Supplement

Analysis of Experiments on a Single Dataset

To provide a more comprehensive evaluation of our approach, Table 1 presents a detailed comparison between our method and others. Evidently, our method achieves state-of-the-art performance across all evaluated moethods.

Table 1. Micro and macro AUC scores of several state-of-the-art methods on the single dataset. We have marked the first, second, and third places in the results with red, orange, and blue respectively.

Method	Ped2		Avenue		Shanghai Tech	
	Mic AUC	Mac AUC	Mic AUC	Mac AUC	Mic AUC	Mac AUC
Ristea *et al.*[16]	-	-	91.6%	92.5%	83.8%	90.5%
Doshi *et al.*[1]	97.8%	-	86.4%	-	71.6%	-
Georgescu *et al.*[2]	97.5%	99.8%	91.5%	92.8%	82.4%	90.2%
BA Framework[3]	98.7%	99.7%	92.3%	90.4%	82.7%	89.3%
Hirschorn *et al.*[4]	-	-	-	-	85.9%	-
OCAE[5]	94.3%	97.8%	87.4%	90.4%	78.7%	84.9%
Ionescu *et al.*[6]	-	-	88.9%	-	-	-
BMAN[8]	96.6%	-	90.0%	-	-	-
Wu *et al.*[21]	-	-	-	-	80.4	-
Yan *et al.*[22]	-	-	90.1%	-	78.6%	-
Zaheer *et al.*[24]	-	-	-	-	78.9%	-
Tur *et al.*[19]	-	-	-	-	76.1%	-
Tang *et al.*[18]	96.3%	-	85.1%	-	73.0%	-
SCR[17]	97.3%	-	89.6%	-	74.7%	-
Bipoco[7]	98.4%	-	80.2%	-	73.7%	-
Liu *et al.*[9]	99.3%	-	89.9%	93.5%	74.2%	83.2%
Zheng *et al.*[10]	-	-	91.8%	92.3%	83.8%	87.8%
Madan *et al.*[11]	-	-	93.2%	91.8%	83.3%	89.3%
Wang *et al.*[20]	99.0%	-	92.2%	-	84.3%	-
Yu *et al.*[23]	97.3%	-	89.6%	-	74.8%	-
Fastano[12]	96.3%	-	85.3%	-	72.2%	-
Park *et al.*[13]	97.0%	-	82.8%	86.8%	68.3%	79.7%
Ramachandra *et al.*[15]	88.3%	-	72.0%	-	-	-
Vatsavai *et al.*[14]	93.0%	-	87.2%	-	-	-
Ours	99.3%	99.8%	93.6%	93.1%	86.1%	90.3%

Analysis of Ablation Study

In the SVAD task, camera viewpoints do not switch within the same video. Additionally, the SVAD dataset frequently includes multiple targets performing different actions within the same frame, which contrasts with mainstream action

recognition datasets that typically classify a video clip as a whole, often incorporating multiple viewpoint transitions. Fig. 1 illustrates examples from both the pre-training datasets and the SVAD dataset. These domain differences can cause difficulties for the model in accurately extracting action-related features without pre-processing.

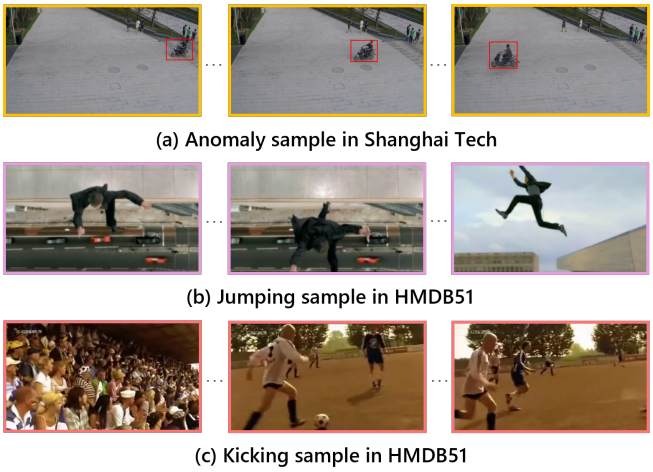

Fig. 1. An example of an anomaly from the Shanghai Tech dataset, as well as examples from the original HMDB51 dataset that include shot transitions and interference from irrelevant targets, are presented.

Limitations

Although our model has achieved satisfactory performance, it still has some limitations. Firstly, we pre-train our model on action recognition datasets to focus on action-based features while ignoring appearance features. However, due to the inherent differences between VAD datasets and action recognition datasets, we must preprocess both types of datasets in the same way. This approach incurs additional computational overhead and affects the generalizability of the method itself. Secondly, the clustering model we currently use is relatively simple, and incorporating methods such as deep evidence learning might yield better results. In the future, we will leverage large language models and multimodal models to further explore more general and concise SVAD methods.

References

1. Doshi, K., Yilmaz, Y.: Any-shot sequential anomaly detection in surveillance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 934–935 (2020) 1

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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12742–12752 (2021) 1
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. IEEE transactions on pattern analysis and machine intelligence **44**(9), 4505–4523 (2021) 1
4. Hirschorn, O., Avidan, S.: Normalizing flows for human pose anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13545–13554 (2023) 1
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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7842–7851 (2019) 1
6. Ionescu, R.T., Smeureanu, S., Popescu, M., Alexe, B.: Detecting abnormal events in video using narrowed normality clusters. In: 2019 IEEE winter conference on applications of computer vision (WACV). pp. 1951–1960 (2019) 1
7. Kanu-Asiegbu, A.M., Vasudevan, R., Du, X.: Bipoco: Bi-directional trajectory prediction with pose constraints for pedestrian anomaly detection. arXiv preprint arXiv:2207.02281 (2022) 1
8. Lee, S., Kim, H.G., Ro, Y.M.: Bman: Bidirectional multi-scale aggregation networks for abnormal event detection. Transactions on Image Processing **29**, 2395–2408 (2019) 1
9. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13588–13597 (2021) 1
10. Liu, Z., Wu, X.M., Zheng, D., Lin, K.Y., Zheng, W.S.: Generating anomalies for video anomaly detection with prompt-based feature mapping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24500–24510 (2023) 1
11. Madan, N., Ristea, N.C., Ionescu, R.T., Nasrollahi, K., Khan, F.S., Moeslund, T.B., Shah, M.: Self-supervised masked convolutional transformer block for anomaly detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 1
12. Park, C., Cho, M., Lee, M., Lee, S.: Fastano: Fast anomaly detection via spatio-temporal patch transformation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2249–2259 (2022) 1
13. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14372–14381 (2020) 1
14. Ramachandra, B., Jones, M., Vatsavai, R.: Learning a distance function with a siamese network to localize anomalies in videos. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2598–2607 (2020) 1
15. Ramachandra, B., Jones, M.J.: Street scene: A new dataset and evaluation protocol for video anomaly detection. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) (2020) 1
16. Ristea, N.C., Croitoru, F.A., Ionescu, R.T., Popescu, M., Khan, F.S., Shah, M.: Self-distilled masked auto-encoders are efficient video anomaly detectors. arXiv preprint arXiv:2306.12041 (2023) 1

17. Sun, C., Jia, Y., Hu, Y., Wu, Y.: Scene-aware context reasoning for unsupervised abnormal event detection in videos. In: Proceedings of the 28th ACM international conference on multimedia. pp. 184–192 (2020) 1
18. Tang, Y., Zhao, L., Zhang, S., Gong, C., Li, G., Yang, J.: Integrating prediction and reconstruction for anomaly detection. Pattern Recognition Letters **129**, 123–130 (2020) 1
19. Tur, A.O., Dall'Asen, N., Beyan, C., Ricci, E.: Exploring diffusion models for unsupervised video anomaly detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 2540–2544 (2023) 1
20. Wang, G., Wang, Y., Qin, J., Zhang, D., Bao, X., Huang, D.: Video anomaly detection by solving decoupled spatio-temporal jigsaw puzzles. In: European Conference on Computer Vision. pp. 494–511 (2022) 1
21. Wu, J.C., Hsieh, H.Y., Chen, D.J., Fuh, C.S., Liu, T.L.: Self-supervised sparse representation for video anomaly detection. In: European Conference on Computer Vision. pp. 729–745 (2022) 1
22. Yan, C., Zhang, S., Liu, Y., Pang, G., Wang, W.: Feature prediction diffusion model for video anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5527–5537 (2023) 1
23. 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: Proceedings of the 28th ACM international conference on multimedia. pp. 583–591 (2020) 1
24. Zaheer, M.Z., Mahmood, A., Khan, M.H., Segu, M., Yu, F., Lee, S.I.: Generative cooperative learning for unsupervised video anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14744–14754 (2022) 1