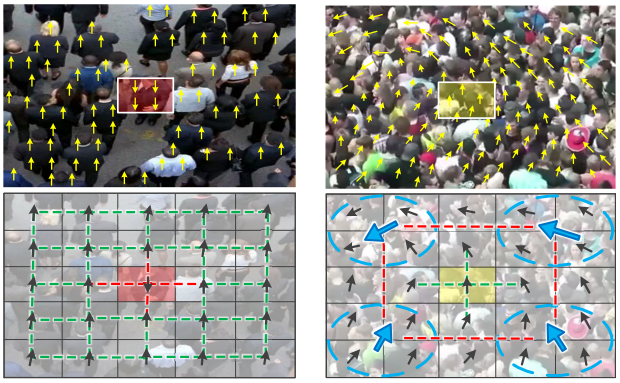
Crowd-Level Abnormal Behavior Detection via Multi-Scale Motion Consistency Learning

Linbo Luo[1]*, Yuanjing Li[1], Haiyan Yin[2], Shangwei Xie[1], Ruimin Hu[1], Wentong Cai[3]

[1] School of Cyber Engineering, Xidian University
[2] Sea AI Lab
[3] School of Computer Science and Engineering, Nanyang Technological University
lbluo@xidian.edu.cn, yjli98@stu.xidian.edu.cn, yinhaiyan@outlook.com,
swxie@stu.xidian.edu.cn, hrm1964@163.com, aswtcai@ntu.edu.sg

Abstract

Detecting abnormal crowd motion emerging from complex interactions of individuals is paramount to ensure the safety of crowds. Crowd-level abnormal behaviors (CABs), e.g., counter flow and crowd turbulence, are proven to be the crucial causes of many crowd disasters. In the recent decade, video anomaly detection (VAD) techniques have achieved remarkable success in detecting individual-level abnormal behaviors (e.g., sudden running, fighting and stealing), but research on VAD for CABs is rather limited. Unlike individual-level anomaly, CABs usually do not exhibit salient difference from the normal behaviors when observed locally, and the scale of CABs could vary from one scenario to another. In this paper, we present a systematic study to tackle the important problem of VAD for CABs with a novel crowd motion learning framework, *multi-scale motion consistency network* (MSMC-Net). MSMC-Net first captures the spatial and temporal crowd motion consistency information in a graph representation. Then, it simultaneously trains multiple feature graphs constructed at different scales to capture rich crowd patterns. An attention network is used to adaptively fuse the multi-scale features for better CAB detection. For the empirical study, we consider three large-scale crowd event datasets, UMN, Hajj and Love Parade. Experimental results show that MSMC-Net could substantially improve the state-of-the-art performance on all the datasets.



(a) Individual-level anomaly (b) Crowd-level anomaly

Figure 1: An illustrative example of individual-level vs. crowd-level abnormal behaviors: (a) individual-level counter-direction pedestrian, whose motion in the local region (red rectangle) exhibits salient difference from neighbouring regions; (b) crowd-level turbulence, in which individuals' motions in the local region (yellow rectangle) do not differ much from neighbouring regions. (Legend: yellow arrows represent optical flow fields; black/blue arrows show the average velocity in each unit-region/circled-region; red/-green dashed lines imply low/high motion consistency.)

Introduction

In real-world crowd events, many self-organizing crowd behaviors could emerge from the complex interactions of individuals, where some of those behaviors, such as counter flow, stop-and-go waves and crowd turbulence, can bring excessive contact forces, making people lose balance, get crushed, or even suffocated (Helbing and Mukerji 2012; Li et al. 2020). Such hazardous self-organizing behaviors are often referred to as crowd-level abnormal behaviors (CABs). Many post-disaster analyses have revealed that CABs are the causes of fatalities in many crowd disasters (Helbing, Johansson, and Al-Abideen 2007; Helbing and Mukerji 2012; Helbing 2013; Ma et al. 2013; Zhao et al. 2020). In the past decade, more than ten thousand people have lost their lives or got injured in the crowd disasters recorded worldwide (Wikipedia 2022). Thus, it is crucial to detect these CABs during large-scale crowd events to reduce crowd risks.

In recent decades, video anomaly detection (VAD) research has gained tremendous momentum. The task of VAD is to detect behaviors or events that are rare and/or have significantly different characteristics from the normal ones in videos. Despite the remarkable success achieved by the existing VAD methods, most of them are designed and testified for detecting individual-level abnormal behaviors, such as sudden running, fighting and stealing, thanks to the full range of publicly available datasets of this kind (e.g., UCSD (Mahadevan et al. 2010), ShanghaiTech (Liu et al. 2018), UCF Crime (Sultani, Chen, and Shah 2018), etc.). However, VAD for CABs is still very limited. We

*Corresponding author

Appendix available at https://arxiv.org/abs/2212.00501

argue that it is challenging to directly apply the existing VAD methods to detect CABs due to the intrinsic difference between individual-level and crowd-level behaviors. First, crowd behavior patterns emerge in crowd motion at macroscopic level (Helbing 2013). Unlike individual-level behaviors whose anomaly could be distinguished by the appearance or motion of abnormal individual(s) in a local region of a crowd, crowd-level anomaly does not always exhibit salient differences from normal ones when observed locally (see an example in Figure 1). Thus, the existing VAD models, which are mostly designed to distinguish anomaly patterns by learning features related to local appearance (e.g., raising arm) or local motion (e.g., an unusual acceleration), are insufficient for detecting CABs. Second, the scale of crowd-level behaviors can vary more considerably under different environmental conditions and crowd densities (Solmaz, Moore, and Shah 2012) as compared to individual-level behaviors, which often have a relatively unified scale (e.g., in the range of one or a small group of pedestrians). Therefore, developing the VAD model against the varying scale of CABs is needed.

To the best of our knowledge, this paper is the first work that formally considers bridging deep learning-based VAD techniques and the CAB detection tasks. To tackle the aforementioned challenges of CAB detection, our work has the following two distinguishing properties. First, instead of detecting anomalies locally like conventional VAD methods, we advocate the analysis of global patterns of collective crowd motion to distinguish CABs from normal behaviors. To this end, we consider the modeling of *crowd motion consistency*, an informative feature to quantify the collectiveness of crowd (Zhou, Tang, and Wang 2013; Li, Chen, and Wang 2020), for CAB detection tasks. Specifically, we introduce a graph-based crowd motion consistency representation, which aims to capture both spatial and time-varying characteristics of crowd motion based on the optical flows extracted from videos. Second, to make the detection robust to the varying scales of the CABs, we design a novel multi-scale motion learning framework, where the model could receive rich crowd behavior information from various feature graphs extracted under different scales for pattern recognition. We also introduce an attentional decoding module to effectively synthesize the multi-scale feature graphs for learning high-quality features for CAB detection.

In a nutshell, our paper has the following contributions:

(i) We motivate and introduce the first work that aims to develop a VAD method specifically for tackling the detection of large-scale crowd-level abnormal behaviors.

(ii) We propose a novel graph-based method, *multi-scale motion consistency network* (MSMC-Net), which comes with a *crowd motion consistency* representation learning module to capture both spatial and temporal motion consistency, as well as a multi-scale decoding module that leverages multiple feature graphs at different scales to capture the crowd behaviors with varying scales.

(iii) We present an extensive empirical evaluation study, where we implement five related baselines, adopt three datasets and demonstrate that our method leads to superior performance consistently across all the datasets.

Related Work

Video Anomaly Detection. Given the ambiguity and diversity of abnormal behaviors in crowd videos, the mainstream of the current VAD research is the data-driven approach, which learns the normal behavior patterns in the training phase using only normal data and detecting the abnormal behavior in the testing phase by evaluating the deviation from the normal patterns. Most existing deep learning methods fall into either reconstruction-based (Nguyen and Meunier 2019; Zhou et al. 2019; Park, Noh, and Ham 2020) or prediction-based (Liu et al. 2018; Ye et al. 2019; Cai et al. 2021; Chen et al. 2022) strands, in which often testing samples with high reconstruction or prediction errors are regarded as anomalies. More recently, the weakly-supervised approach of adding the video-level anomaly label data for training has also gained popularity (Zhong et al. 2019; Feng, Hong, and Zheng 2021; Li, Liu, and Jiao 2022). In this work, we adopt the unsupervised learning approach for detecting CABs. The weakly-supervised learning approach is not considered because crowd-level abnormal behaviors occur less frequently (usually during crowd disasters) than individual-level anomalies, and the crowd anomaly data is thus too scarce to be sufficiently utilized for training. Among recent unsupervised VAD methods (Cai et al. 2021; Chen et al. 2022), the concept of consistency has been considered. However, their consistency modeling is mainly used to capture the individual-level correlation patterns (e.g., the consistency between detected objects' appearance and motion). In contrast, our work exploits the global spatial-temporal consistency of crowd motion and considers the multi-scale issue unique for crowd-level abnormal behaviors.

Crowd Behavior Analysis. Research on crowd behavior in public space has been an intriguing topic for the past decades (Sánchez et al. 2020; Luo et al. 2022). However, the analysis of crowd behaviors during crowd disasters is still very limited due to the scarcity of data. Based on the Hajj 2006 disaster videos, the local density, local velocity, and crowd pressure are measured to analyze the transitions from the normal crowd to stop-and-go wave and crowd turbulence (Helbing, Johansson, and Al-Abideen 2007). Histograms of optical flow extracted from Love Parade 2010 disaster videos are used to cluster the motion patterns, and the magnitude and standard deviation of optical flow motion are combined to assess shock waves in crowd turbulence (Krausz and Bauckhage 2012). The temporal patterns in the Love Parade stampede are analyzed based on distance-based and point process representations of pedestrian movements using the extracted trajectories (Lian et al. 2017). The measurements used in the above methods can capture certain aspects of abnormal behavior patterns. For instance, a high standard deviation of magnitude of optical flow is used to reflect the co-existence of moving people due to pushing and non-moving people in crowd turbulence (Krausz and Bauckhage 2012). However, other normal situations (e.g., visiting market) could also show high variation in speed. Thus, such measurements adopted to evaluate certain characteristics of anomalies are difficult to capture the full range of behav-

ior patterns that are sufficient to distinguish CABs from all normal behaviors. In our work, we leverage the fine-grained feature learning capability of graph convolutional network (GCN) to comprehensively capture the correlations in crowd motion to distinguish CABs from normal ones.

Methodology

In this section, we first present how the spatial and temporal features of crowd motion consistency are extracted from videos and represented in a graph form. Then, we introduce the multi-scale motion consistency network, denoted as **MSMC-Net**, which utilizes the multi-scale motion consistency information to train the network and detect CABs through attention-based unsupervised learning.

Crowd Motion Consistency Representation

To capture global features of crowd motion, we first present the measurements of the spatial consistency and temporal consistency, respectively. Then, we introduce a graph-based representation that encompasses both spatial and temporal consistency features at multiple scales.

Spatial Crowd Motion Consistency. To measure the spatial consistency of crowd motion, we first extract the optical flow field (Farnebäck 2003) for every two consecutive frames and then divide the optical flow field of $H \times W$ area of the video frame into $h \times w$ number of regions. The optical flow vectors within each region are used to calculate an average velocity over each region. The average velocity field of this frame can be expressed as a matrix $\mathbf{M}_t = \{\bar{\mathbf{v}}_{t,c_i}\}_{i=1}^{h \times w}$, where $\bar{\mathbf{v}}_{t,c_i}$ represents the average velocity at region c_i of frame t.

We propose to use spatial-inner consistency (Ω^{sp}) to measure the uniformity degree within one region and spatial-inter consistency (Γ^{sp}) to measure the similarity of the average velocities of two adjacent regions. The spatial-inner consistency is used to capture the internal disorganization, such as the case of escape in different directions (Zhao et al. 2019). The spatial-inter consistency is used to evaluate the relationship between neighbor regions. For example, two opposite crowds in counter flows can produce the lowest spatial-inter consistency (Crociani et al. 2017).

The spatial-inner consistency is calculated using the spatial velocity entropy within a region. Based on the optical flow vectors of a region, the vector direction space is discretized into D classes: $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$, such as up, down, left, right, etc. $H^{\text{sp}}_{t,c_i}(\mathbf{v}_p)$ is used to represent the number of optical flow vectors whose direction belongs to $\mathbf{v}_p$ at region c_i of frame t. The distribution probability of optical vectors at region c_i of frame t can be computed by $P^{\text{sp}}_{t,c_i}(\mathbf{v}_p) = H^{\text{sp}}_{t,c_i}(\mathbf{v}_p)/n$, where n is the total number of the pixel-level optical flow vectors at the region c_i of frame t. The spatial-inner consistency at region c_i of frame t can be calculated as follows:

$$\Omega^{\text{sp}}_{t,c_i} = -\sum_{p=1}^{D} P^{\text{sp}}_{t,c_i}(\mathbf{v}_p) \log P^{\text{sp}}_{t,c_i}(\mathbf{v}_p). \quad (1)$$

For the spatial-inter consistency, the difference in the average velocities of two regions is measured based on the adjusted cosine similarity, which measures the angular difference and the absolute value difference of vectors. The spatial-inter consistency can be calculated as follows:

$$\Gamma^{\text{sp}}_{t,(c_i,c_j)} = \cos(\bar{\mathbf{v}}_{t,c_i}, \bar{\mathbf{v}}_{t,c_j})\left(1 - \frac{\left|\|\bar{\mathbf{v}}_{t,c_i}\| - \|\bar{\mathbf{v}}_{t,c_j}\|\right|}{\|\bar{\mathbf{v}}_{t,c_i}\| + \|\bar{\mathbf{v}}_{t,c_j}\|}\right), \quad (2)$$

where c_i and c_j are two adjacent regions and $\bar{\mathbf{v}}_{t,c_i}, \bar{\mathbf{v}}_{t,c_j}$ are their average velocities at frame t.

Temporal Crowd Motion Consistency. To measure the temporal consistency of crowd motion over a period of time, a sliding window approach is adopted to separate the given video of T frames into sliding snippets. Each sliding snippet contains m frames, and the window is slid by τ frames each time. An average velocity field sequence of the snippet starting with frame t can be expressed as $\{\mathbf{M}_t, ..., \mathbf{M}_{t+m}\}$. Temporal features are then extracted over each snippet.

Similar to spatial consistency, we propose temporal-inner consistency (Ω^{tp}) to measure the uniformity degree of one region's average velocity during its change over time and temporal-inter consistency (Γ^{tp}) to measure the uniformity degree of two adjacent regions' average velocities over time. The temporal consistency compensates for the spatial consistency for obtaining time-varying features. For instance, the temporal-inner consistency can help detect the frequent velocity change over time when crowd turbulence starts (Helbing, Johansson, and Al-Abideen 2007; Ma et al. 2013). The temporal-inter consistency can capture pedestrians' synchronized movement over time in turbulence regions (Lian et al. 2016).

Based on the average velocity field sequence of a snippet starting with frame t, temporal velocity entropy is calculated to measure the temporal-inner consistency. Similar to the calculation of spatial-inner consistency, the average velocities are divided into D categories, and $H^{\text{tp}}_{t+m,c_i}(\mathbf{v}_p)$ measures the number of times, in which the direction of average velocity of region c_i belongs to $\mathbf{v}_p$ during frame t to $t+m$. The velocity distribution probability over this period can be computed by $P^{\text{tp}}_{t+m,c_i}(\mathbf{v}_p) = H^{\text{tp}}_{t+m,c_i}(\mathbf{v}_p)/m$, where m is the length of the snippet. Temporal velocity entropy at region c_i of frame t is calculated as follows:

$$\Omega^{\text{tp}}_{t+m,c_i} = -\sum_{p=1}^{D} P^{\text{tp}}_{t+m,c_i}(\mathbf{v}_p) \log P^{\text{tp}}_{t+m,c_i}(\mathbf{v}_p). \quad (3)$$

To measure the temporal-inter consistency between two regions, we leverage the concept of mutual information to describe the correlations between two regions' motion over time. The distribution probability of two adjacent regions' velocities $P^{\text{tp}}_{t+m,c_i}(\mathbf{v}_p), P^{\text{tp}}_{t+m,c_j}(\mathbf{v}_q)$ and their joint distribution probability $P^{\text{tp}}_{t+m}(\mathbf{v}_p, \mathbf{v}_q)$ are first obtained. The temporal-inter consistency between region c_i and c_j from frame t to frame $t+m$ can be calculated as follows:

$$\Gamma^{\text{tp}}_{t+m,(c_i,c_j)} = \sum_{p,q}^{D,D} P^{\text{tp}}_{t+m}(\mathbf{v}_p, \mathbf{v}_q) \log \frac{P^{\text{tp}}_{t+m}(\mathbf{v}_p, \mathbf{v}_q)}{P^{\text{tp}}_{t+m,c_i}(\mathbf{v}_p) P^{\text{tp}}_{t+m,c_j}(\mathbf{v}_q)}. \quad (4)$$

Construction of Motion Consistency Graphs. To characterize the crowd motion consistency of frame t, the spatial consistency feature of frame t and the temporal consistency feature of the snippet ending with frame t are used.

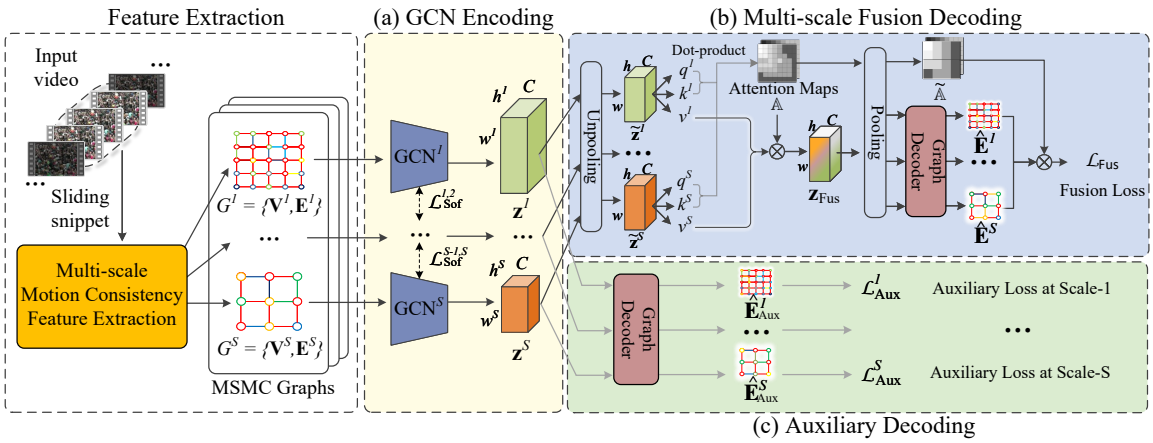


Figure 2: Overview of our MSMC-Net. Part (a) receives the extracted MSMC graphs to generate multi-scale embedding vectors. Two decoding processes in parts (b) and (c) are then performed to reconstruct the MSMC graphs for anomaly detection.

A graph structure capturing both spatial and temporal consistency information of snippet t is proposed as follows: $G_t = \{\mathbf{V}_t, \mathbf{E}_t\}$, where each divided region in the frame is considered as a vertex, adjacent regions are connected by edges, and $\mathbf{V}_t$ and $\mathbf{E}_t$ represent the set of vertexes and edges at frame t, respectively. Each vertex for a region c_i contains the information of the spatial-inner consistency and temporal-inner consistency in the form of vectors:

$$\mathbf{V}_{t,c_i} = [\Omega^{\text{sp}}_{t,c_i}; \Omega^{\text{tp}}_{t,c_i}].$$

The weight of the edge connecting two adjacent regions c_i and c_j contains the information of the spatial-inter consistency and temporal-inter consistency in the form of vectors:

$$\mathbf{E}_{t,(c_i,c_j)} = [\Gamma^{\text{sp}}_{t,(c_i,c_j)}; \Gamma^{\text{tp}}_{t,(c_i,c_j)}].$$

For a given video, a sequence of motion consistency graphs is generated, and each graph is generated per snippet. The first graph G_m is constructed based on the first sliding snippet from frame 1 to m. The subsequent graphs are generated when the window of snippet is slid by τ frames.

However, using only a unified scale to analyze crowd motion makes it difficult to adapt to the variable range of crowd behaviors. Therefore, it is necessary to extract multi-scale crowd motion features for adaptive learning of crowd behaviors. To this end, we define the baseline scale using the average size of the pedestrians in a given video. The video frame of $W * H$ size is divided into $h^1 \times w^1$ number of regions, where W/w^1 and H/h^1 are ensured to be closest to the pedestrians' average shoulder pixel count as the baseline scale (1x-scale). Based on s times the baseline scale, we can perform a sx-scale division to obtain $\left\lceil \frac{h^1}{s} \right\rceil \times \left\lceil \frac{w^1}{s} \right\rceil$ number of regions. The motion consistency graphs above can be extracted at different scales. The multi-scale motion consistency (MSMC) graphs at frame t containing 1x-scale to Sx-scale can be expressed as $\{G^s_t\}^S_{s=1}$ and a sequence of MSMC graphs can then be generated for the entire video.

Multi-scale Motion Consistency Network

Based on the construction of the MSMC graphs described above, we aim to learn the behavioral patterns of the normal crowd under a proper scale level so that crowd-level abnormal behaviors can be detected by examining the deviation from the normal patterns. To this end, our MSMC-net is proposed to perform a multi-scale fusion-based reconstruction of the MSMC graphs in training and reconstruction-based anomaly detection in testing.

Network Architecture. Figure 2 shows our proposed crowd motion learning framework, which consists of (a) GCN encoding, (b) multi-scale fusion decoding, and (c) auxiliary decoding. Part (a) receives the MSMC graphs extracted from the input video, and the GCN-based encoders (Welling and Kipf 2017) are used to exploit the structural features in these graphs to capture the correlation in crowd motion. The GCN-based encoders produce a set of multi-scale embedding vectors. To solve the scale variation problem, part (b) takes the multi-scale embedding vectors and fuses them into a multi-scale fusion vector based on a self-attention mechanism. Fusion-based decoding is then performed to reconstruct the MSMC graphs. To prevent the fusion-based reconstruction from falling into a local optimum, part (c) introduce an auxiliary decoding process that reconstructs the MSMC graphs at each scale separately during training.

In part (a), the MSMC graphs $\{G^s\}^S_{s=1}$ extracted from a given sliding snippet are encoded into multi-scale embedding vectors $\{\mathbf{z}^s\}^S_{s=1}$. Since the edges in an MSMC graph are presented as two-dimensional vectors containing spatial and temporal consistency information, we adopt two GCNs for encoding each graph. One GCN aggregates spatial and temporal features of nodes using only spatial features of edges, and another uses only temporal features of edges. The embedding vectors from the two GCNs are concatenated to obtain the graph embedding vector $\mathbf{z}^s \in \mathbb{R}^{w^s \times h^s \times 2C}$ at each scale s, where $w^s \times h^s$ equals to the number of vertexes in the motion consistency graph of scale s and C is the embedding dimension of the encoder. The benefit of using two

GCNs for multidimensional edge problem lies in its simple implementation and capability to tune GCN parameters separately for each dimension. For knowledge sharing between different scales, the reconstruction tasks are considered as a multi-task operation, and soft sharing constraints ($\mathcal{L}_{\text{Sof}}^{s_1,s_2}$) are applied between the encoders at two different scales.

In part (b), the dimensions of multi-scale vectors are first reshaped to a unified one through nearest-neighbor unpooling. A self-attention mechanism is then leveraged to generate the self-attention maps representing each scale's relative importance. The multi-scale fusion process is performed via aggregating the reshaped vectors based on the attention maps, and a multi-scale fusion vector $\mathbf{z}_{\text{Fus}}$ is generated. Finally, the MSMC graphs are reconstructed with the multi-scale fusion vector, and a fusion loss ($\mathcal{L}_{\text{Fus}}$) is defined to assess the multi-scale fusion-based reconstruction.

In part (c), the MSMC graphs are reconstructed at each scale using only the corresponding scale's embedding vectors obtained from part (a). For a given scale s, the decoder receives the embedding vector $\mathbf{z}^s$ at each scale s and reconstructs the graph's edges ($\hat{\mathbf{E}}_{\text{Aux}}^s$) via inner-product operation. This auxiliary reconstruction task helps to prevent the fusion-based reconstruction process in part (b) from falling into a local optimal solution. This is because if only fusion loss is used to perform both the scale selection and reconstruction learning, the scale that is preferred in scale selection will be more beneficial for reconstruction learning. This may lead to the optimization at a particular scale too much. The auxiliary loss ($\mathcal{L}_{\text{Aux}}^s$) is thus used in part (c) to learn the reconstruction at each scale separately so as to prevent the neglect of some scales that do not perform well initially.

Self-attention based Multi-scale Fusion. In different crowd scenarios, the scale of crowd behaviors can vary, which affects the quantification of crowd behavior patterns and detecting anomalies. To solve the scale variation problem, we leverage a self-attention mechanism, which is used to aggregate the motion consistency features at different scales to perform the multi-scale fusion-based reconstruction in part (b) of our MSMC-Net. During the reconstruction training process, the scale of the fusion model is automatically tuned such that the fusion loss (see in the next section) yields a minimum value.

To perform self-attention-based multi-scale fusion, the graph embedding vectors $\{\mathbf{z}^s\}_{s=1}^S$ are first unsampled across multiple scales to have the same size as the maximum scale through nearest-neighbor unpooling. The unified-scale vectors $\{\tilde{\mathbf{z}}^s\}_{s=1}^S$ are obtained. Then, the embedding vectors are used to guide the generation of attention maps through a cross-scale attention mechanism. The cross-scale attention map a_{xy}^s for scale s at position x, y of the map is calculated as follows:

$$a_{xy}^s = \frac{\left(W_{\text{qry}}\tilde{z}_{xy}^s\right)\left(W_{\text{key}}\tilde{z}_{xy}^s\right)^T}{\left\|W_{\text{qry}}\tilde{z}_{xy}^s\right\|\left\|W_{\text{key}}\tilde{z}_{xy}^s\right\|}, \tag{5}$$

where $s \in \{1, 2, \ldots, S\}$, $x \in \{1, 2, \ldots, w\}$, $y \in \{1, 2, \ldots, h\}$, $\|\cdot\|$ represents the norm function, $W_{\text{qry}}, W_{\text{key}}$ are the trainable weight matrices for generating query ($q_{xy}^s = W_{\text{qry}}\tilde{z}_{xy}^s$) and key ($k_{xy}^s = W_{\text{key}}\tilde{z}_{xy}^s$) for scale s at position x, y. It can be further normalized to denote the current relative importance of scales:

$$\hat{a}_{xy}^s = \frac{\exp\left(a_{xy}^s\right)}{\sum_{s'=1}^S \exp\left(a_{xy}^{s'}\right)}. \tag{6}$$

The set of normalized attention maps for all the scales $\mathbb{A} = \{\mathbf{A}^s\}_{s=1}^S$ can be utilized as the weight for the fusion vector:

$$\mathbf{A}^s = \begin{bmatrix} \hat{a}_{11}^s & \cdots & \hat{a}_{w1}^s \\ \vdots & \ddots & \vdots \\ \hat{a}_{1h}^s & \cdots & \hat{a}_{wh}^s \end{bmatrix} = \left[\hat{a}_{xy}^s\right], \tag{7}$$

$$z_{\text{Fus}}^{xy} = \sum_{s=1}^S \hat{a}_{xy}^s W_{\text{val}}\tilde{z}_{xy}^s, \tag{8}$$

where W_{val} is the trainable weight matrix for generating value ($v_{xy}^s = W_{\text{val}}\tilde{z}_{xy}^s$) for scale s at position x, y.

To achieve fusion-based reconstruction, the fusion vector $\mathbf{z}_{\text{Fus}} = \left[z_{\text{Fus}}^{xy}\right]$, and normalized attention map $\mathbb{A}$ are passed through pooling layers to reshape into the original scales. The reshaped fusion vectors $\{\widetilde{\mathbf{z}}_{\text{Fus}}^s\}_{s=1}^S$ are used for reconstructing the MSMC graphs. The set of reshaped attention maps $\widetilde{\mathbb{A}} = \{\widetilde{\mathbf{A}}^s\}_{s=1}^S$ is used to represent the importance of each scale for the aggregation of multi-scale fusion loss.

Summary of reconstruction procedure. The reconstruction procedure of our MSMC-Net is shown in Algorithm 1.

Algorithm 1: Reconstruction procedure of our MSMC-Net

Input: MSMC graphs $\{G^s\}_{s=1}^S$

for $s = 1 \rightarrow S$ **do**
  $\mathbf{z}^s = \text{GCN}^s\left(G^s\right)$ # GCN encoding
  $\tilde{\mathbf{z}}^s = \text{Unpooling}^s\left(\mathbf{z}^s\right)$ # reshape into a unified scale
  Obtain attention map $[a_{xy}^s]$ based on Eq. 5
end for
Obtain normalized attention maps $\mathbb{A}$ based on Eq. 6-7
Obtain fusion vector $\mathbf{z}_{\text{Fus}}$ using Eq. 8
for $s = 1 \rightarrow S$ **do**
  $\tilde{\mathbf{A}}^s = \text{Pooling}^s\left(\mathbf{A}^s\right)$ # reshape into original scales
  $\tilde{\mathbf{z}}_{\text{Fus}}^s = \text{Pooling}^s\left(\mathbf{z}_{\text{Fus}}\right)$
  $\hat{\mathbf{E}}^s = \text{Decode}\left(\tilde{\mathbf{z}}_{\text{Fus}}^s\right)$ # part (b) reconstruction
  $\hat{\mathbf{E}}_{\text{Aux}}^s = \text{Decode}\left(\mathbf{z}^s\right)$ # part (c) reconstruction
end for
return $\{\hat{\mathbf{E}}^s\}_{s=1}^S$, $\{\hat{\mathbf{E}}_{\text{Aux}}^s\}_{s=1}^S$ and $\{\tilde{\mathbf{A}}^s\}_{s=1}^S$.

Loss Functions. In the training phase, the MSMC-Net is optimized by minimizing the fusion loss $\mathcal{L}_{\text{Fus}}$, auxiliary losses $\mathcal{L}_{\text{Aux}}$ and soft sharing losses $\mathcal{L}_{\text{Sof}}$:

$$\mathcal{L} = \lambda_{\text{Fus}}\mathcal{L}_{\text{Fus}} + \lambda_{\text{Aux}}\sum_{s=1}^S \mathcal{L}_{\text{Aux}}^s + \lambda_{\text{Sof}}\sum_{s_1,s_2}^{S,S} \mathcal{L}_{\text{Sof}}^{s_1,s_2}, \tag{9}$$

where the hyper-parameters λ_{Fus}, λ_{Sof}, λ_{Aux} are used to tune the importance of each part.

To achieve multi-scale fusion-based reconstruction, the reshaped attention map $\widetilde{\mathbb{A}}$ is utilized to weigh different

dataset / method	UMN		Hajj		Love Parade	
	AUC (%, ↑)	EER (%, ↓)	AUC (%, ↑)	EER (%, ↓)	AUC (%, ↑)	EER (%, ↓)
AMC	85.4 ± 1.3	20.3 ± 1.3	65.3 ± 0.5	33.5 ± 0.8	59.6 ± 1.4	39.4 ± 1.2
FramePred	87.9 ± 1.7	19.6 ± 1.4	70.2 ± 1.3	32.2 ± 0.9	58.7 ± 2.2	40.9 ± 1.7
AMMC-Net	87.5 ± 2.3	20.3 ± 1.9	71.4 ± 1.8	30.1 ± 1.8	53.7 ± 1.6	45.4 ± 2.7
MNAD (recons)	85.1 ± 0.7	24.2 ± 1.1	82.4 ± 2.6	24.3 ± 1.8	57.6 ± 1.6	42.3 ± 2.1
MNAD (pred)	88.6 ± 0.4	19.3 ± 0.7	73.2 ± 0.7	31.8 ± 0.9	56.8 ± 1.1	46.9 ± 1.7
MSMC-Net (Ours)	$\mathbf{94.4 \pm 0.5}$	$\mathbf{12.1 \pm 1.0}$	$\mathbf{92.3 \pm 0.5}$	$\mathbf{18.0 \pm 1.8}$	$\mathbf{82.2 \pm 0.9}$	$\mathbf{22.5 \pm 0.5}$

Table 1: Frame-level detection results of our method and the compared baseline methods. All the results support the statistically significant improvement of our method over the baseline methods by a two-sample t-test at a 0.05 significance level.

scales' losses to obtain the fusion loss as follows:

$$\mathcal{L}_{\text{Fus}} = \sum_{s=1}^{S} \frac{1}{w^s \times h^s} \sum_{x_i y_i, x_j y_j} \widetilde{a}^s_{x_i y_i} \widetilde{a}^s_{x_j y_j} \mathcal{L}^s_{(x_i y_i, x_j y_j)}, \quad (10)$$

where $\mathcal{L}^s_{(x_i y_i, x_j y_j)} = \left\| \mathbf{E}^s_{(x_i y_i, x_j y_j)} - \hat{\mathbf{E}}^s_{(x_1 y_1, x_2 y_2)} \right\|$ denotes the ℓ_2 distance between the fusion-based reconstructed edge and the original edge at scale s.

To avoid the pitfall in multi-scale fusion-based decoding as described previously, the auxiliary loss is used in auxiliary decoding, which is defined as the ℓ_2 distance between the auxiliary reconstructed edges and the original at each scale s as follows:

$$\mathcal{L}^s_{\text{Aux}} = \left\| \mathbf{E}^s - \hat{\mathbf{E}}^s_{\text{Aux}} \right\|, \quad (11)$$

where $\hat{\mathbf{E}}^s_{\text{Aux}}$ denotes the reconstructed edges via auxiliary decoding at scale s. To share knowledge between the GCN encoders at two different scales, the soft sharing loss is defined as:

$$\mathcal{L}^{s_1, s_2}_{\text{Sof}} = \| W^{s_1} - W^{s_2} \|, \quad (12)$$

where W^{s_1}, W^{s_2} are the sets of parameters of the encoders at scale s_1 and s_2 , respectively.

Fusion-based Anomaly Score. In the testing phase, given the learned normal correlation patterns of crowd motion, anomalies are detected by examining the deviation from these normal correlation patterns. Since the scale of crowd behavior has been learned during the training phase, only the multi-scale fusion loss $\mathcal{L}_{\text{Fus}}$ is used for detecting anomaly in the testing phase. Considering that crowd-level abnormal behaviors tend to last for a period of time, the moving average is used for the anomaly score $\mathcal{S}_t$ at each frame t as follows:

$$\mathcal{S}_t = (1 - \lambda_{\text{Mov}}) \times \mathcal{S}_{t-1} + \lambda_{\text{Mov}} \times \mathcal{N}(\mathcal{L}_{\text{Fus}}), \quad (13)$$

where λ_{Mov} is the weight of moving average, $\mathcal{N}(\cdot)$ denotes the min-max normalization, $\mathcal{S}_{t-1}$ is the anomaly score in the previous frame and $\mathcal{S}_0 = \mathcal{N}(\mathcal{L}_{\text{Fus}})$. A higher anomaly score (in the range of 0 to 1) indicates a higher degree of anomaly.

Experiments

Datasets

We evaluate the performance of our method on three publicly available datasets, **UMN**, **Hajj** and **Love Parade**, which contain crowd-level abnormal behaviors including crowd escaping, counter flow and crowd turbulence. **UMN** consists of walking and escaping captured by CCTV cameras in three wild scenes. **Hajj** is derived from surveillance videos of the annual religious pilgrimage in Saudi Arabia. **Love Parade** contains surveillance videos of the 2010 Love Parade crowd disaster. Note that our work is the first to introduce **Hajj** and **LoveParade** for VAD study. More details on these datasets are described in Appendix.

Baseline Methods

Four state-of-the-art unsupervised VAD methods are considered as baselines, which can be categorized as *prediction* and *reconstruction*-based methods. Two prediction-based methods include **AMC** (Nguyen and Meunier 2019) and **FramePred** (Liu et al. 2018). Two reconstruction-based methods include **AMMC-Net** (Cai et al. 2021) and **MNAD** (Park, Noh, and Ham 2020). Note that **MNAD** also offers a variant of prediction-based version, and we evaluate both versions of **MNAD** (recons) and **MNAD** (pred). Details on these baseline methods and our settings are in Appendix.

Evaluation Metrics

We use two commonly adopted metrics for the evaluation of frame-level anomaly detection. **AUC** measures the *area under ROC curve*, and the value is larger the better (↑). **EER** measures the *equal error rate* in terms of the location on ROC curve, where the false acceptance rate and false rejection rate are equal, and the value is smaller the better (↓).

Benchmark Results

In our evaluation, ten independent training and testing runs for each method are performed. The average results of different VAD methods in terms of AUC and EER are shown in Table 1. It can be seen that our model achieves the highest average AUC indicating the best overall performance, and the lowest EER indicating that our method generates fewer false and missed alarms. It can also be seen that different datasets exhibit different complexities for anomaly detection. In the Love Parade dataset that contains crowd turbulence, our proposed method exceeds the existing methods significantly. This is probably because pedestrians are packed tightly and move coherently in crowd turbulence, which makes it more difficult to distinguish from the normal congested crowd by their appearance or motion. However, our method can discover its difference by examining both spatial and temporal crowd motion consistency. It should also be noted that the original UMN videos contain a text tag

	Ablated Method Spec	UMN		Hajj		Love Parade	
		AUC (%, ↑)	EER (%, ↓)	AUC (%, ↑)	EER (%, ↓)	AUC (%, ↑)	EER (%, ↓)
Consistency modules	**Spatial** + Multi-scale	88.4 ± 3.7	17.5 ± 2.8	73.6 ± 1.1	38.1 ± 0.7	51.6 ± 2.1	47.9 ± 0.5
	Temporal + Multi-scale	91.6 ± 1.9	19.1 ± 3.8	86.0 ± 0.8	24.8 ± 1.0	80.3 ± 1.7	25.1 ± 2.6
Single scale	Spatial+Temporal (**1x**)	89.6 ± 0.8	20.6 ± 0.9	90.5 ± 1.1	22.5 ± 0.4	67.6 ± 0.2	34.6 ± 0.3
	Spatial+Temporal (**2x**)	73.2 ± 1.4	32.1 ± 2.1	84.6 ± 0.4	27.5 ± 1.3	74.9 ± 0.3	32.3 ± 0.7
	Spatial+Temporal (**4x**)	68.6 ± 1.8	33.1 ± 2.0	72.3 ± 0.6	38.1 ± 0.7	76.9 ± 0.5	30.2 ± 0.2
Our full method	$\surd$	$\mathbf{94.4 \pm 0.5}$	$\mathbf{12.1 \pm 1.0}$	$\mathbf{92.3 \pm 0.5}$	$\mathbf{18.0 \pm 1.8}$	$\mathbf{82.2 \pm 0.9}$	$\mathbf{22.5 \pm 0.5}$

Table 2: Ablation study on (1) crowd motion representation and (2) multi-scale fusion learning in our proposed method.

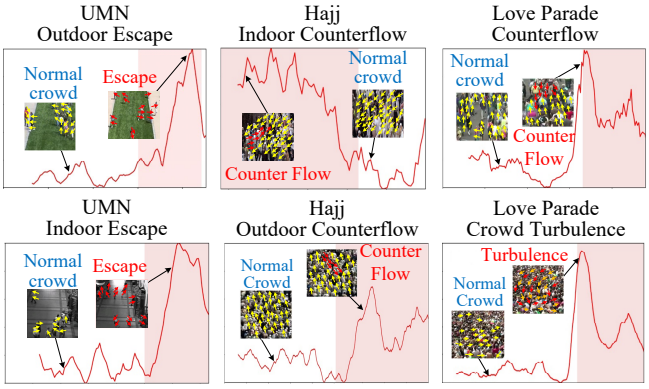


Figure 3: Frame-level anomaly score of our MSMC-Net on three datasets in the testing phase. Each pink area indicates the time interval that the anomaly occurs, and the red curve shows how the anomaly score changes over time.

whenever an anomaly appears in the video (see Appendix), which unnecessarily makes the detection task easier. To prevent the influence of the tag, the UMN results in Table 1 are based on the trimmed videos removing those tags.

In Figure 3, we show how our anomaly scores change over time. It can be observed that our method can effectively produce low anomaly scores for normal situations and high ones for abnormal situations containing various CABs. The changes in anomaly score also reflect how crowd-level abnormal behavior evolves over time. For instance, in UMN, the anomaly score is low at the very beginning of the crowd escape event. It is because only pedestrians' directions change when they just start escaping, while their speed changes take a while due to inertia. The corresponding anomaly score is thus low when escape starts and gradually increases, as shown in Figure 3. In Love Parade, the anomaly score fluctuates in crowd turbulence, since pedestrians are sometimes pushed to move or halted in crowd waves. More results on the evolution of anomaly scores in training and average running time in testing are in Appendix.

Ablation Study

Effectiveness of Motion Consistency Representation. Two variants of our model are trained using the spatial and temporal motion consistency features to construct MSMC graphs, respectively. It can be seen in Table 2 that our full method gives the best results indicating that both the spatial and temporal motion consistency are useful for detecting CABs. It can also be seen in Table 2 that the model with only the temporal consistency feature outperforms the one with only the spatial consistency feature. This indicates that the unusual time-varying patterns occur more frequently than the unusual instantaneous patterns in some CABs. For instance, pedestrians in crowd turbulence frequently change their speeds, causing incongruence in crowd motion over time, whereas instantaneous patterns show subtle differences from the high-density normal crowd. However, the model with only the temporal consistency feature is not enough to achieve the best result. Adding spatial consistency can bring an AUC improvement of 2.8%, 6.3%, and 1.9% for the three datasets.

Effectiveness of Multi-scale Learning. To validate the significance of multi-scale learning, three single-scale variants were trained using motion consistency graphs constructed at 1x, 2x, and 4x scales, respectively. The reconstruction loss at a single scale as defined in our auxiliary loss is used for each ablated model. It can be observed in Table 2 that the scales achieving the best detection result are different for the three datasets. This demonstrates that the scale of crowd behaviors indeed varies for different scenarios. The 1x-scale single-scale model achieves the best result in UMN and Hajj, whereas the 4x-scale model performs best in Love Parade. A larger scale is required for Love Parade probably because the crowd turbulence patterns emerge from the interactions of many pedestrians in the high-density crowd. Moreover, the best scale for detecting CABs can change slightly over time for a given scenario. For example, it is found that the vortex size of crowd turbulence changes as the crowd moves (Ivancevic and Reid 2012). Since our multi-scale fusion mechanism can adaptively estimate the proper scale over time when continuously reading the video input, our method can thus achieve better results than any single-scale variant, demonstrating the ability of scale adaptation.

Conclusion

This paper proposes a detection method for crowd-level abnormal behaviors, which have not yet been fully investigated in the VAD research. Our method exploits the macroscopic crowd motion patterns via multi-scale motion consistency learning and a deep learning network is introduced to learn crowd motion features extracted at multi-scale and estimate the proper scale of crowd behavior for detecting CABs. Evaluations on real-world datasets show the effectiveness of our method. Our future work may include validation on more datasets and cross-scenario detection.

Ethical Statement

This work involves the use of crowd videos containing human subjects. However, the proposed method does not rely on the extraction of individual identity (e.g., face recognition) but rather the motion of the crowd for anomaly detection. Since the optical flow based on the tracking of moving pixels (e.g., head area) is used as our primary input, the proposed method should work even if the human faces in the video are obfuscated. Thus, privacy can be preserved to some extent.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61872282, in part by the Key Research and Development Program of Shaanxi under Program 2022KW-06, in part by the National Natural Science Foundation of China under Grant U22A2035, and in collaboration of the project supported by MoE Tier 1 Grant RG12/21 (Singapore).

References

Cai, R.; Zhang, H.; Liu, W.; Gao, S.; and Hao, Z. 2021. Appearance-motion memory consistency network for video anomaly detection. In *AAAI*, 938–946.

Chen, C.; Xie, Y.; Lin, S.; Yao, A.; Jiang, G.; Zhang, W.; Qu, Y.; Qiao, R.; Ren, B.; and Ma, L. 2022. Comprehensive Regularization in a Bi-directional Predictive Network for Video Anomaly Detection. In *AAAI*, 230–238.

Crociani, L.; Gorrini, A.; Feliciani, C.; Vizzari, G.; Nishinari, K.; and Bandini, S. 2017. Micro and macro pedestrian dynamics in counterflow: The impact of social group. In *International Conference on Traffic and Granular Flow*, 151–158.

Farnebäck, G. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, 363–370.

Feng, J.-C.; Hong, F.-T.; and Zheng, W.-S. 2021. MIST: Multiple instance self-training framework for video anomaly detection. In *CVPR*, 14009–14018.

Helbing, D. 2013. Globally networked risks and how to respond. *Nature*, 497(7447): 51–59.

Helbing, D.; Johansson, A.; and Al-Abideen, H. Z. 2007. Dynamics of crowd disasters: An empirical study. *Physical Review E*, 75: 046109.

Helbing, D.; and Mukerji, P. 2012. Crowd disasters as systemic failures: analysis of the Love Parade disaster. *EPJ Data Science*, 1(1): 1–40.

Ivancevic, V. G.; and Reid, D. J. 2012. Turbulence and shock-waves in crowd dynamics. *Nonlinear Dynamics*, 68(1): 285–304.

Krausz, B.; and Bauckhage, C. 2012. Loveparade 2010: Automatic video analysis of a crowd disaster. *Computer Vision and Image Understanding*, 116(3): 307–319.

Li, S.; Liu, F.; and Jiao, L. 2022. Self-training multi-sequence learning with Transformer for weakly supervised video anomaly detection. In *AAAI*, 1395–1403.

Li, X.; Chen, M.; and Wang, Q. 2020. Quantifying and detecting collective motion in crowd scenes. *IEEE Transactions on Image Processing*, 29: 5571–5583.

Li, X.; Song, W.; Xu, X.; Zhang, J.; Xia, L.; and Shi, C. 2020. Experimental study on pedestrian contact force under different degrees of crowding. *Safety Science*, 127: 104713.

Lian, L.; Song, W.; Ma, J.; and Telesca, L. 2016. Correlation dimension of collective versus individual pedestrian movement patterns in crowd-quakes: A case-study. *Physica A: Statistical Mechanics and its Applications*, 452: 113–119.

Lian, L.; Song, W.; Richard, Y. K. K.; Ma, J.; and Telesca, L. 2017. Long-range dependence and time-clustering behavior in pedestrian movement patterns in stampedes: the Love Parade case-study. *Physica A: Statistical Mechanics and its Applications*, 469: 265–274.

Liu, W.; Luo, W.; Lian, D.; and Gao, S. 2018. Future Frame Prediction for Anomaly Detection - A New Baseline. In *CVPR*, 6536–6545.

Luo, L.; Zhang, B.; Guo, B.; Zhong, J.; and Cai, W. 2022. Why They Escape: Mining Prioritized Fuzzy Decision Rule in Crowd Evacuation. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 19456–19470.

Ma, J.; Song, W. G.; Lo, S. M.; and Fang, Z. M. 2013. New insights into turbulent pedestrian movement pattern in crowd-quakes. *Journal of Statistical Mechanics: Theory & Experiment*, 2013(02): P02028.

Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly detection in crowded scenes. In *CVPR*, 1975–1981.

Nguyen, T.-N.; and Meunier, J. 2019. Anomaly detection in video sequence with appearance-motion correspondence. In *ICCV*, 1273–1283.

Park, H.; Noh, J.; and Ham, B. 2020. Learning Memory-guided Normality for Anomaly Detection. In *CVPR*, 14372–14381.

Sánchez, F. L.; Hupont, I.; Tabik, S.; and Herrera, F. 2020. Revisiting crowd behaviour analysis through deep learning: Taxonomy, anomaly detection, crowd emotions, datasets, opportunities and prospects. *Information Fusion*, 64: 318–335.

Solmaz, B.; Moore, B. E.; and Shah, M. 2012. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10): 2064–2070.

Sultani, W.; Chen, C.; and Shah, M. 2018. Real-World Anomaly Detection in Surveillance Videos. In *CVPR*, 6479–6488.

Welling, M.; and Kipf, T. N. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Wikipedia. 2022. List of human stampedes and crushes. https://en.wikipedia.org/wiki/List_of_human_stampedes_and_crushes. Accessed: 2022-07-25.

Ye, M.; Peng, X.; Gan, W.; Wu, W.; and Qiao, Y. 2019. AnoPCN: Video anomaly detection via deep predictive coding network. In *ACM MM*, 1805–1813.

Zhao, H.; Thrash, T.; Kapadia, M.; Wolff, K.; Hölscher, C.; Helbing, D.; and Schinazi, V. R. 2020. Assessing crowd management strategies for the 2010 Love Parade disaster using computer simulations and virtual reality. *Journal of the Royal Society Interface*, 17(167): 20200116.

Zhao, R.; Hu, Q.; Liu, Q.; Li, C.; Dong, D.; and Ma, Y. 2019. Panic propagation dynamics of high-density crowd based on information entropy and aw-rascle model. *IEEE Transactions on Intelligent Transportation Systems*, 21(10): 4425–4434.

Zhong, J.-X.; Li, N.; Kong, W.; Liu, S.; Li, T. H.; and Li, G. 2019. Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In *CVPR*, 1237–1246.

Zhou, B.; Tang, X.; and Wang, X. 2013. Measuring crowd collectiveness. In *CVPR*, 3049–3056.

Zhou, J. T.; Du, J.; Zhu, H.; Peng, X.; Liu, Y.; and Goh, R. S. M. 2019. AnomalyNet: An anomaly detection network for video surveillance. *IEEE Transactions on Information Forensics and Security*, 14(10): 2537–2550.