liu. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46(7):1851–1864, 2013. 2

[11] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018. 2

[12] 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. 6

[13] Fei Dong, Yu Zhang, and Xiushan Nie. Dual discriminator generative adversarial network for video anomaly detection. *IEEE Access*, 8:88170–88176, 2020. 1, 6

[14] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 934–935, 2020. 1, 6

[15] 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. 1

[16] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019. 2, 3

[17] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. *arXiv preprint arXiv:2011.07491*, 2020. 1, 2

[18] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 5

[19] 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. 1, 2, 6, 7

[20] 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. 2

[21] 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. 2, 6

[22] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pages 4116–4126. PMLR, 2020. 2

[23] Roei Herzig, Elad Levi, Huijuan Xu, Eli Brosh, Amir Globerson, and Trevor Darrell. Classifying collisions with spatio-temporal action graph networks. *arXiv preprint arXiv:1812.01233*, 2, 2018. 2

[24] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3619–3627, 2017. 6

[25] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019. 1, 2, 3, 6

[26] Radu Tudor Ionescu, Sorina Smeureanu, Marius Popescu, and Bogdan Alexe. Detecting abnormal events in video using narrowed normality clusters. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1951–1960. IEEE, 2019. 1, 2, 6

[27] Xiangli Ji, Bairong Li, and Yuesheng Zhu. Tam-net: Temporal enhanced appearance-to-motion generative network for video anomaly detection. In *2020 International Joint Con-*

ference on Neural Networks (IJCNN), pages 1–8. IEEE, 2020. 6

[28] Jaechul Kim and Kristen Grauman. 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*, pages 2921–2928. IEEE, 2009. 1

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[30] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2, 4

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 3

[32] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1811–1820, 2019. 1, 2

[33] Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Stan: Spatiotemporal adversarial networks for abnormal event detection. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1323–1327. IEEE, 2018. 6

[34] 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. 1

[35] 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. 2

[36] Yingying Li, Jie Wu, Xue Bai, Xipeng Yang, Xiao Tan, Guanbin Li, Shilei Wen, Hongwu Zhang, and Errui Ding. Multi-granularity tracking with modularlized components for unsupervised vehicles anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 586–587, 2020. 1

[37] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[39] 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. 2, 3, 6, 7

[40] Yusha Liu, Chun-Liang Li, and Barnabás Póczos. Classifier two sample test for video anomaly detections. In *BMVC*, page 71, 2018. 6

[41] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 2, 5

[42] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *European Conference on Computer Vision*, pages 125–141. Springer, 2020. 1

[43] 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. 5, 6

[44] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010. 2, 5

[45] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10539–10547, 2020. 2, 3

[46] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942. IEEE, 2009. 1, 2

[47] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019. 2, 3

[48] 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. 1, 2, 3, 6

[49] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020. 2, 3, 4

[50] Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12173–12182, 2020. 1, 2

[51] 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. 1, 2, 6, 7, 8

[52] Mantini Pranav, Li Zhenggang, et al. A day on campus-an anomaly detection dataset for events in a single camera. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 5, 6, 7

[53] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video

anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2569–2578, 2020. 5, 6, 7

[54] 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. 6

[55] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. 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)*, pages 1689–1698. IEEE, 2018. 1, 6

[56] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. Abnormal event detection in videos using generative adversarial nets. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017. 2, 6

[57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 2, 4

[58] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2626–2634, 2020. 6

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[60] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 1

[61] Mohammad Sabokrou, Mahmood Fathy, and Mojtaba Hoseini. Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder. *Electronics Letters*, 52(13):1122–1124, 2016. 2

[62] Mohammad Sabokrou, Mahmood Fathy, Guoying Zhao, and Ehsan Adeli. Deep end-to-end one-class classifier. *IEEE transactions on neural networks and learning systems*, 2020. 1

[63] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes. *IEEE Transactions on Image Processing*, 26(4):1992–2004, 2017. 1, 2

[64] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018. 2, 6, 7

[65] Mohammad Sabokrou, Masoud Pourreza, Mohsen Fayyaz, Rahim Entezari, Mahmood Fathy, Jürgen Gall, and Ehsan

Adeli. Avid: Adversarial visual irregularity detection. In *Asian Conference on Computer Vision*, pages 488–505. Springer, 2018. 1, 7

[66] Mohammadreza Salehi, Atrin Arya, Barbod Pajoum, Mohammad Otoofi, Amirreza Shaeiri, Mohammad Hossein Rohban, and Hamid R Rabiee. Arae: Adversarially robust training of autoencoders improves novelty detection. *arXiv preprint arXiv:2003.05669*, 2020. 2

[67] Mohammadreza Salehi, Ainaz Eftekhar, Niousha Sadjadi, Mohammad Hossein Rohban, and Hamid R Rabiee. Puzzle-ae: Novelty detection in images through solving puzzles. *arXiv preprint arXiv:2008.12959*, 2020. 2, 3, 6

[68] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 1

[69] Sorina Smeureanu, Radu Tudor Ionescu, Marius Popescu, and Bogdan Alexe. Deep appearance features for abnormal behavior detection in video. In *International Conference on Image Analysis and Processing*, pages 779–789. Springer, 2017. 1, 6

[70] 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. 1, 6

[71] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 184–192, 2020. 2, 3

[72] Yao Tang, Lin Zhao, Shanshan Zhang, Chen Gong, Guangyu Li, and Jian Yang. Integrating prediction and reconstruction for anomaly detection. *Pattern Recognition Letters*, 129:123–130, 2020. 2, 3, 6

[73] Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, and Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10433, 2019. 2, 3

[74] Radu Tudor Ionescu, Sorina Smeureanu, Bogdan Alexe, and Marius Popescu. Unmasking the abnormal events in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2895–2903, 2017. 6

[75] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[76] Petar Velickovic, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. In *ICLR (Poster)*, 2019. 2, 3, 4

[77] Hung Vu, Tu Dinh Nguyen, Trung Le, Wei Luo, and Dinh Phung. Robust anomaly detection in videos using multilevel representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5216–5223, 2019. 6

[78] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018. 2, 3, 4

[79] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gang-shan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019. 2, 4

[80] Peng Wu, Jing Liu, and Fang Shen. A deep one-class neural network for anomalous event detection in complex scenes. *IEEE transactions on neural networks and learning systems*, 31(7):2609–2622, 2019. 1, 6

[81] Shandong Wu, Brian E Moore, and Mubarak Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2054–2060. IEEE, 2010. 1

[82] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 5

[83] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117–127, 2017. 6

[84] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Ella Atkins, and David Crandall. When, where, and what? a new dataset for anomaly detection in driving videos. *arXiv preprint arXiv:2004.03044*, 2020. 1

[85] Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. When does self-supervision help graph convolutional networks? In *International Conference on Machine Learning*, pages 10871–10880. PMLR, 2020. 2

[86] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017. 2, 3

[87] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020. 2

[88] Muhammad Zaigham Zaheer, Jin-ha Lee, Marcella Astrid, and Seung-Ik Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2020. 1, 2

[89] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognition*, page 107706, 2020. 2, 3

[90] Ying Zhang, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Shun Sakai. Video anomaly detection based on locality sensitive hashing filters. *Pattern Recognition*, 59:302–311, 2016. 1, 6

[91] Yiru Zhao, Bing Deng, Chen Shen, Yao Liu, Hongtao Lu, and Xian-Sheng Hua. Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1933–1941, 2017. 2

[92] Jia-Xing Zhong, Nannan Li, Weijie Kong, Shan Liu, Thomas H Li, and Ge Li. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019. 2, 3

[93] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020. 2