applied sciences

MDPI

Article

A Self-Attention Augmented Graph Convolutional Clustering Networks for Skeleton-Based Video Anomaly Behavior Detection

Chengming Liu [1], Ronghua Fu [1], Yinghao Li [1], Yufei Gao [1], Lei Shi [1,*] and Weiwei Li [2]

[1] School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450002, China; cmliu@zzu.edu.cn (C.L.); furh1996@163.com (R.F.); lyh@zzu.edu.cn (Y.L.); yfgao@zzu.edu.cn (Y.G.)
[2] Principal's Office, Zhengzhou College of Finance and Economics, Zhengzhou 450044, China; yb@zzife.edu.cn
* Correspondence: shilei@zzu.edu.cn

Abstract: In this paper, we propose a new method for detecting abnormal human behavior based on skeleton features using self-attention augment graph convolution. The skeleton data have been proved to be robust to the complex background, illumination changes, and dynamic camera scenes and are naturally constructed as a graph in non-Euclidean space. Particularly, the establishment of spatial temporal graph convolutional networks (ST-GCN) can effectively learn the spatio-temporal relationships of Non-Euclidean Structure Data. However, it only operates on local neighborhood nodes and thereby lacks global information. We propose a novel spatial temporal self-attention augmented graph convolutional networks (SAA-Graph) by combining improved spatial graph convolution operator with a modified transformer self-attention operator to capture both local and global information of the joints. The spatial self-attention augmented module is used to understand the intra-frame relationships between human body parts. As far as we know, we are the first group to utilize self-attention for video anomaly detection tasks by enhancing spatial temporal graph convolution. Moreover, to validate the proposed model, we performed extensive experiments on two large-scale publicly standard datasets (i.e., ShanghaiTech Campus and CUHK Avenue datasets) which reveal the state-of-art performance for our proposed approach when compared to existing skeleton-based methods and graph convolution methods.

Keywords: video anomaly detections; skeleton; self-attention; graph convolutional networks

Citation: Liu, C.; Fu, R.; Li, Y.; Gao, Y.; Shi, L.; Li, W. A Self-Attention Augmented Graph Convolutional Clustering Networks for Skeleton-Based Video Anomaly Behavior Detection. *Appl. Sci.* **2022**, *12*, 4. https://doi.org/10.3390/app12010004

Academic Editors: Antonio Fernández-Caballero, Hugo Pedro Proença and Byung-Gyu Kim

Received: 21 November 2021
Accepted: 14 December 2021
Published: 21 December 2021

1. Introduction

Video anomaly detection is a highly challenging task in unsupervised video analysis. In recent years, surveillance video anomaly detection has gained widespread attention owing to its applications in public security, social security management, and the rising trends in deep learning and computer vision. Inherently, the abnormal events are also complex in nature due to various reasons such as messy background/objects, motion in the scene, etc. Therefore, the complexity of the abnormal events creates a bottleneck issue in the detection of such events from real-world video data. Additionally, handling and modeling of video data itself are difficult because of its high dimensionality, noise, and a diversity of events and interactions involved. So far, many efforts have been reported in literatures that provide in-depth studies on video anomaly detection by mainly focusing on appearance features, depth features, optical flow modeling, etc., but very less attention has been paid to skeleton-based video anomaly detection models. Likewise, we explicitly use a common structure of surveillance video, i.e., people and objects moving on a static background where most abnormal phenomena are caused by the humans. However, most of these models are primarily based on the image level and instead of focusing on normal pattern modeling of humans, emphasize more on background hence increasing the burden on background model. Therefore, to mitigate the above stated issues, we employ

skeleton features and take advantage of their compactness, strong structure, semantically rich properties, and strong description of human behavior and movement. In this way, the analysis can be free from any interference caused by the factors such as illumination and busy background.

Nowadays, graph convolutional networks (GCN) are listed among the most popular methods available for analyzing Non-Euclidean Structure Data. As an effective representation of Non-Euclidean Structure Data, they can effectively capture spatial (intra-frame) and temporal (inter-frame) information. While referring to skeleton-based action recognition, Yan et al. [1] proposed the spatial temporal graph convolutional networks (ST-GCN), which first apply GCN to model skeleton data. The ST-GCN model has been proven to perform well on skeleton data [2–4], but as spatiotemporal graph convolution operation only operates on a local neighborhood node and is restricted by the size of the convolution kernel, it lacks the global information. Moreover, the correlation between body joints in the human skeleton that are not directly connected are also underestimated, e.g., the left hand and right foot. Transformer self-attention [5] was originally applied in natural language processing tasks to encode the short-distance and long-distance correlations between the words in sentences. Likewise, considering the sequential nature and hierarchical structure of the human skeleton sequences, this mechanism can be extended to the skeleton data. Self-attention can resolve the major shortcoming of ST-GCN (i.e., it can only capture the local features of the spatial dimension) because of its flexibility in dealing with long dependencies. Recently, the self-attention method is used in one of the works to solve the locality of the convolution operator by capturing the global context of pixels in the image [6]. The proposed novel spatial temporal self-attention augmented graph convolutional network (SAA-Graph) contains a new graph convolution operator by combining improved spatial graph convolution operator with a modified transformer self-attention operator to capture both local and global information of the joints. The improved spatial graph convolution operator uses a data-driven approach to improve the flexibility of the model building graphs and brings in more versatility to align with various data samples. Our work uses self-attention mechanism on skeleton data to enhance the graph convolution. We capture the information of local and global joints by combining the operator of the improved spatial graph convolution with the modified transformer self-attention operator. Spatiotemporal graph convolution operation only operates on a local neighborhood node and is restricted by the size of the convolution kernel, it lacks the global information. Therefore, the autoencoder constructed with spatiotemporal graph convolution also lacks global information. We use self-attention to solve the locality of the graph convolution operator by capturing the global information in the skeleton data.

Specifically, the extracted spatiotemporal graph of skeleton features is encoded to generate a latent vector using the encoder part of a spatial temporal self-attention augmented graph convolutional autoencoder (SAA-STGCAE). The deep embedded clustering layer is used to softly assign the latent vector to the clusters. We use the Dirichlet process mixture model to measure their distribution. We can obtain the normality score for each sample and determine whether the action should be classified as normal or not. An overview of proposed method can be viewed in Figure 1.

The key contributions of this work are summarized in this paper as follows: (1) We propose a novel spatial temporal self-attention augmented graph convolutional clustering networks for skeleton-based video anomaly detection tasks by employing the spatial temporal self-attention augmented graph convolutional autoencoder to extract the relevant features and embedded clustering; (2) We design a new spatial self-attention enhancement graph convolution operator to understand the intra-frame interaction between different body parts and capture the local and global features of a skeleton in the frame; (3) Our model achieves state-of-the-art AUC of 0.789 for the ShanghaiTech Campus anomaly detection datasets and also exhibits excellent performance metrics for CUHK Avenue datasets.

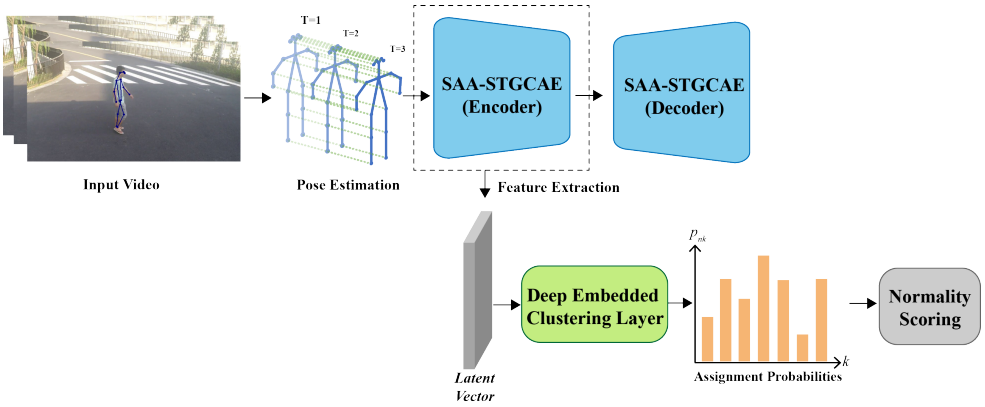

Figure 1. Framework Diagram: First, we perform pose estimation algorithm to extract skeletons for each frame in each video. The extracted pose of skeleton is encoded to generate a latent vector using the encoder part of a spatial temporal self-attention augmented graph convolutional autoencoder. The deep embedded clustering layer is used to softly assign the latent vector to the clusters, and P_{nk} represents the probability of the sample being assigned to the cluster k.

2. Related Work

2.1. Video Anomaly Detection

Video anomaly detection is defined as a way to find abnormal patterns or actions in the data. These abnormalities are defined as infrequent or rare events. Traditional methods for abnormal event detection that extract and analyze the hand-crafted low-level visual features are unable to characterize the more complex behaviors. Additionally, the extracted features by such methods are relatively single, which demonstrates the fact that the generalization ability of hand-crafted features is usually weak and is not robust to crowd scenes. For instance, trajectory [7,8] is used to describe the trajectory of moving objects. Similarly, Histogram of Oriented Gradient (HOG) [9] and Histogram of Flow (HOF) [10] can characterize the shape and contour information of the human body in a static image. Accordingly, optical flow [11] can describe the changes in the gray value of pixels between adjacent frames and is often used to characterize the motion information. Zhang et al. [12] associated optical flows to capture short-term trajectories between multiple frames and described short-term trajectories by histogram-based shape descriptor. However, the mentioned methods revealed only a suboptimal performance when subjected to complex surveillance scenarios and large-scale video anomaly detection datasets.

Recently, various works have used deep learning-based models to address the problem of video anomaly detection. Such models can be roughly categorized into reconstructive models, predictive models, and generative models. The reconstruction model uses the difference between reconstructed image and the original image as a basis for scoring and positioning of anomaly detection, and often relies on autoencoders [13,14]. The prediction model, on the other hand, utilizes the recurrent neural networks [15–17] or 3D convolutions [14] and emphasizes on addition of prediction and generation of future frames based on the original reconstruction to calculate the loss. Finally, the generative models primarily use the variational autoencoders (VAEs) or GANs to reconstruct, predictor model the distribution of the data. The early anomaly detection work of Leo et al. [18] has been used for human activity recognition in wide-area automatic visual surveillance. A method proposed by Liu et al. [19] uses a future frame prediction model by combining U-Net and Beyond-MSE. Wu et al. [20] proposed a Fast Sparse Coding Network based on High-level Features to discriminate spatio-temporal fusion features for video anomaly detection to achieve higher accuracy. In another work, Morais et al. [21] adhered two RNN branches together to form global and local features, using a message-passing encoder–decoder RNN architecture. The work by Luo et al. [22] is the first one which applies graph convolutional networks on skeleton-based video anomaly detection to analyze the graph connection of human joints. Progressively, Markovitz et al. [23] proposed an approach to use embedded

pose graphs and a Dirichlet process mixture for video anomaly detection with a new coarse-grained setting for exploring broader aspects of video anomaly detection.

2.2. Skeleton-Based Action Recognition

Most of the conventional techniques for skeleton-based action recognition generally rely on hand-crafted features to model the human body [24–26]. However, it is evident from the literature that hand-crafted features can only perform well on some certain types of datasets [27], which further illustrates the fact that the hand-crafted features are extracted from one data set cannot be always transferred to other data set. Moreover, deep learning has revolutionized the activity recognition by proposing techniques which can directly improve the robustness through data-driven approaches to achieve unprecedented performance metrics, where the most widely used models are RNNs and CNNs.

The RNN-based method is suitable for processing time series data due to its unique structure while skeleton sequences are natural time series of joint coordinate positions, but its spatial modeling ability is weak. Alternatively, many CNN-based researches encode the skeleton joints to multiple 2D Pseudo images to learn useful features [28,29]. However, existing CNN-based models largely fail to capture the various aspects of a skeleton sequence. Banerjee et al. [30] extracted four feature representations from the angle information and kinematics information of human movements, which then captured the complementary features of key joint sequences. Even so, neither RNNs nor CNNs can fully represent the structure of skeletal data because skeletal data are naturally embedded as graphs rather than vector sequences or two-dimensional grids.

Lately, GNN-based methods have been proposed which demonstrate a better performance by considering the fact that human skeleton data is a natural topological graph data structure (joints and bones can be treated as vertices and edges, respectively) rather than images and sequences vector. In order to retain the skeleton spatial information and improve the feature generalization ability, Yan et al. [1] proposed the ST-GCN to directly model the human skeleton data as the spatiotemporal graph structure by realizing the automatic extraction of robust spatiotemporal features from human skeleton data. It has strong expression and generalization capabilities, thus achieves better performance than previously reported methods. Inspired by this work, we used an improved ST-GCN block to construct a spatial temporal self-attention graph convolutional autoencoder, named SAA-STGCAE. We encoded to generate a latent vector using the encoder part of SA-STGCAE.

2.3. Transformer

The Transformer was originally proposed for natural language processing. It uses the attention mechanism to achieve parallel capturing of sequence dependencies and to process tokens at each position of the sequence simultaneously. The transformer follows and encoder–decoder structure and only relies on multi-head self-attention [5]. Recently, the self-attention mechanism has also been implemented for visual tasks [6] to enhance the standard convolution. Likewise, Our work uses self-attention mechanism on skeleton data to enhance the graph convolution.

2.4. Graph Convolutional Neural Networks

The implementation methods for graph convolutional neural networks (GCN) are mainly divided into two categories: (1) Spectral-based method and (2) Spatial-based method. Spectral-based method uses graph Fourier transform to convert the graph data into frequency domain data and then performs the calculation by exploiting the fundamental property of time domain convolution being equivalent to frequency domain multiplication. On the other hand, Spatial-based methods construct a convolution kernel directly in the spatial domain for feature extraction. In this work, we adopt the spatial-based graph convolutional neural network (GCN) method to extract features from structured graph data composed of human skeleton sequences.

3. Proposed Method

We propose a framework called SAA-STGCN for skeleton-based anomaly detection. The overall framework diagram of proposed method is illustrated in Figure 1. The suggested method focuses on human behavior detection while searching for anomaly detection. First, we directly perform the pose estimation algorithm to extract the human skeletons in each frame of the input video to generate spatiotemporal graphs (Section 3.1). This step makes the algorithm robust to complex backgrounds, lighting changes, human scales, and dynamic camera views. Next, we use the encoder part of SAA-STGCAE as a feature extractor to embed data and generate latent vectors (Sections 3.2 and 3.3). The deep embedding clustering layer (Section 3.4) is used to softly assign latent vectors to the clusters, and then each sample is represented by the probability that it is assigned to k cluster. Later on, we use the Dirichlet process mixture model (Section 3.5) to evaluate a set of distribution parameters in the estimation stage and uses the fitted model to provide a score for each embedding sample. The normality score provided by the model is used to determine whether the action is normal or not.

3.1. Spatiotemporal Graph Connection Configuration for Skeleton

The original skeleton data that can be obtained from pretrained video pose estimation algorithms or motion capture devices are provided as a sequence of vectors. We define N as the number of joints in skeleton and T as the total number of frames. For each person, a spatiotemporal graph is established as $G = (V, E)$, where $V = \{v_{tn} \mid t = 1, 2, \cdots, T; n = 1, 2, \cdots, N\}$ is the set of all the joint nodes as the vertices of the graph, and E represents the set of all the edges describing natural connections in the human body structure and time as the edge of the graph. Further more, E consists of two subsets E_s and E_t, where $E_s = \{(v_{si}, v_{sj}) \mid s = 1, 2, \cdots, T; i, j = 1, 2, \cdots, N\}$ represents the connection of any pair of joints (i, j) in each frame t. $E_t = \{(v_{tn}, v_{(t+1)n}) \mid t = 1, 2, \cdots, T; n = 1, 2, \cdots, N\}$ represents the connection between each frame along the continuous time. Figure 2a shows an example of the constructed spatiotemporal graph, where the joints are represented as vertices and their natural connections in the human body are represented as spatial edges (the blue lines in the Figure 2a) and the corresponding between two adjacent frames are connected as temporal edges (the green lines in the Figure 2a).

We adopt the spatial configuration partition [1] to divide the neighborhood of a node into three subsets according to graph distance. First, the center of gravity is determined as the average coordinate of all joints of the skeleton in the frame, then the first subset is the root node itself (red node in Figure 2b), the second subset is the neighbor nodes closer to the center of gravity than the root node (green node in Figure 2b), and the third subset is adjacent nodes away from the center of gravity (blue node in Figure 2b).

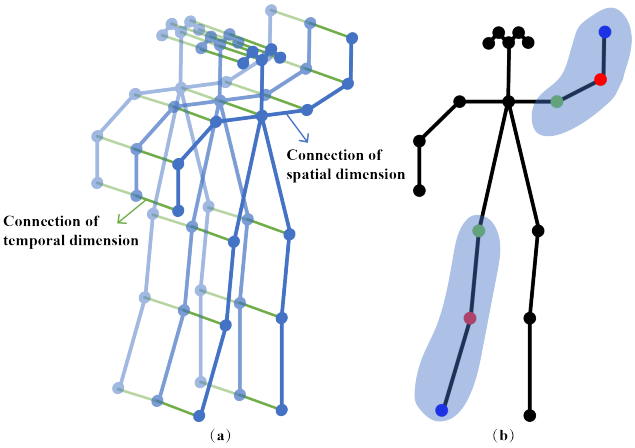

Figure 2. (**a**) The description of the spatiotemporal graph follows ST-GCN [1]; (**b**) The configuration of spatial configuration partitioning. The three colors represent three different subsets.

3.2. Feature Extraction

The proposed SAA-STGCAE uses a spatial self-attention augmented graph convolution module (SAA-Graph) presented next and the temporal convolution module (TCN) to embed the spatiotemporal graphs as shown in Figure 3. We employ the same temporal convolution module as ST-GCN and execute a $1 \times K_t$ convolution on the feature map obtained from the spatial dimension, where K_t is the kernel size in the time dimension. Then, we use encoder part of SAA-STGCAE to embed the extracted skeleton pose into the spatiotemporal graph to generate latent vectors for clustering branch.

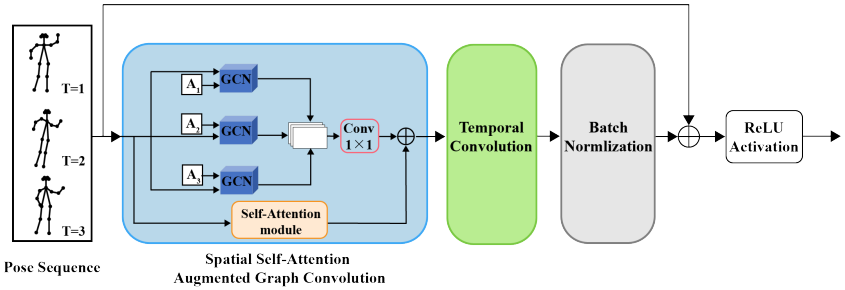

Figure 3. Spatial temporal self-attention augmented graph convolutional block. It is internally composed of a spatial self-attention augmented graph convolution (SAA-Graph, as shown in Figure 4) followed by a temporal convolution (TCN) [1] and batch normalization.

3.3. Spatial Self-Attention Augmented Graph Convolution

We propose a new graph convolution operator called Spatial Self-Attention Augmented Graph Convolution (SAA-Graph), which is based on the improved ST-GCN block and uses the Self-Attention module to enhance spatial graph convolution, as shown in Figure 4.

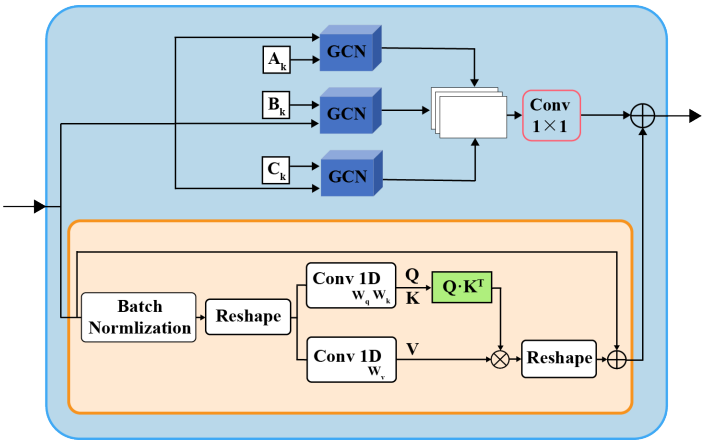

Figure 4. Spatial self-attention augmented graph convolution combines an improved ST-GCN block and a self-attention module to enhance spatial graph convolution.

3.3.1. Spatial Graph Convolution

For the spatial dimension, we use adjacency matrices of three types: static adjacency matrices (A_1), globally-learned adjacency matrices (A_2), and adaptive adjacency matrices (A_3). A_1 is a $N \times N$ hard-coded adjacency matrix of graph representing the physical structure of the human intra-body connections, A_2 is also an $N \times N$ adjacency matrix, which is learned by initializing a fully-connected graph according completely to the training data. The matrix and the parameters of the model are optimized together during training process. The matrix element can be any value, which can not only indicate whether there is a connection between two joints, but also the strength of the connection. A_3 is learned an adaptive graph for each sample to represent the strength of the connection between

two vertices. We embed the input twice by using two sets of learned weights, then we transpose one of the embedded matrices and take the dot product between the two and normalize to get the adaptive adjacency matrix, similar as [4].

Each adjacency type is applied with its own graph convolution operation (GCN) by using individual weights instead of replacing the original A_1 with A_2 or A_3. Then, the output of GCN applies a 1×1 convolution as a learnable reduction measure for weighting the stack output and provides the required number of output channels. In this way, the model can increase flexibility without reducing the original performance.

For the spatial dimension, the graph convolution operation is formulated as

$$GCN_l(f_l) = \sum_{i=1}^{3} (D_k^{-\frac{1}{2}}(A_l + I)D_k^{-\frac{1}{2}}) f_{in} W_i, l = 1, 2, 3; \quad (1)$$

$$GCN(f_{out}) = Concat(GCN_1, GCN_2, GCN_3). \quad (2)$$

where A_l is adjacent matrixs, D_k is a degree matrix, I is an identity matrix describing the self-connection of joints, f_{in} is the set of joints, and W_i is trainable parameter of the neighbor subset. $(D_k^{-\frac{1}{2}}(A_l + I)D_k^{-\frac{1}{2}})$ means a normalization of the $A_l + I$.

3.3.2. Spatial Self-Attention (SAA) Module

The transformer model employing self-attention originally designed to operate on words in NLP tasks. The self-attention mechanism reduces the dependence on external information and is better at capturing the internal correlation of data or features. The SAA applies a modified self-attention operator, as depicted in Figure 5, to capture the spatial features of different joints in the same frame and dynamically build spatial relationships within and between joints to strengthen the correlation of body joints that are not directly connected in human skeletons.

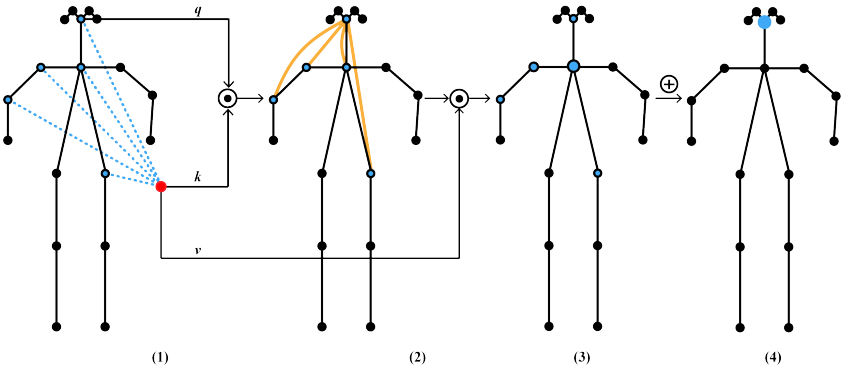

Figure 5. Self-attention of skeleton joints. (**1**) We calculate a query q a key k and a value vector v; (**2**) The query of the joint and the key of all the other joints is performed by dot product ($\odot$), and a weighted value is obtained to represent the strength of the connection between each pair of joints; (**3**) Each joint is scaled to a new node due to its correlation; (**4**) The new features are added ($\oplus$) the weighted nodes together.

The relations between joints are dynamically generated in SAA, thus the relevant structure of skeleton is adaptively generated, not fixed for all actions. The SAA is achieved by independently calculating the correlation between each pair of joints in each frame, as shown in Figure 6. When the source node that calculates the weight needs to calculate the weighted results, all the other nodes are required to participate in the calculation, which is a manifestation of the ability to capture the global characteristics.

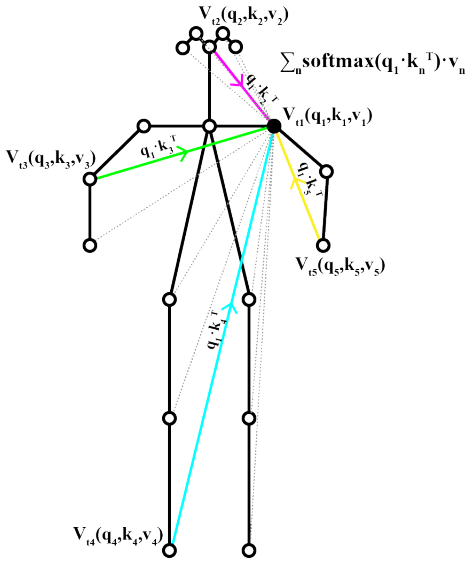

Figure 6. Illustration of Spatial Self-Attention (SAA). For ease of explanation, the process is only shown with a group of five joints as an example, but in fact it runs on all joints.

For each joint v_{tn} of the skeleton at time t, we first calculate the query vector $q_n^t \in \mathbb{R}^{d_q}$, the key vector $k_n^t \in \mathbb{R}^{d_k}$, and the value vector $v_n^t \in \mathbb{R}^{d_v}$ by applying trainable linear transformations to the joint features $j_n^t \in \mathbb{R}^{C_{in}}$ with parameters $W_q \in \mathbb{R}^{C_{in} \times d_q}$, $W_k \in \mathbb{R}^{C_{in} \times d_k}$, $W_v \in \mathbb{R}^{C_{in} \times d_v}$, shared by all nodes. Where C_{in} is the number of input features and d_k, d_q, d_v are the channel dimensions of the key vectors, the query vector and the value vector, respectively. Then, for each pair of body joints (V_{tn}, V_{tm}), the score α_{nm}^t, which represents the strength of the correlations between the two joints, is determined by $\alpha_{nm}^t = q_n^t \cdot (k_m^t)^T \in \mathbb{R}, \forall t \in T$, then it is used to weight each joint value v_m^t, and a weighted sum is calculated to get a new embedding $z_n^t \in \mathbb{R}^{C_{out}}$ for joint v_{tn}, as shown in Equation (3).

$$z_n^t = \sum_m softmax_m(\frac{\alpha_{nm}^t}{\sqrt{d_k}}) v_m^t. \tag{3}$$

Multi-head attention [5] is multiple independent self-attention calculations applied by repeating the process many times, each time with a diverse set of learnable parameters as an integrated function to prevent overfitting.

$$head_{N_h}(X_N) = Softmax(\frac{(X_N W_q)(X_N W_k)^T}{\sqrt{d_k^{N_h}}})(X_N W_V), \tag{4}$$

where X_N is the reshaped input, and $W_q \in \mathbb{R}^{C_{in} \times N_h \times d_q^h}$, $W_k \in \mathbb{R}^{C_{in} \times N_h \times d_k^h}$, and $W_v \in \mathbb{R}^{C_{in} \times N_h \times d_v^h}$ are learned linear transformations. Then the outputs of all heads are concatenated as

$$SA_N = Concat(head_1, \ldots, head_{N_h}) W^o, \tag{5}$$

where W^o is a learnable linear transformation combining outputs of all heads.

3.4. Deep Embedded Clustering

The beginning of clustering layer is the embedding of SAA-STGCAE. We adjust the deep embedded clustering [31] and use our proposed SAA-STGCAE architecture for soft clustering spatiotemporal graphs. The embedding is fine-tuned based on the initial

reconstruction to obtain a cluster-optimized embedding, then each sample is represented by its probability p_{nk} assigned to each cluster

$$p_{nk} = Pr(y_n = k|Z_n, \Theta) = \frac{exp(\theta_k^T Z_n)}{\sum_{k'}^{K} exp(\theta_k^T Z_n)}. \tag{6}$$

where Z_n is the latent embedding generated by the encoder part of SAA-STGCAE, y_n is the soft cluster assignment, and Θ is the clustering layer's parameters with cluster number k.

We perform an algorithm optimization following the clustering goal [31] to minimize the Kullback–Leibler (KL) divergence between the current model probability cluster prediction P and the target distribution Q

$$q_{nk} = \frac{p_{nk}/(\sum_{n'} p_{n'k})^{\frac{1}{2}}}{\sum_{k'} p_{nk'}/(\sum_{n'} p_{n'k})^{\frac{1}{2}}}, \tag{7}$$

$$L_{cluster} = KL(Q||P) = \sum_{n}\sum_{k} q_{nk} log\frac{q_{nk}}{p_{nk}}. \tag{8}$$

In the process of expectation, we fixed the model and updated the target distribution Q, and during the maximization step, the model is optimized to minimize the clustering loss, $L_{cluster}$.

3.5. Normality Score

The Dirichlet process mixture model [32] is a useful measure for assessing the distribution of proportional data and theoretically ideal for processing large, unlabeled dataset. It evaluates a set of distribution parameters in the estimation stage, and uses a fitted model to provide a score for each embedded sample in the inference stage. In the testing phase, the fitted model is used to score each sample with logarithmic probability. The normality score provided by the model is used to determine if the action is normal or not.

4. Experiment

We evaluate the performance of our method for video anomaly detection on two public datasets: ShanghaiTech Campus [17] and CUHK Avenue [33], which can easily identify pedestrians and extract human skeleton data. Figure 7 shows some normal and abnormal events in the dataset used in our experiment. We compare our proposed network with appearance-based [13,17,19] and skeleton-based [21,23] methods. All experiments are evaluated on the frame-level AUC measurement.

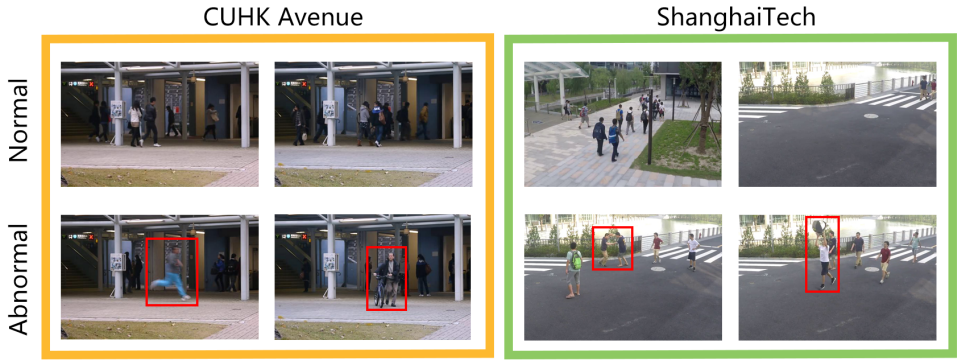

Figure 7. Some normal and abnormal event frames in CUHK Avenue and ShanghaiTech datasets. The abnormal event in the abnormal frame is displayed by the red box.

4.1. Dataset

ShanghaiTech Campus dataset [17] is a new complex and large-scale anomaly detection dataset. The video data of the dataset were collected under 13 scenes with complex lighting conditions and different camera angles in campus. Most of anomalous events in the dataset can be caused by humans, which are the target of our method. We conduct more detailed experiments on this dataset. The previous work [21] divides a subset from ShanghaiTech Campus which contains only anomalous events related to human, denominated HR-ShanghaiTech. We also evaluate our method on this subset.

CUHK Avenue [33] contains 16 training and 21 testing video clips including 47 abnormal events such as movement of pedestrians, the wrong direction of movement, the appearance of abnormal objects. The clips are captured in CUHK campus avenue with a single view.

4.2. Implementation Details

We use Alpha-Pose algorithm [34] to extract skeletons for each frame in each clip in the dataset. For video streams of unknown length, we divide the input pose sequence into fixed-length clips with the sliding window method. For more than one person in the clip, each person is scored individually and we take the highest score of each person in the frame. As done in work [35], the number of heads of multi-head attention is set to 8, and the embedding dimensions of d_q, d_k, and d_v in each layer are $0.25 \times C_{out}$ in all these experiments.

The training of the model includes two stages, the pre-training stage of an autoencoder and the optimization stage of the refinement embedding and clustering adjustment. The pre-training stage of the autoencoder learns to encode and reconstruct sequences by minimizing reconstruction loss, named $L_{reconstruction}$, which is the L_2 loss between the original spatiotemporal graph and the reconstruction of SAA-STGCAE. The optimization stage combines reconstruction loss and clustering loss and the combined loss function is

$$L = L_{reconstruction} + \lambda \cdot L_{cluster}. \tag{9}$$

where λ value is used for weighted clustering loss. The default value is 0.6.

4.3. Comparison with State-of-the-Art Methods

The most popular evaluation metric of video anomaly detection is area under ROC curve (AUC) in previous work [14,15,17,19,33,36]. We report the same metric of frame-level AUC results in Table 1. following the previous work for performance evaluation. A higher value indicates better anomaly detection performance.

Table 1. Anomaly Detection Results. The performance of different methods on ShanghaiTech Campus Dataset, Human-Related ShanghaiTech Campus Dataset (HR-ShanghaiTech) and CUHK Avenue (Avenue). The results are frame-level AUC scores.

	Method	ShanghaiTech Campus	HR-ShanghaiTech Campus	Avenue
Appearance	Conv-AE [13]	0.704	0.698	0.848
	TSC sRNN [17]	0.680	N/A	N/A
	Liu et al. [19]	0.702	0.727	0.862
Skeleton	MPED-RNN [21]	0.734	0.754	0.863
	Normal Graph [22]	0.734	0.765	0.873
	GEPC [23]	0.749	0.756	0.876
	Ours	0.789	0.793	0.884

According to AUC indicator, we compare our method with appearance-based methods and skeleton-based ones. In general, the skeleton-based methods perform better than the appearance-based methods, especially on the HR-ShanghaiTech Campus subset on the HR-ShanghaiTech Campus subset where the anomaly is only related to humans. The reason is that these algorithms only focus on human posture instead of irrelevant features, such as complex background, illumination changes, dynamic camera views, etc. As to the skeleton-based methods, the GCN-based method [22,23] performs better than the RNN-based methods [21], because skeleton can be naturally defined as a graph structure and Graph convolutional networks have advantages over RNN networks in processing Non-Euclidean Structure Data. In addition, our method performs better than GEPC [23] which builds an autoencoder with ST-GCN can only capture local features in spatial dimensions to model the relationship of skeletons, while our method utilizes self-attention to capture global features of skeletons to enhance graph convolution. Therefore, the SAA-Graph can understand the intra-frame interaction of different body parts, and can dynamically establish the relationship between the bones and joints to represent the various parts of the human body.

4.4. Ablation Study

We conduct some experiments to evaluate the effectiveness of our model with the SAA-Graph module by comparing it with the graph convolution baseline (Graph) and self-attention single module (SA), where all these methods adopt the same temporal convolution (TCN). We follow the GEPC [23] settings to implement the graph convolution baseline (Graph) and the results of simplified modules of our method are listed in Table 2. Regarding SA, it conditionally depends on movement and is independent of natural human body structure. From Table 2, we can see the performance of the self-attention module can achieve a similar effect to that of the graph convolution baseline, which demonstrates that self-attention module can replace the graph convolution baseline. The experimental results confirm that the self-attention module is effective, and the best result can be obtained by the SAA-Graph in the control experiments.

Table 2. Ablation study of SAA-Graph component.

	ShanghaiTech Campus	**HR-ShanghaiTech Campus**
SAA-Graph/Graph	0.749	0.756
SAA-Graph/SA	0.746	0.749
SAA-Graph	0.789	0.793

4.5. The Visualization of SAA-Graph

In order to further analyze cases of success and failure, each sample is scored by using the logarithmic probability of the fitted model and visualize the video clips on the ShanghaiTech dataset. As shown in Figure 8, our model SAA-Graph can effectively detect human-related abnormal events in most cases. SAA-Graph can produce high regularity scores in normal activities and low regularity scores in abnormal activity. Abnormal conditions will produce a strong drop peak.

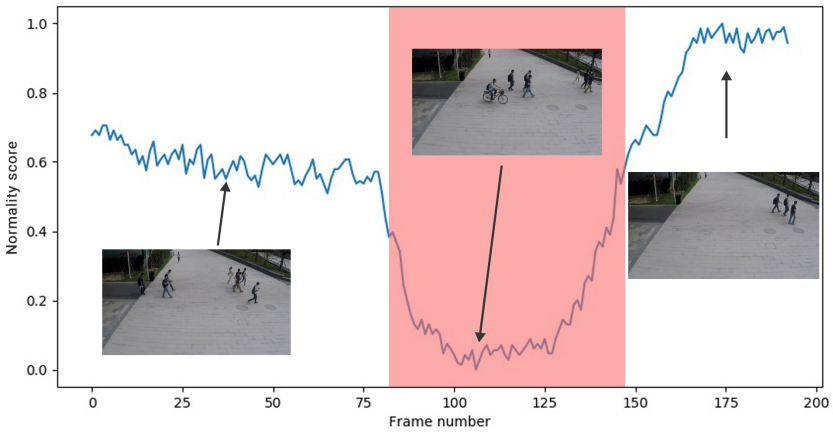

Figure 8. Normality Score of video clips from ShanghaiTech dataset. Regularity score is normalized to [0, 1] and the red areas in the figure represent anomalies.

4.6. Fail Cases Analysis

The performance of SAA-Graph is better than the related methods, but there are still some failure cases. Figure 9a shows the vehicle appearing in the video, which is non-human related incidents and can not be processed by our method because no skeleton is extracted. Figure 9b shows that the abnormal motion tracking skeleton may be lost when obstructed by obstacles. The main reason of this error is the inaccuracy of skeleton detection and tracking. We tested the current advanced skeleton detection methods, all of them have inaccurate skeleton phenomena, such as low-resolution areas of the target person or obstruction by obstacles. Figure 9c shows the pattern of a slow cyclist misjudged to walk due to similar speed and posture to walking due that all appearance features are filtered out. Although individual movements and postures can reflect anomalies in most cases, they do not include the interactions between multiple people and between people and objects in the event. We will consider using visual features to enhance the skeleton structure as a future work to solve this problem.

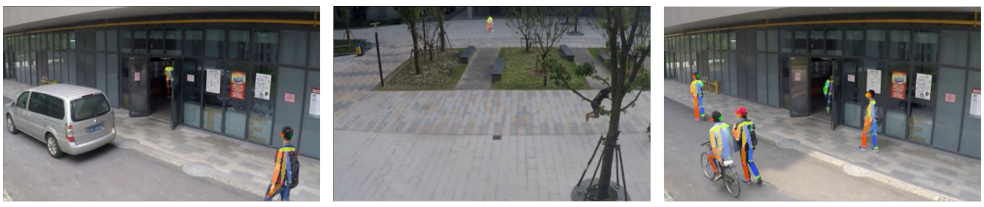

Figure 9. The failure cases on the ShanghaiTech Campus dataset. (**a**) Vehicle appears in the clip, which is not processed by our method; (**b**) Abnormal motion tracking skeleton may be lost when obstructed by obstacles; (**c**) False negative situation is a misjudgment by a slow cyclist.

5. Conclusions

In this work, we propose a novel spatial temporal self-attention augmented graph convolutional clustering networks for skeleton-based video anomaly detection tasks by employing the SAA-STGCAE to extract features and embedded clustering. We proved that the SAA-Graph can achieve a more flexible and dynamic representation between skeletons while overcoming the locality of graph convolution. This data-driven approach increases the flexibility of the graph convolutional network and brings more versatility to adapt to various data samples. To the best of our knowledge, we are the first to consider using self-attention for video anomaly detection tasks as an enhancement of spatial temporal graph convolution to capture global features. Our proposed model achieves the excellent performance on both two anomaly detection datasets, ShanghaiTech Campus and CUHK Avenue. Future work includes detecting abnormal phenomena between humans and

human–object interactions, enhancing skeleton features with appearance features, and looking for a fully self-attentional solution, which leads to improved network performance and reduces the number of parameters.

Author Contributions: Project administration, L.S.; Conceptualization, C.L. and R.F.; Methodology, R.F. and C.L.; Formal analysis, R.F., Y.L. and Y.G.; Funding acquisition, L.S., C.L. and Y.G.; Resources, Y.L.; Validation, Y.G.; Writing, R.F. and C.L.; Writing—review and editing, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program of China under grant of 2018YFC0824402 and 2020YFB1712401, and was supported in part by the Nature Science Foundation of China (62006210).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yan, S.; Xiong, Y.; Lin, D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
2. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
3. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-Based Action Recognition with Directed Graph Neural Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 7912–7921.
4. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
5. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems. In Proceedings of the First 12 Conferences, 2017; pp. 5998–6008. Available online: https://mitpress.mit.edu/books/advances-neural-information-processing-systems (accessed on 13 December 2021).
6. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention Augmented Convolutional Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 3286–3295.
7. Clausi, Z. Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance. *Image Vis. Comput.* **2011**, *29*, 230–240.
8. Anjum, N.; Cavallaro, A. Multifeature Object Trajectory Clustering for Video Analysis. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1555–1564. [CrossRef]
9. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, San Diego, CA, USA, 20–25 June 2005.
10. Dalal, N.; Triggs, B.; Schmid, C. Human Detection Using Oriented Histograms of Flow and Appearance. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006.
11. Fischer, P.; Dosovitskiy, A.; Ilg, E.; Husser, P.; Hazrba, C.; Golkov, V.; Patrick, V.; Cremers, D.; Brox, T. FlowNet: Learning Optical Flow with Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Amsterdam, The Netherlands, 11–14 October 2016.
12. Zhang, X.; Yang, S.; Zhang, J.; Zhang, W. Video anomaly detection and localization using motion-field shape description and homogeneity testing. *Pattern Recognit.* **2020**, *105*, 107394. [CrossRef]
13. Hasan, M.; Choi, J.; Neumann, J.; Roy-Chowdhury, A.K.; Davis, L.S. Learning Temporal Regularity in Video Sequences. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
14. Chong, Y.S.; Tay, Y.H. *Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder*; Springer: Cham, Switzerland, 2017.
15. Medel, J.R.; Savakis, A. Anomaly Detection in Video Using Predictive Convolutional Long Short-Term Memory Networks. *arXiv* **2016**, arXiv:1612.00390.
16. Luo, W.; Wen, L.; Gao, S. Remembering history with convolutional LSTM for anomaly detection. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017.
17. Luo, W.; Wen, L.; Gao, S. A Revisit of Sparse Coding Based Anomaly Detection in Stacked RNN Framework. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

18. Leo, M.; D'Orazio, T.; Spagnolo, P.; D'Orazio, T. Human activity recognition for automatic visual surveillance of wide areas. In Proceedings of the ACM International Workshop on Video Surveillance & Sensor Networks, New York, NY, USA, 15 October 2014.
19. Liu, W.; Luo, W.; Lian, D.; Gao, S. Future Frame Prediction for Anomaly Detection—A New Baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018; pp. 6536–6545. Available online: https://openaccess.thecvf.com/content_cvpr_2018/papers/Liu_Future_Frame_Prediction_CVPR_2018_paper.pdf (accessed on 13 December 2021).
20. Wu, P.; Liu, J.; Li, M.; Sun, Y.; Shen, F. Fast Sparse Coding Networks for Anomaly Detection in Videos. *Pattern Recognit.* **2020**, *107*, 107515. [CrossRef]
21. Morais, R.; Le, V.; Tran, T.; Saha, B.; Mansour, M.; Venkatesh, S. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020.
22. Luo, W.; Liu, W.; Gao, S. Normal Graph: Spatial Temporal Graph Convolutional Networks based Prediction Network for Skeleton based Video Anomaly Detection. *Neurocomputing* **2020**, *444*, 332–337. [CrossRef]
23. Markovitz, A.; Sharir, G.; Friedman, I.; Zelnik-Manor, L.; Avidan, S. Graph Embedded Pose Clustering for Anomaly Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
24. Vemulapalli, R.; Arrate, F.; Chellappa, R. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
25. Hussein, M.E.; Torki, M.; Gowayyed, M.A.; El-Saban, M. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. In Proceedings of the International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
26. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Modeling video evolution for action recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
27. Lei, W.; Du, Q.H.; Koniusz, P. A Comparative Review of Recent Kinect-based Action Recognition Algorithms. *IEEE Trans. Image Process.* **2019**, *29*, 15–28.
28. Xu, Y.; Cheng, J.; Wang, L.; Xia, H.; Feng, L.; Tao, D. Ensemble One-dimensional Convolution Neural Networks for Skeleton-based Action Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1044–1048. [CrossRef]
29. Ding, Z.; Wang, P.; Ogunbona, P.O.; Li, W. Investigation of Different Skeleton Features for CNN-based 3D Action Recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2017; pp. 617–622. Available online: https://arxiv.org/pdf/1705.00835.pdf (accessed on 13 December 2021).
30. Banerjee, A.; Singh, P.K.; Sarkar, R. Fuzzy Integral-Based CNN Classifier Fusion for 3D Skeleton Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2206–2216. [CrossRef]
31. Xie, J.; Girshick, R.; Farhadi, A. Unsupervised Deep Embedding for Clustering Analysis. In Proceedings of the International Conference on Machine Learning, 2016; pp. 478–487. Available online: http://proceedings.mlr.press/v48/xieb16.pdf (accessed on 13 December 2021).
32. Blei, D.M.; Jordan, M.I. Variational inference for Dirichlet process mixtures. *J. Bayesian Anal.* **2006**, *1*, 121–143. [CrossRef]
33. Lu, C.; Shi, J.; Jia, J. Abnormal Event Detection at 150 FPS in MATLAB. In Proceedings of the IEEE International Conference on Computer Vision, 2013; pp. 2720–2727. Available online: https://www.cv-foundation.org/openaccess/content_iccv_2013/papers/Lu_Abnormal_Event_Detection_2013_ICCV_paper.pdf (accessed on 13 December 2021).
34. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. RMPE: Regional Multi-person Pose Estimation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
35. Plizzari, C.; Cannici, M.; Matteucci, M. Skeleton-based Action Recognition via Spatial and Temporal Transformer Networks. *Comput. Vis. Image Underst.* **2021**, *208*, 103219. [CrossRef]
36. Yang, C.; Yuan, J.; Ji, L. Sparse reconstruction cost for abnormal event detection. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011.