International Journal of Computer Vision
https://doi.org/10.1007/s11263-020-01295-1

Visual Social Relationship Recognition

Junnan Li[1] · Yongkang Wong[2] · Qi Zhao[3] · Mohan S. Kankanhalli[2]

Received: 25 May 2018 / Accepted: 22 January 2020

Abstract

Social relationships form the basis of social structure of humans. Developing computational models to understand social relationships from visual data is essential for building intelligent machines that can better interact with humans in a social environment. In this work, we study the problem of visual social relationship recognition in images. We propose a dual-glance model for social relationship recognition, where the first glance fixates at the person of interest and the second glance deploys attention mechanism to exploit contextual cues. To enable this study, we curated a large scale People in Social Context dataset, which comprises of 23,311 images and 79,244 person pairs with annotated social relationships. Since visually identifying social relationship bears certain degree of uncertainty, we further propose an adaptive focal loss to leverage the ambiguous annotations for more effective learning. We conduct extensive experiments to quantitatively and qualitatively demonstrate the efficacy of our proposed method, which yields state-of-the-art performance on social relationship recognition.

Keywords Social relationship · Label ambiguity · Context-driven analysis · Attention

1 Introduction

Since the beginning of early civilizations, social relationships derived from each individual fundamentally form the basis of social structure in our daily life. Today, apart from social interactions that occur in physical world, people also communicate through various social media platforms, such as Facebook and Instagram. Large amount of images and videos have been uploaded to the internet that explicitly and implicitly capture people's social relationship information. Humans can naturally interpret the social relationships of people in a scene. In order to build machines with intelligence, it is necessary to develop computer vision algorithms that can interpret social relationships.

Enabling computers to understand social relationships from visual data is important for many applications. First, it enables users to pose a socially meaningful query to an image retrieval system, such as 'Grandma playing with grandson'. Second, visual privacy advisor systems (Orekondy et al. 2017) can alarm users about potential privacy risks if the posted images contain sensitive social relationships. Third, robots can better interact with people in daily life by inferring people's characteristics and possible behaviors based on their social relationships. Last but not least, surveillance systems can better analyse human behaviors with the understanding of social relationships.

In this work, we aim to build computational models that address the problem of visual social relationship recognition in images. We start by defining a set of social relationship categories. With reference to the *relational models* theory (Fiske 1992) in social psychology literature, we define a hierarchical social relationship categories which embed the coarse-to-fine characteristic of common social relationships (as illustrated in Fig. 1). Our definition follows a prototype-based approach, where we are interested in finding exemplars that parsimo-

Communicated by Ivan Laptev.

✉ Junnan Li
lijunnan@u.nus.edu

Yongkang Wong
yongkang.wong@nus.edu.sg

Qi Zhao
qzhao@cs.umn.edu

Mohan S. Kankanhalli
mohan@comp.nus.edu.sg

[1] Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore 117456, Singapore

[2] School of Computing, National University of Singapore, Singapore 117417, Singapore

[3] Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Published online: 03 February 2020

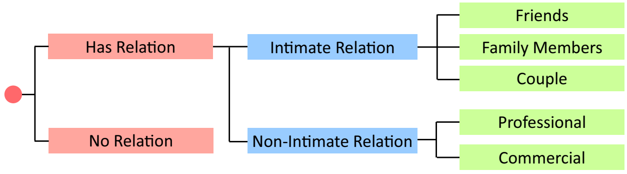


Fig. 1 Defined hierarchical social relationship categories

niously describe the most common situations, rather than an abstract definition that could cover all possible cases.

Social relationship recognition from images is a challenging task for several reasons. First, images have wide variations in scale, scene, human pose and appearance, as well as occlusions. Second, humans infer social relationships not only based on the physical appearance (e.g., color of clothes, gender, age, etc.), but also from subtler cues (e.g., expression, proximity, and context) (Alletto et al. 2014; Ramanathan et al. 2013; Zhang et al. 2015b). Third, a pair of people in an image might have multiple plausible social relationships, as shown in Fig. 2. While previous works on social relationship recognition only consider the majority consensus (Li et al. 2017a; Sun et al. 2017), it remains a challenging issue to make use of the ambiguity in social relationship labels.

A preliminary version of this work was published earlier (Li et al. 2017a). We have extended this work in the following manner: First, we propose a novel adaptive focal loss, that addresses label ambiguity challenge and class imbalance problem in training. Second, we improve the dual-glance model in (Li et al. 2017a) with network modifications (see Sect. 3.2). Third, we conduct additional experiments on two dataset [i.e. People in Social Context (Li et al. 2017a) and Social Domain and Relation (Sun et al. 2017)], and achieve significant performance improvement over previous methods.

The key contributions can be summarized as:

- We propose a dual-glance model, that mimics the human visual system to explore useful and complementary visual cues for social relationship recognition. The first glance fixates at the individual person pair of interest, and performs prediction based on its appearance and geometrical information. The second glance exploits contextual cues from regions generated by Region Proposal Network (RPN) (Ren et al. 2015) to refine the prediction.
- We propose a novel Attentive R-CNN. Given a person pair, the attention is selectively assigned on the informative contextual regions. The attention mechanism is guided by both bottom-up and top-down signals.
- We propose a novel adaptive focal loss. It leverages the embedded ambiguity in social relationship annotations to adaptively modulate the loss and focuses training on hard examples. Performance is improved compared to using other loss functions.
- To study social relationships, we collected the People in Social Context (PISC) dataset. It consists of 23,311 images and 79,244 person pairs with manually labeled social relationship labels. In addition, PISC consists of 66 annotated occupation categories.
- We perform experiments with ablation studies on PISC and the Social Domain and Relation (SDR) (Sun et al. 2017) dataset, where we quantitatively and qualitatively validate the proposed method.

The remainder of the paper is organized as follows. First, we review the related work in Sect. 2. Then we elaborate on the proposed dual-glance model in Sect. 3, and the adaptive focal loss in Sect. 4. Section 5 details the PISC dataset, whereas the experiment details and results are delineated in Sect. 6. Section 7 concludes the paper.

2 Related Work

2.1 Social Relationship

The study of social relationships lies at the heart of social sciences. Social relationships are the cognitive sources for generating social action, for understanding individual's social behavior, and for coordinating social interaction (Haslam and Fiske 1992). There are two forms of representations for relational cognition. The first approach represents relationship with a set of theorized or empirically derived dimensions (Conte and Plutchik 1981). The other form of representation proposes implicit categories for relation cognition (Haslam 1994). One of the most widely accepted categorical theory is the *relational models* theory (Fiske 1992). It offers a unified account of social relations by proposing four elementary prototypes, namely *communal sharing*, *equality matching*, *authority ranking*, and *market pricing*. In this work, inspired by the relational models theory, we identify 5 exemplar relationships that are common in daily life and visually distinguishable (i.e. friends, family members, couple, professional and commercial). We group them into two relation domains, namely intimate relation and non-intimate relation, as illustrated in Fig. 1.

In the computer vision literature, social information has been widely adopted as supplementary cues in several tasks. Gallagher and Chen (2009) extract features describing group structure to aid demographic recognition. Shao et al. (2013) use social context for occupation recognition in photos. Qin and Shelton (2016) exploit social grouping for multi-target tracking. For group activity recognition, social roles and relationship information have been implicitly embedded into the inference model (Choi and Savarese 2012; Deng et al. 2016;

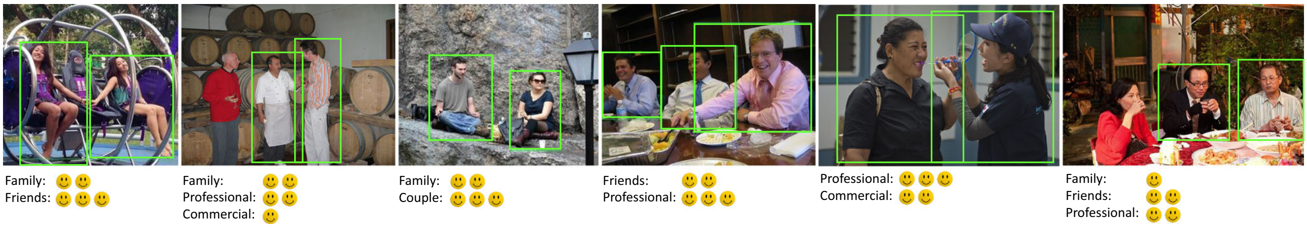


Fig. 2 Example of images where annotators do not agree on a single social relationship class

Direkoglu and O'Connor 2012; Lan et al. 2012a; Lan et al. 2012b). Alletto et al. (2014) define 'social pairwise feature' based on F-formation and use it for group detection in egocentric videos. Recently, Alahi et al. (2016) and Robicquet et al. (2016) model social factor for human trajectory prediction.

Many studies focus on relationships among family members, such as siblings, husband-wife, parent-child and grandparent-grandchild. Such studies include kinship recognition (Wang et al. 2010; Chen et al. 2012; Guo et al. 2014; Shao et al. 2014; Xia et al. 2012) and kinship verification (Fang et al. 2010; Xia et al. 2012; Dibeklioglu et al. 2013) in group photos. Most of these works leverage facial information to infer kinship, including the location of faces, facial appearance, attributes and landmarks. Zhang et al. (2015b) discover relation traits such as "warm", "friendly" and "dominant" from face images. Another relevant topic is intimacy prediction (Yang et al. 2012; Chu et al. 2015) based on human poses.

For video based social relation analysis, Ding and Yilmaz (2014) discover social communities formed by actors in movies. Marín-Jiménez et al. (2014) detect social interactions in TV shows, whereas Yun et al. (2012) study human interaction in RGBD videos. Ramanathan et al. (2013) study social events and discover pre-defined social roles in a weakly supervised setting (e.g. birthday child in a birthday party). Lv et al. (2018) propose to use multimodal data for social relation classification in TV shows and movies. Fan et al. (2018) analyze shared attention in social scene videos. Vicol et al. (2018) construct graphs to understand the relationships and interactions between people in movies.

Our study also partially overlaps with the field of social signal processing (Vinciarelli et al. 2012), which aims to understand social signals and social behaviors using multiple sensors. Such works include interaction detection, role recognition, influence ranking, personality recognition, and dominance detection in group meeting (Gan et al. 2013; Hung et al. 2007; Rienks et al. 2006; Salamin et al. 2009; Alameda-Pineda et al. 2016).

Very recently, Li et al. (2017a) and Sun et al. (2017) studied social relationship recognition in images. We (Li et al. 2017a) propose a dual-glance model with Attentive R-CNN to exploit contextual cues, whereas Sun et al. (2017) leverage semantic attributes learnt from other dataset as intermediate representation to predict social relationships. Two datasets have been collected, namely the PISC dataset (Li et al. 2017a) and the SDR dataset (Sun et al. 2017) (see detailed comparison in Sect. 5). In this paper, we extend our work (Li et al. 2017a) with adaptive focal loss, improved dual-glance model, and additional experiments on both datasets.

2.2 Region-Based Convolutional Neural Networks

The proposed Attentive R-CNN incorporates Faster R-CNN (Ren et al. 2015) pipeline with attention mechanism to extract information from multiple contextual regions. The Faster R-CNN pipeline has been widely exploited by many researchers. Gkioxari et al. (2015) propose R*CNN, that makes use of a secondary region in an image for action recognition. Johnson et al. (2016) study dense image captioning that focuses on the regions. Li et al. (201b) adopt the Faster R-CNN pipeline as basis framework to study the joint task of object detection, scene graph generation and region captioning.

Attention model has been recently proposed and applied to image captioning (Xu et al. 2015; You et al. 2016), visual question answering (Yang et al. 2016) and fine-grained classification (Xiao et al. 2015). In this work, we employ attention mechanism on the contextual regions, so that each person pair can selectively focus on its informative regions to better exploit contextual cues. Our attentive R-CNN can also be viewed as a soft Multiple-Instance Learning (MIL) approach (Maron and Lozano-Pérez 1997), where the model receives bags of instances (contextual regions) and bag-level labels (relationship class), and learns to discover informative instances for correct prediction.

2.3 Focal Loss

The proposed adaptive focal loss is inspired by the Focal Loss (Lin et al. 2017) for object detection. Focal loss is designed to address the imbalance in samples between foreground and background classes during training, where a modulating factor is introduced to down-weight the easy examples. Our adaptive focal loss not only addresses class imbalance, but more importantly, takes into account the uncertainty in visually identifying social relationship labels.

3 Proposed dual-glance Model

Given an image I and a target person pair highlighted by bounding boxes $\{b_1, b_2\}$, our goal is to infer their social relationship r. In this work, we propose a dual-glance relationship recognition model, where the first glance module fixates at b_1 and b_2, and the second glance module explores contextual cues from multiple region proposals $\mathbf{P}_\mathsf{I}$. The final score over possible relationships, $\mathbf{S}$, is computed via

$$\mathbf{S} = \mathbf{S}_1(\mathsf{I}, b_1, b_2) + \mathbf{w} \otimes \mathbf{S}_2(\mathsf{I}, b_1, b_2, \mathbf{P}_\mathsf{I}), \tag{1}$$

where $\mathbf{w}$ is a weight vector, and $\otimes$ is the element-wise multiplication of two vectors. We use softmax to transform the final score into a probability distribution. Specifically, the probability that a given pair of people having relationship r is calculated as

$$p_r = \frac{\exp(S_r)}{\sum_r \exp(S_r)}. \tag{2}$$

An overview of the proposed dual-glance model is shown in Fig. 3.

3.1 First Glance Module

The first glance module takes in input image I and two human bounding boxes. First, we crop three patches from I and refer them as p_1, p_2, and $\mathsf{p}_\cup$. p_1 and p_2 each contains one person, and $\mathsf{p}_\cup$ contains the union region that tightly covers both people. The three patches are resized to 224 × 224 pixels and fed into three CNNs, where the CNNs that process p_1 and p_2 share the same weights. The outputs from the last convolutional layer of the CNNs are flattened and concatenated.

We denote the geometry feature of the human bounding box b_i as $\mathbf{b}_i^{\text{loc}} = \{x_i^{min}, y_i^{min}, x_i^{max}, y_i^{max}, area_i\} \in \mathbb{R}^5$, where all the parameters are relative values, normalized to zero mean and unit variance. $\mathbf{b}_1^{\text{loc}}$ and $\mathbf{b}_2^{\text{loc}}$ are concatenated and processed by a fully-connected (fc) layer. We concatenate its output with the CNN features for p_1, p_2 and $\mathsf{p}_\cup$ to form a single feature vector, which is subsequently passed through another two fc layers to produce first glance score, $\mathbf{S}_1$. We use $\mathbf{v}_{\text{top}} \in \mathbb{R}^k$ to denote the output from the penultimate fc layer. $\mathbf{v}_{\text{top}}$ serves as a top-down signal to guide the attention mechanism in the second glance module. We set $k = 4096$ with the same dimension as the regional features in Attentive R-CNN.

3.2 Attentive R-CNN for Second Glance Module

For the second glance module, we adapt Faster R-CNN (Ren et al. 2015) to make use of multiple contextual regions. Faster R-CNN processes the input image I with Region Proposal Network (RPN) to generate a set of region proposals $\mathbf{P}_\mathsf{I}$ with high objectness. For each person pair with bounding boxes b_1 and b_2, we select the set of contextual regions $\mathbf{R}(b_1, b_2; \mathsf{I})$ from $\mathbf{P}_\mathsf{I}$ as

$$\mathbf{R}(b_1, b_2; \mathsf{I}) = \{c \in \mathbf{P_I} : \max(G(c, b_1), G(c, b_2)) < \tau_u\} \tag{3}$$

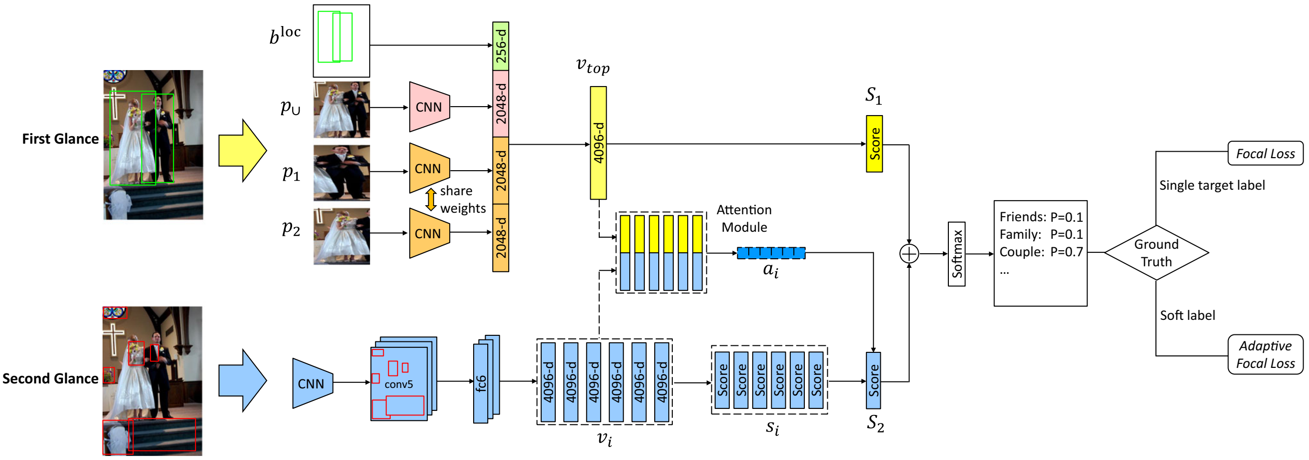


Fig. 3 An overview of the proposed dual-glance model. The first glance module fixates at the target person pair and outputs a score. The second glance module explores contextual regions, allocates attention to each region, and aggregates regional scores in a weighted manner. The attention is guided by both top-down signal from the first glance, and bottom-up signal form the local region. During training stage, if the supervision is a hard label (majority vote), we use the focal loss. If the supervision is a a soft label (distribution over classes), we use the proposed adaptive focal loss

where $G(b_1, b_2)$ computes the Intersection-over-Union (IoU) between two regions, and τ_u is the upper threshold for IoU overlap. The threshold encourages the second glance module to explore cues different from that of the first glance module.

We then process I with a CNN to generate a convolutional feature map $conv(\mathsf{I})$. For each contextual region $c \in \mathbf{R}$, ROI pooling is applied to extract a fixed-length feature vector from $conv(\mathsf{I})$, which is then processed by a fc layer to generate regional feature $\mathbf{v} \in \mathbb{R}^k$. We denote $\{\mathbf{v}_i | i = 1, 2, \dots, N\}$ as the bag of N regional feature vectors for $\mathbf{R}$. Each regional feature is then fed to another fc layer to generate a score for the ith region proposal:

$$\mathbf{s}_i = \mathsf{W}_s \mathbf{v}_i + \mathbf{b}_s. \tag{4}$$

Not all contextual regions are informative for the target person pair's relationship. Therefore we assign different attention to the region scores so that more informative regions could contribute more to the final prediction. In order to compute the attention, we first take each local regional feature $\mathbf{v}_i$, and combine it with the top-down feature from the first glance module $\mathbf{v}_{\text{top}}$ (which contains semantic information of the person pair) into a vector $\mathbf{h}_i \in \mathbb{R}^k$ via

$$\mathbf{h}_i = ReLU(\mathbf{v}_i + \mathbf{w}_{\text{top}} \otimes \mathbf{v}_{\text{top}}), \tag{5}$$

where $\mathbf{w}_{\text{top}} \in \mathbb{R}^k$, and $\otimes$ is the element-wise multiplication.

Then, we calculate the attention $a_i \in [0, 1]$ over the ith regional score with the sigmoid function:

$$a_i = \frac{1}{1 + \exp(-(\mathsf{W}_{h,a}\mathbf{h}_i + b_a))}, \tag{6}$$

where $\mathsf{W}_{h,a} \in \mathbb{R}^{1\times k}$ is the weight matrix, and $b_a \in \mathbb{R}$ is the bias term.

Given the attention, the output score of the second glance module is computed as a weighted average of all regional scores:

$$\mathbf{S}_2 = \frac{1}{N}\sum_{i=1}^{N} a_i \mathbf{s}_i. \tag{7}$$

Note that the dual-glance model described above has several differences compared with our previously proposed model (Li et al. 2017a): (i) We add a new fc6 layer in the Attentive R-CNN model to increase the depth of the network. (ii) We add *ReLU* non-linearity to compute $\mathbf{h_i}$, which introduces sparse representation that is more robust. (iii) We modify (1) to use element-wise weighting instead of a scalar weight, so that the network can learn to better fuse the scores. Those modifications can individually improve the performance, and together they lead to +0.7% improvement in mAP for relationship recognition while other settings remain the same as Li et al. (2017a).

4 Adaptive Focal Loss

Given a target person pair, our proposed dual-glance model outputs a probability distribution $\mathbf{p}$ over the relationships. In order to train the model to predict higher probability p_t for the ground truth target relationship t, the standard loss function adopted by Li et al. (2017a) and Sun et al. (2017) is the cross entropy (CE) loss defined as

$$\text{CE}(\mathbf{p}, t) = -\log p_t. \tag{8}$$

In the task of social relationship recognition, there often exists class imbalance in the training data. The classes with more samples can overwhelm the loss and lead to degenerate models. Previous work addresses this with a heuristic sampling strategy to maintain a manageable balance during training (Li et al. 2017a). Recently, in the field of object detection, focal loss (FL) has been proposed (Lin et al. 2017), where a modulating factor $(1 - p_t)^\gamma$ is added to the cross entropy loss:

$$\text{FL}(\mathbf{p}, t) = -(1 - p_t)^\gamma \log p_t. \tag{9}$$

The modulating factor down-weights the loss contribution from the vast number of well-classified examples, and focuses on the fewer hard examples, where the focusing parameter γ adjusts the rate at which easy examples are down-weighted.

In a wide range of visual classification tasks [e.g. image classification (Russakovsky et al. 2015), object detection (Lin et al. 2014), visual relationship recognition Krishna et al. 2017,etc.], the common approach to determine the ground truth class of a sample is to take the majority vote from human annotations. While this approach has been effective, we argue that social relationship recognition is different from other tasks. The annotation of social relationship has a higher level of uncertainty (as suggested by the agreement rate in Sect. 5.2), and the minority annotations are not necessarily wrong (as shown in Fig. 2). Therefore, taking the majority vote and ignoring other annotations has the potential disadvantage of neglecting useful information.

In this work, we propose an adaptive focal loss that takes into account the ambiguity in social relationship labels. For each sample, instead of using the hard label from majority voting, we transform the annotations into a soft label $\mathbf{p}^y$, which is a distribution calculated by dividing the number of annotations for each relation r with the total number of annotations for that sample. Then we define the adaptive FL as

$$\text{Ada-FL}(\mathbf{p}, \mathbf{p}^y) = -\sum_r \max((p_r^y - p_r), 0)^\gamma \log p_r. \quad (10)$$

The adaptive FL inherits the ability to down-weight easy examples from the FL, and extends the FL with two properties to address label ambiguity: (i) Instead of considering only the single target class, the adaptive FL takes the sum of losses from all classes, so that all annotations can contribute to training. (ii) The modulating factor $\max((p_r^y - p_r), 0)$ is adaptively adjusted for each class based on the ground truth label distribution. The loss still demands the model to predict high probability for the predominant class, but the constraint is relaxed if not all annotations agree. For example, if 4 out of the 5 annotators agree on *friends* as the label, the adaptive FL term for $r = \textit{friends}$ will decrease to 0 if output $p_{\textit{friends}} \geq 0.8$, hence it will push $p_{\textit{friends}}$ to 0.8 instead of 1. Note that if the ground truth annotations all agree on the same class t, then $p_t^y = 1$ and $p_r^y = 0$ for $r \neq t$, the adaptive FL is the same as the FL.

The same philosophy of learning from ambiguous label distributions has also been studied by Gao et al. (2017), where they use the Kullback-Leibler (KL) divergence loss defined as

$$\begin{aligned} \text{KL-div}(\mathbf{p}, \mathbf{p}^y) &= \sum_r p_r^y \log \frac{p_r^y}{p_r} \\ &= -\sum_r p_r^y \log p_r + \sum_r p_r^y \log p_r^y \\ &= \text{CE}(\mathbf{p}, \mathbf{p}^y) - \text{H}(\mathbf{p}^y), \quad (11) \end{aligned}$$

where $\text{CE}(\mathbf{p}, \mathbf{p}^y)$ is the cross entropy between the output distribution and the label distribution, and $\text{H}(\mathbf{p}^y)$ is the entropy of the label distribution. Since $\text{H}(\mathbf{p}^y)$ is independent of the parameters of the model, minimizing $\text{KL-div}(\mathbf{p}, \mathbf{p}^y)$ is equivalent to minimizing $\text{CE}(\mathbf{p}, \mathbf{p}^y)$.

The difference between KL divergence and the proposed adaptive focal loss is the per-class modulating factor. While KL divergence uses the ground truth label distribution p_r^y to modulate the per-class loss, adaptive focal loss uses both p_r^y and the model's output p_r to determine modulation, thereby down-weighting the easy examples and focusing training on the hard examples.

In practice, similar as Lin et al. (2017), we use an α-balanced variant of the adaptive FL defined as

$$\text{Ada-FL}(\mathbf{p}, \mathbf{p}^{\mathbf{y}}) = -\sum_r \alpha_r \max((p_r^y - p_r), 0)^\gamma \log p_r. \quad (12)$$

$\alpha_r \in [0, 1]$ is determined by inverse class frequency via

$$\alpha_r = \left(\frac{\min(L_1, L_2, \ldots, L_R)}{L_r}\right)^\beta, \quad (13)$$

where L_r is the total number of annotations for relationship r, and β is set to be 0.5 as a smoothing factor. We find that the α-balanced adaptive focal loss yields slightly better performance over the non-α-balanced form.

5 People in Social Context Dataset

The People in Social Context (PISC) dataset is an image dataset that focuses on social relationship study (see example images in Fig. 4). In this section, we first describe the data curation pipeline. Then we analyze the dataset statistics and provide comparison with another dataset for social relationship study, following the presentation style by Goyal et al. (2017) and Agrawal et al. (2018).

5.1 Curation Pipeline

The PISC dataset was curated through a pipeline of three stages. In the first stage, we collected around 40k images that contain people from a variety of sources, including Visual Genome (Krishna et al. 2017), MSCOCO (Lin et al. 2014), YFCC100M (Thomee et al. 2016), Flickr, Instagram, Twitter and commercial search engines (i.e. Google and Bing). We used a combination of key words search (e.g. co-worker, people, friends, etc.) and people detector (Faster R-CNN Ren et al. 2015) to collect the image. The collected images have high variation in image resolution, people's appearance, and scene type.

In the second and third stage, we hired workers from CrowdFlower platform to perform labor intensive manual annotation task. The second stage focused on the annotation of person bounding box in each image. Following Krishna et al. (2017), each bounding box is required to strictly satisfy the coverage and quality requirements. To speed up the

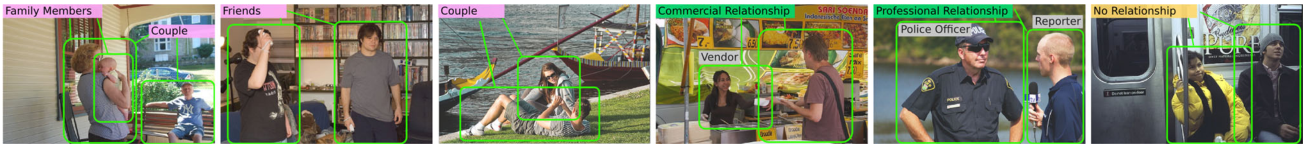


Fig. 4 Example images from the People in Social Context (PISC) dataset

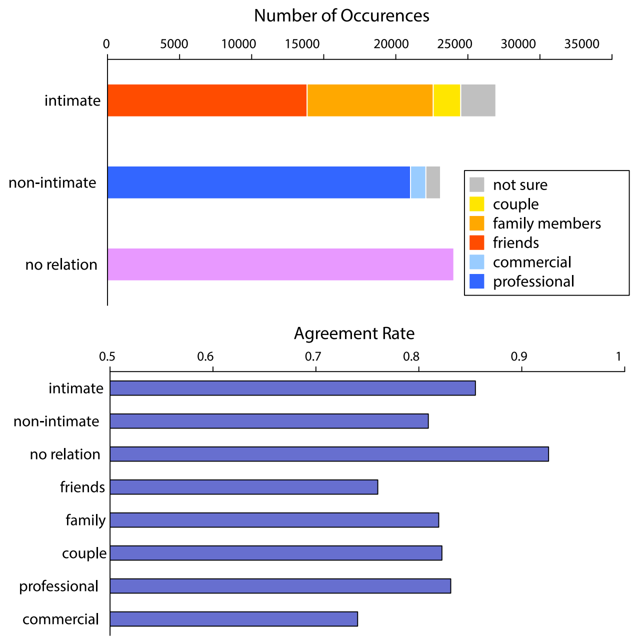


Fig. 5 Annotation statistics of the relationship categories

annotation process, we first deployed Faster R-CNN (Ren et al. 2015) to detect people on all images, followed by asking the annotators to re-annotate the bounding boxes if the computer-generated bounding boxes were inaccurately localized. Overall, 40% of the computer-generated boxes are accepted without re-annotation. For images collected from MSCOCO and Visual Genome, we directly used the provided groundtruth bounding boxes.

Once the bounding boxes of all images had been annotated, we selected images consisting of at least two people, and avoided images that contain crowds of people where individuals cannot be distinguished. In the final stage, we requested the annotators to identify the occupation of all individuals in the image, as well as the social relationships of all person pairs. To ensure consistency in the occupation categories, the annotation is based on a list of reference occupation categories. The annotators could manually add a new occupation category if it was not in the list.

For social relationships, we formulate the annotation task as multi-level multiple choice questions based on the hierarchical structure in Fig. 1. We provide example images to help annotators understand different relationship classes. We also provide instructions to help annotators distinguish between professional[1] and commercial relationship[2]. Annotators can choose the option 'not sure' at any level if they cannot confidently identify the relationship. Each image was annotated

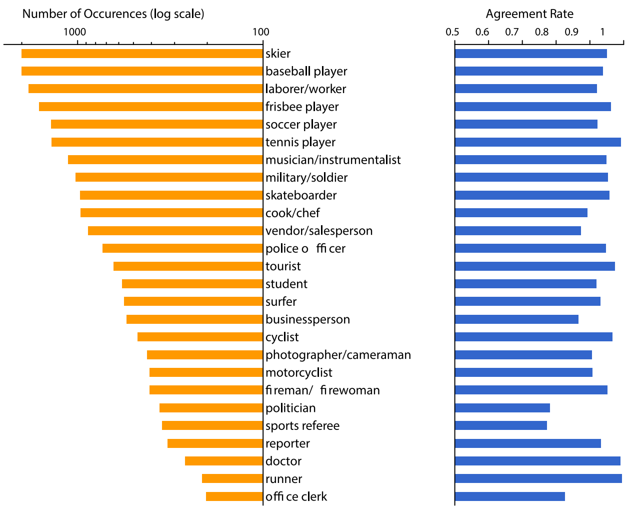


Fig. 6 Annotation statistics of the top 26 occupations

by at least five workers, Overall, 7928 unique workers have contributed to the annotation.

5.2 Dataset Statistics

In total, the PISC dataset consists of 23,311 images with 79,244 pairs of people. For each person pair, if there exists a relationship class which at least 60% of the annotators agree on, we refer it as a 'consistent' example and assign the majority vote as its class label. Otherwise we refer it as an 'ambiguous' example. The top part of Fig. 5 shows the distribution of each type of relationships. We further calculate the agreement rate on the consistent set by dividing the number of agreed human annotations with the total number of annotations. As shown in the bottom part of Fig. 5, the agreement rate reflects how visually distinguishable a social relationship class is. The rate ranges from 74.1 to 92.6%, which indicates that social relationship recognition has certain degree of ambiguity, but is a visually solvable problem nonetheless.

For occupations, 10,034 images contain people that have recognizable occupations. In total, there are 66 identified occupation categories. The occupation occurrence and the agreement rate for the 26 most frequent occupation categories are shown in Fig. 6. Since two source datasets, i.e. MSCOCO and Visual Genome, are highly biased towards 'baseball player' and 'skier', we limit the total number of instances per occupation to 2000 based on agreement rate ranking to ensure there are no bias towards any particular occupation.

5.3 Comparison with SDR Dataset

The Social Domain and Relation (SDR) dataset (Sun et al. 2017) is a subset of the PIPA dataset (Zhang et al. 2015a)

[1] The people are related based on their professions (e.g. co-worker, coach and player, boss and staff, etc).

[2] One person is paying money to receive goods/service from the other (e.g. salesman and customer, tour guide and tourist, etc).

Table 1 Comparison between PISC and SDR (Sun et al. 2017) dataset

Dataset	PISC	SDR (Sun et al. 2017)
Image source	Wide variety (see Sect. 5.1)	Flickr photo album
Number of image	23,311	8570
Number of person pair	79,244	26,915
Person's identity	Different images, different people	Multiple images, same person
Person's bounding box	Full-body	Head only

with social relation annotation. Table 1 provides the details of both datasets. In comparison, our PISC dataset has multiple advantages. First and foremost, the PISC dataset contains more images and more person pairs. Second, the images in SDR dataset all come from Flickr photo albums, while our images are collected from a wide variety of sources. Therefore, the images in PISC dataset are more diverse. Third, since the images in SDR dataset were originally collected for the task of people identification (Zhang et al. 2015a), the same person would appear in multiple images, which further reduce the diversity of the data. Last but not least, our PISC dataset provides full-body person bounding box annotation, while SDR dataset provides the head bounding box and uses that to approximate the body bounding box.

6 Experiment

In this section, we perform experiments and ablation studies to fully demonstrate the efficacy of the proposed method on both PISC and SDR dataset. We first delineate the dataset and training details, followed by experiment details and discussion.

6.1 Dataset Details

PISC On the collected PISC dataset, we perform two tasks, namely domain recognition (i.e. *Intimate* and *Non-Intimate*) and relationship recognition (i.e. *Friends, Family, Couple, Professional* and *Commercial*). We refer to each person pair as one sample. For domain recognition, we randomly select 4000 images (15,497 samples) as test set, 4000 images (14,536 samples) as validation set and use the remaining images (49,017 samples) as training set. For relationship recognition, since there exists class imbalance in the data, we sampled the test and validation split to have balanced class. To do that, we select 1250 images (250 per relation) with 3961 samples as test set and 500 images (100 per relation) with 1505 samples as validation set. The remaining images (55,400 samples) are used as training set.

All the samples used above are selected only from the consistent samples, where each relationship sample are agreed by a majority of annotators. For the relationship recognition task, we enrich the consistent training set with ambiguous samples to create an ambiguous training set. It contains a total of 58,885 samples, or 3445 samples more than the consistent training set.

SDR The SDR dataset is annotated with 5 domains and 16 relationships (Sun et al. 2017). However, the class imbalance is severe for the relationship classes. 7 out of the 16 classes have no more than 40 unique individuals. In the test set, 4 classes have less than 20 samples (person pairs). In the validation set, 6 classes have no more than 5 samples. We tried to re-partition the dataset, but the issue that a same person appears across multiple images makes it very difficult to form a test and validation set with reasonable class balance. Therefore, we only perform domain recognition task, where the imbalance is less severe. The 5 domains include *Attachment, Reciprocity, Mating, Hierarchical power* and *Coalitional groups*. Note that the samples in SDR dataset are all consistent samples.

6.2 Training Details

In the following experiments, we experiment with both *Focal Loss* using hard label as supervision and the proposed *adaptive focal loss* using soft label distribution as supervision. We set the focusing parameter γ to be 2 in focal loss and 1 in adaptive focal loss, which yield best performance respectively. Unless otherwise specified, Sect. 6.3 uses focal loss on the consistent training set, Sect. 6.4 experiments with various loss functions, Sects. 6.5–6.7 use adaptive focal loss on the ambiguous training set.

We employ pre-trained CNN models to initialize our dual-glance model. For the first glance, we fine-tune the ResNet-101 model (He et al. 2016). For the second glance, we fine-tune the Faster R-CNN model with VGG-16 as backbone (Ren et al. 2015). We employ two-stage training, where we first train the first-glance model until the loss converges, then we freeze the first-glance model, and train the second-glance model. We train our model with Stochastic Gradient Descent and backpropagation. We set learning rate as 0.01, batch size as 32, and momentum as. During training, we use two data augmentation techniques: (1) horizontally flipping the image, and (2) reversing the input order of a person pair

Table 2 Mean average precision (mAP%) and per-class recall of baselines and proposed dual-glance model on PISC dataset

Method	Domain			Relationship					
	mAP	Intimate	Non-intimate	mAP	Friends	Family	Couple	Professional	Commercial
Union (Lu et al. 2016)	75.2	81.5	75.3	49.3	42.7	52.5	45.0	70.2	49.4
Location	45.5	77.2	35.5	24.1	19.4	11.9	46.3	72.1	3.5
Pair (Sun et al. 2017)	76.9	82.1	76.5	54.9	58.1	58.5	47.3	72.7	52.3
Pair + Loc.	77.7	82.7	77.2	56.9	42.9	60.4	61.7	80.3	54.8
Pair + Loc. + Union (first-glance)	**80.2**	83.4	78.6	**58.7**	45.2	68.4	67.2	78.3	58.8
Pair + Loc. + Global	79.4	83.1	78.5	58.3	44.1	68.2	65.4	81.1	57.8
R-CNN	76.0	81.7	76.6	53.6	55.7	57.8	31.6	85.5	42.6
All attributes (Sun et al. 2017)	78.1	82.9	77.6	57.5	46.5	59.7	63.2	80.1	55.0
Dual-glance	85.4	85.5	83.1	65.2	60.6	64.9	54.7	82.2	58.0
Dual-glance + occupation	**85.8**	85.8	83.5	**65.9**	60.1	63.6	55.2	87.9	61.1
Dual-glance + all attributes	85.5	85.2	83.6	65.4	58.9	67.8	59.4	81.5	57.7

Best results are given in bold

(i.e. if p_1 and p_2 are a couple, then p_2 and p_1 are also a couple.).

6.3 Baselines Versus Dual-Glance

We evaluate multiple baselines and compare them to the proposed dual-glance model to show its efficacy. Formally, the compared methods are as followed:

1. **Union** Following the predicate prediction model by Lu et al. (2016), we use a CNN model that takes the union region of the person pair as input, and outputs their relationship.
2. **Location** We only use the geometry feature of the two individuals' bounding boxes to infer their relationship.
3. **Pair** The model consists of two CNNs with shared weights. The inputs are two cropped image patches for the two individuals. The model is similar to the **End-to-end Finetuned** double-stream CaffeNet in (Sun et al. 2017), except that Sun et al. (2017) don't share weights.
4. **Pair + Loc.** We extend Pair by using the geometry feature of the two bounding boxes.
5. **Pair + Loc. + Union** first-glance model illustrated in Fig. 3, which combines Pair + Loc. with Union.
6. **Pair + Loc. + Global** Model structure is the same as first-glance, except that we replace the union region with the entire image as global input.
7. **R-CNN** We train a R-CNN using the region proposals $\mathbf{R}(b_1, b_2; \mathrm{I})$ in (3), and use average pooling to combine the regional scores.
8. **All Attributes** (Sun et al. 2017) We follow the method by Sun et al. (2017) and extract 9 semantic attributes (age, gender, location&scale, head appearance, head pose, face emotion, clothing, proximity, activity) using models pre-trained on multiple annotated datasets. Then a linear SVM is used for classification. The SVM is calibrated to produce probabilities for calculating mAP. For attributes that require head bounding boxes (e.g. age, head pose, face emotion, etc.), we use a pre-trained head detector to find the head bounding box within each person's ground-truth body bounding box.
9. **Dual-Glance** Our proposed model (Sect. 3).
10. **Dual-Glance + Occupation** We first train a CNN for occupation recognition using the collected occupation labels. Then during social relationship training, we concatenate the occupation score (from the last layer of the trained CNN) for each person with the human-centric feature $\mathbf{v}_{\text{top}}$ as the new human-centric feature for the first glance.
11. **Dual-Glance + All Attributes** We fuse the score from baseline 8 with the score from the dual-glance model for the final prediction.

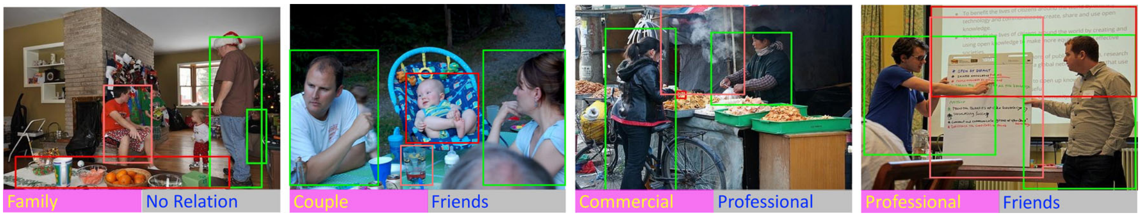
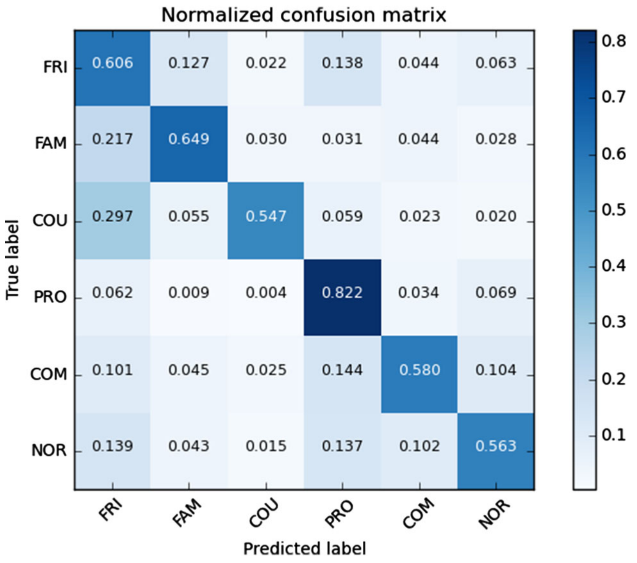


Fig. 7 Examples where dual-glance correctly predict the relationship (yellow label) while first-glance fails (blue label). GREEN boxes highlight target people pair, and the top two contextual regions with highest attention are shown in RED (Color figure online)



	FRI	FAM	COU	PRO	COM	NOR
FRI	0.606	0.127	0.022	0.138	0.044	0.063
FAM	0.217	0.649	0.030	0.031	0.044	0.028
COU	0.297	0.055	0.547	0.059	0.023	0.020
PRO	0.062	0.009	0.004	0.822	0.034	0.069
COM	0.101	0.045	0.025	0.144	0.580	0.104
NOR	0.139	0.043	0.015	0.137	0.102	0.563

Fig. 8 Confusion matrix of relationship recognition task using the proposed dual-glance model trained on PISC dataset

Table 3 Domain recognition result (%) on SDR dataset

Method	Accuracy
End-to-end finetuned (Sun et al. 2017)	59.0
All attributes (Sun et al. 2017)	67.8
First-glance	68.2
Dual-glance	72.1
Dual-glance + all attributes	**72.5**

Best result is given in bold

Table 2 shows the results for both domain recognition task and relationship recognition task on the PISC dataset. We can make several observations from the results. First, **Pair + Loc.** outperforms **Pair**, which suggests that peoples' geometric location in an image contains information useful to infer their social relationship. This is supported by the law of *proxemics* (b) which says people's interpersonal distance reflects their relationship. However, the location information alone cannot be used to predict relationship, as shown by the results of **Location**.

Second, adding **Union** to **Pair + Loc.** improves performance. The performance gain is lesser if we use the global context (entire image) rather than the union region. Third, using contextual regions is effective for relationship recognition. **R-CNN** achieves comparable performance to the first-glance model by using only contextual regions. The proposed **dual-glance** model outperforms the first-glance model by a significant margin (+ 5.2% for domain recognition, + 6.5% for relationship recognition).

Visual attributes also provide useful mid-level information for social relationship recognition. Combining **All Attributes** with **dual-glance** slightly improves performance, while **dual-glance + Occupation** achieves the best performance among all methods. However, **All Attributes** itself cannot outperform the proposed first-glance method. The reason is because of the unreliable attribute detection caused by frequently occluded head/face in the PISC dataset or the domain shift from source datasets (where the attribute detectors are trained) to target dataset (where the attribute detectors are applied, i.e. PISC).

Figure 7 shows some intuitive illustrations where the dual-glance model correctly classifies relationships that are misclassified by the first-glance model.

Figure 8 shows the confusion matrix of relationship recognition with the proposed dual-glance model, where we include *no relation* (NOR) as the 6th class. The model tends to confuse the intimate relationships, especially, misclassifying *family* and *couple* as *friends*.

Table 3 shows the result of domain recognition task on SDR dataset. **End-to-end Finetuned** (Sun et al. 2017) is a double-stream CNN model that uses the person pair as input, similar to our **Pair** except for weight sharing. **All Attributes** is the best-performing method by Sun et al. (2017), where a set of pretrained models from other dataset are used to extract semantic attribute representations (e.g. age, gender, activity, etc.), and a linear SVM is trained to classify relation using the semantic attributes as input. Compared with the results from Sun et al. (2017), both our first-glance and dual-glance yield better performance. While first-glance slightly outperforms **All Attributes**, dual-glance achieves more improvement by utilizing contextual regions.

Table 4 Relationship recognition result (mAP%) on PISC dataset with various loss functions and different level of ambiguity in training data

Loss function	Training set	Training supervision	First-glance	Dual-glance
Cross entropy	Consistent	Single label t	57.4	63.9
Focal loss[a]		Single label t	58.7	65.2
KL divergence		Soft label $\mathbf{p}^y$	58.7	65.1
Adaptive focal loss		Soft label $\mathbf{p}^y$	**59.7**	**66.4**
KL divergence	Ambiguous	Soft label $\mathbf{p}^y$	59.1	65.8
Adaptive focal loss		Soft label $\mathbf{p}^y$	**61.2**	**68.3**

Best results are given in bold

[a] Is the optimal results in Table 2

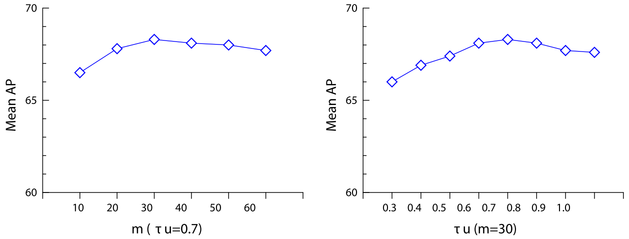


Fig. 9 Performance of dual-glance model on PISC dataset over variations in maximum number of region proposals (Left) and upper threshold of overlap between proposals and the person pair (Right)

6.4 Efficacy of Adaptive Focal Loss

In this section, we conduct relationship recognition experiment on the PISC dataset using various loss functions and two training data. We experiment with cross entropy loss (8), focal loss (9), KL divergence loss (11) and the proposed adaptive focal loss (10) on both consistent training set and ambiguous training set (see Sect. 6.1 for dataset details). Note that we use the α-balanced version for all losses while α is computed as in (13).

Table 4 shows the result. There are several observations we can make. First, comparing cross entropy loss and focal loss that both use single target label as training supervision, focal loss yields better performance (+ 1.3%). Second, adaptive focal loss achieves further improvement on focal loss. With dual-glance model, the improvement is + 1.2% in mAP if we use the same consistent training set. If we train on the ambiguous set, the improvement boosts to + 3.1%. Third, KL-divergence loss produces similar performance compared to focal loss on consistent set, and slight improvement on ambiguous set. On both training sets, KL-divergence gives lower mAP compared to adaptive focal loss. And last but not least, compared with the cross entropy loss by Li et al. (2017a), the proposed adaptive focal loss with ambiguous training set increases mAP by **+4.4%** using dual-glance model. The results demonstrate that the minority social relationship annotations do contain useful information, and the proposed adaptive focal loss can effectively exploit the ambiguous annotations for more accurate relationship recognition.

6.5 Variations in Contextual Regions

In order to encourage the attentive R-CNN to explore contextual cues that are not used by first-glance, we set a threshold τ_u in (3) to suppress regions that highly overlap with the person pair. Another influence factor in attentive R-CNN is the number of region proposals m from RPN, which can be controlled by a threshold on the objectness score. In this section, We experiment with different combinations of m and τ_u with the dual-glance model trained using adaptive focal loss on PISC dataset. As shown in Fig. 9, $m = 30$ and $\tau_u = 0.7$ produce the best performance on relationship recognition.

6.6 Ground Truth Versus Automatic People Detection

In this section, we study the propose method using ground truth annotation of person's bounding box. In other words we assume to possess a person detector that works as well as human. In this section, we test the robustness of our proposed method with automatic person detector. We employ Faster R-CNN (Ren et al. 2015) person detector pre-trained on MSCOCO dataset. Same as Ren et al. (2015), for each person in the test set, we treat all output boxes with $\not< 0.5$ IoU overlap with the ground truth box as positives, and apply greedy non-maximum suppression to select the highest scoring box as final prediction. In total, 3171 out of 3961 person pairs have been detected, while the average IoU overlap between detection boxes and ground truth is 79.7%.

Table 5 shows the relationship recognition result. Using automatic person detector leads to −1.5% decrease in mAP for first-glance model. The decrease is slighter for dual-glance model (− 0.8%), because the attentive R-CNN is less

Table 5 Relationship recognition result (mAP%) using different person bounding box on PISC dataset

Method	Ground truth	Faster R-CNN
First-glance	61.2	59.7
Dual-glance	68.3	67.5

Table 6 Relationship recognition result (mAP%) of the proposed dual-glance model with and without attention mechanism using various aggregation functions on PISC dataset

Without attention		With attention	
avg(·)	max(·)	avg(·)	max(·)
64.0	65.5	**68.3**	67.1

Best results is given in bold

affected by person's bounding box. The relatively insignificant performance decrease indicates that our proposed model is robust to person detection noise, and can be applied in a fully automatic setting.

6.7 Analysis on Attention Mechanism

In this section we demonstrate the importance of the attention mechanism on the proposed dual-glance model. We remove the attention module and experiment with two functions to aggregate regional scores, which are avg(·) and max(·).

Table 6 shows the relationship recognition result on PISC dataset. Adding attention mechanism leads to improvement for both avg(·) and max(·). The performance improvement is more significant for avg(·). For dual-glance without attention, max(·) performs best, While for dual-glance with attention, avg(·) performs best. This is because max(·) assumes that there exists a single contextual region that is most informative of the relationship, but sometimes there is no such region. On the other hand, avg(·) consider all regions, but could be distracted by irrelevant ones. However, with properly guided attention, avg(·) can better exploit the collaborative power of relevant regions for more accurate inference.

6.8 Visualization of Examples

The attention mechanism enables different person pairs to exploit different contextual cues. Some examples are shown in Fig. 10. Taking the images on the second row as an example, the little girl in red box is useful to infer that the other girl on her left and the woman on her right are family, but her existence indicates little of the couple in black.

Figure 11 shows examples of the misclassified cases. The model fails to pick up gender cue (misclassifies *friends* as *couple* in the image at row 3 column 3), or picks up the wrong cue (the white board instead of the vegetable in the image at row 2 column 3). Figure 12 shows examples of correct recognition for each relationship category in the PISC test set. We can observe that the proposed model learns to recognize social relationship from a wide range of visual cues including clothing, environment, surrounding people/animals, contextual objects, etc. For intimate relationships, the contextual cues varies from *beer* (friends), *gamepad* (friends), *TV* (family), to *cake* (couple) and *flowers* (couple). In terms of non-intimate relationships, the contextual cues are mostly related to the occupations of the individuals. For instance, *goods shelf* and *scale* indicate commercial relationship, while *uniform* and *documents* imply professional relationship.

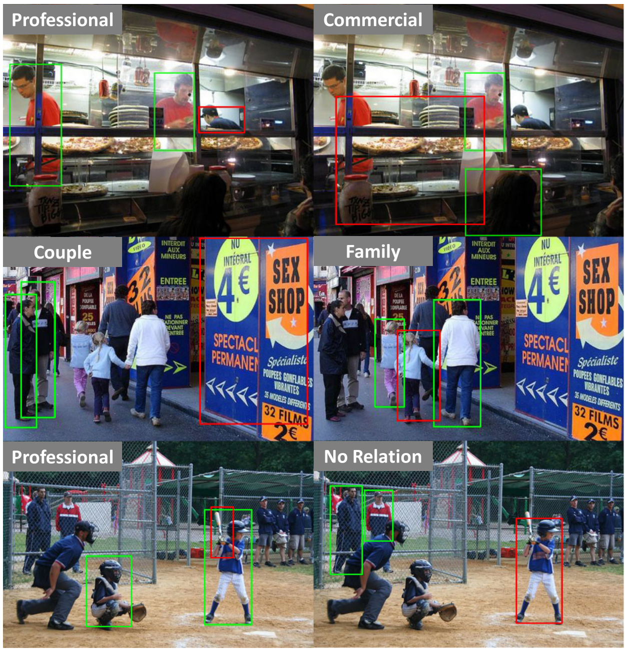


Fig. 10 Illustration of the proposed attentive RCNN. GREEN boxes highlight the target pair of people, and RED box highlights the contextual region with the highest attention. For each target pair, the attention mechanism fixates on different region (Color figure online)

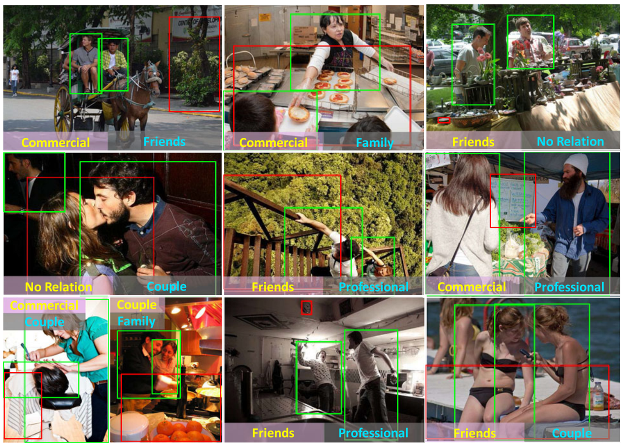


Fig. 11 Examples of incorrect predictions on PISC dataset. Yellow labels are the ground truth, and BLUE labels are the model's predictions (Color figure online)

7 Conclusion

In this study, we address the problem of social relationship recognition, a key challenge to bridge the social gap towards

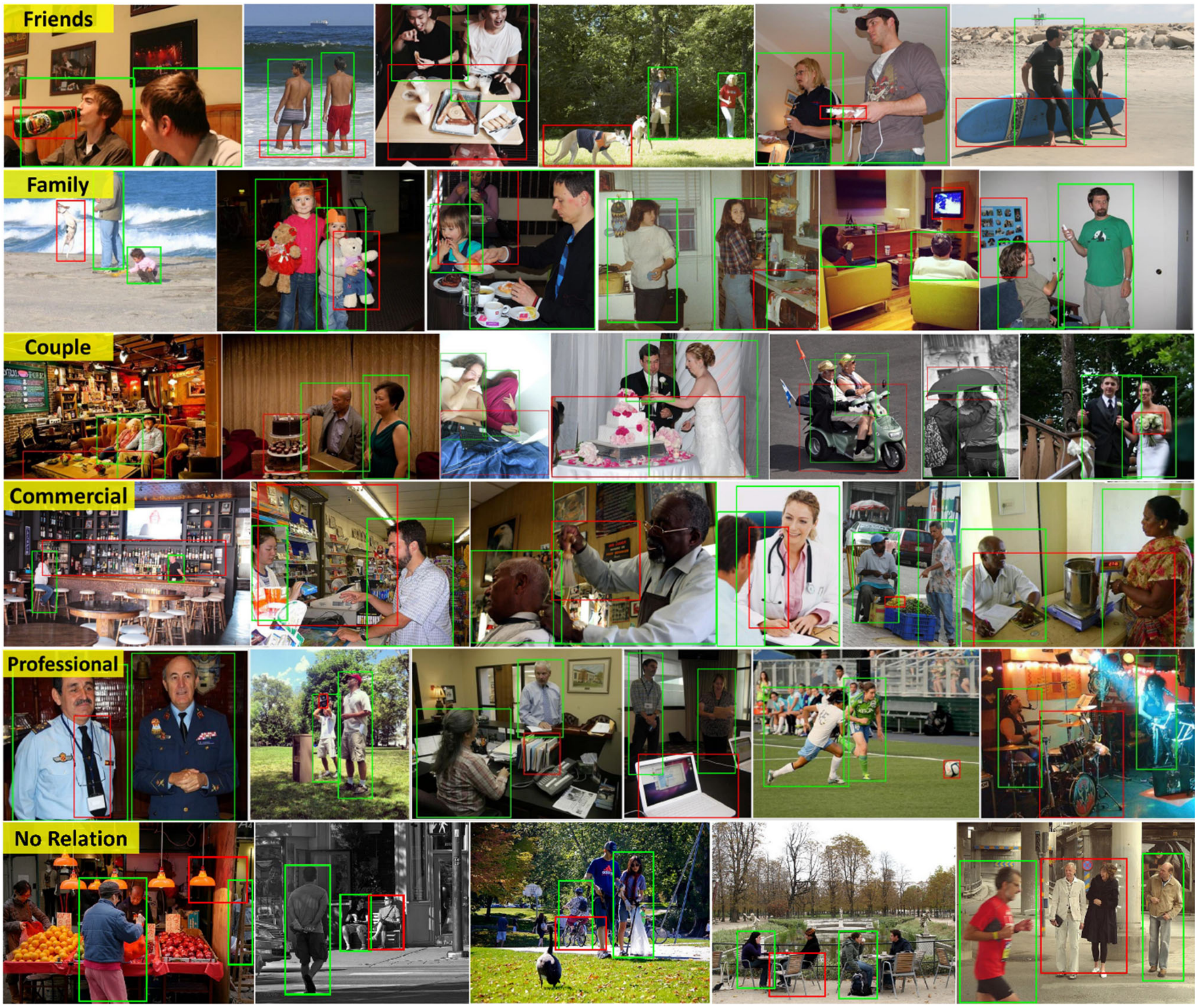


Fig. 12 Example of correct predictions on PISC dataset. GREEN boxes highlight the targets, and RED box highlights the contextual region with highest attention (Color figure online)

higher-level social scene understanding. To this end, we propose a dual-glance model, which exploits useful information from the person pair of interest as well as multiple contextual regions. We incorporate attention mechanism to assess the relevance of each region instance with respect to the person pair. We also propose an adaptive focal loss, that leverages the ambiguity in social relationship labels for more effective learning. The adaptive focal loss can be potentially used in a wider range of tasks that have a certain degree of subjectivity, such as sentiment classification, aesthetic prediction, image style recognition, etc.

In order to facilitate research in social scene understanding, we curated a large-scale PISC dataset. We conduct extensive experiments and ablation studies, and demonstrate both quantitatively and qualitatively the efficacy of the proposed method. Our code and data are available at https://doi.org/10.5281/zenodo.831940.

Our work builds a state-of-the-art computational model for social relationship recognition. We believe that our work can pave the way to more studies on social relationship understanding, and social scene understanding in general.

Acknowledgements This research was carried out at the NUS-ZJU SeSaMe Centre. It is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its International Research Centre in Singapore Funding Initiative.

References

Agrawal, A., Batra, D., Parikh, D.,& Kembhavi, A .(2018). Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR* (pp. 6904–6913).

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In *CVPR* (pp. 961–971).

Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L. M., Ricci, E., Lepri, B., et al. (2016). SALSA: A novel dataset for multimodal group behavior analysis. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(8), 1707–1720.

Alletto, S., Serra, G., Calderara, S., Solera, F., & Cucchiara, R. (2014). From ego to nos-vision: Detecting social relationships in first-person views. In *CVPR workshops* (pp. 594–599).

Chen, Y., Hsu, W. H., Liao, H. M. (2012). Discovering informative social subgraphs and predicting pairwise relationships from group photos. In *ACMMM* (pp. 669–678).

Choi, W., & Savarese, S. (2012). A unified framework for multi-target tracking and collective activity recognition. *ECCV, Lecture Notes in Computer Science*, *7575*, 215–230.

Chu, X., Ouyang, W., Yang, W., & Wang, X .(2015). Multi-task recurrent neural network for immediacy prediction. In *ICCV* (pp. 3352–3360).

Conte, H. R., & Plutchik, R. (1981). A circumplex model for interpersonal personality traits. *Journal of Personality and Social Psychology*, *40*(4), 701.

Deng, Z., Vahdat, A., Hu, H., & Mori, G. (2016). Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In *CVPR* (pp. 4772–4781).

Dibeklioglu, H., Salah, A. A., & Gevers, T. (2013). Like father, like son: Facial expression dynamics for kinship verification. In *ICCV* (pp. 1497–1504).

Ding, L., & Yilmaz, A. (2014). Learning social relations from videos: Features, models, and analytics. In *Human-Centered Social Media Analytics* (pp. 21–41).

Direkoglu, C., & O'Connor, N. E. (2012). Team activity recognition in sports. *ECCV, Lecture Notes in Computer Science*, *7578*, 69–83.

Fan, L., Chen, Y., Wei, P., Wang, W., & Zhu, S. C. (2018). Inferring shared attention in social scene videos. In *CVPR* (pp. 6460–6468).

Fang, R., Tang, K. D., Snavely, N., & Chen, T. (2010). Towards computational models of kinship verification. In *ICIP* (pp. 1577–1580).

Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological Review*, *99*(4), 689.

Gallagher, A. C., & Chen, T. (2009). Understanding images of groups of people. In *CVPR* (pp. 256–263).

Gan, T., Wong, Y., Zhang, D.,& Kankanhalli, M. S. (2013). Temporal encoded F-formation system for social interaction detection. In *ACMMM* (pp. 937–946).

Gao, B., Xing, C., Xie, C., Wu, J., & Geng, X. (2017). Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, *26*(6), 2825–2838.

Gkioxari, G., Girshick, R. B., Malik, J. (2015). Contextual action recognition with R*CNN. In *ICCV* (pp. 1080–1088).

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR* (pp. 6325–6334).

Guo, Y., Dibeklioglu, H., van der Maaten, L. (2014). Graph-based kinship recognition. In *ICPR* (pp. 4287–4292).

Hall, E. T. (1959). *The silent language* (Vol. 3). New York: Doubleday.

Haslam, N. (1994). Categories of social relationship. *Cognition*, *53*(1), 59–90.

Haslam, N., & Fiske, A. P. (1992). Implicit relationship prototypes: Investigating five theories of the cognitive organization of social relationships. *Journal of Experimental Social Psychology*, *28*(5), 441–474.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Hung, H., Jayagopi, D. B., Yeo, C., Friedland, G., Ba, S. O., Odobez, J., et al. (2007). Using audio and video features to classify the most dominant person in a group meeting. In *ACMMM* (pp. 835–838).

Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). Densecap: Fully convolutional localization networks for dense captioning. In *CVPR* (pp. 4565–4574).

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer vision*, *123*(1), 32–73.

Lan, T., Sigal, L., & Mori, G. (2012a). Social roles in hierarchical models for human activity recognition. In *CVPR* (pp. 1354–1361).

Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., & Mori, G. (2012b). Discriminative latent models for recognizing contextual group activities. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(8), 1549–1562.

Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. S. (2017a). Dual-glance model for deciphering social relationships. In *ICCV* (pp. 2650–2659).

Li, Y., Ouyang, W., Zhou, B., Wang, K., & Wang, X. (2017b). Scene graph generation from objects, phrases and region captions. In *ICCV* (pp. 1261–1270).

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. *ECCV, Lecture Notes in Computer Science*, *8693*, 740–755.

Lin, T., Goyal, P., Girshick, R. B., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *ICCV* (pp. 2980–2988).

Lu, C., Krishna, R., Bernstein, M. S., & Fei-Fei, L. (2016). Visual relationship detection with language priors. *ECCV, Lecture Notes in Computer Science*, *9905*, 852–869.

Lv, J., Liu, W., Zhou, L., Wu, B., & Ma, H. (2018). Multi-stream fusion model for social relation recognition from videos. In *MMM* (pp. 355–368).

Marín-Jiménez, M. J., Zisserman, A., Eichner, M., & Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, *106*(3), 282–296.

Maron, O., & Lozano-Pérez, T. (1997). A framework for multiple-instance learning. In *NIPS* (pp. 570–576).

Orekondy, T., Schiele, B., & Fritz, M. (2017). Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *ICCV* (pp. 3686–3695).

Qin, Z., & Shelton, C. R. (2016). Social grouping for multi-target tracking and head pose estimation in video. *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(10), 2082–2095.

Ramanathan, V., Yao, B., Fei-Fei, L. (2013). Social role discovery in human events. In *CVPR* (pp. 2475–2482).

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS* (pp. 91–99).

Rienks, R., Zhang, D., Gatica-Perez, D., & Post, W. (2006). Detection and application of influence rankings in small group meetings. In *ICMI* (pp. 257–264).

Robicquet, A., Sadeghian, A., Alahi, A., & Savarese, S. (2016). Learning social etiquette: Human trajectory understanding in crowded scenes. *ECCV, Lecture Notes in Computer Science*, *9912*, 549–565.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Salamin, H., Favre, S., & Vinciarelli, A. (2009). Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, *11*(7), 1373–1380.

Shao, M., Li, L., & Fu, Y. (2013). What do you do? Occupation recognition in a photo via social context. In *ICCV* (pp. 3631–3638).

Shao, M., Xia, S., & Fu, Y. (2014). Identity and kinship relations in group pictures. In *Human-centered social media analytics* (pp. 175–190).

Sun, Q., Schiele, B., & Fritz, M. (2017). A domain based approach to social relation recognition. In *CVPR* (pp. 3481–3490).

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., et al. (2016). YFCC100M: The new data in multimedia research. *Communications of the ACM*, *59*(2), 64–73.

Vicol, P., Tapaswi, M., Castrejon, L., & Fidler, S. (2018). Moviegraphs: Towards understanding human-centric situations from videos. In *CVPR* (pp. 8581–8590).

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., et al. (2012). Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *The IEEE Transactions on Affective Computing*, *3*(1), 69–87.

Wang, G., Gallagher, A. C., Luo, J., & Forsyth, D. A. (2010). Seeing people in social context: Recognizing people and social relationships. *ECCV, Lecture Notes in Computer Science*, *6315*, 169–182.

Xia, S., Shao, M., Luo, J., & Fu, Y. (2012). Understanding kin relationships in a photo. *IEEE Transactions on Multimedia*, *14*(4), 1046–1056.

Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., & Zhang, Z. (2015). The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR* (pp. 842–850).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., & Salakhutdinov, R., et al. (2015) Show, attend and tell: Neural image caption generation with visual attention. In *ICML* (pp. 2048–2057).

Yang, Y., Baker, S., Kannan, A., & Ramanan, D. (2012). Recognizing proxemics in personal photos. In *CVPR* (pp. 3522–3529).

Yang, Z., He, X., Gao, J., Deng, L., Smola, A. (2016). Stacked attention networks for image question answering. In *CVPR* (pp. 21–29).

You, Q., Jin, H., Wang, Z., Fang, C., Luo, J. (2016). Image captioning with semantic attention. In *CVPR* (pp. 4651–4659).

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., & Samaras, D. (2012). Two-person interaction detection using body-pose features and multiple instance learning. In *CVPR workshops* (pp. 28–35).

Zhang, N., Paluri, M., Taigman, Y., Fergus, R., & Bourdev, L. D. (2015a). Beyond frontal faces: Improving person recognition using multiple cues. In *CVPR* (pp. 4804–4813).

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015b). Learning social relation traits from face images. In *ICCV* (pp. 3631–3639).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.