Enhancing Social Relation Inference with Self-Supervised Learning Based on Image Scene Classification

Anonymous CVPR 2021 submission

Paper ID 8168

Abstract

*There has been a recent surge of research interest in attacking the problem of social relation inference based on images. Existing works mainly learn the social relation of two persons by taking advantage of human interaction or knowledge graph of persons and objects in an image from small-scale labeled datasets, leading to two significant drawbacks. One is the neglect of subtle information from scenes, and the other is the lack of exploration in unlabeled data. In this paper, in order to solve the two drawbacks in a unified framework, we propose a deep model to enhance social relation inference by self-supervised learning based on image scene classification, which is named as **S**elf-supervised **E**nhanced **R**elational **G**raph **C**onvolutional **N**etwork (SERGCN). Intuitively, the task of image scene classification captures the basic feature of holistic scene context and enhances the downstream task of social relation understanding among people in the same image. Besides, we design a simple and fast relational graph convolution network (RGCN) to learn interactive features among persons in one image. To boost the performance in social relation inference, we collect and distribute a new large-scale dataset (named as PISC-extension) for social relation inference, 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., SERGCN achieves 6.8% improvement for domain classification in PIPA dataset.*

1. Introduction

Social relations reflect the connections that exist between people, which are fundamental to human daily life [18]. Based on [4], common social relations include family, couple, friends, colleagues, professional, etc. Nowadays, billions of people share images in social media platforms such as Facebook[1], Twitter[2] and Flickr[3]. There has been an increasing interest in understanding social relations among persons in an image due to the broad applications including group behavior analysis [15], image caption generation [17] and human trajectory prediction [1].

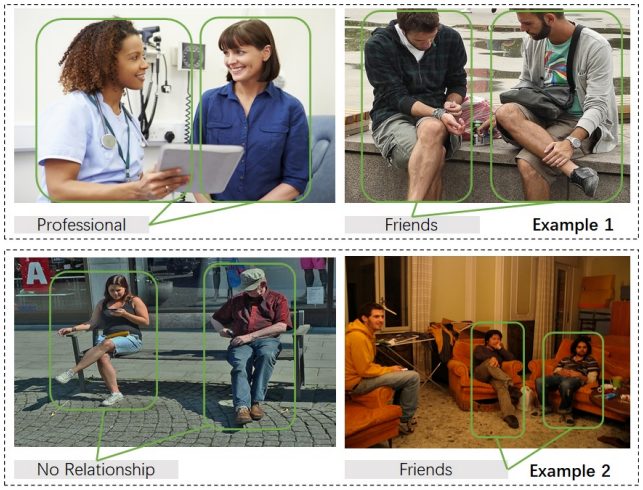

Figure 1. The comparison of inference on social relations under different scenes, with images taken from PISC dataset [20]. Each example consists of two images to show different relationship due to the scene context. In particular, Example 1 shows the professional relationship in the context of hospital. Example 2 shows the significant implication of close relationship in the context of staying indoors.

For the problem of social relation understanding, there is mounting evidence that the scene information in an image greatly influences the social relation between people [11, 20, 34, 38]. For instance, Goel et al. [11] proposed a two-branch model, which consists of a channel of body interaction and a channel of the whole image, to classify social relations among people in an end-to-end manner. Basically, the channel of the whole image is expected to char-

[1]https://www.facebook.com/
[2]https://twitter.com
[3]https://www.flickr.com/

acterize some scene information. Wang et al. [20] showed that region proposals of an image provide significant improvements for classification of social relations. To better demonstrate the importance of scene, we show two examples of inferring different social relations on the basis of different scene information, which are illustrated in Figure 1. Each example consists of two images to show different relationship 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.

It is surprising that the scene information has not been systematically and methodically investigated for the problem of social relation inference based on images. Specifically, in [20], Li et al. focused on modelling local context near the detected persons. In [34], a knowledge graph, which emphasizes on the interaction between persons and objects, was extracted from an image. In addition, the prior works usually adopt small-scale labeled datasets. The resultant drawbacks are two-fold. First, the neglect of information from scenes leads to bias in social relationship understanding. Second, unlabeled data have not been explored in the learning models of prior works.

It is urgent and significant to attack the problem of social relation inference in a broader view where the scene information, from not only labeled data but also unlabeled data, provides essential hints for the classification of relationships among people. In this paper, to address the above two challenges, we propose **S**elf-supervised **E**nhanced **R**elational **G**raph **C**onvolutional **N**etwork (SERGCN), a deep model to enhance social relation inference by self-supervised learning based on image scene classification.

We first design a simple and fast relational graph convolution network (RGCN) to learn the features of human interaction in one image. By utilizing the local and global context feature from ImageNet pre-trained CNN model, RGCN performs better compared to the prior works in learning human interaction, which achieves 4.0% improvement for domain classification of PIPA dataset in ablation study as compared to GR^{2}N.

Then, to boost the performance in social relation inference, a pretext task of image scene classification is incorporated into the proposed SERGCN using self-supervised approach. Specifically, the learned scene feature from the self-supervised model are concatenated with the local and global context features, and the interactive features from RGCN for classification of social relation. Intuitively, the task of image scene classification helps to extract features in providing holistic scene context, and enhances the downstream task of social relation classification. We demonstrate that SERGCN achieves a significant improvement in social relation inference compared to the state-of-the-art methods, e.g., 6.8% improvement for domain classification in PIPA dataset.

For training the self-supervised model, we collect and distribute a new large-scale dataset for social relation inference, which consists of about 240 thousand unlabeled images. The usefulness of the new large-scale dataset can be extended to other tasks in computer vision, such as group behaviour analysis.

We summarize our contributions in this work as follows.

- We design a simple and fast relational graph convolutional network to capture the features of human interaction in one image.
- We systematically develop a novel deep model, i.e., SERGCN, for social relation inference by taking advantage of self-supervised learning to utilize scene information. We construct image scene classification as the pretext task for self-supervised learning.
- We collect and distribute a new large-scale dataset for social relation inference, which is named as PISC-extension. Extensive experiments demonstrate the effectiveness our proposed model, which achieves better performance than the state-of-the-art methods.

2. Related Work

In order to assess our contributions in social relation inference, it is important to consider two streams of research: social relation recognition and self-supervised learning.

2.1. Social Relation Recognition

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 [31], kinship recognition [23, 22, 24] and image caption generation [35]. Shu et al. [31] addressed the problem of human interaction recognition by modeling the long-term inter-related dynamics among a group of persons for image frames from videos (e.g., UT dataset [30]). Lu et al. [23] proposed a new method to learn hierarchical non-linear features for face and kinship verification in images. Liang et al. [22] developed models to learn multiple metrics in graphs of persons by characterizing the intra-class compactness and inter-class separability. Xu et al. [35] trained a model with automatic attentions to the content of images, especially for the social interaction among multiple persons.

The pioneering work on social relation recognition dates back to 2010 from [33], where the authors developed a model to characterize the interaction between multi-person actions, facial appearances and identities. The model recognized the family relationships, such as husband-wife, siblings, grandparent-child, father-child, or mother-child. Later, Zhang et al. [40] developed a deep neural network

to learn social relation traits from rich facial attributes, such as expression, head pose, gender, and age. In [40], the social relation traits were defined in terms of psychological studies [14, 12], which consists of eight types, e.g., trusting and friendly.

Recently, Zhang et al. [39] introduced a new dataset to evaluate the problem of social relation recognition, which is named as People In Photo Albums (PIPA). Besides, in [20], another dataset, which is People in Social Context (PISC), was published for social relation recognition and a novel dual-glance model was developed based on deep neural networks. With PIPA and PISC, several interesting works move forward along the research line of social relation understanding [32, 34, 11, 21].

In light of domain based theory from social psychology, Sun et al. [32] presented a model with semantic attributes to classify social relations and domains. Wang et al. [33] modeled a novel structured knowledge graph with proper message propagation and attention to learn the social relations among people in an image. Recently, in [11], Goel et al. proposed an end-to-end neural network to learn the graph information in an image. In [21], a social graph was proposed to constrain the relations among multiple pairs of persons, which achieved the state-of-the-art results in the research line of social relation understanding.

2.2. Self-supervised Learning

In the recent decades, a number of self-supervised learning methods for visual feature learning have been developed without using any human-annotated labels [25, 19, 28, 10, 29, 8, 27, 5, 37, 26]. Supervised learning methods usually require a data pair x_i and y_i to train a model, where x_i is the feature of the data sample and y_i is human-annotated information. By contrast, self-supervised learning (SSL) methods, which belong to unsupervised learning, train a model with data x_i along with its pseudo label p_i, where p_i is automatically generated for a pre-defined pretext task without involving any human annotation [16]. It is worth noting that as long as the pseudo labels do not involve human annotations, then the methods belong to SSL.

Large-scale labeled data are generally required to train deep neural networks in order to obtain better performance in many practical applications of computer vision [7]. To avoid extensive cost of collecting and annotating large-scale datasets, it is essential to design models which can learn from limited labeled data and also boost the performance by utilizing large-scale unlabeled data. The paradigm of SSL fits this setting well and has shown great potential in solving the problem of learning with unlabeled data [6].

Early work in the pre-training of SSL for deep neural networks aimed to effectively train stacked auto-encoder without lables [2]. Recently, greedy layer-wise unsupervised learning has fallen out of fashion in favor of end-to-end learning where the whole deep model is trained in one operation [13]. After training a model on the pretext task, it can then be adapted to the target task through transfer learning. Pre-training tasks come in many forms. They usually involve transforming or imputing the input data with the goal of forcing the model to predict missing parts of the data or through introducing some information bottleneck.

In this paper, we design a pretext task of image scene classification for self-supervised learning with unlabeled data. The downstream task is to classify the social relation between two persons in an image.

3. Methodology

In this section, we present the technical details of our proposed SERGCN model. We first introduce the approach that converts persons in an image into an interaction graph, and then apply the RGCN model on the graph to learn interactive features among people in the same image. Finally, to better utilize scene information for social relation understanding, we propose a self-supervised learning approach. The overall pipeline of SERGCN is shown in Figure 2.

3.1. Graph-based Approach

We adopt the graph-based approach proposed in [21]. Namely, 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 $\mathcal{G} = (\mathcal{V}, \xi)$ 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 ResNet101 model. These three types of features include RoI features of single person, union region of person pairs (a.k.a. local context feature) as well as features of the whole image (a.k.a global context 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 [21]. Denote the feature representation of the i-th person in image I as x_i, we have

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

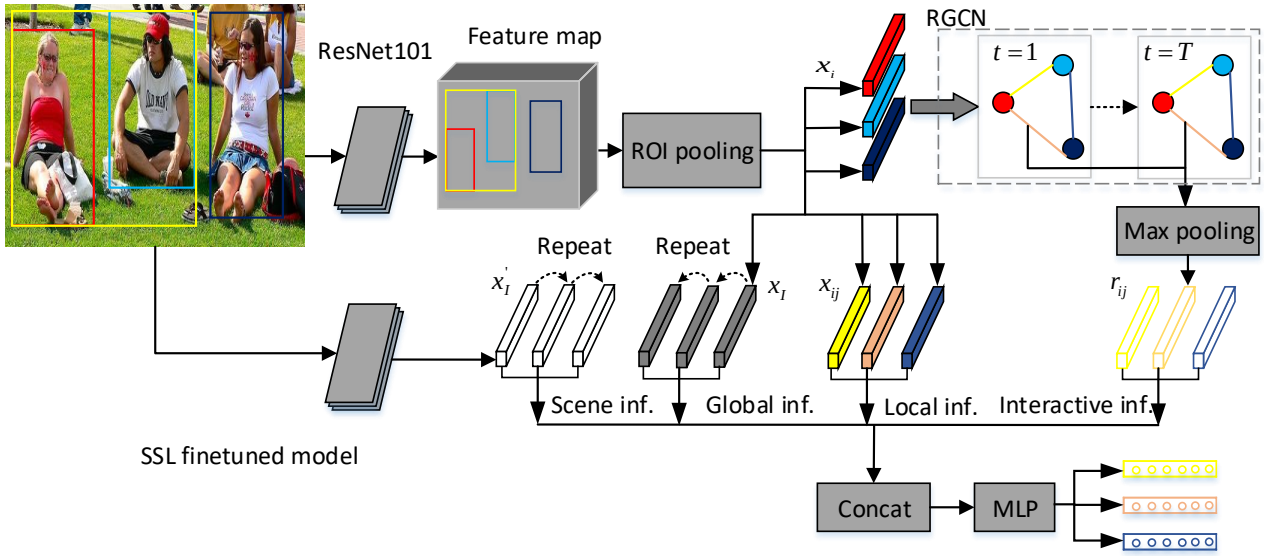

Figure 2. The overall pipeline of our proposed SERGCN network. Given an input image I, we use an ImageNet pre-trained CNN model (ResNet101) to extract RoI features of people in the image x_i, local context feature x_{ij} and global context feature x_I from the last shared feature map. In addition, another pre-trained CNN model (ResNet50) that was finetuned using SSL approach is used in a similar way to extract scene feature x'_I. We construct a graph for each image. The relational graph convolution network is used to obtain interactive feature between person pair $r_{ij}, i = 1, 2, ..., N, j = 1, 2, ..., N$. 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 global features the same as the number of person pairs when there exist more than two persons in an image.

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 pairs 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:

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

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 $f_{CNN-RoI}$.

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

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 context information in a local region, while the feature of whole image x_I encodes the global context information of all persons in the image. Thus, all these features can provide useful information for social relation understanding.

3.2. Relational Graph Convolution Network

In this section, we introduce relational graph convolution network (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, which is obtained from Section 3.1 to generate interactive features.

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 [9]. Similar to Residual GateGCN [3], our RGCN has residual connections on the node feature representations, and explicitly maintains edge feature at each layer. Intuitively, the edge feature represen-

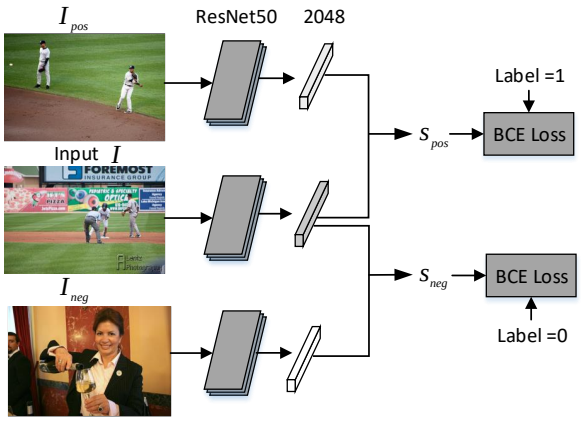

Figure 3. An overview structure of our proposed SSL pretext task. For a given image I, we first sample a similar image I_{pos} and a dissimilar image I_{neg} from the image dataset. All these three images are passed through the pre-trained ResNet50 model to obtain a feature representation. The score is calculated using bi-linear scoring function and the network is trained using binary cross entropy loss function.

tations at different layers encode the pair-wise human interaction information. Following similar ideas in JK-Net [36], 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. SSL for Scene Understanding

The scene of an image provides important 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 [11]. To utilize the scene information for social relation understanding, authors in [11] use CNN model that was pre-trained on *Places365* dataset [41], and apply it on the image as a feature extractor for downstream social relation understanding task. However, we empirically found that such method provides little improvements, as will be shown in the experiment section. To harvest the power of pre-trained CNN model and unlimited amount of unlabeled images, in this paper, we propose an SSL approach.

We construct an image scene classification task as the pretext task for self-supervised learning. As a pre-process step, we use the pre-trained ResNet50 model [41] to obtain the top-5 scene classes for each 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.

Following the contrastive learning paradigm, the scene classification task is designed to distinguish between similar and dissimilar images. The structure of SSL pretext 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 scores of samples are calculated using a simple bilinear scoring function with sigmoid activation function as follows:

$$s_{pos} = \sigma(xWx_{pos}), \tag{7}$$
$$s_{neg} = \sigma(xWx_{neg}), \tag{8}$$

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.

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

$$x'_I = f_{SSL-RoI}(I, b_I), \tag{9}$$

where $f_{SSL-RoI}(\cdot)$ represents the 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, the feature of their union region x_{ij}, global contextual feature x_I extracted from ImageNet pre-trained CNN model and scene feature x'_I extracted from SSL 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. In contrast, our model directly learns the joint distribution of social relations for multiple people. Given an image as input, SERGCN extracts features of multiple people in the image and directly outputs the relational class distributions for all person pairs.

4. Experiments

In this section, we conduct extensive experiments based on PIPA and PISC, as well as a new large-scale unlabeled

dataset. Specifically, we firstly introduce the description of datasets. Secondly, we present the details of implementation in experiments. Thirdly, we compare the results of our proposed model with benchmarks. Fourthly, we conduct an ablation study to show the power of SSL with image scene classification for social relation inference. Fifthly, we visualize the results from the SSL finetuned model. Finally, we discuss the effectiveness of SERGCN. All the codes and experimental results are publicly available on github[4].

4.1. Datasets

4.1.1 Social Relation Datasets

We conduct experiments on two social relation datasets, i.e., the PIPA dataset [39] and the PISC dataset [20]. The PIPA dataset partitions social life into 5 social domains and 16 social relations, which is divided into 13,729 person pairs for training, 709 for validation, and 5,106 for testing. 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). For fair comparisons, we follow the standard train/val/test split released by [20]. The per-class recalls and *mean Average Precision (mAP)* are used to evaluate all methods in PISC dataset.

4.1.2 Self-supervised Learning Dataset

For self-supervised learning, 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 Figure 4. 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

4.2.1 Self-supervised Learning

In self-supervised learning, we use the pre-trained ResNet50 model [41] to obtain the top-5 scene categories of an input image. We then construct positive and negative sample pairs based on the scene category. Specifically, for any two images, if they have more than two scene categories that are the same, the two images constitute a positive sample. Otherwise, they constitute a negative sample. In this way, 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 [5].

[4]https://github.com/IFBigData/SERGCN

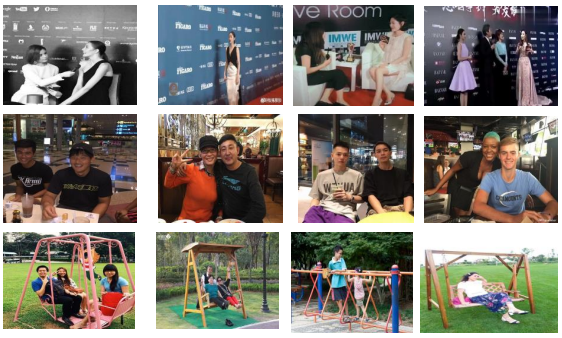

Figure 4. Visualization of the PISC-extension dataset. The images in the first column are from PISC dataset, and images from following columns are from Google image search engine.

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 in an end-to-end style by 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.

4.2.2 Training of Networks

Our SERGCN 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., $T = 2$. We use the SSL finetuned model to capture the feature of scene information. In our experiments, we observe that fixing the parameters of all the CNN models results in better performance.

4.3. Comparisons with Benchmarks

Table 1. Comparisons of the accuracy between our SERGCN and other state-of-the-art methods in PIPA dataset.

Methods	domain	relation
Pair CNN [20]	65.9%	58.0%
Dual-Glance [20]	-	59.6%
SRG-GN [11]	-	53.6%
GRM [34]	-	62.3%
MGR [38]	-	64.4%
GR^{2}N [21]	72.3%	64.3%
SERGCN	**77.2%**	**69.5%**

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

In our experiments, we compare SERGCN with the following existing methods.

Pair CNN [20]. 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 [20]. 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 [11]. Scene and human attribute context features are extracted by five CNNs. Then these features are utilized to update GRUs.

GRM [34]. 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 [38]. This model employs two graph neural network (GNN) to extract the relationship between people and the relationship between people and objects.

GR^{2}N [21]. 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 social relations on the same image separately. Unlike our method, GR^{2}N do not use any scene or attribute context cues. Besides, Dual-Glance, GRM and MGR use object information in an image to assist in relation inference, while SRG-GN uses scene information as SERGCN does in this paper. However, we note that in SRG-GN, they directly apply the pre-trained scene classification model as feature extractor, while in our SERGCN, we first use the SSL approach to finetune the pre-trained CNN model, and then apply the finetuned model for feature extraction.

The experimental results of social domain recognition and social relationship recognition in PIPA dataset are shown in Table 1. We observe that our SERGCN 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, outperforming 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. On one hand, this result highlights the benefit of considering scene and context information. On the other hand, it shows the effectiveness of our SERGCN model.

Table 2 shows the experimental comparison with the recent state-of-the-art methods in PISC dataset. We observe that our method achieves an mAP of 83.3% for the coarse-level recognition and 73.8% for the fine-level recognition, which are new state-of-the-art. Compared with GR^{2}N, SERGCN achieves competitive performance for both coarse and fine relationship recognition with a much simpler GCN structure. This result further validates the advantage of considering scene and context information in social relation inference.

4.4. Ablation Study

In this section, we examine the performance of our SERGCN model variations in PIPA dataset. The comparisons of different variants by accuracy are listed in Table 3. We evaluate the importance of scene context in predicting relationships in our SERGCN model by considering two variants. The first variant, denoted as *SERGCN (w.o. scene inf.)* in Table 3, is to remove the scene information x'_I in our model. The second variant, denoted as *SERGCN (pre-trained scene)* in Table 3, is to replace the SSL finetuned model with the pre-trained CNN model. We note that the way we use the scene information in the second variant is the same as that in SRG-GN. We observe that both methods exhibit inferior performance as compared to our SERGCN. On one hand, this result shows the benefit of considering scene information in social relation understanding. On the other hand, it suggests the superiority of our SSL approach to utilize the scene information as compared to the method in SRG-GN. In addition, even without scene information, our model achieves 4.0% improvement for domain classification of PIPA dataset, as compared to GR^{2}N in Table 1.

4.5. Visualization of SSL Finetuned Model

Through observation, we found that scene information plays a very important role in the identification of social relationships. In order to demonstrate that the scene information learned by our self-supervised finetuned model, we calculate the similarity between the features of different images obtained by the SSL finetuned model. The results are shown in Figure 5. We can observe that the feature similarities are very high if two images have the same scene, but this value is very low if two images have completely different scenes. The result directly shows that the features learned by our self-supervised finetuned model is sensitive to the scene, i.e., the scene information contained in the feature is distinguishable.

4.6. Effectiveness of SERGCN

4.6.1 RGCN: Simpler and Faster

We note that graph convolution network has been used in social relation inference such as GR^{2}N by Li et al. [21]. In GR^{2}N, each social relation has a virtual relation graph, each of which contains a set of trainable parameters. For a social relation inference problem with K categories of social relation, Li et al. 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.

Table 2. Comparisons of the per-class recall for each relationship and the mAP over all relationship (in %) between our SERGCN 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 [20]	70.3	80.5	38.8	65.1	30.2	59.1	69.4	57.5	41.9	34.2	48.2
Dual-Glance [20]	73.1	84.2	59.6	79.7	34.4	68.1	76.3	70.3	57.6	60.9	63.2
SRG-GN [11]	-	-	-	-	-	-	-	-	-	-	71.6
GRM [34]	81.7	73.4	65.5	82.8	59.6	64.4	58.6	76.6	39.5	67.7	68.7
MGR [38]	-	-	-	-	64.6	67.8	60.5	76.8	34.7	70.4	70.0
GR^{2}N [21]	81.6	74.3	70.8	83.1	60.8	65.9	84.8	73.0	51.7	70.4	72.7
SERGCN	73.3	79.2	71.8	**83.3**	47.1	74.7	76.6	73.2	70.3	68.2	**73.8**

Table 3. Ablation study of the SERGCN model in PIPA dataset.

Methods	domain	relation
SERGCN (pretrained scene)	75.6%	68.9%
SERGCN (w.o. scene inf.)	75.2%	68.0%
SERGCN	**77.2%**	**69.5%**

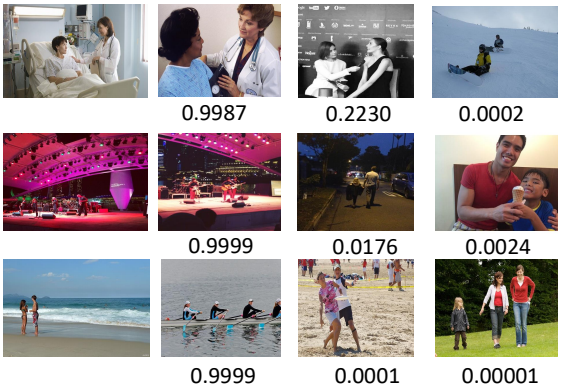

Figure 5. Visualization of images and corresponding feature similarity from SSL finetuned model. The images in the first column are the query images, each of which is followed by three images, sorted by the feature similarity in descending order.

To further compare GR^{2}N and RGCN in terms of performance and inference time, we conduct experiments by replacing RGCN model with GR^{2}N in our SERGCN while keeping other components the same as SERGCN. 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 SERGCN with this new interactive feature. We denote this setting as SEGR^{2}N. The experimental results on the accuracy and inference time on test set of both algorithms in PIPA dataset are shown in Table 4. We can observe that the performance of SEGR^{2}N is comparable to SERGCN. However, in terms of testing time, the model with GR^{2}N is much slower as compared to SERGCN. These results strongly support our conclusion that the proposed RGCN model is much simpler and faster as compared to GR^{2}N.

Table 4. Comparisons of GR^{2}N and RGCN in PIPA dataset.

Methods	domain		relation	
	accuracy	time	accuracy	time
SEGR^{2}N	75.6%	3.33s	69.1%	6.23s
SERGCN	**77.2%**	1.91s	**69.5%**	1.82s

4.6.2 SERGCN: Power of SSL

Experiment results show the power of SSL in SERGCN. For instance, in the setting of SERGCN without scene information in Table 3, our model outperforms SERGCN without scene information by 2.7% and SERGCN (pretrained scene) by 2.1% in domain recognition in PIPA dataset. We observe similar results in PISC dataset.

5. Conclusion

In this paper, we have originally proposed a deep model (i.e., SERGCN) to enhance social relation inference by self-supervised learning based on image scene classification. The task of image scene classification captures the basic feature of the whole context and thus enhances the downstream task of social relation understanding between two persons. To learn human interaction among a group of people, we also designed a simple and fast graph convolutional network called RGCN. Because the common two datasets are small-size in social relation inference, we have collected and distributed a new large-scale dataset consisting of about 240 thousand unlabeled images for self-supervised learning. Our proposed model achieved superior performance than state-of-the-art methods for classification in regular datasets of social relation.

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] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19:153–160, 2006. 3

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

[4] 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

[5] Uta Buchler, Biagio Brattoli, and Bjorn Ommer. Improving spatiotemporal self-supervision by deep reinforcement learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 770–786, 2018. 3

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[8] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2051–2060, 2017. 3

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

[10] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 3

[11] 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. 1, 3, 5, 6, 7, 8

[12] John Gottman, Robert Levenson, and Erica Woodin. Facial expressions during marital conflict. *Journal of Family Communication*, 1(1):37–57, 2001. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[14] 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. 3

[15] 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

[16] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[17] 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

[18] 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

[19] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Advances in Neural Information Processing Systems*, pages 7763–7774, 2018. 3

[20] 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, 6, 7, 8

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

[22] 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

[23] 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

[24] 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

[25] Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Cross pixel optical-flow similarity for self-supervised learning. In *Asian Conference on Computer Vision*, pages 99–116. Springer, 2018. 3

[26] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 3

[27] T Nathan Mundhenk, Daniel Ho, and Barry Y Chen. Improvements to context based self-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9339–9348, 2018. 3

[28] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 3

[29] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–771, 2018. 3

[30] Michael S Ryoo and Jake K Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th international conference on computer vision*, pages 1593–1600. IEEE, 2009. 2

[31] 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

[32] 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. 3

[33] 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, 3

[34] 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. 1, 2, 3, 6, 7, 8

[35] 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

[36] 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. 5

[37] Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2555, 2019. 3

[38] 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. 1, 6, 7, 8

[39] 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. 3, 6

[40] 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, 3

[41] 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