University of Reading

Crowd descriptors and interpretable gathering understanding

Article

Accepted Version

Zhou, Y., Liu, C., Ding, Y., Yuan, D., Yin, J. and Yang, S.-H. (2024) Crowd descriptors and interpretable gathering understanding. IEEE Transactions on Multimedia. ISSN 1941-0077 doi: https://doi.org/10.1109/TMM.2024.3381040 Available at https://centaur.reading.ac.uk/115864/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1109/TMM.2024.3381040

Publisher: IEEE

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Crowd Descriptors and Interpretable Gathering Understanding

Yuxin Zhou, Chenguang Liu, Yulong Ding, Diping Yuan, Jiyao Yin, Shuang-Hua Yang, *Senior Member, IEEE*

***Abstract*—Crowd gathering events deeply affect public safety. To enhance city management and avoid potential risks, many algorithms are designed for crowd analysis and deployed on video surveillance. Widely applied deep learning models also can be trained for crowd analysis. However, there are still few works focusing on crowd gathering behavior. Furthermore, as a result of the lack of interpretability of deep learning models, which also brings potential risk of being rejected by the users. In this paper, we categorize crowd behaviors into wandering, merging, walking gathering, standing gathering, and dispersing. Also, we propose an interpretable framework for crowd gathering understanding based on crowd density estimation model and proposed crowd descriptors, named Irregularity, Sparsity, Randomness, and Volatility. The experiments on the PETS2009 dataset demonstrate our method has outperformed the previous works on the crowd gathering understanding task. Moreover, we further analyze the framework performance with different crowd feature extraction models and the relations between our descriptors and crowd behavior. Besides, an ablation study is conducted to investigate the effectiveness of the descriptors and differences between density estimation models. The results demonstrate the effectiveness and the much better interpretability of our framework. Our descriptors also show significant contributions to the quantification of crowd gathering behaviors.**

***Index Terms*—Crowd gathering understanding, crowd descriptor, crowd density estimation, interpretable framework.**

I. Introduction

HUMAN activities deeply affect public safety in cities, which may cause stampedes, riots, and other negative incidents. Many people have lost their lives in these negative crowd events in the past decades. Internet of Things (IoT) development brings many technologies to facilitate crowd management, including smart video surveillance, which can monitor pedestrians and help avoid potential risks.

Crowd analysis is an important task for smart video surveillance. Computer vision algorithms are widely adopted to address crowd analysis tasks, such as crowd counting [1]–[5], crowd anomaly detecting [6]–[8], pedestrians tracking [9]–[12]. Crowd anomaly detecting pays attention to abnormal behaviors, such as fighting [13], robbing [14], and sudden running [6]. Crowd gathering is also an important type of anomaly, which usually occurs in advance before crowd events lead to crowd disasters or accidents. However, very few studies pay attention to crowd gathering.

There is little consensus on what crowd gathering is in existing research. The absence of consensus is caused by the variety of crowd gathering behaviors in various aspects. For instance, whether a gathering crowd should be well-organized and how many people should a gathering crowd have. A definition of crowd gathering is given to clarify the crowd gathering behavior discussed in this paper. Crowd gathering is a crowd behavior, in which a group of well-organized people share a common purpose, and appear in a common physical location. Besides, we categorize crowd behaviors into five stages, named wandering, merging, walking gathering, standing gathering, and dispersing, respectively.

Recently, researchers begin to put their interest in crowd gathering understanding. Liu *et al*. [15] proposed foreground stillness model to detect the crowd gathering behaviors, and Yang *et al*. [16] further improved the detection accuracy by the motion model. Xu *et al*. [17] addressed the task by crowd counting model, which detects the crowd gathering behavior according to the number of people and outperforms previous works in detection accuracy. Moreover, deep learning models have been used to address various image and video analysis tasks and demonstrate a promising pattern recognition performance [18]–[22]. The models can also be trained to detect crowd gathering and outperform the models designed for crowd gathering understanding. A well-functioning crowd analysis system can inform the users possible accident. However, as a black-box model, deep learning suffers from the lack of interpretability [23], which is important for crowd analysis. The lack of interpretability comes with potential risks of being rejected by the users. Specifically, the model can be affected by small perturbations and falsely alert [24], even be embedded with malwares [25]. For untrustable models, the alert may be perceived as untrustworthy and could consequently be rejected by the users, as shown in Fig. 1. The users are generally public

This research is supported in part by the National Natural Science Foundation of China (Grant No. 92067109, 61873119, 62211530106), in part by Shenzhen Science and Technology Program (Grant No. ZDSYS20210623092007023, JCYJ20200109141218676), in part by the Science and Technology Planning Project of Guangdong Province(Grant No. 2021A0505030001), and in part by the Educational Commission of Guangdong Province (Grant No. 2019KZDZX1018). (*Corresponding author: Shuang-Hua Yang.)*

Yuxin Zhou is with the Department of Computer Science and Engineering, Southern University of Science and Technology (SUSTech), Shenzhen, China (e-mail: zhouyx2020@mail.sustech.edu.cn).

Chenguang Liu is with the School of Engineering, University of Warwick, Coventry, UK, CV4 7AL (e-mail: Chenguang.Liu@warwick.ac.uk).

Yulong Ding is with Shenzhen Key Laboratory of Safety and Security for Next Generation of Industrial Internet, and the Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China (email: dingyl@sustech.edu.cn).

Diping Yuan is with Shenzhen Research Institute, China University of Mining and Technology, Shenzhen, China (email: dpyuan2002@aliyun.com)

Jiyao Yin is with the Technology and Information Center, Shenzhen Urban Safety Monitoring and Early Warning Technology Co., Ltd., Shenzhen, China (email: yinjy@szsti.org)

Shuang-Hua Yang is with Shenzhen Key Laboratory of Safety and Security for Next Generation of Industrial Internet, Southern University of Science and Technology, Shenzhen, China, and also with the Department of Computer Science, University of Reading, UK (email: shuang-hua.yang@reading.ac.uk).

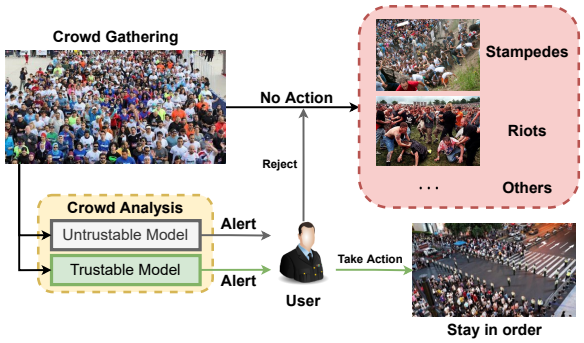

Fig. 1. The use of crowd analysis model in public management

managers. A powerful interpretable model is more human-understandable and trustable for the users [26], which is also required for crowd gathering understanding.

Inspired by the works on measuring crowd collectiveness [27], [28], we propose a set of crowd descriptors to describe intra-group properties. Based on the proposed descriptors, we further build a crowd gathering understanding framework, which extracts a crowd density map and an optical flow map for each frame of the input video sequence as the crowd features. Then the aforementioned descriptors are calculated based on these features. Finally, a classifier outputs which stage the crowd is in.

Our contributions can be summarised as below:

- We propose a novel crowd gathering understanding framework. Our proposed framework considers information of crowd density and crowd motion, and further analyzes the relations between people. It outperforms the existing methods on the task of crowd gathering understanding.
- A set of crowd descriptors, Irregularity, Sparsity, Randomness, and Volatility, are proposed to describe intra-group properties for crowd gathering understanding. To the best of the author's knowledge, crowd density map is first used in crowd property quantification by the proposed descriptors. Our descriptors can improve the interpretability of the proposed framework.
- To describe the transition between non-gathering to gathering, we divide the crowd gathering process into five stages, wandering, merging, walking gathering, standing, gathering, and dispersing respectively. This method specifically defines the process of crowd gathering. Based on the proposed descriptors, we further reveal the crowd motion patterns of each gathering stage.

The rest of this paper is organized as follows. Section II briefly reviews the related works on the analysis of crowd motion patterns, crowd collectiveness measuring, and crowd behavior understanding. In Section III the proposed descriptors are first introduced. Then in Section IV the proposed framework based on our descriptors is presented and followed by an experiment on crowd gathering understanding task in Section V. Further analysis and ablation study are in Section VI. Section VII finally summarizes our work and discusses possible future improvements.

II. Related Work

During the past decades, many researchers have analyzed the underlying principles of pedestrian movement to prevent crowd disasters. In this section, we briefly review the previous works on the analysis of crowd motion patterns, collectiveness measuring, and behavior understanding.

A. Analysis of Crowd Motion Patterns

In early studies, researchers understand crowd behavior based on pedestrian dynamics [29]–[33]. In 1995, Helbing *et al*. [29] described the patterns of pedestrian motion by their social force model, which models the dynamics of pedestrian behavior and formulates the force guiding individual behavior. They also analyzed the Hajj Stampede in 2006 [30]. In this study, they tracked the heads of pedestrians to extract their trajectories and then calculated global and local densities, speeds, and flows. According to these properties, the motion states named laminar, stop-and-go wave, turbulence, and transition among them were discovered. Afterward, an analysis of the factors leading to the Love Parade disaster in 2010 was conducted [31]. Based on the social force model, Yu *et al.* [32] extended its repulsive force term to reproduce the crowd turbulence phenomenon. Moussaïd *et al.* [33] proposed a behavioral heuristics based cognitive science approach to model crowd motion, which overcomes the inconsistency with observation. However, the error of these methods is still large, which makes them difficult to be applied.

B. Crowd Collectiveness Measuring

Many biologists and sociologists have studied collective motion in animal and human society [34]–[37]. Computer scientists also study the crowd behavior quantitatively [27], [28], [38]–[42], which is one type of collective motion. Zhou *et al.* [27] first proposed a descriptor of collectiveness, which quantifies crowd collectiveness by its topological relations among individuals. Shao *et al.* [28] further proposed their group descriptors including descriptors of collectiveness, stability, uniformity, and conflict. Based on *Agent-based Motion Models* (AMM) [29], [43], [44], Liu *et al.* [38] adopted multiple exemplar-AMMs for recognizing crowd motion. In this research, they also proposed a crowd movement numerical measuring framework, which combines all entropy descriptors from exemplar-AMMs to compute a crowd movement feature. Furthermore, they proposed individual holistic features to describe crowd motion [39]. Zou *et al.* [40] proposed a novel method that leverages macroscopic and microscopic features to quantify crowd motion consistency. Wang *et al.* [41] developed a structural context descriptor representing crowd motion dynamics. Pai *et al.* [42] quantified the structuredness in crowd scenes by *Histogram of Angular Deviations* (HAD), which is a structuredness index proposed by them. Li *et al.* [45] designed trajectory-based descriptors profiling the crowd motions for group detection. Zou *et al.* [46] proposed a two-part motion model based on the shortest path principle to simulate and classify the behaviors of pedestrians. Behera *et al.* [47] mapped crowd characterization to a graph classification problem to classify movements based on order parameter,

active force components, and steadiness. Simon *et al.* [48] understanded the crowd movements by analysing the dynamics of motion with the Lagrangian approach. Japar *et al.* [49] studied the collectiveness by analysing temporal information and topological relationship propagation among individuals.

C. Crowd Behavior Understanding

Recently, deep learning obtains great results in many fields including crowd behavior understanding. Feng *et al.* [50] extract event features with PCANet [51] and identify abnormal events by a deep *Gaussian Mixed Model* (GMM). *Convolutional Neural Network* (CNN) based [52] and *Generative Adversarial Networks* (GAN) based [53] models are also proposed for anomalous event detection in crowd scenes. In [54], the author labeled the video data as eight types of crowd behaviors and classified the data by their proposed convolutional *Long Short-Term Memory* (LSTM) based network. Gupta *et al.* [55] proposed a framework for crowd disaster prediction. The framework adopts random forest [56] as the classifier and AlexNet [18] as the backbone net with the input of optical flow and saliency flow. Yang *et al.* [57] proposed a novel deep learning based architecture named DeepSDAE to detect anomalies, which can be trained by reinforcement learning. Zhang *et al.* [58] predict trajectory of individuals and ongoing group behaviors using a novel LSTM-based framework modeling the interaction between pedestrians and environments. Su *et al.* [59] detect social groups according to interpersonal distances and spatio-temporal trajectories of pedestrians. Alafif *et al.* [60] learned a GAN for individual-level abnormal behavior detection. In this work, they trained a GAN by inputting normal samples only. The generator of GAN imitates normal inputs and the discriminator identifies if its input is a real sample or a generated one. Therefore, the discriminator could find the abnormal samples when they are fed in. Some researchers [6], [7] also trained models only with normal data to make the models overfitted. Besides, several studies focus on the crowd gathering understanding task [15]–[17]. Liu *et al.* [15] proposed an image processing based framework for crowd gathering detection. The framework adopts the leaky bucket model and the foreground stillness model to detect crowd gathering behavior in public with video surveillance data. Yang *et al.* [16] improved the frameworks by introducing the motion model and the improved background subtraction algorithm. However, the potential misrecognition is not considered in the stillness-based frameworks. The frameworks cannot identify whether the target is a human. Xu *et al.* [17] overcame the problem by detecting crowd gathering with the crowd counting model. They calculated the number of people in a selected region and compared it with that out of the region to determine the crowd gathering location. However, the authors assumed that a gathering event had occurred in the input video in this study.

Most existing collectiveness approaches extract crowd features by calculating optical flow or tracking feature points. However, they only consider the motion of selected feature points without considering what object these feature points belong to. Moreover, deep learning based methods have achieved significant performance for crowd behavior understanding, whereas the training and inference are treated as a black box, which cannot be explained by the users.

(a) (b) (c)

Fig. 2. Examples with different values of proposed descriptors. (a) A parade troop. (b) A jogging crowd, some people are running and others are walking. (c) A view of a station and people walking in all directions. These examples are from CUHK Crowd Dataset [28]. The red points denote the pedestrians' position, the green arrows denote the movement speeds and directions, and the color blocks indicate different groups.

TABLE I
THE RELATIVE DESCRIPTOR LEVELS CORRESPONDING TO THE EXAMPLES.

Example	$g(G_n)$	$s(G_n)$	$r_v(G_n)$	$r_d(G_n)$	$v(\mathbb{G}_t)$
Fig.2(a)	Low	Low	Low	Low	Low
Fig.2(b)	High	High	High	Low	High
Fig.2(c)	High	High	High	High	High

III. Crowd Descriptors

To describe the intra-group properties of the crowd, we propose a set of descriptors to characterize the features of the crowd behaviors. We focus on capturing the spatial, temporal, and stability features of crowds when designing the descriptors. Taking these features into consideration, we define the proposed descriptors as Irregularity g, Sparsity s, Randomness r, and Volatility v. Besides, the temporal variations of the descriptors are also considered. This section gives the details of each descriptor. Before calculating these descriptors, the crowd should be divided into several groups according to their motion feature.

A. Irregularity

In crowd behaviors, the degree of regularity indicates whether a group is strongly organized or not. Specifically, three examples with different regularity structures are demonstrated in Fig. 2. As the first example, a military parade group with a highly regular structure in Fig. 2(a) illustrates that it is strongly organized. A relatively less regular structure than the military parade is indicated in Fig. 2(b), which is a group of joggers. The third example depicts an unorganized group of wandering pedestrians in Fig. 2(c). To describe the property of the group's behavior, we propose the Irregularity descriptor $g(G_n)$ to represent the degree of the irregularity for the group G_n. The descriptor is calculated by the variance of the average distances between each individual in the group and its K nearest neighbors, which can be formulated as

$$g(G_n) = \sqrt{\sum_{i=1}^{|G_n|}(d_{n,i} - \bar{d_n})^2} \tag{1}$$

$$d_{n,i} = \frac{1}{K} \sum_{c' \in knn(c_{n,i},K)} \|c_{n,i} - c'\| \tag{2}$$

$$\bar{d_n} = \frac{1}{|G_n|} \sum_{i=1}^{|G_n|} d_{n,i} \tag{3}$$

where $|\cdot|$ denotes the count operation for the individuals in the group, and $\|\cdot\|$ denotes the calculation of the Euclidean distance. $c_{n,i}$ is the i-th element in group G_n, and $knn(c_{n,i}, K)$ denotes a set of the K nearest neighbors of $c_{n,i}$. $d_{n,i}$ is the average distance between $c_{n,i}$ and its nearest neighbors, and $\bar{d_n}$ is the average of $d_{n,i}$. A greater K can consider more nearest neighbors. We adopt $K = 100$ in our experiments.

Irregularity is used to describe the spatial features of the crowds. A group has a high Irregularity when it is loose and unorganized. Its structure is not uniform and the distances between individuals fluctuate. Conversely, when Irregularity g is low, this demonstrates that the individuals in the group tend to have a similar mutual distance. For example, in Fig. 2(a), troops show a high degree of uniformity, resulting in a low level of Irregularity.

B. Sparsity

Normally, people in an organized group demonstrate a dense distribution, whereas individuals in an unorganized group with a common destination often show a relatively sparser distribution. However, Irregularity cannot adequately present this distribution, and regular structure is not necessary for a organized gathering group. Therefore, we propose Sparsity descriptor $s(G_n)$ to address this problem, which is formulated as below,

$$s(G_n) = \frac{1}{|G_n|^2 - |G_n|} \sum_{i=1}^{|G_n|} \sum_{c' \in G_n, c' \neq c_{n,i}} \|c_{n,i} - c'\| \tag{4}$$

Sparsity describes the group property as the mean value of the average distances from each individual to other individuals. The descriptor measures the sparsity of a group's position distribution, which represents the spatial features. A group might be an unorganized gathering group when it is sparse. Conversely, a group would raise a potential gathering event when it has a low Sparsity, which indicates the group has a dense distribution. For instance, the soldiers in Fig. 2(a) have a low Sparsity due to their smaller mutual distances than the other two examples.

C. Randomness

Individuals would be assigned to a common group when they have the same motion feature. However, this cannot demonstrate whether or not they are related or organized. In other words, they may have different destinations and objectives with overlapped parts on their paths. Generally, the motion of individuals in a organized group can be more consistent. Thus, to further investigate the difference between an organized group and a randomly gathering group, we propose Randomness $r(G_n)$. Randomness consists of Velocity Randomness $r_v(G_n)$ and Direction Randomness $r_d(G_n)$ expressed as shown in the following formula,

$$r_v(G_n) = -\sum_{x \in C'_t[G_n]} P(x_v) log P(x_v) \tag{5}$$

$$r_d(G_n) = -\sum_{x \in C'_t[G_n]} P(x_d) log P(x_d) \tag{6}$$

where $C'_t[G_n]$ denotes a set of elements belonging to the group G_n in the crowd flow map C'_t. x_v and x_d denote the velocity and direction of x respectively. The two descriptors are expressed by velocity entropy and direction entropy of individuals in the group respectively.

The Randomness is used to describe the temporal features of the crowd group. When two Randomness descriptors have high values, the group is likely to be randomly gathering. Each person in the group can have a very different moving speed and direction. On the contrary, when the values of two Randomness descriptors are low, the group could be potentially organized. The individuals in the group tend to move with a similar speed and direction. For example, the individuals in Fig. 2(c) have a wide range of movement speeds and directions, leading to high Randomness levels.

Moreover, the two descriptors can identify whether the group is standing gathering or walking gathering. For the standing gathering, individuals move at a very slow speed in various directions, which can be regarded as vibrating in the group. Therefore, standing gathering shows a high value in Direction Randomness due to the constantly changing directions. For the walking gathering, individuals move slightly faster with less direction changing. Therefore, compared with standing gathering, walking gathering tends to have a higher Velocity Randomness and a lower Direction Randomness.

D. Volatility

Usually, a well-organized group shows high stability in a gathering event by not dispersing within a period. The number of groups would change in a small range or not change when a gathering event occurs.

We propose the Volatility descriptor $v(\mathbb{G}_t)$ to represent this crowd property, which can be expressed by,

$$v(\mathbb{G}_t) = -\sum_{i=t-m+1}^{t} P(|\mathbb{G}_i|) log P(|\mathbb{G}_i|) \tag{7}$$

where $|\mathbb{G}_t|$ denotes the number of groups in $\mathbb{G}_t$, which is the set of groups in the t-th frame. The Volatility descriptor is expressed as the entropy of the number of groups $|\mathbb{G}_t|$ in the past m video frames, which indicates the instability of $\mathbb{G}_t$ in a period of past time.

Volatility is used to illustrate the stability feature for the crowd group. A high Volatility indicates that the merging or dispersing can be frequently happening to the crowd, and the

number of groups can increase or decrease significantly. In this state, the crowd might be weakly organized and individuals join or leave a group freely. Conversely, a low Volatility demonstrates that there can be less merging and dispersing, thus the number of crowd groups tends to be constant. This shows that the crowd is potentially well organized. For example, the consistent and stable group structure of the soldiers in Fig 2(a) leads to a low Volatility.

E. Temporal Variation

Besides the four descriptors, their corresponding variations are also taken into consideration. The behavior of people is continuous in the time dimension, i.e. it would not only happen at a moment or change suddenly. Therefore, we further include the variations to quantify group behaviors, which can represent the transition process and the underlying trend. We compute the variation according to the formula,

$$\Delta o = o_t - o_{t-1}, o \in \{g, s, r_v, r_d, v\} \tag{8}$$

where o represents one of proposed descriptors, and o_t, o_{t-1} are the corresponding descriptor in t-th and $t-1$-th frame.

IV. Framework for Crowd Gathering Understanding

To address the task of crowd gathering understanding, we propose a novel framework for crowd gathering understanding based on the descriptors in Section III. The details of our method will be introduced in this section. We first introduce the gathering behavior categories and outline the proposed framework, and then go into the details of each module respectively.

A. Gathering Behavior Categories

Generally, a complete crowd gathering process can be described as follows. People walk towards the destination or enter the camera sight, and then the size of the gathering group expands gradually until it stops growing. After gathering for a while, individuals in the group start to disperse or quit the camera sight, which indicates that the gathering event ends. In practical gathering events, the number of people in the group might grow discontinuously. Moreover, the group might disperse for a short time and then become larger again. As the description presents, simply categorizing the crowd behaviors into non-gathering and gathering comes with the challenge of data labeling. It is difficult to determine the threshold between non-gathering and gathering. To overcome this challenge, crowd behaviors are further categorized into five stages, wandering, merging, walking gathering, standing gathering, and dispersing, respectively. The examples are shown in Fig.3. These stages are described as below:

- Wandering: It refers to a stage where the crowd in camera sight has different motion features or the crowd has similar motion patterns without strong organization. People in this stage usually show inconsistent movement.
- Merging: It refers to a stage where the crowd in sight is gathering and the gathering group is still growing. People have a common moving direction or destination, and the group size increases.
- Walking Gathering: It refers to a stage where people form a group with similar motion features for a common destination. In this stage, the number of people in the group stops increasing. People have a common moving direction or destination, but the number of people in the gathering group does not show significant change.
- Standing Gathering: It refers to a stage where people form a group by staying in the same position. In this stage, the number of people in the group stops increasing. People gather and stand in a common area. Likewise, the number of people in the group does not show significant change.
- Dispersing: It refers to a stage where people in the gathering group start to leave the group and the size of the group diminishes. The number of people in the gathering group decreases, and the moving directions radiate from the center of the gathering area.

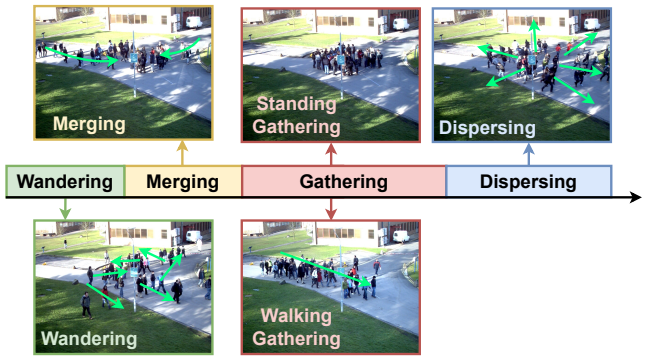

Fig. 3. The examples for five stages of crowd gathering process. The movement of pedestrians is marked by green arrows.

Usually, the above stages occur in order and can be recognized by their movement in a video clip. With the above crowd gathering stages, the challenge of determining the threshold is avoided.

B. Framework Architecture

The proposed framework consists of three modules, crowd feature extraction, crowd behavior quantification, and classification. The outline is shown in Fig.4. According to the architecture of the proposed framework, the current and the previous video frame are put into the framework as the input data to the framework. The framework calculates the crowd density map and dense optical map for the input data. We formulate the crowd density estimation as

$$\hat{D}_t = \mathcal{DE}(F_t) \tag{9}$$

where F_t denotes the t-th frame of the input video sequence, and $F_t \in \mathbb{R}^{W \times H \times 3}$. $\mathcal{DE}(\cdot)$ denotes the operation for crowd density estimation. $\hat{D}_t$ is the estimated density map, which is a matrix having the same width and height as the input data, i.e. $\hat{D} \in \mathbb{R}^{W \times H}$.

The number of people in the input frame can be obtained by summing up all elements in the density map, which is

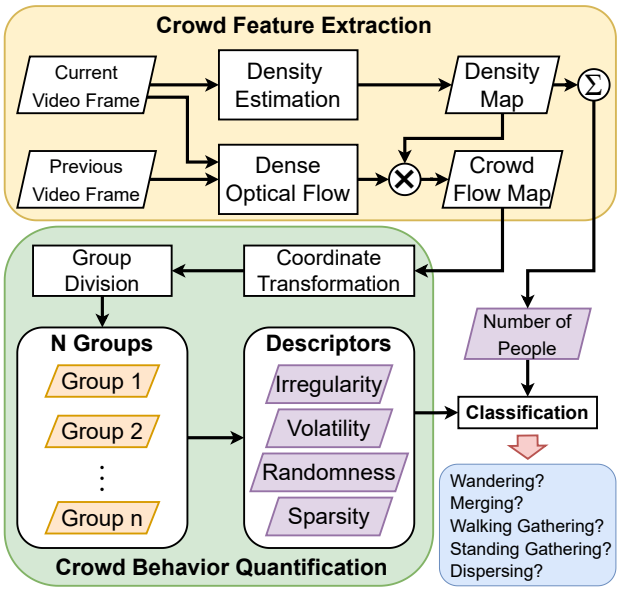

Fig. 4. The Proposed Crowd Gathering Understanding Framework.

used as a spatial feature in the classification module. For the calculation of dense optical flow, the current frame and the previous frame are used. The optical flow calculation module also outputs a map with the same size as the input data. The crowd flow map can be obtained by an element-wise product between the optical flow and the density map. The process can be formulated as:

$$O_t = \mathcal{OF}(F_t, F_{t-1}) \tag{10}$$

$$C_t = \hat{D}_t \cdot O_t \tag{11}$$

where the $\mathcal{OF}(\cdot)$ denotes the function for dense optical flow calculation. O_t denotes the optical flow map calculated from F_t and F_{t-1}, which meets $O_t \in \mathbb{R}^{W\times H\times 2}$. The crowd flow map $C_t \in \mathbb{R}^{W\times H\times 2}$ is obtained by an element-wise product between $\hat{D}_t$ and O_t. O_t includes two channels, which represent the optical flow of each pixel in vertical and horizontal directions respectively. Likewise, C_t contains this information of the crowd flow.

In the crowd behavior quantification module, we first convert C_t from the Cartesian coordinate system to the polar coordinate system to get the polar crowd flow map C'_t. Therefore, every element in C'_t is a pair of two values of its corresponding pixel in the input data. One is the crowd flow magnitude and the other is the direction. This conversion could potentially improve the performance of the group division algorithm. Afterward, the group division algorithm divides the crowd in the camera sight into N groups. The set of groups is $\mathbb{G} = \{G_1, G_2, \cdots, G_N\}$, in which every group $G_n \in \mathbb{G}$ is a set of frame pixels with similar crowd flow values. Then our proposed descriptors detailed in Section III are calculated to describe the property of the group. After the calculation, the classifier determines which stage the crowd is in based on the descriptors. The crowd behavior quantification and classification modules can be formulated as below,

$$\mathbb{G} = \{G_1, G_2, \cdots, G_N\} = \mathcal{GD}(C'_t) \tag{12}$$

$$Result = \mathcal{CF}(g(\mathbb{G}), v(\mathbb{G}), r(\mathbb{G}), s(\mathbb{G})) \tag{13}$$

where $\mathcal{GD}(\cdot)$ denotes the group division algorithm and $\mathcal{CF}(\cdot)$ denotes the classifier function. The $Result$ is the prediction of crowd behavior.

C. Crowd Feature Extraction

Existing methods usually extract crowd features by tracking feature points [27], [28], [40], which is efficient and effective. However, these methods only extract the movement of feature points, which causes the information about the object category to be lost. Therefore, the extracted movement might be a movement of non-human objects, such as cars and pets. To address the problem, our framework utilizes a crowd density estimation model to extract crowd features. Zhang *et al.* [61] first proposed counting crowds by crowd density, which trains a CNN for regressing the density map of the crowd.

In this framework, *Context-Aware Network* (CAN) [2] is adopted as the crowd density estimation model. CAN is a fully convolutional network adopting VGG-16 [62] as the front-end network followed by a *Spatial Pyramid Pooling* [63] to calculate scale-aware features. Finally, seven convolutional layers are applied to the scale-aware features to compute the density map corresponding to the input data. For optical calculation, we adopt Farneback optical flow algorithm [64] to calculate the optical flow map. Afterward, an element-wise product is performed between the density map and the optical flow map to calculate the crowd flow map.

Ideally, the crowd density should be 0 in no-man areas. However, these areas usually have a very small density in the estimated density map due to the error of the model. Therefore, we revise the formula to remove the error on crowd flow calculation. The revised formula can be expressed as,

$$C_t = \Theta(\hat{D}_t; \theta) \cdot O_t \tag{14}$$

$$\Theta(x;\theta) = \begin{cases} x, & x \geq \theta \\ 0, & x < \theta \end{cases} \tag{15}$$

where $\Theta(\cdot)$ is a threshold function and θ denotes the corresponding threshold value. Empirically, an ideal density map is calculated when $\theta = 0.055$.

D. Crowd Behavior Quantification

Individuals in a gathering group share highly uniform motion features when a crowd gathering event occurs. The individuals having similar moving velocities and directions would be assigned to a common group. The velocities and directions of these individuals would vary in a small range due to the perspective. After extracting crowd features, a group division algorithm is applied to divide the crowd in the camera sight into several groups according to their velocities and directions. The clustering algorithm is adopted as our group division algorithm. However, for samples in the Cartesian coordinate system, the algorithm would underperform.

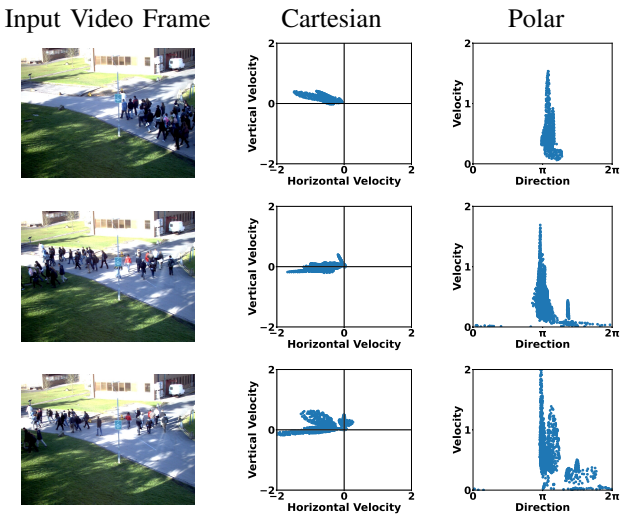

Fig. 5. Extract crowd features of these three frames, their crowd flow pixels' distribution in the Cartesian coordinate and polar coordinate.

Every pixel of the crowd flow map can be expressed as a tuple of (c_x, c_y), and expressed as (c_ρ, c_θ) in the polar coordinate. The pixel scatter diagrams of the crowd flow map in the Cartesian and the polar coordinate system are shown in Fig. 5. As shown in the figures, pixels radiate from the origin in the Cartesian coordinate. For the polar coordinate, every group demonstrates the shape of one peak.

All types of clustering algorithms cannot distinguish groups from the whole crowd in the Cartesian coordinate. The samples for slow individuals in different groups have a short distance between each other. Therefore, the algorithm tends to assign all samples into one cluster. The problem can be solved in the polar coordinate. There are relatively larger gaps between different groups.

The knowledge of the number of groups is lacked, which is not required for DBSCAN [65] clustering algorithm. Therefore, DBSCAN is adopted as the group division algorithm, which assigns individuals into several groups by their velocities and directions. However, the individuals in a common group cannot demonstrate that they are gathering. We further compute the proposed descriptors to identify the gathering groups. Then, the extracted features are input to the classification module to predict the group behavior.

E. Classification

The scales of the descriptors vary due to the perspective and angle variety of the camera. The descriptors also have different scales. The scale variety could affect the performance of the classifier. To address this problem, random forest is adopted as our classifier. The random forest predicts the crowd behavior based on the crowd descriptors and the number of people. More groups provide more features, whereas the number of input features for the classifier is fixed. To address this conflict, only the features of the group with the most people are fed in the random forest. Furthermore, we apply the mean filter to Randomness descriptors and the median filter to other descriptors to reduce noise. The classifier outputs the prediction of whether the group is wandering, merging, standing gathering, walking gathering, or dispersing.

V. Experiments

A. Experiment settings

The crowd behavior dataset we used in the experiments is PETS2009 Dataset [66]. The dataset was collected for crowd counting, pedestrian tracking, behavior recognition, and other crowd analysis tasks. The dataset includes 129 video clips containing more than 40000 frames in total and varieties of crowd behaviors. The frames in the dataset have a resolution of 768×576 and a *Frame Per Second* (FPS) of 7. The videos were captured in eight different views with different camera angles. We conduct our experiment in the views of *View_001*, *View_002* and *View_003*, as shown in Fig. 6, which are proper for crowd gathering understanding. We mark each frame as one of the five labels: wandering, merging, standing gathering, walking gathering, and dispersing. Moreover, the dataset is preprocessed by removing duplicate video sequences. The finally used dataset contains 43 video sequences consisting of 10832 frames captured in three views.

In this section, the proposed framework is implemented. Then the experiment of crowd gathering understanding is conducted on a computer with Intel(R) Xeon(R) CPU E5-2678 V3@2.50GHz and NVIDIA GeForce GTX 1080 Ti.

B. Numerical Result

In this part, the proposed framework is tested on PETS2009, which has been preprocessed. We train the proposed framework to predict crowd behavior categories in video frames as their corresponding labels. Besides testing in all views, we also respectively test our framework in each view. Because different views are captured by cameras with different camera parameters, the descriptors have different scales, which make the features more difficult to learn for our classifier. We compute Micro-F1 scores and Macro-F1 scores of the experiment results as shown in Table II. The Micro-F1 score indicates how accurate the prediction result is, which is mainly affected by the category with more samples. The Macro-F1 score considers the precision and recall of each class, thus, it represents the comprehensive ability of the framework. The confusion matrices of the result are shown in Fig. 7, in which each value in the grids is a rate of classifying vertical label samples as the horizontal label.

The experiment result demonstrates that the proposed framework obtains a significant accuracy in all four settings

Fig. 6. The views we adopt in PETS2009: (a) View_001; (b) View_002; (c) View_003.

TABLE II
MICRO-F1 AND MACRO-F1 SCORES OF CLASSIFYING INPUT VIDEO FRAMES EXPERIMENT

Model Name		Video View			
		View_001	View_002	View_003	All Views
ResNet [19]	Micro-F1	0.9219	0.9278	0.9110	0.9449
	Macro-F1	0.7084	0.8167	0.9156	0.9140
R3D [22]	Micro-F1	0.8466	0.9400	0.9163	0.9513
	Macro-F1	0.7799	0.8194	0.7879	0.9168
Swin [20]	Micro-F1	**0.9781**	**0.9682**	0.9682	**0.9849**
	Macro-F1	**0.9559**	**0.9557**	0.9495	**0.9792**
Ours	Micro-F1	0.9524	0.9645	**0.9741**	0.9534
	Macro-F1	0.8628	0.9197	**0.9648**	0.9003

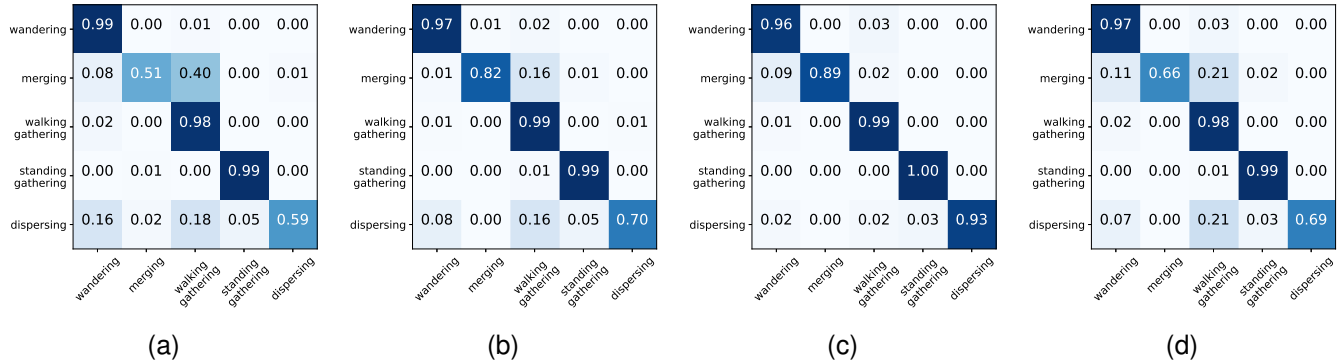

Fig. 7. Normalized confusion matrices of experiments with different test data: (a) View_001; (b) View_002; (c) View_003; (d) All views. The true labels are on the vertical axis, and the predicted labels are on the horizontal axis.

of test data, especially in the recognition of wandering and two types of gathering behaviors. Our framework achieves the purpose of understanding crowd gathering behavior and can identify crowd gathering events accurately with input video frames. The framework also misclassifies a part of merging and dispersing samples, which is caused by two reasons. One is that the end moment of merging and the start moment of dispersing are difficult to choose when labeling the data. The other one is that these misclassified behaviors are the transition process between gathering and wandering, which have similar features to gathering behavior. Thus, their features can be confusing for the classifier. Among the four views, our framework achieves the highest scores in *View_003*, and the lowest scores are obtained in *View_001*. For *All Views* and *View_001*, the Micro-F1 scores are almost the same, and the Macro-F1 score in *All Views* is 0.0375 larger than that in *View_001*. Moreover, the Macro-F1 in *All Views* is 0.0645 lower than in *View_003*. The most likely reason why *All Views* gets intermediate scores is that the classifier trained in *All Views* trades off the performance in different views.

TABLE III
EXPERIMENT RESULT AFTER COMBINING BEHAVIORS FOR COMPARISON

Model Name	Accuracy			
	View_001	View_002	View_003	All Views
Liu *et al.* [15]	0.6837	0.4492	0.6306	0.6031
Yang *et al.* [16]	0.6646	0.6499	0.6893	0.6403
Xu *et al.* [17]	0.6036	0.6585	0.6859	0.5882
ResNet [19]	0.9647	0.9523	0.9628	0.9571
R3D [22]	0.8647	0.9504	0.9354	0.9576
Swin [20]	**0.9805**	0.9706	**0.9804**	**0.9875**
Ours	0.9762	**0.9786**	0.9765	0.9673

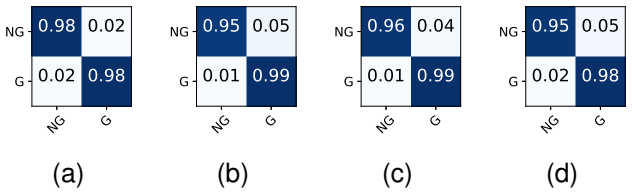

Fig. 8. Confusion matrices of experiments with different test data after combining behaviors: (a) View_001; (b) View_002; (c) View_003; (d) All views. The true labels are on the vertical axis, the predicted labels are on the horizontal axis. NG—Non-Gathering; G—Gathering.

There is no available baseline, due to the different categorizing method of crowd behaviors in existing works. Therefore, we fine-tune three pre-trained deep learning networks on image classification and video classification respectively, namely ResNet [19], R3D [22] and *Swin Transformer* (Swin) [20] as our baselines. The scores of the baselines are also shown in Table. II. In comparison to the baseline models, our method exhibits a slight performance advantage over ResNet and R3D, with the Swin Transformer achieving the highest scores. Notably, our approach maintains a higher level of interpretability when contrasted with the baseline methods.

Specifically, we model the aspects of crowd gathering behaviors we concern as a set of descriptors, which have been introduced in Section III. These methods are more interpretable and easier to understand for users than deep learning methods. Furthermore, our proposed framework leverages the representation capabilities of deep learning based methods. Therefore, our framework combines the advantages of deep learning based methods with those of model-based methods. It emerges as a powerful solution, surpassing most model-based methods in effectiveness while maintaining a higher level of interpretability than deep learning based approaches. As a result, our method achieves comparable performance to deep learning-based models while concurrently providing enhanced interpretability.

Moreover, to compare with existing methods, we respectively combine the result of merging, standing gathering, and walking gathering as gathering behavior, and combine the result of wandering and dispersing as non-gathering behavior. The performance comparison and confusion matrices are presented in Table III and Fig. 8. As the result shows, our framework outperforms the existing methods and the CNN-based baselines, and is slightly outperformed by the Swin.

VI. Analysis and Ablation Study

In this section, more experiments are conducted. Afterward, the effectiveness of each framework component is analyzed. We select three typical video clips from PETS2009 as our examples, which contain all types of crowd behavior. The examples can be described as follows:

- Video 1: In this clip, people enter the camera sight and go to the center from three directions. Then all people meet at the center and merge as a standing gathering group. After having gathered for a period, all people disperse suddenly and run out of the camera sight. The video clip's path in the dataset is S3/High_Level/Time_14-33.
- Video 2: In this clip, people walk into the camera sight from left to right as a team. When the pedestrian at the head of the team arrives at the center of sight, he begins to run and other people follow until the video ends. The video clip's path in the dataset is S1/L3/Time_14-17.
- Video 3: In this clip, no gathering event occurs. Some groups having massive and loose people, appear in the camera sight, which would confuse detection models. The video clip's path in the dataset is S2/L2/Time_14-55.

Besides, we also set three different models for our experiments. The setting details of the models are described below:

- Model 1: The model has similar architecture to our proposed framework, except for the crowd feature extraction module. The crowd feature extraction module is replaced with a traditional method. It adopts YOLO [67] to detect pedestrians and uses Deep SORT [10] to track detected pedestrians.
- Model 2: The model replaces the proposed framework's crowd density estimation model with MCNN [1]. MCNN has a limited ability compared with CAN.
- Model 3: This model is the same as the model described in Section IV.

These three models are tested on Video 1, and then Model 3 is tested on Video 2 and Video 3. The curves for the features in these experiments are shown in Fig. 9. The curves for the Volatility descriptor start from the twentieth frame due to we set $m = 20$ in (7). Moreover, the background is painted in different colors according to the labels of the crowd behaviors. The green background represents the crowd in this part wandering; the yellow parts represent merging; the red parts represent walking gathering in Video 1 and represent standing gathering in Video 2; the blue parts represent dispersing. Fig. 9(a)-9(o) are used to analyze the crowd feature extraction modules' effects on the descriptors. Fig. 9(k)-9(y) are used to analyze the relation between crowd behavior patterns and the proposed descriptors.

A. Feature Extraction Effects on Descriptors

In our framework, the crowd density estimation model is adopted as the crowd feature extraction method. Therefore, the proposed framework can exclude the influence of non-human objects. Besides, the crowd density estimation model also enhances the crowd feature extraction ability of our framework. To investigate how the crowd density estimation model affects our framework, Model 1, 2, and 3 are tested on Video 1. The result for different models are compared, and the effects of the crowd feature extraction models are analyzed. The result can be found in Fig. 9(a)-9(o).

As shown in the figures, the descriptors output by Model 3 have identifiable features in each part. The curves for the descriptors of Model 2 present a roughly consistent trend with those of Model 1. The curves can indicate the crowd behaviors with small errors. However, more outliers and fluctuations appear in the curves of Model 2, due to the limited ability of crowd feature extraction, which demonstrates a noisy curve in Fig. 9(f)-9(j). For Model 1, only the Irregularity and Sparsity descriptors have similar trends to other models, and the other descriptors cannot depict the crowd behaviors. One potential reason is that Model 1 uses YOLO and Deep SORT together to extract crowd features. The YOLO model would miss some objects when people overlap each other. The extracted features cannot accurately represent the crowd behavior due to the missing objects. Therefore, the curves contain many biases.

The comparison demonstrates that the crowd feature extraction module plays a crucial role in the whole framework, and provides the framework with significant performance.

B. Analysis of Crowd Behavior Pattern

This paper proposes four descriptors for representing crowd behavior. In this section, we test Model 3 on Video 1, 2, and 3 respectively, and record the curves for the descriptors. The background is also colored according to the behaviors. Experiment results are as shown in Fig. 9(k)-9(y).

According to the figures, the curves for the descriptors of different behaviors demonstrate different tendencies. For the crowd, which is not standing gathering or walking gathering, its Irregularity and Volatility descriptors are great and the number of people is small. This situation indicates an organized

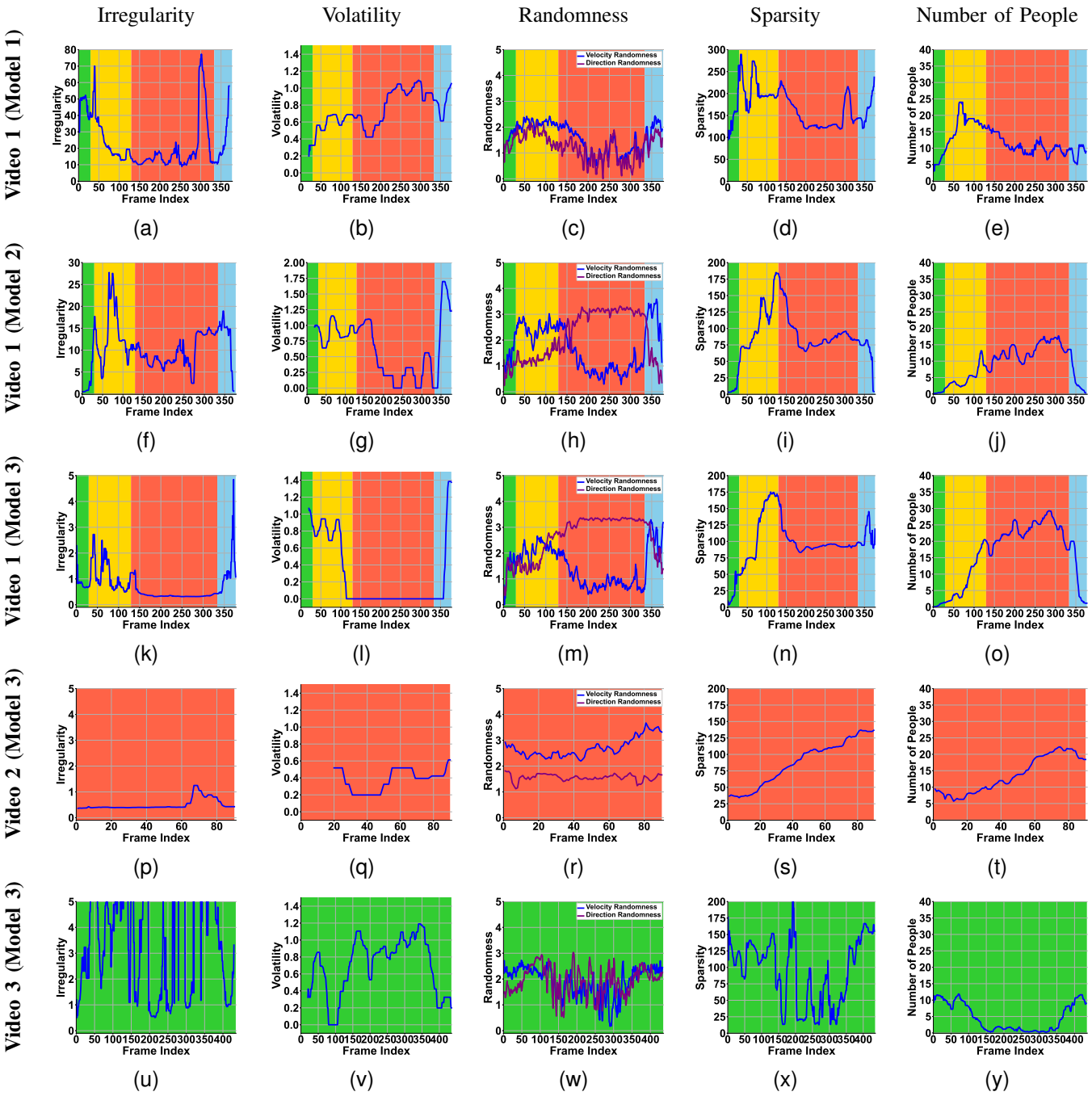

Fig. 9. The experiment models are tested on three example video clips. The curves for the descriptors are drawn here. The background is painted in different colors according to the behavior labels: green–wandering; yellow–merging; red–standing or walking gathering; blue–dispersing. The (a)-(o) are compared in Section VI-A. The (k)-(y) are analyzed in Section VI-B.

group appears in the camera sight, which has a small number of people.

For uncrowded scenes, all groups produced by the division algorithm would have a small number of people, due to the people in the crowd demonstrating different motions. Therefore, the Randomness and Sparsity of the groups are small. On the contrary, for crowded scenes, the group division algorithm would assign all people into one group, which has a massive number of people. The people in the group have different motions, and some people move in opposite directions. Therefore, the Randomness can be large. Moreover, the Sparsity of the group is large as a result of the wide distribution of the group in the camera sight.

When the gathering event occurs, the Irregularity and Volatility of the group would decrease and become flat. Meanwhile, the number of people increases. When a standing gathering event occurs, two peaks would appear before and after the event respectively in the Sparsity curve. During the process of people merging, the group size grows gradually. The group could be loose at the moment some people begin to be assigned to the gathering group, thus, its Sparsity is large, which forms the first peak. After a short time, the group

gathers more closely, and the Sparsity decreases. Likewise, the group has a large Sparsity at the beginning of dispersing. At this moment, the group could become loose, but the algorithm still assigned them to one group. Therefore, the other peak of the Sparsity is formed.

Furthermore, Randomness can identify gathering types. The Velocity Randomness and Direction Randomness can be relatively flat and have a parallel trend with small errors when the group is walking gathering. The Velocity Randomness would be larger than the Direction Randomness. Standing gathering demonstrates a small Velocity Randomness and a large Direction Randomness, which may be caused by the vibration of individuals in the group. The vibration stems from the small movement of individuals, such as shaking and rotating. Therefore, we can identify the behavior of standing gathering according to the Randomness. For a wandering crowd, the Randomness shows noisy curves.

In summary, the proposed descriptors can describe crowd gathering behavior adequately. The crowd behavior can be identified only with the curves of the descriptors. Consequently, the proposed framework can achieve remarkable performance based on the descriptors.

C. Ablation Study

TABLE IV
F1 SCORE VARIATIONS COMPARED WITH THE ORIGIN MODEL.

Setting	Micro-F1(%)	Macro-F1(%)
Without Number Of People	-2.74	-5.10
Without Irregularity	-0.49	-0.75
Without Volatility	-0.59	-2.47
Without Randomness	-4.02	-4.21
Without Sparsity	-2.34	-6.17

1) Descriptors: In this part, the effect of each descriptor on our framework is studied. We set five models by removing these five features and their corresponding variations respectively. The models are tested with the same experimental setting as described in Section V and the data of *All Views*. The result is shown in Table IV. As shown in Table IV, all modified frameworks demonstrate a limited prediction ability. Removing the number of people, Randomness and Sparsity affect the framework performance the most, and removing Irregularity and Volatility have relatively small influences. When a gathering event occurs, the velocity and direction of the crowd can be highly consistent. Thus, Randomness descriptors affect the most. Moreover, the gathering group is tight, which has a small Sparsity and a massive number of people. Therefore, the Sparsity descriptor and the number of people also have a large influence on the framework. On the other hand, not all gathering groups have a large Irregularity, due to the unnecessity of being like a troop for a gathering group. Besides, few samples have highly regular groups in the PETS2009 dataset. Because of this, the test result demonstrates a small influence when Irregularity is removed. Volatility is computed by the number of groups. Therefore, Volatility is also sensitive to the changes of the wandering individuals. Removing each descriptor shows different performance reductions. All descriptors show significant contributions to the quantification of the crowd gathering behaviors.

TABLE V
ABLATION STUDY FOR DIFFERENT DENSITY ESTIMATION MODELS.

Model Name	Front-end	Micro-F1	Macro-F1
CAN [2]	VGG [62]	**0.9534**	**0.9003**
CAN [2]	ResNet [19]	0.9139	0.8183
CAN [2]	Swin [20]	0.8890	0.8487
CCTrans [5]	Twins [68]	0.9415	0.8652
SASNet [4]	VGG [62]	0.9456	0.8540

2) Density Estimation Model: The CAN originally utilizes VGG as its front-end network. In recent years, there have been significant advancements in backbone networks, outperforming the VGG. To assess the impact of the front-end and the density estimation performance on our framework, we modify the CAN by replacing its front-end network with ResNet and Swin Transformer, and train them on the dataset named ShanghaiTech Part B [1]. The trained model are adopted as the density estimation model in our proposed framework.

For the CANs with different front-end network, according to the result, the VGG-based model achieves the highest score, and the Swin-based model performs the least effectively. The original CAN (i.e., VGG-based model), using the first ten convolutional layers of VGG-16 as the front-end network, is adopted as a benchmark. In order to guarantee a fair comparison, we only use the first nine convolutional layers of ResNet-18. Firstly, an output with the same resolution as the VGG-based model can be obtained in this way. Secondly, like VGG-16, ResNet-18 has a comparable number of layers, meaning that removing some layers would result in a similar reduction of model capability. Meanwhile, due to the similar number of layers, the ResNet-based and VGG-based CANs would have comparable numbers of parameters and FLOPs. Moreover, when the network is shallow, the residual connection only makes the ResNet more complex, which is designed to mitigate degradation in deeper networks. As a result, ResNet cannot fully demonstrate its performance. Conversely, VGG boasts a simpler and more flexible structure, allowing for easy customization to suit various applications, and all convolutional layers of VGG can be used. Therefore, the VGG-based model outperforms the ResNet-based one. As for the Swin-based model, CAN would underperform when paired with a Transformer-based front-end network. Transformer-based and CNN-based models generally employ distinct feature spaces and methods for encoding spatial features. Thus, when the CNN-based back-end decoder in CAN attempts to learn from feature maps encoded by a Swin Transformer, the task can be more challenging. Furthermore, the Swin-based models often have a more intricate structure and necessitate larger training datasets compared to CNN-based models, leading to more powerful generalization capabilities. However, in cases where the task involves a limited dataset, CNN-based models would yield better performance. Therefore, in this specific case, the VGG-based model can outperform the Swin Transformer-

based model. Consequently, VGG is the preferred choice for CAN due to its considerable performance.

Moreover, we also test and compare the frameworks with CCTrans [5] and SASNet [4] as the density estimation model, both of which represent the current state-of-the-arts of the Transformer-based and the CNN-based models. The comparisons indicate that the highest score is achieved when the framework adopts CAN as the density estimation model. One potential explanation for this superiority is that the CAN considers the perspective effect, a factor not taken into account by CCTrans and SASNet. The consideration of information about the perspective effect proves the beneficial in enhancing the feature extraction capability of the model. Nevertheless, the state-of-the-art models significantly contribute to the considerable performance of their respective frameworks.

VII. Conclusion

In this paper, we define four crowd descriptors Irregularity, Sparsity, Randomness, and Volatility, respectively. Besides, we further divide the crowd gathering behavior into wandering, merging, standing gathering, walking gathering, and dispersing to improve the crowd gathering understanding. Based on the crowd descriptors, we propose a novel framework for crowd gathering understanding, which consists of three modules. The crowd feature extraction estimates the crowd flow map by performing an element-wise product between the extracted crowd density map and the dense optical flow map. The crowd behavior quantification calculates the crowd descriptors based on the groups divided according to their motion. The classification predicts the crowd behavior based on the descriptors and the number of people. To verify our framework for crowd gathering understanding, we conduct experiments on the PETS2009 dataset. The experiments demonstrate the effectiveness and the much better interpretability of the proposed framework. Moreover, to compare with existing methods, we process the results of the experiments by integrating the five behaviors into two most commonly used states of non-gathering and gathering, and then compute the accuracy score to compare with other methods. The numerical results of the experiments demonstrate our framework outperforms the existing works on the crowd gathering understanding. The framework also achieves a further understanding of crowd gathering behavior.

Our method has shown considerable performance, nevertheless, there are still some limitations to be addressed. Our framework cannot deal with large-scale scenes such as railway stations and other occasions with a large number of groups. The framework only considers the group with the most people, whereas complex scenes usually contain many groups having different states. Therefore, in future work, its performance would be improved in more complex scenes. Also, the group division algorithm works currently based on the motion features of pedestrians. However, the density based clustering algorithm (the DBSCAN algorithm we used) tends to cluster all people into one group in complex scenes. A general clustering can not adequately complete the work. Thus, a specific group division algorithm is required in future work. Finally, non-human objects are ignored, such as backgrounds and banners. To understand crowd gathering behavior better, future work could further mine more semantic information conveyed by video frames.

References

[1] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 589–597.

[2] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5099–5108.

[3] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1091–1100.

[4] Q. Song, C. Wang, Y. Wang, Y. Tai, C. Wang, J. Li, J. Wu, and J. Ma, "To choose or to fuse? scale selection for crowd counting," in *Proc. Conf Artificial intell.*, vol. 35, no. 3, 2021, pp. 2576–2583.

[5] Y. Tian, X. Chu, and H. Wang, "Cctrans: Simplifying and improving crowd counting with transformer," *arXiv preprint arXiv:2109.14483*, 2021.

[6] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, and S. Gao, "Video anomaly detection with sparse coding inspired deep neural networks," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 43, no. 3, pp. 1070–1084, 2021.

[7] D. Chen, L. Yue, X. Chang, M. Xu, and T. Jia, "Nm-gan: Noise-modulated generative adversarial network for video anomaly detection," *Pattern Recognit.*, vol. 116, p. 107969, 2021.

[8] M. Ravanbakhsh, E. Sangineto, M. Nabi, and N. Sebe, "Training adversarial discriminators for cross-channel abnormal event detection in crowds," in *IEEE Winter Conf. Appl. Comput. Vis.* IEEE, 2019, pp. 1896–1904.

[9] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *Proc. IEEE Int. Conf. Image Process.* IEEE, 2016, pp. 3464–3468.

[10] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *Proc. IEEE Int. Conf. Image Process.* IEEE, 2017, pp. 3645–3649.

[11] Z. Wang, L. Zheng, Y. Liu, Y. Li, and S. Wang, "Towards real-time multi-object tracking," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 107–122.

[12] X. Zhou, X. Xu, W. Liang, Z. Zeng, and Z. Yan, "Deep-learning-enhanced multitarget detection for end–edge–cloud surveillance in smart iot," *IEEE Internet Things J.*, vol. 8, no. 16, pp. 12 588–12 596, 2021.

[13] S. Blunsden and R. Fisher, "The behave video dataset: ground truthed video for multi-person behavior classification," *Ann. BMVA*, vol. 4, no. 1-12, p. 4, 2010.

[14] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6479–6488.

[15] C.-Y. Liu, W.-H. Liao, and S.-J. Ruan, "Crowd gathering detection based on the foreground stillness model," *IEICE Trans. Inf. Syst.*, vol. 101, no. 7, pp. 1968–1971, 2018.

[16] D.-S. Yang, C.-Y. Liu, W.-H. Liao, and S.-J. Ruan, "Crowd gathering and commotion detection based on the stillness and motion model," *Multimedia Tools Appl.*, vol. 79, no. 27, pp. 19 435–19 449, 2020.

[17] J. Xu, H. Zhao, W. Min, Y. Zou, and Q. Fu, "Dgg: A novel framework for crowd gathering detection," *Electronics*, vol. 11, no. 1, p. 31, 2021.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.*, vol. 25, 2012.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.

[21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[22] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

[23] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao *et al.*, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld Ubiquitous Intell. Comput. Adv. Trusted Computed Scalable Comput. Commun. Cloud Big Data Comput., Internet People Smart City Innov.* IEEE, 2017, pp. 1–6.

[24] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[25] Z. Wang, C. Liu, and X. Cui, "Evilmodel: hiding malware inside of neural network models," in *2021 IEEE Symp. Comput. Commun.* IEEE, 2021, pp. 1–7.

[26] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, p. eaay7120, 2019.

[27] B. Zhou, X. Tang, and X. Wang, "Measuring crowd collectiveness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3049–3056.

[28] J. Shao, C. C. Loy, and X. Wang, "Learning scene-independent group descriptors for crowd understanding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 6, pp. 1290–1303, 2016.

[29] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical Rev. E*, vol. 51, no. 5, p. 4282, 1995.

[30] D. Helbing, A. Johansson, and H. Z. Al-Abideen, "Dynamics of crowd disasters: An empirical study," *Physical Rev. E*, vol. 75, no. 4, p. 046109, 2007.

[31] D. Helbing and P. Mukerji, "Crowd disasters as systemic failures: analysis of the love parade disaster," *EPJ Data Sci.*, vol. 1, no. 1, pp. 1–40, 2012.

[32] W. Yu and A. Johansson, "Modeling crowd turbulence by many-particle simulations," *Physical Rev. E*, vol. 76, no. 4, p. 046105, 2007.

[33] M. Moussaïd, D. Helbing, and G. Theraulaz, "How simple rules determine pedestrian behavior and crowd disasters," *Proc. Nat. Acad. Sci.*, vol. 108, no. 17, pp. 6884–6888, 2011.

[34] N. W. Bode, A. J. Wood, and D. W. Franks, "Social networks and models for collective motion in animals," *Behav. Ecol. Sociobiol.*, vol. 65, no. 2, pp. 117–130, 2011.

[35] D. J. Sumpter, "The principles of collective animal behaviour," *Philos. Trans. Royal Soc. B: Biol. Sci.*, vol. 361, no. 1465, pp. 5–22, 2006.

[36] A. Deutsch, G. Theraulaz, and T. Vicsek, "Collective motion in biological systems," *Interface Focus*, vol. 2, no. 6, pp. 689–692, 2012.

[37] W. H. Warren, "Collective motion in human crowds," *Current Directions Psychol. Sci.*, vol. 27, no. 4, pp. 232–240, 2018.

[38] W. Liu, R. W. Lau, X. Wang, and D. Manocha, "Exemplar-amms: Recognizing crowd movements from pedestrian trajectories," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2398–2406, 2016.

[39] W. Liu, R. W. Lau, and D. Manocha, "Robust individual and holistic features for crowd scene classification," *Pattern Recognit.*, vol. 58, pp. 110–120, 2016.

[40] Y. Zou, X. Zhao, and Y. Liu, "Measuring crowd collectiveness by macroscopic and microscopic motion consistencies," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3311–3323, 2018.

[41] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 42, no. 1, pp. 46–58, 2018.

[42] A. K. Pai, P. Chandrahasan, U. Raghavendra, and A. Karunakar, "Motion pattern-based crowd scene classification using histogram of angular deviations of trajectories," *Vis. Comput.*, vol. 39, no. 2, pp. 557–567, 2023.

[43] C. W. Reynolds, "Flocks, herds and schools: A distributed behavioral model," in *Proc. 14th Annu. Conf. Comput. Graph. Interact. Techn.*, 1987, pp. 25–34.

[44] J. v. d. Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robot. Res.* Springer, 2011, pp. 3–19.

[45] M. Li, T. Chen, H. Du, N. Ma, and X. Xi, "Social group detection based on multi-level consistent behaviour characteristics," *Transportmetrica A: Transp. Sci.*, vol. 19, no. 1, p. 1976877, 2023.

[46] Y. Zou and Y. Liu, "Modeling pedestrian motion in crowded scenes based on the shortest path principle," *Appl. Sci.*, vol. 12, no. 1, p. 381, 2021.

[47] S. Behera, D. P. Dogra, M. K. Bandyopadhyay, and P. P. Roy, "Crowd characterization in surveillance videos using deep-graph convolutional neural network," *IEEE Trans. Cybern.*, vol. 53, no. 6, pp. 3428–3439, 2023.

[48] M. Simon, E. Bochinski, M. Küchhold, and T. Sikora, "Utilizing crowd collectiveness to enhance bottleneck detection based on the lagrangian framework," in *2022 18th IEEE Int. Conf. Adv. Video Signal Based Surveillance*. IEEE, 2022, pp. 1–8.

[49] N. Japar, V. J. Kok, and C. S. Chan, "Collectiveness analysis with visual attributes," *Neurocomputing*, vol. 463, pp. 77–90, 2021.

[50] Y. Feng, Y. Yuan, and X. Lu, "Learning deep event models for crowd anomaly detection," *Neurocomputing*, vol. 219, pp. 548–556, 2017.

[51] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5017–5032, 2015.

[52] M. Ravanbakhsh, M. Nabi, H. Mousavi, E. Sangineto, and N. Sebe, "Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.* IEEE, 2018, pp. 1689–1698.

[53] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *Proc. IEEE Int. Conf. Image Process.* IEEE, 2017, pp. 1577–1581.

[54] Y. Li, "A deep spatiotemporal perspective for understanding crowd behavior," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3289–3297, 2018.

[55] T. Gupta, V. Nunavath, and S. Roy, "Crowdvas-net: A deep-cnn based framework to detect abnormal crowd-motion behavior in videos for predicting crowd disaster," in *Proc. IEEE Int. Conf. Syst. Man Cybern.* IEEE, 2019, pp. 2877–2882.

[56] L. Breiman, "Random forests," *Mach. learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[57] M. Yang, S. Tian, A. S. Rao, S. Rajasegarar, M. Palaniswami, and Z. Zhou, "An efficient deep neural model for detecting crowd anomalies in videos," *Appl. Intell.*, vol. 53, no. 12, pp. 15 695–15 710, 2023.

[58] B. Zhang, R. Zhang, N. Bisagno, N. Conci, F. G. De Natale, and H. Liu, "Where are they going? predicting human behaviors in crowded scenes," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 17, no. 4, pp. 1–19, 2021.

[59] J. Su, J. Huang, L. Qing, X. He, and H. Chen, "A new approach for social group detection based on spatio-temporal interpersonal distance measurement," *Heliyon*, vol. 8, no. 10, 2022.

[60] T. Alafif, B. Alzahrani, Y. Cao, R. Alotaibi, A. Barnawi, and M. Chen, "Generative adversarial network based abnormal behavior detection in massive crowd videos: a hajj case study," *J. Ambient Intell. Humanized Comput.*, pp. 1–12, 2021.

[61] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 833–841.

[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[64] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conf. Image Anal.* Springer, 2003, pp. 363–370.

[65] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[66] J. Ferryman and A. Shahrokni, "Pets2009: Dataset and challenge," in *IEEE Int. Workshop Perform. Eval. Tracking Surveillance*. IEEE, 2009, pp. 1–6.

[67] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[68] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 9355–9366, 2021.

Yuxin Zhou received his B.E. degree in robotics engineering from Harbin Institute of Technology, Weihai, P.R.China, in 2020. He is currently studying as a Master of Engineering research student at the Southern University of Technology and Science, Shenzhen, P.R.China. His research interests include deep learning, image processing, and crowd analysis.

Chenguang Liu received his B.E. degree in software engineering from Dalian University of Technology, Dalian, P.R.China, in 2016 and MSc degree in advanced computer science from The University of Manchester, U.K., in 2017. He is currently studying as a PhD student at the University of Warwick, U.K. His research interests include deep learning, wireless communications and signal detection.

Yulong Ding received the BSc and MSc degrees in chemical engineering from Tsinghua University, Beijing, China, in 2005 and 2008, respectively, and the PhD degree in chemical engineering from The University of British Columbia, Canada, in 2012. He is currently a Research Assistant Professor with the Department of Computer Science and Engineering, Southern University of Science and Technology. His main interests are industrial Internet of Things and low-power wide area networks(LPWANs).

Diping Yuan is currently a professor and dean of Shenzhen Research Institute, China University of Mining and Technology, Shenzhen, China. He is also the deputy head of the Expert Commission of Emergency Management of Guangdong Province, the director of China Fire Protection Association, and a member of Emergency Management and Disaster Reduction and Response (SAC/TC307). He is mainly engaged in research on big data application for urban public safety and emergency management, safety risk monitoring and early warning, intelligent emergency decision-making and simulation, and smart fire protection.

Jiyao Yin is currently the director of Informatization Commission, Shenzhen Urban Public Safety and Technology Institute, Shenzhen, China. He is also the deputy director of Informatization Special Commission of Emergency Governance Society of Shenzhen. His current research interests are smart emergency construction, big data analysis for urban safety, and intelligent agents for urban safety.

Shuang-Hua Yang received the B.S. degree in instrument and automation and the M.S. degree in process control from the China University of Petroleum (Huadong), Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree in intelligent systems from Zhejiang University, Hangzhou, China, in 1991. He was awarded DSc from Loughborough University in 2014 to recognize his academic contribution to wireless monitoring research. He is currently a professor and the Head of Department of Computer Science at the University of Reading, the UK and the Director of Shenzhen Key Laboratory of Safety and Security for Next Generation of Industrial Internet, Southern University of Science and Technology (SUSTech), Shenzhen, China. His current research interests include cyber-physical system safety and security, Internet of Things. He is a Fellow of IET and a Fellow of InstMC, U.K. He is an Associate Editor of the IET Journal Cyber-Physical Systems Theory and applications.