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

Hongyu Ye, Ke Xu, Xinghao Jiang, Tanfeng Sun*

Information Cognition Lab, Shanghai Jiao Tong University

ICASSP 2024 KOREA

INTRODUCTION

Motivation

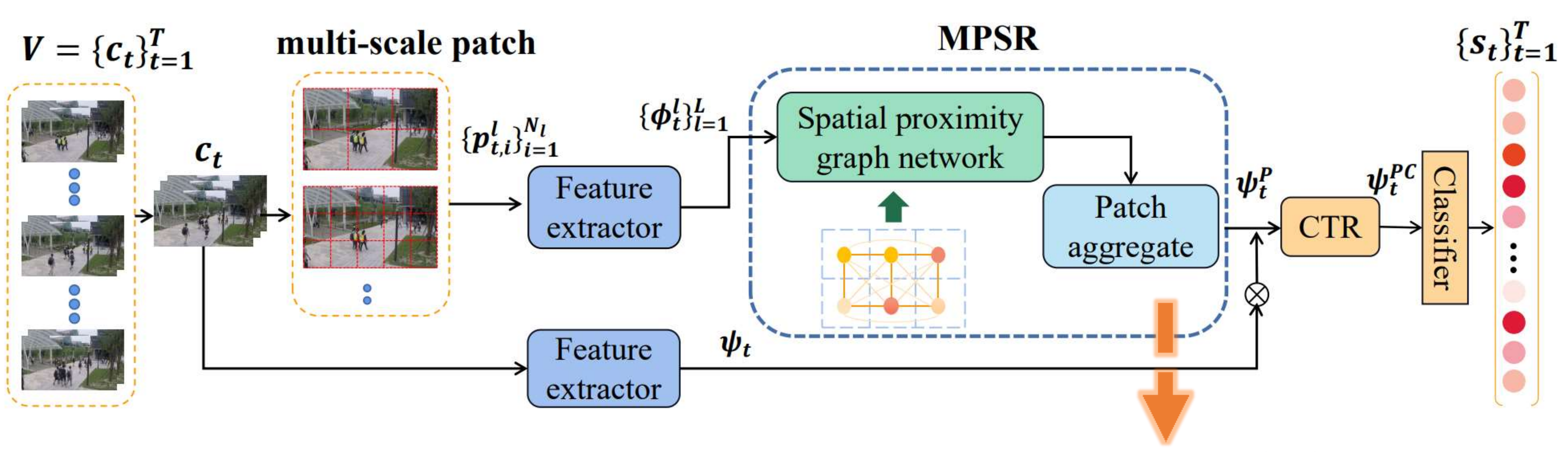
- Anomalous events in surveillance videos tend to occur in restricted, even compact regions, possibly causing background dominance and thus increasing recognition difficulty.
- Anomalies vary in size and position. Current methods struggle to simultaneously detect potential anomalies across various scales.

Contributions

- Propose a multi-scale integrated perception learning method to capture scale-varying anomalies.
- Propose a MPSR network and a HGC block to model the relations among multi-scale patches concurrently.

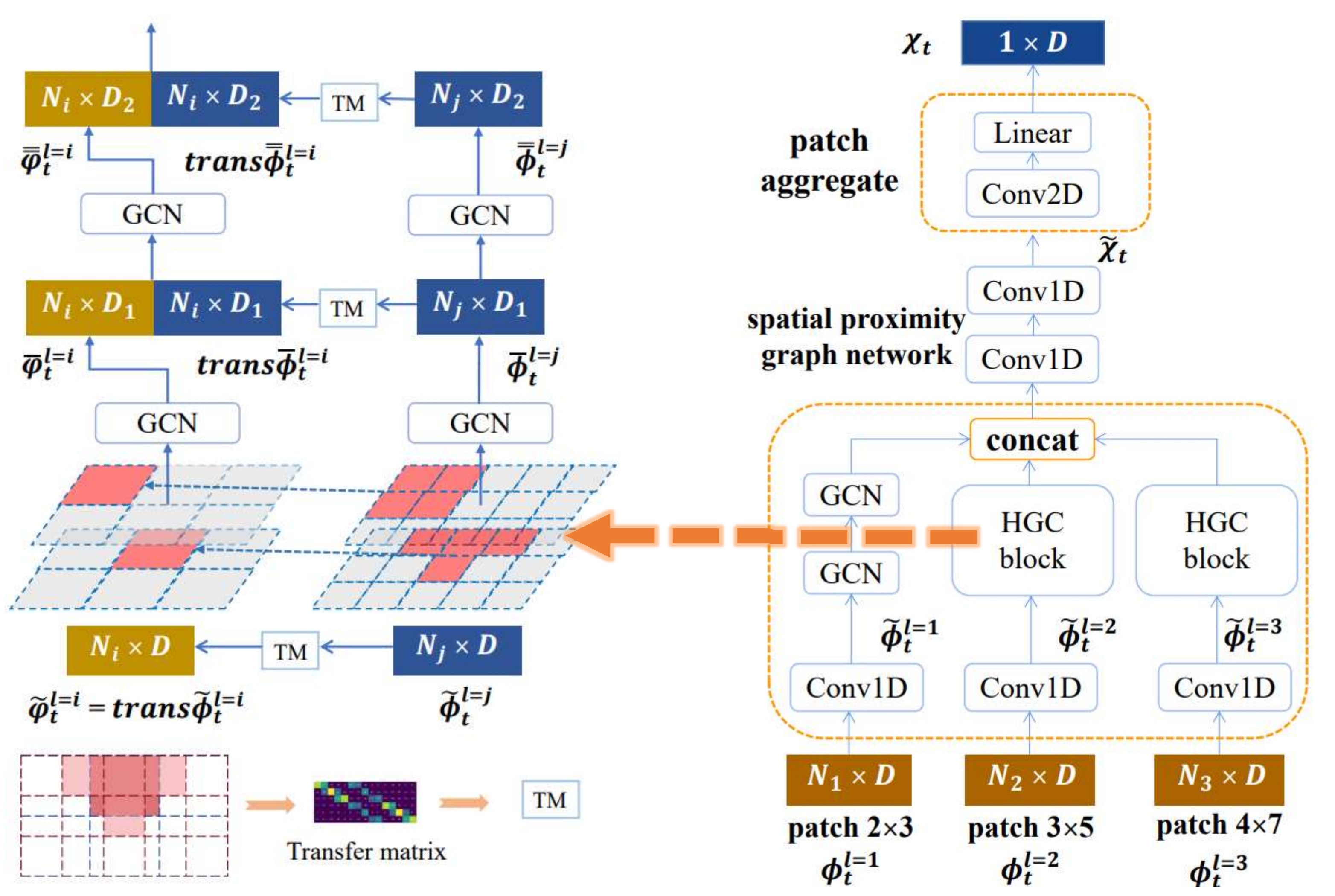
METHOD

The Overall Pipeline



HGC Block

MPSR Module



Procedure

- Segment 480×840 clips into three sets of patches with varying scales: 240×280, 160×168, and 120×120.
- Leverage the pre-trained I3D to extract features of clips and patches.
- Pass the features of multi-scale patches into the multi-scale patch spatial relation (MPSR) network.
- Implement cross-scale feature learning and fusion using the hierarchical graph convolution(HGC) block.
- Fuse the output of MPSR with the clip feature, and then input them into an existing clip temporal relation(CTR) module.
- Send the output of CTR into a classifier to generate anomaly scores.

Training

- Training is divided into two stages. Stage one trains the CTR module and classifier until convergence, while stage two freezes the network and separately trains the MPSR module.
- the total loss function of our model is defined as follows:

$$\mathcal{L} = \mathcal{L}_{BCE} + \lambda_{fm}\mathcal{L}_{FM} + \lambda_1 \sum_t |s_t^+| + \lambda_2 \sum_t (s_t^+ - s_{t-1}^+)^2$$

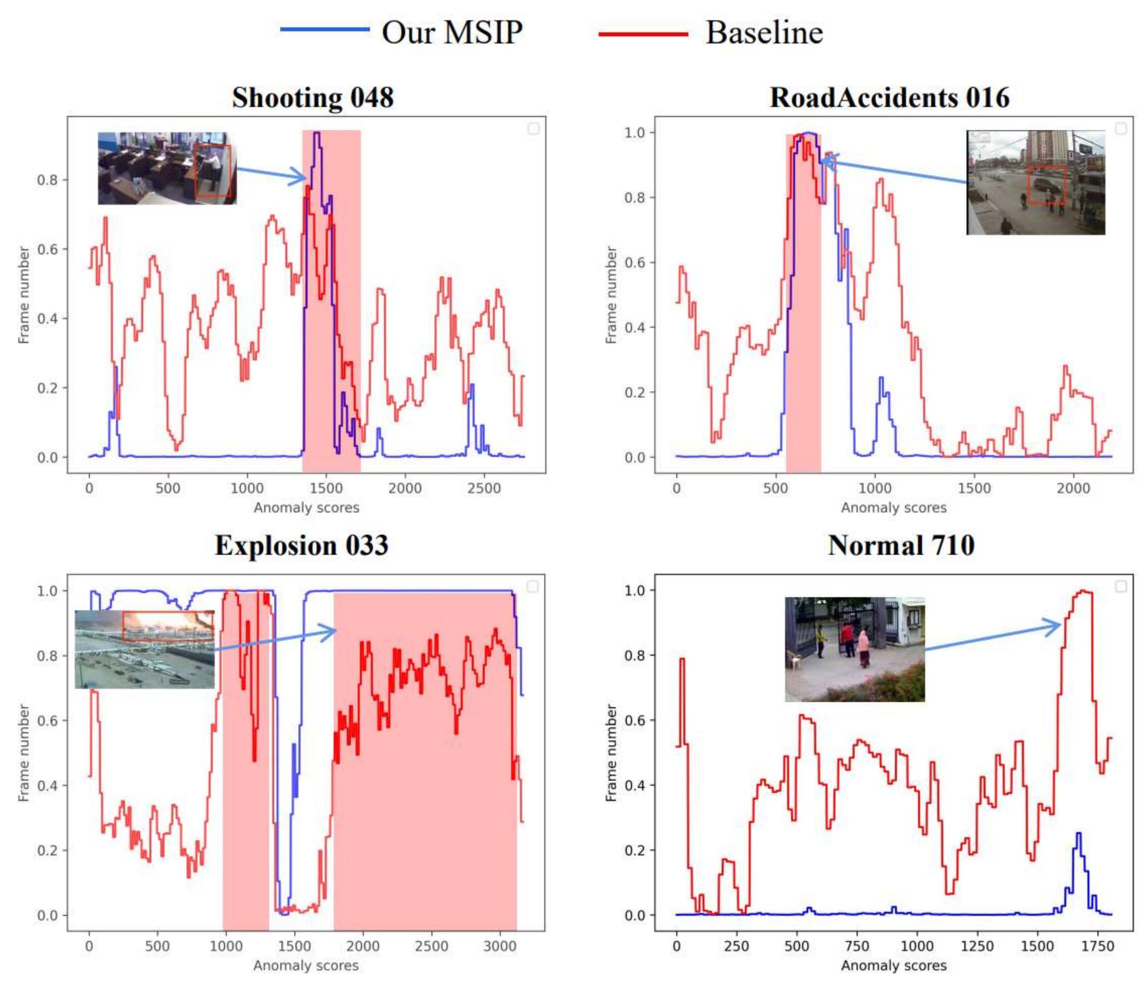
where $\mathcal{L}_{FM}$ is the feature magnitude based MIL ranking loss proposed by previous work.

RESULTS

Comparisons on two datasets. Our method achieves SOTA on ShanghaiTech and competitive results on UCF-Crime.

Sup.	Method	Venue	Feature	AUC@ROC ↑ ShanghaiTech	AUC@ROC ↑ 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	C3D	89.67	83.03
	RTFM [8]	ICCV'21	C3D	91.57	83.28
	Sultani et al. [2]	CVPR'18	I3D	85.33	77.92
	Wu et al. [13]	ECCV'20	I3D	–	82.44
	MIST [1]	CVPR'21	I3D	94.83	82.30
	RTFM [8]	ICCV'21	I3D	97.21	84.30
	S3R [24]	ECCV'22	I3D	97.48	85.99
	SSRL [9]	ECCV'22	I3D	97.98	**87.43**
	SSRL(share parameters) [9]	ECCV'22	I3D	97.84	86.85
	UR-DMU [25]	AAAI'23	I3D	–	86.97
	CLAV [21]	CVPR'23	I3D	97.60	86.10
	Ours: MSIP		I3D	**98.00**	86.98

Visual results on UCF-Crime test videos. Pink areas are temporal ground truths of anomalies.



Ablation study on MPSR and HGC

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

Computational 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

Ablation study on patch data scale variations in inputs.

Patch Size 240 × 280	Patch Size 160 × 168	Patch Size 120 × 120	UCF-Crime (AUC@ROC ↑)	ShanghaiTech (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**