A SCENE-ADAPTIVE FRAMEWORK FOR POSE-ORIENTED ABNORMAL EVENT DETECTION

Yuxing Yang[1], Zeyu Fu[2] and Syed Mohsen Naqvi[1]

[1] Intelligent Sensing and Communications Research Group, Newcastle University, UK
[2] Department of Computer Science, University of Exeter, UK

ABSTRACT

For intelligent surveillance systems, abnormal event detection (AED) automatically analyses monitoring video sequences and detects abnormal objects or strange human actions at the frame level. Due to the shortage of labelled data, most approaches for AED are based on reconstruction or prediction models in a semi-surprised manner. However, these methods may not generalize well to an unseen scene context. To address this, we present a pose-oriented scene-adaptive framework for AED. In this framework, we propose synergistic pose estimation and object detection, which integrates human poses and object detection information well to improve pose information accuracy. Subsequently, the enhanced pose sequences are taken into a spatial-temporal graph convolutional network to extract the geometric features. Finally, the features are embedded in a clustering layer to classify the type of actions and calculate the normality scores. For evaluation, the proposed framework is tested on video sequences with unseen scene context across from UCSD PED1 & PED2 and ShanghaiTech Campus datasets. The performance analysis and the results compared with other state-of-the-art works confirm the robustness and effectiveness of our proposed framework for cross-scene AED.

Index Terms— Abnormal event detection, scene-adaptive, pose estimation, object detection, graph convolutions

1. INTRODUCTION

Abnormal event detection (AED) detects and analyses human-related activities and classifies the anomalies at the frame level from video sequences. With the widespread application of surveillance cameras, this technology has become significant in intelligent surveillance systems, health care, robotics, and human-computer interface [1, 2, 3, 4]. Due to the limitations of labelled data, most current approaches for AED are based on reconstruction or prediction models via a semi-supervised learning manner. The reconstruction methods, such as [5, 6, 7], model the feature distributions of normal events by an auto-encoder-based neural network and use reconstruction errors to discriminate between normal and abnormal events during inference. The prediction methods, such as [8, 9, 10], learn the probability distribution of normal events by predicting the future frames and use prediction errors to recognize the differences between normal and abnormal events.

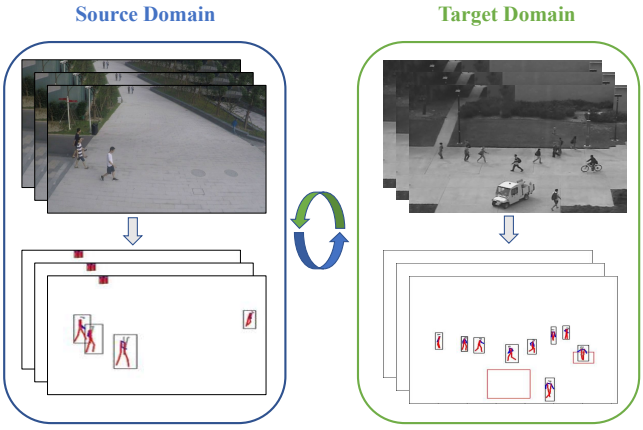

Fig. 1. **Problem review.** Scene-adaptive works for AED are trained in the source domain and tested in the target domain. Compared with the image-level works, our pose-oriented AED is more explainable and efficient.

However, these methods lack the ability for cross-scene generalization. For instance, there are several video sequences from different scenes; if a model is trained on one scene and can be generalized to the other scene context without additional training, it will effectively reduce the computational requirement for intelligent surveillance systems. A cross-scene AED evaluation was discussed in [2], which was trained on ShanghaiTech Campus[9] dataset and tested on CUHK Avenue[11] dataset with similar scenes at the image level. Due to the intervention of background information and various definitions of abnormal events on different datasets, the performance of the proposed cross-scene training framework for AED is unsatisfactory. This paper presents a scene-adaptive framework to address the pose-oriented AED, where human poses are considered a better scene-agnostic feature for cross-scene learning, as shown in Fig 1. To this end, we propose synergistic pose estimation and object detection, which well integrates human poses and object

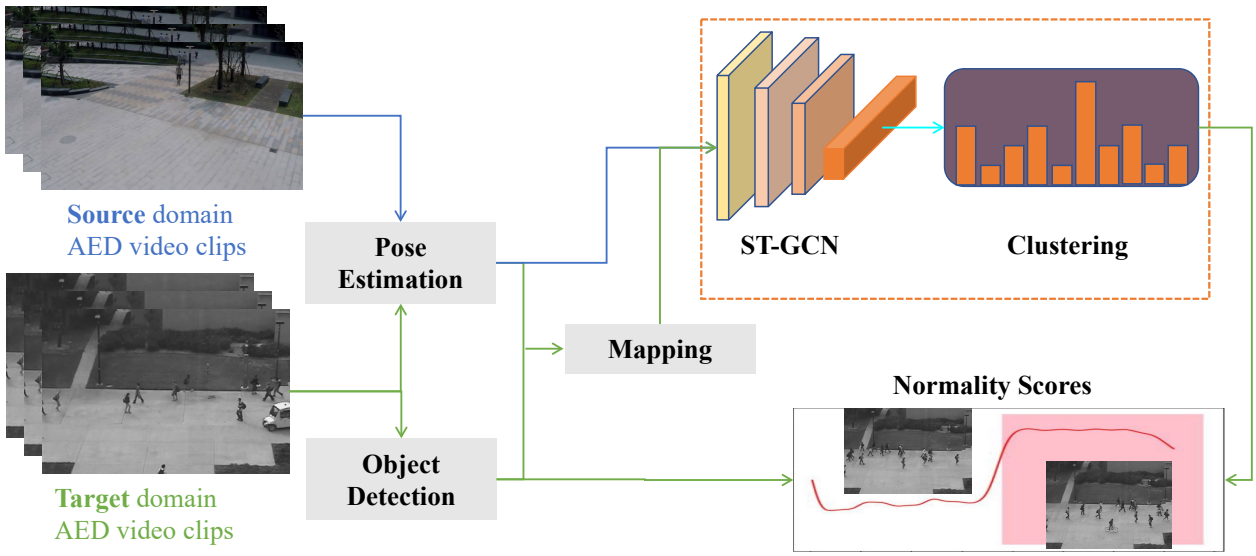

Fig. 2. The overall architecture of the proposed framework. Firstly, video sequences are clipped at the frame level and imported into object detection and pose estimation blocks. The extracted class and pose data are mapped to enhance the data accuracy. Then, in training, the improved pose data are fed into ST-GCN networks to extract the latent vectors and further processed by a deep action-based clustering block. In testing, the videos are resized to adapt to the background size of training samples. The class and pose data are also extracted and mapped. The improved testing pose samples are put into the neural network to calculate normality scores frame by frame. Finally, the results of clustering and detection blocks are fused to do a binary classification to decide whether the frames are normal or abnormal. Best viewed in color.

detection information To improve pose information accuracy by mapping the prior pose and class results. The resulting pose sequences are fed into spatial and temporal graph convolutional networks (ST-GCN) to learn spatio-temporal feature representations. Then a fully connected layer is appended to clustering the human-related activities. Final decisions at the frame level are made based on the integrated normalities from the ST-GCN and object detection, as shown in Fig. 2.

The contributions of this work are threefold: 1). A pose-oriented scene-adaptive framework is proposed for abnormal event detection on surveillance videos. 2) A synergistic pose estimation and object detection is developed to improve the pose information by integrating context class results, which is demonstrated to be effective for scene-adaptive AED. 3) The effectiveness of the proposed framework is validated on two public AED datasets with comparisons with other state-of-the-art methods.

2. PROPOSED METHOD

2.1. Synergistic Pose Estimation and Object Detection

Given a video sequence $\mathbf{V}$, a ready-made model for object detection: YOLOv3 [12] is being used in the frame level to extract the moving objects' classes, bounding boxes, and confidences, which are represented as T, $X_{1,2}$, $Y_{1,2}$ and C, separately. In the $i-$th frame, if there are J objects detected, the image is represented as:

$$\mathbf{V}_J^i = \{[(T, X_{1,2}, Y_{1,2}, C)_{i,j} \,|\, j = 1, ..., J\} \qquad (1)$$

The prior texture class information is efficient in the initial target improvement and the final AED steps.

Inspired by our pose-level human tracking [13] and human action recognition[14, 15], we exploit a pose-driven action recognition framework for scene-adaptive AED. There are two essential factors of pose graphs: nodes and edges. Nodes are the physical location under the coordinate system. Moreover, edges are the vectors between two nodes. As a type of no-grid data, pose graphs are estimated to be fed into ST-GCN[16] to capture the latent vectors of normal actions and to cluster abnormal actions on testing videos. Meanwhile, pose information can effectively decrease the influence of background illumination and different camera viewpoint on different datasets, which create favourable conditions for cross-scene AED. AlphaPose network [17] is conducted to extract the related pose graphs on different datasets. Each pose graph is represented in a 3-D dictionary. The results contain the pose identifications, the frame numbers, the 17 body landmarks and the corresponding confidences. The m-th frame's output with K pose graphs detected after pose estimation is:

$$\mathbf{P}_K^m = \{[m, k, X_n, Y_n, C_n] \,|\, k = 1, ..., K; n = 1, ..., 17\} \qquad (2)$$

Since the proposed framework for scene-adaptive AED is two-stage, the accuracy of pose information is essential for the classification performance of the second-stage neural network. There are two major methods to refine the estimated pose graphs. One method is to improve the accuracy of joint landmarks, and the other is to remove false pose detection and identification. Since our work focuses on abnormal human behaviours, the head landmarks are less important and

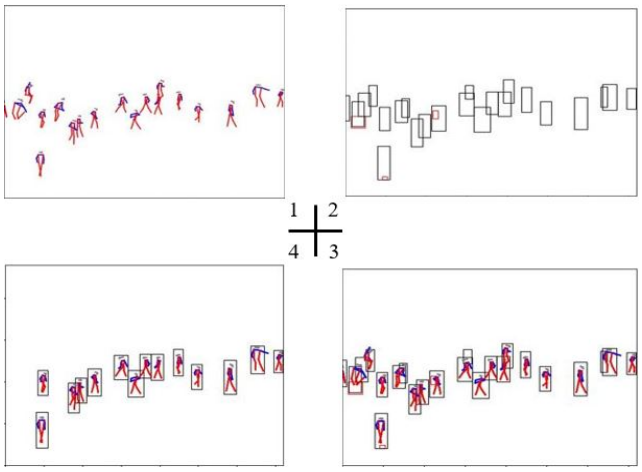

Fig. 3. 1. A pose estimation sample 2. The corresponding object detection sample 3. The combining of pose and bounding box information 4. The enhanced pose graphs after mapping. Best viewed in color.

merged into one landmark. Moreover, combined with the prior object class information, the pose data improved after mapping. There are plenty of false alarms in object and pose detection. As shown in Fig 3, after capturing the pose graphs and bounding boxes, the pose and bounding boxes with low confidence are removed. Then, the K pose graphs are merged with the J bounding boxes. Only the pose graphs that correspond to bounding boxes are retained.

2.2. ST-GCN and Clustering Implementation

After the refinement of pose information, a spatio-temporal graph convolutional network (ST-GCN) [16] is exploited to extract the features of normal events. A dictionary of underlying actions is built using a deep-embedded clustering step. In the spatial domain, the physical localization and connection of body landmarks under the coordinate system are analyzed. Meanwhile, the movement of body landmarks during consecutive frames is considered in the temporal domain. The k_{th} pose graph in the $i-th$ frames on video clips is represented as $\Gamma_k^i = \{N_k^i, E_k^i\}$. $N_k^i = \{\mu_{j,k}^i \in S^D | j = 1, ..., 17\}$ is the nodes of the pose graph, which describe the localization of all landmarks. S are the set of all joints, while D is the dimension of the joints. Moreover, $E_k^i = \{\epsilon_{j,k}^i\}$ is the edges, which means the connection of neighbouring landmarks. To implement the graph neural network, a fixed matrix A is set for all layers [18]. The implementation formula:

$$G(N^i) = \Lambda^{-\frac{1}{2}}(A+I)\Lambda^{-\frac{1}{2}}N^iW^i \tag{3}$$

where N^i is the set of nodes, Λ is the degree matrix, I is the identity matrix representing self-connections, and W^i is trainable weights.

The next step is to analyze and cluster the spatio-temporal features from the neural network. The clustering model is constructed to calculate the probability distribution of normal data. In the training step, the Kullback–Leibler (KL) divergence is minimized between the clustering probability distribution P and the distribution of the detected objects Q [1]. The loss function of the clustering step is:

$$L = KL(Q\|P) = \sum_m \sum_n q_{mn} log\left(\frac{q_{mn}}{p_{mn}}\right) \tag{4}$$

where m, n mean the m_{th} pose assigned to n_{th} cluster.

During inference, anomaly scores are calculated with the combining results of pose normality scores N_p from graph neural network and the class normality scores from object detection. The formula is:

$$N_t = \delta\frac{\sum_{k=1}^K N_{pt}^k}{K} + (1-\delta)\frac{\sum_{k=1}^K R_t^k C_t^k W_i}{U} \tag{5}$$

where δ is the weight for pose information, R is the area of bounding boxes, C is the corresponding confidence, W_i is the weights of objects, and U is the frame size.

3. EXPERIMENTS

3.1. Experimental Setup

We evaluate our proposed method on the following two benchmark datasets, **UCSD PED [19]**: this dataset contains two parts: UCSD PED1 & PED2. PED1 consists of 34 training videos and 36 testing videos with a frame resolution of 238 $\times$ 158, and PED2 consists of 16 training videos and 12 testing videos with a frame resolution of 360 $\times$ 240. The datasets are recorded with fixed viewpoints, which means the scene context of the training and testing sets is consistent. **ShanghaiTech Campus [9]** (SHTC): This dataset is one of the most challenging datasets for AED. There are 330 training video sequences and 107 testing video sequences with complex illuminations and 13 different scenes of a frame resolution of 856 $\times$ 480.

We utilize ST-GCN[16] as our backbone architecture and a deep embedded clustering [20] for the action classification. The clustering parameter K is set to default as 10. The PyTorch framework implements the experiments on a GeForce GTX 1080Ti GPU.

3.2. Performance Analysis

Firstly, the results of scene-adaptive AED are presented. The proposed framework is trained in single-scene video sequences of the SHTC dataset and tested on the rest of the video sequences from the same dataset. Table. 1 details the AUC performance on different scenes and compares it with two related works for AED on the SHTC dataset. Compared with two related works for cross-scene AED, our framework performs better on almost all the testing sequences.

Fig. 4 visualizes some example results on the SHTC dataset for the same-scene, single-scene, and cross-scene

Table 1. AUC performance on the SHTC Dataset when training in one-camera scene video sequence and testing in the rest.

Method	Cam-1	Cam-2	Cam-3	Cam-4	Cam-5	Cam-6	Cam-7	Cam-8	Cam-9	Cam-10	Cam-11	Cam-12
Liu et al.[9]	0.678	0.618	0.663	0.659	0.698	0.735	0.681	0.619	0.674	0.679	0.655	0.651
Doshi et al.[2]	0.753	0.707	0.761	0.681	0.784	**0.814**	**0.789**	0.626	0.706	0.663	0.753	0.719
Proposed	**0.792**	**0.826**	**0.839**	**0.830**	**0.824**	0.792	0.782	**0.763**	**0.787**	**0.749**	**0.779**	**0.738**

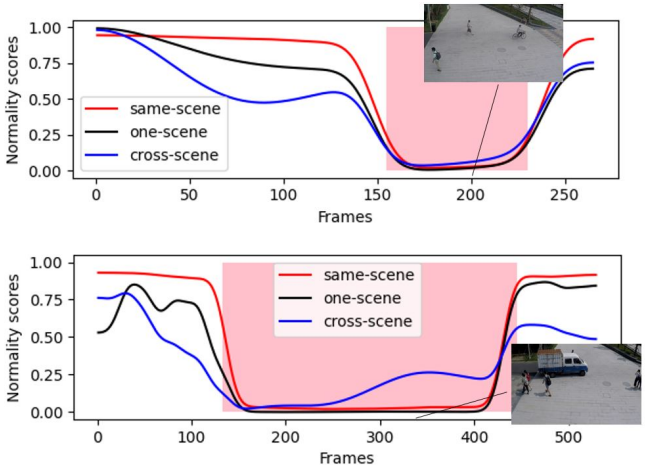

Fig. 4. Normality scores for different scene training. Same-scene: SHTC⟶SHTC. One-scene: one scene on SHTC ⟶ the rest on SHTC. Cross-scene: PED2 ⟶ SHTC.

approaches. The figures present the normality scores of abnormal actions: cycling and driving. In general, all abnormal frames are separated with normal samples approximately. One-scene and cross-scene training performance are slightly worse than same-scene training.

Table. 2 details the performance of our models for cross-scene AED on the UCSD and SHTC datasets. The AUC performance decreased on cross-scene evaluation compared with same-scene evaluation. In our proposed framework, the performance of training on the SHTC dataset and testing on the PED2 dataset slightly drops. On the contrary, the AUC performance drops dramatically. Table. 2 also demonstrates the advantages of the pose-based model for scene-adaptive AED, as pose-level action classification models are robust for cross-scene evaluation with the combination of object detection. Compared with other image-based AED work results, the proposed framework is more suitable for scene-adaptive AED tasks.

Table 2. AUC performance on different datasets for scene adaptability

Train⟶ Test	Recon-based[21]	Frame-Pred[9]	**Proposed**
PED2⟶ PED2	0.865	**0.954**	0.944
SHTC⟶ PED2	0.798	0.895	**0.938**
SHTC⟶ SHTC	0.650	0.728	**0.839**
PED2⟶ SHTC	0.575	0.717	**0.785**

Table 3. AUC Performance for AED compared with some state-of-the-art methods.

	SHTC	PED2	PED1
ConvLSTM-AE [5]	-	0.881	0.755
Conv-AE [6]	0.609	0.811	0.750
Yang *et al.* [22]	-	0.940	-
Luo *et al.*[8]	0.680	0.922	-
Liu *et al.* [9]	0.728	**0.954**	0.831
Zhao *et al.* [10]	-	0.912	**0.923**
Markovitz1 *et al.* [1]	0.761	-	-
Morais *et al.* [23]	0.734	-	-
Proposed (same-scene)	**0.839**	0.944	0.748
Proposed (cross-scene)	-	0.938	0.745

3.3. Comparison with State-of-the-art

Table. 3 compares the AUC performance between the proposed scene-adaptive framework and other state-of-the-art methods. The proposed approach achieves the highest AUC score of 0.839 on the SHTC dataset and a competitive AUC score of 0.944 on the UCSD PED2 dataset. Existing solutions for abnormal event problems are mainly reconstruction models such as ConvAE [6] and ConvLSTM-AE [5] and prediction models such as future frame prediction[9]. Compared with the results of these approaches, our proposed framework has a comparable performance for cross-scene evaluation across from SHTC and UCSD PED2 datasets, which suggests that our pose-oriented framework is suitable for cross-scene AED and that the synergistic pose estimation and object detection strategy effectively integrates the pose and class information for feature extraction. However, for the UCSD PED1 dataset, the AUC performance of our proposed model is underperforming, which is the limitation of the low-resolution dataset for the detection model. Future work will address this matter by including motion features.

4. CONCLUSIONS

We presented a pose-oriented scene-adaptive framework for AED in surveillance videos. With the synergistic pose estimation and object detection, pose data is extracted and enhanced to feed into ST-GCN for action classification and a downstream AED task. The solution of the novel cross-scene training for AED has huge potential in real-world applications. The results of pose-level cross-scene training for AED are feasible on different datasets with the same tasks, similar activities and different backgrounds. Compared with image-level cross-scene training for AED, the scene-agnostic pose information is more effective for the cross-scene AED task.

Through the results of cross-scene training decrease, the AUC performances of training on complicated datasets and testing on simple datasets are acceptable.

5. REFERENCES

[1] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] K. Doshi and Y. Yilmaz, "Multi-task learning for video surveillance with limited data," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022.

[3] F. Angelini, J. Yan, and S. M. Naqvi, "Privacy-Preserving Online Human Behaviour Anomaly Detection Based On Body Movements and Objects Positions," *ICASSP*, 2019.

[4] J. Yan, F. Angelini, and S. M. Naqvi, "Image segmentation based privacy-preserving human action recognition for anomaly detection," *ICASSP*, 2020.

[5] W. Luo, W. Liu, , and S. Gao, "Remembering history with convolutional lstm for anomaly detection," *IEEE International Conference on Multimedia and Expo*, 2017.

[6] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[7] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[8] W. Luo, W. Liu, and S.Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] W. Liu, D. Lian W. Luo, and S. Gao, "Future frame prediction for anomaly detection – a new baseline," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Y. Zhao, B. Deng, C. Shen, Y. Liu, H. Lu, and X. Hua, "Spatio-temporal autoencoder for video anomaly detection," in *ACM international conference on Multimedia*, 10 2017, pp. 1933–1941.

[11] J. Shi C.Lu and J. Jia, "Abnormal event detection at 150 fps in matlab," *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[12] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv*, 2018.

[13] F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, "2d pose-based real-time human action recognition with occlusion-handling," *IEEE Transactions on Multimedia*, vol. 22, no. 6, 2020.

[14] Z. Fu, F. Angelini, J.A. Chambers, and S. M. Naqvi, "Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2277–2291, 2019.

[15] Y. Yang, Z. Fu, and S. M. Naqvi, "A two-stream information fusion approach to abnormal event detection in video," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5787–5791, 2022.

[16] S.Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition.," *AAAI Conference on Artificial Intelligence*, 2018.

[17] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose Flow: Efficient online pose tracking," *British Machine Vision Conference (BMVC)*, 2018.

[18] W. Luo, W. Liu, and S. Gao, "Graph convolutional neural network for skeleton-based video abnormal behavior detection," *Generalization with Deep Learning*, pp. 139–155, 2021.

[19] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 30(5)*, pp. 909–926, 2008.

[20] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," *International Conference on Machine Learning*, vol. 48, pp. 478–487, 2016.

[21] Y. Chong and Y. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Lecture Notes in Computer Science*. 2017, vol. 10262, pp. 189–196, Springer.

[22] Y. Yang, Y. Xian, Z. Fu, and S. M. Naqvi, "Video anomaly detection for surveillance based on effective frame area," *IEEE 24th International Conference on Information Fusion (FUSION)*, 2021.

[23] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," *Computer Vision and Pattern Recognition (CVPR)*, 2019.