Video Anomaly Detection by Solving Decoupled Spatio-Temporal Jigsaw Puzzles (Supplementary Material)

Guodong Wang[1,2], Yunhong Wang[2], Jie Qin[3], Dongming Zhang[4], Xiuguo Bao[4], and Di Huang[1,2⋆]

[1] SKLSDE, Beihang University, Beijing, China
[2] SCSE, Beihang University, Beijing, China
[3] CCST, NUAA, Nanjing, China
[4] CNCERT/CC, Beijing, China
{wanggd,yhwang,dhuang}@buaa.edu.cn, qinjiebuaa@gmail.com, zhdm@cert.org.cn, baoxiuguo@139.com

The supplementary material provides:

- detailed configuration of the network architecture.
- comparison in terms of macro-averaged AUROC metric [3].
- running time analysis.
- visual results on UCSD Ped2 [11], CUHK Avenue [9] and ShanghaiTech Campus (STC) [10].
- multi-label regression loss *vs.* multi-label classification loss for training.
- action recognition experiment on UCF-101 [14] using linear probing evaluation.

1 Network Architecture

The detailed configuration of the network is presented in Table 1. The network consists of a shared convolutional part and two independent heads. The shared convolutional neural network (CNN) consists of 3D convolutions (conv) to extract spatio-temporal representations and 2D conv to aggregate spatial representations, while the two individual heads are fully connected (fc) layers. The shared part consists of three 3D blocks and one 2D block. Each 3D block comprises two 3D convolutional layers with the filters of $3 \times 3 \times 3$ and a 3D max-pooling layer. Each convolutional layer is followed by an instance normalization (IN) layer [15], a ReLU activation layer. We perform 3D max-pooling along the spatial dimension in the first two blocks while the last 3D max-pooling layer performs global temporal pooling. The 2D block consists of a 2D convolutional layer, followed by an IN layer, a ReLU activation layer, a 2D dropout layer, and a 2D max-pooling layer. Both heads share the same configuration with two fc layers. We employ IN layers in the network since spatial and temporal jigsaw puzzles are instance-specific and independent of each other. Generally, we adopt the similar architecture (expect for the normalization layer) with the "deep+wide" 3D CNN in [2] for fair comparison.

⋆ Corresponding author (ORCID: 0000-0002-2412-9330).

Table 1: The detailed network architecture. Global temporal pooling is denoted by ":". n^2 and l denote the number of patches in space dimension and the number of frames in time dimension, respectively.

3D	$3 \times 3 \times 3$ conv 32	
	$3 \times 3 \times 3$ conv 32	
	$1 \times 2 \times 2$ max-pooling	
	$3 \times 3 \times 3$ conv 64	
	$3 \times 3 \times 3$ conv 64	
	$1 \times 2 \times 2$ max-pooling	
	$3 \times 3 \times 3$ conv 64	
	$3 \times 3 \times 3$ conv 64	
	: $\times 2 \times 2$ max-pooling	
2D	3×3 conv	
	dropout	
	2×2 max-pooling	
Head	512 fc	512 fc
	$(n^2)^2$ fc	l^2 fc

2 Macro-averaged AUROC Comparison

We note that most of the existing works [1, 4, 7, 8, 16, 18] report the micro-averaged AUROC by concatenating all frames in the dataset then computing the score while some [2, 5] report macro-average AUROC by first computing the AUROC for each video then averaging these scores. Note that we report the micro-averaged AUROC in our main paper by default. Here, we also report the macro-averaged AUROC in Table 2. Clearly, we also achieve the best performance.

3 Running Time

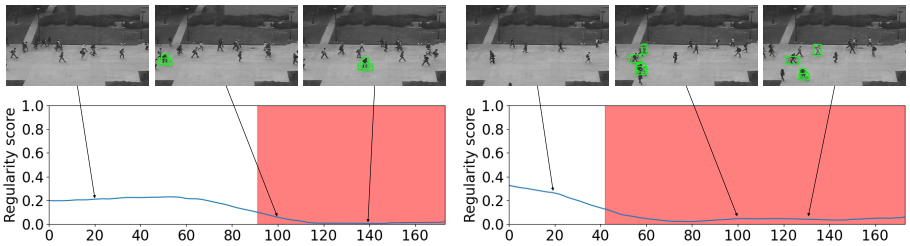
All experiments are conducted on an NVIDIA RTX 2080 Ti GPU and an Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz. For object detection, the YOLOv3 model [13] takes about 20 milliseconds (ms) per frame. In the anomaly detection phase, our lightweight model infers the anomaly scores in 3 ms. With all components considered, our method runs at 28 FPS with an average of 5 objects per frame while the running speed of HF^2-VAD is about 10 FPS. The run-time bottleneck of our framework principally lies in object detection and spatio-temporal cube construction.

4 Visual Results

We provide visual results on UCSD Ped2 [11], CUHK Avenue [9] and STC [10], shown in Figure 1, Figure 2 and Figure 3, respectively. Clearly, the regularity scores correlate strongly with the ground-truth temporal segments of the abnormal events, indicating the effectiveness of our method.

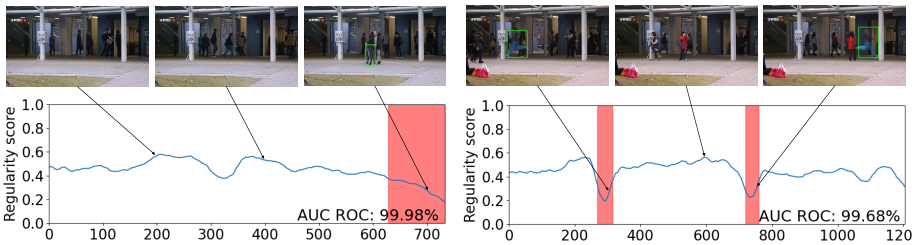
Table 2: Comparison with state-of-the-art methods on frame-level performance in terms of macro-averaged AUROC (%). The best and second-best results are bold and underlined, respectively. * denotes the results taken from [3].

Year	Method	Ped2	Avenue	STC
2018	Frame-Pred.* [7]	98.1	81.7	80.6
2019	CAE-SVM* [5]	97.8	90.4	84.9
2021	SS-MTL [2]	<u>99.8</u>	<u>91.9</u>	<u>89.3</u>
2022	**Ours**	**99.9**	**93.0**	**90.6**



(a) Test video 02 from UCSD Ped2. (b) Test video 07 from UCSD Ped2.

Fig. 1: Regularity score curves by our method on UCSD Ped2. The light red shaded regions represent the ground-truth segments of abnormal events.

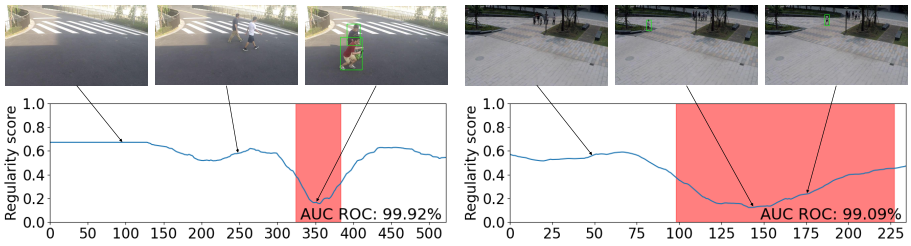


(a) Test video 16 from CUHK Avenue. (b) Test video 02 from CUHK Avenue.

Fig. 2: Regularity score curves by our method on CUHK Avenue. The light red shaded regions represent the ground-truth segments of abnormal events.

5 Classification *vs.* Regression

We first convert each position label to one-hot format and then use mean square error (MSE) to regress entries of the one-hot label for each frame/patch. We obtain 83.9% on STC *vs.* 84.3% (ours), showing that multi-label formulation is robust to different losses.



(a) Test video 03_0031 from STC. (b) Test video 12_0149 from STC.

Fig. 3: Regularity score curves by our method on STC. The light red shaded regions represent the ground-truth segments of abnormal events.

6 Action Recognition

Both action recognition and VAD require learning spatio-temporal features for classification. But the features they require are different - VAD expects the features sensitive to more subtle changes leading to higher discrimination, while our pretext task also benefits action recognition (as shown by the preliminary results under the fine-tuning protocol on UCF-101 [14] in Table 3).

Table 3: Results on UCF-101.

Method	**Accuracy**
Shuffle & Learn [12]	50.2
OPN [6]	56.3
VCOP [17]	64.9
Ours	**67.7**

References

1. Feng, X., Song, D., Chen, Y., Chen, Z., Ni, J., Chen, H.: Convolutional transformer based dual discriminator general adversarial networks for video anomaly detection. In: ACM MM (2021)
2. Georgescu, M.I., Barbalau, A., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: Anomaly detection in video via self-supervised and multi-task learning. In: CVPR (2021)
3. Georgescu, M.I., Ionescu, R.T., Khan, F.S., Popescu, M., Shah, M.: A background-agnostic framework with adversarial training for abnormal event detection in video. arXiv preprint arXiv:2008.12328 (2020)
4. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: ICCV (2019)

5. Ionescu, R.T., Khan, F.S., Georgescu, M.I., Shao, L.: Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: CVPR (2019)
6. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: ICCV (2017)
7. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: CVPR (2018)
8. Liu, Z., Nie, Y., Long, C., Zhang, Q., Li, G.: A hybrid video anomaly detection framework via memory-augmented flow reconstruction and flow-guided frame prediction. In: ICCV (2021)
9. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: ICCV (2013)
10. Luo, W., Liu, W., Gao, S.: A revisit of sparse coding based anomaly detection in stacked rnn framework. In: ICCV (2017)
11. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: CVPR (2010)
12. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV (2016)
13. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
14. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
15. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
16. Wang, Z., Zou, Y., Zhang, Z.: Cluster attention contrast for video anomaly detection. In: ACM MM (2020)
17. Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y.: Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR (2019)
18. Yu, G., Wang, S., Cai, Z., Zhu, E., Xu, C., Yin, J., Kloft, M.: Cloze test helps: Effective video anomaly detection via learning to complete video events. In: ACM MM (2020)