BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 04/06 pp120–129
Volume 5, Number 2, June 2022
DOI: 10.26599/BDMA.2021.9020022

Understanding Social Relationships with Person-Pair Relations

Hang Zhao, Haicheng Chen, Leilai Li, and Hai Wan*

Abstract: Social relationship understanding infers existing social relationships among individuals in a given scenario, which has been demonstrated to have a wide range of practical value in reality. However, existing methods infer the social relationship of each person pair in isolation, without considering the context-aware information for person pairs in the same scenario. The context-aware information for person pairs exists extensively in reality, that is, the social relationships of different person pairs in a simple scenario are always related to each other. For instance, if most of the person pairs in a simple scenario have the same social relationship, "friends", then the other pairs have a high probability of being "friends" or other similar coarse-level relationships, such as "intimate". This context-aware information should thus be considered in social relationship understanding. Therefore, this paper proposes a novel end-to-end trainable Person-Pair Relation Network (PPRN), which is a GRU-based graph inference network, to first extract the visual and position information as the person-pair feature information, then enable it to transfer on a fully-connected social graph, and finally utilizes different aggregators to collect different kinds of person-pair information. Unlike existing methods, the method—with its message passing mechanism in the graph model—can infer the social relationship of each person-pair in a joint way (i.e., not in isolation). Extensive experiments on People In Social Context (PISC)- and People In Photo Album (PIPA)-relation datasets show the superiority of our method compared to other methods.

Key words: social relationship understanding; person-pair relations; Person-Pair Relation Network (PPRN)

1 Introduction

Social relationships in either physical or virtual forms are the basis of social networks in our daily lives[1–3]. Previous studies have shown that implicit social relationships can be discovered from texts[4], images[2, 5–8], and videos[9–12]. The current paper focuses on still images.

The goal of understanding social relationships is to infer the social relations among people in a given scenario, such as a still image. This process has been demonstrated to have a wide range of practical value in reality. Nowadays, with the increasing dependence of humans on machines, understanding social relationships enables the latter to blend in and make better responses in different situations. Furthermore, social relationship understanding is also helpful in avoiding potential privacy risks generated by automatically parsing information that may reveal such relationships in many forms of media (e.g., texts[4]) and informing users about this.

In past years, social relationship understanding has drawn increasing research interest. For example, Sun et al.[13] proposed an approach that uses the information of head regions, body regions, and human attributes to predict social relationships. Li et al.[5] proposed a dual-

• Hang Zhao is with the Guizhou Post and Telecommunications Planning and Design Institute Co., Ltd., Guiyang 550003, China. E-mail: 18985021970@189.cn.
• Haicheng Chen and Hai Wan are with School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China. E-mail: chenhch8@mail2.sysu.edu.cn; wanhai@mail.sysu.edu.cn.
• Leilai Li is with Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518049, China. E-mail: lileilai446@pingan.com.cn.
* To whom correspondence should be addressed.
Manuscript received: 2021-10-20; accepted: 2021-11-03

glance model that utilizes the visual information of a person-pair and the contextual information of region proposals around the pair to make predictions. Wang et al.[6] proposed a Graph Reasoning Model (GRM) that incorporates common sense knowledge of the correlation between objects and person-pairs.

However, existing methods infer the social relationship of each person pair in isolation, without taking into account the context-aware information for such pairs in the same scenario. In fact, context-aware information for person-pairs exists extensively in reality, that is, the social relationships of different person-pairs in a simple scenario are always related to each other. Taking Fig. 1a as an example, in a simple scenario, most of the person-pairs have the same social relation (friends), and although there is one different social relation (grandma-grandchild), these two relationships belong to one coarse-level relationship (intimate). Intuitively, if we want to infer the social relationships of a person-pair and already know several others are the same relationship (e.g., friends), then the person-pair has a high probability

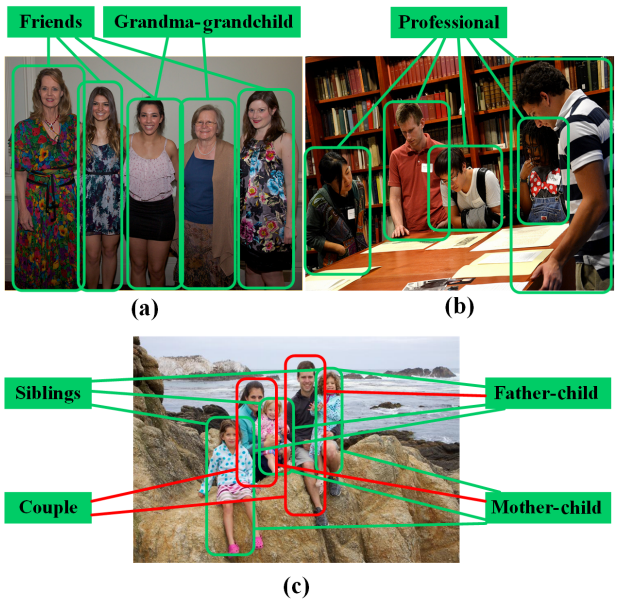


Fig. 1 Examples from PIPA-relation datasets[13]. Most person pairs in a simple scenario have the same or similar coarse-level social relations. For instance, almost all person-pairs in (a) have the same relation "friends" and all person-pairs in (b) have the same relation, "professional". Furthermore, all four social relations in (c) belong to the similar coarse-level social relation, "intimate". There are many examples in the dataset that have the same or similar social relationships in a simple scenario, and the detailed statistics can be found in Section 4.2.

of being "friends" or having the same coarse-level relationship. Moreover, we find that the context-aware information for person-pairs exists extensively in People In Social Context (PISC)- and People In Photo Album relation (PIPA)-relation datasets (see Section 4.2). This context-aware information for person-pairs should be taken into account in social relationship understanding. Therefore, this paper proposes a novel end-to-end trainable Person-Pair Relation Network (PPRN), a GRU-based graph inference network to first extract the visual and position information as the person-pair feature information, and then enable the information to transfer on a fully-connected social graph before finally utilizing various aggregators to collect different person-pair information. Unlike existing works, this method—with the message passing mechanism in the graph model—can infer the social relationship of each person-pair in a joint way (i.e., not in isolation).

To date, understanding the social relationship between two persons in an image remains a challenging task. First, we should design a mechanism to encode the context-aware information for person-pairs. Second, in previous methods, for a given person-pair, it is difficult to automatically detect its contextual regions correctly due to the lack of annotation information about the regions, which can be quite harmful when conducting inference. Therefore, whether to use the information of contextual regions has also become an important consideration among researchers in this field, including the authors of the current paper.

To address the above problems, we propose a novel end-to-end trainable PPRN. This network not only reduces the negative influence of the wrong information of contextual regions, but also encodes the context-aware information for person-pairs in the same scenario. PPRN can make inferences in social graphs and consists of three modules: feature extraction, message passing, and relation inference. The feature extraction module extracts the person-pair visual feature. First, for a given person-pair, we use one ResNet-101 to extract the visual features of each person and another ResNet-101[14] for the visual feature of the person-pair within an image. Next, we obtain the person-pair feature by concatenating these three visual features and their position information. Then, we use the person-pair feature to initialize the node in the social graph.

Meanwhile, the message passing module encodes the interaction cue among person-pair nodes on the social graph. Then, we utilize GRU[15], a recurrent neural

network, to pass messages among different person-pair nodes in an iterative way. To pool the messages from neighbor nodes in this module, we use three different pooling mechanisms. Finally, the relation inference module infers a social relation for a given person-pair. We use a Multi-Layer Perceptron (MLP) to calculate the probability distribution of all social relationships, and the resulting relationship with the largest probability is the final prediction result.

Our model is evaluated using two benchmark datasets, namely, PISC-[5] and PIPA-relation[13]. The experimental results verify the superiority of our model over previous methods. The contributions of this work to the literature are as follows:

• To the best of our knowledge, this is the first attempt to infer the social relationship of each person-pair in a joint way (i.e., in the field of social relationship understanding.

• We design a novel message passing mechanism to model the context-aware information for person-pairs in the same scenario, thereby achieving the best performance in this task compared to other methods.

• We analyze and verify that the role of information of contextual regions in the scenario is not that huge, which is a very novel idea in the field of social relationship understanding.

2 Related Work

In this section, we introduce the related works on social relationship understanding and message passing.

2.1 Social relationship understanding

The foundation of a social network is the understanding of social relationships, which is an important multidisciplinary problem that has attracted increasing attention among computer vision researchers in recent years. With the rise of deep learning, a large number of studies have emerged to detect social relations from texts[4], images[5–8], and videos[10–12]. For instance, motivated by psychological studies, Dibeklioglu et al.[16] and Zhang et al.[8] investigated social relationships based on affective behavior analysis and facial attributes, such as expression and head pose. Li et al.[5] first introduced the contextual regions detected by Faster RCNN[17] in this task. Then, they proposed a dual-glance model for social relationships, in which the first glance makes a coarse relationship prediction for a given person-pair, and then the second one refines the prediction by utilizing the regions surrounding the pair. Wang et al.[6] constructed a semantic-aware knowledge graph by detecting objects and used a Gated Graph Neural Network (GGNN)[18] to integrate the graph into the Graph Reasoning Model (GRM). This is a graph reasoning network wherein a proper message propagation and graph attention mechanism were both introduced to explore person-pair interactions and the contextual objects .

Unlike the aforementioned methods that infer the social relationship of each person-pair in isolation, our proposed PPRN approach can infer the relationship in a joint way. To the best of our knowledge, it is also the first to introduce context-aware information for person-pairs and to feature a message-passing mechanism to encode the information.

2.2 Message passing

The message passing mechanism provides a way by which to update nodes' states and exchange information by passing messages to their neighbor nodes until they reach a stable equilibrium. This has already been successfully applied in many graph inference tasks. Before the rise of message passing, the use of Conditional Random Fields (CRFs)[19] is a common method used in graph inference tasks. For example, Johnson et al.[20] used CRF to infer scenario graph grounding distributions for image retrieval. Yatskar et al.[21] used a deep CRF model to propose situation-driven object and action prediction, while Xu et al.[22] used the GRU-RNNs to solve the scenario graph generation problem iteratively. Our work is related to Graph-LSTM[23] and the work of Xu et al.[22], who formulated the message passing problem using RNN models. The same study in Ref. [22] also designed a primal graph and a dual graph in their model. In comparison, we merely simplified the model and just used one graph to pool social relationship messages, thereby achieving a better result. Similar to that study in Ref. [22], our model iteratively refines the social relationship predictions through relationship message passing in the graph, whereas the structural RNN model only makes one-time prediction along the temporal dimension; thus, it is unable to refine its past predictions[22].

3 PPRN Model

We introduce in this section the proposed PPRN for social relationship understanding. We formulate the social relationships among persons in an image as a social graph, in which each node denotes the relationship

of a person-pair. The framework consists of three modules, namely, feature extraction, message passing, and relation inference. In the first module, following the approach in GRM[6], we first extract the person-pair visual features from a given image, and then employ these features to initialize the nodes in the graph. In the second module, we use GRU[15] and an attention mechanism to explore the interaction cues of multiple person-pair relationships in the graph. Finally, we output the prediction result in the third module. Our proposed framework is shown in Fig. 2.

3.1 Task formulation

Let I denote an image and $B = \{(x_i^0, y_i^0; x_i^1, y_i^1) : i = 1, 2, \dots\}$ denote the bounding boxes of all persons in I, where (x_i^0, y_i^0) and (x_i^1, y_i^1) are the upper left and lower right coordinates of the i-th person's box, respectively. Let $G = \{V, E\}$ denote the social graph, where V is the set of all person-pair nodes, E is the edge set of any two nodes in I, and G is a fully-connected undirected graph. The task is to infer the social relationship between any two persons in I, which we formulate as a graph inference task.

3.2 Feature extraction module

Given the i-th and j-th persons in I, we crop three patches p_i, p_j, and p_{ij}, with the first two patches covering each person and the last one covering their union region. These patches are resized to 224 × 224 pixels and fed into Convolution Neural Networks (CNNs) accordingly. The feature map of the last convolutional layer of the CNNs is flattened to obtain visual vectors $\boldsymbol{p}_i$, $\boldsymbol{p}_j$, and $\boldsymbol{p}_{ij}$, where $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ share the same CNN, and $\boldsymbol{p}_{ij}$ uses another CNN. Moreover, the geometry feature $\boldsymbol{b}_{ij}$ of the i-th and j-th persons is complementary to the visual information. This is because it is not easy to infer by only using the information when the social relationship is "no relation", as stated below:

$$\begin{aligned}
\boldsymbol{b}_{ij} &= [x_{ij}^{min}, x_{ij}^{max}, y_{ij}^{min}, y_{ij}^{max}, area_{ij}], \\
x_{ij}^{min} &= \min\{x_i^0, x_j^0\}, \\
x_{ij}^{max} &= \max\{x_i^1, x_j^1\}, \\
y_{ij}^{min} &= \min\{y_i^0, y_j^0\}, \\
y_{ij}^{max} &= \max\{y_i^1, y_j^1\}, \\
area_{ij} &= (x_{ij}^{max} - x_{ij}^{min}) \times (y_{ij}^{max} - y_{ij}^{min}) \qquad (1)
\end{aligned}$$

where $[\cdot]$ is the concatenation operation, and $\boldsymbol{p}_i$, $\boldsymbol{p}_j$, and $\boldsymbol{p}_{ij}$ are concatenated and fed into an MLP to produce a person-pair visual feature vector. The resulting vector is given by

$$\boldsymbol{v}_{ij} = \text{MLP}([\boldsymbol{p}_i, \boldsymbol{p}_j, \boldsymbol{p}_{ij}, \boldsymbol{b}_{ij}]) \qquad (2)$$

where $\text{MLP}(\cdot) = \text{Linear}(\text{ReLU}(\text{Linear}(\cdot)))$. $\boldsymbol{v}_{ij} \in R^{4096}$ is used to initialize the node in the social graph. It is worth mentioning that, the CNNs is the ResNet-101 in our framework[14].

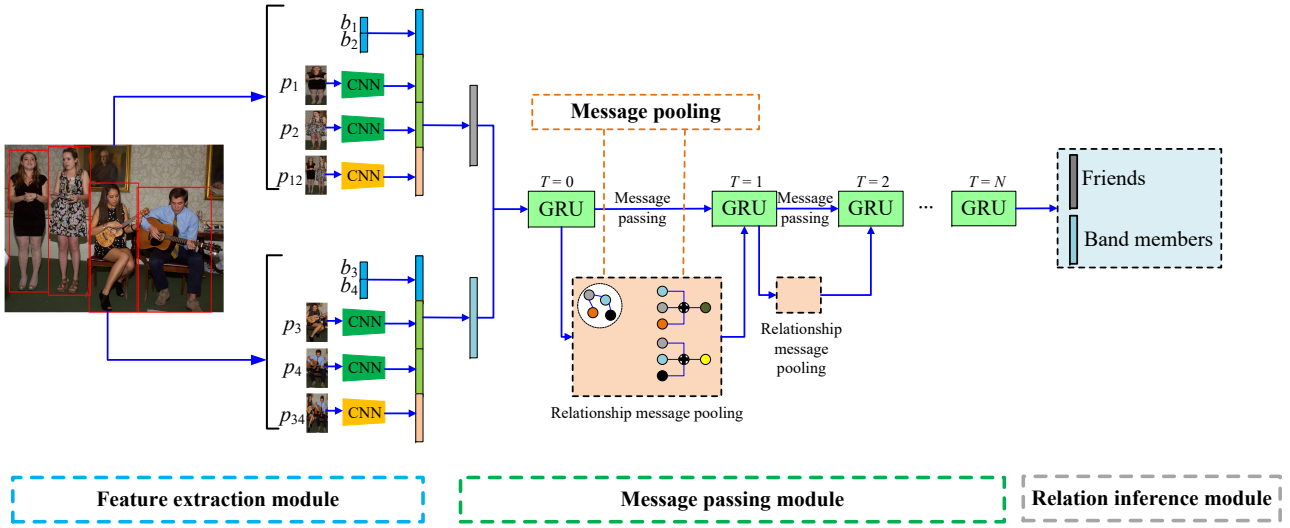


Fig. 2 Architecture of our proposed PPRN. In the feature extraction module, the model first extracts the features of each person and person-pairs. p_i denotes the region of the i-th person, p_{ij} denotes the union region of the i-th and j-th persons, and b_i denotes the position of the i-th person. In the message passing module, an attention pooling method is used to propagate the relationship messages among nodes in the graph (the learned weighted sum operation is denoted by the ⊕ symbol). Moreover, we iteratively update the hidden states of nodes by using the GRUs. In the relation inference module, we infer the relation of the nodes by using the hidden state of each node at the last step.

3.3 Message passing module

This module is used to encode the context-aware person-pair information to facilitate social relationship understanding. Specifically, we use GRU[15] with a pooling mechanism to pass information among person-pair nodes in the social graph. GRU is an RNN with GRUs[15]. Similar to GGNN[24], messages from each node in the graph are iteratively propagated across the graph as follows:

$$\begin{aligned}\boldsymbol{r}_v^t &= \sigma(\boldsymbol{W}_r[\boldsymbol{h}_v^{(t-1)}, \boldsymbol{x}_v^{(t)}]),\\ z_v^t &= \sigma(\boldsymbol{W}_z[\boldsymbol{h}_v^{(t-1)}, \boldsymbol{x}_v^{(t)}]),\\ \hat{\boldsymbol{h}}_v^{(t)} &= \tanh(\boldsymbol{W}[\boldsymbol{r}_v^t \odot \boldsymbol{h}_v^{(t-1)}, \boldsymbol{x}_v^{(t)}]),\\ \boldsymbol{h}_v^{(t)} &= (1-\boldsymbol{z}_t)\odot \boldsymbol{h}_v^{(t-1)} + \boldsymbol{z}_v^t \odot \hat{\boldsymbol{h}}_v^{(t)}\end{aligned} \tag{3}$$

where $\sigma(\,)$ and $\tanh(\,)$ are the logistic sigmoid and hyperbolic tangent functions, respectively; $\odot$ denotes the element-wise multiplication operation; $\boldsymbol{r}_v^t$ and $\boldsymbol{z}_v^t$ denote the reset gate and update gate of node $\boldsymbol{v}$ at the t-th step, respectively; $\boldsymbol{W}_r$, $\boldsymbol{W}_z$, and $\boldsymbol{W}$ are the trainable parameters; $\boldsymbol{h}_v^{(t)}$ is the hidden state of $\boldsymbol{v}$ at the t-th step; and $\boldsymbol{h}_v^{(0)}$ is initialized as the person-pair visual feature $\boldsymbol{v}$ from Eq. (2).

Meanwhile, $x_v^{(t)}$ in Eq. (3) is the new state of node $\boldsymbol{v}$ obtained by pooling the message from neighbor nodes of $\boldsymbol{v}$. Here we employ three different pooling mechanisms (Attention, Max, and Mean) to encode the context-aware information in the message pooing module.

3.3.1 Attention aggregator

Here, we use the representation of the node $\boldsymbol{v}$ to attend the hidden states of its neighbor nodes and obtain the aggregated state $\boldsymbol{x}^{(t)}$, which is given as follows:

$$\boldsymbol{x}_v^{(t)} = \sum_{v'\in N(\boldsymbol{v})} \psi(\boldsymbol{w}^{\mathrm{T}}[\boldsymbol{h}_v^{(t-1)}, \boldsymbol{h}_{v'}^{(t-1)}])\boldsymbol{h}_{v'}^{(t-1)} \tag{4}$$

where $\boldsymbol{w}$ is the trainable parameter, $N(\boldsymbol{v})$ is the neighbor nodes of $\boldsymbol{v}$, and $\psi(\cdot)$ is the attention weight between $\boldsymbol{v}$ and another node $\boldsymbol{v'}$.

3.3.2 Max aggregator

The Max aggregator performs the element-wise Max operation among hidden states, and is expressed as follows:

$$\boldsymbol{x}_v^{(t)} = \mathrm{Max}(h_{v_0}^{(t-1)}, h_{v_1}^{(t-1)}, \dots) \tag{5}$$

where $h_{v_i}^{(t-1)} \in N(\boldsymbol{v})$.

3.3.3 Mean aggregator

The mean aggregator performs the element-wise Mean operation among hidden states, and which is expressed as follows:

$$\boldsymbol{x}_v^{(t)} = \mathrm{Mean}(h_{v_0}^{(t-1)}, h_{v_1}^{(t-1)}, \dots) \tag{6}$$

where $h_{v_i}^{(t-1)} \in N(\boldsymbol{v})$.

3.4 Relation inference module

This module is an MLP. Its input is the hidden state of the nodes at the last step T, and its output is the probability distribution of all social relationships. It is formulated as follows:

$$\boldsymbol{s}_v = \mathrm{Softmax}(\mathrm{MLP}(\boldsymbol{h}_v^{(T)})) \tag{7}$$

where $\mathrm{MLP}(\cdot) = \mathrm{Linear}(\mathrm{ReLU}(\mathrm{Linear}(\cdot)))$, $\boldsymbol{s}_v \in R^{\mathcal{C}}$, and $\mathcal{C}$ represents the number of social relationships.

3.5 Optimization

We use the cross entropy loss function to optimize our model. The objective function is expressed as follows:

$$\mathcal{L} = -\frac{1}{\sum\limits_{I\in\mathcal{I}}|V_I|}\sum_{I\in\mathcal{I}}\sum_{v\in V_I}\mathrm{CrossEntropy}(y_v, \boldsymbol{s}_v) \tag{8}$$

where $\mathcal{I}$ is the image set. V_I is the node set of the social graph in image I, $|V_I|$ is the size of V_I. y_v is the ground-true label of node $\boldsymbol{v}$ and $\boldsymbol{s}_v$ is obtained based on Eq. (7).

4 Experiment

4.1 Experiment setting

4.1.1 Dataset

In this work, two datasets are used to evaluate our proposed method. The first one is the large-scale PISC[5], which has 22 670 images and contains two-level recognition tasks: **three coarse-level relationships**, namely, "no relation, intimate relation, and none-intimate relation", and **six fine-level relationships**, namely, "friend, family, couple, professional, commercial, and no relation". The second one is the PIPA-relation[13], an extended version of PIPA[25] with 37 107 images. It also annotates 26 915 person-pairs on two-level recognition tasks: **five social domains** and **16 social relations** based on these domains. The train/val/test in PISC involves 13 142/4000/4000 images with 14 536/25 636/15 497 person-pairs on the coarse-level relationship, and 16 828/500/1250 images with 55 400/1505/3691 person-pairs on the fine-level relationship. In PIPA-relation, we follow a previous study in Ref. [6] and focus on recognizing its 16 relationships in the experiment. Its train/val/test involves 13 729/709/5106 person-pairs.

4.1.2 Implementation details

We adopted the same strategy as that in previous works[5, 6] to train our model. First, we used the SGD optimizer with learning rate 0.0001 to fine-tune the

ResNet-101 model[14]. Next, we froze the feature extraction module and used the ADAM optimizer with a learning rate of 0.0001 to train the last two modules jointly. For the message passing module, the dimension of hidden size was 512 and the iteration times, T, was 4.

4.2 Datasets analysis

Based on the data statistic shown in Table 1, we have two conclusions about the social relationship understanding task. First, according to the columns of "Multiple person-pairs" and "Single person-pairs", there is always more than one person-pair in most images. Second, based on the "Single relations" and "Multi relations" columns, most person-pair social relationships in a simple scenario are always the same or similar. These conclusions suggest that the interaction cue of person-pairs always exists in the task.

4.3 Comparisons with state-of-the-art methods

We compared our proposed model with existing state-of-the-art methods on both the PISC- and PIPA-relation datasets. Formally, results of the comparison are as follows:

4.3.1 Performance on the PISC dataset

We compared our method with six existing methods on this dataset.

Union-CNN[26]: This uses a single CNN model to classify the union region of the individual pair of interest, following the predicate prediction model in Ref. [26].

Pair-CNN[5]: This consists of two equivalent CNNs with shared weights to extract features of the cropped image patches for two individuals.

Pair-CNN+BBox+Union[5]: Based on pair-CNN and Union-CNN, this incorporates the spatial location information of two bounding boxes as supplementary information.

Pair-CNN+BBox+Global[5]: Instead of the union region, this uses the entire image as the input to Union-CNN.

Dual-glance[5]: It implements coarse and fine predictions, which include three and six relationships, respectively. Dual-glance employs Pair-CNN+BBox+Union and refines the prediction by utilizing the surrounding region proposal.

GRM[6]: It proposes a graph reasoning model that unifies the frequency of co-occurrence of each relationship-object pair to facilitate social relationship understanding.

Similar to GRM, we adopted the per-class recall and mAP to evaluate our model. The experimental results are shown in Table 2. To extract the local contextual cues (object proposal), Pair-CNN+BBox+Global, dual-glance, and GRM use extra Faster-RCNN[17], which is pre-trained on the COCO dataset[27]. GRM utilizes the object proposal to construct a semantic-aware knowledge graph for the social relationship reasoning. Notably, both

Table 1 Statistics on PISC- and PIPA-relation. "Multiple person-pairs" is the proportion of the images with more than one person-pair, while "Single person-pairs" is the proportion of images with just one person-pair. "Single relations" is the proportion of the images that have only one category of social relationship, while "Multi relations" is the proportion that has more than one category of social relationship.

(%)

Dataset		Multi person-pairs	Single person-pairs	Single relations	Multi relations
PISC	Coarse	87.1	12.9	79.9	20.1
	Fine	83.9	16.1	86.4	13.6
PIPA-relation 16		71.9	28.1	94.9	5.1

Table 2 Comparison of our PPRN and previous methods on recall-per-class and mean Average Precision (mAP) evaluation in PISC.

(%)

Method	Coarse relationship				Fine relationship						
	Intimate	Non-intimate	No relation	mAP	Friends	Family	Couple	Professional	Commercial	No relation	mAP
Union-CNN[26]	72.1	81.8	19.2	58.4	29.9	58.5	70.7	55.4	43.0	19.6	43.5
Pair-CNN[5]	70.3	80.5	38.8	65.1	30.2	59.1	69.4	57.5	41.9	34.2	48.2
Pair-CNN+ BBox+Union[5]	71.1	81.2	57.9	72.2	32.5	62.1	73.9	61.4	46.0	52.1	56.9
Pair-CNN+ BBox+Global[5]	70.5	80.0	53.7	70.5	32.2	61.7	72.6	60.8	44.3	51.0	54.6
Dual-glance[5]	73.1	**84.2**	59.6	79.7	35.4	**68.1**	76.3	70.3	57.6	60.9	63.2
GRM[6]	81.7	73.4	65.5	**82.8**	59.6	64.4	**58.6**	76.6	39.5	67.7	68.7
Ours	**81.9**	67.3	**74.7**	81.8	**61.0**	67.1	56.2	**76.9**	**46.0**	**68.1**	**69.7**

of them incur extra detection annotations containing noises. Specifically, our model achieves an accuracy rate of 75.1% and mAP of 81.8% in the coarse-level recognition. The model also achieves an accuracy rate of 65.6% and mAP of 69.7% in the fine-level recognition. Our model outperforms the previous best model in terms of fine-level recognition, but shows slightly worse performance on mAP than the best model in coarse-level recognition.

4.3.2 Performance on the PIPA-relation dataset

On this dataset, we compared our proposed model with two-stream CNN[13], dual-glance[5], and GRM[6]. Before our method was proposed, GRM achieved the best performance among these existing methods. Specifically, we directly reprinted the experimental results of several baselines from the GRM and the results are shown in Table 3. Notably, our PPRN significantly outperforms the previous methods and beats the best of them with a 2.4% accuracy when used on the PIPA-relation dataset.

4.4 Analysis of the experimental results

In this section, we first implemented different pooling mechanisms in the message passing module and compared their performance on the PIPA-relation dataset. Then, we conducted a conditional experiment to investigate the effectiveness of the factor of contextual object regions and the context-aware information for person-pairs. The experimental results are presented in Table 4.

Table 3 Comparison of our PPRN and previous methods in terms of accuracy evaluation in PIPA-relation.

(%)

Method	Accuracy
Two stream CNN[25]	57.2
Dual-glance[5]	59.6
GRM[6]	62.3
Ours	**64.7**

Table 4 mAP and accuracy result of RCNN, our model, and our model with contextual region, implemented in the same way as the dual-glance approach.

(%)

Method	PISC coarse		PISC fine	
	Accuracy	mAP	Accuracy	mAP
RCNN	–	63.5	–	48.4
Ours (Max)	74.3	80.8	64.1	68.3
Ours (Mean)	74.6	80.1	63.8	68.3
Ours (Attention)	**75.1**	**81.8**	**65.7**	**69.7**
Ours (Attention) + objects region	74.9	81.2	65.3	69.1

4.4.1 Significance of message passing

The core component in our proposed method is the introduction of the message passing mechanism, and one of its key components is the message pooling mechanism, which uses the attention-pooling mechanism to aggregate hidden states of other relationship nodes into an aggregated representation. We also implemented and evaluated other standard pooling mechanisms in our model to further investigate the advantage of the attention-pooling mechanism in our proposed method. First, we used the mean-pooling mechanism to replace the attention-pooling mechanism. Then, we use the max-pooling mechanism to replace it. As shown in Table 4, we can see that all scores with the attention mechanism increased by over 1% compared to the mean-pooling and the max-pooling mechanisms. This suggests that the attention-pooling mechanism has better performance in terms of aggregating the context-aware information into an aggregated representation.

4.4.2 Analysis of contextual information

Dual-glance and GRM spouse the idea that the contextual information of the object region plays an important role in social relationship understanding because of the co-occurrence of object regions and social relations. However, this is only suitable for the scenario in which only a single social relationship can be found in an image. This is because contextual information can only select the most relevant image when there are various social relationships in one image. Moreover, the object regions are always misidentified due to the defects of the current object detection model and the complexity of the image scenario, which may provide misinformation in the process of inferring the social relationship. Therefore, we believe it is harmful to incorporate the contextual information of the object region into our method, which is also proven in Table 4. Finally, we compared the performance of our method with and without object regions and found that it performed worse than the other methods when using the regions.

4.5 Case study

Four examples in Fig. 3 are shown to illustrate the capability of our PPRN to infer social relationships. We compared our PPRN with GRM[6] in these examples. The results reveal that compared to the social relation graphs in Fig. 3c, the graphs in Fig. 3b are very similar to those in Fig. 3a, which means that our PPRN performs better than GRM. In addition, similar to Fig. 3a, over

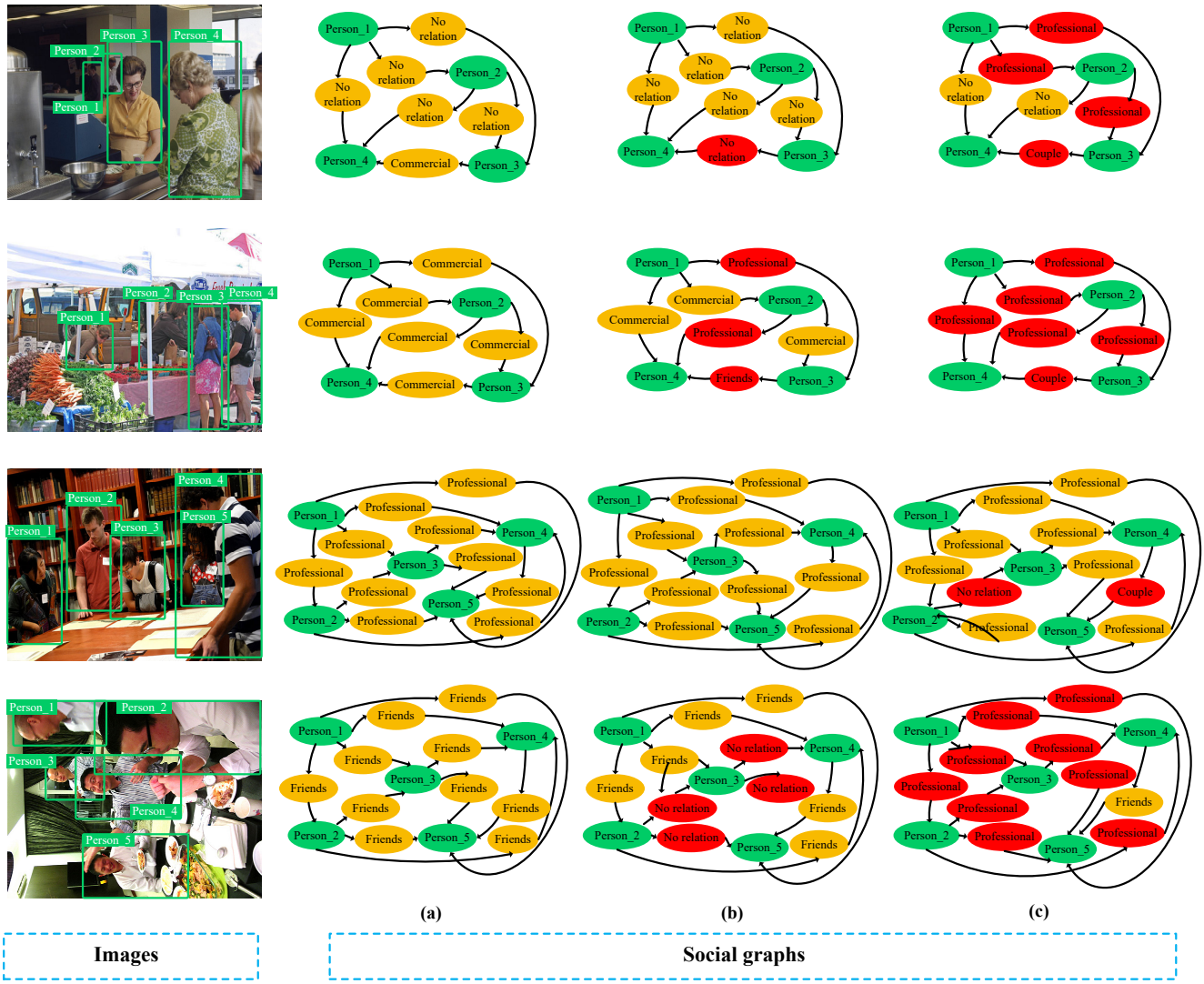


Fig. 3 Qualitative examples of social relation graphs from different sources: (a) PISC fine, (b) our proposed PPRN, and (c) GRM. The red oval denotes the social relationship that is wrongly predicted. These examples show that our method performs better than GRM.

half of the edges in Fig. 3b are identical, which strongly suggests that the context-aware information for person-pairs plays a significant role in social relationship understanding. For the third example, the accuracy of PPRN is 100%, while that for GM is 80% in GRM, thus proving the superiority of our method over GRM.

5 Conclusion

In this study, we propose a PPRN method to solve social relationship understanding in an image. Our proposed model is a novel graph inference network that incorporates context-aware information for person-pairs to infer the relations from these pairs. The key challenge is to design a model to explain the interactions among social relationships. The PPRN features a message-passing mechanism designed to propagate relationship messages through the RNNs, which can help improve the performance of relationship prediction. Moreover, we also analyzed the influence of contextual relationships and contextual object regions, and found the problem of information of contextual object regions. To our best knowledge, this is the first attempt to infer the social relationship of each person-pair in a joint way (i.e., not in isolation) to improve social relationship understanding. The results of extensive experiments on two large-scale benchmarks (PISC- and PIPA-relation) prove the superiority of our method over previous methods.

Our future work will incorporate external knowledge into our method for further improvement.

Acknowledgment

This paper was supported by the National Natural Science

Foundation of China (Nos. 61976232 and 51978675), Humanities and Social Science Research Project of Ministry of Education (No. 18YJCZH006), and All-China Federation of Returned Overseas Chinese Research Project (No. 17BZQK216).

References

[1] L. Yang, X. Wang, and M. M. M. Luo, Trust and closeness: A mixed method for understanding the relationship of social network users, *J. Int. Technol. Inf. Manage.*, vol. 30, no. 1, p. 4, 2021.

[2] J. Mou, W. L. Zhu, M. Benyoucef, and J. Kim, Understanding the relationship between social media use and Depression: A systematic review, in *Proc. of the 26*th *Americas Conf. on Information Systems, Virtual Conference*, 2020, p. 15.

[3] C. X. R. Shen, Z. C. Lu, T. Faas, and D. Wigdor, The labor of fun: Understanding the social relationships between Gamers and paid gaming teammates in China, in *Proc. 2021 CHI Conf. on Human Factors in Computing Systems* (*CHI*), Yokohama, Japan, 2021, p. 140.

[4] N. Fairclough, *Analysing Discourse*: *Textual Analysis for Social Research*. London, UK: Routledge, 2003.

[5] J. N. Li, Y. K. Wong, Q. Zhao, and M. S. Kankanhalli, Dual-glance model for deciphering social relationships, in *Proc. IEEE Int. Conf. on Computer Vision* (*ICCV*), Venice, Italy, 2017, pp. 2669–2678.

[6] Z. X. Wang, T. S. Chen, J. Ren, W. H. Yu, H. Cheng, and L. Lin, Deep reasoning with knowledge graph for social relationship understanding, in *Proc. 27th Int. Joint Conf. on Artificial Intelligence* (*IJCAI*), Stockholm, Sweden, 2018, pp. 1021–1028.

[7] G. Wang, A. Gallagher, J. B. Luo, and D. Forsyth, Seeing people in social context: Recognizing people and social relationships, in *Proc. 11th European Conf. on Computer Vision* (*ECCV*), Heraklion, Greece, 2010, pp. 169–182.

[8] Z. P. Zhang, P. Luo, C. C. Loy, and X. O. Tang, Learning social relation traits from face images, in *Proc. IEEE Int. Conf. on Computer Vision* (*ICCV*), Santiago, Chile, 2015, pp. 3631–3639.

[9] P. L. Dai, J. N. Lv, and B. Wu, Two-stage model for social relationship understanding from videos, in *Proc. 2019 IEEE Int. Conf. on Multimedia and Expo* (*ICME*), Shanghai, China, 2019, pp. 1132–1137.

[10] V. Ramanathan, B. P. Yao, and L. Fei-Fei, Social role discovery in human events, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Portland, OR, USA, 2013, pp. 2475–2482.

[11] L. Ding and A. Yilmaz, Learning relations among movie characters: A social network perspective, in *Proc. 11th European Conf. on Computer Vision* (*ECCV*), Heraklion, Greece, 2010, pp. 410–423.

[12] A. Vinciarelli, M. Pantic, and H. Bourlard, Social signal processing: Survey of an emerging domain, *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[13] Q. R. Sun, B. Schiele, and M. Fritz, A domain based approach to social relation recognition, in *Proc. IEEE. Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 435–444.

[14] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[15] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, in *Proc. SSST-8, 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, 2014, pp. 103–111.

[16] H. Dibeklioglu, A. A. Salah, and T. Gevers, Like father, like son: Facial expression dynamics for kinship verification, in *Proc. IEEE Int. Conf. on Computer Vision* (*ICCV*), Sydney, Australia, 2013, pp. 1497–1504.

[17] S. Q. Ren, K. M. He, R. B. Girshick, and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in *Proc. 28th Int. Conf. Advances in Neural Information Processing Systems*, Montreal, Canada, 2015, pp. 91–99.

[18] L. J. Li, D. A. Shamma, X. N. Kong, S. Jafarpour, R. Van Zwol, and X. H. Wang, CelebrityNet: A social network constructed from large-scale online celebrity images. *ACM Trans. Multimed. Comput., Commun. Appl.*, vol. 12, no. 1, p. 3, 2015.

[19] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Z. Su, D. L. Du, C. Huang, and P. H. S. Torr, Conditional random fields as recurrent neural networks, in *Proc. IEEE Int. Conf. on Computer Vision*, Santiago, Chile, 2015, pp. 1529–1537.

[20] J. Johnson, R. Krishna, M. Stark, L. J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, Image retrieval using scene graphs, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, USA, 2015, pp. 3668–3678.

[21] M. Yatskar, L. Zettlemoyer, and A. Farhadi, Situation recognition: Visual semantic role labeling for image understanding, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, 2016, pp. 5534–5542.

[22] D. F. Xu, Y. K. Zhu, C. B. Choy, and L. Fei-Fei, Scene graph generation by iterative message passing, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 2017, pp. 3097–3106.

[23] X. D. Liang, X. H. Shen, J. S. Feng, L. Lin, and S. C. Yan, Semantic object parsing with graph LSTM, in *Proc. 14thEuropean Conf. on Computer Vision* (*ECCV*), Amsterdam, The Netherlands, 2016, pp. 125–143.

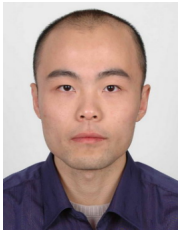
[24] Y. J. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel, Gated graph sequence neural networks, in *Proc. 4th Int. Conf. on Learning Representations*, San Juan, Puerto Rico, 2018, pp. 273–283.

[25] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, Beyond frontal faces: Improving person recognition using multiple cues, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Boston, MA, USA, 2015, pp. 4804–4813.

[26] C. W. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, Visual

relationship detection with language priors, in *Proc. 14th European Conf. on Computer Vision* (*ECCV*), Amsterdam, The Netherlands, 2016, pp. 852–869.

[27] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, Microsoft COCO: Common objects in context, in *Proc. 13th European Conf. on Computer Vision* (*ECCV*), Zurich, Switzerland, 2014, pp. 740–755.



Zhao Hang received the MEng degree in software engineering from Sun Yat-sen University, China in 2009. He is currently the deputy general manager of Guizhou Posts & Telecommunications Planning & Design Institute Co., Ltd. His current research interests include knowledge graph, scene graph, and artificial intelligence applications.

Haicheng Chen received the BEng degree in software engineering from Sun Yat-sen University, China in 2018. He is currently a master student at the School of Computer Science and Engineering, Sun Yat-sen University. His research focuses on natural language processing and reinforcement learning network.

Leilai Li received the MEng degree in software engineering from Sun Yat-sen University, China in 2019. He is currently a software engineer at Ping An Technology (Shenzhen) Co., Ltd. His current research interests include natural language processing and knowledge graph.

Hai Wan received the PhD degree in computer software and theory from Sun Yat-sen University, China in 2006. He is the director of the Department of Software Engineering, a PhD supervisor, and an associate professor at the School of Computer Science and Engineering, Sun Yat-sen University. His research focuses on natural language processing, as well as knowledge representation and reasoning.