Enhancing Social Relation Inference with Concise Interaction Graph and Discriminative Scene Representation

Anonymous ICCV submission

Paper ID 8211

Abstract

*There has been a recent surge of research interest in attacking the problem of social relation inference based on images. Existing works classify social relations mainly by creating complicated graphs of human interactions, or learning the foreground and/or background information of persons and objects, but ignore holistic scene context. The holistic scene refers to the functionality of a place in images, such as dinning room, playground and office. In this paper, by mimicking human understanding on images, we propose an approach of **PR**actical **I**nference in **S**ocial r**E**lation (PRISE), which concisely learns interactive features of persons and discriminative features of holistic scenes. Technically, we develop a simple and fast relational graph convolutional network to capture interactive features of all persons in one image. To learn the holistic scene feature, we elaborately design a contrastive learning task based on image scene classification. To further boost the performance in social relation inference, we collect and distribute a new large-scale dataset, which consists of about 240 thousand unlabeled images. The extensive experimental results show that our novel learning framework significantly beats the state-of-the-art methods, e.g., PRISE achieves 6.8% improvement for domain classification in PIPA dataset.*

1. Introduction

Social relations describe the connections among two or more individuals, which are fundamental to daily life of human beings [17]. Nowadays, billions of people share images in social media platforms such as Facebook and Twitter. In light of [3], common social relations include family, couple, friends, colleagues, professional, etc. There has been an increasing interest in understanding social relations among persons from still images due to the broad applications including group behavior analysis [13], image caption generation [15] and human trajectory prediction [1].

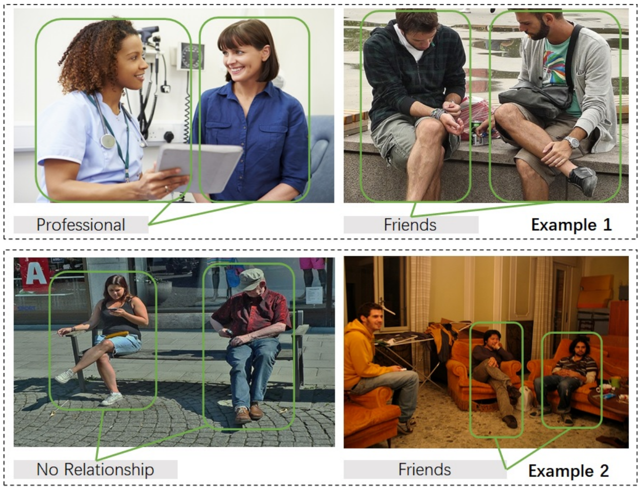

Figure 1. The comparison of inferring social relations under different scenes, with images taken from PISC dataset [19]. Example 1 shows relations of professional and friends corresponding to hospital and park, respectively. Example 2 significantly implies a close relation in the context of staying indoors.

The problem of social relation inference is challenging and complicated because it requires high-level semantic understanding of images. Inspired by the cognition process of human beings on images, we summarize three steps for classification of social relations. First, we take a whole view of the scene, which represents the functionality of a place. Second, we identify the background objects and persons, and the foreground union regions of person pair in images. Third, we observe the interaction of persons, such as hugging and handshaking. With these information in mind, we infer the category of relationships for all persons.

To demonstrate the importance of scene at inferring social relations, we present two examples in Figure 1. Each example consisting of two images shows different relationships mainly due to the scene context. For instance, Example 1 shows the professional relationship in the context of hospital and the relationship of friends in the context of a park. It is clear that the scene information should be carefully taken into consideration for social relation inference.

We note that prior works put more effort into learning from human interactions, the foreground union regions of person pair, and the background information of persons and objects, but missed the importance of holistic scene context, especially in inferring relationships from different scenes with similar human interactions. Goel et al. [10] adopted a pre-trained model to directly output the feature of foreground regions, which missed the whole context and may get a less representative feature. Zhang et al. [34] utilized a pre-trained model with ImageNet to generate a feature at the level of object classification. The resultant feature cannot semantically summarize high-level scene features.

A large number of studies proposed models to learn social relations based on interaction graph of persons in an image [10, 20, 29, 34]. Wang et al. [29] proposed graphs of persons and objects to infer social relations. The significant drawback of [29] is that graphs of persons and objects can only characterize the connection between two persons, which leads to complicated calculations for cases of three or more persons in an image.

It is urgent and essential to attack the problem of social relation inference in a broader view where an interaction graph works for two or more persons and simultaneously the scene feature provides holistic hints for the classification of relationships. In this paper, inspired by the understanding process of human beings, we propose an approach of **PR**actical **I**nference in **S**ocial r**E**lation (PRISE), which synthesizes three streams of information, i.e., holistic scenes, foreground and background information of persons and objects, and interaction of persons. To the best of our knowledge, we are the first to methodically and systematically develop a model to learn the holistic scene feature in social relation inference based on contrastive learning.

In PRISE, we first technically design a concise relational graph convolutional network (RGCN) to extract the interactive features of all persons in one image. Then, to boost the performance in social relation inference, a contrastive learning task for capturing holistic scene is incorporated into the proposed PRISE. Intuitively, the contrastive learning task helps to extract discriminative features of the holistic scene context. We demonstrate that PRISE achieves a significant improvement in social relation inference compared to the state-of-the-art methods.

We summarize our contributions in this work as follows.

- We systematically develop a novel approach, i.e., PRISE, for social relation classifications. PRISE significantly beats the state-of-the-art methods in social relation inference.
- We design a concise relational graph convolutional network to capture the interactive features for all persons. The proposed RGCN is simpler and faster than the graph model in [20].
- We construct a contrastive learning task to learn discriminative representation of holistic scene. We distribute a new large-scale dataset for contrastive learning, which is named as PISC-extension. The usefulness of PISC-extension can be extended to other tasks in computer vision, such as group behaviour analysis.
- Extensive experiments including a comprehensive ablation study demonstrate the effectiveness the PRISE, and show the significance of interaction graph and scene information in social relation inference.

2. Related Work

To assess our contributions in classification of social relations, it is important to consider three streams of studies: social relation inference, graph neural networks and contrastive learning.

2.1. Social Relation Inference

For a large number of scenarios in computer vision, social information has played an important role by providing additional cues in tasks of image understanding, e.g., human interaction [26], kinship recognition [22, 21, 23] and image caption generation [31].

The pioneering work on social relation inference dates back to 2010 from [28], where the authors developed a model to characterize the interaction between multi-person actions, facial appearances and identities. Zhang et al. [37] developed a deep neural network to learn social relation traits from rich facial attributes, such as expression, gender, and age. In [37], the social relation traits were defined based on psychological studies [12, 11], consisting of eight types, e.g., trusting and friendly.

For datasets in social relations, Zhang et al. [35] distributed a dataset to evaluate classification of social relations, which is named as People In Photo Albums (PIPA). Besides, another dataset, which is People in Social Context (PISC), was published in [19].

With PIPA and PISC, several interesting works move forward along the research line of social relation understanding [27, 29, 10, 20]. In light of domain based theory from social psychology, Sun et al. [27] presented a model with semantic attributes to classify social relations and domains. Wang et al. [28] modelled a knowledge graph with proper messages propagation and attention to learn the social relations among people in an image. Recently, in [10], Goel et al. proposed an end-to-end neural network to learn the interaction graph of persons. In [20], a social graph was proposed to restrict logical connections of persons, which achieved the state-of-the-art results in social relation inference. In Table 1, we present the differences between our PRISE and the prior studies in terms of feature information.

Table 1. Comparisons between our PRISE and previous methods in features for social relation inference. "Fore." is short for "Foreground" and "Back." is short for "Background".

Methods	Interaction Graph	Fore. & Back.	Holistic Scene
Pair CNN [19]	No	Yes	No
Dual-Glance [19]	No	Yes	No
SRG-GN [10]	Yes	Yes	No
GRM [29]	Yes	Yes	No
MGR [34]	Yes	Yes	No
GR^{2}N [20]	Yes	No	No
PRISE	Yes	Yes	Yes

2.2. Graph Neural Networks

Inspired by the success of convolutional networks in the computer vision domain, GNNs are proposed to re-define the notation of convolution for graph structured data [16]. Most recently, GNNs have been adopted to social relation reasoning [29, 34, 20]. For instance, Zhang et al. [34] designed person-object graph and person-pose graph, and conducted social relation reasoning on these two graphs by GNN. Li et al. [20] proposed a graph relational reasoning network to jointly infer social relations by building a graph for each image, where the nodes represent the persons and the edges represent the relations. In this paper, we follow the similar graph-based approach proposed in [20], and design a concise relational graph convolutional network to extract interactive features among people in the image.

2.3. Contrastive Learning

Over the last few years, contrastive representation learning based on deep learning models has shown the power in many practical tasks [4, 6, 7, 24, 30, 33, 36], especially for natural language and computer vision domains. Contrastive learning usually maximizes similarity and dissimilarity over data samples which are organized into similar and dissimilar pairs, respectively.

A significant challenge in contrastive learning is how to select the similar (or positive) and dissimilar (or negative) pairs. The main difference among different approaches of contrastive learning lies in their strategy for obtaining sample pairs [5]. To generate sample pairs without additional human labels, many researchers create models with multiple views of each sample. Besides, for complicated tasks, some studies also construct positive and negative sample pools from pre-trained models [9, 18]. In this paper, by utilizing the pseudo-labels from a pre-trained model of image scene classification, we construct negative and positive sample pools, and design a contrastive learning task to learn discriminative scene representations.

3. Methodology

In this section, we first introduce the approach that converts all persons in an image into an interaction graph, and then apply the RGCN model on the graph to learn interactive features of people in the same image. Finally, to better utilize scene information for social relation understanding, we propose a contrastive learning approach to learn discriminative scene representation. The overall pipeline of PRISE is shown in Figure 2.

3.1. Graph-based Approach

Inspired by [20], we adopt graph-based approach. We build a graph for each image, where each person in an image is modeled as a node in the graph. The edge between two nodes represents the social relation between the corresponding two persons. For simplicity, we consider the fully connected graph, i.e., each pair of persons in the image has an edge. Denote $\mathcal{G} = (\mathcal{V}, \xi)$ as the fully connected graph with node set $\mathcal{V}$ and edge set ξ in an image.

For each image, we extract three types of features using a ImageNet pre-trained model (i.e., ResNet101). These three types of features include RoI features of single person, union region of person pairs (a.k.a. foreground feature), and persons and objects of the whole image (a.k.a background feature). In the following, we will introduce the detailed ways to generate these features.

Following traditional approach in detection, the feature representation of each person is extracted directly from the last convolutional feature map of the input image. Specifically, given input image I with N bounding boxes $b_1, b_2, ..., b_N$ for N persons, we obtain the feature representations of all people in the image using a pre-trained ResNet101 model, where an RoI pooling layer is constructed based on the last convolutional feature map. Note that the RoI pooling layer is a common trick in social relation learning with graph representation [20]. Denote the feature representation of the i-th person in image I as x_i,

$$x_i = f_{CNN-RoI}(I, b_i) \in \mathbb{R}^F, i = 1, 2, ..., N, \quad (1)$$

where $F = 2048$ is the feature dimension for each person. For simplicity, we denote the set of feature representations for people in image I as $X = \{x_1, x_2, ..., x_N\}$.

In addition, we obtain the features of union regions of person pair using the same approach. For person i and j, we first compute the bounding box of their union region b_{ij}. Then we get its feature as follows:

$$x_{ij} = f_{CNN-RoI}(I, b_{ij}). \quad (2)$$

Besides, we also obtain x_I, the feature representation for the whole image, by setting the bounding box to cover the whole image and passing it to $f_{CNN-RoI}$.

$$x_I = f_{CNN-RoI}(I, b_I), \quad (3)$$

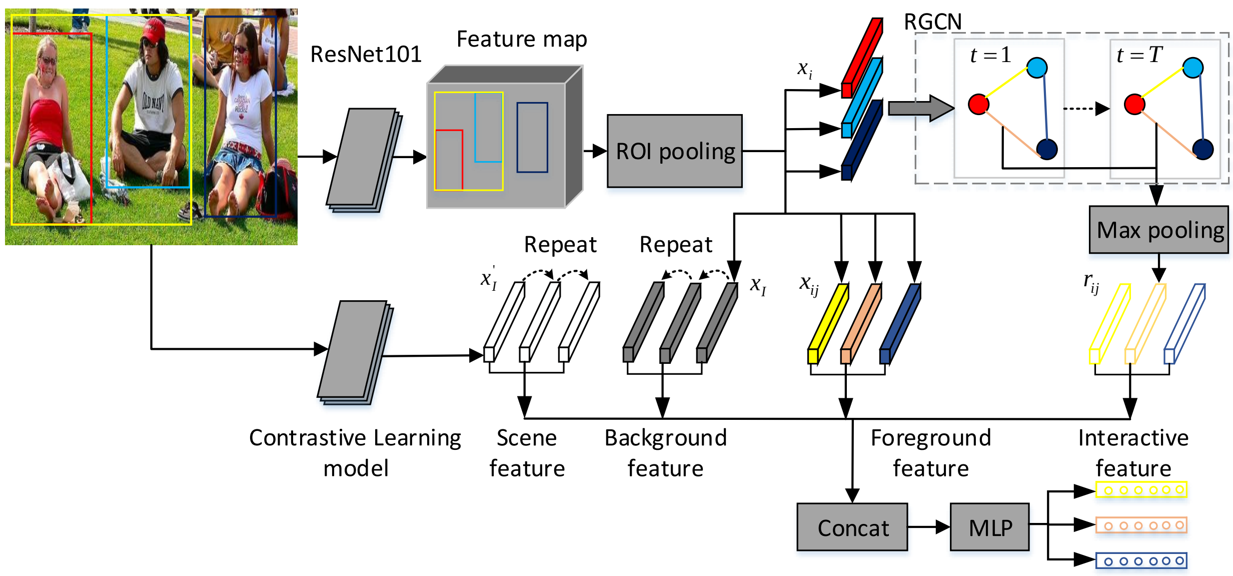

Figure 2. The overall pipeline of PRISE model. Given an input image I, we use ResNet101 to extract RoI features of people in the image x_i, foreground feature x_{ij} and background feature x_I. In addition, another pre-trained ResNet50 that was finetuned using contrastive learning approach to extract discriminative scene feature x_I'. The RGCN is used to obtain interactive feature between person pair r_{ij}. Finally, $r_{ij}, x_I, x_{ij}, x_I'$ are concatenated and passed to a MLP layer for relation classification. The network outputs relational class distribution for all person pairs in the image. The operation 'Repeat' is a must to keep the number of scene and background features the same as the number of person pairs when there exist more than two persons in an image.

where b_I is the bounding box for the whole image. Intuitively, the feature of single person x_i encodes personalized information of each person, the feature of union region x_{ij} encodes the pair-wise foreground information, while the feature of whole image x_I encodes the background of all persons and objects. Thus, all these features can provide useful information for social relation understanding.

3.2. Relational Graph Convolutional Network

In this section, we introduce RGCN, an end-to-end trainable network architecture, that can learn pair-wise interactive features given arbitrary graph structured data. We apply RGCN on the fully connected graph $\mathcal{G}$ with features X.

Given $\mathcal{G}$ and X, for each node $i \in \mathcal{V}$, we set its initial node feature vectors as $h_i^0 = wx_i \in \mathbb{R}^F, \forall i \in \mathcal{V}$, where $w \in \mathbb{R}^{F \times F}$ is a learnable parameter that maps input feature vectors to the new feature space. Correspondingly, each edge has a feature vector, and we denote the initial edge feature vector between node i and node j as $r_{ij}^0 \in \xi$. In RGCN with T layer, the edge and node feature vectors are updated iteratively for T times. Specifically, at t-th layer the edge and node representations can be expressed as follows:

$$r_{ij}^t = \sigma(W^t h_i^t + W^t h_j^t), \tag{4}$$

$$h_i^{t+1} = h_i^t + \sigma(W^t h_i^t + \sum_{j \in \mathcal{N}_i} r_{ij}^t \odot W^t h_j^t), \tag{5}$$

where $\mathcal{N}_i$ is the set of neighbors for node i, $W^t \in \mathbb{R}^{F \times F}, t = 1, 2, ..., T$ are the learnable parameters at each layer, and $\sigma(\cdot)$ is the ReLU function.

We note that the RGCN defined in (4)-(5) is an anisotropic variant of GCN [8]. Similar to Residual GateGCN [2], our RGCN has residual connections on the node feature representations, and explicitly maintains edge feature at each layer. Intuitively, the edge feature representations at different layers encode the pair-wise human interaction information. Following similar ideas in JK-Net [32], we obtain the final interactive features by using a max pooling on the edge representations from different RGCN layers. Formally, the final interactive feature between person i and j, denoted as r_{ij}, can be expressed as

$$r_{ij} = f_{max}(r_{ij}^0, r_{ij}^1, ..., r_{ij}^T), \tag{6}$$

where $f_{max}(\cdot)$ is an element-wise max function.

3.3. Discriminative Scene Representation Learning

The scene of an image provides important visual clues for social relation understanding. For instance, given a party scene, the group of people are more likely to be friends than colleagues, and a group of athletes running on a track are much more likely to be sports team members than band members [10]. To harvest the power of pre-trained CNN model and unlimited amount of unlabeled images, in this paper, we propose a contrastive learning (CL) approach for discriminative scene representation learning.

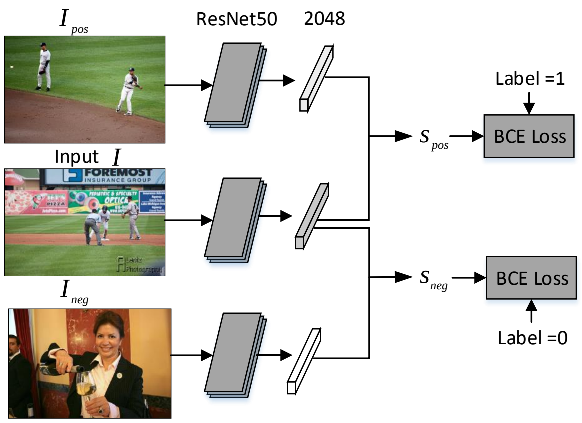

Figure 3. An overview structure of CL task. For a given image I, we first sample a similar image I_{pos} and a dissimilar image I_{neg} from the image dataset. All these three images are passed through the pre-trained ResNet50 model to obtain a feature representation.

Following the CL paradigm, we design a scene classification task to distinguish between similar and dissimilar images. As a pre-process step, we use the pre-trained ResNet50 model [38] to obtain the top-5 scene classes for each unlabeled image. Two images are defined as similar in scene if there are more than K scene classes that are the same among the top-5 scene classes. Otherwise, they are dissimilar. Based on this definition, we can have for each image a pool of similar images and a pool of dissimilar images. The structure of CL task is shown in Figure 3. For each input image I, we randomly sample one image I_{pos} from its pool of similar images to construct positive sample, and another image I_{neg} from its pool of dissimilar images to construct negative sample. We then apply the pre-trained ResNet50 model on these three images I, I_{pos}, I_{neg} to extract features, denoted as $x, x_{pos}, x_{neg} \in \mathbb{R}^F$, respectively. The similarity scores of samples are calculated using a simple bilinear scoring function with sigmoid activation function as follows:

$$s_{pos} = \sigma(xWx_{pos}), \quad s_{neg} = \sigma(xWx_{neg}),$$

where $W \in \mathbb{R}^{F\times F}$ is learnable parameter, σ is the sigmoid function. The pre-trained ResNet50 model is finetuned using the binary cross-entropy loss function. Namely, the loss function can be expressed as

$$L_{cl} = \frac{1}{|I|}\sum_{I}\left(-log(s_{pos}) - log(1 - s_{neg})\right). \quad (7)$$

We note that the structure of the CL task is similar to Triplet network [14]. Intuitively, the task is designed to map images with similar scenes closely to each other and dissimilar scenes separable as farther apart as possible. Thus images with similar scenes could have high similarity scores and vice versa. This contrastive learning paradigm enables the model to learn discriminative scene representations.

After finetuning the pre-trained CNN model, we use it as a scene feature extractor in our downstream social relation inference task. Namely, given an image I, we obtain the scene feature of the image as follows:

$$x'_I = f_{CL-RoI}(I, b_I), \quad (8)$$

where $f_{CL-RoI}(\cdot)$ represents the CL finetuned CNN model with RoI pooling layer.

Finally, to predict the relational class distribution of person i and j in the image, we concatenate their interactive feature r_{ij} extracted from RGCN model, foreground feature x_{ij}, background feature x_I extracted from ImageNet pre-trained CNN model and discriminative scene feature x'_I extracted from CL finetuned model together. The concatenated features are fed as input to the MLP layer for relation classification. The network outputs relational class distribution for all person pairs in the image.

We note that most of the previous methods, such as Pair CNN, Dual-Glance, SRG-GN, etc, consider the social relations on the same image separately. Namely, their model outputs relational class distribution for single pair of person *only*, even if there are multiple people in the image. This may cause occurrence of some obviously unreasonable and contradictory relationships in one image. In contrast, our model directly learns the joint distribution of social relations for multiple people. Given an image as input, PRISE extracts features of multiple people in the image and directly outputs the relational class distributions for *all* person pairs. This enables our model to generate reasonable and consistent social relationships in one image.

4. Experiments

In this section, we conduct extensive experiments based on PIPA and PISC, as well as a new large-scale unlabeled dataset. We first present the description of datasets and the implementation details. Then we evaluate the performance of our proposed model through comparisons with benchmarks and ablation study. Finally, we visualize the results from the CL model with discussions. All the codes and experimental results are publicly available on github[1].

4.1. Datasets

Social Relation Datasets. We conduct experiments on two social relation datasets, i.e., the PIPA dataset [35] and the PISC dataset [19]. The *accuracy* over all classes is used to evaluate all methods in PIPA dataset. The PISC dataset has a hierarchy of three coarse-level relations (intimate, non-intimate, no relation) and six fine-level relations (friend, family, couple, professional, commercial, and no relation).

[1] https://github.com/IFBigData/PRISE

Table 2. Comparisons of the accuracy (in %) between our PRISE and other state-of-the-art methods in PIPA dataset.

Methods	domain	relation
Pair CNN [19]	65.9	58.0
Dual-Glance [19]	-	59.6
SRG-GN [10]	-	53.6
GRM [29]	-	62.3
MGR [34]	-	64.4
GR^{2}N [20]	72.3	64.3
PRISE	**77.2**	**69.5**

For fair comparisons, we follow the standard train/val/test split in [19]. The *per-class recalls* and *mean Average Precision (mAP)* are used to evaluate all methods in PISC dataset. **Contrastive Learning Dataset.** For CL, we extend PISC dataset to a new dataset with 240,200 images by using google image search engine. Specifically, we search for 10 similar images on Google for every image in PISC dataset, thus extending PISC dataset by approximately 10 times. We combine the extended dataset with PIPA and PISC images and name it as PISC-extension dataset. We show some examples from this dataset in supplementary materials. We take 80% of samples in the PISC-extension dataset as the training set, and the remaining samples as the test set.

4.2. Implementation Details

Contrastive Learning. We use the pre-trained ResNet50 model [38] to obtain the top-5 scene categories of an input image. We then construct positive and negative sample pairs based on the scene category. Each image has a pool of positive samples and a pool of negative samples. For simplicity, we limit the maximum number of images in a pool for each image. In this paper, we set the maximum number as 50 [2]. In the training phase, we randomly select one image from the positive and negative sample pool respectively. We set the batch size to be 32, the learning rate to be 1×10^{-5}. The ResNet50 model is finetuned end-to-end using the Adam optimizer. For the performance of the finetuned ResNet50 model, the accuracy and AUC on the test set are 91.0% and 96.7%, respectively. After training, the network parameters are saved for the downstream task.
Training of PRISE. Our PRISE is trained with a learning rate of 5×10^{-5}. We resize the input image into 448×448, and train the network for 20 epochs with a batch size of 32. The number of layers in RGCN is set to be 2, i.e., $T = 2$.

4.3. Comparisons with Benchmarks

In experiments, we compare PRISE with the following existing methods. For fair comparisons, we report the best results in experiments for Tables 2 and 3 following the routine in this research field.

Pair CNN [19]. Two cropped image patches of the two persons are fed into two CNNs with sharing weights to extracted features for social relation classification.

Dual-Glance [19]. The first glance focuses on the pair of people. The second glance extracts the information of objects in the context to refine the prediction.

SRG-GN [10]. Scene and human attribute context features are extracted by five CNNs.

GRM [29]. This model represents the person and objects existing in an image as a weighted graph, and then using a gated graph network to predict social relation.

MGR [34]. This model employs two graph neural network (GNN) to extract the relationship between people and the relationship between people and objects.

GR^{2}N [20]. This model uses GNN to model all relationships in one graph which can provide strong logical constraints among different types of social relations.

It is worth noting that all of the above methods, except GR^{2}N, are person pair-based, which means that they consider the pair-wised social relations on the same image separately. In contrast, both PRISE and GR^{2}N consider the social relations among all people in one image jointly. Unlike our method, GR^{2}N do not use foreground, background and scene features. Besides, Dual-Glance, GRM and MGR use object information in an image to assist in relation inference. We note that in SRG-GN, they directly apply the pre-trained scene classification model as feature extractor for foreground ground information, while in our PRISE, we first use the CL approach to finetune the pre-trained model, and then apply the model for scene feature extraction.

The experimental results of social domain recognition and social relationship recognition in PIPA dataset are shown in Table 2. We observe that our PRISE outperforms other methods by a significant margin. Specifically, our method achieves an accuracy of 77.2% for social domain recognition and 69.5% for social relation recognition, beating all the person pair-based methods. This shows the benefit of graph-based approach that jointly models all the social relationships among people in an image. Besides, our method improves the current state-of-the-art method, i.e., GR^{2}N, by 6.8% for social domain recognition and 8.1% for social relation recognition, respectively.

Similar results can be found in Table 3, where we shows the experimental comparison with prior methods in PISC dataset. We observe that our method achieves an mAP of 83.4% for the coarse-level recognition and 73.8% for the fine-level recognition, which are new state-of-the-art. We note that PRISE takes full advantage of holistic scene and thus makes better predictions for non-intimate relation. For degradation in 'Int' and 'Fri', we argue that the similar scenes of 'Fri' and 'Fam' misleads our model.

[2] We have considered other values (e.g., 30, and 80) and found that this parameter is insensitive to the results.

Compared with GR^{2}N, PRISE achieves competitive performance for both coarse and fine relationship recognition with a much simpler GCN structure. The above results highlight the benefits of concise interaction graph and discriminative scene feature in PRISE.

4.4. Ablation Study

We conduct ablation study to show how much each component of PRISE contributes to the performance. Specifically, we remove the interactive feature, the scene feature, the foreground and background features from PRISE, denoted as *w/o Int.*, *w/o Scene*, *w/o Fore.*, *w/o Back.*, respectively. In addition, to show the effectiveness of discriminative scene representation, we consider a variant denoted by *PRISE*|*Pretrained*, where we replace the CL finetuned model with the ResNet50 that was pretrained on Place365 dataset. The results are summarized in Table 4.

As we can see in Table 4, among all the four components, the interactive feature is the most important, followed by the scene feature. Without interactive feature, the mean of mAP in PISC-coarse and PISC-Fine dataset drops 7.4% and 11.3% in absolute value, respectively. These two numbers become 0.8% and 0.9% if we remove the scene feature from PRISE. On one hand, this result demonstrates the effectiveness of RGCN to extract interactive feature. On the other hand, it shows the benefit of considering scene information in social relation understanding.

Besides, by comparing the results of *PRISE*|*Pretrained* and PRISE, we clearly find that the discriminative scene representation provides significant hints for social relation classifications, especially for fine relationships with 1.8% relative improvement.

4.5. Visualization of Scene Representations

To demonstrate the discriminative scene representation learned by our CL finetuned model, we randomly choose 4000 images from PISC-coarse test dataset, and conduct a clustering task based on the learned features. Specifically, we first use the CL finetuned model to generate the 2048-dimensional scene representations for each image. Then, we use spectral clustering to cluster these features. The 4000 test images are divided into 6 categories. We visualize sample images in different clusters in supplementary materials. We use TSNE to reduce the features dimension from 2048 to 2, and visualize them in Figure 4. We can observe that images from different categories are separated. These results directly show that the features learned by our CL finetuned model are discriminative.

4.6. Discussions on Effectiveness of PRISE

RGCN as Interactive Feature Extractor: Simpler and Faster. We would have used GR^{2}N [20] as the interactive

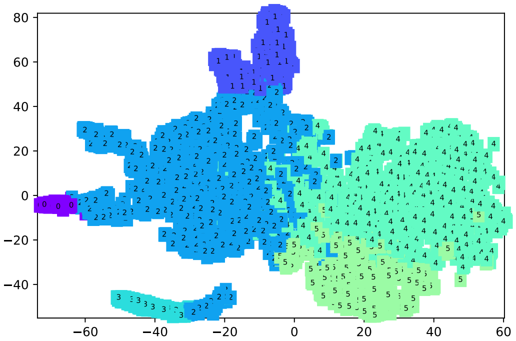

Figure 4. TSNE visualization of image scene features obtained from CL finetuned model on PISC test dataset. Different colors are used to represent clusters of different categories.

feature extractor, however, we note that it is too complicated. For a social relation inference problem with K categories of social relation, GR^{2}N introduced K sets of trainable parameters. In contrast, the number of parameters of our RGCN does not depend on the number of social relation categories, which makes RGCN much simpler.

To further compare GR^{2}N and RGCN as interactive feature extractors in terms of performance and inference time, we conduct experiments by replacing RGCN with GR^{2}N in PRISE while fixing other components. Specifically, in GR^{2}N each category of social relation has a representation. We apply a max pooling operator on representations of different social relations to obtain the interactive feature, and replace r_{ij} in our PRISE with this new interactive feature. We denote this setting as *PRISE*|*GR*^{2}*N*. The experimental results on the accuracy and inference time [3] on test set of both algorithms in PIPA dataset are shown in Table 5. We can observe that the performance of PRISE|GR^{2}N is comparable to PRISE. However, in terms of inference time, the model with GR^{2}N is much slower as compared to PRISE. These results strongly support our conclusion that the proposed RGCN model as interactive feature extractor is much simpler and faster as compared to GR^{2}N.

Benefits of Utilizing the Scene Information. Scene information is important for social relation understanding. In Figure 5, we visualize two sample images from PISC test dataset, where *PRISE w/o Scene* makes wrong predictions while PRISE makes correct predictions. In this example, people (in the left image) in a home environment tends to have intimate relation, while people (in the right image) in public environment tends to have non-intimate relation. Without scene information, PRISE makes wrong predictions by predicting the two persons in the left image to have non-intimate relation, and the two persons in the right im-

[3] The inference time reported here does not include the time needed to extract features using ResNet101 model and CL finetuned model.

Table 3. Comparisons of the per-class recall for each relationship and the mAP over all relationship (in %) between our PRISE and other state-of-the-art methods in PISC dataset. Int: Intimate, Non: Non-Intimate, NoR: No Relation, Fri: Friend, Fam: Family, Cou: Couple, Pro: Professional, Com: Commercial, NoR: No Relation.

Methods	Coarse relationships				Fine relationships						
	Int	Non	NoR	mAP	Fri	Fam	Cou	Pro	Com	NoR	mAP
Pair CNN [19]	70.3	80.5	38.8	65.1	30.2	59.1	69.4	57.5	41.9	34.2	48.2
Dual-Glance [19]	73.1	84.2	59.6	79.7	34.4	68.1	76.3	70.3	57.6	60.9	63.2
SRG-GN [10]	-	-	-	-	-	-	-	-	-	-	71.6
GRM [29]	81.7	73.4	65.5	82.8	59.6	64.4	58.6	76.6	39.5	67.7	68.7
MGR [34]	-	-	-	-	64.6	67.8	60.5	76.8	34.7	70.4	70.0
GR^{2}N [20]	81.6	74.3	70.8	83.1	60.8	65.9	84.8	73.0	51.7	70.4	72.7
PRISE	73.3	79.2	71.8	**83.4**	47.1	74.7	76.6	73.2	70.3	68.2	**73.8**

Table 4. Ablation study of the PRISE model in PISC dataset. We report the mean and standard deviation of mAP (in %) among 50 random runs in PISC dataset.

Methods	Coarse	Fine
PRISE w/o Int.	75.3 ± 0.2	61.4 ± 0.4
PRISE w/o Scene	81.9 ± 0.3	71.8 ± 0.4
PRISE w/o Fore.	82.2 ± 0.4	71.9 ± 0.5
PRISE w/o Back.	82.5 ± 0.3	72.5 ± 0.4
PRISE|Pretrained	82.2 ± 0.4	71.4 ± 0.4
PRISE	**82.8 ± 0.3**	**72.8 ± 0.5**

Table 5. Comparisons of GR^{2}N and RGCN in PIPA dataset. We report the accuracy (in %) and inference time.

Methods	domain		relation	
	accuracy	time	accuracy	time
PRISE|GR^{2}N	75.6	3.33s	68.1	6.23s
PRISE	**77.2**	1.91s	**69.5**	1.82s

age to have intimate relation. More examples are illustrated in supplementary materials.

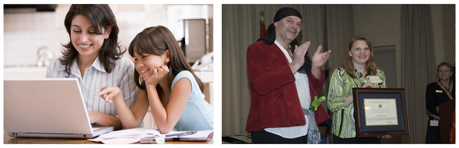

Figure 5. Visualization of sample images from PISC test dataset, where *PRISE w/o Scene* makes wrong predictions while PRISE makes correct predictions.

Besides, the scene representations extracted from different models could be very different. In Figure 6, we show the heat map and gradient map [25] of the pre-trained models from ImageNet and Place365 [38] dataset, respectively. We can observe that the feature extracted by ImageNet pre-trained model focuses more on the persons, while the feature extracted by Place365 pre-trained model focuses more on the holistic scene.

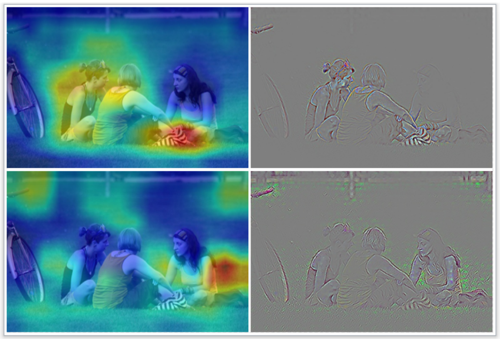

Figure 6. Heat map and gradient map of scene representations extracted from different models. The top two images are the pre-trained model on Imagenet. The bottom two images are from the pre-trained model on Place365.

5. Conclusion

In this paper, we have originally proposed PRISE to enhance social relation inference. PRISE synthesizes three streams of information, i.e., holistic scenes, foreground and background information of persons and objects, and interaction of persons. Technically, we have developed a RGCN model to extract interactive features and designed a CL task to learn discriminative scene representations. The RGCN model in PRISE is concise in terms of learning the interaction for all persons in an image and the running time of feature extraction. Extensive experiments demonstrate that PRISE is superior than prior methods, and achieves new state-of-the-art results in PIPA and PISC datasets. The contrastive learning task in PRISE sheds new lights on improving performance of more complicated tasks in computer vision, such as behaviour analysis and image captioning.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 1

[2] Xavier Bresson and Thomas Laurent. Residual gated graph convnets. *ICLR*, 2018. 4

[3] David G Bromley and Bruce C Busching. Understanding the structure of contractual and covenantal social relations: Implications for the sociology of religion. *SA. Sociological analysis*, pages 15S–32S, 1988. 1

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[5] Ching-Yao Chuang, Joshua Robinson, Lin Yen-Chen, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *arXiv preprint arXiv:2007.00224*, 2020. 3

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[7] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 3

[8] Vijay Prakash Dwivedi, Chaitanya K Joshi, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Benchmarking graph neural networks. *TSP*, 12:50–500. 4

[9] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning from multi-domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3245–3255, 2019. 3

[10] Arushi Goel, Keng Teck Ma, and Cheston Tan. An end-to-end network for generating social relationship graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11186–11195, 2019. 2, 3, 4, 6, 8

[11] John Gottman, Robert Levenson, and Erica Woodin. Facial expressions during marital conflict. *Journal of Family Communication*, 1(1):37–57, 2001. 2

[12] Ursula Hess, Sylvie Blairy, and Robert E Kleck. The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal behavior*, 24(4):265–283, 2000. 2

[13] Minh Hoai and Andrew Zisserman. Talking heads: Detecting humans and recognizing their interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 875–882, 2014. 1

[14] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition*, pages 84–92. Springer, 2015. 5

[15] Mahmoud Khademi and Oliver Schulte. Image caption generation with hierarchical contextual visual spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1943–1951, 2018. 1

[16] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 3

[17] Shinobu Kitayama and Hazel Rose Markus. The pursuit of happiness and the realization of sympathy: Cultural patterns of self, social relations, and well-being. *Culture and subjective well-being*, 1:113–161, 2000. 1

[18] Wonhee Lee, Joonil Na, and Gunhee Kim. Multi-task self-supervised object detection via recycling of bounding box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4984–4993, 2019. 3

[19] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2659, 2017. 1, 2, 3, 5, 6, 8

[20] Wanhua Li, Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Graph-based social relation reasoning. *arXiv preprint arXiv:2007.07453*, 2020. 2, 3, 6, 7, 8

[21] Jianqing Liang, Qinghua Hu, Chuangyin Dang, and Wangmeng Zuo. Weighted graph embedding-based metric learning for kinship verification. *IEEE Transactions on Image Processing*, 28(3):1149–1162, 2018. 2

[22] Jiwen Lu, Junlin Hu, and Yap-Peng Tan. Discriminative deep metric learning for face and kinship verification. *IEEE Transactions on Image Processing*, 26(9):4269–4282, 2017. 2

[23] Jiwen Lu, Xiuzhuang Zhou, Yap-Pen Tan, Yuanyuan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):331–345, 2013. 2

[24] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[26] Xiangbo Shu, Jinhui Tang, Guojun Qi, Wei Liu, and Jian Yang. Hierarchical long short-term concurrent memory for human interaction recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[27] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3481–3490, 2017. 2

[28] Gang Wang, Andrew Gallagher, Jiebo Luo, and David Forsyth. Seeing people in social context: Recognizing people and social relationships. In *European conference on computer vision*, pages 169–182. Springer, 2010. 2

[29] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge

graph for social relationship understanding. *arXiv preprint arXiv:1807.00504*, 2018. 2, 3, 6, 8

[30] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. *arXiv preprint arXiv:2008.05659*, 2020. 3

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2

[32] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi, and Stefanie Jegelka. Representation learning on graphs with jumping knowledge networks. *arXiv preprint arXiv:1806.03536*, 2018. 4

[33] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763, 2019. 3

[34] Meng Zhang, Xinchen Liu, Wu Liu, Anfu Zhou, Huadong Ma, and Tao Mei. Multi-granularity reasoning for social relation recognition from images. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1618–1623. IEEE, 2019. 2, 3, 6, 8

[35] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4804–4813, 2015. 2, 5

[36] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017. 3

[37] Zhanpeng Zhang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3639, 2015. 2

[38] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5, 6, 8