# Linking People across Text and Images Based on Social Relation Reasoning

**Yang Lei[1], Peizhi Zhao[1], Pijian Li[1], Yi Cai[3,4,5], Qingbao Huang[1,2] ***

[1]School of Electrical Engineering, Guangxi University, Nanning, China
[2]Guangxi Key Laboratory of Multimedia Communications and Network Technology
[3]School of Software Engineering, South China University of Technology, Guangzhou, China
[4]Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
[5]Peng Cheng Laboratory, Shenzhen, China
{2012391019, 2112391073, 1912302005}@st.gxu.edu.cn, ycai@scut.edu.cn, qbhuang@gxu.edu.cn

## Abstract

As a sub-task of visual grounding, linking people across text and images aims to localize target people in images with corresponding sentences. Existing approaches tend to capture superficial features of people (e.g., dress and location) that suffer from the incompleteness information across text and images. We observe that humans are adept at exploring social relations to assist identifying people. Therefore, we propose a Social Relation Reasoning (SRR) model to address the aforementioned issues. Firstly, we design a Social Relation Extraction (SRE) module to extract social relations between people in the input sentence. Specially, the SRE module based on zero-shot learning is able to extract social relations even though they are not defined in the existing datasets. A Reasoning based Cross-modal Matching (RCM) module is further used to generate matching matrices by reasoning on the social relations and visual features. Experimental results show that the accuracy of our proposed SRR model outperforms the state-of-the-art models on the challenging datasets Who's Waldo and FL: MSRE, by more than 5% and 7%, respectively. Our source code is available at https://github.com/VILAN-Lab/SRR.

## Introduction

As a core task of multi-modal researches, visual grounding (a.k.a. referring expression comprehension) aims to localize an object in an image with a corresponding sentence, which becomes a hot and widespread concern topic for its great potential in bridging the gap between visual perception and textual expression. Since the original work (Yu et al. 2016), a surge of subsequent models (Mao et al. 2016a; Liu et al. 2019; Li, Bu, and Cai 2021; Ye et al. 2022) have emerged, and promote significant development of this task.

Previous datasets (Yu et al. 2016; Mao et al. 2016b) permit models to match the features of a salient object in an image-sentence pair, however, they ignore considering matching multiple objects in actual needs. To localize multiple people in an image-sentence pair, Cui et al. (2021) propose a person-centric sub-task of linking people across text and images, which encourages models to learn the capability of reasoning on textual cues and visual features. It is a meaningful task that can be widely used in reading comprehension and people identification in the newspaper.

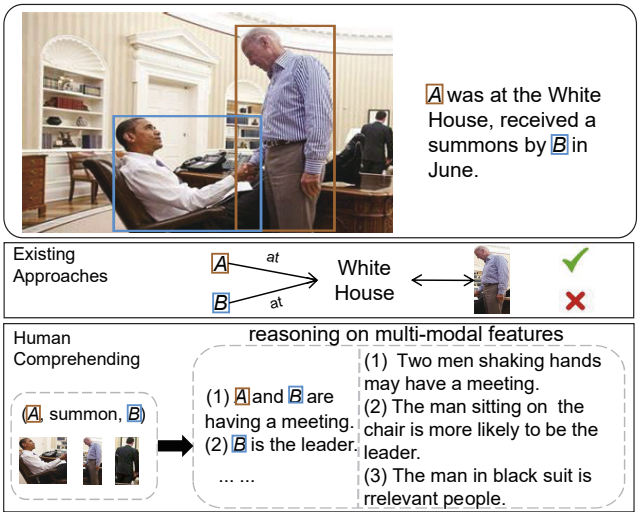


Figure 1: A case of linking people across text and images. Existing approaches capture superficial features such as "at the White House". However, it is difficult for them to match corresponding people with the information above and lead inaccurately linking. Humans are adept at comprehending image-sentence pairs by extracting social relations $(B, summon, A)$, and then matching corresponding people by reasoning on multi-modal features. In addition, $summon$ is a social relation but not defined in the existing datasets, which requires zero-shot learning methods.

However, existing models tend to capture superficial features (e.g., dress and location) of people and establish a one-to-one mapping. As shown in Figure 1, these models are capable of capturing superficial textual and visual features, such as "at the White House", "shaking hands", "sitting on a chair", and "standing behind". Obviously, it is still difficult for them to link people merely depending on the information above. Moreover, when these models cannot comprehend the image-sentence pair, they are more likely to link both $A$ and $B$ to the man in blue and white stripes shirt for he is close to the center of the image. Although Cui et al.

*Corresponding author: Qingbao Huang

(2021) modify the UNITER (Chen et al. 2020) for this task and preliminarily establish multiple people matching across text and images, it still suffers from the deficiency caused by incompleteness information across text and images. For example, "White House" is a significant location feature in the sentence but not discerned in the image. Thus, we consider that exploring rich interactions between people may assist models in comprehending image-sentence pairs.

We observe that humans are adept at identifying people by reasoning on social relations and visual features even if they are unfamiliar with the specific people. For example, from the sentence (cf. Figure 1), we can extract a social relation triplet $(B, summon, A)$. It denotes that $A$ and $B$ are having a meeting, and $B$ is the leader. Then we look at the image to match corresponding people with the social relation. Specificly, the man on a chair is deemed more likely to be the leader $B$, the man shaking hands with him is regarded as $A$, and the man standing behind is not described in the sentence. Furthermore, we notice that social relations are diversity in the real world and not all of them (e.g., $summon$ and $interview$) have been defined in the existing social relation extraction dataset (Gui et al. 2014). We thereby consider that extracting good social relation representations based on zero-shot learning can effectively address the key issue of incompleteness information across text and images.

In this paper, we propose a Social Relation Reasoning (SRR) model to address the aforementioned issues. Firstly, we design a zero-shot learning based Social Relation Extraction (SRE) module to enhance the interaction across text and images. The SRE module encodes input sentences into social relation representations, and then optimizes them from global to local by a two-stage process. The first stage is used to increase the discrepancy between different social relations to achieve accurate classification, while the second stage is used to reduce the semantic gap in the same class to generate better representations. Following the way of people comprehending an image-sentence pair, we further propose a Reasoning based Cross-modal Matching (RCM) module. The RCM module fuses the multi-modal features to obtain reasoning vectors and generates matching matrix by parsing the vectors. Specially, scores of matching matrix could be updated to distinguish indiscernible people. We conduct experiments on two challenging datasets, Who's Waldo (Cui et al. 2021) and FL: MSRE (Wan et al. 2021). Experimental results show that SRR model outperforms existing models.

Our contributions can be summarized as follows:

- We propose a Social Relation Reasoning (SRR) model to link people across text and images. Our SRR model simulates the comprehending process of linking people by reasoning on social relations and visual features.
- We propose a zero-shot learning based Social Relation Extraction (SRE) module that extracts social relation representations to address the issue caused by the incompleteness information across text and images.
- The accuracy of our SRR model outperforms existing state-of-the-art models on the Who's Waldo and FL: MSRE datasets by more than $5\%$ and $7\%$, respectively.

# Related Works

## Visual Grounding

Existing models can be mainly divided into two groups: joint embedding models and modular-based models.

Joint embedding models utilize a CNN-LSTM framework, which obtains textual and visual features respectively, and then encodes them into a same feature space. Specially, Ref-NMS (Chen et al. 2021) proposes AFS and GAR modules to address the problem of limited feeling field. ISREG (Sun, Xiao, and Lim 2021) proposes an iterative contraction mechanism to explore visual context information guided by the sentence. CMS-Net (Li et al. 2022) proposes a cross-modal Transformer to generate interactive visual features and text representation. However, these models encode images and sentences independently, that suffer the limitation of huge semantic gap between monolithic textual representations and visual features.

Modular-based models follow a one-step matching framework which decomposes the sentences into different components and then matches each component with the corresponding visual region through a modular network. Specially, USTS (Rong, Yi, and Tian 2020) extracts attributes from sentences as the guidance to comprehend semantic features of images. BBA (Li, Bu, and Cai 2021) proposes a multi-step bidirectional potential referred pairs to align different granularity level by pyramid visual and textual features. LGREC (Cheng et al. 2021) extends a logical matching module based on CM-Att-Erase (Liu et al. 2019), which performs logical matching over them with explicit logical sentences. CMRE (Yang, Li, and Yu 2021) proposes a cross-modal relation extractor to generate a semantic graph guided by sentences and images. However, all of the above approaches are based on sentence decomposition and match directly with the detected visual regions, that they are arduous to adapt to the multiple objects in complex image-sentence pairs. In actual needs, to identify multiple people in an image-sentence pair, Who's Waldo (Cui et al. 2021) proposes a people-centric sub-task of linking people across text and images, which encourages models to learn the capability of reasoning on textual cues and visual features.

## Zero-Shot Relation Extraction

Relevant studies on zero-shot relation extraction are quite limited. Levy et al. (2017) treat zero-shot relation extraction as a question answering task. Chen et al. (2021) treat the task as a cluster task, in which they extract relation representations between two entities and cluster by calculating the distances between every representations. However, they concentrate on the difference between relations, ignoring the semantic gap of the same relations, which leads to poorly performance of them on downstream tasks.

# Model

## Overview

The task of linking people across text and images can be formulated as follows: given an image $I$ and a sentence $S$ as inputs, the task aims to localize the same people pictured

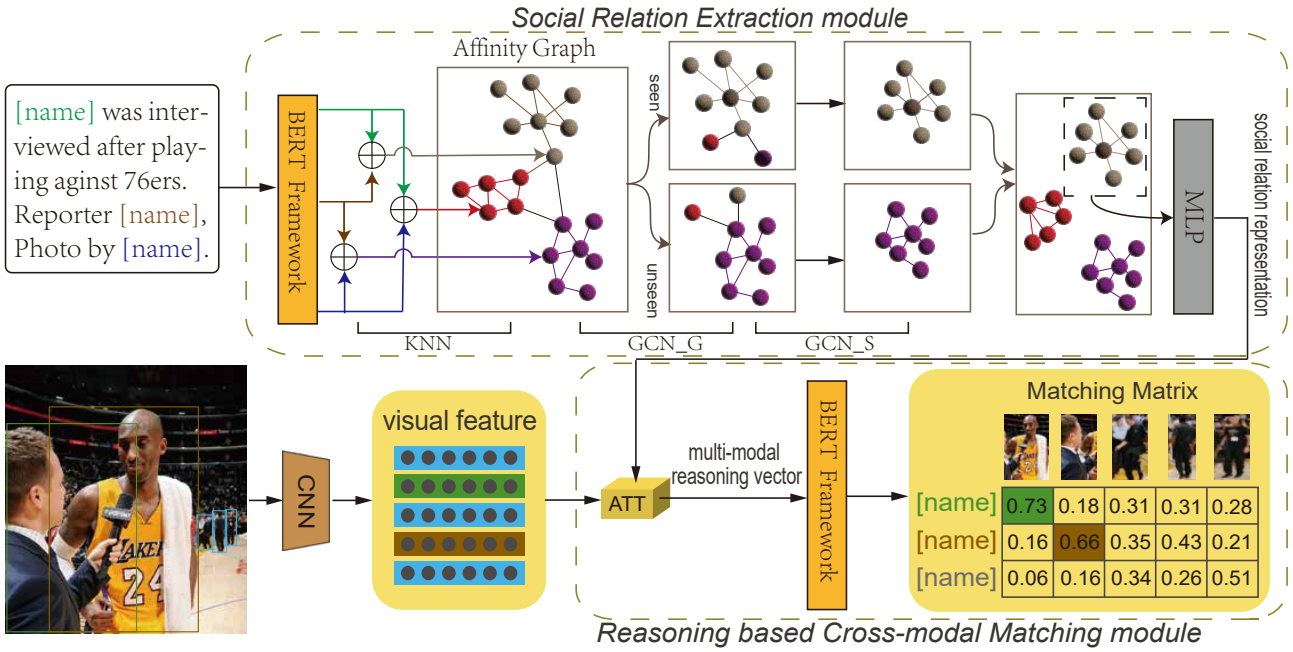


Figure 2: The overall framework of our proposed Social Relation Reasoning (SRR) model.

in the image and described in the sentence. Specifically, we denote the input sentence as $S = \{s_i\}_{i=0}^{l^S-1}$, where $s_i$ represents the $i^{\text{th}}$ word in the sentence and $l^S$ is the length of $S$. The input image $I$ contains several people $P = \{p_i\}_{i=0}^{l^P-1}$, where $p_i$ is the $i^{\text{th}}$ person detected in the image.

As shown in Figure 2, our proposed Social Relation Reasoning (SRR) model mainly contains two modules: a Social Relation Extraction (SRE) module and a Reasoning based Cross-modal Matching (RCM) module.

## Detect People in Images

We first extract the features of people in image following the bottom-up & top-down attention model (Anderson et al. 2018). For each person region proposal $p_i$ in the input image $I$, we obtain its visual feature by a pre-trained Faster RCNN model (Ren et al. 2017), which could be denoted as:

$$v_i = Linear(Conv(p_i)), \tag{1}$$

where $v_i \in R^{d^I}$.

## Social Relation Extraction Module

To explore interactions between people in sentence, we extract social relation representations based on zero-shot learning and optimize them from global to local (cf. Figure 2).

We start with the notations. Social relations are specific relations and defined as the associations between two people in either physical or virtual world. Social relations could be denoted as $G = \{E, R, F\}$, where $E$, $R$ and $F$ indicate the sets of people entities, social relations and facts, respectively. A social relation triplet $(h, r, t) \in F$ indicates that there is a social relation $r \in R$ between head-people entity $h \in E$ and tail-people entity $t \in E$.

**Extract Initial Social Relation Representations** Given an input sentence $S$, we first obtain its word embedding vectors $w_i \in R^{d^S}$ by the GloVe word2vec model (Pennington, Socher, and Manning 2014). To encourage our model pay more attention to the context instead of learning straightforward association between faces and names, we use unified symbols to indicate people entities. Specially, we record locations of people as column vectors $C = \{c_i\}$, where $c_i$ denotes the location of $i^{\text{th}}$ pair of people in the sentence.

We extract social relations from by a transformer encoder (Vaswani et al. 2017). By sequentially feeding the word embedding vectors $W = \{w_i\}_{i=0}^{l^S-1}$ into the encoder, we map the discrete word embedding vectors into initial social relation representation $r_i$ of $i^{\text{th}}$ pair of people:

$$r_i = tanh(\theta_1 W + b_1)c_i^\top, \tag{2}$$

where $\theta_1 \in R^{d^S \times d^S}$, $b_1 \in R^{d^S}$ are trainable parameters, and $r_i \in R^{d^S}$.

For social relation representations $R = \{r_i\}$, we assign highest entailment probability to the verbalizations of the same classes, and low entailment probabilities to the rest of the hypothesis. Specifically, we treat the word embedding vectors of label words as the target $T = \{t_i\}$ and the entailment probability $o$ is defined as:

$$o_{r_i,t_j} = \frac{r_i t_j^\top}{\| r_i \| \| t_j^\top \|}, \tag{3}$$

where $i = j$ denotes $r_i, t_j$ belong to the same classes and $i \neq j$ denotes they belong to the different. The loss $L_{SRE}$ of the transformer encoder can be defined as:

$$L_{SRE} = (1 - o_{r_i,t_i}) + \lambda \frac{\sum_{i=1}^{\eta} o_{r_i,t_j}}{\eta}, \tag{4}$$

where $\lambda$ is set to $0.3$.

**Optimize from Global to Local** Intuitively, social relation representations with dense and pure connections are more likely to be the same class, while representations with sparse connections belong to different classes. Therefore, to obtain better social relation representations, we conduct optimization from global to local by a two-stage process.

Conducting global optimization, we accurately distinguish social relations by increasing the discrepancy between different classes. Firstly, we set an undirected graph, in which its nodes $N = \{n_i\}$ consist of initial social relation representations $R = \{r_i\}$ and embedding vectors of label words $T = \{t_i\}$. Then we calculate the cosine-similarity between $i^{\text{th}}$ and $j^{\text{th}}$ node as the edges $k_{i,j}$:

$$k_{i,j} = exp(-\frac{\parallel n_i - n_j \parallel_2}{2}), \tag{5}$$

where $\parallel \cdot \parallel_2$ denotes L2 norm.

We utilize a KNN (Abeywickrama, Cheema, and Taniar 2016) to suggest an affinity graph to obtain base classification proposals of social relations. GCNs (Kipf and Welling 2017) are proved to work effectively on no annotation data. Thus, to increase the discrepancy of different social relations, a GCN is employed to perform classification detection to optimize the whole graph (GCN_G). For a node with high confidence, its neighboring nodes tend to belong to the same class while a node away from it usually belongs to another class. Thus, we define the confidence $s_i$ for node $n_i$:

$$s_i = \frac{1}{|N_i|} \sum k_{i,j} \times \begin{cases} 1 & y_j = y_i \\ 0 & y_j \neq y_i \end{cases}, \tag{6}$$

where $N_i$ are the adjacency nodes of $n_i$, $|\cdot|$ denotes the number of nodes and $y_i$ is the initial class of $n_i$.

The computation of $l^{\text{th}}$ layer $F_l$ can be formulated as:

$$F_l = ReLU(g(\widetilde{A}, F_{l-1})\ M_l), \tag{7}$$

$$\widetilde{A} = \widetilde{D}^{-1}\ (A + I), \tag{8}$$

where $\widetilde{D}_{ii} = \sum_j (A + I)_j$ is a diagonal degree matrix, $M_l \in R^{d^S \times d^S}$ is a trainable matrix and $g$ is the concatenation of input embedding vectors and them after neighborhood aggregation $g(\widetilde{A}, F_l) = [(F_l)^\top, (\widetilde{A}F_l)^\top]^\top$.

To increase the discrepancy between different social relations of whole nodes $N$ in the graph, the loss $L_{GCN\_G}$ of GCN_G is defined as:

$$L_{GCN\_G} = \frac{1}{|N|} \sum_{i=1}^{|N|} \parallel (\widetilde{s_i}, s_i) \parallel_2, \tag{9}$$

where the new confidence $\widetilde{s} = Linear(F_l)$ is predicted by a fully-connected layer.

After global optimizing, we conduct local optimization by reducing the semantic gap inside sub-graph with another GCN (GCN_S), which has the same structure of GCN_G.

Similar to the setting of traditional zero-shot learning task, we divide all sub-graphs into two groups: seen and unseen sub-graph. For the seen sub-graph, which nodes therein contain the target social relation representation $t$, we reduce the semantic gap between the target word and its adjacency nodes by maximizing its confidence. The loss of seen sub-graphs is defined as:

$$L_{seen} = \frac{1}{|N_1|} \sum_{i=1}^{|N_1|} exp(-\frac{\parallel n_i - t \parallel_2}{2}), \tag{10}$$

where $N_1$ denotes the seen sub-graphs.

For the unseen sub-graph, we maximize the confidence of each node for better semantic representation. The loss of unseen sub-graph is defined as:

$$L_{unseen} = \frac{1}{|N_2|^2} \sum_{i=1}^{|N_2|} \sum_{j=1}^{|N_2|} exp(-\frac{\parallel n_i - n_j \parallel_2}{2}), \tag{11}$$

where $N_2$ denotes the unseen sub-graphs. The loss of GCN_S is defined as $L_{GCN\_S} = L_{seen} + L_{unseen}$.

The loss $L$ of SRE module is defined as:

$$L = L_{SRE} + a_1\ L_{GCN\_G} + a_2\ L_{GCN\_S}, \tag{12}$$

where hyperparameters $a_1 = 0.7$ and $a_2 = 0.5$ are employed to control the tradeoff among these losses.

Then, a MLP layer is used to obtain the final social relation representation $\widetilde{r_i} = Linear(N_i)$, where $N_i$ denotes nodes in the corresponding sub-graph of $r_i$ and $\widetilde{r_i} \in R^{d^S}$.

## Reasoning Based Cross-Modal Matching Module

To consider a single person with a single social relation at a time, we divide people detected in the image into three classes: head-people entity of the social relation, tail-people entity of the social relation and irrelevant people.

Following humans comprehending an image-sentence pair, we further propose a Reasoning based Cross-modal Matching (RCM) module to adopt reasoning on multi-modal features. RCM module mainly includes two processes: generating multi-modal reasoning vectors by attention mechanism and parsing these vectors to generate matching matrix.

Considering that there is still a huge semantic gap between visual features and social relation representations, we obtain multi-modal reasoning features $X = \{x_{i,j}\}$ by an attention mechanism:

$$x_{i,j} = v_i U + \widetilde{r_j}^\top v_i U \widetilde{r_j}, \tag{13}$$

where $U \in R^{d^I \times d^S}$ is a trainable matrix and $x_{i,j} \in R^{d^S}$.

Then, a transformer decoder is used to parse multi-modal reasoning feature $x_{i,j}$ to the probability of people in the image belonging to different classes:

$$m_{i,j} = Linear(tanh(\theta_2 x_{i,j} + b_2)), \tag{14}$$

where $\theta_2 \in R^{d^S \times d^S}$, $b_2 \in R^{d^S}$ are trainable parameters, and $m_{i,j} \in R^3$. Then we generate a matching matrix $M$ by filling the probability $m_{i,j}$ into corresponding positions.

We obverse that multiple people in one modality tend to link to a same person in another, thus, we propose a updating mechanism. First, we adopt a row normalization to distinguish indiscernible people in the sentence:

$$\hat{M}_r = M_r((q^\top M)^\top)^{-1}, \tag{15}$$

| | Models | visual features | Who's Waldo | | FL:MSRE | |
|---|---|---|---|---|---|---|
| | | | val | test | val | test |
| **Modular-Based Models** | | | | | | |
| 1 | CM-Att-Erase (CVPR 2019) | Darknet53 | 59.64 ±1.10 | 59.72 ±1.05 | 59.82 ±0.77 | 61.06 ±0.94 |
| 2 | USTS (TIP 2020) | Resnet101 | 61.09 ±1.08 | 62.14 ±1.15 | 66.40 ±1.01 | 64.32 ±1.01 |
| 3 | UNITER-M (ICCV 2021) | Resnet101 | 63.50 ±1.03 | 61.90 ±1.04 | 66.81 ±1.16 | 64.50 ±1.13 |
| 4 | RealGIN (TNNLS 2021) | Darknet53 | 63.91 ±1.02 | 64.21 ±1.03 | 65.46 ±1.00 | 64.32 ±0.91 |
| 5 | CMRE (TPAMI 2021) | Darknet53 | 65.45 ±0.86 | 63.36 ±1.32 | 65.20 ±1.06 | 66.07 ±1.05 |
| 6 | TransVG (CVPR 2021) | Resnet101 | 67.31 ±1.22 | 66.93 ±1.31 | 64.18 ±1.05 | 65.84 ±1.02 |
| 7 | LGREC† (MM 2021) | Resnet101 | 66.05 ±1.37 | 65.23 ±1.19 | 63.39 ±1.20 | 63.42 ±1.15 |
| 8 | BBA† (MM 2021) | Darknet53 | 67.09 ±0.82 | 67.56 ±1.44 | 68.37 ±1.06 | 66.91 ±1.13 |
| 9 | MDGT (CVPR 2022) | Resnet101 | 66.47 ±1.01 | 67.39 ±1.11 | 66.46 ±1.04 | 66.23 ±1.33 |
| 10 | QRNET (CVPR 2022) | Resnet101 | 67.20 ±1.28 | 68.79 ±0.97 | 69.24 ±1.05 | 66.30 ±1.35 |
| **Joint-Embedding Models** | | | | | | |
| 11 | MMI (CVPR 2016) | VGG16 | 54.80 ±1.05 | 57.10 ±1.02 | 57.71 ±1.27 | 56.36 ±1.02 |
| 12 | ISREG (CVPR 2021) | Resnet101 | 61.42 ±1.07 | 62.01 ±0.90 | 62.87 ±1.11 | 65.82 ±0.94 |
| 13 | Ref-NMS (AAAI 2021) | Resnet101 | 64.77 ±1.18 | 65.44 ±0.96 | 68.19 ±1.01 | 64.13 ±1.04 |
| 14 | CMS-Net† (Neurocomputing 2022) | Darknet53 | 65.11 ±1.12 | 66.52 ±1.05 | 68.25 ±1.13 | 66.62 ±1.00 |
| **Ours** | | | | | | |
| | **SRR** | Resnet101 | **74.46** ±1.19 | **74.49** ±1.10 | **75.74** ±0.96 | **76.26** ±1.11 |

Table 1: Comparison with the state-of-the-art models on Who's Waldo and FL:MSRE. † denotes we reproduct the codes.

where $M_r$ is the $t^{\text{th}}$ row in the matching matrix $M$ and $q$ is a row vector of 1. Then, a column normalization is proposed to distinguish indiscernible people in the image:

$$\hat{M_c} = ((M^{\omega} q^{\top})^{\top})^{-1} {M_c}^{\omega}, \tag{16}$$

where $M_c$ is the $c^{\text{th}}$ column in the matching matrix, and $\omega$ is set to 0.7 to reduce the impact of matching scores in lower confidence. We obtain the final matching matrix $\hat{M}$ by conducting the row and column normalization twice.

### Training and Inference

During training, we adopt a step-by-step training strategy to ensure the training effect is achieved without taking up too much storage space. Specially, parameters of the Faster R-CNN are frozen, which are only pretrained on COCO (Lin et al. 2014) and finetuned on COCO-WholeBody (Jin et al. 2020). The SRE module is pretrained on the dataset provided by Gui ea al. (2014), then the SRE module is frozen during training the whole model.

During inference, matching score is recovered to decode single-person matching vectors. A person may linked to multiple people or not appeared in another modality. Thus, we consider a final matching when the score is at least 0.6.

## Experiment

### Dataset

**Who's Waldo** (Cui et al. 2021) consists of 19.2K image-sentence pairs in Wikimedia Commons by people sifting, which are split into 17.9K training, 6.7K validation, and 6.7K test image–sentence pairs. It is originally designed for linking people across text and images. Currently, this is the largest dataset on this task.

**FL: MSRE** (Wan et al. 2021) consists of $3,716$ images and $6,485$ sentences. To explore the influence of visual features on the task of social relation extraction, Wan et al. (Wan et al. 2021) collect images from the TV adaptations of the four Chinese masterpieces based on Du et al. (Du et al. 2019). FL: MSRE regards image-sentence pairs and locations of corresponding people as input, while social relations between them as output. We transform FL: MSRE for this task by removing the locations of people in images from the input and regard it as the ground-truth.

### Experimental Setup

**Evaluation Metric** Following the setting of the prior work (Cui et al. 2021), we adopt $Acc@0.5$ as evaluation metric, which considers a right prediction if the IoU is at least 0.5.

**Implementation Details** Visual features of dimension 2048 are extracted from Faster R-CNN (Ren et al. 2017), which is pretrained on COCO (Lin et al. 2014) and finetuned on COCO-WholeBody (Jin et al. 2020). Words are encoded into dimension 300 by a pretrained GLoVe word2vec model (Pennington, Socher, and Manning 2014) and SRE module are pretrained on (Gui et al. 2014). Specifically, during pretraining the SRE module, learning-rate is setting to $1e-3$, dropout rate is setting to 0.5 for a total of 80 epochs. The visual features are then mapped into dimension 512 by a fully-connected layer. Inside the SRE module, the number of multi-head attention is set to 6 with dimension 300 for each head in the 3 layers transformer encoder. Layers of GCNs is set to 2. Inside the RCM module, visual features and social relation representations are merged into dimension 300 multi-modal reasoning vectors by an attention mechanism and are then parsed by transformer decoder with the same structure below. Inside SRR model, $d^S$ is set to 300 and $d^I$ is set to 512, respectively.

| Approaches | | | $Acc$@0.5 |
|---|---|---|---|
| | random | | 36.94 |
| **Zero-Shot Learning Models** | | | |
| 1 | ReCLIP | M$\rightarrow F^{\star}$ | 60.72 |
| 2 | Pseudo-Q | M$\rightarrow F^{\star}$ | 61.05 |
| **Joint Embedding and Modular-Based Models** | | | |
| 3 | MMI | W$\rightarrow F^{\star}$ | 46.90 |
| 4 | ISREG | W$\rightarrow F^{\star}$ | 48.56 |
| 5 | Ref-NMS | W$\rightarrow F^{\star}$ | 46.23 |
| 6 | CM-Att-Erase | W$\rightarrow F^{\star}$ | 48.74 |
| 7 | USTS | W$\rightarrow F^{\star}$ | 47.51 |
| 8 | UNITER-M | W$\rightarrow F^{\star}$ | 47.09 |
| 9 | RealGin | W$\rightarrow F^{\star}$ | 50.03 |
| 10 | CMRE | W$\rightarrow F^{\star}$ | 49.89 |
| 11 | LGREC | W$\rightarrow F^{\star}$ | 47.32 |
| 12 | BBA | W$\rightarrow F^{\star}$ | 46.41 |
| 13 | TransVG | W$\rightarrow F^{\star}$ | 49.04 |
| 14 | MDGT | W$\rightarrow F^{\star}$ | 49.57 |
| 15 | QRNET | W$\rightarrow F^{\star}$ | 49.75 |
| 16 | CMS-Net | W$\rightarrow F^{\star}$ | 47.25 |
| | **SRR(ours)** | W$\rightarrow F^{\star}$ | **60.13** |

Table 2: Cross-Dataset Experiments. $M \rightarrow F^{\star}$ denotes the models are pretrained beforehand and test on FL: MSRE, $W \rightarrow F^{\star}$ denotes we train the models on Who's Waldo and test on FL: MSRE.

Fully connected layers have the same dropout rate $0.3$. Gradients are clipped to $0.25$. Batch size is set as $8$. Adam optimizer (Kingma and Ba 2015) is used with initial learning rate of $2e-4$. The learning rate is halved every 10 epochs for a total of 50 epochs.

## Results and Analysis

As shown in Table 1, we conduct experiments on Who's Waldo and FL: MSRE datasets. The top part of the table lists results of recently modular-based models, the middle part of the table lists results of recently joint embedding models and the bottom part of the table lists result of our model. The "visual features" column is the backbone of respective models. We evaluate the recently models from respective papers when codes are available and reproduct some models when they are not available. We run all these models ten times and report their average and variance results.

It is worth noting that the accuracy of proposed SRR model outperforms existing state-of-the-art models across all the datasets, by more than $5\%$ and $7\%$, respectively.

We obtain several observations. First, we compare our SRR model with LGREC, a modular-based model which is optimized for complex sentences. To be remarkable, our SRR model outperforms it more than $8\%$ on Who's Waldo dataset, of which average length of the sentences achieve $21$. A possible explanation is that exploring the logic relationship in sentences and matching them with visual features, ignoring the incompleteness information across text and images. The performance of our SRR model validates the effectiveness of the SRE module of learning interactions in complex sentences.

| backbone | pretrained and GCN | $Acc$@0.5 |
|---|---|---|
| **LSTM** | non-pretrained | 64.2 |
| | non GCN | 70.7 |
| | GCN_G | 69.5 |
| | GCN_G + GCN_S | 71.2 |
| **BiLSTM** | non-pretrained | 65.1 |
| | non GCN | 69.1 |
| | GCN_G | 69.5 |
| | GCN_G + GCN_S | 71.1 |
| **Transformer** | non-pretrained | 63.4 |
| | non GCN | 70.1 |
| | GCN_G | 72.3 |
| | **SRR(ours)** | **74.5** |

Table 3: Ablation study on structures of the SRE module

Moreover, we compare our SRR model with CMS-Net, a joint-embedding model designed for multi-scale matching, also, inferiors to ours. We consider that ignoring huge semantic gap between textual and visual features makes it difficult to comprehend a image-sentence pair with multiple people. Our proposed SRR model achieves significant improvement on the two datasets, which demonstrates effectiveness of our model.

## Cross-Dataset Experiment

As shown in Table 2, we conduct a cross-dataset experiment to prove the generalization of our SRR model. We train the joint-embedding models and modular-based models on Who's Waldo training set and evaluate them on FL: MSRE test set ($M \rightarrow F^{\star}$). As for zero-shot learning models, we evaluate them on FL: MSRE test set directly ($W \rightarrow F^{\star}$).

It is worth noting that the accuracy of our proposed SRR model outperforms all modular-based and joint embedding models, by more than $10\%$, and achieves an approximate level with zero-shot learning models. Inside the two datasets, they are similar in image types, but there is a huge gap in sentence format. Thus, it is difficult for these models to learn a unified sentence representation only based on superficial features. For zero-shot learning models, they have been pretrained on multifarious sentence format, which makes them more adaptable than these models.

We think that obtaining superficial features in the sentence rely on similar textual format, while social relations are hardly changed. The performance of our SRR model can be an effective evidence that learning social relations is a meaningful way to address the incompleteness information across text and images.

## Ablation Study

In this section, we design different structures for ablation studies to understand our SRR model better. We conduct experiments on Who's Waldo validation set, which is more challenging for its complex and long image-sentence pairs.

The SRE module is proposed to extract social relations between people in sentences. Specially, a pertained transformer encoder is used to extract social relation representa-

| Compositions | $Acc@0.5$ |
|---|---|
| multi-modal features match directly | 61.2 |
| stack without attention meshanism | 70.8 |
| matching matrix without updating | 73.2 |
| **SRR(ours)** | **74.5** |

Table 4: Importance of compositions in the RCM module

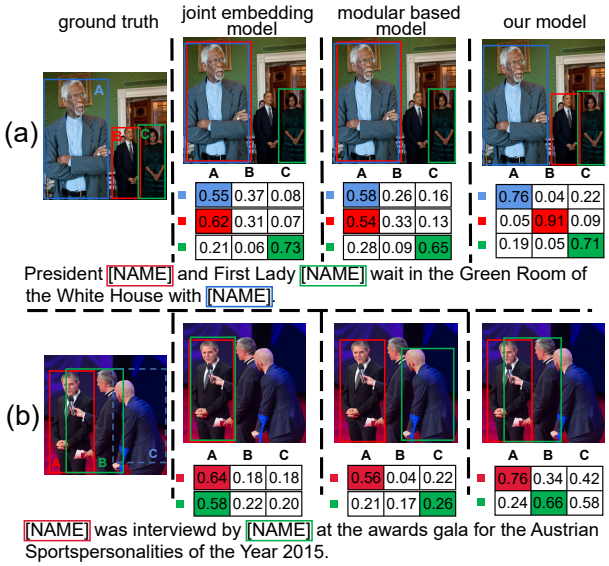


Figure 3: Case studies of our SRR model. Corresponding people are marked in same color boxes.

tions and two GCNs is used to optimize the representations from global to local. Therefore, we first conduct experiments with non-pertrained models to prove learning can effectively get over the incompleteness information across text and images and improve accuracy. Then, we design a series of models with different numbers of GCNs to compare with the full model (the last row in Table 3) to show the importance of these compositions in the SRE module.

Considering the advantage of modular-based models on comprehending multi-modal features, we propose a RCM module to decode multi-modal features. Specially, an attention mechanism is used to combine multi-modal features, a transformer decoder is used to parse reasoning vectors and generate matching matrices. Therefore, we remove different compositions of the RCM module to compare with the full model, including matching multi-modal features directly, stacking multi-modal features without attention mechanism and generating matching matrix without updating. Table 4 shows the importance of each composition.

## Case Study

Figure 3 shows that failure cases matched by previous joint embedding model CMS-Net and modular-based model LGREC can be comprehended by our SRR model. As shown in Case (a), all existing models correctly link the person pictured at lower right corner to superficial features *lady*, but fail to deal with the other two people. Especially for Case

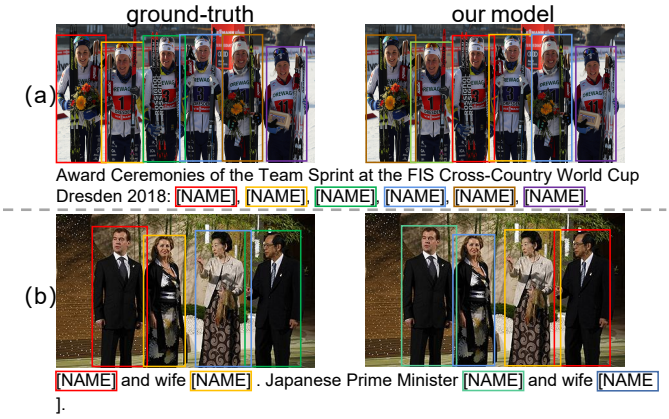


Figure 4: Bad cases of our model on Who's Waldo dataset. Corresponding people are marked in same color boxes.

(b), there is not any direct or indirect description of superficial features in the sentence which leads wrong linking of existing models. A possible explanation is that social relations between multiple people help to learn associations between them in images. The results demonstrate that reasoning on multi-modal features such as social relations and visual features helps to make more accurate and consistent matching.

## Error Analysis

Although our SRR model achieves remarkable performance, there are still some cases that cannot be solved. To explore further directions for future studies, we select some cases as shown in Figure 4 and summarize the two main error types. The first (Case (a)) is the difficulty of identifying people without facilitates reasoning information. The second is caused by the weakness to reason on multiple people-pairs with the same social relations, e.g., two couples in Case (b).

## Conclusion

In this paper, we propose the Social Relation Reasoning (SRR) model that progressively links people across text and images. It can address the key issue of misalignment caused by the incompleteness information across text and images. In this paper, we capture the interactions between multiple people by extracting social relations based on zero-shot learning, which can be supplements of sentences to address the issue of incompleteness information. Following the way of human comprehending an image-sentence pair, we further match corresponding people by reasoning on multi-modal features. Extensive experiments and ablation studies show the powerful performance of our model. Our SRR model achieves significant improvement on Who's Waldo and FL: MSRE datasets, which validates the robustness of our model. We will explore the task of temporal grounding with zero-shot relation extraction and multi-modal reasoning in the future.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (62276072, 62076100, and

62261003), the Guangxi Natural Science Foundation (No. 2022GXNSFAA035627), Guangxi Scientific and Technological Bases and Talents Special Projects (Application No. 2022AC21300, 2022AC21254) the Open Research Fund of Guangxi Key Laboratory of Multimedia Communications and Network Technology, the Fundamental Research Funds for the Central Universities, SCUT (D2210010, D2200150, and D2201300), CAAI-Huawei MindSpore Open Fund and the Science and Technology Planning Project of Guangdong Province (2020B0101100002).

## References

Abeywickrama, T.; Cheema, M. A.; and Taniar, D. 2016. k-Nearest Neighbors on Road Networks: A Journey in Experimentation and In-Memory Implementation. *Proc. VLDB Endow.*, 9(6): 492–503.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, 6077–6086.

Chen, C.; and Li, C. 2021. ZS-BERT: Towards Zero-Shot Relation Extraction with Attribute Representation Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, 3470–3479.

Chen, L.; Ma, W.; Xiao, J.; Zhang, H.; and Chang, S. 2021. Ref-NMS: Breaking Proposal Bottlenecks in Two-Stage Referring Expression Grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 1036–1044.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12375, 104–120.

Cheng, Y.; Wang, R.; Yu, J.; Zhao, R.; Zhang, Y.; and Feng, R. 2021. Exploring Logical Reasoning for Referring Expression Comprehension. In *MM '21: ACM Multimedia Conference*, 5047–5055.

Cui, C. Y.; Khandelwal, A.; Artzi, Y.; Snavely, N.; and Averbuch-Elor, H. 2021. Who's Waldo? Linking People Across Text and Images. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 1354–1364.

Du, J.; Pan, J. Z.; Wang, S.; Qi, K.; Shen, Y.; and Deng, Y. 2019. Validation of Growing Knowledge Graphs by Abductive Text Evidences. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, 2784–2791.

Gui, F.; Zhang, F.; Ma, Y.; Liu, M.; and Shen, W. 2014. Social relation extraction of large-scale logistics network based on mapreduce. In *2014 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2014*, 2273–2277.

Jin, S.; Xu, L.; Xu, J.; Wang, C.; Liu, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Whole-Body Human Pose Estimation in the Wild. In *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12354, 196–214.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017*.

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 333–342.

Li, L.; Bu, Y.; and Cai, Y. 2021. Bottom-Up and Bidirectional Alignment for Referring Expression Comprehension. In *MM '21: ACM Multimedia Conference*, 5167–5175.

Li, Q.; Zhang, Y.; Sun, S.; Wu, J.; Zhao, X.; and Tan, M. 2022. Cross-modality synergy network for referring expression comprehension and segmentation. *Neurocomputing*, 467: 99–114.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference*, volume 8693 of *Lecture Notes in Computer Science*, 740–755.

Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, 1950–1959.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016a. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 11–20.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016b. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, 11–20.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, 1532–1543.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6): 1137–1149.

Rong, X.; Yi, C.; and Tian, Y. 2020. Unambiguous Scene Text Segmentation With Referring Expression Comprehension. *IEEE Trans. Image Process.*, 29: 591–601.

Sun, M.; Xiao, J.; and Lim, E. G. 2021. Iterative Shrinking for Referring Expression Grounding Using Deep Reinforcement Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 14060–14069.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information*

*Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 5998–6008.

Wan, H.; Zhang, M.; Du, J.; Huang, Z.; Yang, Y.; and Pan, J. Z. 2021. FL-MSRE: A Few-Shot Learning based Approach to Multimodal Social Relation Extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, 13916–13923.

Yang, S.; Li, G.; and Yu, Y. 2021. Relationship-Embedded Representation Learning for Grounding Referring Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(8): 2765–2779.

Ye, J.; Tian, J.; Yan, M.; Yang, X.; Wang, X.; Zhang, J.; He, L.; and Lin, X. 2022. Shifting More Attention to Visual Backbone: Query-Modulated Refinement Networks for End-to-End Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15502–15512.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *Computer Vision - ECCV 2016 - 14th European Conference*, 69–85.