ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447603

LEARNING SPATIO-TEMPORAL RELATIONS WITH MULTI-SCALE INTEGRATED PERCEPTION FOR VIDEO ANOMALY DETECTION

Hongyu Ye[1], *Ke Xu*[1,*], *Xinghao Jiang*[1], *Tanfeng Sun*[1]

[1]Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

In weakly supervised video anomaly detection, it has been verified that anomalies can be biased by background noise. Previous works attempted to focus on local regions to exclude irrelevant information. However, the abnormal events in different scenes vary in size, and current methods struggle to consider local events of different scales concurrently. To this end, we propose a multi-scale integrated perception (MSIP) learning approach to perceive abnormal regions of different scales simultaneously. In our method, a frame is partitioned into several groups of patches with varying scales, and a multi-scale patch spatial relation (MPSR) module is further proposed to model the inconsistencies among multi-scale patches. Specifically, we design a hierarchical graph convolution block in the MPSR module to improve the integration of patch features by implementing cross-scale feature learning. An existing clip temporal relation network is also introduced to enable spatio-temporal encoding in our model. Experiments show that our method achieves new state-of-the-art performance on the ShanghaiTech and competitive results on UCF-Crime benchmarks.

Index Terms— video anomaly detection, multi-scale perception, weakly supervised, spatio-temporal relation

1. INTRODUCTION

Video anomaly detection (VAD) is the task of detecting abnormal events that differ from usual patterns and determining the time window of the occurring anomaly [1, 2, 3], which can be applied in many real-world scenarios [3].

Due to the high cost of manual annotations, most of the existing methods treat VAD as an unsupervised [4, 5, 6] or weakly supervised [1, 2] problem. A noticeable drawback of unsupervised VAD is the lack of prior knowledge of abnormality, resulting in the inability to capture all normalcy variations [7]. Therefore, unsupervised VAD generally has worse performance than weakly supervised methods, which can produce more reliable results using coarse-grained video-level labels to maintain a relatively small annotation effort.

*Corresponding author. This work is funded by the National Natural Science Foundation of China (62002220,62372295).

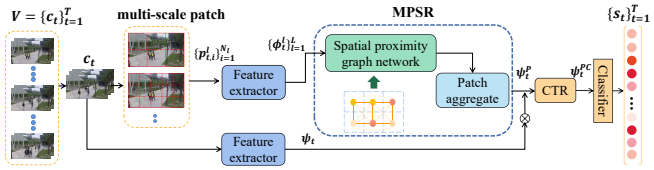

Fig. 1: The overall pipeline of our proposed MSIP.

In recent research, weakly supervised VAD commonly employs the multiple instance learning (MIL) framework [2, 8], which can alleviate the imbalance between abnormal and normal samples. However, weakly supervised anomaly detection still poses many unresolved challenges. One of the challenges is how to reduce the background noise (i.e., normal regions in a frame) that increases the difficulty in detecting anomalous events. Guoqiu Li et al. [9] divided the input video frames into several sets of non-overlapping patches with different scales and proposed a scale-aware model to explore anomalous patches. They trained the model in multiple stages, each dedicated to training a separate branch of the model to handle patches of a specific size. Another work chose [10] to apply an object detector to extract the object proposals, removing the background noise. However, the above methods either focus excessively on the objects and may miss the information about abnormal events, or ignore the connections between patches at different scales. Additionally, the step-by-step training strategy in [9] leads to a complex network and high computational demands.

To address the problems above, we propose a novel multi-scale integrated perception (MSIP) learning method with cascaded spatio-temporal relation networks and a cross-scale learning block. A concise multi-scale patch spatial relation (MPSR) network is proposed to model the inconsistencies among multi-scale patches simultaneously, enabling the identification of scale-varying anomalous regions. An existing clip temporal relation (CTR) module is introduced [8] to explore the temporal dependencies among clips, enabling spatio-temporal feature learning.

Graph convolutional networks (GCNs) have been applied to action recognition [11, 12] and video anomaly detection [3, 13, 14]. Furthermore, [15] proposed a hierarchical GCN for traffic forecasting. Inspired by their works, we design a hierarchical graph convolution (HGC) block with a similar structure to [15] to achieve cross-scale feature learning. Our

contributions are summarized below:

- We propose a novel multi-scale integrated perception learning method for weakly supervised video anomaly detection, which can perceive multi-scale patches simultaneously to capture scale-varying anomalies in video clips.
- A spatio-temporal relation network is introduced, comprising our proposed MPSR module and a CTR module.
- A hierarchical graph convolution block is proposed to achieve cross-scale feature learning in the MPSR module.
- Experiments on UCF-Crime and ShanghaiTech datasets show that our method can achieve competitive results with a simple training process and fewer parameters.

2. PROPOSED METHOD

2.1. Overview

The overall pipeline of our proposed MSIP is illustrated in Fig 1, including a multi-scale patch spatial relation (MPSR) module, a clip temporal relation (CTR) module, and a classifier. Given an input video V, and the video level annotation $Y \in \{0, 1\}$, which indicates whether the video contains anomalous events ($Y = 1$ for abnormal videos). Following the previous approach [2, 13, 8, 9], we divide the video along the temporal sequence into T non-overlapping clips $\{c_t\}_{t=1}^{T}$. For each clip $c_t \in \mathbb{R}^{H\times W\times F\times 3}$ ($H = height, W = width, F = frames$), we further subdivide it into several sets of non-overlapping patches with L different sliding window sizes $\{(h_l, w_l)\}_{l=1}^{L}$, where $l \in \{1, \ldots, L\}$ is the index of patch set l. These patch sets are represented as $\mathbf{P}_t^l = \{p_{t,i}^l\}_{i=1}^{N_l}$, and N_l is the number of patches in patch set l, $p_{t,i}^l \in \mathbb{R}^{h_l\times w_l\times F\times 3}$ denotes a patch of patch set l in the t_{th} clip. Each clip and corresponding multi-scale patches will be fed into a pre-trained I3D network to extract features. Feature of the t_{th} clip is represented as $\psi_t \in \mathbb{R}^D$, and the features of patches $p_{t,i}^l$ are stacked as $\mathbf{\Phi}_t = \{\phi_t^l\}_{l=1}^{L}, \phi_t^l \in \mathbb{R}^{N_l\times D}$, where D denotes the feature dimension. The features of multi-scale patches are then passed into our MPSR module, producing the aggregated patch feature $\chi_t \in \mathbb{R}^D$, which will be element-wise added to the clip feature ψ_t to obtain the patch-enhanced clip feature ψ_t^P. Finally, $\{\psi_t^P\}_{t=1}^{T}$ of T video clips will be input into the CTR module and a classifier to generate anomaly scores $\{s_t\}_{t=1}^{T}$.

2.2. Multi-scale Patch Spatial Relation Module

The multi-scale patch spatial relation module aims to recognize and spotlight the scale-varying anomalies by capturing the correlations between multi-scale patches. As shown in Fig 2, the patch features $\mathbf{\Phi}_t = \{\phi_t^l\}_{l=1}^{L=3}$ are sent to vanilla GCNs or a HGC block in spatial proximity graph network, the outputs are then concatenated and passed into patch aggregate module to obtain the aggregated feature $\chi_t \in \mathbb{R}^D$.

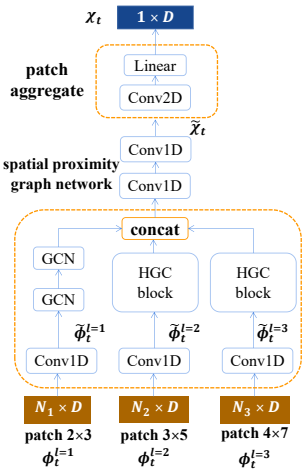

Fig. 2: Illustration of the proposed multi-scale patch spatial relation module.

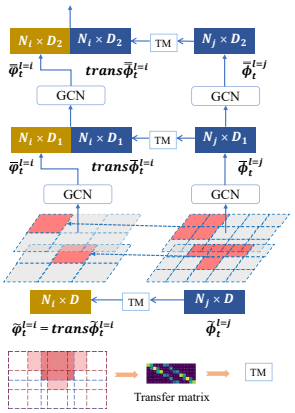

Fig. 3: Illustration of the proposed hierarchical graph convolution block.

Spatial proximity graph network. Motivated by previous work, we utilize the graph convolutional network (GCN) [16] to capture spatial relationships among patches. In our spatial proximity graph, an input patch set with N patches of the same scale is considered as the vertex set, where patch features correspond to the attributes of the vertexes. The adjacency matrix $A^S \in \mathbb{R}^{N\times N}$ of the graph is dependent on the relative proximity prior of the i^{th} and j^{th} patches:

$$A_{ij}^S = \exp(-\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}) \tag{1}$$

where (x_1, y_1) and (x_2, y_2) are the coordinates of the patch center. In the above construction, closer distances indicate a closer connection between patches and vice versa. As indicated in Fig 2 and 3, each branch contains two GCN layers. Specifically, for the l^{th} GCN layer, the graph convolution is implemented by:

$$X^l = A^S X^{l-1} W^l + f^l(X^{l-1}) \tag{2}$$

where $X^{l-1} \in R^{N\times D^{l-1}}$ are the hidden features of patches at layer $l-1$, and W_l is a trainable parametric matrix. Residual connection is adopted to alleviate the over-smoothing problem in GCN, and f^l is an inserted convolution layer.

Algorithm 1 Cross-scale feature transfer function F_{trans}

Input: $\tilde{\phi}_t^{l=j} \in \mathbb{R}^{N_j\times D}$, vertex coordinates of patches in set i $\{coord_r^{l=i} = [x_r^1, y_r^1, x_r^2, y_r^2]\}_{r=1}^{N_i}$, vertex coordinates of patches in set j $\{coord_s^{l=j} = [x_s^1, y_s^1, x_s^2, y_s^2]\}_{s=1}^{N_j}$, Intersection area function f_{ia}, patch area function f_a
Output: $trans\tilde{\phi}_t^{l=i} \in \mathbb{R}^{N_i\times D}$

```
transformation matrix Tran ∈ R^{N_i × N_j}, threshold τ = 1/3
for r = 1 → N_i do
    for s = 1 → N_j do
        S_overlap ⇐ f_ia(coord_r^{l=i}, coord_s^{l=j}); S_j ⇐ f_a(coord_s^{l=j})
        if S_overlap ÷ S_j > τ then
            Tran[r, s] ⇐ 1
        else
            Tran[r, s] ⇐ 0
        end if
    end for
end for
return trans φ̃_t^{l=i} ⇐ Trans × φ̃_t^{l=j}
```

Hierarchical graph convolution block. The varied spatial distributions of multi-scale patches pose a challenge for vanilla GCNs to effectively capture their relations, for which we propose a hierarchical graph convolution block to implement cross-scale feature learning. First, we aim to transform patch features of different scales into a uniform one by employing a transfer function as shown in Alg 1. Referring to Fig 3, the transformation matrix in the function is essentially equivalent to re-partitioning patch set j into overlapping regions that resemble the distribution of set i, and patch features in the same region are aggregated to obtain the transformed features. Given features $\widetilde{\phi}_t^{l=j} \in \mathbb{R}^{N_j \times D}$ of patch set j, we first use the transfer function to convert it into features $\widetilde{\varphi}_t^{l=i} \in \mathbb{R}^{N_i \times D}$ with the same scale as patch set i. $\widetilde{\phi}_t^{l=j}$ and $\widetilde{\varphi}_t^{l=i}$ will be passed to two different branches of GCNs and the output of each GCN layer in the branch of $\widetilde{\phi}_t^{l=j}$ will also be transformed using the above function and combined with that in the branch of $\widetilde{\varphi}_t^{l=i}$ by concatenating. Taking the output of the first layer GCN as an example, this process can be described as:

$$Combine(\bar{\varphi}_t^{l=i}, \bar{\phi}_t^{l=j}) = Concat(\bar{\varphi}_t^{l=i}, F_{trans}(\bar{\phi}_t^{l=j})) \quad (3)$$

where $\bar{\phi}_t^{l=j}$ and $\bar{\varphi}_t^{l=i}$ represent the output of the first layer GCN in the two branches. Thus, the graph convolution on the scale i will be affected by the convolution of the scale j, utilizing the spatial intersection relations between multi-scale patches to enable effective feature integration, which is the unique feature of our HGC block.

Patch aggregate. The patch aggregate module is designed to further aggregate patch features of different spatial regions. As Fig 2 shows, the outputs of different branches are concatenated to get the integrated feature $\widetilde{\chi}_t \in \mathbb{R}^{N_1 \times D}$. Inspired by previous work, we first reshape $\widetilde{\chi}_t$ into $\mathbb{R}^{H/h_1 \times W/w_1 \times D}$ according to the initial spatial location, and then input it into the patch aggregate module with a 2D convolutional and a fully connected layer, producing an aggregated multi-scale patches feature $\chi_t \in \mathbb{R}^D$.

2.3. Clip Temporal Relation Module

An existing clip temporal relation (CTR) module [8] is employed to learn the temporal correlations between video clips. The CTR consists of dilated convolutions [17] and a non-local block [18]. The non-local block is formulated as follows:

$$\tilde{\mathbf{\Psi}}^P = \mathrm{softmax}\left(f_\theta(\bar{\mathbf{\Psi}}^P) \times f_\varphi(\bar{\mathbf{\Psi}}^P)^\top\right) \times f_g(\bar{\mathbf{\Psi}}^P), \quad (4)$$

$$\hat{\mathbf{\Psi}}^P = f_z(\tilde{\mathbf{\Psi}}^P) + \bar{\mathbf{\Psi}}^P \quad (5)$$

where $\mathbf{\Psi}^P = [\psi_1^P, \psi_2^P, \ldots, \psi_T^P] \in \mathbb{R}^{T \times D}$ represents the patch-enhanced features of a T clips video, which is dimensionally reduced to obtain $\bar{\mathbf{\Psi}}^P \in \mathbb{R}^{T \times D/4}$. The final output $\mathbf{\Psi}^{PC} \in \mathbb{R}^{T \times D}$ is calculated as follows:

$$\mathbf{\Psi}^{PC} = [\hat{\mathbf{\Psi}}^P, \mathbf{\Psi}_*^P] + \mathbf{\Psi}^P \quad (6)$$

where $\mathbf{\Psi}_*^P$ denotes the outputs of dilated convolution layers.

2.4. Training Loss

The multiple instance learning (MIL) method is applied to our weakly supervised learning. Following previous works [9, 8], we adopt the feature magnitude learning method and corresponding feature magnitude ranking loss proposed by [8] for training to enhance the discrimination between anomalous and normal clips.

3. EXPERIMENTAL RESULTS

3.1. Datasets And Implementation Details.

Dataset. Experiments are conducted on two public datasets: ShanghaiTech [19] and UCF-Crime [2]. The ShanghaiTech dataset has 437 videos, including 130 anomaly videos. Following Zhong et al. [3], we reorganize the training and testing sets to make them suitable for weakly supervised VAD. UCF-Crime is a large-scale dataset that contains 1900 untrimmed real-world videos of 13 classes of anomalous events.

Implementation details. Following [8, 9], a video is split into 32 video clips ($T = 32$) and each video clip is resized into $480 \times 840 \times 16$ pixels (H, W, F). Then we segment frames into three patch sets of different scales: 240×280, 160×168, and 120×120, where 240×280 corresponds to the input of branch $l = 1$ in Fig 2. The I3D network pretrained on the Kinetic-400 dataset is leveraged to extract features of the full clip and corresponding patches. For hyper-parameters in the loss function and CTR module, we use the same settings as [8]. Our network is implemented in PyTorch and trained using a two-stage strategy. In the first stage, only the CTR module and the classifier are trained until the optimization process converges. In the second stage, the previously trained network is frozen, and the MPSR module is trained separately to obtain the final results. In both stages, we use Adam optimizer [20] with a learning rate of 0.001, a weight decay of 0.0005, and a batch size of 64 for all datasets.

3.2. Results.

Quantitative comparison. The quantitative comparison results between our method and other SOTAs are shown in Table 1. Our MSIP achieves new state-of-the-art on the Shanghaitech dataset with an AUC result of 98.00%, exceeding the SOTA method SSRL [9] by 0.02% and CLAV [21] by 0.40%. For the UCF-Crime dataset, our method still achieves suboptimal results with an AUC result of 86.98%, surpassing most previous methods. Notably, our MSIP outperforms the light version SSRL with shared parameters on both datasets, which is 98.00% compared to 97.84% on Shanghaitech and 86.98% compared to 86.85% on UCF-Crime, showing that our method achieves superior performance while demanding lower computational requirements.

Computational complexity. The comparison of parameter amount and computational cost is in Table 2. Our method

Table 1: Quantitative comparisons on ShanghaiTech [19] and UCF-Crime [2].The best scores are **bolded** and the second best are underlined.

Sup.	Method	Venue	Feature	AUC@ROC ↑	
				ShanghaiTech	UCF-Crime
Un-	GODS [22]	ICCV'19	I3D	–	70.46
	STC-Graph [5]	MM'20	-	74.70	72.70
	GCL_{PT} [23]	CVPR'21	ResNext	78.93	71.04
Weakly-	Zhong et al. [3]	CVPR'19	TSN	84.44	82.12
	GCL_{WS} [23]	CVPR'21	ResNext	86.21	71.04
	Zhong et al. [3]	CVPR'19	C3D	76.44	81.08
	CLAWS [6]	ECCV'20		89.67	83.03
	RTFM [8]	ICCV'21		91.57	83.28
	Sultani et al. [2]	CVPR'18	I3D	85.33	77.92
	Wu et al. [13]	ECCV'20		–	82.44
	MIST [1]	CVPR'21		94.83	82.30
	RTFM [8]	ICCV'21		97.21	84.30
	S3R [24]	ECCV'22		97.48	85.99
	SSRL [9]	ECCV'22		97.98	**87.43**
	SSRL(share parameters) [9]	ECCV'22		97.84	86.85
	UR-DMU [25]	AAAI'23		–	86.97
	CLAV [21]	CVPR'23		97.60	86.10
	Ours: MSIP		I3D	**98.00**	86.98

Table 2: Computational complexity comparisons.

Method	Feature	Param	FLOPs
RTFM(CTR) [8]	I3D	24.7M	7.9G
SSRL [9]	I3D	192.0M	57.7G
SSRL(share parameters) [9]	I3D	79.8M	57.7G
Ours: MSIP	I3D	75.2M	17.0G

minimally raises computational costs compared to the baseline and significantly reduces model parameters by 60.8% and floating-point operations by 70.5% compared to the SOTA method SSRL [9]. Therefore, our method not only achieves comparable results to SOTAs but also features a lightweight and computationally efficient model, striking a superior balance between performance and computational costs.

Visual results. Visual results are shown in Fig 4, which compares the predicted anomaly scores of our MSIP and the baseline [8] on three abnormal videos and one normal video. Evidently, our method distinguishes abnormal and normal events more effectively than the baseline and produces much fewer false positives on normal videos.

Table 3: Ablation study on MPSR and HGC in our method. The results of the baseline are marked with *.

CTR	MPSR	HGC	UCF-Crime (AUC@ROC ↑)	ShanghaiTech (AUC@ROC ↑)
✓	✗	✗	84.30*	97.21*
✓	✓	✗	86.09	97.37
✓	✓	✓	**86.98**	**98.00**

3.3. Ablation Studies.

Ablation studies are conducted on the proposed modules to investigate their influences on our model. As shown in Table 3, we employ the RTFM [8] as our baseline, corresponding to the case of using only CTR. By leveraging the MPSR module but replacing the HGC block with common GCNs, the performances increase to 86.09% on UCF-Crime and 97.37% on ShanghaiTech compared to the baseline. When both the

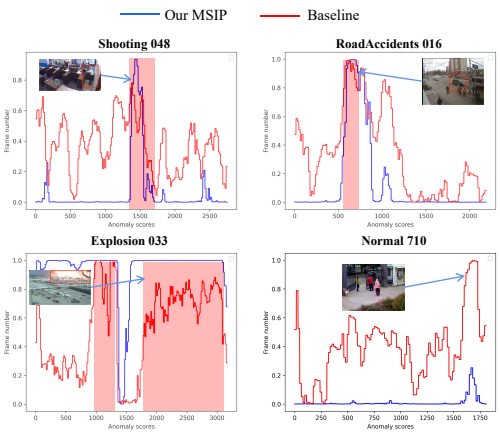

Fig. 4: Visualization of the anomaly scores of our MSIP and the baseline on UCF-Crime test videos. Pink areas are temporal ground truths of anomalies. Anomalous regions are denoted by red boxes.

Table 4: Ablation study on patch data scale variations in inputs. The results of the baseline are marked with *.

Patch Size			UCF-Crime	ShanghaiTech
240 × 280	160 × 168	120 × 120	(AUC@ROC ↑)	(AUC@ROC ↑)
✗	✗	✗	84.30*	97.21*
✓	✗	✗	85.85	97.69
✗	✓	✗	85.83	97.63
✗	✗	✓	85.49	97.58
✓	✓	✗	86.24	97.84
✓	✗	✓	86.71	97.74
✗	✓	✓	86.69	97.75
✓	✓	✓	**86.98**	**98.00**

MPSR module and HGC block are added, the performances further improve to 86.98% on UCF-Crime and 98.00% on ShanghaiTech, respectively. The above results indicate that both the MPSR module and HGC block contribute to our method in producing superior performances.

Ablation studies on multi-scale patch data are also conducted as presented in Table 4. When using patches of one scale, the results show growth within the 1.19% to 1.55% range in UCF-Crime and the 0.37% to 0.48% range in ShanghaiTech, and the results will gradually increase by adding more scales of patches. Eventually, the model achieves its best performance when utilizing patch data of all scales.

4. CONCLUSION

We proposed a novel multi-scale integrated perception learning method for weakly supervised VAD to focus on local anomalous regions with varying scales simultaneously. A relation network is introduced to learn spatio-temporal features, comprising our MPSR and a CTR module. Specifically, we designed a hierarchical graph convolution block in the MPSR module to learn cross-scale features for integration. Comprehensive experiments show that our method achieves significant improvements compared to SOTAs.

5. REFERENCES

[1] Jia-Chang Feng, Fa-Ting Hong, and Wei-Shi Zheng, "MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 14004–14013, IEEE.

[2] Waqas Sultani, Chen Chen, and Mubarak Shah, "Real-World Anomaly Detection in Surveillance Videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6479–6488, IEEE.

[3] 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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 1237–1246, IEEE.

[4] 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 1705–1714, IEEE.

[5] 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*. 2020, pp. 184–192, ACM.

[6] 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 *Computer Vision – ECCV 2020*, vol. 12367, pp. 358–376. Springer International Publishing, 2020.

[7] Nicolae-Catalin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B. Moeslund, and Mubarak Shah, "Self-Supervised Predictive Convolutional Attentive Block for Anomaly Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 13566–13576, IEEE.

[8] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W. Verjans, and Gustavo Carneiro, "Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 4955–4966, IEEE.

[9] Guoqiu Li, Guanxiong Cai, Xingyu Zeng, and Rui Zhao, "Scale-Aware Spatio-Temporal Relation Learning for Video Anomaly Detection," in *Computer Vision – ECCV 2022*, vol. 13664, pp. 333–350. Springer Nature Switzerland, 2022.

[10] Guoqiu Li, Shengjie Chen, Yujiu Yang, and Zhenhua Guo, "A Two-Branch Network for Video Anomaly Detection with Spatio-Temporal Feature Learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2023, pp. 1–5, IEEE.

[11] Xiaolong Wang and Abhinav Gupta, "Videos as Space-Time Region Graphs," in *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds., vol. 11209, pp. 413–431. Springer International Publishing, 2018.

[12] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 12018–12027, IEEE.

[13] 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 *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., vol. 12375, pp. 322–339. Springer International Publishing, 2020.

[14] Congqi Cao, Xin Zhang, Shizhou Zhang, Peng Wang, and Yanning Zhang, "Adaptive Graph Convolutional Networks for Weakly Supervised Anomaly Detection in Videos," *IEEE Signal Processing Letters*, vol. 29, pp. 2497–2501, 2022.

[15] Kan Guo, Yongli Hu, Yanfeng Sun, Sean Qian, Junbin Gao, and Baocai Yin, "Hierarchical graph convolution network for traffic forecasting," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, pp. 151–159, 2021.

[16] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[17] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2016.

[18] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 2018, pp. 7794–7803, Computer Vision Foundation / IEEE Computer Society.

[19] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao, "Future Frame Prediction for Anomaly Detection - A New Baseline," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6536–6545, IEEE.

[20] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[21] MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, and Sangyoun Lee, "Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. 2023, pp. 12137–12146, IEEE.

[22] Jue Wang and Anoop Cherian, "GODS: Generalized One-Class Discriminative Subspaces for Anomaly Detection," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 8200–8210, IEEE.

[23] M. Zaigham Zaheer, Arif Mahmood, M. Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee, "Generative Cooperative Learning for Unsupervised Video Anomaly Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 14724–14734, IEEE.

[24] Jhih-Ciang Wu, He-Yen Hsieh, Ding-Jie Chen, Chiou-Shann Fuh, and Tyng-Luh Liu, "Self-supervised Sparse Representation for Video Anomaly Detection," in *Computer Vision – ECCV 2022*, vol. 13673, pp. 729–745. Springer Nature Switzerland, 2022.

[25] Hang Zhou, Junqing Yu, and Wei Yang, "Dual Memory Units with Uncertainty Regulation for Weakly Supervised Video Anomaly Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, pp. 3769–3777, 2023.