Multi-Scale Video Anomaly Detection by Multi-Grained Spatio-Temporal Representation Learning (Supplementary Materials)

Menghao Zhang, Jingyu Wang, Qi Qi, Haifeng Sun, Zirui Zhuang, Pengfei Ren, Ruilong Ma, Jianxin Liao
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
zhangmenghao,wangjingyu,qiqi8266,hfsun,zhuangzirui,rpf,maruilong@bupt.edu.cn; jxlbupt@gmail.com

In this supplement, we provide the following:

- More ablation results for different non-zero weight combinations for three proxy tasks.
- More experimental results of the proposed proxy tasks with other backbones on the Avenue and ShanghaiTech datasets.
- More qualitative results, including the gap between the anomaly scores of normal and anomalous frames on the Avenue and ShanghaiTech datasets and running time.
- Evaluation of the generalization capacity of the proposed framework.
- More visualization examples on the ShanghaiTech dataset.

1. Different Combinations of Weights

In the implementation details, we empirically set the loss weights of all three proxy tasks as $w_1 = w_2 = w_3 = 1.0$ for training, and we experimented with combinations of more non-zero weights on the Avenue and ShanghaiTech datasets. Table 1 reports the Micro AUC scores for different combinations.

The experimental results corroborate with experience that the model achieves optimal performance when all three weights are set to 1.0 (ID 16). As described previously, the three proxy tasks learn normal model features from different aspects. Continuity judgment learns the global motion pattern and long-range features, thus it can bring significant gains to the model on the Avenue dataset (ID 7-12), which has a relatively homogeneous scene and motion. Discontinuous localization learns motion differences between frames and short-range features in a fine-grained manner. Therefore, it can be effective not only on the Avenue dataset, but also on the ShanghaiTech dataset (ID 1-3 and ID 13-16). Missing frame estimation focuses on understanding the scene and motion, thus significant benefits can be achieved on Campus datasets that contain scene-dependent anomalies. While competitive performance can be achieved by performing a task alone, joint optimization allows for more comprehensive learning of features to reach the optimum(ID 16).

ID	w_1	w_2	w_3	Avenue	ShanghaiTech
1	1.0	0.1	0.1	85.4	78.5
2	1.0	0.5	0.1	88.1	80.7
3	1.0	0.8	0.1	90.5	83.0
4	1.0	1.0	0.1	91.3	83.7
5	1.0	1.0	0.5	91.9	84.2
6	1.0	1.0	0.8	92.1	84.5
7	0.1	1.0	0.1	88.1	83.5
8	0.5	1.0	0.1	90.2	83.7
9	0.8	1.0	0.1	91.1	83.8
10	0.1	0.1	1.0	87.6	81.6
11	0.5	0.1	1.0	88.2	81.9
12	0.8	0.1	1.0	88.7	82.1
13	1.0	0.1	1.0	89.1	82.2
14	1.0	0.5	1.0	90.8	83.4
15	1.0	0.8	1.0	91.9	84.2
16	**1.0**	**1.0**	**1.0**	**92.4**	**85.1**

Table 1. Micro AUC(%) for different combinations of non-zero weights on the Avenue and ShanghaiTech datasets. w_1, w_2, w_3 stand for the weights for continuity judgment, discontinuous localization and missing frame estimation, respectively. The best performing results are marked in bold.

2. Results with Other Backbones

In addition to the I3D network [2] and the multi-task backbone [3], we incorporate the proposed proxy tasks with more backbone, including Unet for frame prediction [5], architecture with memory modules [4, 9], multi-path network [12] and hierarchical learning network [11]. All implementation details and training parameters are set as in

year	Method	Avenue AUC		ShanghaiTech AUC	
		Micro	Macro	Micro	Macro
2018	Liu *et al.* [5]	85.1	-	72.8	-
	+ Estimation Task	87.2	88.5	74.3	75.2
	+ Three Tasks	88.3	89.2	77.0	77.9
2019	Gong *et al.* [4]	83.8	-	71.2	-
	+ Estimation Task	86.1	86.8	75.2	75.7
	+ Three Tasks	87.8	88.3	76.0	76.9
2020	Park *et al.* [9]	88.5	-	70.5	-
	+ Estimation Task	89.8	90.7	72.6	73.8
	+ Three Tasks	90.6	91.2	75.0	76.2
2021	Liu *et al.* [6]	91.1	-	76.2	-
	+ Estimation Task [6]	91.5	92.0	78.0	79.1
	+ Three Tasks	92.1	92.9	79.2	80.4
2022	Wang *et al.* [12]	88.3	-	76.6	-
	+ Estimation Task	89.5	89.6	77.9	78.6
	+ Three Tasks	90.3	90.3	79.2	80.1
2023	Sun *et al.* [11]	92.4	-	83.0	-
	+ Estimation Task	92.5	92.5	83.8	84.4
	+ Three Tasks	92.9	93.0	85.9	86.2

Table 2. Micro and Macro AUC (%) of different backbone on Avenue and ShanghaiTech datasets.

the original paper [4, 5, 9, 11, 12]. When incorporating the proposed proxy tasks, we equally report the performance of both versions, incorporating only the missing frame estimation task and incorporating all tasks. Table 2 shows the Micro AUC (%) of different backbone on Avenue and ShanghaiTech datasets.

For Unet[5], architecture with memory modules [4, 9], and multi-path network [12] incorporating the missing frame estimation task alone and incorporating all three tasks achieve significant improvements on both datasets.

For backbone of Liu *et al.* [6] and hierarchical learning network [11], incorporating our proposed proxy tasks on the Avenue dataset achieves only marginal improvement. This is attributed to the fact that both structures perform object-centered feature learning, and the motion patterns and anomaly scales of This is attributed to the fact that both structures perform object-centric feature learning, especially since both design many branches of feature learning with different aspects. And the motion patterns and anomaly scales the Avenue dataset are relatively homogeneous and easy to learn, the way the original training is done is close to performance saturation. Whereas on the ShanghaiTech dataset, which has large variations in motion patterns and scales, significant improvement can be achieved by combining the agent tasks we designed.

3. More Qualitative Results

3.1. Anomaly Score Gap

Although the comparison with the SOTA methods in the main text highlight the superior performance of our method, we calculate the gap between average scores of normal and abnormal frames to validate the superiority of our architecture. The score gap Δ_{S} in dataset $\mathcal{D}$ is defined by:

$$\Delta_{\mathrm{S}} = \sum_{\mathcal{V}\in\mathcal{D}} \sum_{t\in\{t|\boldsymbol{I}_t\in\mathcal{V}\}} (2y_t - 1)\, S_t \tag{1}$$

where a higher Δ_{S} value indicates a more robust network for distinguishing normal and abnormal events. As shown in Table 3, the large score gap validates the effectiveness of our design.

Method	year	Avenue	ShanghaiTech
Liu *et al.* [5]	2018	0.275	0.175
Georgescu *et al.* [3]	2021	0.368	0.235
Wang *et al.* [12]	2022	0.344	0.182
Zhong *et al.* [14]	2022	0.352	0.153
Yang *et al.* [13]	2023	0.362	0.161
Cao *et al.* [1]	2023	0.360	0.205
Ours		**0.380**	**0.261**

Table 3. The GAP Δ_{S} between average anomaly scores for the normal and abnormal frames. The best performing results are marked in bold.

3.2. Evaluation of the Generalization Capacity

To demonstrate the generalization capacity, we conduct experiments with proposed framework in a cross-dataset setting [7, 8]. In our experiments, we trained on the ShanghaiTech dataset, and then validated on the UCSD ped2 and Avenue datasets. The comparison of our method with the SOTA methods [7, 8] is shown in Table 4. Our proposed method can achieve gains of about 2% on various benchmarks (from 5 shots to 0 shot). In all cases, our method is more general than the meta-learning-based DPU [7] and finetune-based rGAN [8].

Source	Target	Methods	0-Shot	1-Shot	5-shot
Shanghai Tech	UCSD Ped2	rGAN [7]	82.0%	91.2%	91.8%
		DPU [8]	90.2%	94.5%	94.7%
		Ours	**92.5**%	**95.9**%	**96.6**%
	Avenue	rGAN [7]	71.4%	76.6%	77.1%
		DPU [8]	74.0%	78.9%	80.0%
		Ours	**76.3**%	**80.1**%	**81.8**%

Table 4. Comparison of K-shot (K = 0; 1; 5) anomaly detection under the cross-dataset testing setting. Note that K = 0 represents the models are only pre-trained without any adaption.

3.3. Running Time

All our experiments are run on a single NVIDIA RTX3090 GPU, and the backbone of the proposed multiscale framework defaults to the I3D network [2]. Since our method

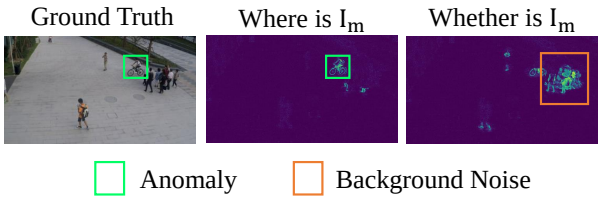

Figure 1. The comparison of reconstruction error maps for the "where is I_m" and "whether is I_m" tasks.

performs proxy tasks at the frame-level by default instead of the object-level [3] (Only perform at the object-level for fair comparisons when using multi-task backbone.), there is no need to resort to the auxiliary network YOLOv3 [10]. For a video clip from the ShanghaiTech dataset, the entire framework runs at approximately 42 frames per second. For reference, real-world surveillance videos are typically saved at frame rates below 30 FPS.

4. More Visualization Examples

In order to highlight the superiority of the proposed method on scenarios with small-scale anomalies more intuitively, we produce a comparison video with the state-of-the-art method ROADMAP[12]. ROADMAP is a multi-scale anomaly detection method based on multi-path ConvGRU. The video for comparison is test video 04-0001 from the ShanghaiTech dataset, where the first part of the video shows a small-scale anomaly that occurs in the upper left portion of the video, and the second part of the video shows a running and jumping anomaly that occurs in the center area.

We mark the anomalous frames and anomalous regions identified by the two methods with red boxes, respectively. As shown in the video, our method can accurately identify anomalies in the part of small-scale anomalies while ROADMAP can only determine a part of the anomalies.

4.1. Visual analysis of discontinuous localization task

The main challenge in small-scale anomaly detection is the interference of background noise. Finding where the missing frame is located in a clip directs the model to focus on subtle changes between adjacent frames rather than the background. The comparison of reconstruction error maps for the tasks "where is I_m" and "whether I_m" are illustrated in Fig. 1. It is evident that the "where is I_m" task enhances the model's sensitivity to subtle changes. The model trained using the "whether I_m" task is more susceptible to background noise. This comparative analysis of visualization results will be incorporated into the final version.

References

[1] Congqi Cao, Yue Lu, Peng Wang, and Yanning Zhang. A new comprehensive benchmark for semi-supervised video anomaly detection and anticipation. In *CVPR*, pages 20392–20401, 2023. 2

[2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. 1, 2

[3] 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. In *CVPR*, pages 12742–12752, 2021. 1, 2, 3

[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 *ICCV*, pages 1705–1714, 2019. 1, 2

[5] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *CVPR*, pages 6536–6545, 2018. 1, 2

[6] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In *ICCV*, pages 13568–13577, 2021. 2

[7] Yiwei Lu, Frank Yu, Mahesh Kumar Krishna Reddy, and Yang Wang. Few-shot scene-adaptive anomaly detection. In *ECCV*, pages 125–141, 2020. 2

[8] Hui Lv, Chen Chen, Zhen Cui, Chunyan Xu, Yong Li, and Jian Yang. Learning normal dynamics in videos with meta prototype network. In *CVPR*, pages 15425–15434, 2021. 2

[9] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, pages 14360–14369, 2020. 1, 2

[10] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. 3

[11] Shengyang Sun and Xiaojin Gong. Hierarchical semantic contrast for scene-aware video anomaly detection. In *CVPR*, pages 22846–22856, 2023. 1, 2

[12] Xuanzhao Wang, Zhengping Che, Bo Jiang, Ning Xiao, Ke Yang, Jian Tang, Jieping Ye, Jingyu Wang, and Qi Qi. Robust unsupervised video anomaly detection by multipath frame prediction. *IEEE Trans. Neural Networks Learn. Syst.*, 33 (6):2301–2312, 2022. 1, 2, 3

[13] Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, and Xiaotao Liu. Video event restoration based on keyframes for video anomaly detection. In *CVPR*, pages 14592–14601, 2023. 2

[14] Yuanhong Zhong, Xia Chen, Yongting Hu, Panliang Tang, and Fan Ren. Bidirectional spatio-temporal feature learning with multiscale evaluation for video anomaly detection. *IEEE Trans. Circuits Syst. Video Technol.*, 32(12):8285–8296, 2022. 2