A Hierarchical Spatio-Temporal Graph Convolutional Neural Network for Anomaly Detection in Videos

Xianlin Zeng, Yalong Jiang, Wenrui Ding, Hongguang Li, Yafeng Hao, and Zifeng Qiu

***Abstract*— Deep learning models have been widely used for anomaly detection in surveillance videos. Typical models are equipped with the capability to reconstruct normal videos and evaluate the reconstruction errors on anomalous videos to indicate the extent of abnormalities. However, existing approaches suffer from two disadvantages. Firstly, they can only encode the movements of each identity independently, without considering the interactions among identities which may also indicate anomalies. Secondly, they leverage inflexible models whose structures are fixed under different scenes, this configuration disables the understanding of scenes. In this paper, we propose a Hierarchical Spatio-Temporal Graph Convolutional Neural Network (HSTGCNN) to address these problems, the HSTGCNN is composed of multiple branches that correspond to different levels of graph representations. High-level graph representations encode the trajectories of people and the interactions among multiple identities while low-level graph representations encode the local body postures of each person. Furthermore, we propose to weightedly combine multiple branches that are better at different scenes. An improvement over single-level graph representations is achieved in this way. An understanding of scenes is achieved and serves anomaly detection. High-level graph representations are assigned higher weights to encode moving speed and directions of people in low-resolution videos while low-level graph representations are assigned higher weights to encode human skeletons in high-resolution videos. Experimental results show that the proposed HSTGCNN significantly outperforms current state-of-the-art models on four benchmark datasets (UCSD Pedestrian, ShanghaiTech, CUHK Avenue and IITB-Corridor) by using much less learnable parameters.**

***Index Terms*— Anomaly detection, human skeletons, graph convolutional network, hierarchical graph representations, weightedly combination, understanding of scenes.**

Manuscript received 24 June 2021; revised 20 August 2021; accepted 6 December 2021. Date of publication 10 December 2021; date of current version 6 January 2023. This work was supported in part by the Aeronautical Science Fund of China under Grant 2020Z071051001 and in part by the National Natural Science Foundation of China under Grant U20B2042 and Grant 62076019. This article was recommended by Associate Editor Y.-P. Tan. *(Corresponding authors: Yalong Jiang; Wenrui Ding.)*

Xianlin Zeng is with the School of Electrical and Information Engineering, Beihang University, Beijing 100191, China (e-mail: zengxianlin@buaa.edu.cn).

Yalong Jiang, Wenrui Ding, and Hongguang Li are with the Institute of Unmanned System, Beihang University, Beijing 100191, China (e-mail: allenyljiang@buaa.edu.cn; ding@buaa.edu.cn; lihongguang@buaa.edu.cn).

Yafeng Hao is with the 54th Research Institute, China Electronics Technology Group Corporation (CETC), Shijiazhuang, Hebei 050081, China (e-mail: datoushr@126.com).

Zifeng Qiu is with the Laboratory of Aerospace Information Applications, China Electronics Technology Group Corporation (CETC), Shijiazhuang, Hebei 050081, China (e-mail: qzf93@qq.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3134410.

Digital Object Identifier 10.1109/TCSVT.2021.3134410

I. Introduction

HUMAN-RELATED anomaly detection is the task of localizing from videos the activities that do not match regular patterns. The sophistication of anomaly semantics as well as the imbalance problem in anomaly-related datasets pose the need for detailed annotations that require expensive human labor. However, even detailed annotations can not ensure the generalization of real-world scenarios. It is, therefore, necessary to formalize the task as an unsupervised learning task that requires automatically discriminating small quantities of behavioral irregularity from the vast majority of normal events and is extremely challenging. This article concentrates on the study of abnormal events that are related to human behaviors and aims at developing models that characterize the feature patterns presenting in human behaviors and leverage the learned features to localize the short clips with behavioral irregularity from videos.

Existing unsupervised deep learning models develop the feature representations describing regular behaviors through training on video sequences with only normal events. The models reconstruct input data with learned embeddings which are optimized to produce low reconstruction errors on normal behaviors and the errors increase on behavioral irregularities. These works can be roughly divided into two categories: i) Reconstruction in the unit of pixels. The approach proposed in [1] extracted features from historical frames with a Generative Adversarial Network (GAN) which then function in predicting future frames. Both temporal ranges and spatial locations of anomalies are determined by the GAN-based method. However, intensity-based features involve low-level clues which are quite noisy. For instance, the entanglement of background regions with foreground regions is difficult to be tackled. More seriously, real-world interferences such as dramatic changes in lightening conditions can lead to high reconstruction errors which are predicted by models as anomalies. As a result, the false alarm rate increases. On the other hand, the GAN-based models for per pixel reconstruction require huge computational burdens and are inefficient; ii) Reconstruction in the unit of body joints. These methods take advantage of the rich semantic features in critical regions and can describe human behaviors under lower noises and with higher efficiency. Paper [2] decomposed the movements of human skeletons into two sub-processes which describe the variation in moving speed and the variances in poses of

each person, respectively, it proposed a model called Message-Passing Encoder-Decoder Recurrent Network (MPED-RNN) which is similar to the LSTM Auto Encoder (LSTM AE) proposed in [3]. However, the RNN-based approaches are limited by the following disadvantages: First of all, the Seq2seq architecture proposed in [4] is fully-supervised and is only suitable for making predictions on the limited types of actions from training data. Secondly, this type of methods detect the abnormal behaviors of each person independently without fully considering the interactions among different people. Recently, Luo et al. [5] proposed Normal Graph for skeleton-based video anomaly detection, it explores the moving patterns of body joints in normal behaviors. Whereas, it only leverages skeleton-related features which are quite noisy in low-resolution videos. As a result, Normal Graph fails on low-resolution datasets, such as UCSD Pedestrian [6]. By integrating high and low-level graph representations which describe the interactions among different people and show single person movements, respectively, the proposed approach achieves an improvement of about 7% over Normal Graph and MPED-RNN with less learnable parameters on both high-resolution datasets such as ShanghaiTech and low-resolution datasets such as UCSD.

To leverage the strength of GCNN in analyzing structured data, we utilize STGCNN to effectively encode the spatial and temporal embeddings of human skeletons and determine the temporal inconsistency in human behaviors based on motion prediction. We name it Hierarchical Spatial-Temporal Graph Convolutional Neural Network (HSTGCNN). The HSTGCNN includes three components: a spatio-temporal graphical feature extractor, a future trajectory predictor, and an outlier arbiter. In detail, we firstly organize the inputs into a spatio-temporal graph whose nodes are human body joints from multiple frames and then feed it to the HSTGCNN to obtain intermediate feature representations that describe the motion vectors of semantic joints and indicate both directions and extents of movements. For instance, the action is more likely to be fighting if limbs are with much more strenuous movements than the head and torso. The HSTGCNN is trained on the input graph representations which correspond to normal activities and achieves the capability to accurately predict the trajectories of human joints in normal behaviors. The coordinates of human joints in historical frames, as well as future frames, are organized in the form of hierarchical graph representations, as is shown in Fig. 1. A high-level graph representation is composed of nodes each of which represents a person, it characterizes the relative positions and interactions among individuals, the moving speed of each identity is also encoded in the graph representations. A low-level graph representation indicates the pose of a person and is leveraged in the detection of abnormal actions of a single individual.

Furthermore, the weighted combination of different levels of graph representations enables the proposed framework to exhibit robustness to the variances in scenes. Different branches conduct inference on different levels of graph representations and are better at handling different scenes some of which are with dense small human objects while others

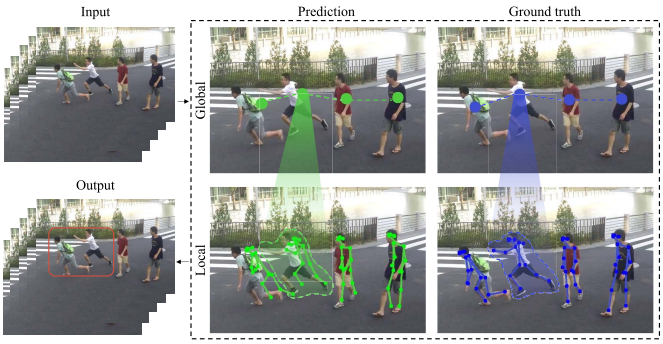

Fig. 1. The proposed scheme for hierarchical graph representations. The top row corresponds to the high-level graph representations where each node represents one individual. The bottom row shows low-level graph representations where each node describes one body joint. Individuals in the input video frame are encoded by high-level and low-level graph representations.

contain sparse and large humans. To acquire an understanding of scenes and determine the weights of branches accordingly, we propose to use optical flow fields and average sizes of human bounding boxes and skeletons to cluster videos into different groups that correspond to different scenes. We determine the weight of branches which minimizes the loss of predictions in normal behaviors. For scenes with dense crowds and small people, higher weights are assigned to high-level graph representations while in sparse scenes, the weights of low-level graph representations are increased. Finally, the outlier arbiter weightedly sums the predictions from independent branches to obtain the final anomaly score. The model can produce effective detections of abnormal events only by being trained in an end-to-end manner on a small number of video sequences without anomalies. Experimental results on standard benchmarks, including UCSD Pedestrian [6], ShanghaiTech [7], CUHK Avenue [8] and IITB-Corridor [9], demonstrate the effectiveness and efficiency of our approach, it outperforms state-of-the-art approaches while using much less learnable parameters.

The main contributions of this article can be summarized as follows:

(1) We propose a Spatio-Temporal Graph Convolutional Architecture that integrates branches each of which corresponds to the graph representations in a certain level. Low-level graph structures represent the spatial and temporal embeddings of human body joints. High-level graph structures not only characterize the speed and direction of each pedestrian but also represent behavioral irregularities by modeling the interactions among multiple identities.

(2) Different branches are better at different scenes (sparsely distributed people with high resolutions and dense small-scale humans) grouped by clustering. The predictions from different branches are weightedly combined to obtain the final anomaly score and achieve an understanding of scenes.

(3) Compared with other STGCNN-based methods, our proposed framework can overcome drastic changes in scenes by using high-level graph representations combined with skeleton-based features and perform well on both high-quality videos and low-quality ones where humans are of low resolutions.

The rest of the paper is organized as follows. Section II describes the related work on human pose estimation, human action recognition, and video anomaly detection. In Section III, the pre-processing of human skeletons is given full description firstly. Then the details of the overall framework and each part are presented. Section IV first introduces dataset and implementation details, then conducts several comprehensive experiments to demonstrate the effectiveness of the proposed model and perform analysis on errors. Finally, Section V summarized.

II. Related Work

A. Human Pose Estimation in Videos

Human pose estimation in computer vision is the task of estimating the pose of articulated bodies. Existing methods for pose estimation are divided into 2D approaches and 3D ones. Among the 2D pose estimation methods, the first type of method only works on images with a single person [10]. These methods firstly localize semantic parts and then leverage the connections between parts to describe gestures. Single-person pose estimations can not be applied to many real-world scenarios. The second type of method conducts multi-person pose estimation by using a human detector to localize humans before estimating the pose of each person. This type of methods are named top-down methods and include Mask-RCNN [11], RMPE [12], CPN [13], etc. The third type of method detects the body joints of all people in an image before grouping them into people. This type of methods are called bottom-up methods and include Open Pose [14], Deep Cut [15], Deeper Cut [16] etc.

B. Human Action Recognition in Videos

Related studies have conducted human action recognition based on pose estimations. Earlier work used simple geometric models in representing the structures of human bodies (such as a 2D contour model [17] and a 3D cylinder model [18]), and focused on the dynamics of external contours. Recently, some methods have been proposed to use interactive recurrent networks for modeling human motions on hard scenarios [19], [20]. The above-mentioned approaches all modeled each human as one rigid object, ignoring the local movements of human body joints. To pay attention on both moving speed and pose variances of people, the approaches [21], [22] took human skeletons as the input to RNN and modeled human movements in this way. Recently, paper [23] proposed to divide human skeletons into five parts and feed them into five independent RNNs for feature extraction. The model has been successfully applied in some cases [24]–[26]. However, the RNN-based models suffer from error accumulation, gradient explosion, and gradient vanishing problems due to complex structures. The problem is avoided in our approach with a shallow structure. These methods can only detect some fixed types of actions pre-defined in training data and are difficult to generalize to unpredictable abnormal behaviors.

Skeleton based action recognition is an emerging topic. Skeleton and joint trajectories of human bodies are more robust to illumination changes and scene variations than pixel-level features. As a result, it has been widely used in video action recognition tasks [27], [28], pose tracking [29], etc. [30] applied the GCN over the graph to establish the relationship between two action proposals to boost the performance of temporal action localization. It achieved good generalization ability and the method [31] also received great attention. [32] also built the GCN on the individual level for action recognition but lacked the modeling of interactions.

However, the above-mentioned models for action recognition suffer from the following problems: i) Most of them [33] take into account the spatial and temporal relations but only work on temporally trimmed videos which include only one fixed type of action across an entire video. ii) The reasons leading to anomalies are diverse and unpredictable, resulting in sophisticated semantics that can hardly be covered by existing training data. Existing supervised action recognition methods only detect a limited number of action types that are annotated in training data [32]. Inspired by the application of unsupervised domain in pedestrian re-identification [34], our proposed approach can detect abnormal behaviors in untrimmed videos through unsupervised training and does not require the types of anomalies to appear in training data.

C. Human Anomaly Detection in Videos

In the community of video processing and computer vision, early algorithms treated each frame as an independent sample, superpixel techniques were employed to figure out the motion orientations of objects and conducted anomaly detection simultaneously [35]. Paper [36] conducted a detailed investigation into the videos and images deep learning based anomaly detection methods. Typical machine learning techniques involve clustering (K-means [37], GMM [38], OC-SVM [39], etc.) and reconstructing discriminant analysis (such as PCA [40]–[42] and SC [1], [43]–[46], etc.). When dealing with large-scale data with complex abnormal patterns, these methods often fail to achieve desired results.

The advancements of deep learning have brought many advantages over traditional methods in feature engineering. However, the determination of anomalies is still based on hand-crafted approaches. Typical deep learning-based feature extractors include Auto Encoder (AE) [47], target detection neural network [48], 3D Convolutional Network (C3D) [49] and so on. Although the CNN-based method has achieved great success in many visual tasks, the applications of CNN in anomaly detection are still limited by two constraints. Firstly, CNN shows a heavy reliance on a large amount of training data with full supervision. Secondly, CNN can not deal with non-euclidean graphical data where the number of neighbors around each node is not fixed, as is often the case in human-related graph representations.

As more research efforts been invested into anomaly detection, recent studies combined the feature extraction step with the model training step and proposed deep learning methods in an end-to-end manner, such as VAE [50], Generative Adversarial Network (GAN) [51], Recurrent Neural Network (RNN) [7] and Long Short-Term Memory (LSTM) [52], etc.

Among them, paper [53] used a U-Net-based model to achieve the cross-modal reconstruction of video frames based on generative approaches and optical flow estimations, the reconstruction error was used to determine anomalies. The network utilized long skip connections to reduce the information loss caused by the reduction in dimensionality and achieved better results than AE. Besides, the GAN-based network improved the performance of the generator through adversarial training. For example, the research [54] used Bidirectional GAN (BiGAN) to perform anomaly detection, pixel intensity loss and discriminator loss were also leveraged. Paper [2] proposed a Message-Passing Encoder-Decoder Recurrent Network (MPED-RNN) with human skeletons as inputs, it suppressed pixel-level noises which commonly exist in optical flow estimations, the interpretability of abnormal semantics was also improved. However, using recursive architecture to extract skeleton features is not the optimal solution in that error accumulating in sequential predictions made by RNNs and too many learnable parameters lead to the consumption of huge memory spaces. Paper [5] proposed a single-level graph structure called Normal Graph for skeleton-based video anomaly detection, which takes body joint locations of each pedestrian as inputs and simply calculates the mean square error between the predicted joints and the ground truth. This method shows a heavy reliance on skeleton estimations which are negatively influenced by the reduction in resolutions of human objects. The high-level graph representations in the proposed framework address this issue. Besides, Normal Graph lacks a description of the interactions among multiple people, as result, it suffers from poor accuracy.

Unlike the above-mentioned works, inspired by the success of graph structures in the question answering system [55] which was based on dynamic image understanding and in the traffic forecast system [56], we propose a Hierarchical Spatial-Temporal Graph Convolutional Neural Network (HSTGCNN) which encodes the spatial and temporal relations while constructing hierarchical graph representations in describing group behaviors in anomaly detection. The model can characterize not only the interactions among multiple individuals but also the local movements of each individual by Spatial-Temporal characteristics of graph representations in various resolution video frames and achieve a comprehensive understanding of scenes.

III. Approach

Human-related anomalies can be detected with unsupervised methods through predicting human motions in future video frames based on historical trajectories. The models are trained to minimize the errors between the predictions and the ground truth only on normal behaviors. Since spatio-temporal graph convolutional networks can comprehensively represent the structures of human bodies and dynamics, we propose a hierarchical graph representation in jointly modeling the inter-connections among identities and the correlations among semantic parts within the same identity. Meanwhile, the weighted combination of multiple branches that are better at handling different scenes is proposed to dynamically adjust the overall model for anomaly detection under various scenes. The feature representations of scenes, including the densities and scales of identities, are proposed to determine the weights of branches. In this section, we will describe the proposed framework in detail.

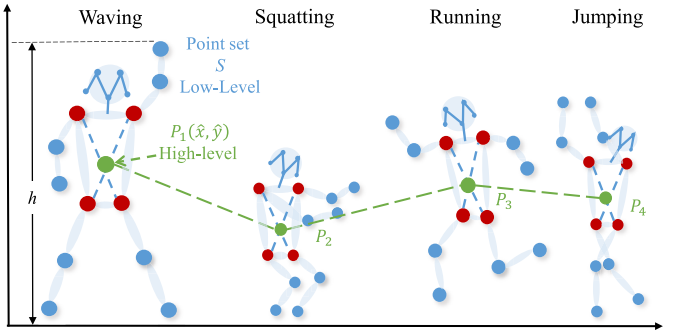

Fig. 2. The demonstration of the high/low-level graph representations. Low-level graph representations are the blue graphs denoted by S and each node corresponds to a body joint. High-level graph representations are shown by the green graph where each node is denoted by P_i, each node represents the feature embeddings of one identity. S denotes the set of all human skeleton trajectories in T consecutive video frames; P_i is the geometric center of the i-th person.

A. Problem Formulation and Skeleton Modeling

Denote the set of human skeleton trajectories in T consecutive video frames as S, expressed as:

$$S = \{f_t \mid t \in \{1, \cdots, T\}\}, \tag{1}$$

where,

$$f = \left\{\left(x_{m,n}, y_{m,n}\right)\right\} \mid m \in \{1, \cdots, M\}; n \in \{1, \cdots, N\}, \tag{2}$$

where m is the index of individuals in a single frame, n is the index of each individual's joints, and f_t represents the set of coordinates in the t-th frame.

The input data denoted by S is organized in a graph structure and is fed to hierarchical graph convolutional networks which describe the motion of single-individuals and the correlations among multiple individuals. The proposed graph convolutional network for anomaly detection learns to predict the future trajectories of each individual, including both the trajectories of human centers and those of body joints. For video sequences with only normal behaviors, the model can accurately predict the trajectories of joints in future frames because the motions in all short clips subject to the same distribution. For the videos with anomalies, there is an obvious difference between joint trajectories and the model's predictions in the occurrences of anomalies. Typical anomalies involve dramatic changes in posture and rapid movements. In order to comprehensively describe the difference between predicted graph structures and the ground truth, we combine high-level graph structures with low-level graph structures, the former regards human bodies as rigid objects while the latter treats different body joints as different nodes.

Some pre-processing methods are required to reduce the influences of task-irrelevant factors on anomaly detection. The most prominent factor is the variances in human sizes, the variances in larger humans' motions are obviously larger and

this is the same with prediction errors, we propose to conduct normalization on joint coordinates with respect to human sizes while maintaining geometric centers. The human skeleton before pre-processing is shown in Eq. 2. The coordinates are normalized with respect to the heights of bounding boxes which are the maximal difference in vertical coordinates of joints' vertical locations: $h_m = \max\left(y_{m,n}\right) - \min\left(y_{m,n}\right)$. Upon normalization, the movement of each body joint is decomposed into the movement of the overall body and the relative motion of the joint with respect to its body center. The body geometric center P is obtained by computing the mean value of the coordinates of the four joints from the human torso, as shown in Eq. 3 and 4:

$$P = \left(\hat{x}_m, \hat{y}_m\right), \tag{3}$$

$$\hat{x}_m = \frac{\sum\limits_{n\in k} x_{m,n}}{4}, \hat{y}_m = \frac{\sum\limits_{n\in k} y_{m,n}}{4}, k \in [5, 6, 11, 12], \tag{4}$$

where k represents the indices of the four selected points among 17 body joints, which represent different human body parts, the set of joints included in k is shown by the four red points in Fig. 2. More details about the definitions of the 17 body joints are provided in HRNet [57].

First of all, we collect all human skeletons in each frame, and calculate the geometric centers P_i of the i-th person according to Eq. 3 and 4 to obtain high-level graph nodes. Then, we construct the adjacency matrix according to Eq. 10 to obtain high-level graph edges. High-level graph representations are denoted as f_m^g, which encodes the relative position of all individuals,

$$f_m^g = \left\{\left(\hat{x}_m, \hat{y}_m\right) \mid m \in \{1, \cdots, M\}\right\}. \tag{5}$$

By subtracting the coordinates of body joints by the geometric centers in P, we can focus on the variances in poses. We construct low-level graph nodes encoding the proposal relative positions of body joints with respect to body center. The low-level graph edges between semantically connected body joints are 1 and others 0. Low-level graph representations are denoted as $f_{m,n}^l$,

$$f_{m,n}^l = \left\{\left(\tilde{x}_{m,n}, \tilde{y}_{m,n}\right)\right\} \mid m \in \{1, \cdots, M\}; n \in \{1, \cdots, N\}, \tag{6}$$

where,

$$\tilde{x}_{m,n} = \frac{x_{m,n}}{\frac{h_m}{2}}, \quad \tilde{y}_{m,n} = \frac{y_{m,n}}{h_m}. \tag{7}$$

Fig. 2. shows the construction of the high/low-level graph representations, the coordinate of each body joint can be expressed as the summation of a global coordinate which denotes the center of the individuals, and a local coordinate which denotes the offset of the joint from the body center. High-level graph representations encode the speed and relative positions of different individuals. The anomalies caused by the outliers in pedestrian movement speed or suddenly dispersed crowds can be detected by high-level graph representations. Low-level graph representations are better at handling posture changes, such as fighting, falling, and other irregular behaviors. Abnormal events are usually contributed by a variety of complex factors and we leverage three prediction branches that are better at expressing and detecting different types of anomalies by concentrating on different levels of graph representations. By adjusting the weights of different branches, we can adjust the overall framework to fit different scenarios and improve the robustness of the method.

B. Network Architecture

1) Overall Network: The HSTGCNN model proposed in this paper consists of three major parts: a spatio-temporal graphical feature extractor, a future frame predictor, and an outlier arbiter. The spatio-temporal graphical feature extractor consists of a spatio-temporal graph convolutional neural network and performs spatio-temporal convolutional operations on the graph representations of all skeletons in the historical frames to extract features. The future frame predictor consists of a temporal convolutional network. Taking the graph representations of the human skeletons as input, the temporal convolutional network is expected to predict the future skeleton trajectories with convolutional operations. The outlier arbiter firstly feeds the outputs from the second part into multiple branches, and then weightedly sum up the predictions from all branches to obtain the anomaly score. Fig. 3. shows the structure of the model.

2) Local Network Structure:

- Construction of graph representations
  In Section III-A, we used the xOy coordinate system to represent the set S of all human skeleton trajectories in a video with t frames and then convert it into a graph structure as the input of the STGCNN. Firstly, construct a set of spatial graphs G_t to represent the graph structure of S. G_t is defined as:

$$G_t = (V(G_t), E(G_t)), \tag{8}$$

  where $V^g(G_t) = \{V_i(G_t) \mid \forall i \in \{1, \cdots, M\}\}$ and $V^l(G_t) = \{V_i(G_t) \mid \forall i \in \{1, \cdots, N\}\}$ denote two sets of vertices in a high-level graph and a low-level graph, respectively. The coordinates of geometric centers P and body joints f compose $V^g(G_t)$ and $V^l(G_t)$, respectively. $E(G_t)$ is a set of edges in the graph G_t, expressed as:

$$E(G_t) = \left\{e_t^{ij} \mid \forall i, j \in \mathbb{N}^*\right\}. \tag{9}$$

  If $V_i(G_t)$ and $V_j(G_t)$ are connected, $e_t^{ij} = 1$, otherwise, $e_t^{ij} = 0$. Besides, an adjacency matrix indicating the relative positions of nodes is integrated in the input to HSTGCNN. HSTGCNN uses a_t^{ij} to represent the distance between the i-th node and the j-th node. The value is determined by the following formula:

$$a_t^{ij} = \begin{cases} \dfrac{1}{\left(V_i(G_t) - V_j(G_t)\right)^2}, & \left(V_i(G_t) - V_j(G_t)\right)^2 \neq 0 \\ 0, & \left(V_i(G_t) - V_j(G_t)\right)^2 = 0, \end{cases} \tag{10}$$

  where $V^g(G_t) = \left\{V_{ij}(G_t) \mid \forall i, j \in \{1, \cdots, M\}\right\}$ and $V^l(G_t) = \left\{V_{ij}(G_t) \mid \forall i, j \in \{1, \cdots, N\}\right\}$ correspond to

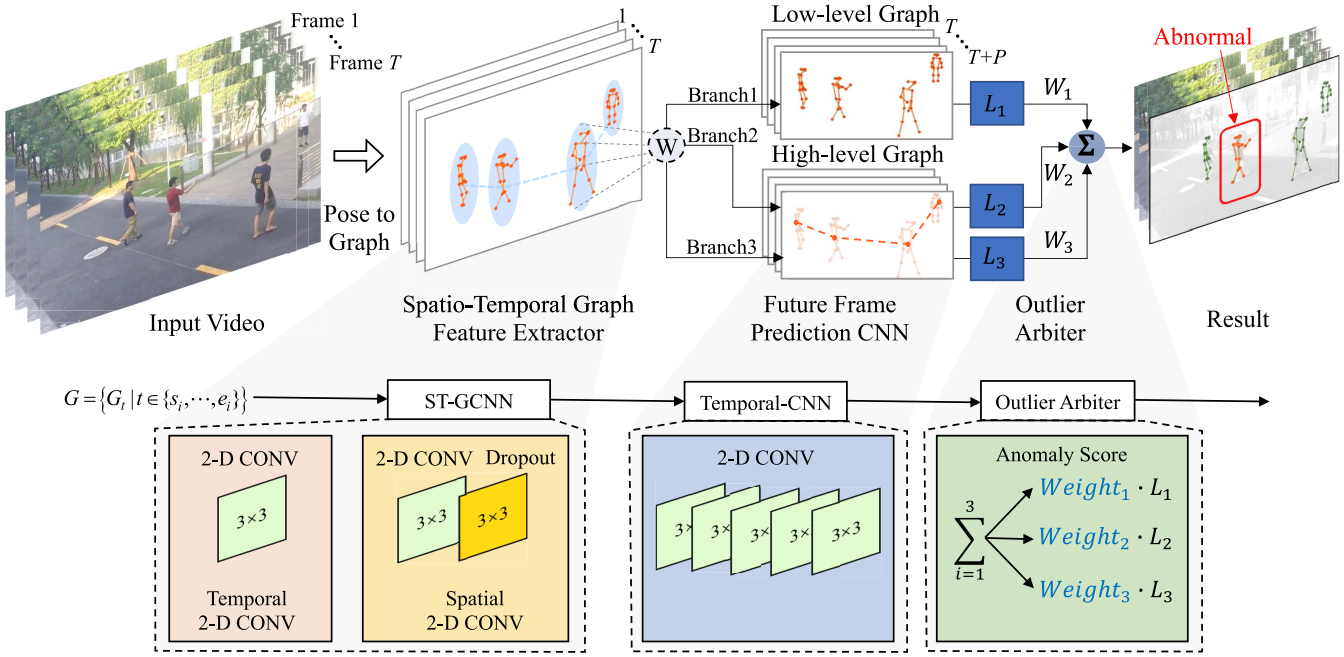

Fig. 3. Structure of the proposed framework. The HSTGCNN model consists of three major parts: a spatio-temporal graphical feature extractor (STGCNN), a future frame predictor (future trajectory prediction CNN), and an outlier arbiter (anomaly score).

high-level graph edges and low-level graph edges respectively, we calculate each e_t^{ij} to construct a weighted adjacency matrix A.

- STGCNN model layout
  In order to facilitate better convergence of the model, we normalize the adjacency matrix. The adjacency matrix A of T frames can be expressed as a set: $A_t = \{A_1, \cdots, A_T\}$, we symmetrically normalize each A_t, according to Eq. 11:

$$A_t = D^{-\frac{1}{2}} \hat{A}_t D^{-\frac{1}{2}}, \tag{11}$$

  where D is the diagonal nodal degree matrix of $\hat{A}_t$. The operation in the $(l+1)$-th layer can be expressed by the formula:

$$H^{(l+1)} = \sigma\left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(l)} W^{(l)}\right), \tag{12}$$

  where $W^{(l)}$ is the matrix of trainable parameters at layer l and σ is the nonlinear activation function.
- Future Frame Prediction CNN model layout
  STGCNN extracts spatio-temporal graph representations that serve the purpose of predicting future trajectories. Future Frame Prediction CNN operates directly on the temporal dimension of graph embeddings V_t and leverages the temporal clues for prediction. It is composed of five residual convolutional layers, it takes in four frames of input and outputs one frame with predicted human poses.

Using the spatio-temporal graphical feature extractor, we get the feature representations in the form of 4D tensors: (f, t, m, n), as detailed in Eq. 1 & 2. We exchange the first and second dimensions of the tensor to produce the input of the future frame predictor. The future frame predictor uses the residual convolutional network [58] as the backbone and predicts the coordinates of future samples. The difference between predicted trajectories and ground truth is the loss for training the model. The outlier arbiter does not require training, it is only responsible for calculating the anomaly score to judge the level of anomalies as will be discussed later in Eq. 14 to 17.

C. Model Training

1) Fixed Number of Frames as Input: Based on the above-mentioned approaches, we propose to track multiple individuals in multiple video frames. However, HSTGCNN needs to be trained on a sequence with a fixed length of T. We use a sliding window strategy to extract video clips with $T = 5$ frames. The reason for not selecting a too large T is due to some limitations. For instance, due to occlusions and other problems, long trajectories may contain inconsistencies in human appearances or movements. The determination of T is with reference to [1]. An appropriate T ensures stable training. During training, we propose to use the human skeletons in four consecutive video frames as the input of the reconstruction model to predict all human skeletons in the fifth video frame.

2) Model Configuration: This section focuses on the spatio-temporal graphical feature extractor and the future frame predictor. We use PReLU [59] as the activation function σ in all layers. According to Mohamed et al.'s [31] research, when the number of STGCNN layers increases, the performance decreases, so spatio-temporal graphical feature extractor includes one STGCNN layer and the future frame predictor has five Temporal-CNN layers.

3) Loss Function: In order to better reconstruct the trajectory of individuals in normal mode in unsupervised learning, we choose the mean square error (MSE) loss function to calculate the loss between the model output and the ground

truth. Such as the formula:

$$MSE = \frac{1}{\mathrm{T}} \sum_{t=1}^{T} \left(\hat{f}_t^{g+l} - f_t^{g+l} \right)^2 , \tag{13}$$

where $\hat{f}^{g+l}$ represents the global and local locations of prediction and f_t^{g+l} represents the ground truth.

D. Video Clustering

Under different scenarios, different strategies are adopted with different weights of branches. In this work, we divide training videos into different groups and use clustering methods to group videos with similar scenes. In implementations, we collect the number of individuals, average sizes of human bounding boxes and skeletons, optical flow fields and other information as the features of the scenarios in the video as the input of the K-means for initialization. Next, we determine a set of appropriate weight coefficients, (W_1, W_2, W_3), so as to minimize the loss of weightedly summing the three prediction branches in each group on the training set. For instance, for scenes with dense crowds, higher weights are assigned to high-level graph representations while in sparse scenes, the weights of low-level graph representations are increased.

E. Anomaly Detection

In order to calculate the anomaly score of each frame in the video of the test dataset, we propose the third part of the model: the outlier arbiter. Based on the hierarchical graph representation inference introduced in Section III-B, three independent branches are built to provide predictions on the level of anomalies. The outlier arbiter combines the branches to obtain the anomaly scores through the weighted summation, the process is divided into the following several steps:

1. In the same way as the training set, we take a sliding window with a stride of 1 and a window size of 5 frames on the test set. Every 4 frames are leveraged by the predictor to predict the fifth frame.
2. L_1 denotes the prediction error on the poses of each independent individual, which is computed with low-level graph representations in branch 1, according to Eq. 6 & 7:

$$L_1 = \frac{1}{T_\mathrm{e} - T_s} \sum_{t=T_s}^{T_e} \left(\hat{f}_t^{l} - f_t^{l} \right)^2 . \tag{14}$$

   L_2 is the error in jointly predicting the center points of multiple people. It is computed with high-level graph representations in branch 2, according to Eq. 5:

$$L_2 = \max \left[\left(\hat{f}_t^{g} - f_t^{g} \right)^2 \mid t \in \{T_s, \cdots, T_e\} \right]. \tag{15}$$

   L_3 is the prediction error on the motion vectors of multiple people, it is computed by high-level graph representations in branch 3, according to Eq. 5:

$$L_3 = \max \left[\left(\hat{f}_{t_1}^{g} - \hat{f}_{t_2}^{g} \right)^2 - \left(f_{t_1}^{g} - f_{t_2}^{g} \right)^2 \right] \quad t_1, t_2 \in \{T_s, \cdots, T_e\}\,; t_1 \neq t_2. \tag{16}$$

   Obviously, if there is only one individual in the frame, the score L_3 is 0. $\hat{f}_t$ represents the predicted locations and f_t represents the ground truth in Eq. 14 to 16.
3. Based on the calculation in step 2, assuming a video frame has a duration of T, We obtain anomaly scores on $T-4$ frames, the score on each frame is the summation of 3 terms. Here, the three branches independently characterize anomalies from different point-of-views. Finally, after assigning the same set of weights to each set of videos grouped by clustering, the overall anomaly score is obtained as is demonstrated in Eq. 17:

$$L = W_1 \cdot L_1 + W_2 \cdot L_2 + W_3 \cdot L_3 , \tag{17}$$

   where W represents the individual weight coefficient of each branch.

We use Algorithm 1 to specifically describe the anomaly score calculation process of the above algorithm.

Algorithm 1 Calculate Anomaly Score

Input: The high-level graph representations ${f_t}^g$, the low-level graph representations ${f_t}^l$, the weights of branches W.
Output: Final anomaly score L.

```
Initialize f_0^g, f_0^l and others;
for T_s = 0 ~ T − 5 do
    T_e = T_s + 5;
    for all t ∈ {T_s, ··· , T_e} do
        Calculate L_1, L_2, L_3 with Eq. 14, 15, and 16;
        Minimize W_1 · L_1 + W_2 · L_2 + W_3 · L_3 with respect
        to W_1, W_2, W_3 on training set;
        Share the weights of branches W in the same
        group of videos;
        Calculate anomaly score L with Eq. 17;
    end
end
```

IV. Experiment

A. Datasets and Pre-Processing

1) Datasets: To demonstrate the effectiveness of the proposed method, we conduct experiments on four public datasets: the UCSD Pedestrian dataset [6], the ShanghaiTech Campus dataset [7], the CUHK Avenue dataset [8] and the IITB-Corridor dataset [9]. The training set of these data sets contains only normal events, while the test set contains normal and abnormal events.

- UCSD Pedestrian dataset is acquired with a stationary camera mounted at an elevation and pedestrian walkways. UCSD includes two subsets: Ped1 and Ped2 which contain 7200 frames with 40 abnormal events and 2010 frames with 12 abnormal events, respectively. Videos are from the outdoor scene, recorded with a static camera at 10 fps. All other objects except for pedestrians are considered as irregularities.
- ShanghaiTech Campus dataset is considered to be one of the most comprehensive and realistic video anomaly detection data sets currently available. It contains

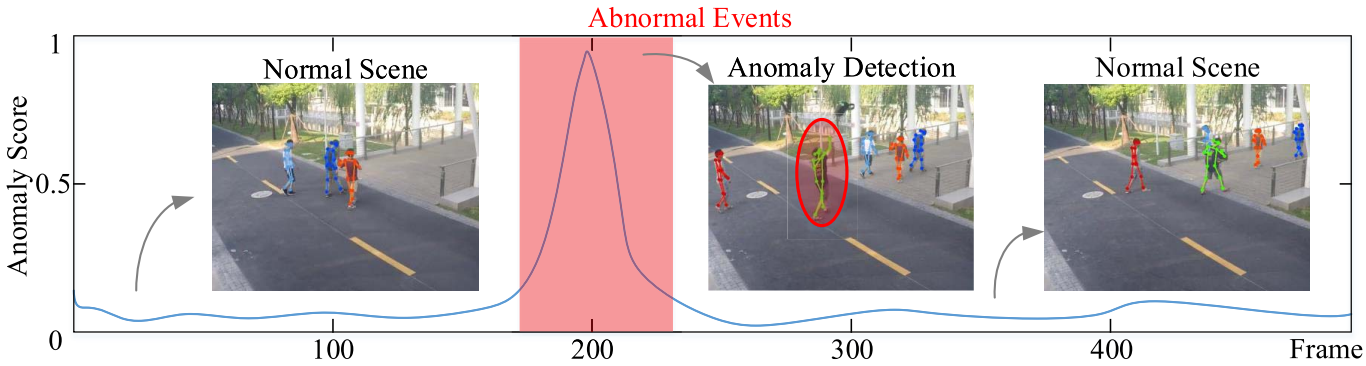

Fig. 4. Anomaly score graph for a testing video in the HR-ShanghaiTech dataset. The peak in anomaly scores corresponds to the frame with anomalies which are marked by red circles indicate frame-level ground truth of anomaly score.

330 training videos and 107 test videos with 130 abnormal events on the campus of Shanghai University of Science and Technology. A total of 13 different scenarios and various types of anomalies are included. Due to the complexity of abnormal semantics, it is extremely challenging for current methods.

- CUHK Avenue dataset is another representative dataset for video anomaly detection. It contains 16 and 21 video clips captured from a single camera for training and test, respectively. These videos were taken on Campus Avenue, a total of 30652 frames (15328 of training, 15324 of the test), mainly to detect pedestrians' abnormal movements, wrong movement directions, and the characteristics of abnormal objects.
- IITB-Corridor dataset is a large scale surveillance dataset with 483566 frames (301999 of training, 181567 of the test), which consists of group activities such as protest, chasing, fighting, sudden running as well as single person activities such as hiding face, loitering, unattended baggage, carrying a suspicious object and cycling (in a pedestrian area).

In order to facilitate comparison with other existing methods, we follow Morais et al.'s [2] strategy to manually filter out a set of video frames where the main anomalies are not related to humans, or where the individuals are not visible, or where objects can not be detected or tracked. The revised datasets are reserved as "Human-Related (HR) ShanghaiTech dataset" and "Human-Related (HR) Avenue dataset", respectively. Similarly, we adopt Jain et al.'s [60] strategy to filter IITB-Corridor to generate "Human-Related (HR) IITB-Corridor dataset"

2) Pose Tracking: The multi-person pose tracking algorithm is a crucial part of the framework. Firstly, the intensity of the input frame is normalized to the range of $[-1, 1]$, and each frame is adjusted to be with a resolution of 256×256 pixels. We use the object detection algorithm (YOLOv5x) [61] to locate each person. Next, the reID [62] algorithm is used to extract the features of each pedestrian, including the clues such as body shape and clothing. Finally, the similarity score calculated by the bounding box coordinates from adjacent frames, and the score calculated by the reID algorithm is used together with the Hungarian [63] algorithm to track people.

The occlusion or sudden abnormal behavior of pedestrians cause the tracking ID of the human to be missing in some video frames. Existing algorithms assign a new tracking ID after the previously occluded person re-appears. We propose a "track-back" module that uses the similarity score calculated by the reID [62] and motion vectors to judge whether to assign a new tracking ID or re-use original IDs to the identities in new frames. We use HRNet [57] to independently detect the skeleton in each video frame, by using the backbone of different sizes to balance the relationship between speed and accuracy, and explore the impact of accuracy on the effect of subsequent anomaly detection algorithms. Moreover, we utilize RAFT [64] to extract the optical flow of each frame as auxiliary information, which is suitable for low-resolution datasets.

3) Experimental Setting: We use a training batch size of 64 and the model was trained by using Stochastic Gradient Descent (SGD) with $\beta = 0.9$, with 20, 60, 30 and 80 epochs on UCSD [6], ShanghaiTech [7], CUHK Avenue [8], and IITB-Corridor [9], respectively. The initial learning rate is 0.5, the learning rate is adjusted according to a cosine annealing method [65]. We train the model for one time in each scene, finally, different weight files are generated. All models are trained in an end-to-end manner using PyTorch [66] with an Nvidia GTX 2070.

B. Evaluation

1) Comparisons With Existing Methods on Accuracy: We train HSTGCNN on the training dataset and obtain weights for different scenarios. Fig. 4. shows a curve describing the anomaly scores on different frames. Among them, for scenes with only normal behaviors, the anomaly score keeps low while for a frame with abnormal events, the anomaly score has a relatively high value.

In the literature of anomaly detection, a popular evaluation method is to use the metrics of the Area Under Curve of the Receiver Operating Characteristic (ROC AUC). A higher ROC AUC value indicates better anomaly detection performance. In this paper, we leverage frame-level ROC AUC for performance evaluation following the work [1].

Table I lists the comparison between the proposed HSTGCNN model and the latest state-of-the-art methods using ROC AUC. The ten methods are Frame-Pred [1], MPED-RNN [2], w/Mem [67], ST-GCAE [68], Multi-timescale [9], PoseCVAE [60], LSA [69], Ano-Graph [70],

TABLE I
COMPARISON OF ROC AUC BETWEEN HSTGCNN AND OTHER METHODS

Methods	UCSD ped1	UCSD ped2	HR-Avenue	Avenue	HR-ShanghaiTech	ShanghaiTech	HR-IITB-Corridor	IITB-Corridor
Frame-Pred [1]	83.10%	95.40%	-	84.90%	-	72.80%	-	64.65%
MPED-RNN [2]	-	-	86.30%	-	75.40%	73.40%	68.07%	64.27%
w/ Mem [67]	-	97.00%	-	**88.50%**	-	70.50%	-	-
ST-GCAE [68]	-	-	-	-	-	71.60%	-	-
Multi-timescale [9]	-	-	88.33%	82.85%	77.04%	76.03%	-	67.12%
PoseCVAE [60]	-	-	87.78%	82.10%	75.70%	74.90%	70.60%	67.34%
LSA [69]	-	95.40%	-	-	-	72.50%	-	-
Ano-Graph [70]	-	96.68%	-	86.20%	-	74.42%	-	-
AnomalyNet [71]	**83.50%**	94.90%	-	86.10%	-	-	-	-
Normal Graph [5]	-	-	-	87.30%	76.50%	74.10%	-	-
HSTGCNN	83.39%	**97.73%**	**88.65%**	87.51%	**83.40%**	**81.80%**	**73.92%**	**70.46%**

TABLE II
COMPARISON OF MODEL PARAMETERS AND INFERENCE SPEED BETWEEN HSTGCNN AND OTHER METHODS (HR-SHANGHAITECH DATASET)

Methods	Model parameters	Inference time (s)
Frame-Pred [1]	7.7M	0.234
YOLOv5x [61] + ST-GCAE [68]	87.8M + 793.4K	0.021 + 0.084
YOLOv5x [61] + HRNet-w48 [57] + MPED-RNN [2]	87.8M + 63.6M + 25.46K	0.021 + 0.063 + 0.025
YOLOv5x [61] + HRNet-w32 [57] + RAFT [64] + HSTGCNN	87.8M + 28.5M + 5.3M + 0.126K	**0.021 + 0.029 + 0.019 + 0.0013**
YOLOv5x [61] + HRNet-w48 [57] + RAFT [64] + HSTGCNN	87.8M + 63.6M + 5.3M + 0.126K	0.021 + 0.063 + 0.019 + 0.0013

AnomalyNet [71], and Normal Graph [5], some of them integrate a model focusing on appearance and motion with others dealing with the trajectories of human skeletons. From experiments, we can conclude that HSTGCNN outperforms the ten methods mentioned above on four public datasets, including Human-Related (HR) and original datasets. Although there are anomalies unrelated to humans in the video segment of the dataset, it still achieves the highest frame-level ROC AUC. As a result, compared with methods based on pixel reconstruction [1], [71] and [69], the model shows a stronger robustness. The noises of intensity-based features can be reduced by extracting the features describing human skeletons and motion instead of pixels. By levering a novel structure, our approach achieves advantages over RNNs and improves understanding of the global scenes in contrast to the coarse-grained STGCNN [5], [70]. As expected, anomalous events about humans can be correctly detected in different scenes in the ShanghaiTech dataset, as shown in Fig. 5. In addition, skeleton-based sequences and motion-related features are integrated by high-level graph representations to accommodate different resolution datasets. To conclude, effective structures, comprehensive feature representations involving high/low-level embeddings and adjustable weights according to scenes contribute to the significant improvement of the overall accuracy and robustness of the HSTGCNN method.

2) Runtime: In terms of computational burdens, as shown in Table II, the average detection time is measured for different network designs, and the model parameters and the amount of computation are compared. These models are implemented on Ubuntu systems with Nvidia GTX 2070 GPU. For the sake of fairness, the methods for comparison are with the same inputs and outputs. Our model for anomaly detection performs inference at a speed of 14.22 PFS and 9.59 FPS by using different sizes of HRNet for skeleton detection, respectively. The inference speed of other latest technologies

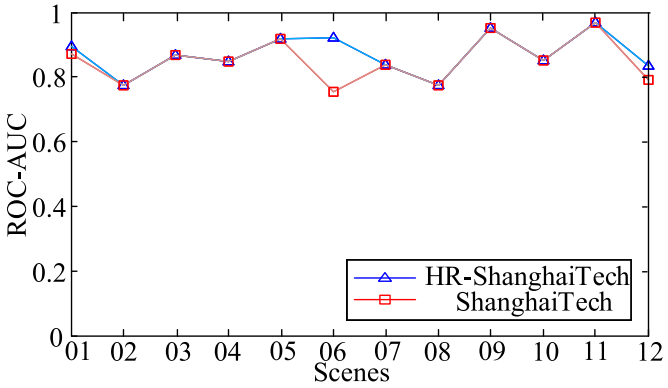

Fig. 5. Independent ROC AUC for each scene in the HR-ShanghaiTech dataset and ShanghaiTech dataset.

ST-GCAE [68] is 9.52 FPS, Frame-Pred [1] is 4.27 FPS and MPED-RNN [2] is 9.17 FPS. Compared with the four modules mentioned in Table II, the runtime of tracking module can be ignored.

As a result, the HSTGCNN network has a much higher efficiency than the above three methods. The design of the HSTGCNN network structure does not contain a large number of recursive structures (such as MPED-RNN [2]). At the same time, the number of learnable parameters in the proposed HSTGCNN is much less than in other models.

3) Qualitative Results: Fig. 6. shows the qualitative results of our model for future frame prediction and the other two latest methods on the ShanghaiTech dataset [7]. The w/ Mem [67] predicts anomalies based on reconstruction error. Although it can effectively detect abnormal regions, the result is noisy in that normal and abnormal regions are entangled. MPED-RNN [2] is an anomaly detection method based on the human skeleton. Even if it reduces the noise in the image and improves the detection accuracy, the anomaly regions are not accurately predicted in the video. Our hierarchical graphical

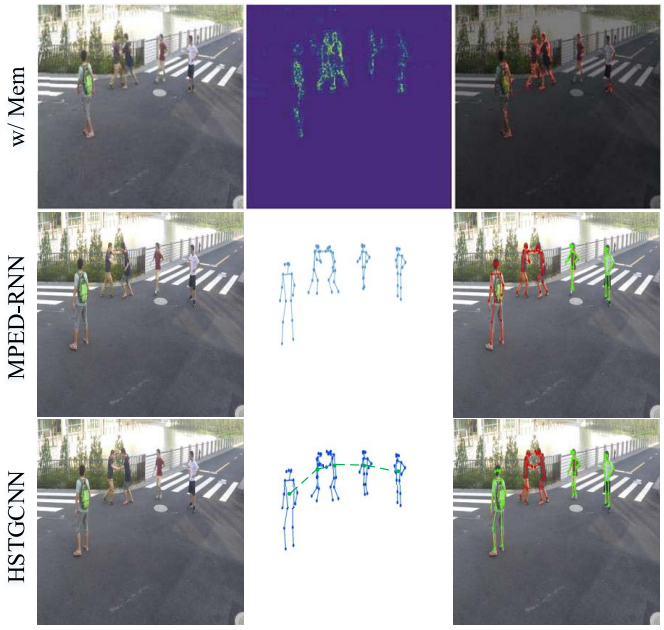

Fig. 6. Qualitative results for future frame prediction of (top to bottom): w/ Mem [67] model, MPED-RNN [2] model and our HSTGCNN model. Among them, input frames (left); prediction error (middle); abnormal regions (right). We can see that our model accurately localizes the regions with abnormal events.

TABLE III
INFLUENCE OF THREE BRANCHES COMBINATION

Methods	L_1	L_2	L_3	HR-ShanghaiTech	HR-IITB-Corridor
(a)	√	×	×	68.60%	68.97%
(b)	×	√	×	81.20%	67.52%
(c)	×	×	√	78.00%	69.75%
(d)	√	√	×	79.00%	68.67%
(e)	×	√	√	81.70%	69.98%
(f)	√	×	√	79.00%	70.16%
(g)	√	√	√	**83.40%**	**73.92%**

inference method can more accurately detect abnormal regions in video frames, while also ensuring a higher frame-level detection accuracy.

C. Ablation Study

1) Influence of Three Branches Combination: We analyze the network performance by altering the hyper-parameters of the outlier arbiter in the network. The outlier arbiter uses three prediction branches, (L_1, L_2, L_3), for hierarchical graphical inference, and these three branches represent three anomaly scores for each frame, the scores from different branches are weightedly summed to calculate the final anomaly score. In Section III-E, we explained in detail the meaning of the three prediction branches, which have a corresponding mapping relationship with high/low-level graph representations and affect the model detection strategy and provide different anomaly detection effects.

Table III shows the influence of hierarchical graph representations and combining branches. Experiments show that hierarchical graph representations are necessary, which covers more diverse abnormal events and improves the ability to handle abnormal semantics.

TABLE IV
INFLUENCE OF WEIGHT COEFFICIENTS IN DIFFERENT GROUPS ON ACCURACY (HR-SHANGHAITECH DATASET)

Number of K	W_1, W_2, W_3	Total loss during training	ROC AUC
3	$(0.2, 0.5, 0.3)$ $(0.1, 0.8, 0.1)$ $(0.2, 0.7, 0.1)$	41.2396	80.60%
6	$(0.3, 0.5, 0.2)$ $(0.1, 0.5, 0.4)$ …… $(0.6, 0.3, 0.1)$	37.3425	82.20%
12	$(0.6, 0.1, 0.3)$ $(0.1, 0.7, 0.2)$ …… $(0.1, 0.5, 0.4)$ $(0.1, 0.1, 0.8)$	**34.6117**	**83.40%**

2) Scene Clustering and Determination of Weight Coefficients: We divided training and test videos including 12 scenes into 3, 6, and 12 groups by clustering to explore the impact of the weight coefficients which are adapted to scene groups on the accuracy of the anomaly detection algorithm. Similar scenes correspond to the same group. In Section III-D, we clarified in detail the advantage of video clustering, which adjusts the weight coefficients of high/low level graph representations according to dense or sparse scenes, respectively, to enhance the robustness of the method.

Table IV reports the influence of the weight coefficients in different groups on the accuracy of anomaly detection algorithm. In each group, we iterate repeatedly to select the best set of weight coefficients that minimize total output loss without labels in the training set. As can be seen, the lowest total loss and highest ROC AUC are usually achieved for $K = 12$, that is, each scene corresponds to a set of weight coefficients and we use that value through all our experiments.

Fig. 7. shows that with the usage of different strategies with different weights of branches under different scenarios among the datasets. Due to the size of pedestrians becomes smaller in dense scenes, the weight coefficient corresponding to high-level graph representations increases. On the contrary, the weight coefficient corresponding to low-level graph representations increases in sparse scenes including large pedestrians. The accuracy of our approach has been improved, which proves the weighted combinations contribute to an understanding of scenes.

3) Influence of Using Different Tracking Modules: In order to demonstrate the effectiveness of different components of the proposed framework, we divide the framework into a tracking module (including the different sizes of skeleton detection models) and a module for anomaly detection. We combine our tracking module with the anomaly detection module in Morais et al.'s [2] by using different sizes of HRNet [57] for skeleton detection. Finally, the above methods are compared on the HR-ShanghaiTech dataset [7] and the HR-IITB-Corridor dataset [9]. As shown in Table V, the improvements in the accuracy of our method are contributed by larger HRNet for skeleton detection and our proposed HSTGCNN for anomaly detection.

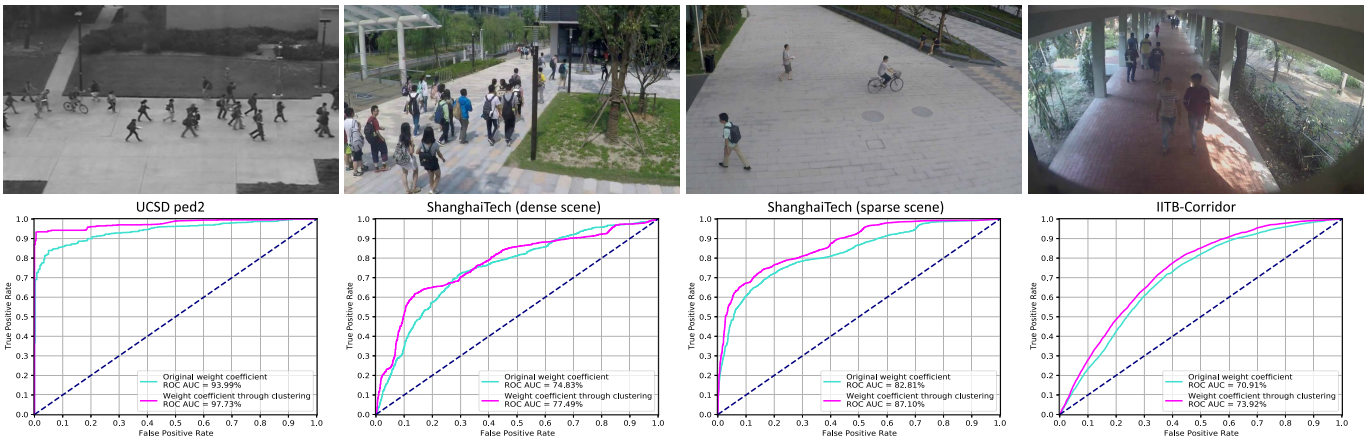

Fig. 7. Comparing the original weight coefficient and the weight coefficient through clustering, the ROC AUC curve shows significant differences in the crowd or sparse scenarios on various resolutions datasets.

TABLE V

ROC AUC PERFORMANCE OF TWO METHODS WITH DIFFERENT PRE-PROCESSING WORK

Skeleton Trajectory Tracking	Anomaly Detection Modules	HR-ShanghaiTech	HR-IITB-Corridor
Morais et al.'s [2]	MPED-RNN [2]	75.40%	68.07%
YOLOv5x [61] + HRNet-w32 [57]	MPED-RNN [2]	74.47%	66.73%
YOLOv5x [61] + HRNet-w48 [57]	MPED-RNN [2]	75.00%	69.05%
YOLOv5x [61] + HRNet-w32 [57]	HSTGCNN	81.31%	70.08%
YOLOv5x [61] + HRNet-w48 [57]	HSTGCNN	**83.40%**	**73.92%**

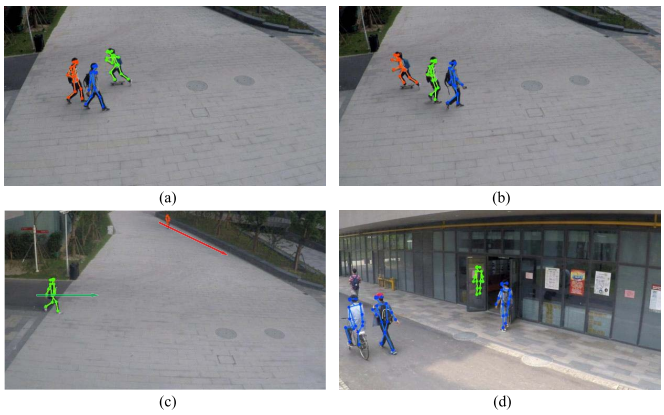

Fig. 8. Some misrepresentations that may occur in pose tracking.

D. Error Analysis

Although the proposed method based on human skeleton detection and tracking effectively improves the detection accuracy of abnormal events, it strongly depends on the performance of the model for detecting human skeletons. In some cases in the test set, the changes in the resolution or contrast of the picture cause the anomaly detection algorithm to fail. When multiple individuals are severely occluded by each other, the tracking algorithm produces wrong IDs, because the predictions on anomaly scores are negatively influenced. As a result, the problems that occur in pose estimation and tracking influence the performance of anomaly detection.

As shown in Fig. 8.(a) & (b), the pedestrian's exchange IDs due to occlusion. At the same time, the partial occlusion also leads to dramatic changes in poses which are judged as anomalies. In Fig. 8.(c), due to the variances in scales, the moving speed of remote bicycles (small) is the same as that of a human in the vicinity (large), this influences the determination of anomalies based on moving speed. Even if we normalized the sizes of pedestrians, some false positives sometimes can not be avoided. In the case of the small-scale pedestrians of low resolutions, motion consistency among consecutive frames will be utilized for improvement. In Fig. 8.(d), the algorithm erroneously detects the contour from the glass reflection, this introduces some interference for the subsequent tracking.

V. CONCLUSION

In this paper, we propose to use skeleton-based sequences and motion-related features to detect human-related anomalies without relying on any annotations about abnormal events. Specifically, we use a spatio-temporal convolutional network as our graphical feature extractor which is superior to other existing models in accuracy, memory consumption, and efficiency. The HSTGCNN we proposed integrates high-level graph representations with low-level graph representations. Low-level graph structure focuses on encoding the spatial and temporal embeddings of human body joints in high-resolution videos. High-level graph structure leverages the speed and directions of individuals and the interactions among multiple identities to describe abnormality in low-resolution videos. Meanwhile, high-level graph structure distinguishes the dense scenes including small people from sparse scenes including large ones by bounding boxes, skeletons, and optical flow. The weighted combinations of multiple branches that are better at handling different scenes achieve state-of-the-art performance. Additionally, a clustering method is leveraged to group scenes.

Future research directions include the detection of human abnormal events with the assistance of semantic information and the improvement of pose tracking approaches.

REFERENCES

[1] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–A new baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6536–6545.

[2] R. Morais, V. Le, T. Tran, B. Saha, M. Mansour, and S. Venkatesh, "Learning regularity in skeleton trajectories for anomaly detection in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11996–12004.

[3] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.

[4] J. Martinez, M. J. Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2891–2900.

[5] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, Jul. 2021.

[6] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 1, pp. 18–32, Jan. 2014.

[7] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked RNN framework," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 341–349.

[8] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2720–2727.

[9] R. Rodrigues, N. Bhargava, R. Velmurugan, and S. Chaudhuri, "Multi-timescale trajectory prediction for abnormal human activity detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2626–2634.

[10] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 483–499.

[11] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[12] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[13] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*.

[15] L. Pishchulin et al., "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4929–4937.

[16] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 34–50.

[17] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[18] N. Zouba, B. Boulay, F. Bremond, and M. Thonnat, "Monitoring activities of daily living (ADLs) of elderly based on 3D key human postures," in *Proc. Int. Workshop Cognit. Vis.* Berlin, Germany: Springer, 2008, pp. 37–50.

[19] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 961–971.

[20] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.

[21] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4346–4354.

[22] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," 2017, *arXiv:1704.05831*.

[23] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[24] N. Zheng, J. Wen, R. Liu, L. Long, J. Dai, and Z. Gong, "Unsupervised representation learning with long-term dynamics for skeleton based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[25] X. Liu, Y. Li, and Q. Wang, "Multi-view hierarchical bidirectional recurrent neural network for depth video sequence based action recognition," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 10, Oct. 2018, Art. no. 1850033.

[26] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 816–833.

[27] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[28] W. Zhu et al., "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[29] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1746–1753.

[30] R. Zeng et al., "Graph convolutional networks for temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7094–7103.

[31] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14424–14432.

[32] M. Wang, B. Ni, and X. Yang, "Learning multi-view interactional skeleton graph for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Oct. 21, 2020, doi: 10.1109/TPAMI.2020.3032738.

[33] C. Gu et al., "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.

[34] F. Zhao, S. Liao, G.-S. Xie, J. Zhao, K. Zhang, and L. Shao, "Unsupervised domain adaptation with noise resistible mutual-training for person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 526–544.

[35] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2017.

[36] B. Mohammadi, M. Fathy, and M. Sabokrou, "Image/Video deep anomaly detection: A survey," 2021, *arXiv:2103.01739*.

[37] D.-G. Lee, H.-I. Suk, S.-K. Park, and S.-W. Lee, "Motion influence map for unusual human activity detection and localization in crowded scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 10, pp. 1612–1623, Oct. 2015.

[38] R. Leyva, V. Sanchez, and C.-T. Li, "Fast detection of abnormal events in videos with binary features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1318–1322.

[39] J. Yin, Q. Yang, and J. J. Pan, "Sensor-based abnormal human-activity detection," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 8, pp. 1082–1090, Aug. 2008.

[40] E. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?: Recovering low-rank matrices from sparse errors," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop*, Oct. 2010, pp. 201–204.

[41] L. Xiong, X. Chen, and J. Schneider, "Direct robust matrix factorizatoin for anomaly detection," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 844–853.

[42] M. Debruyne and T. Verdonck, "Robust kernel principal component analysis and classification," *Neural Comput.*, vol. 4, no. 2, pp. 151–167, 2010.

[43] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. CVPR*, Jun. 2011, pp. 3313–3320.

[44] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proc. CVPR*, Jun. 2011, pp. 3449–3456.

[45] X. Zhu, J. Liu, J. Wang, C. Li, and H. Lu, "Sparse representation for robust abnormality detection in crowded scenes," *Pattern Recognit.*, vol. 47, no. 5, pp. 1791–1799, 2014.

[46] Y. Yuan, Y. Feng, and X. Lu, "Structured dictionary learning for abnormal event detection in crowded scenes," *Pattern Recognit.*, vol. 73, pp. 99–110, Jan. 2018.
[47] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *Proc. Int. Symp. Neural Netw.* Sapporo, Japan: Springer, 2017, pp. 189–196.
[48] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3619–3627.
[49] W. Chu, H. Xue, and C. Yao, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 246–255, Jan. 2019.
[50] T. Wang et al., "Generative neural networks for anomaly detection in crowded scenes," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 5, pp. 1390–1399, May 2019.
[51] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3379–3388.
[52] S. Lee, H. G. Kim, and Y. M. Ro, "STAN: Spatio-temporal adversarial networks for abnormal event detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1323–1327.
[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. (ICIP)*, Sep. 2017, pp. 1577–1581.
[54] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-based anomaly detection," 2018, *arXiv:1802.06222*.
[55] D. Huang, P. Chen, R. Zeng, Q. Du, M. Tan, and C. Gan, "Location-aware graph convolutional networks for video question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, 2020, pp. 11021–11028.
[56] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.
[57] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5693–5703.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[60] Y. Jain, A. K. Sharma, R. Velmurugan, and B. Banerjee, "PoseCVAE: Anomalous human activity detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2927–2934.
[61] G. Jocher, K. Nishimura, T. Mineeva, and R. Vilariño. (2020). *YOLOv5*. Accessed: Jun. 20, 2021. [Online]. Available: https://github.com/ultralytics/yolov5
[62] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.
[63] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logist. Quart.*, Vol. 2, nos. 1–2, pp. 83–97, May 1955.
[64] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 402–419.
[65] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[66] A. Paszke et al., "Automatic differentiation in PyTorch," in *Proc. Adv. Neural Inf. Process. Syst. Workshop*, Los Angeles, CA, USA, 2017, pp. 1–2.
[67] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.
[68] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10539–10547.
[69] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 481–490.
[70] M. Pourreza, M. Salehi, and M. Sabokrou, "Ano-graph: Learning normal scene contextual graphs to detect video anomalies," 2021, *arXiv:2103.10502*.
[71] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

Xianlin Zeng received the B.S. degree in electronic and information engineering from Xidian University, Xi An, China. He is currently pursuing the Ph.D. degree in information and communication engineering with Beihang University. His research interests include computer vision, machine learning, and anomaly detection.

Yalong Jiang received the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. Since April 2020, he has been with Beihang University, where he is currently an Assistant Professor at the Institute of Unmanned System. His research interests include pattern recognition, computer vision, and machine learning.

Wenrui Ding received the Ph.D. degree in electrical and information engineering from Beihang University. She is currently in charge of information transmission and processing data link at the Institute of Unmanned System, Beihang University. Her research interests include the command and control of aerial vehicles, image processing, and pattern recognition.

Hongguang Li received the Ph.D. degree in aerospace science and technology from Beihang University, Beijing, China. He currently works at the Institute of Unmanned System, Beihang University. His research interests include intelligent image processing and end-side applications of unmanned systems, including: optical image restoration and enhancement, object detection and tracking, geometric correction and target positioning, landing visual guidance, image stitching and fusion, and AI end-side hardware systems design and application.

Yafeng Hao received the B.S. degree in electronic and information engineering from the Harbin Institute of Technology, Harbin, China. He currently works at the 54th Research Institute, China Electronics Technology Group Corporation (CETC), China. His research interests include intelligent image processing, pattern recognition, computer vision, and machine learning.

Zifeng Qiu received the master's degree in electronic and information engineering from Beijing Jiaotong University, Beijing, China. He currently works at the Laboratory of Aerospace Information Applications, China Electronics Technology Group Corporation (CETC), China. His research interests include image processing and pattern recognition.