PRN: Psychology-Inspired Relation Network for Detecting Social Interaction Groups from Single Images

Jiaqi Yu
jiaqiyu@sjtu.edu.cn
Jinhai Yang
youngjh@sjtu.edu.cn
Hua Yang ✉
hyang@sjtu.edu.cn
Guangtao Zhai
zhaiguangtao@sjtu.edu.cn

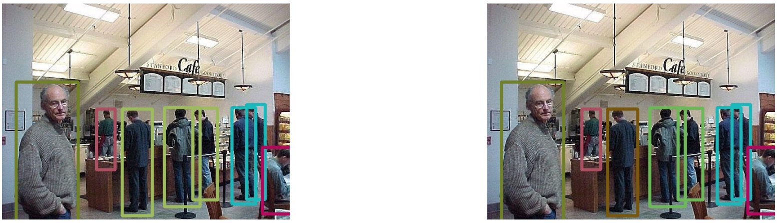
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China
Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai, China
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
✉corresponding author.

Abstract

Detecting interaction groups is an essential task for understanding human behaviours and social activities. However, it is still challenging to identify social interactions and the resulting crowd groups using purely visual cues, especially from single images. Prior works either require additional statistics, such as interpersonal angles and kinaesthetic information, or simply deduce the group memberships with the similarity of individual actions. In this paper, we present the Psychology-inspired Relation Network (PRN) to comprehensively understand the static social scenes and effectively model the interaction relations between individuals. More concretely, stimulated by recent advances in social psychology, we first predict the keypoint heatmap from an image with the human bounding boxes as the visual representations of the key factors determining interaction groups: distance, orientation and postural openness. We then incorporate the personal and mutual influences together to compute the interaction strength matrix via self-attention, and finally utilise a perception to convert this matrix into dyadic interaction probability. Moreover, we devise two loss functions, the dyad loss to optimise the dyadic interaction probability and the group loss to enhance the distinguishability among different social groups. To evaluate the performance of PRN, we introduce a novel dataset containing various scenes with different crowd densities, by merging representative databases and relabeling the group labels. Our method achieves outstanding results on the proposed dataset.

1 Introduction

Humans are by nature social animals as they are innately prone to interact with each other and thus form social groups. Automatic detection of interaction groups in social scenes has broad application prospects, such as group re-identification [45, 46, 48] and crowd anomaly detection [7, 24, 26]. During a public health emergency like COVID-19, this technology



(a) Example of the Interaction Group (IG).

(b) Example of the Social Interaction Group (SIG).

Figure 1: Illustration of the IG versus the SIG. Colours denote different groups. (a) Prior works usually treat the three persons in front of the counter as a single group. In [4], the group category of them is defined as "standing facing same direction" . (b) In this paper, they are identified into two groups according to the interpersonal social interactions. Two people with green bounding boxes are talking while another person has no obvious intention to interact with them.

could also be helpful for epidemiological investigation and contact tracing. Detecting Interaction Groups (IGs) [2, 4] has gained growing interest from the fields of computer vision, sociology and psychology. Gestalt psychology has identified several principles of perceptual groupings, such as proximity, similarity, and common fate [41]. However, most prior studies on visual group discovery are limited to interpersonal proximity and action similarity. It focuses on the consistency of group actions. Facing the same direction, queuing and other actions are considered as IGs. In this paper, we are motivated to identify Social Interaction Groups (SIGs) from single images, which is a relatively new task yet to be fully explored. As a psychological research [31] emphasised, the definition of social interaction is "*behavior that tries to influence or take into account another's subjective experiences or intentions*". The group formed by social interaction is called Social Interaction Group. Our work focuses more on the study of SIGs, that is, the mutual influence between intentions, such as communication, students listening to teachers in class, and the interaction between defensive players and offensive players on the court. The common pattern of these actions is that when individuals in the group produce an action or intention, others respond to it. Our grouping criteria is stricter than that of the preceding studies [2, 4], resulting in a more fine-grained group partition. For example, as shown in Fig. 1, the person waiting alone in front of the counter does not influence or get influenced by the others' subjective intentions, so we identify him as an isolated group. Consequently, our task is more challenging than existing works of visual group discovery.

To tackle this challenging task, we construct an end-to-end network called Psychology-inspired Relation Network (PRN). Recent advances in social psychology [51] found that closer interpersonal distances [13], more coherent orientations [1], and greater postural openness [11, 38] suggest higher probability of social interactions. Although [51] also provided a quantitative model regarding these key factors, it is infeasible to obtain precise statistics from static images. Instead, we propose to utilise the keypoint heatmaps [36] generated with the pose estimation technology to represent the postural and directional information, and use the human bounding boxes (b-boxes) to provide the location cues. The rich information encoded in the heatmaps can capture the relative relation between keypoints better, rather than direct regression of the keypoints [37]. The b-boxes can directly reflect the position of each person in the image, which are elemental for the relative distance calculation. We use a convolutional layer and a fully-connected layer to further extract the features of heatmaps, which reflect the **personal influence** of individuals in the social interaction space. On the other hand, people will affect each other in the interaction space [12], and distance is the most important incarnation of the **mutual influence**. Due to the impact of shooting angles, the distance relationship can not be well captured by simply calculating the centre distance of b-boxes [43]. We combine the relative position encoding proposed in [39] with a perception to learn the distance relation between persons while maintaining the relative position invariance of each person. Then, a self-attention mechanism is adopted to fuse the personal influence and the mutual influence to compute the interaction strength matrix. Finally, we employ a perception to convert the obtained strength matrix into interaction probabilities between any dyads. To optimise the PRN, we also propose two loss functions that constraint the network at the pair-wise level and the group-wise level, respectively. The Dyad Loss directly optimises the output interaction probability between each pair, while the Group Loss focuses more on the intra-group collectiveness and inter-group separations among the detected groups in an image from a global perspective.

In order to verify the effectiveness of the proposed model, we collected a new dataset, dubbed as Social Interaction Dataset (SID). In this dataset, some images are selected from the existing databases, SGD [4] and CAD [3]. Numerous pictures of some NBA games are also included to enrich the scene diversity of SID. The group labels are re-labelled to fit the above-discussed group definition of SIG in this paper. The SID dataset has various crowd densities and assorted social scenes, and the experimental results indicate that the proposed PRN achieves substantial improvements over all the competitors.

2 Related work

Human Interaction Recognition. (1) Sociology-based methods: Early research on visual perception of human interactions is commonly inspired by sociological studies. A particularly important notion is the F-formation, which are defined as the intrinsic spatial patterns that humans maintain during social interactions [19]. In practice, [5] exploited some typical arrangements of the predefined forms of social spaces to find interaction groups in static images with a Hough voting strategy [23]. [16] uses modularity cut method to estimation F-formation. [32] adopts a multi-scale method to adaptively discover the F-formation in the image. [33] proposed to detect F-formations using a computational model for clustering individuals, with the efficient graph-cut based optimisation [22]. [34] uses F-formation to find social groups in images or videos combined with estimating the person distribution in space. However, this kind of methods usually requires proxemic information such as head orientations and positions. In actual scenarios, it may not be easy to obtain directly. **(2) Action-based methods:** Recent works tend to detect interactions by action similarity. [21] fed an interaction model with the action features of individuals to identify the interactive relations between dyads. [47] combine CNN with LSTM [14] to extract the temporal and spatial features of each persons from video sequences. [42] introduced the action compatibility to constrain a graph network with a logic-aware reasoning module. These methods commonly attend to a limited set of actions, which are sub-optimal for generic interaction recognition where an infinite variety of actions may take place. **(3) Other methods:** [49] identifies human interactions leveraging both geometric and social relations. However, facial information is indispensable for this model. In many crowded scenes, it may be difficult to detect faces due to occlusions. Different from these methods, inspired by the theory in [51], this paper uses the combination of position information and personal posture information to

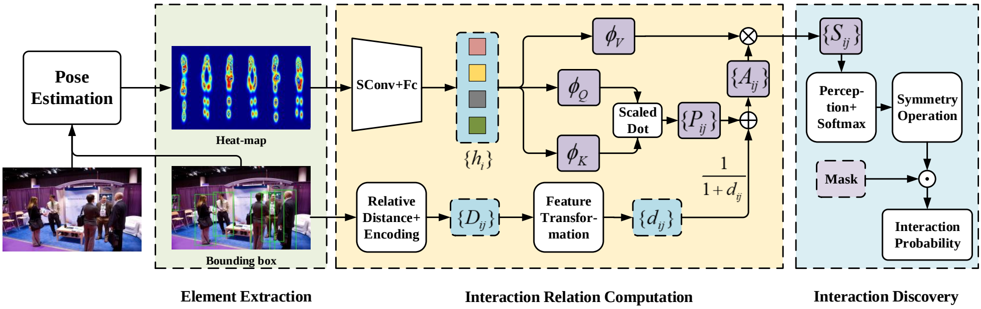


Figure 2: Overview of the proposed PRN. The given human b-boxes reflect the position information, while the heatmaps obtained by pose estimation reflect the postural openness and orientation information. We employ a self-attention mechanism to compute the interaction strength according to the extracted features of the interaction elements. For the operations represented by $\oplus$, $\otimes$ and $\odot$, please refer to Eq. (5), Eq. (6), and Eq. (7) respectively.

realize the detection of interaction groups in images.

Group Relation Analysis. Understanding group relations is essential for interaction recognition and group discovery. Many studies of group activity recognition have explored to analyse the relations among individuals. Rather than directly recognising group activities, a common methodology is to introduce an intermediate representation referred to as structure groups [4], which models how people interact spatially. [17] proposed a hierarchical network by stacking multiple relational layers to represent interpersonal relations. [30] inferred the relations based on spatio-temporal attention and semantic graph. [43] also constructed the relation graphs to capture the underlying interactions between actors, and employed the graph convolutional network [20] with sparse temporal sampling strategy for the relational reasoning. A recent study [6] attempted to use the graph attention networks [40] for directly learning the potential interactions and meanwhile capturing the global activity context. The discovery of these relationships is more to help identify group activities, while this paper only focuses on the discovery of interaction, which is a kind of fundamental research.

3 Approach

Given an image and the associated b-boxes, our target is the interaction probability matrix $\mathbf{R} = (R_{ij})_{N\times N}$ (N represents the the number of people in an image), where $R_{ij} \in [0,1]$ represents the interaction probability between the i-th and the j-th person. Naturally, $\mathbf{R}$ should be a symmetric matrix with zeros on its diagonal. Subsequently, the social groups are determined via probability thresholding. As shown in Fig. 2, the proposed PRN mainly constitutes three modules, as detailed in the following.

3.1 Element Extraction

Psychological research [8] has indicated that the human visual system typically identifies people first during group identification in complex scenes. Therefore, we first discover the interaction elements of individuals. More concretely, we adopt the human b-boxes to provide the positional information, and leverage the keypoint heatmaps to represent the postural openness and the orientation information. In this paper, thanks to recent progress in pose estimation, we adopt the Simple Baseline [44] pretrained on the COCO [25] dataset to predict the keypoint heatmaps of each person. For the i-th person, the corresponding keypoint heatmaps can be denoted by $\mathbf{heat}_i \in \mathbb{R}^{C\times W_0\times H_0}$, where C is the number of keypoints, and H_0 and W_0 are the fixed size of the output heatmap.

3.2 Interaction Relation Computation

The proxemics theory [10] in social psychology describes the basic principle of human interaction as human-centred spaces where people affect each other. Our model excavates the personal and mutual influence in a computational manner, and adopts a self-attention mechanism to learn the pair-wise relations to calculate the interaction strength.

Personal Influence. Each person has their own social influence in interactive scenarios. A convolutional layer with spectral normalisation [27] and a fully-connected (FC) layer is used to learn the underlying information of postural openness and orientation from the extracted interaction elements, which are closely related to the personal influence. Let *SConv* denote the convolutional layer with spectral normalisation, this process can be defined by:

$$\mathbf{h}_i = \mathbf{W}_h^T SConv(heat_i) + \mathbf{m}_h \tag{1}$$

where W_h and m_h are the weight and bias of the FC layer, and $\mathbf{h}_i \in \mathbb{R}^{dh}$ represents personal influence feature.

Mutual Influence. Interpersonal distance is a crucial factor of the mutual influence on interaction [13]. Simply calculating the pixel distance between b-boxes cannot adapt to the evolving interaction situation. Instead, our model calculates the distance relations between individuals using relative distances and position coding. For the b-boxes $\mathbf{b}_i = [b_i^x, b_i^y, b_i^w, b_i^h]$, $\mathbf{b}_j = [b_j^x, b_j^y, b_j^w, b_j^h]$ of the i-th and the j-th person, the relative distance is computed in a way similar to [15]: $B_{ij} = \left[\log\left(\frac{\left|b_i^x - b_j^x\right|}{b_i^w}\right), \log\left(\frac{\left|b_i^y - b_j^y\right|}{b_i^h}\right), \log\left(\frac{b_i^w}{b_j^w}\right), \log\left(\frac{b_i^h}{b_j^h}\right)\right]^T$. Inspired by Transformer [39], the relative distance is coded by relative position encoding to obtain D_{ij}, the relative distance relation of this dyad. Given the coding frequency $L = \left[\frac{1}{1000^{\frac{k-1}{d_{model}}}}\right]_K^T$. Calculate the outer product of the relative distance and coding frequency $E_{ij} = B_{ij} \otimes L$. Here, $\otimes$ represents the outer product. The dimension of E_{ij} is $\mathbb{R}^{4\times K}$. Change E_{ij} to an vector. The final relative position coding is obtained from sin and cos coding: $D_{ij} = [sin(E_{ij}), cos(E_{ij})]$, where $D_{ij} \in \mathbb{R}^{dd}$ and $dd = 8 \times K$.

Interaction Relation Computation. Since the correlation between elements can be captured by the self-attention mechanism, it is used to construct the interaction strength between dyads in an image. Using the Scaled Dot Product Attention [39], the interaction strength caused by personal influence $\mathbf{P} = (P_{ij})_{N\times N}$ is calculated by:

$$P_{ij} = \frac{\phi_Q(\mathbf{h_i})^T \phi_K(\mathbf{h_j})}{\sqrt{dim}} \tag{2}$$

where $\phi_Q(\mathbf{h}_i) = \mathbf{W}_Q^T\mathbf{h}_i + \mathbf{m}_Q$, $\phi_K(\mathbf{h}_i) = \mathbf{W}_K^T\mathbf{h}_i + \mathbf{m}_K$. $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{dh\times dim}$ are learnable weights and $\mathbf{m}_Q, \mathbf{m}_K \in \mathbb{R}^{dim}$ are learnable bias.

The other is mutual influence. For the relative position coding, feature transformation is performed on it to align it with the dimension of P_{ij}. Mutual influence features $\mathbf{d} = (d_{ij})_{N\times N}$ are:

$$d_{ij} = Max\left(\mathbf{W}_b^T D_{ij} + m_b, 0\right) \tag{3}$$

where $\mathbf{W}_b \in \mathbb{R}^{dd}$ is projection weight and $m_b \in \mathbb{R}$. Considering the inverse relation between the distance and the interaction probability, an inverse transformation is made on the distance relation. To prevent infinity, a constant term λ is added to the denominator.

$$d_{ij}^{'} = \frac{1}{\lambda + d_{ij}} \tag{4}$$

Combining Eq. (2) and (4), the interaction weight $\mathbf{A} = (A_{ij})_{N\times N}$ can be got as follows:

$$A_{ij} = \frac{exp(P_{ij} + logd_{ij}^{'})}{\sum_j exp(P_{ij} + logd_{ij}^{'})} = \frac{exp(P_{ij}) \cdot exp(logd_{ij}^{'})}{\sum_j exp(P_{ij}) \cdot exp(logd_{ij}^{'})} = \frac{d_{ij}^{'} exp(P_{ij})}{\sum_j d_{ij}^{'} exp(P_{ij})} \tag{5}$$

The interaction strength is obtained by multiplying each person's interaction weight between others with their personal influence features. An affine transformation $\phi_V(\mathbf{h}_i) = \mathbf{W}_V^T h_i + \mathbf{m}_V$ is applied to get personal influence feature ϕ_{V_i} of i-th person and $\phi_{V_i} \in \mathbb{R}^{dim}$. Combined with interaction weight, the global interaction strength $\mathbf{S}$ in an image can be acquired as follows:

$$S_{i,:,:} = (A_{i,:})^T \otimes \phi_{V_i} \tag{6}$$

where $\mathbf{S} \in \mathbb{R}^{N\times N\times dim}$.

3.3 Interaction Discovery

Interaction Probability. To intuitively reflect the pairwise relations in the image, the interaction strength is converted into probabilities to measure the interaction between dyads. The probability matrix is obtained by $\mathbf{X} = Softmax(\mathbf{W}_s^T \mathbf{S} + m_s)$, where W_s and m_s are the weight and bias. $Softmax$ is utilised to limit the matrix in $[0,1]$. A symmetric operation is performed to ensure the symmetry of the interaction probability matrix $\mathbf{R}$ as follows:

$$\mathbf{R} = \frac{\mathbf{X} + \mathbf{X}^T}{2} \odot \mathbf{Mask} \tag{7}$$

Here, $\mathbf{Mask} = \mathbf{J} - \mathbf{I}$, where $\mathbf{J}$, $\mathbf{I}$ are all-ones matrix and identity matrix of order N respectively. $\odot$ represents Hadamard product to let the diagonal of $\mathbf{R}$ be 0.

Loss Function. Loss function is another core of our designed model. According to the group label of given data, a binary matrix $\mathbf{G} = (G_{ij})_{N\times N}$ is established as the ground truth with the entry G_{ij} indicating whether there exists an interaction between the i-th and the j-th person. To optimise our model comprehensively, we build two loss functions from two perspectives.

(1) From the individual perspective, we minimise the difference between the predicted interaction probability matrix and the ground truth. Assuming that the interactions of different pairs are independent, the likelihood of predicting correctly can be defined by:

$$P(\mathbf{G} \mid \mathbf{R}) = \prod_{G_{ij}=1} R_{ij} \prod_{G_{ij}=0} (1 - R_{ij}) \tag{8}$$

Subsequently, we can use the maximum likelihood estimation and convert it to the standard optimization form to derive the Dyad Loss:

$$loss_{dyad} = -\sum_{i,j} [G_{ij} log(R_{ij}) + (1 - G_{ij}) log(1 - R_{ij})] \tag{9}$$

(2) From the global perspective, it is expected to optimise the intra-group collectiveness and the inter-group discrimination among the predicted groups. We introduce the modularity [28, 29] which is primarily used in community detection to measure the quality of group division. Previous work [16, 50] also utilised the modularity to detect social networks or groups. However, it is usually used as a metric for unsupervised heuristic solutions. In this work, we modified it into a supervised metric named grouping quality, which reflects the distinguishability of different groups, as defined as follows:

$$Q_G = \frac{1}{2n} \sum_{i,j} (R_{ij} - \frac{k_i k_j}{2n}) G_{ij} \tag{10}$$

where $n = \sum_{i,j} R_{ij}$, k_i and k_j represent the degree of i-th and j-th persons in R_{ij}. The higher the degree of grouping quality is, the closer the interaction probability matrix is to the real situation. Since the standard form of optimisation is usually to minimise the objective function, we define the Group Loss as $loss_{group} = 1 - Q_G$. And the joint loss is defined as:

$$\mathcal{L} = loss_{dyad} + \beta loss_{group} \tag{11}$$

where $\beta > 0$ is a hyper-parameter.

4 Experiments

4.1 Experimental Setup

Dataset. Although there have been some datasets of group discovery, they do not apply to the problem investigated in this paper. **Coffee Break** [5] mainly contain images from surveillance videos. As a result, the shooting angles, shooting distances and scenarios change very little. Similarly, the **Volleyball** dataset [18] are also limited in scenarios. However, the actual interaction may take place in multiple scenarios. The scenes in **Structured Group Dataset** (SGD) [4], and **Collective Activity Dataset** (CAD) [3] are much more diverse, but the definition of groups in these datasets pays more attention to the similarity of actions, rather than intimate interactions. The dataset proposed in [49] better meets our task, but it is not publicly accessible. Therefore, a novel dataset named **Social Interaction Dataset** (SID) is reconstructed, including the b-boxes and group labels of each people in all images.

Table 1: The statistics of interaction scene in SID.

Scenes	Bus stop	Cafe	Classroom	Conference	Court	Library	Park	Slidewalk	Others
Pics	21	72	66	90	140	47	94	158	28
Groups	32	186	115	263	293	108	225	242	45
Cliques	68	295	146	403	628	182	318	482	58

In this dataset, 390 images were selected from SGD [4]. We relabeled these images, considering the groups with similar actions but without intimate interactions as non-interactive groups. Other 186 images were from CAD [3], and we also assigned group labels to each person. In addition, 140 images from NBA highlights were added to the dataset to provide more complex interactive scenes. For the annotation of data, people with obvious interaction relationship are regarded as a group, and specific interaction categories are not used, that is, only whether there is interaction relationship between people in the image is judged without being divided into specific categories. There are 716 images with 5453 persons in total, with an average of 7.616 persons per image, involving various scenes such as classrooms, basketball courts, and sidewalks. There are 1509 interaction groups (no single person) and 2580 cliques (with single person) in total. The interaction scenes are also changeable, as shown in the Tab 1. The number of persons in an image ranges from 3 to 27, making the interactions more abundant and group discovery more challenging. According to the crowd density, 424 of the images are sparse scenes (3-7 persons), 275 of the images are of medium density (8-17 persons), and the rest 17 images are dense scenes (more than 17 persons).

Implementation Details. Our experiments are implemented using PyTorch. For the pose estimation of each person, the pixels in the corresponding b-box are resized to 192×256 to obtain the keypoint heatmaps containing rich postural information. Following the common setting in pose estimation, the size of the heatmaps is set to 48×64, and the number of keypoints are set to 17. That is, $W_0 = 48$, $H_0 = 64$, and $C = 17$. The dimension of the personal features $\mathbf{h}$ is 1024, i.e. $dh = 1024$. For ϕ_Q, ϕ_K, and ϕ_V, $dim = 64$. The dimension of coding frequency K is 8. Also, for the mutual features $\mathbf{d}$, $dd = 64$. The two hyper-parameters λ and β are set to 1 and 10, respectively. For model training, the initial learning rate is 0.0005, and the learning rate is decreased to 0.1 of the initial value at epoch 60. The whole training process is completed within 150 epochs. All experiments are conducted on a single TITAN-X GPU. The threshold of determining interactions from the probability matrix is 0.1. We also adopt the transitivity oracle defined in [42]. For example, if both (A, B) and (B, C) are judged as interactive dyads, while (A, C) may not, the triplet of (A, B, C) will also be considered as an SIG.

Evaluation Metrics. Judging whether there exists interactions can be seen as a binary classification task. Therefore, we adopt precision, recall, F1-score, and accuracy that commonly used in classification as the indicator to evaluate the proposed method. We also use the receiver operating characteristic (ROC) curve and area under the curve (AUC) for comparison. To draw ROC, the prediction is the interaction probability matrix without diagonal derived by each method, and the ground truth is the actual interaction matrix without diagonal.

4.2 Comparison with Previous Methods

Competitors. Some recent works on interaction analysis and group activity recognition are selected as comparison algorithms. (1) ARG [43] uses the appearance features extracted from each person combined with the distance relations to construct a relation graph between each person. To adapt to our task, we directly treat this relation graph as the interaction matrix. (2) To facilitate the learning of interactive relations, we apply the Dyad Loss to supervised the resulting relation graph of ARG, setting up an enhanced baseline (referred to as "DL+ARG"). (3) JS [6] improves group activity recognition by adding social group labels. As our target is to detect SIGs from single images, the I3D branch [9] of JS is removed, which only works for video sequences. Instead, we apply an Inception v3 [35] to extract feature maps from images. (4) LAGNet [42] combines action recognition with interaction inference in a unified network. The predicted interaction relations are taken as the result of interaction group discovery. For the action labels required by these competitors, we utilise the six categories defined in CAD [3], and annotate the other actions as "others".

Results. As can be observed from the indicators listed in Tab. 2 and the ROC curves shown in Fig. 3, our model achieves substantial improvements against all the competitors. In terms of accuracy, the proposed PRN outperforms other methods by at least 9.68% with the guidance of social psychology. This suggests that the model based on the combination of pose features and distance features inspired by psychological knowledge has achieved better results. The

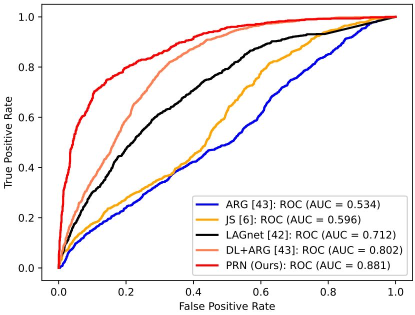


Figure 3: ROC curves of different methods.

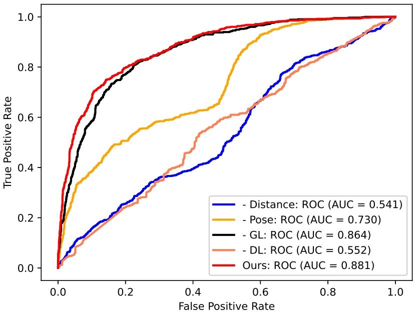


Figure 4: ROC curves of ablation study.

Table 2: Performances of different methods.

Method	P	R	F1	Acc
ARG[43]	41.87	54.14	47.22	69.83
JS[6]	40.02	61.62	48.53	67.41
LAGNet[42]	38.95	70.70	50.23	65.07
DL+ ARG[43]	47.09	73.57	57.43	72.81
PRN(Ours)	63.89	68.47	66.10	82.49

Table 3: Performances of ablation study.

Method	P	R	F1	Acc
PRN(Ours)	63.89	68.47	66.10	82.49
- Distance	26.31	42.36	32.46	56.05
- Pose	37.47	64.97	47.52	64.23
- DL	35.59	64.33	45.83	62.09
- GL	50.38	83.92	62.96	75.39

self-attention mechanism successfully integrates the personal and mutual influence, and can learn the interaction relations in an adaptive manner. Some detection results of PRN and the competitors are visualised in Fig. 5.

4.3 Ablation Study

To investigate the effects of the interaction elements and the loss functions, we carry out ablation studies by removing different components from our model at each experiment.

Comparison of different element combinations. PRN represents the key factors of determining interaction groups suggested by the social interaction field model in social psychology [51] with visual interaction elements. To validate the effectiveness of the combination of interaction elements, the branch involving the keypoint heatmaps (denoted as "Pose" in Tab. 3) and the branch involving the b-boxes (denoted as "Distance" in Tab. 3) are removed respectively. In other words, the heatmaps and the b-boxes are used to identify the interaction group separately.

Comparison of different loss combinations. We also conduct experiments to evaluate the effectiveness of the two loss functions. The Dyad Loss (denoted as "DL" in Tab. 3) pays more attention to the interactions between each pair from the individual perspective, while the Group Loss (denoted as "GL" in Tab. 3) pays more attention to the intra-group collectiveness and the inter-group discrimination from the global perspective. These two loss functions are disabled respectively, compared to the full model which employs a linear combination of them to measure their respective contributions.

The experimental results are shown in Tab. 3 and Fig. 4. An obvious conclusion is that

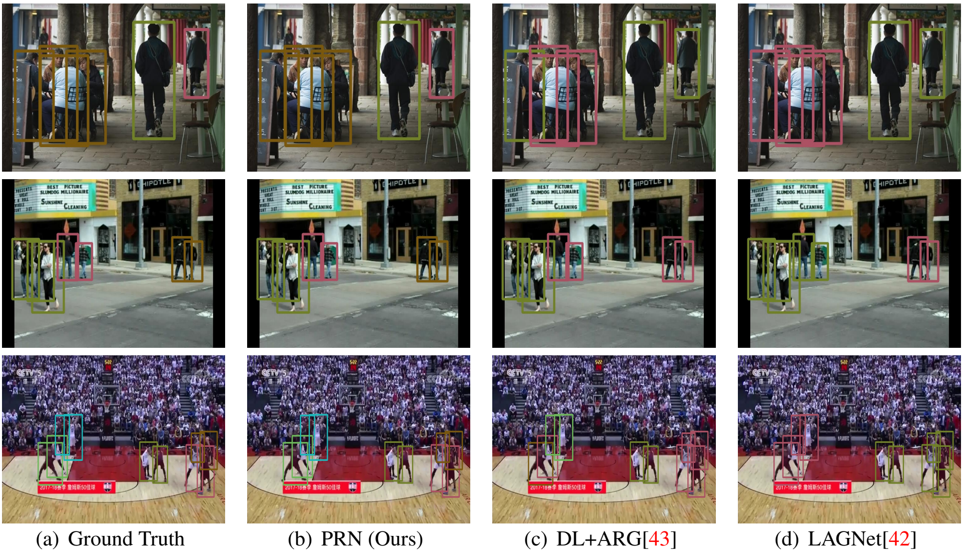

(a) Ground Truth (b) PRN (Ours) (c) DL+ARG[43] (d) LAGNet[42]

Figure 5: Visualisation of the predictions of each model. Different groups in each image are marked with different colours. From left to right, each column represents the ground truth, the prediction results of our model, and the results predicted by the DL + ARG and LAGNet.

combining the two interaction elements effectively improves the performance to recognise social interaction groups, which also coincides the finding of the social psychology studies [1, 11, 13, 38]. With only postural or positional features, the information is insufficient and the model tends to overlook the interactive relations between persons. By combining them, the pairwise interaction relations can be effectively understood in a more comprehensive manner, which is in line with the theory proposed in [51]. Also, combining both of the loss functions is more effective than using a single loss. These two loss functions constraint the PRN complementarily. Since the prior assumption of the Dyad Loss is that the interactions of different dyads are independent, it ignores the mutual influence inside each clique. The Group Loss can remedy this drawback by considering the grouping quality from the whole, and the grouping results better fit the ground truths.

5 Conclusion

In this paper, we propose the Psychology-inspired Relation Network (PRN) to detect social interaction groups in an end-to-end fashion under the guidance of recent advances in social psychology. Using the self-attention mechanism, PRN captures interactive information from the keypoint heatmaps and the bounding boxes of each person in an image, and measure the pairwise interaction relations to calculate the interaction probability of each dyad. Moreover, we present the Dyad Loss and the Group Loss to optimise PRN from complementary perspectives. In addition, a novel dataset is constructed to facilitate the research of social interaction group discovery, which involves diverse scenarios of various density. Extensive experiments and ablation studies have demonstrated the effectiveness of our method.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC, Grant No. 62171281), Science and Technology Commission of Shanghai Municipality (STCSM, Grant Nos. 19DZ1209303, 20DZ1200203, 2021SHZDZX0102), and SJTU Yitu/Thinkforce Joint Laboratory for Visual Computing and Application.

References

[1] Michael Argyle. *Bodily communication*. Routledge, 2013.

[2] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 215–230. Springer, 2012.

[3] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*, pages 1282–1289, 2009.

[4] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 417–433. Springer, 2014.

[5] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, volume 2, page 4. Citeseer, 2011.

[6] Mahsa Ehsanpour, Alireza Abedin, Fatemeh Saleh, Javen Shi, Ian Reid, and Hamid Rezatofighi. Joint learning of social groups, individuals action and sub-group activities in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 177–195, 2020.

[7] Yachuang Feng, Yuan Yuan, and Xiaoqiang Lu. Learning deep event models for crowd anomaly detection. *Neurocomputing*, 219:548–556, 2017.

[8] Sue Fletcher-Watson, John M Findlay, Susan R Leekam, and Valerie Benson. Rapid detection of person information in a naturalistic scene. *Perception*, 37(4):571–583, 2008.

[9] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[10] Edward Twitchell Hall. *The hidden dimension*, volume 609. Garden City, NY: Doubleday, 1966.

[11] Judith A Hall, Erik J Coats, and Lavonia Smith LeBeau. Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. *Psychological bulletin*, 131(6): 898, 2005.

[12] Leslie A Hayduk. Personal space: An evaluative and orienting overview. *Psychological bulletin*, 85(1):117, 1978.

[13] Leslie A Hayduk. The shape of personal space: An experimental investigation. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 13(1):87, 1981.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018.

[16] Hayley Hung and Ben Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 231–238, 2011.

[17] Mostafa S Ibrahim and Greg Mori. Hierarchical relational networks for group activity recognition and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*, pages 721–736, 2018.

[18] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016.

[19] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[20] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[21] Yu Kong, Yunde Jia, and Yun Fu. Interactive phrases: Semantic descriptionsfor human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(9):1775–1788, 2014.

[22] L'ubor Ladickỳ, Chris Russell, Pushmeet Kohli, and Philip HS Torr. Inference methods for crfs with co-occurrence statistics. *International journal of computer vision*, 103(2): 213–225, 2013.

[23] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on statistical learning in computer vision, ECCV*, pages 17–32, 2004.

[24] Shuheng Lin, Hua Yang, Xianchao Tang, Tianqi Shi, and Lin Chen. Social mil: Interaction-aware for crowd anomaly detection. In *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2019.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010.

[27] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[28] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[29] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[30] Mengshi Qi, Jie Qin, Annan Li, Yunhong Wang, Jiebo Luo, and Luc Van Gool. stagnet: An attentive semantic rnn for group activity recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 104–120, 2018.

[31] R.J. Rummel. *Understanding conflict and war.* Sage Publications, ISBN 0470745010., 1981.

[32] Francesco Setti, Oswald Lanz, Roberta Ferrario, Vittorio Murino, and Marco Cristani. Multi-scale f-formation discovery for group detection. In *2013 IEEE International Conference on Image Processing*, pages 3547–3551. IEEE, 2013.

[33] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation detection: Individuating free-standing conversational groups in images. *PloS one*, 10(5): e0123783, 2015.

[34] Hyun Soo Park and Jianbo Shi. Social saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[36] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014.

[37] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[38] Tanya Vacharkulksemsuk, Emily Reit, Poruz Khambatta, Paul W Eastwick, Eli J Finkel, and Dana R Carney. Dominant, open nonverbal displays are attractive at zero-acquaintance. *Proceedings of the National Academy of Sciences*, 113(15):4009–4014, 2016.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010. Curran Associates, Inc., 2017.

[40] Petar Velikovi, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[41] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[42] Zhenhua Wang, Jiajun Meng, Jin Zhou, Dongyan Guo, Guosheng Lin, Jianhua Zhang, Javen Qinfeng Shi, and Shengyong Chen. Lagnet: Logic-aware graph network for human interaction understanding. *arXiv preprint arXiv:2011.10250v2*, 2020.

[43] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9964–9974, 2019.

[44] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.

[45] Hao Xiao, Weiyao Lin, Bin Sheng, Ke Lu, Junchi Yan, Jingdong Wang, Errui Ding, Yihao Zhang, and Hongkai Xiong. Group re-identification: Leveraging and integrating multi-grain information. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 192–200, 2018.

[46] Qiling Xu, Hua Yang, Lin Chen, and Guangtao Zhai. Group re-identification with hybrid attention model and residual distance. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1217–1221. IEEE, 2019.

[47] Yichao Yan, Bingbing Ni, and Xiaokang Yang. Predicting human interaction via relative attention model. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3245–3251, 2017.

[48] Yichao Yan, Jie Qin, Bingbing Ni, Jiaxin Chen, Li Liu, Fan Zhu, Wei-Shi Zheng, Xiaokang Yang, and Ling Shao. Learning multi-attention context graph for group-based re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[49] Haanju Yoo, Taekyu Eom, Jeongmin Seo, and Sang-Il Choi. Detection of interacting groups based on geometric and social relations between individuals in an image. *Pattern Recognition*, 93:498–506, 2019.

[50] Ting Yu, Ser-Nam Lim, Kedar Patwardhan, and Nils Krahnstoever. Monitoring, recognizing and discovering social networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1462–1469. IEEE, 2009.

[51] Chen Zhou, Ming Han, Qi Liang, Yi-Fei Hu, and Shu-Guang Kuai. A social interaction field model accurately identifies static and dynamic social groupings. *Nature human behaviour*, 3(8):847–855, 2019.