Towards Grouping in Large Scenes with Occlusion-aware Spatio-temporal Transformers

Jinsong Zhang, Lingfeng Gu, Yu-Kun Lai, *Member, IEEE*, Xueyang Wang, Kun Li✉, *Member, IEEE*

Abstract*—Group detection, especially for large-scale scenes, has many potential applications for public safety and smart cities. Existing methods fail to cope with frequent occlusions in large-scale scenes with multiple persons, and are difficult to effectively utilize spatio-temporal information. In this paper, we propose an end-to-end framework, *GroupTransformer*, for group detection in large-scale scenes. To deal with the frequent occlusions caused by multiple people, we design an occlusion encoder to detect and suppress severely occluded person crops. To explore the potential spatio-temporal relationship, we propose spatio-temporal transformers to simultaneously extract trajectory information and fuse inter-person features in a hierarchical manner. Experimental results on both large-scale and small-scale scenes demonstrate that our method achieves better performance compared with state-of-the-art methods. On large-scale scenes, our method significantly boosts the performance in terms of precision and F1 score by more than 10%. On small-scale scenes, our method still improves the performance of F1 score by more than 5%. *We will release the code for research purposes.

***Index Terms*—Group detection, Large-scale scenes, Spatio-temporal transformers.**

I. Introduction

GROUP detection is a fundamental task in computer vision that involves identifying groups of people from images or videos, which has numerous applications in human-centric analysis tasks such as abnormal detection [1], [2], trajectory prediction [3]–[6] and group activity recognition [7]–[10]. In this work, our primary focus is on group detection in large-scale scenes, which has significant implications for public safety [11], dynamic environments [12], and smart cities [13], [14].

Existing group detection methods mainly focused on small-scale scenes with limited persons or interactions. Some works focused on group detection on a single image. Traditional methods [11], [16], [17], [18] utilized some hand-crafted features to identify F-formations [19] in crowds, which needed to detect the interaction spaces of groups of people. However, they only detected groups of F-formations, while other spatial patterns of groups were ignored. Furthermore, detecting groups based solely on a single image is not always reasonable, considering the dynamic nature of human relationships in real life. Consequently, researchers have explored group detection in videos. Some methods [20] directly input video clips into an I3D backbone [21] to extract spatial and temporal information simultaneously. However, it is important to note that spatial and temporal concepts are learned by different cognitive mechanisms in our brains [22], indicating the challenge of effectively learning both concepts within the same network. Other methods [23] have utilized trajectory information to detect groups in crowds. However, these approaches overlook the impact of image information, such as interactions between individuals, on group detection. Moreover, the aforementioned existing methods are designed for group detection in short

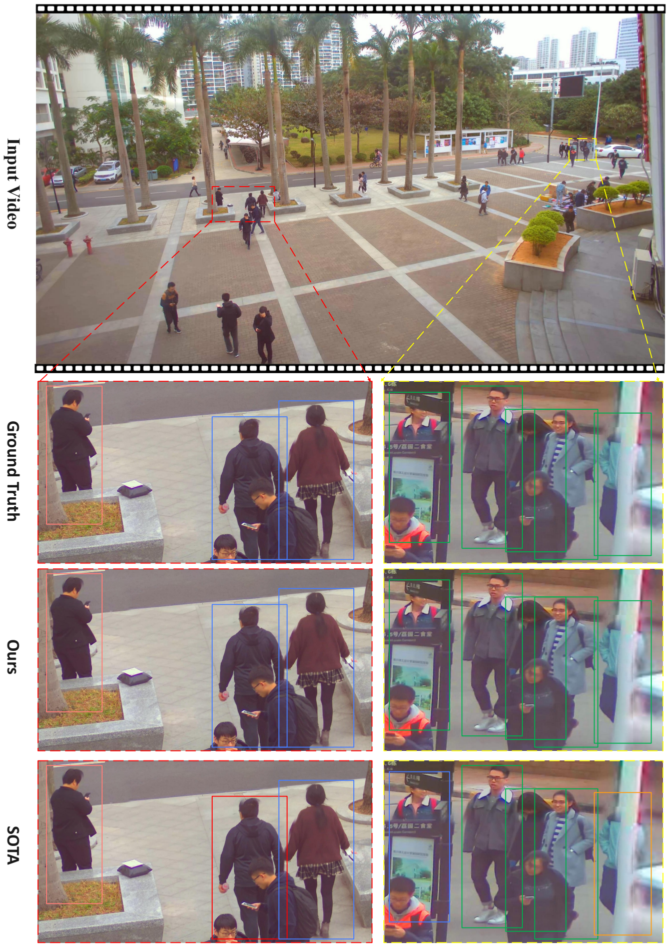

Figure 1. Given a video of a large-scale multi-person scene, our method can detect reasonable groups, despite the large number of people and frequent occlusions in the scene. While the state-of-the-art method [15] fails to predict right results. Different groups are shown with different color boxes.

Jinsong Zhang, Lingfeng Gu and Kun Li are with the College of Intelligence and Computing, Tianjin University, Tianjin 300350, China.

Yu-Kun Lai is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom.

Xueyang Wang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100190, China.

Corresponding author: Kun Li (Email: lik@tju.edu.cn).

videos of small-scale scenes, which are not suitable for large-scale multi-person scenes due to the large computation cost.

Group detection on large-scale multi-person videos holds great research significance and offers a wide range of potential applications. However, there is a scarcity of works in this area primarily due to the limited availability of datasets designed specifically for large-scale multi-person scenes. Wang *et al.* [24] addressed this gap by introducing the gigaPixel-level humANcentric viDeo dAtaset (PANDA), which features a wide field-of-view, gigapixel-level image resolution, and temporally long-term crowd activities. They also proposed a global-to-local zoom-in framework to detect groups in large-scale multi-person videos using the multi-modal inputs, *i.e.*, images and trajectories. However, they ignored the frequent occlusions in crowd videos, *i.e.*, a person can be occluded sometimes, and the appearance features at those time instances are unreliable for group detection. Furthermore, the method extracted image and trajectory features independently, employing 3D ConvNet [25] for spatio-temporal image features and LSTM (Long Short-Term Memory) [26] for temporal trajectory features. It is worth noting that 3D ConvNet faces similar challenges as I3D, and LSTM fails to incorporate spatial information from trajectories. Recently, Li *et al.* [15] proposed a novel group detection method by pre-training the model using a self-supervised method, which produced promising results. However, this method also disregarded the occlusion problem and processed spatial and temporal information separately, which means that the spatial information and the temporal information are processed in sequence and not synchronized [22], limiting its performance.

In this paper, we propose GroupTransformer, an end-to-end spatio-temporal framework designed to address the challenges of frequent occlusions and complex spatio-temporal interactions in large-scale multi-person scenes. To tackle the issue of occlusions prevalent in large-scale scenes, we introduce an occlusion encoder that focuses on better extracting individual features. It leverages inter-frame similarity to identify and suppress features influenced by severe occlusions in specific frames, thereby enhancing the robustness of our model to occlusions. The occlusion encoder improves the overall performance by effectively considering the inter-frame similarities of individuals. Inspired by [22], we design a hierarchical scheme for the spatio-temporal transformers, which addresses the challenge of processing spatial and temporal information in a series computation manner. Our scheme incorporates a spatial branch and a temporal branch, enabling synchronized processing of spatial and temporal information during data-driven training. This design allows for comprehensive exploration of the relationship between trajectory and appearance, with high-level semantic features from the temporal branch guiding the learning of appearance features. The temporal information is primarily extracted from trajectory features using densely connected convolutional layers, while the spatial information is fused and obtained through a transformer encoder. We evaluate our method on both small-scale scene datasets and large-scale scene datasets, and the experimental results demonstrate its superiority over state-of-the-art methods. Figure 1 illustrates an example of the grouping results compared our method with the state-of-the-art method, S3R2 [15]. To facilitate further research, we will release the code for GroupTransformer for academic purposes.

The main contributions of this work are summarized as follows:

- We propose an end-to-end framework which fuses spatio-temporal information from multi-modal inputs to detect groups in a large-scale multi-person scene, which effectively explores the multi-modal information and well deals with the occlusion problem.
- We propose spatio-temporal transformers, comprising a spatial branch and a temporal branch, which facilitate the hierarchical fusion of appearance and trajectory features.
- To further enhance the personalized individual feature extraction process, we propose an occlusion encoder. This component effectively suppresses features affected by severe occlusions in specific frames, thereby improving the anti-occlusion ability of our model.
- Experimental results on both large-scale and small-scale benchmark datasets demonstrate the superior performance of our proposed method.

We organize the remainder of this paper as follows: in Section II, we give a brief review of related work, including static methods, dynamic methods for group detection, occlusion-aware methods, and spatio-temporal methods. In Section III, we introduce our proposed GroupTransformer, including the training strategy and loss functions. In Section IV, we first validate the effectiveness of our method through qualitative and quantitative experimental results, comparing it with several state-of-the-art methods. Subsequently, we conduct ablation studies to assess the impact of different components in our model. Additionally, we evaluate the robustness of our method by introducing noise to the bounding boxes and randomly reducing the bounding boxes of each person. Finally, we conclude and discuss our work in Section V.

II. Related Work

In this section, we review group detection on still images and videos. The methods can be classified into static methods and dynamic methods according to the input. Besides, we review the existing works about occlusion-aware methods and spatio-temporal methods to learn about the motivation of our model.

A. Static Methods

Group detection in still images has been studied extensively in the literature. These methods aim to detect interactions and spatial arrangements among individuals in a single image.

Early static methods focused on detecting F-formations, which are spatial patterns formed by free-standing conversational individuals. Kendon *et al.* [19] proposed the concept of F-formations and identified specific spatial configurations that tend to emerge during conversations. Traditional static methods for detecting F-formations relied on hand-crafted rules and mathematical models [27], [28]. With the advancement of deep learning, recent static methods have benefited from the power of deep neural networks to improve performance in group detection. For example, some methods have adopted Graph Neural Networks (GNNs) [29] to model the mutual features and relationships among individuals in a single image. These methods leverage position and pose features of individuals and construct a fully connected graph to transfer messages and capture the group structure [30].

Static methods are suitable for closed scenes with a limited number of people, such as social gatherings or small group activities [17], [31]. However, they have limitations in real-world scenarios where the number of individuals is large and dynamic. Additionally, static methods often rely on camera parameters and depth information to reconstruct world coordinates, which may not be readily available in practical settings. Moreover, grouping results obtained from a single image may not be reliable, as the relationships and interactions between individuals are inherently dynamic and may change over time. Therefore, there is a need for methods that can leverage temporal information from video sequences to improve group detection performance.

B. Dynamic Methods

Dynamic methods leverage the temporal information from consecutive frames to infer the relationships among individuals in a video sequence. Early dynamic methods focused on using trajectory features for group detection. Ge *et al.* [23] proposed a classic trajectory-based method that utilized pedestrian detection and tracking techniques to extract trajectories from video frames. They then applied hierarchical clustering to detect groups of people with similar trajectories. However, relying solely on trajectory information may not be sufficient to determine the relationships between individuals. For example, in a crowd walking on a sidewalk, most people may have similar trajectories but have no direct relationship with each other.

To overcome this limitation, appearance features obtained from images, which contain meaningful interactive information, are incorporated to obtain reliable grouping results. Ehsanpour *et al.* [20] proposed a novel framework for small-scale videos group detection in small-scale videos by utilizing appearance features. They used the I3D network [32] as a video backbone to extract spatial and temporal features across frames. However, this type of feature extractor may not be suitable for large-scale scenes, especially for videos with gigapixel-level resolution. Additionally, using the same network to extract both spatial and temporal features from the inputs may not be reliable.

To address the challenges of large-scale scenes, researchers have explored different approaches. Wang *et al.* [24] proposed a global-to-local zoom-in framework for group detection in large-scale scenes. They utilized both appearance and trajectory features. However, in their appearance-based model, the fusion of spatial and temporal information from appearance input was not well addressed. In their trajectory-based model, they used LSTM to capture temporal information from the trajectory input but ignored the spatial information. Furthermore, the occlusion of appearance features of different persons in long-duration crowd videos was not properly considered, leading to inaccurate grouping results. Recently, Li *et al.* [15] proposed a two-stage method that pre-trains the model on unsupervised tasks before fine-tuning for group detection. While achieving promising performance, this approach neglects frequent occlusions in crowd videos and relies on a gated recurrent unit model for temporal information aggregation, which may limit its effectiveness.

In this paper, we propose an end-to-end framework for group detection in large-scale multi-person scenes by extracting personalized individual features with an occlusion encoder and exploring the relationship between trajectory and appearance features using spatio-temporal transformers. By explicitly considering occlusion and leveraging both spatial and temporal information, our method aims to improve group detection performance in challenging large-scale scenes.

C. Occlusion-aware Methods

Occlusion problems exist in many computer vision tasks, including person re-identification [33], optical-flow estimation [34], instance segmentation [35] and so on. Existing occlusion-aware methods aim to recover the occluded information through different approaches. For person re-identification, Wang *et al.* [33] proposed a feature erasing and diffusion network to recover the occluded information by data augmentation to achieve intrinsic representation. For flow estimation, Wang *et al.* [34] addressed occlusion by estimating an occlusion map to post-process the estimated flow. For instance segmentation, Ke *et al.* [35] presented a novel perspective that segments single images as double-layer images, and adopted a transformer-based network to recover the occluded information, which meant it also tries to deal with the occlusion problem by recovering the occluded information.

In our work, instead of recovering the occluded information, we adopt a transformer-like architecture [36] as our occlusion encoder to extract more useful and precise information, while ignoring unreliable information caused by occlusions. By focusing on the essential information, our model aims to achieve accurate and reliable group detection in the presence of occlusions.

D. Spatio-temporal Methods

The fusion of spatial and temporal information is a well-studied problem, particularly in video-based tasks [37]–[46]. Some methods used optical flow as the temporal information to cope with video instance segmentation [47], [48]. However, these approaches heavily rely on accurate optical flow estimation, which can be challenging and memory-intensive, making them less suitable for large-scale scenes. Some methods [42], [43] utilize CNN-based networks to extract frame-wise features and employ RoIAlign to extract individual features for group activity recognition. However, in the context of large-scale scenes with images at the giga-pixel level, extracting frame-wise features becomes impractical. Ehsanpour *et al.* [20] employed video-based backbones, such as I3D [21], for video activity recognition. However, using the same architecture to extract both spatial and temporal features may not be optimal [22]. Zheng *et al.* [38] proposed a novel temporal attention mechanism for abnormal event detection. They treated temporal information as a sequence and used non-local attention to extract and aggregate it. Similar with [38], some methods [49]–[51] adopted transformer-like architectures to cope with spatio-temporal information. These methods consider the temporal information as a sequence, and apply attention mechanism to extract and aggregate information for each frame, which is suitable for video-based tasks. However, as discussed in [22], the spatial information and the temporal information should be learned by different cognitive mechanisms, and should be synchronized to process the sequential information. The conventional approaches [41], [45], [46] that first extract spatial information from each frame and then fuse it with temporal transformers may lead to a lack of information synchronization. Besides, in group detection, the temporal information *e.g.*, trajectories, is not just a sequence, it is also the important location information for distinguishing the group results. LSGD [24] and S3R2 [15] handle temporal information and spatial information respectively, which can result in suboptimal performance.

Motivated by [22], we propose to update the temporal information and the spatial information in a hierarchical manner. Specifically, we propose the spatio-temporal transformer with a temporal branch and a spatial branch, which first extracts temporal information, and then concatenates with the spatial information. With this operation, the spatial information can be synchronized with temporal information. We stack several spatio-temporal transformers to extract and aggregate spatial and temporal information effectively.

III. Method

The objective of our work is to detect groups in large-scale multi-person scenes. To achieve this, we represent individuals and their relationships as vertices and edges in a graph $G = (V, E)$. We then formulate the group detection task as an edge classification task, and exploit features from multiple modalities (the trajectory and the video) for edge classification. The proposed framework consists of three key components: 1) **occlusion encoder**, which takes the appearance feature of individuals as inputs and aims to alleviate the influence of occluded person crops in specific frames; 2) **spatio-temporal transformers (STT)**, which encode the features from both modalities, *i.e.*, the trajectory and the video, and fuse information across both spatial and temporal dimensions; 3) **edge classifier**, which classifies the interested edges in the graph G according to the fused individual features from spatio-temporal transformers. Figure 2 shows the framework of our method, depicting the flow of information and the interaction between the different components.

A. Occlusion Encoder

Previous works [15], [24] of group detection in large-scale scenes ignored the issue of frequent occlusions in multi-person scenes. As a result, the extracted features may contain noise and unreliable information. To address this problem, we propose the occlusion encoder, inspired by the attention mechanism [36]. Instead of recovering occluded information [33]–[35] in other vision tasks, the occlusion encoder aims to filter out the occluded information in the input sequence, allowing us to obtain more reasonable and meaningful information from the inputs.

The occlusion encoder consists of two learnable functions, denoted as $f(\cdot)$ and $g(\cdot)$. The input is the sequential appearance features $\mathcal{X} \in \mathbb{R}^{N\times D\times T}$ extracted by a pre-trained ResNet50 [52], where N, D, T represent the number of persons, the dimension of appearance features, and the number of frames, respectively. For each person and its sequential appearance features $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T] \in \mathbb{R}^{D\times T}$, we assume that most of the person crops are not occluded. Then, we leverage the fact that appearance features for the same non-occluded person crops exhibit inter-frame similarity, while occluded ones are likely to have little similarity with other frames.

To capture the inter-frame similarity, we calculate the affinity between frames using the inner product in the normalized feature space. Specifically, the similarity between frame i and frame j of a person can be computed as:

$$s_{i,j} = \frac{f(\mathbf{x}_i)f(\mathbf{x}_j)^\top}{\|f(\mathbf{x}_i)\|\|f(\mathbf{x}_j)\|}. \tag{1}$$

Here, $f(\cdot)$ is a learnable function that maps appearance features to a feature space, and $\mathbf{x}_i$ represents the appearance feature at frame i. The similarity values range from 0 to 1, as $f(\cdot)$ outputs ReLU-activated feature vectors.

To determine the attention value a_i for frame i, we calculate the mean similarity between frame i and all other frames:

$$a_i = \frac{1}{T}\sum_{j=1}^{T} s_{i,j}. \tag{2}$$

For a severely occluded frame i, it will have little similarity with the other frames, resulting in a small attention value a_i. In practice, if a person does not appear in all frames, the mean value is calculated only based on the visible frames. Finally, we obtain the processed appearance feature with the attention mask by:

$$\mathbf{z}_i = g(\mathbf{x}_i) \times a_i. \tag{3}$$

The embedding functions $f(\cdot)$ and $g(\cdot)$ are implemented by a linear layer and a ReLU activation function. The output of this module is the refined appearance features for each person, denoted as $Z_{app} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T]$.

B. Spatio-temporal Transformers

Previous works on video understanding [20], [23] adopted RNN/LSTM-based or 3D ConvNet-based approaches to extract temporal information. However, these methods typically process spatial and temporal information in a sequential manner, leading to a lack of synchronization between spatial and temporal features [22]. This limitation hinders the effective representation of extracted features. In order to exploit the potential relationship between the multi-modal inputs in both

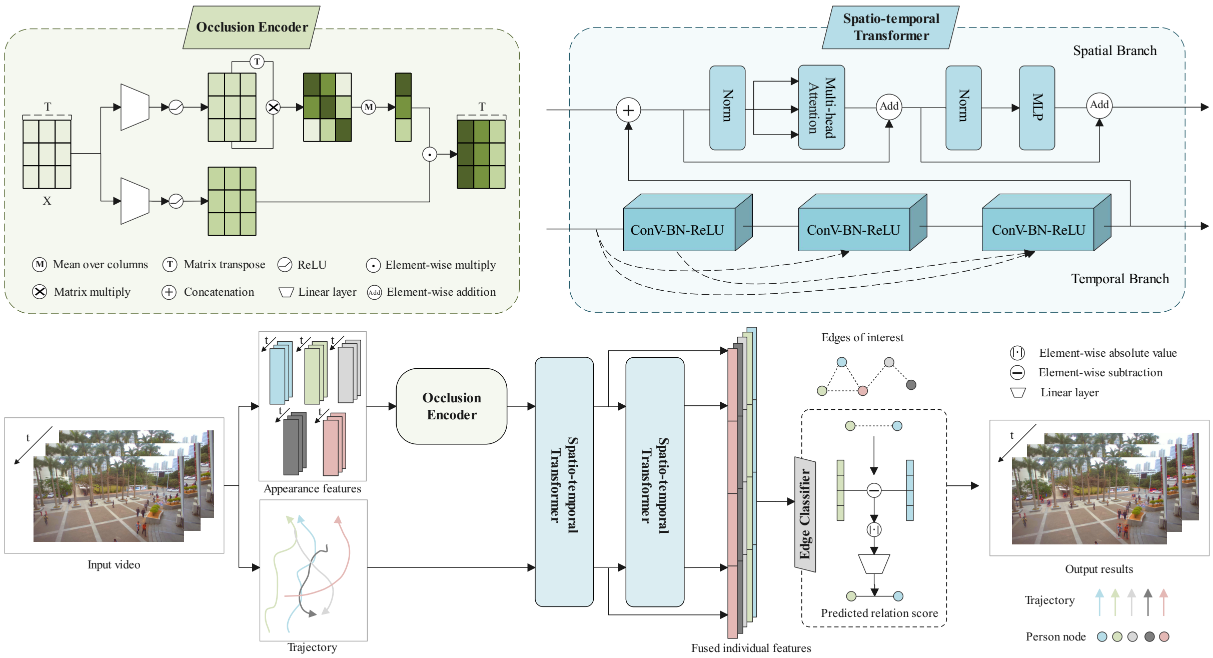

Figure 2. Overview of the proposed GroupTransformer. We initially adopt a pre-trained ResNet50 network to extract individual appearance features. Subsequently, the proposed occlusion encoder is utilized to filter out occluded information. Finally, spatio-temporal transformers are employed to fuse both trajectory and appearance features. In output results, the individuals encapsulated within red bounding boxes, who are interconnected by blue edges, are classified into the same group.

the spatial and temporal dimensions, we propose the spatio-temporal transformers (STT) to fuse spatio-temporal information from trajectory and appearance features.

The STT module contains two branches: a temporal branch and a spatial branch. The temporal branch extracts high-level temporal semantics from trajectory features, such as the moving velocity at each moment. We design the temporal branch based on the DenseNet [53], where 1D convolutional layers are utilized, and each layer is connected to every other layer in a feed-forward manner. This architecture enables the preservation of both low and high-level trajectory semantics.

The spatial branch employs a transformer encoder to capture the spatial patterns from both trajectory and appearance features. Specifically, given the input appearance feature of all persons $\mathcal{Z}_{app_m} = \{\mathcal{Z}_{app_m^n} | n = 1, 2, ..., N\}$ at depth m, it is firstly concatenated with the processed trajectory features $\mathcal{Z}_{traj_{m+1}}$ to form the raw embedding features for individual persons:

$$\mathcal{Z}_m = \mathcal{Z}_{app_m} \oplus \mathcal{Z}_{traj_{m+1}}. \tag{4}$$

We view the temporal dimension as the batch dimension and apply a transformer encoder to exploit spatial context for all persons. The process of embedding spatial context for frame i can be formulated as:

$$Q_m^i = \mathcal{Z}_m^i W_{q,m}, K_m^i = \mathcal{Z}_m^i W_{k,m}, V_m^i = \mathcal{Z}_m^i W_{v,m}, \tag{5}$$

$$V_m'^i = softmax(\frac{Q_m^i K_m^i}{\sqrt{D_1}})V_m^i + V_m^i, \tag{6}$$

$$V_m''^i = \mathrm{MLP}(V_m'^i), \tag{7}$$

where $W_{q,m}$,$W_{k,m}$,$W_{v,m}$ are learnable parameters, D_1 is the dimension of Q_m, and MLP is the Multi-Layer Perceptron in the canonical transformer. The features of all persons at all time instances $\{V_m''^i | i = 1, 2, ..., T\}$ are packed together as $\mathcal{Z}_{app_{m+1}}$. Finally, the STT module outputs the extracted features from the two branches, $\mathcal{Z}_{traj_{m+1}}$ and $\mathcal{Z}_{app_{m+1}}$, which can be further used as input of the next STT module. We stack M STT modules to form a deep model inspired by [54].

C. Edge Classifier

Previous spatio-temporal transformers [42], [43] designed for group activity recognition are not suitable for grouping. Adopting these approaches to generate the complete output results simultaneously makes it necessary to predict all pair-wise relation scores. This way not only imposes additional computational overhead but also introduces training complexities owing to the abundance of zero scores, given that most individuals are not part of the same group. To deal with this problem, we propose an edge classifier to predict the relation scores for edges in the graph $G(V, E)$ using the features of individual persons. The individual features are collected by concatenating all appearance and trajectory features from the STT modules of different depths, which can be expressed as:

$$\mathcal{Z}_{all} = \mathcal{Z}_{app_1} \oplus ... \oplus \mathcal{Z}_{app_M} \oplus \mathcal{Z}_{traj_1} \oplus ... \oplus \mathcal{Z}_{traj_M}. \tag{8}$$

Note that $\mathcal{Z}_{all}$ preserves the temporal dimension, resulting in $\mathcal{Z}_{all} = [\mathbf{z}_{all}^1, \mathbf{z}_{all}^2, ...\mathbf{z}_{all}^T]$. Then, we convert the individual features to inter-person edge features. Specifically, for an edge

TABLE I: Architecture of GroupTransformer. The hyper-parameters for linear layers are denoted by L(input dimension, output dimension); the hyper-parameters for convolutional layers are denoted by C(input dimension, output dimension, kernel size); and the hyper-parameters for transformer encoders are denoted by T(hidden layer dimension, hidden layer number, the number of heads). The variables N, T, N_e represent the number of persons, the number of frames, and the number of edges, respectively.

Module	Input	Parameter	Value
Occlusion Encoder	$N \times T \times 16384$	$f(x)$	$L(16384, 1024)$
		$g(x)$	$L(16384, 512)$
STT1	$N \times 5 \times T$	Conv1	$C(5, 64, 3)$
	$N \times 69 \times T$	Conv2	$C(69, 64, 3)$
	$N \times 133 \times T$	Conv3	$C(133, 128, 3)$
	$T \times N \times 640$	Transformer Encoder	$T(128, 2, 4)$
STT2	$N \times 128 \times T$	Conv1	$C(128, 64, 3)$
	$N \times 192 \times T$	Conv2	$C(192, 64, 3)$
	$N \times 256 \times T$	Conv3	$C(256, 128, 3)$
	$T \times N \times 256$	Transformer Encoder	$T(128, 2, 4)$
Edge Classifier	$N_e \times T \times 512$	Linear	$L(512, 1)$

$(u, v) \in E$, we construct its feature $F_{(u,v)}$ by taking the absolute difference between the features of the two individuals:

$$F_{(u,v)} = \left| \mathcal{Z}_{all}^{u} - \mathcal{Z}_{all}^{v} \right| . \tag{9}$$

After obtaining the edge features, they are fed into a Multi-Layer Perceptron (MLP) to generate frame-wise classification logits. Subsequently, a global average pooling is applied along the temporal dimension to obtain the overall prediction. This can be expressed as:

$$R_{u,v} = \mathrm{MLP}_{\mathrm{R}}(F_{u,v}), \; c_{u,v} = \frac{1}{T} \sum_{t=1}^{T} R_{u,v}^{t}, \tag{10}$$

where $R_{u,v}$ represents the classification logits for edge u and edge v, $c_{u,v}$ represents the average prediction score over the temporal dimension, and $\mathrm{MLP}_{\mathrm{R}}$ is the Multi-Layer Perceptron layer. For the person pairs with social interactions, the classifier will assign positive scores for the corresponding edges, while for those without interactions, negative scores are supposed.

D. Training

We train the proposed model in an end-to-end manner. Instead of sampling a fixed number of persons in a scene, we sample a fixed number of groups, which ensures a certain number of positive edges to guarantee sufficient training. However, the number of negative edges is still much higher than that of positive edges. This leads to two issues: (1) classifying a large number of edges becomes computationally inefficient; and (2) most of the negative edges are too easy to discriminate, limiting the performance of the model.

To cope with these problems, we propose a pre-processing strategy that filters out edges that are not worth training during the construction of the relation graph. We keep all positive edges due to their rarity, while removing unnecessary negative edges. Specifically, we first filter out edges where the two persons never appear simultaneously in the frames. Then, we ignore person pairs with a minimal distance on trajectories larger than a threshold δ_{train}. This means that only hard negative edges are retained for training. By applying this strategy, we construct the edge set E corresponding to the sampled persons. Simultaneously, the edges are assigned binary labels $y_{u,v} \in 0, 1, \forall (u, v) \in E$ according to the ground-truth group annotation. During training, we utilize binary cross-entropy loss to train the model in a supervised manner:

$$\begin{aligned} \mathcal{L} = - \sum_{(u,v) \in E} & (1 - \lambda) y_{u,v} \times \log(\sigma(c_{u,v})) + \\ & \lambda(1 - y_{u,v}) \times \log(1 - \sigma(c_{u,v})), \end{aligned} \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\lambda = \frac{|E_{positive}|}{|E|}$ is a balance coefficient to account for the difference in the number of positive and negative edges. Here, $E_{positive}$ and E are the sets positive edges and all edges, and $| \cdot |$ is the cardinality of the set.

E. Inference

During the training phase, the input persons are controlled to form a limited number of groups. However, during inference, all the persons from a scene are fed into the model simultaneously. In the case of large scenes, this can result in thousands of persons and millions of edges in the graph, which is computationally expensive. To improve efficiency, we adopt two strategies to filter out obvious negative edges during classification.

The first strategy is similar to the training phase. We filter out edges where the distance between two persons exceeds a threshold δ_{test}. It is important to note that δ_{test} should be larger than δ_{train} to avoid mistakenly removing positive edges. The second strategy involves removing edges where two persons do not appear simultaneously for most of the frames. We calculate the Intersection over Unions (IoU) at visible frames of each person. If the IoU is less than a threshold γ, the edge is assumed to represent no interaction behavior and is removed from the edge set E. After applying these filtering strategies, we feed the remaining edges into our model and construct an affinity matrix using the predicted relation scores. The ignored edges are assigned a score of 0 by default. Finally, we solve a clustering problem on the affinity matrix to detect groups. Various clustering algorithms can be applied for this task (see the next subsection for implementation details).

F. Implementation Details

The appearance features are extracted using a ResNet50 [52] pre-trained on ImageNet. We extract feature vectors of dimension $4\times2\times2048$ from the layer preceding the max-pooling operation in ResNet50. These vectors are then flattened to form 16384-dimensional feature vectors, which serve as the appearance features. Our models are trained using a stochastic gradient descent optimizer with no momentum, and the learning rate is set to 0.1 initially. We sample 8 groups in each iteration and apply the gradient descent every 10 iterations. For large-scale datasets, the models are trained for 200 epochs, and the learning rate drops by a factor of 5 at 50, 100 and 150 epochs. For small-scale datasets, the models are trained for 20 epochs without learning rate decay. In data preprocessing, we set $\delta_{train} = 0.1, \delta_{test} = 0.2, \gamma = 0.3$ for large-scale scenes, and $\delta_{train} = 0.5, \delta_{test} = 0.75, \gamma = 0.001$ for small-scale scenes. It is important to note that the positions in trajectories are normalized, so larger scenes with high image resolution require smaller threshold values. For group inference, we apply the same clustering algorithm as the compared methods to ensure fair comparison. We use label propagation [55] for large-scale scenes and spectral clustering [56] for small scenes. We train and test our model on a desktop with an Intel(R) Xeon(R) CPU and a GeForce RTX 2080 Ti GPU. The training time for PANDA dataset is about 20 hours, while it takes around 3 hours for JRDB dataset. The inference time of our model is related to the number of edges in the test video. On average, the inference time on JRDB test set is 0.11 seconds, while the inference time on PANDA test set is 0.48 seconds. The number of model parameters is about 33.73M.

Detailed Framework. The detailed architecture of our GroupTransformer is defined in Table I. We stack 2 Spatio-Temporal Transformers (STT), namely STT1 and STT2. The $f(x)$ and $g(x)$ functions are followed by a ReLU activation function. All the Conv1, Conv2, and Conv3 are followed by a Batch Normalization layer [57] and a ReLU activation function.

TABLE II: Quantitative comparison on *PANDA* dataset.

Method	PANDA		
	Precision	Recall	F1
G2L w/o Local [24]	0.237	0.120	0.160
G2L w/ Random [24]	0.244	0.133	0.172
G2L w/ Uncertainty [24]	0.293	0.16	0.207
Dis.Mat+ [58]	0.429	0.120	0.188
GNN w/ GRU	0.419	0.173	0.245
ARG [59]	0.349	0.200	0.254
S3R2 [15]	0.559	0.507	0.532
Ours	**0.750**	**0.545**	**0.632**

TABLE III: Quantitative comparison on *JRDB-Group* dataset.

Method	JRDB-Group		
	Precision	Recall	F1
Joint [20]	0.300	0.284	0.291
JRDB-Group [60]	0.390	0.379	0.384
Dis.Mat+ [58]	0.573	0.235	0.334
GNN w/ GRU	0.434	0.286	0.345
ARG [59]	0.325	0.384	0.352
S3R2 [15]	0.577	0.562	0.569
Ours	**0.662**	**0.606**	**0.633**

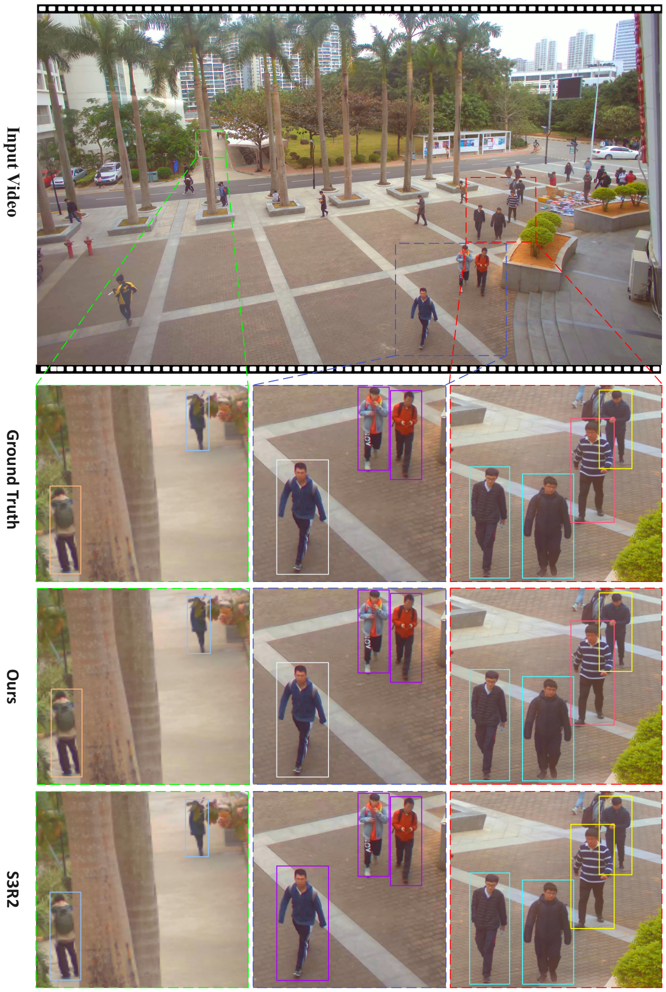

Figure 3. Qualitative results compared with S3R2 [15] on *PANDA* benchmark. More results can be found in the supplementary video.

IV. Experimental Results

A. Datasets

To validate the effectiveness of our method, we conduct our experiments on a large-scale dataset *PANDA* [24] and a small-scale dataset *JRDB-Group* [60] following [15]. The details of two datasets are given in the following.

PANDA benchmark. PANDA is a gigapixel-level human-centric video dataset with a wide field-of-view (up to $1km^2$) and a very dense crowd (up to 4k subjects in a frame). It consists of 9 videos for group detection, providing rich and hierarchical ground-truth annotations, including bounding boxes, fine-grained labels, trajectories, and interactions. Following [15], [24], we adopt 8 videos as the training set and 1 video as the test set for fair comparison. On average, each training video has 2713 frames and $1070.4k$ bounding boxes, while each test video has 3500 frames and $335.2k$ bounding boxes. In the training and test video sets, the average group sizes per video are 144.6 and 75, respectively. Same with [15], [24], we use the ground-truth bounding boxes and trajectories to validate the effectiveness of our method in training and test phases.

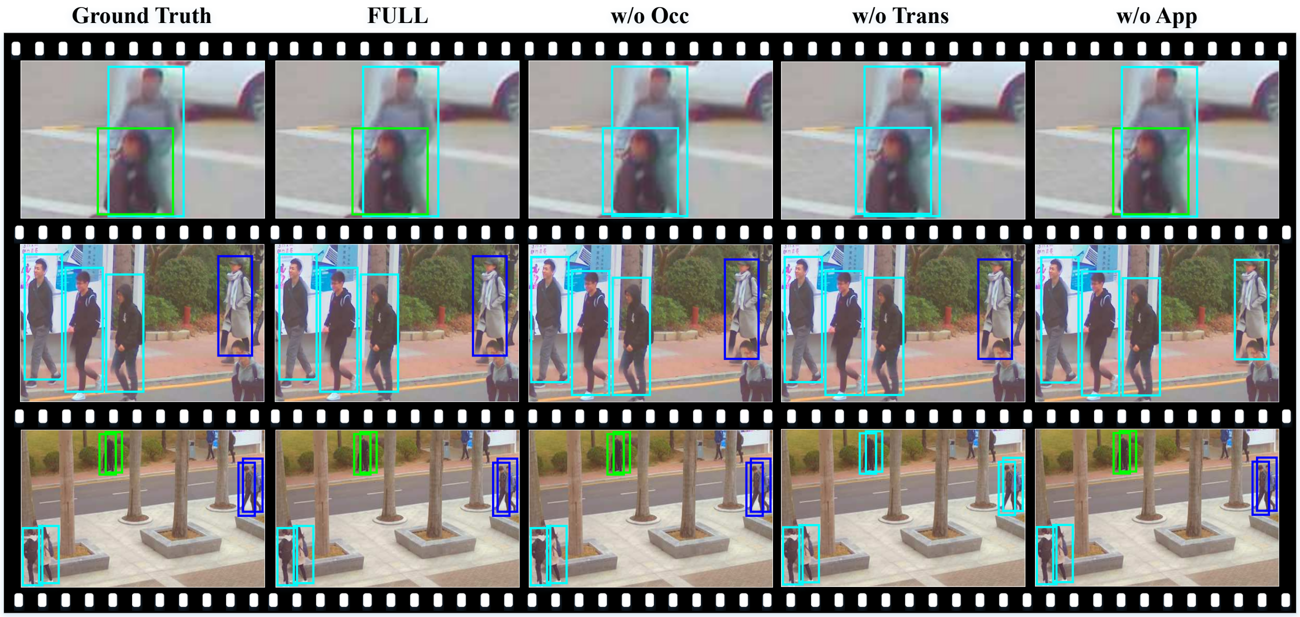

Figure 4. Qualitative results of ablation study on *PANDA* benchmark.

JRDB-Group benchmark. JRDB-Group [60] is a multi-person video dataset captured by a panoramic camera equipped on a robot, which contains outdoor and indoor crowd scenes. Following [15], we take 20 videos as the training set and 7 videos as the test set, and use key frames to validate the performance of group detection. The key frames are sampled every 15 frames, and we get 1419 samples for training and 404 samples for testing.

B. Metrics

To evaluate the performance of our method, we use the same metrics as the compared state-of-the-art methods [15], [24]. For large-scale scenes, we calculate precision, recall, and F1-score using the half metric [61] with a group member IoU threshold of 0.5. The half metric determines whether a detected group is considered positive or negative based on the intersection over union (IoU) between the individuals in the detected group and the ground-truth group. This can be formulated as

$$\frac{|Gp_{det} \cap Gp_{gt}|}{max(|Gp_{det}|\,, |Gp_{gt}|)} > 0.5, \tag{12}$$

where Gp_{det} denotes the detected groups, and Gp_{gt} denotes the ground-truth groups. That means if the number of individuals in the intersection is greater than half the number of individuals in the ground-truth group, the detected group is considered a positive sample. Precision, recall, and F1-score are then calculated based on these positive and negative samples.

C. Comparison Results

We try to conduct as much comparative experiments as possible to demonstrate the performance of our model. However, the source code of most group detection methods are not available, so we implement all models based on [15] to give a clear comparison results. Here, we give a brief introduction of compared methods.

Dis.Mat+ [58]. This method is a straightforward method that first measures the distance of each subject pair in the scene and adopts label propagation algorithm [15] to obtain the final group results.

GNN w/ GRU. This method is designed by [15], which uses a gated recurrent unit model to aggregate the temporal information and adopts a graph neural network to model the graph relation.

ARG [59]. This method is a state-of-the-art method for group activity recognition. We follow the modifications applied in [15].

G2L [24]. This method is the baseline method proposed with the PANDA benchmark. It models the group relation of subjects as a graph, and adopts a global-to-local strategy to leverage the visual cues for group detection. The group results can be obtained after label propagation.

Joint [20] and JRDB-Group [60]. The Joint method leverages group detection results to get better results for group activity recognition, which is suitable to be a compared method. The JRDB-Group is built on Joint by introducing spatial information.

S3R2 [15]. This method is the most relevant method and is also the state-of-the-art method of group detection. It first trains the model using its self-supervised method, and then adds a group detection head to tune the model to get the group detection result using supervised training.

The quantitative results presented in Table II and Table III demonstrate the superior performance of our method compared to other state-of-the-art methods on the PANDA and JRDB-Group datasets.

On the PANDA dataset, our method achieves the best performance in terms of precision, recall and F1-score. The

TABLE IV: Quantitative comparison with three alternative designs.

Method	*PANDA*			*JRDB-Group*		
	Precision	Recall	F1	Precision	Recall	F1
w/o Occ.	0.686	0.530	0.598	0.568	0.522	0.544
w/o Trans.	0.574	0.470	0.517	0.560	0.508	0.533
w/o App.	0.679	0.545	0.605	0.585	0.570	0.577
FULL	**0.750**	**0.545**	**0.632**	**0.662**	**0.606**	**0.633**

baseline method G2L and other straightforward methods struggle to predict correct group detection results for the large-scale dataset, resulting in unsatisfactory performance. S3R2, which adopts self-supervised learning for pre-training, performs better than the baseline methods but fails to consider occlusion circumstances in crowd scenes. Our method addresses the occlusion problem through the proposed occlusion encoder and leverages the spatio-temporal transformer to model temporal information. As a result, our method significantly improves the performance compared to S3R2, with an increase in the F1-score from 0.532 to 0.632. These results highlight the effectiveness of our approach in handling large-scale group detection challenges. Similarly, on the JRDB-Group dataset, our method outperforms the compared methods, including Joint and JRDB-Group, as well as Dis.Mat+, GNN w/ GRU, ARG, and S3R2. Our method achieves a higher precision, recall, and F1-score, indicating its superiority in group detection on both large-scale and small-scale scenes. Overall, the comparison results demonstrate that our method surpasses other methods in terms of performance metrics on both datasets, proving its effectiveness and suitability for group detection tasks.

Figure 3 presents qualitative results on the PANDA benchmark, comparing our approach with the state-of-the-art method S3R2 [15][1]. It can be observed that S3R2 fails to detect groups in cases where individuals are walking towards the same direction, such as the three pedestrians in the second column. With our spatio-temporal transformer, our model can distinguish whether pedestrians with similar trajectories belong to the same group. Dynamic results and additional results can be found in the supplementary video.

TABLE V: Quantitative results with different Gaussian noises.

Metrics	$\sigma = 0$	$\sigma = 0.1$	$\sigma = 0.3$
precision	0.750	0.750	0.715
recall	0.545	0.545	0.409
F1	0.632	0.623	0.519

TABLE VI: Quantitative results with different Missing detection rate.

Metrics	*MDR*=0	*MDR*=0.1	*MDR*=0.2
precision	0.750	0.793	0.700
recall	0.545	0.348	0.212
F1	0.632	0.484	0.326

D. Ablation Study

We evaluate our method with three alternative models to assess the factors that contribute to achieving better group detection results on both small-scale videos and large-scale videos.

The Model without Occlusion Encoder (w/o Occ.). We delete the mask branch in occlusion encoder and the appearance features are simply processed with a linear projection layer.

The Model without Transformer (w/o Trans.). We use a MLP to fuse appearance and trajectory feature instead of a transformer encoder.

The Model without Appearance Feature (w/o App.). We input only trajectories with the spatial branch deleted, and consequently, the fused individual features are obtained by only concatenating the features from the temporal branch.

Table IV shows quantitative results compared with three alternative models on PANDA and JRDB-Group benchmarks, respectively. The full model outperforms all the alternatives on both large-scale PANDA benchmark and small-scale JRDB-Group benchmark, which verifies the effectiveness of different modules. The model with only trajectory as input works better than the model with appearance feature but without occlusion encoder. The possible reason is that frequent occlusions make the appearance features full of noise, leading to a negative effect on performance.

To give a more intuitive reasoning, Figure 4 shows the performance of different ablation models on several typical cases. We zoom in the persons in the large-scale scenes, and the persons belonging to the same group are marked with the same color. Our full model predicts the groups accurately. In the top row, the model without occlusion encoder and the model without transformer fail to distinguish the two separate persons. The overlap of the two bounding boxes makes their appearance features similar, and therefore the edge between them is misclassified as positive. We also notice that the two persons can be separated simply by the trajectory, demonstrating that improper use of appearance features or insufficient fusion model may cause confusion. In the middle row, four persons follow the same trajectory along the sidewalk, and thus the model without appearance feature tends to detect the four persons as a group. Appearance features are required to make the correct prediction. In the bottom row, six persons are wrongly detected as a group by the model without the transformer. They have the same destination and similar posture, which will lead to the wrong detection if the features are insufficiently extracted.

E. Robustness

To evaluate the robustness of our method with a raw video input, we simulate the detection errors by adding noise to the bounding boxes or randomly dropping some detections on PANDA test set.

[1] The public repository only contains the source code for the PANDA benchmark.

Adding Noise to Bounding Boxes. Denote σ as the noise intensity and $\mathbf{u} = [x_0, y_0, x_1, y_1]$ as a bounding box, where x_0, y_0 and x_1, y_1 are the top-left and bottom-right coordinates of the bounding box. We define w and h as the width and the height of $\mathbf{u}$. Taking top-left coordinates as an example, we add the disturbances δ_x and δ_y to top-left coordinates x_0 and y_0 by sampling them from Gaussian distributions with standard deviation $\sigma \times w$ and $\sigma \times h$, respectively. Table V shows the quantitative results with different levels of Gaussian noises. Even with noise $\sigma = 0.3$, our method still outperforms LSGD [24] that uses the clean input (Table II).

Missing Detections. Denote *MDR* as the missing detection rate of the bounding box to each person in the whole video. Table VI shows the quantitative results with different missing detection rates. With increasing *MDR*, although the performance of our method degrades, our method still outperforms LSGD without missing detections [24],

F. Limitations

Although our method effectively detects groups in most large-scale scenes and small-scale scenes, it cannot cope with the cases of extremely low resolution and long-time occlusion. Figure 5 shows some failure cases on large-scale scenes. The middle row shows ground-truth groups and the bottom row shows our grouping results. Our method fails to group persons due to the low resolution of images and the serve occlusion of persons in a long time. This can be coped in our future work.

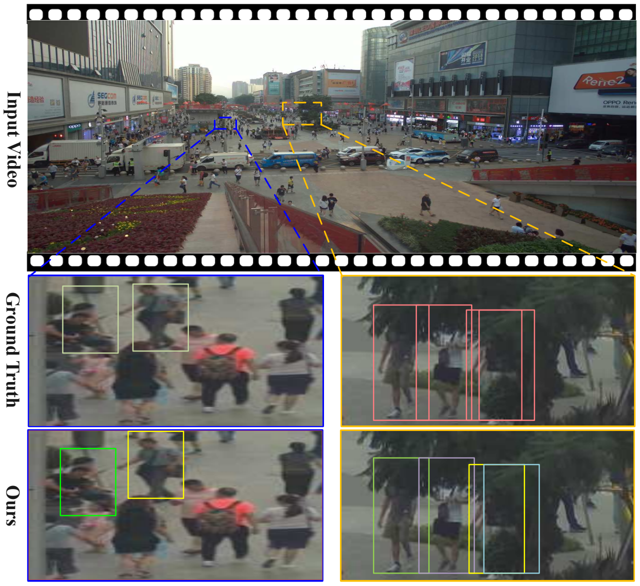

Figure 5. Some failure cases on large-scale scenes.

V. Conclusion

In this paper, we propose a novel end-to-end spatio-temporal framework for group detection in large-scale video scenes. The proposed framework addresses the challenges of frequent occlusions and complex spatio-temporal interactions. The occlusion encoder effectively extracts individual features by considering occlusion patterns, while the spatio-temporal transformers capture the dynamic relationships among individuals in a hierarchical manner. The comprehensive evaluation on the PANDA and JRDB-Group benchmarks confirms the superior performance of our method, and the ablation study demonstrates the effectiveness of each proposed module. The proposed framework opens up possibilities for further research in group detection and understanding. Future work could focus on exploring more advanced occlusion modeling and extending the framework to handle more complex scenarios with variable group sizes and activities.

References

[1] Y. Zhang, L. Qin, R. Ji, H. Yao, and Q. Huang, "Social attribute-aware force model: exploiting richness of interaction for abnormal crowd detection," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 25, no. 7, pp. 1231–1245, 2014.

[2] S. Coşar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Brémond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, 2016.

[3] A. D. Berenguer, M. Alioscha-Perez, M. C. Oveneke, and H. Sahli, "Context-aware human trajectories prediction via latent variational model," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 31, no. 5, pp. 1876–1889, 2020.

[4] X. Wang, Z. Chen, J. Tang, B. Luo, Y. Wang, Y. Tian, and F. Wu, "Dynamic attention guided multi-trajectory analysis for single object tracking," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 31, no. 12, pp. 4895–4908, 2021.

[5] C. Choi and B. Dariush, "Looking to relations for future trajectory forecast," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019.

[6] X. Liu, J. Yin, J. Liu, P. Ding, J. Liu, and H. Liu, "Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 31, no. 6, pp. 2133–2146, 2020.

[7] Y. Xing and J. Zhu, "Deep learning-based action recognition with 3d skeleton: a survey," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 80–92, 2021.

[8] W. Choi, K. Shahid, and S. Savarese, "What are they doing? : Collective activity classification using spatio-temporal relationship among people," in *Proc. IEEE Int. Conf. Comput. Vis. Worksh*, 2009.

[9] L. Lu, Y. Lu, R. Yu, H. Di, L. Zhang, and S. Wang, "Gaim: Graph attention interaction model for collective activity recognition," *IEEE Trans. Multimedia*, vol. 22, pp. 524–539, 2020.

[10] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered hmms," *IEEE Trans. Multimedia*, vol. 8, pp. 509–520, 2006.

[11] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-Formation Detection: Individuating free-standing conversational groups in images," *PLOS ONE*, vol. 10, no. 5, p. e0123783, 2015.

[12] N. Bain and D. Bartolo, "Dynamic response and hydrodynamics of polarized crowds," *Science*, vol. 363, no. 6422, pp. 46–49, 2019.

[13] J. Rios-Martinez, A. Spalanzani, and C. Laugier, "From proxemics theory to socially-aware navigation: A survey," *Int. J. Social Robot.*, vol. 7, no. 2, pp. 137–153, 2015.

[14] M. Bendali-Braham, J. Weber, G. Forestier, L. Idoumghar, and P.-A. Muller, "Recent trends in crowd analysis: A review," *Machine Learning with Applications*, vol. 4, p. 100023, 2021.

[15] J. Li, R. Han, H. Yan, Z. Qian, W. Feng, and S. Wang, "Self-supervised social relation representation for human group detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022.

[16] S. Thompson, A. Gupta, A. W. Gupta, A. Chen, and M. Vázquez, "Conversational group detection with graph neural networks," in *Proc. Int. Conf. on Multimodal Interaction*, 2021.

[17] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of F-formations." in *Proc. Brit. Mach. Vis. Conf.*, 2011.

[18] H. Hedayati, D. Szafir, and S. Andrist, "Recognizing F-formations in the open world," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2019.

[19] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*, 1990, vol. 7.

[20] M. Ehsanpour, A. Abedin, F. Saleh, J. Shi, I. Reid, and H. Rezatofighi, "Joint learning of social groups, individuals action and sub-group activities in videos," in *Proc. Eur. Conf. Comput. Vis.*, 2020.

[21] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[22] B. Pang, K. Zha, H. Cao, J. Tang, M. Yu, and C. Lu, "Complex sequential understanding through the awareness of spatial and temporal concepts," *Nature Mach. Intell.*, vol. 2, pp. 245–253, 2020.

[23] W. Ge, R. T. Collins, and R. B. Ruback, "Vision-based analysis of small groups in pedestrian crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 1003–1016, 2012.

[24] X. Wang, X. Zhang, Y. Zhu, Y. Guo, X. Yuan, L. Xiang, Z. Wang, G. Ding, D. Brady, Q. Dai, and L. Fang, "PANDA: A gigapixel-level human-centric video dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[25] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D cnns retrace the history of 2D cnns and imagenet?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[26] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inform. Process. Syst.*, 2015.

[27] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in *Proc. ACM Int. Conf. Multimodal Interfaces*, 2011.

[28] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "Detecting conversational groups in images and sequences: A robust game-theoretic approach," *Comput. Vis. Image Understand.*, vol. 143, pp. 11–24, 2016.

[29] P. Battaglia, J. B. C. Hamrick, V. Bapst *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[30] M. Swofford, J. Peruzzi, N. Tsoi, S. Thompson, R. Martín-Martín, S. Savarese, and M. Vázquez, "Improving social awareness through dante:Deep affinity network for clustering conversational interactants," *Proc. ACM Hum.-Comput. Interaction*, vol. 4, no. CSCW1, pp. 1–23, 2020.

[31] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks: Towards socially and personality aware visual surveillance," in *Proc. ACM Int. Workshop Multimodal Pervasive Video Anal.*, 2010.

[32] C. Gu, C. Sun, D. A. Ross *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[33] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[34] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[35] L. Ke, Y.-W. Tai, and C.-K. Tang, "Occlusion-aware instance segmentation via bilayer network architectures," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Adv. Neural Inform. Process. Syst.*, 2017.

[37] J. Cui, L. Zheng, Y. Yu, Y. Lin, H. Ni, X. Xu, and Z. Zhang, "Deeply-recursive attention network for video steganography," *CAAI Transactions on Intelligence Technology*, 2023.

[38] X. Zheng, Y. Zhang, Y. Zheng, F. Luo, and X. Lu, "Abnormal event detection by a weakly supervised temporal attention network," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 3, pp. 419–431, 2022.

[39] Y. Fang, B. Luo, T. Zhao, D. He, B. Jiang, and Q. Liu, "St-sigma: Spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting," *CAAI Transactions on Intelligence Technology*, vol. 7, no. 4, pp. 744–757, 2022.

[40] J. Yang, X. Guo, K. Li, M. Wang, Y.-K. Lai, and F. Wu, "Spatio-temporal reconstruction for 3d motion recovery," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 30, no. 6, pp. 1583–1596, 2019.

[41] Q. Zhou, X. Li, L. He, Y. Yang, G. Cheng, Y. Tong, L. Ma, and D. Tao, "Transvod: End-to-end video object detection with spatial-temporal transformers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7853–7869, 2023.

[42] X. Zhu, Y. Zhou, D. Wang, W. Ouyang, and R. Su, "Mlst-former: Multi-level spatial-temporal transformer for group activity recognition," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 33, no. 7, pp. 3383–3397, 2023.

[43] S. Li, Q. Cao, L. Liu, K. Yang, S. Liu, J. Hou, and S. Yi, "Groupformer: Group activity recognition with clustered spatial-temporal transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021.

[44] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, and M. Y. Yang, "Spatial-temporal transformer for dynamic scene graph generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021.

[45] Z. Li, J. Li, Y. Ma, R. Wang, Z. Shi, Y. Ding, and X. Liu, "Spatio-temporal adaptive network with bidirectional temporal difference for action recognition," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 33, no. 9, pp. 5174–5185, 2023.

[46] H. Liu, Y. Liu, Y. Chen, C. Yuan, B. Li, and W. Hu, "Transkeleton: Hierarchical spatial–temporal transformer for skeleton-based action recognition," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 33, no. 8, pp. 4137–4148, 2023.

[47] Y.-H. Tsai, M.-H. Yang, and M. J. Black, "Video segmentation via object flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[48] M. Ding, Z. Wang, B. Zhou, J. Shi, Z. Lu, and P. Luo, "Every frame counts: Joint learning of video segmentation and optical flow," in *Proc. AAAI*, 2020.

[49] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, and H. Li, "Decoupled spatial-temporal transformer for video inpainting," *arXiv preprint arXiv:2104.06637*, 2021.

[50] Z. Geng, L. Liang, T. Ding, and I. Zharkov, "Rstt: Real-time spatial temporal transformer for space-time video super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[51] X. Li, J. Huang, J. Zhang, X. Sun, H. Xuan, Y.-K. Lai, Y. Xie, J. Yang, and K. Li, "Learning to infer inner-body under clothing from monocular video," *IEEE Trans. Vis. Comput. Graph.*, 2022.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[53] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[54] J. Zhang, X. Liu, and K. Li, "Human pose transfer by adaptive hierarchical deformation," *Computer Graphics Forum*, vol. 39, no. 7, pp. 325–337, 2020.

[55] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "WiderPerson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Trans. Multimedia*, vol. 22, pp. 380–393, 2020.

[56] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inform. Process. Syst.*, 2002, pp. 849–856.

[57] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, July 2017.

[58] X. Zhan, Z. Liu, J. Yan, D. Lin, and C. C. Loy, "Consensus-driven propagation in massive unlabeled data for face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018.

[59] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, "Learning actor relation graphs for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[60] M. Ehsanpour, F. Saleh, S. Savarese, I. Reid, and H. Rezatofighi, "Jrdb-act: a large-scale multi-modal dataset for spatio-temporal action, social group and activity detection," *arXiv preprint arXiv:2106.08827*, 2021.

[61] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese, "Discovering groups of people in images," in *Proc. Eur. Conf. Comput. Vis.*, 2014.