Dance with Self-Attention: A New Look of Conditional Random Fields on Anomaly Detection in Videos

Didik Purwanto, Yie-Tarng Chen, and Wen-Hsien Fang
National Taiwan University of Science and Technology, Taiwan, R.O.C.
Email: {d10602806,ytchen,whf}@mail.ntust.edu.tw

Abstract

This paper proposes a novel weakly supervised approach for anomaly detection, which begins with a relation-aware feature extractor to capture the multi-scale convolutional neural network (CNN) features from a video. Afterwards, self-attention is integrated with conditional random fields (CRFs), the core of the network, to make use of the ability of self-attention in capturing the short-range correlations of the features and the ability of CRFs in learning the inter-dependencies of these features. Such a framework can learn not only the spatio-temporal interactions among the actors which are important for detecting complex movements, but also their short- and long-term dependencies across frames. Also, to deal with both local and non-local relationships of the features, a new variant of self-attention is developed by taking into consideration a set of cliques with different temporal localities. Moreover, a contrastive multi-instance learning scheme is considered to broaden the gap between the normal and abnormal instances, resulting in more accurate abnormal discrimination. Simulations reveal that the new method provides superior performance to the state-of-the-art works on the widespread UCF-Crime and ShanghaiTech datasets.

1. Introduction

Anomaly detection seeks to recognize an event that deviates from normal behaviors and identify the instant of the abnormal event occurring from a sequence of images. It has received growing interests owing to its ubiquitous applications like criminal detection [1], intelligent surveillance [2], and violent alerting [3], *etc*.

Detecting anomalies within real world videos can be demanding for a number of factors like a variety of illuminations, multiple camera angles, indoor and outdoor conditions, and inter- and intra- classes variation problem. Moreover, the abnormal activities often occur in a short period of time.

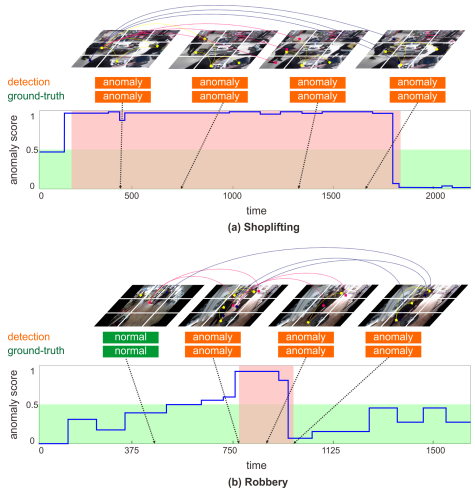


Figure 1: Challenges of the anomaly detection: (a) anomaly is formed by the spatial-temporal interactions among the actors; (b) abnormal behaviors can occur in a short period of time.

Since many commonly used datasets only provide video-level annotations, a plethora of weakly supervised anomaly detection algorithms has been considered. For instance, Sultani *et al.* [1] employed a multi-instance learning (MIL), which incorporated both normal and abnormal samples with weakly labeled annotations to detect the abnormal events. Lin *et al.* [3] used a dual-branch module to learn the context information. Also, Zhang *et al.* [4] considered an inner bag loss to minimize the distance between the negative instances in each bag. To better learn motion cue information, some latest approaches made use of deep motion features [5, 6]. For example, Zhu *et al.* [5] proposed a temporal augmented network to model the motion aware feature representation. Liu *et al.* [7] eradicated the effect of the background information to learn the anomalies from specific areas. However, [1, 5, 7] focused on the present information and did not fully take advantage of the temporal dependencies across frames. Recently, Zhong *et al.* [2] employed a graph convolutional network (GCN) to iteratively refine the noisy anomaly labels. However, this iterative process trades performance for complexity. Zaheer et al. [8] proposed a self-reasoning net-

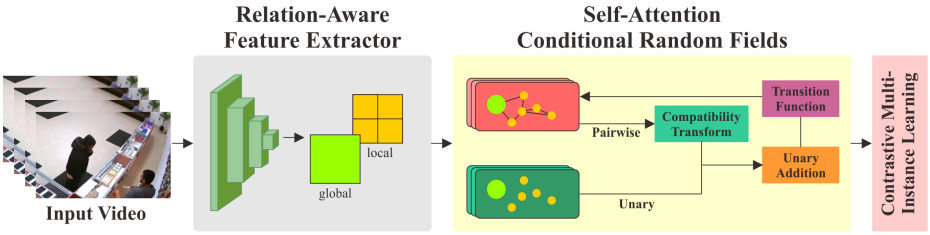


Figure 2: The overall pipeline of the proposed approach.

work that invokes binary clustering to mitigate the noisy labels of video snippets. Recently, Zaheer *et al.* [9] used a normalcy suppression scheme to prevent the inter-batch correlation problem. Wu *et al.* [10] incorporated the audio-visual modality to complement the widely used two-stream architectures.

In light of the importance of adequately modelling the spatial-temporal structural relationship across frames, as illustrated in Fig. 1 (a), where the '*robbery*' event is formed by the dynamic interaction between the '*cashiers*' and '*robber*' across frames, this paper presents a potent weakly supervised method for anomaly detection. The new method first uses a relation-aware feature extractor to capture the multi-scale convolutional neural network (CNN) features from a video. This extractor is an extension of the renowned temporal relational network (TRN) [11] with multi-scale partitions and the use of inner product operation, aiming at providing more discriminative global and local features and their short-range correlations. Afterwards, central to the new network is to integrate self-attention with conditional random fields (CRFs) to make use of the ability of self-attention in learning the short-range correlations of the features and the ability of CRFs in learning the inter-dependencies of these features.

The combination of conditional random fields (CRFs) and self-attention is inspired by [12]. However, there are some crucial differences: i) our problem is weakly supervised learning, while [12] is supervised learning; ii) the nodes of the constructed spatial-temporal graphs, from which we compute the self-attention and conduct CRF inference, are generated by a new relation-aware feature extractor to capture the anomalies in either the global or local regions; iii) we jointly learn both of the spatial and temporal relationships between the nodes and consider the reasoning of multiple actors' nodes to account for their diverse relationships, while [12] modeled the nodes' spatial and temporal relationships independently, as depicted in Fig. 3; iv) our definition of self-attention is different from [12] by taking a set of cliques with different temporal localities to learn the complex movements and their short- and long-term dependencies. Simulations show that the new approach has superior performance to the state-of-the-art works on the widespread UCF-Crime and ShanghaiTech datasets.

The main contributions of this paper include: i) we develop a relation-aware feature extractor, incorporating TRN [11] with multi-scale partitions and an inner product operation to generate salient features; ii) we combine self-attention with CRFs to effectively learn the dynamic behaviours of the actors with different temporal localities in videos; iii) we devise an effective contrastive MIL scheme to broaden the margin between the normal and abnormal instances in videos.

2. Related Work

Unsupervised Anomaly Detection. One way to capture the abnormal events within videos is through unsupervised learning schemes such as optimizing one-class classification model, auto-encoder decoder [13, 14], and feature reconstruction with generative adversarial learning [15]. For example, Hasan *et al.* [14] used an auto-encoder to capture abnormal events using reconstruction error. Yu *et al.* [13] proposed an adversarial event prediction to detect anomalies based on the chronological event predictions. However, these methods can produce more false alarm due to the deficiency of the variations of normal samples [14].

Temporal Reasoning. Temporal reasoning, learning the information from the past data to benefit the current detection and future prediction, has been of importance in computer vision [16, 17]. Santoro *et al.* [16] used a relation network to exploit the foreknowledge information in visual question-answering tasks. Also, Sermanet *et al.* [17] considered a time contrast network, generating the relational features of the objects and actors, to conduct self-supervised learning in action recognition.

Attention Mechanism. With its effective learning capability, attention mechanism has been widely employed in a variety of applications [18, 19]. Zhao *et al.* [18] proposed a feature attention module to boost the low-level spatial features and high-level context for saliency detection. Zhang *et al.* [20] combined generative adversarial networks with an attention network to localize some specific attributes in face images. Purwanto *et al.* [21] employed self-attention to better extract the long-term temporal dependencies of the actions in low-resolution videos.

Graphical Models. The graphical models have received lots of attention in recent years [22, 23]. For instance, Si *et al.* [23] employed an attention augmented graph convolu-

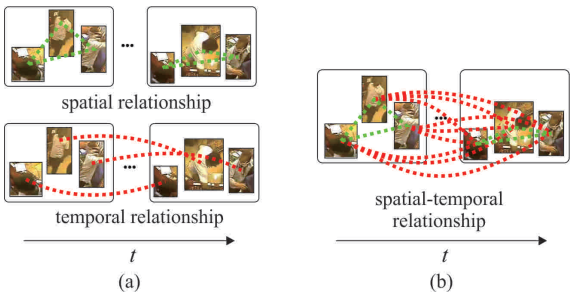


Figure 3: An illustration of the key difference of the spatial-temporal graphs between [12] and our work: (a) [12] models the spatial and temporal relationships between nodes independently; (b) ours jointly learn them.

tional LSTM network to capture the actions from skeleton data. Li *et al.* [24] combined deep relational modelling and feature learning to detect people from a low-dimensional feature representation.

3. Proposed Method

3.1. Overall Methodology

For reference, an overview of the proposed work is depicted in Fig. 2, which encompasses three main building blocks: relation-aware feature extractor in Sec. 3.2, self-attention CRF in Sec. 3.3, and a contrastive MIL scheme in Sec. 3.4. The first block, relation-aware feature extractor, is employed to capture the multi-scale CNN features that are essential to detecting global or local anomalous behavior in each frame. The second block is the self-attention CRF, where self-attention is used to capture the short-range correlations of the features, while CRF is employed to model the inter-dependencies of these features. Finally, as the problem is weakly supervised learning, the third block combines MIL with the contrastive loss to widen the gap between normal and abnormal instances, resulting in more accurate detection.

3.2. Feature Extraction

This section considers a new feature extractor, an extension of the TRN in [11]. TRN can well learn and reason the temporal dependencies across frames at multiple time scales and has been employed in a variety of computer vision tasks [11]. However, the input to TRN is the whole frame, for which either the foreground or the background are treated equally. Thereby, it is inadequate to reason some objects appearing within specific regions, which is beneficial for anomaly detection problems [25, 26].

To resolve this dilemma, we modify TRN in two aspects: First, we partition every frame into multiple scales; each of the partitioned image is referred to as an image patch. Depending on the partitioned scale, an image patch corresponds to either global or local features, which provide context and fine-grained appearance information, respectively. Specifically, from the output of the convolutional blocks, we can obtain the global features, which correspond to the whole frame representation, and the local features that are based on the multiple-scale grid partitions. This multi-scale image patches allow the network to better learn the anomalous patterns in some specific regions [25].

Second, to fully exploit the connections between the objects within neighboring frames where the anomalies can be similar, we make use of the inner-product operation [27], an effective scheme to encode a sequence of data [23, 19], to unleash the short-range correlations of the image patches. More specifically, let the intermediate feature maps of the network be $\mathbf{D} \in \mathbb{R}^{K\times W\times H\times C}$, where K is the length of the video, C is the dimension of the channel, and W and H denote the width and height of the features maps, respectively. The inner-product, $\gamma(a, a') = \psi_1(\mathbf{d}_a)^T\psi_2(\mathbf{d}'_a)$, is now used to decide the short-term dependencies between the positions a and a' in the feature maps in the neighboring frames, where $\mathbf{d}_a, \mathbf{d}'_a \in \mathbf{D}$ denote the features at positions a and a', respectively, and ψ_1 and ψ_2 are linear embedding layers. Collecting all of these dependencies at position a yields $[\gamma(a, 1), \cdots, \gamma(a, KWH)]$, which is concatenated with the intermediate feature maps and then transformed into a scalar attention weight at the corresponding position by multiplying with a trainable weight to re-weigh the output of the convolutional blocks.

To reduce the complexity, we apply the global average pooling to image patches. These features are then concatenated into $\mathbf{B} = [\mathbf{b}_{1,1}, \cdots, \mathbf{b}_{K,G}]$, where $\mathbf{b}_{i,j}$ denotes the feature of image patch i in frame j, and G is the total number of image patches in each frame. As a results, the corresponding image patches' features can produce a fruitful semantic information about the abnormal activities.

3.3. Self-Attention Conditional Random Field

3.3.1 Spatial-Temporal Graph Model

Spatio-temporal graphs are an effective approach to modeling the long-term dependencies among the objects to encapsulate their dynamic interactions [28, 29, 30], and learning the short-term dependencies to capture the abnormal events occurring in a short-time interval, as illustrated in Fig. 1. For a temporal window of K frames, say frames 1 to K, with G image patches in each frame, we can establish a fully-connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{(i, j)|i \in \{1, \cdots, G\}, j \in \{1, \cdots, K\}\}$ is the set of nodes, each of which corresponds to an image patch; $\mathcal{E}$ is the set of edges that connect each pair of node in the graphs. Each node (i, j) is related to a feature $\mathbf{b}_{i,j} \in \mathbf{B}$ derived from the relation-aware feature extractor in Sec. 3.2. As an illustration in Fig. 4, a spatial-temporal graph is established, where the partition scale is 3, leading to 14 image patches in each frame.

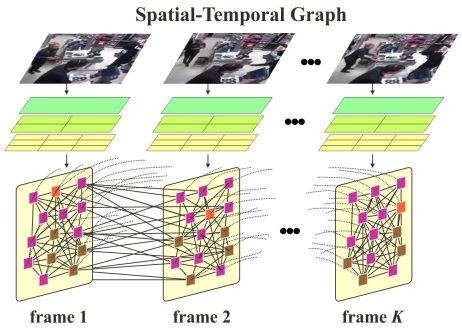


Figure 4: An illustration of a spatio-temporal graph, where the partition scale is 3, resulting in 14 image patches in each frame.

The weight of the edge connecting node (i, j) with node (i', j'), $p(\mathbf{b}_{i,j}, \mathbf{b}_{i',j'})$, is now assigned by the pairwise similarity between the corresponding nodes' features:

$$p(\mathbf{b}_{i,j}, \mathbf{b}_{i',j'}) = \exp(\theta_1(\mathbf{b}_{i,j})^T \theta_2(\mathbf{b}_{i',j'})), \quad (1)$$

where θ_1 and θ_2 are linear embedding functions. With this function, two nearby objects which closely interact with each other such as '*robber*' and '*cashier*' will have a stronger relation than those that are not related to each other like '*robber*' and '*foodstuff cabinet*'.

3.3.2 Conditional Random Fields

As an anomaly can be formed by spatio-temporal interactions among several objects, understanding the intrinsic relationships among the objects in the neighboring frames is thus of importance for anomaly detection [25]. Thereby, CRFs, which possess the advantages of graphical modeling and discriminative classification, is invoked here to model the interactions among the global and local features across frames to capture their context relationships.

Consider a set of node labels $\mathcal{X} = \{0, 1\}$, in which 0 denotes the label for normal patterns while 1 the abnormal ones, and a set of random variables $\mathbf{z} = \{z_{1,1}, \cdots, z_{G,K}\}$, where $z_{i,j}$ is a random variable associated with node (i, j) and is assigned with a label in $\mathcal{X}$. The graph can be learnt by CRFs conditioned on all nodes' features $\mathbf{B}$, which can be characterized by the Gibbs distribution $P(\mathbf{z}|\mathbf{B}) = \frac{1}{n(\mathbf{B})}\exp(-E(\mathbf{z}|\mathbf{B}))$, in which $E(\mathbf{z}|\mathbf{B})$ denotes the energy of the label assignment and $n(\mathbf{B})$ is the partition function [31]. For a fully-connected CRF model, the total energy can be written as [32]

$$\begin{aligned} E(\mathbf{z}|\mathbf{B}) = &\sum_{i,j} \phi_u(z_{i,j}|\mathbf{b}_{i,j}) \\ &+ \sum_{i,j} \sum_{(i',j') \neq (i,j)} \phi_p(z_{i,j}, z_{i',j'}|\mathbf{b}_{i,j}, \mathbf{b}_{i',j'}), \end{aligned} \quad (2)$$

where $\phi_u(z_{i,j}|\mathbf{b}_{i,j})$ is the unary energy, the cost of assigning a label to node (i, j); $\phi_p(z_{i,j}, z_{i',j'}|\mathbf{b}_{i,j}, \mathbf{b}_{i',j'})$ is the pairwise energy, the cost of assigning labels to node (i, j) by considering its relationship with node (i', j'). However, the conventional CRFs are not easy to be amalgamated with CNN networks [33]. Moreover, the non-local dependencies among the nodes is not fully leveraged, resulting in inaccurate detection when the abnormal events involve many objects, such as in the crowded scenarios. To overcome these setbacks, we resort to self-attention to be discussed next.

3.3.3 New Self-Attention

Compared with convolutional networks or RNNs, self-attention can attend the response at each position with other distant positions directly without encountering vanishing gradients. Additionally, self-attention can render faster computations compared to RNNs [34] with even fewer parameters. As such, we make use of self-attention to model the relationships of the nodes, local or non-local, in the spatio-temporal graphs described in Sec. 3.3.1.

To deal with both short- and long-term relationships of the features, we consider a set of complete sub-graphs, cliques, with different temporal localities. A clique $\mathcal{C}_j^\tau$ is said to be with a *temporal locality* τ if it is constructed by connecting every pair of nodes in frame j and its $(\tau - 1)$ adjacent frames. Note that $\mathcal{C}_j^\tau$ amounts to a (τG)-clique [35], both of which consist of τG nodes. While a clique with a small temporal locality is intended to underscore the short-term dependencies between the neighboring frames, a large one can highlight the non-local connections between the nodes. Thereby, if we apply the self-attention mechanism to relate node (i, j) with the other nodes in $\mathcal{C}_j^\tau$, we can obtain the self-attention output with temporal locality τ, $\bar{\mathbf{h}}_{i,j}^\tau \in \mathbb{R}^{1\times F}$, given by [27]

$$\bar{\mathbf{h}}_{i,j}^\tau = \sum_{\forall (i',j') \in \mathcal{C}_j^\tau} p(\mathbf{b}_{i,j}, \mathbf{b}_{i',j'}) \theta_3(\mathbf{b}_{i,j}), \quad (3)$$

where $p(\cdot)$ is the pairwise similarity function defined in (1) and θ_3 is a linear embedding layer [27].

To exploit various non-local relationships of the features, the resulting self-attention output of node (i, j), $\mathbf{h}_{i,j}$ is defined as a superposition of the self-attention outputs of $\bar{\mathbf{h}}_{i,j}^\tau$, $\tau \in \mathcal{K}$, and can be expressed as

$$\mathbf{h}_{i,j} = \sum_{\tau \in \mathcal{K}} w_\tau \bar{\mathbf{h}}_{i,j}^\tau, \quad (4)$$

where $\mathcal{K}$ is a prescribed set of temporal localities, and w_τ is the trainable scalar weight. The self-attention output of all of the nodes in these K consecutive frames can then be represented as $\mathbf{H} = [\mathbf{h}_{1,1}, \cdots, \mathbf{h}_{G,K}] \in \mathbb{R}^{KG\times C}$.

3.3.4 Conditional Random Fields with Self-Attention

In light of the advantage of self-attention discussed above, we utilize it to offset the deficiencies of CRFs described

in Sec. 3.3.2 in modelling the non-local relationships of the nodes, where self-attention is employed to calculate the total energy of the label assignment in (2).

As [33], the unary energy of assigning the label $z_{i,j} \in \mathbf{z}$ to node (i, j) can be determined by invoking a linear feed-forward classifier, $f_u(\cdot)$, to the node features $\mathbf{b}_{i,j}$ derived from the relation-aware feature extractor in Sec. 3.2

$$e^u_{i,j} = f_u(\mathbf{b}_{i,j}). \tag{5}$$

For a more compact representation, the unary energy of all nodes, $\mathbf{E}_u = [e^u_{i,j}, \cdots, e^u_{G,K}]$, can be computed by $\mathbf{E}_u = f_u(\mathbf{B})$.

Likewise, in light of the relational reasoning of graphs based on feature similarity in [22, 32], the total pairwise energy of assigning the same label to node (i, j) based on its correlations with all of the other nodes, $e^p_{i,j} = \sum_{(i',j')\neq(i,j)} \phi_p(z_{i,j}, z_{i',j'}|\mathbf{b}_{i,j}, \mathbf{b}_{i',j'})$, can be modelled by applying the self-attention mechanism to the cliques with different temporal localities as follows

$$e^p_{i,j} = \sum_{(i',j')\neq(i,j)} u(z_{i,j}, z_{i',j'})\hat{p}(\mathbf{b}_{i,j}, \mathbf{b}_{i',j'})f_p(e_3(\mathbf{b}_{i,j})), \tag{6}$$

where f_p is a linear feed-forward layer [32] that transforms the node representation into a label prediction; $u(z_{i,j}, z_{i',j'})$ is the compatibility function which, as [33, 36], can be decided by the Potts model, where $u(z_{i,j}, z_{i',j'}) = 1$ if $z_{i,j} \neq z_{i',j'}$ and equal 0, otherwise; we have also used the fact that $\hat{p}(\mathbf{b}_{i,j}, \mathbf{b}_{i',j'}) = \sum_{\tau\in\mathcal{K}} \sum_{i'\in\mathcal{C}^\tau_j} w_\tau p(\mathbf{b}_{i,j}, \mathbf{b}_{i',j'})$ [37]. To mitigate the computational overhead in (6), we have also confined the estimation of pairwise energy based on a prescribed set of cliques.

Again, the total pairwise energy of all nodes, $\mathbf{E}_p = [\epsilon^p_{1,1}, \cdots, \epsilon^p_{G,K}]$, can be compactly represented as

$$\mathbf{E}_p = f_p(\mathbf{H})\mathbf{U}, \tag{7}$$

where $\mathbf{H}$ is the self-attention outputs of all nodes and $\mathbf{U}$ is a symmetric matrix that can be trained to provide the data-dependent penalty [33], so it can incur low cost for assigning the same label to a pair of nodes with similar properties. The minimization of the total energy of the labelling can be conducted by the mean-field inference.

3.3.5 Mean-Field Inference

Next, we approximate the Gibbs distribution of the labels via the mean-field inference using a product of independent marginal distributions of all nodes, $W(\mathbf{z}) = \prod_{i,j} W_{z_{i,j}}$, where $W_{z_{i,j}}$ is determined by the unary and pairwise energy given by [37]:

$$W_{z_{i,j}} = \frac{1}{Z_{i,j}} \exp\Big(-(e^u_{i,j} + e^p_{i,j})\Big), \tag{8}$$

Algorithm 1 Mean-field inference of the self-attention CRF

Input: $\mathbf{B}$, max_iter ▷ nodes features, number of iterations
Output: $\hat{\mathbf{E}}$ ▷ marginal distribution of all nodes
1: $l = 0$
2: $\mathbf{E}_u \leftarrow \mathrm{f_u}(\mathbf{B})$ ▷ Computation of the unary energy
3: $\mathbf{E} \leftarrow \frac{1}{Z_{i,j}} \exp(-\mathbf{E}_u)$ ▷ initialize the marginal distribution by the unary energy
4: $\hat{\mathbf{E}} \leftarrow \mathrm{softmax}(\mathbf{E})$
5: **while** $l \leq$ max_iter **do**
6: $\quad \mathbf{E}_p \leftarrow f_p(\mathbf{H})$ ▷ computation of pairwise energy
7: $\quad \mathbf{E}_p \leftarrow \mathbf{E}_p\mathbf{U}$ ▷ compatibility transform
8: $\quad \mathbf{E} \leftarrow \hat{\mathbf{E}} - \mathbf{E}_p$ ▷ unary addition
9: $\quad \mathbf{E} \leftarrow \frac{1}{Z_{i,j}}(\mathbf{E})$ ▷ normalization of the marginal distribution
10: $\quad \hat{\mathbf{E}} \leftarrow \mathrm{softmax}(\mathbf{E})$ ▷ update of the marginal distribution
11: $\quad l \leftarrow l + 1$
12: **end while**

where $Z_{i,j}$ is the normalization constant [37]. Unlike the common inference scheme that stacks CNN kernels to revise the marginal distributions [33, 36, 38], here we consider a new mean-field inference algorithm which learns the non-local relationships of the nodes with self-attention by cast the the mean-field inference as a self-attention network.

The overall iterations in the new mean-field inference are summarized in Algorithm 1, where the marginal distribution $\hat{\mathbf{E}}$ is first initialized by the unary energy $\mathbf{E}_u$ in Step 3, whose element $e^u_{i,j}$ is derived by passing every node feature $\mathbf{b}_{i,j}$ through a linear transformation $f_u(\cdot)$ by (5). At each iteration, the pairwise energy $\mathbf{E}_p$ is attained by message passing in Step 6, which applies the linear transformation $f_p(\cdot)$ to the self-attention outputs $\mathbf{H}$ derived in Sec. 3.3.4. Thereafter, in Step 7, the compatibility transform is conducted by post-multiplying $\mathbf{E}_p$ by $\mathbf{U}$ where $\mathbf{U}$ is a trainable matrix that is employed to learn the correlation of the binary label assignment of different nodes. Subsequently, the marginal distribution is refined by subtracting the unary energy in Step 8 and then taking normalization in Step 9. Lastly, the marginal distribution is obtained by using a softmax layer [33] in Step 10. The resulting $\hat{\mathbf{E}}$ after convergence is taken as the final marginal distribution.

3.4. Contrastive Multi-Instance Learning

Since most of the commonly used benchmarks are only with video level annotations, we take advantage of MIL, which has been shown to be effective to learn the anomaly-norm from normal and abnormal bags in weakly-supervised manner [39]. To train an MIL model, we use the vectors in $\mathbf{H}, \mathbf{h}_{i,j}$, each of which corresponds to either a normal or an abnormal sample. Next, the normal and abnormal samples are congregated as negative and positive bags, $\mathcal{B}_n$ and $\mathcal{B}_p$, respectively. An MIL model f_s is then trained with a regression neural network to generate the anomaly score, $v_{i,j}$, for each $\mathbf{h}_{i,j}$ by $v_{i,j} = f_s(\mathbf{h}_{i,j})$. However, the traditional MIL [1] has two main limitations in this problem: i) it does not take account of the underlying temporal contexts of the

abnormal events; ii) most videos are composed of normal events while the anomalies occur only in few segments, resulting in an inaccurate detection [5].

In light of this, we consider a new loss function, which incorporates the contrastive loss with the conventional MIL [1]. Instead of only calculating the hinge loss, which has some limitations discussed above, the new loss function is given by

$$\mathcal{L}_{total} = \mathcal{L}_{bn} + \alpha_1 \mathcal{L}_{sp} + \alpha_2 \mathcal{L}_{ts} + \mathcal{L}_{cs}, \tag{9}$$

where $\mathcal{L}_{bn}$ is the binary cross entropy loss [7] based on the marginal distribution of all samples $\hat{\mathbf{E}}$ obtained by the mean-field inference in Sec. 3.3.5; $\mathcal{L}_{sp}$ and $\mathcal{L}_{ts}$ denote, respectively, the sparsity and temporal smoothness losses [1], employed based on the assumption that the anomalies rarely occur within a video to smooth out the anomaly scores between the adjacent video segments, and α_1, α_2 are the balancing parameters.

The last term $\mathcal{L}_{cs}$ is a contrastive loss. In contrast to [40] that is based on complex data augmentation to effectively learn various image representation, our contrastive loss is aiming at broadening the distance between the embedding of $\mathcal{B}_n$ and $\mathcal{B}_p$, and can be expressed as

$$\mathcal{L}_{cs} = \frac{1}{(|B_p|.|B_n|)^2} \sum_{(i,j)\in\mathcal{B}_p,(i',j')\in\mathcal{B}_n} ||\mathbf{h}_{i,j} - \mathbf{h}_{i',j'}||_2^2, \tag{10}$$

where $|B_p|$ and $|B_n|$ denote the cardinality of the positive and negative bags, respectively, and $||\mathbf{h}_{i,j} - \mathbf{h}_{i',j'}||_2$ is the Euclidean distance between the two samples $\mathbf{h}_{i,j}$ and $\mathbf{h}_{i',j'}$.

4. Experimental Results

4.1. Datasets and Evaluation Metric

UCF-Crime [1]. This dataset is the large-scale anomaly detection dataset that contains 1,900 videos with 13 types of anomaly events captured by CCTV camera indoors and outdoors during day and night scenarios. The activities consist of *Abuse*, *Arrest*, *Assaults*, *Shooting*, *Arson*, *Stealing*, *Explosion*, *Road Accidents*, *Shoplifting*, *Fighting*, *Robbery*, *Vandalism*, and *Burglary*.

ShanghaiTech [41]. It comprises of 437 videos, ranging in duration from 15 seconds to more than a minute in a variety of circumstances and illuminations, such as complex lighting conditions and multiple camera angles, recorded by CCTV camera in an outdoor location.

Evaluation Metrics. The simulations mainly follow the protocols provided by UCF-Crime [1] and ShanghaiTech [2]. We quantify the detection performance in terms of the area under the curve (AUC) of the corresponding frame based on the receiver operating characteristic curves [42]. Moreover, we utilize the false alarm rate (FAR) as another metric in which the model is tested only on normal videos with the threshold of 0.5% [1].

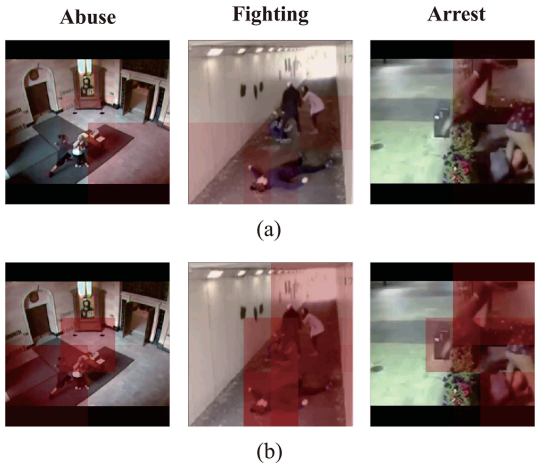


Figure 5: Visualization of the heatmaps (a) before and (b) after employing the new self-attention CRF.

4.2. Implementation Details

We use TRN [11] with a pretrained ResNet-50 model [43] as our backbone architecture with the momentum, the weight decay, and the base learning rate being set as 0.9, 0.0005, and 0.0001, respectively. The SGD optimizer is used to optimize this network. For UCF-Crime, the warming up learning rate in [44] is used for the first 10 epochs and linearly increases to the base learning rate to handle the over-fitting problem. The dropout is set as 0.8 and the partial batch normalization strategy [45] is invoked to train the models. The models are trained for 100 and 225 epochs for UCF-Crime and ShanghaiTech, respectively. The batch size is set as 18 and 16 for UCF-Crime and ShanghaiTech, respectively. The partition scale is set as 3, implying a total of $G = 14$ image patches in each frame and the global max pooling is applied to each of it to reduce the complexity in constructing the graph model. The dynamic halting is employed to decided an appropriate number of iterations for each video in the mean-field inference. The contrastive MIL is trained with Adam optimiser with a weight decay of 0.00001 and a batch size of 32 for 50 epochs. The learning rate is set as 0.0001. α_1 and α_2 are set as $8e^{-5}$. For UCF-Crime, we mainly follow the evaluation protocols provided by [1] while for ShanghaiTech, we adopt the binary-classification split-set provided by [2].

4.3. Ablation Studies

Impact of Each Module. We investigate the detection performance using different combinations of modules, as shown in Table 1, from which we can see that the proposed relation-aware network surpasses TRN by 1% and 0.42% on the UCF-Crime and ShanghaiTech datasets, respectively. This is because the new relation-aware feature extractor can learn the fine-grained information, which is essential to detecting the abnormal behaviours of small objects. We can

Table 1: Comparison with different combinations of modules. The best results are bold-faced.

TRN	Relation-aware network	Self-Attention CRF	Contrastive Loss	Datasets UCF-Crime	Datasets ShanghaiTech
✓	-	-	-	81.52	95.01
-	✓	-	-	82.43	95.43
-	✓	✓	-	83.89	96.67
-	✓	✓	✓	**85.00**	**96.85**

Table 2: Performance comparison using different sets of cliques in deciding the new self-attention. The best results are bold-faced.

Number of Cliques	Datasets UCF-Crime	Datasets ShanghaiTech
{1}	83.03	95.32
{1,2}	83.42	95.84
{1,2,4}	84.06	96.36
{1,2,4,6}	84.98	96.83
{1,2,4,6,8}	**85.00**	**96.85**

also notice that by employing our self-attention CRF to model the dynamic behaviours of the global and local abnormal features with multiple scales of temporal locality, the detection performance can be further enhanced by 1.4% and 1.1% on UCF-Crime and ShanghaiTech, respectively. Finally, with the use of the new contrastive loss, the performance can be further boosted by 0.18% to 1.01% due to its capability to broaden the gap between the normal and abnormal samples in weakly supervised training.

To further demonstrate the effectiveness of the combination of self-attention and CRF, we also provide the heatmaps before and after employing this new scheme, as shown in Fig. 5, from which we can see that by employing the proposed self-attention CRF, dynamic behaviours of the multiple actors involving in anomaly events can be more precisely identified.

Impact of the New Self-Attention. Next, we examine the performance of the proposed method using different sets of cliques, $\mathcal{C}_j^\tau$, in the computation of the new self-attention, as shown in Table 2, from which we can note that the performance improves incrementally by using a set containing more cliques. This is because with more cliques the self-attention can more substantially highlight the local and non-local relations between the two actors to render more precise detection performance.

4.4. Performance Analysis

To provide further insights into our approach, we also provide some successful and failure detection results, as shown in Figs. 6 and 7, respectively. For UCF-Crime, which contains more involved action scenarios, we can see from Fig. 6(a) that our approach can well detect human anomaly activity such as '*Shoplifting*'. However, our method cannot well distinguish a man approaching the car from the incidence of the arson, as shown in Fig. 7(a). This is because the dark situation resulting in a substantial loss of visual information. For ShanghaiTech, which mostly contains small objects in outdoor scenarios, we can observe

Table 3: Comparison with the state-of-the-art works on the UCF-Crime dataset. The best results are bold-faced. † indicates use optical flow, ◇ uses two-stream network, while others only use RGB.

Supervision	Method	Source	Backbone	Performance AUC	Performance FAR
Unsupervised	Hasan *et al.* [14]	CVPR16	-	50.6	27.2
	Sun *et al.* [46]	MM20	TCN	72.7	-
	Yu *et al.*◇ [13]	TNLSS21	3DCNN	81.84	-
Fully supervised	Liu et al. [7]	MM19	NLN	82	-
Weakly supervised	Sultani *et al.* [1]	CVPR18	C3D	75.41	1.9
	Lin *et al.* [3]	AVSS19	C3D	78.28	-
	Hao *et al.* [6]◇	SCN20	ResNet	78.51	-
	Zhang *et al.* [4]	ICIP19	TCN	78.66	-
	Zhu *et al.* [5]†	BMVC19	I3D	79	-
	Zaheer *et al.* [8]	SPL21	C3D	79.54	-
	Zhong *et al.* [2]	CVPR19	TSN	82.12	0.1
	Wu *et al.* [10]	ECCV20	I3D	82.44	-
	Zaheer *et al.* [9]	ECCV20	C3D	83.03	-
	Ours		Relation-aware	**85.00**	**0.024**

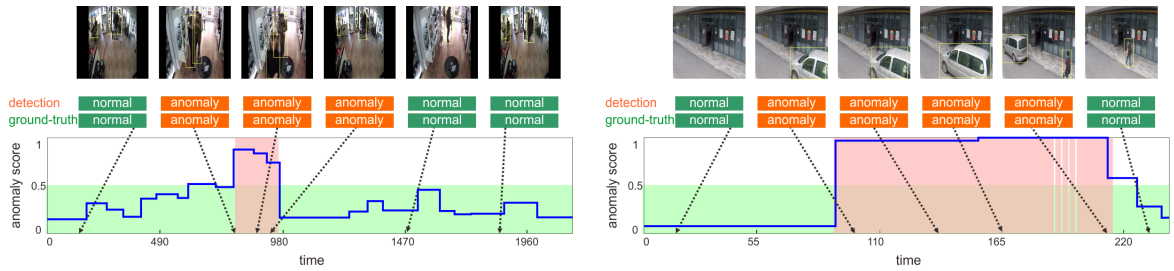
Table 4: Comparison with the state-of-the-art works on the ShanghaiTech dataset. The best results are bold-faced. ◇ uses two-stream network, while others only use RGB.

Supervision	Method	Source	Backbone	Performance AUC	Performance FAR
Unsupervised	Hasan *et al.* [14]	CVPR16	-	60.85	-
	Gong *et al.* [15]	ICCV19	-	71.2	-
	Yu *et al.* [47]◇	MM20	-	74.48	-
Weakly supervised	Zhang *et al.* [4]	ICIP19	TCN	83.5	0.1
	Zaheer *et al.* [8]	SPL21	C3D	84.16	-
	Zhong *et al.* [2]	CVPR19	TSN	84.44	-
	Zaheer *et al.* [9]	ECCV20	C3D	89.67	-
	Wan *et al.* [48]	ICME20	I3D	91.24	0.27
	Hao *et al.* [6]◇	SCN20	ResNet	94.2	-
	Ours		Relation-aware	**96.85**	**0.004**

from Fig. 6(b) that the '*car is passing*' can be well detected. On the other hand, as shown in Fig. 7(b), since the motorbike is located far from the surveillance camera at the beginning of the video, the appearance becomes unclear and thus the proposed method can only detect the anomaly activity when the motorbike is passing nearby the camera.

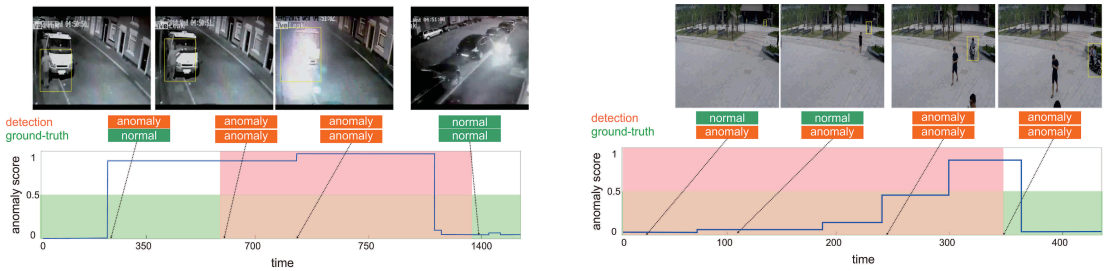
4.5. Comparison with the State-of-the-Art Works

This subsection compares the proposed approach with the main state-of-the-art works in terms of AUC and FAR on the UCF-Crime and ShanghaiTech datasets. For UCF-Crime, our comparison is as shown in Table 3, from which we can see that except [13], the unsupervised-based methods, [14, 46] in general provide inferior performance due to a lack of a variety of training data. The performance of [13] is boosted by using chronological data that learn both of the past and future frames extracted by a 3DCNN model. [7], a fully supervised method, provides better performance by using an anomaly-guided network to learn the abnormal patterns. For weakly supervised approaches, we can note that [3] attains better performance than [1] by employing a dual-branch network to effectively learn semantic information across video frames. Similarly, [6] obtains slight improvement by considering a two-stream network to better learn the motion cue information. [4] and [5]



(a) Shoplifting (b) Car is passing

Figure 6: Successful cases on the (a) UCF-Crime and (b) ShanghaiTech datasets. The red block indicates the ground truth of the anomalies while the blue lines indicate the anomaly scores across time.



(a) Arson (b) Motorcycle is passing

Figure 7: Failure cases on the (a) UCF-Crime and (b) ShanghaiTech datasets. The red block indicates the ground truth of the anomalies while the blue lines indicate the anomaly scores across time.

can obtain slightly better detection performance by modelling the motion aware features. [8] can boost the performance of [4, 5] by utilizing a self-reasoning network to mitigate the noisy labels from anomaly data. A considerable gain is obtained in [2] by using a graph convolutional network to iteratively refine the output detection labels. [10], which exploits the long-term temporal dependencies by a holistic module, can slightly enhance the detection performance. Further improvement is achieved in [9] by employing a clustering scheme with a normalcy suppression module to maximize the distance between normal and abnormal data. Our method excels the above methods by modelling the global and local feature representation with a new self-attention strengthened CRF. Also, we can the note that our new method achieves the lowest FAR compared with other baselines which report this performance, as our method can minimize the outliers that lead to false detection in the normal samples through the contrastive loss.

Our comparison on ShanghaiTech is shown in Table 4, from which we can note that [15], augmenting the auto-encoder with a memory module, attains better detection compared with [14], in which the convolutional feed forward auto-encoder is considered. A substantial gain is achieved in [47] by exploiting the spatial information and motion cues to localize anomalous region. [2] surpasses [8, 4] by invoking self-reasoning during the network training. Impressive result is achieved in [9] that refines the noisy label with a deep clustering-based mechanism. [48] provides substantial improvement by using a dynamic loss incorporated with a regression network to yield more discriminative anomalous features. [6] attains better performance with a two-stream architecture to better model spatial and motion information. The new approach again outperforms the other methods by reasoning short- and long-term temporal dependencies across video frames using a self-attention CRF. Again, we can see that our method yields the lowest FAR compared with the state-of-the-art works.

5. Conclusions

This paper has developed a new network for weakly supervised anomaly detection. The network starts with learning the multi-scale features with a new relation-aware feature extractor. Afterwards, a CRF is employed to model the relationships of the global and local features with a newly devised self-attention. With such a combination, not only the anomalous patterns can be well identified, but the short- and long-term temporal dependencies across video frames can also be effectively learned. Moreover, a contrastive multi-instance learning scheme is considered to further boost the performance. Conducted experiments reveal the superiority of the new method on two popular anomaly detection datasets.

Acknowledgment

This work was supported by the Ministry of Science and Technology, R.O.C. under contracts MOST 110-2221-E-011-071-MY2 and MOT110-2221-E-011 -116.

References

[1] 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, 5, 6, 7

[2] 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 Conference on Computer Vision and Pattern Recognition*, pages 1237–1246, 2019. 1, 6, 7, 8

[3] Shuheng Lin, Hua Yang, Xianchao Tang, Tianqi Shi, and Lin Chen. Social mil: Interaction-aware for crowd anomaly detection. In *Proceedings of the IEEE International Conference on Advanced Video and Signal-based Surveillance*, pages 1–8, 2019. 1, 7

[4] Jiangong Zhang, Laiyun Qing, and Jun Miao. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *Proceedings of the IEEE International Conference on Image Processing*, pages 4030–4034, 2019. 1, 7, 8

[5] Yi Zhu and Shawn Newsam. Motion-aware feature for improved video anomaly detection. In *Proceedings of the British Machine Vision Conference*, pages 1–8, 2019. 1, 6, 7, 8

[6] Wangli Hao, Ruixian Zhang, Shancang Li, Junyu Li, Fuzhong Li, Shanshan Zhao, and Wuping Zhang. Anomaly event detection in security surveillance using two-stream based model. *Security and Communication Networks*, 2020, 2020. 1, 7, 8

[7] Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the ACM Multimedia*, pages 1490–1499, 2019. 1, 6, 7

[8] Muhammad Zaigham Zaheer, Arif Mahmood, Hochul Shin, and Seung-Ik Lee. A self-reasoning framework for anomaly detection using video-level labels. *IEEE Signal Processing Letters*, 27:1705–1709, 2020. 1, 7, 8

[9] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *Proceedings of the European Conference on Computer Vision*, pages 1–8, 2020. 2, 7, 8

[10] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European Conference on Computer Vision*, pages 322–339, 2020. 2, 7, 8

[11] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision*, pages 803–818, 2018. 2, 3, 6

[12] Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. In *Proceeding of the European Conference on Computer Vision*, pages 77–90, 2020. 2, 3

[13] Jongmin Yu, Younkwan Lee, Kin Choong Yow, Moongu Jeon, and Witold Pedrycz. Abnormal event detection and localization via adversarial event prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2, 7

[14] 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, 7, 8

[15] 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 International Conference on Computer Vision*, pages 1705–1714, 2019. 2, 7, 8

[16] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Proceeding of the Advances in Neural Information Processing Systems*, pages 4967–4976, 2017. 2

[17] Pierre Sermanet, Corey Lynch, Jasmine Hsu, and Sergey Levine. Time-contrastive networks: Self-supervised learning from multi-view observation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 486–487, 2017. 2

[18] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3085–3094, 2019. 2

[19] Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Extreme low resolution action recognition with spatial-temporal multi-head self-attention and knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–8, 2019. 2, 3

[20] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European Conference on Computer Vision*, pages 417–432, 2018. 2

[21] Didik Purwanto, Rizard Renanda Adhi Pramono, Yie-Tarng Chen, and Wen-Hsien Fang. Three-stream network with bidirectional self-attention for action recognition in extreme low resolution videos. *IEEE Signal Processing Letter Letters*, 26(8):1187–1191, 2019. 2

[22] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 2, 5

[23] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1227–1236, 2019. 2, 3

[24] Wei-Hong Li, Fa-Ting Hong, and Wei-Shi Zheng. Learning to learn relation for important people detection in still images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5003–5011, 2019. 3

[25] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981, 2010. 3, 4

[26] 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. 3

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceeding of the Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 3, 4

[28] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8561–8568, 2019. 3

[29] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016. 3

[30] Mahdi Khodayar and Jianhui Wang. Spatio-temporal graph deep neural network for short-term wind speed forecasting. *IEEE Transactions on Sustainable Energy*, 10(2):670–681, 2018. 3

[31] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, 2001. 4

[32] Hongchang Gao, Jian Pei, and Heng Huang. Conditional random field enhanced graph convolutional neural networks. In *Proceedings of the ACM International Conference on Knowledge Discovery & Data Mining*, pages 276–284, 2019. 4, 5

[33] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015. 4, 5

[34] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 4

[35] Richard D Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1):113–126, 1973. 4

[36] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *Proceeding of the European Conference on Computer Vision*, pages 524–540, 2016. 5

[37] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Proceeding of the Advances in Neural Information Processing Systems*, pages 109–117, 2011. 5

[38] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1831–1840, 2017. 5

[39] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The knowledge engineering review*, 25(1):1–25, 2010. 5

[40] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, pages 1597–1607, 2020. 6

[41] 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. 6

[42] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2720–2727, 2013. 6

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[44] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 6

[45] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on International Conference on Machine Learning*, pages 448–456, 2015 . 6

[46] Che Sun, Yunde Jia, Yao Hu, and Yuwei Wu. Scene-aware context reasoning for unsupervised abnormal event detection in videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 184–192, 2020. 7

[47] 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 ACM International Conference on Multimedia*, pages 583–591, 2020. 7, 8

[48] Boyang Wan, Yuming Fang, Xue Xia, and Jiajie Mei. Weakly supervised video anomaly detection via center-guided discriminative learning. In *Proceedings of the IEEE International Conference on Multimedia Expo*, pages 1–6, 2020. 7, 8