Neural Networks 173 (2024) 106216

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

Graph-based social relation inference with multi-level conditional attention

Xiaotian Yu [a,*,1], Hanling Yi [a,1], Qie Tang [a], Kun Huang [a], Wenze Hu [a], Shiliang Zhang [b], Xiaoyu Wang [a,c]

[a] Department of AI Technology Center, Shenzhen Intellifusion Ltd., China
[b] Department of Computer Science, Peking University, China
[c] The Chinese University of Hong Kong (Shenzhen), China

ARTICLE INFO

Keywords:
Social relation inference
Multi-level conditional attention
Transformer

ABSTRACT

Social relation inference intrinsically requires high-level semantic understanding. In order to accurately infer relations of persons in images, one needs not only to understand scenes and objects in images, but also to adaptively attend to important clues. Unlike prior works of classifying social relations using attention on detected objects, we propose a MUlti-level Conditional Attention (MUCA) mechanism for social relation inference, which attends to scenes, objects and human interactions based on each person pair. Then, we develop a transformer-style network to achieve the MUCA mechanism. The novel network named as Graph-based Relation Inference Transformer (*i.e.*, GRIT) consists of two modules, *i.e.*, a Conditional Query Module (CQM) and a Relation Attention Module (RAM). Specifically, we design a graph-based CQM to generate informative relation queries for all person pairs, which fuses local features and global context for each person pair. Moreover, we fully take advantage of transformer-style networks in RAM for multi-level attentions in classifying social relations. To our best knowledge, GRIT is the first for inferring social relations with multi-level conditional attention. GRIT is end-to-end trainable and significantly outperforms existing methods on two benchmark datasets, *e.g.*, with performance improvement of 7.8% on PIPA and 9.6% on PISC.

1. Introduction

Social relations, which are fundamental to the daily life of human beings (Kitayama & Markus, 2000), characterize the connections among two or more individuals. Nowadays, billions of people upload images in social media platforms, such as Facebook, Instagram or Snapchat, to share news and broadcast activities in daily life. There have been dramatically increasing interests in understanding social relations among persons in still images due to the broad computer vision applications including personalized social recommendations (Fan et al., 2021; Zhang, Yang, Zhuo, Tian & Liang, 2019), group behavior analysis (Alameda-Pineda et al., 2016; Hoai & Zisserman, 2014; Yan, Xie, Tang, Shu, & Tian, 2020), image caption generation (Chen & Lawrence Zitnick, 2015; Kim, Oh, Choi, & Kweon, 2021; Stefanini et al., 2022) and human trajectory prediction (Alahi et al., 2016; Marchetti, Becattini, Seidenari, & Del Bimbo, 2020; Zhang, Xue, Zhang, Zheng, & Ouyang, 2020).

Based on Li et al. (2017), Zhang, Paluri, Taigman, Fergus and and Bourdev (2015), the input of a model for social relation inference is images with annotated bounding boxes of all persons, and the model is required to predict the social relation of each person pair. The problem of social relation inference is challenging and complicated because it requires high-level semantic understanding of images. In order to accurately infer social relations of persons, one needs not only to identify the category of scenes and background objects, but also to apply attention on-demand on the scenes, objects and human interactions for each person pair, *e.g.*, hugging and handshaking.

To demonstrate the high-level semantic understanding when inferring social relations, we present two examples in Fig. 1. Each example consists of two images. It shows that the task of inferring relationships requires attention on various levels (object, interaction, scene context, *etc*). For instance, Example 1 shows both images presents two people talking to each other. However, it is the scene context and body interaction that classified them into different relation categories. It is clear that models for inferring social relations need to have on-demand attention with a global view.

We note that there are two studies conducting social relation inference with attention. A dual-glance model was developed in Li et al. (2017). The first glance obtains the feature of person pair and the second glance attends to detected objects in images. In Wang et al. (2018), a knowledge graph with persons and objects was proposed

* Corresponding author.
E-mail address: xiaotianyu.ac@gmail.com (X. Yu).
[1] Denotes equal contributions.

https://doi.org/10.1016/j.neunet.2024.106216
Received 16 March 2023; Received in revised form 15 January 2024; Accepted 26 February 2024
Available online 27 February 2024

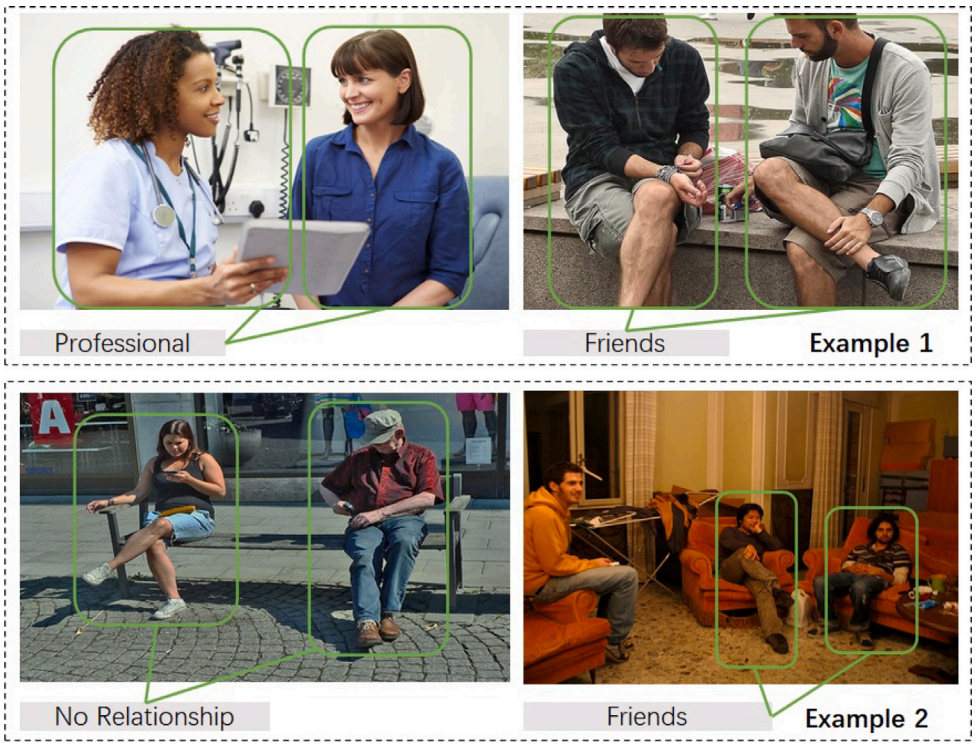


Fig. 1. Example illustration of social relation inference on the PISC dataset (Li, Wong, Zhao, & Kankanhalli, 2017). Example 1 shows two different social relations (friends versus professional) indicated by different background scenes and body interactions. Example 2 shows social relations could be different even if similar interactions are presented. The difference can be implied from image backgrounds (indoor versus outdoor public playground).

to infer social relations. The graph model learns to classify relations with weights on contextual objects. Both above methods rely on object detection as a pre-processing step and only attend to regions with objects. Thus, we call the aforementioned attention mechanism for social relation inference as object attention.

Clearly, the power of object attention in prior methods is heavily restricted by the performance of object detection models. Besides, additional efforts are required to enlarge the number of object categories for attention when inferring social relations. For example, the number of object categories in backbone models for object detection based on COCO dataset is only 80, which might miss plenty of less frequent but important objects or clues affecting social relation inference. These limitations extremely hinder broader applications of prior methods.

It is essential and urgent to develop a new attention mechanism to achieve multi-level attention conditional to each person pair for social relation inference. On one hand, the new attention mechanism should remove the step of object detection. On the other hand, the new attention mechanism is multi-level. It adaptively attends to scenes, objects or human interactions not only for different relations within an image but also the same relation in different images.

Methods equipped with a preliminary object detection step for inferring social relations may directly re-use the trained models for object detection. However, they are limited not only by the number of objects but also by the performance of detection algorithm. In contrast, our novel attention mechanism offers two distinct benefits. The first advantage is its capacity to adopt a global perspective, enabling focused analysis of scenes, objects, or human interactions to uncover diverse relationships within an image. The second benefit lies in its nuanced attention mechanism toward objects, scenes, and human interactions. This nuance allows for contextual differentiation within the same social relation across different images.

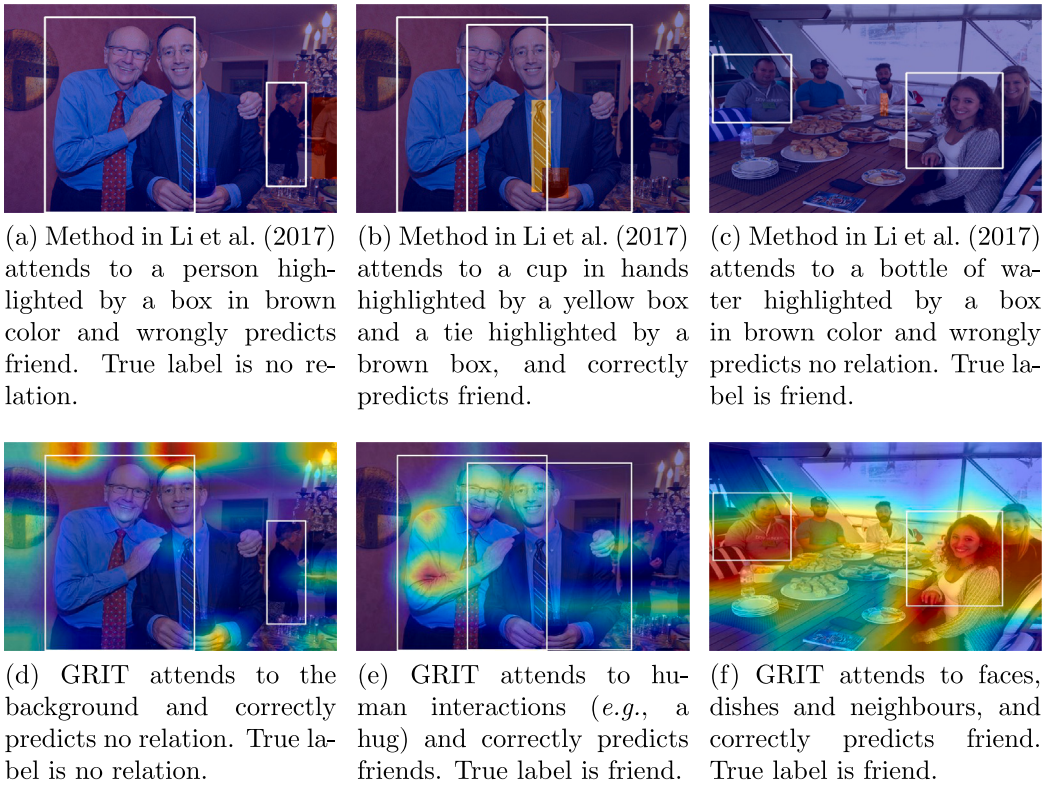
To demonstrate the significance of our novel attention approach and its differentiation from the methods currently in existence for this task, we provide a visual representation of four instances showcasing the attention patterns utilized by the dual-glance model from Li et al. (2017) and our model, as illustrated in Fig. 3. Our multi-level attention mechanism enhances the inference of social relations by considering scenes, objects, or human interactions, each of which plays a distinctive role in determining relations across various scenarios. Conversely, the models proposed in Li et al. (2017), Wang et al. (2018) were significantly constrained by the effectiveness of the object detection process.

Besides, although GCN has been applied in social relation inference (Wang et al., 2018; Zhang et al., 2019), most of them treat the problem as pairwise social relation recognition. This clearly limits the power of GCN. We take fully advantage of graph-based network for social relation inference by classifying relations of all person pairs in one image.

Motivated by effectively solving the problem of social relation inference, we propose a transformer-style network for the social relation graph of person pairs. We note that transformer networks are powerful in global self-attention, and thus may help achieve multi-level attention in this task. Also, the transformer-style network directly provides attention heatmap to explain the predicted category of a person pair.

In particular, we develop a MUlti-level Conditional Attention (MUCA) mechanism for social relation inference, and show its comparison with prior studies in Fig. 2. The MUCA mechanism enables models to attend to multiple clues simultaneously, including scenes, objects and human interactions. It is clear that MUCA benefits social relation inference because scenes, objects or human interactions may contribute to the inference of relations in different ways for various scenarios. By contrast, the object attention mechanism in the model of Li et al. (2017) was heavily restricted by the number of categories in object detection models, as shown in Fig. 2.

In this paper, we propose Graph-based Relation Inference with Transformer (GRIT), which is trained in an end-to-end manner. GRIT consists of a novel graph-based Conditional Query Module (CQM) and a Relation Attention Module (RAM). The proposed CQM is a graph



(a) Method in Li et al. (2017) attends to a person highlighted by a box in brown color and wrongly predicts friend. True label is no relation.

(b) Method in Li et al. (2017) attends to a cup in hands highlighted by a yellow box and a tie highlighted by a brown box, and correctly predicts friend.

(c) Method in Li et al. (2017) attends to a bottle of water highlighted by a box in brown color and wrongly predicts no relation. True label is friend.

(d) GRIT attends to the background and correctly predicts no relation. True label is no relation.

(e) GRIT attends to human interactions (*e.g.*, a hug) and correctly predicts friends. True label is friend.

(f) GRIT attends to faces, dishes and neighbours, and correctly predicts friend. True label is friend.

Fig. 2. Comparison of procedures on the prior studies with object attention and our MUCA mechanism for social relation inference. The MUCA mechanism adaptively attends to scenes, objects or human interactions for each person pair, and removes the pre-task of object detection.

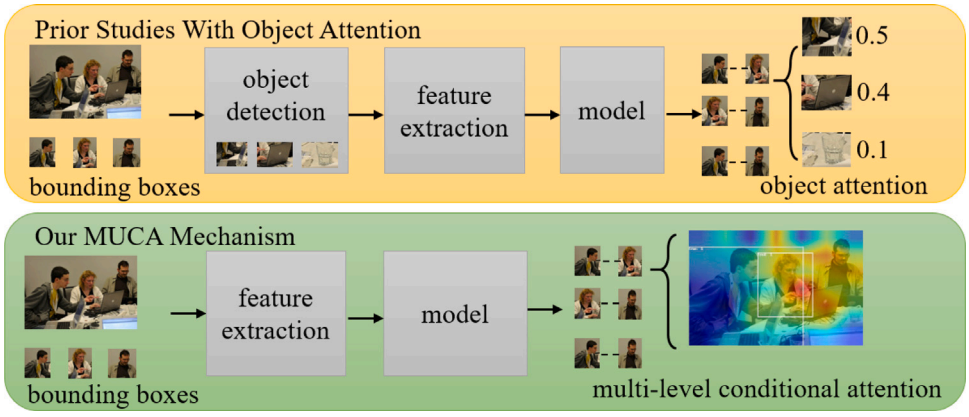


Fig. 3. Comparisons of attention in social relation inference between the dual-glance model (Li et al., 2017) and our GRIT. The pair of persons with two white boxes in each image is the target to infer social relation. The dual-glance model (Li et al., 2017), which predicts a relation for a target pair of persons based on weights on detected objects, actually performs object attention. In contrast, the proposed GRIT achieves MUCA in various scenarios. Figs. 3(d) and 3(e) show MUCA on the background and human interactions for different relations within an image. Fig. 3(f) shows multi-level attention for the same social relation. The attention region in each image is highlighted with heatmap.

neural network that is designed to fuse local person features and global context to generate informative relation queries for each person pair. The relation queries may help to capture logical constraints among different types of social relations. Inspired by Vaswani et al. (2017), we design RAM with an encoder–decoder transformer for classifying social relations, which contains powerful self-attention units. The combination of CQM and RAM achieves MUCA in GRIT, i.e., a relational query for each person pair in CQM and a global transformer decoder in RAM produce MUCA. To the best of our knowledge, we are the first to tackle the social relation inference problem with multi-level conditional attention. GRIT in this study achieves new state-of-the-art results on two benchmark datasets for social relation inference.

We summarize our contributions in this work as follows.

- We develop a novel framework which is named GRIT by taking advantage of global self-attention units and graph representation learning for social relation inference. The proposed GRIT achieves MUCA without a separate object detection in classifying social relations.
- We design a graph-based CQM in the framework, which is experimentally verified to be effective for social relation inference.

The graph-based CQM enables GRIT to classify the relations of all person pairs in an image within a single pass.

- GRIT establishes new state-of-the-art results on two benchmark datasets of social relation inference. The proposed GRIT significantly outperforms previous object attention methods, *e.g.*, 6.3% absolute improvement for fine relation on PISC dataset with Resnet101 backbone. It is worth mentioning that GRIT outperforms the current state-of-the-art by 7.8% and 9.6% on the fine relation inference task on PIPA and PISC datasets respectively.

2. Related work

To assess our contributions in the classification of social relations, we report three streams of studies: social relation inference, graph neural networks and transformer.

2.1. Social relation inference

Recently, there has been a recent surge of research interest in social relation inference based on images. There are two benchmark datasets for this task. Specifically, one dataset is named as People In Photo Albums (PIPA) in Zhang, Paluri, et al. (2015) and the other is named as People in Social Context (PISC) in Li et al. (2017).

The pioneering work on social relation inference dates back to 2010 from (Wang, Gallagher, Luo, & Forsyth, 2010), where the authors developed a model to characterize the interaction between multi-person actions, facial appearances and identities. The model recognized the family relationships, such as husband–wife, siblings, grandparent–child, father–child, or mother–child. Zhang, Luo, Loy, and Tang (2015) developed a deep neural network to learn social relation traits from rich facial attributes, such as expression, gender, and age. In Zhang et al. (2015), the social relation traits were defined based on psychological studies, consisting of eight types, e.g., trusting and friendly (Gottman, Levenson, & Woodin, 2001; Hess, Blairy, & Kleck, 2000). Recently, social relation inference was studied from viewpoints of multi-modal information and few-shot learning (Wan et al., 2021).

The availability of datasets plays an important role in technology advancements. In the field of social relations, Zhang, Paluri, et al. (2015) distributed a dataset to evaluate classification of social relations, which is named as People In Photo Albums (PIPA). Besides, another dataset, which is People in Social Context (PISC), was published in Li et al. (2017).

With PIPA and PISC, several interesting works move forward along the research line of social relation understanding (Goel, Ma, & Tan, 2019; Li, Duan, Lu, Feng, & Zhou, 2020; Sun, Schiele, & Fritz, 2017; Wang et al., 2018). In light of domain based theory from social psychology, Sun et al. (2017) presented a model with semantic attributes to classify social relations and domains. Wang et al. (2010) modeled a knowledge graph with proper messages propagation and attention to learn the social relations among people in an image. Recently, Goel et al. (2019) proposed an end-to-end neural network to learn the interaction graph of persons. In Li et al. (2020), a social graph was proposed to restrict logical connections of persons. In Li et al. (2021), a new framework for extraction and fusion of the hybrid features in social relation classification is proposed, and the method is named as HF-SRGR. HF-SRGR achieved the state-of-the-art results in social relation inference. In Table 1, we present the differences between our GRIT and the prior studies in terms of attention format, as well as the requirement of object detection as a pre-task. It is clear that our GRIT makes a big progress in attention for social relation inference.

Table 1
Comparisons between our GRIT and previous methods in attention for social relation inference. *IG* means interaction graph and *OD* means object detection.

Methods	IG	Attention	OD as Pre-task
Pair CNN (Li et al., 2017)	No	No	No
Dual-Glance (Li et al., 2017)	No	Object	Yes
SRG-GN (Goel et al., 2019)	Yes	No	No
GRM (Wang et al., 2018)	Yes	Object	Yes
MGR (Zhang et al., 2019)	Yes	No	No
GR^2N (Li et al., 2020)	Yes	No	No
HF-SRGR (Li et al., 2021)	Yes	Object	Yes
Our GRIT	Yes	Multi-level	No

2.2. Graph Neural Networks (GNNs)

Sperduti and Starita (1997) first applied the structure of neural networks into learning patterns of complicated data, motivating the early studies of Graph Neural Networks (GNNs). The concept of GNN was formed in Gori, Monfardini, and Scarselli (2005), and GNNs were firstly investigated for learning relation data with generalization ability in Scarselli, Gori, Tsoi, Hagenbuchner, and Monfardini (2008).

Recently, inspired by the success of deep convolutional networks in the computer vision domain, Kipf and Welling proposed a novel design of convolutional networks for graph-structured data (Kipf & Welling, 2017). Basically, deep convolutional networks effectively learns internal representation with spatial information (LeCun, Kavukcuoglu, & Farabet, 2010). For graph-structured data, it is essential to learn the feature representation of nodes and edges by propagating neighbor information (Cai, Li, Wang, & Ji, 2021; Defferrard, Bresson, & Vandergheynst, 2016; Isufi, Gama, & Ribeiro, 2021; Jiang, Zhu, Li, & Ji, 2020; Levie, Monti, Bresson, & Bronstein, 2018; Priebe, Shen, Huang, & Chen, 2021; Wang, Yan, & Yang, 2020). Later, efficient training of deep Graph Convolutional Networks (GCNs) has been widely studied, such as the model of GraphSAGE in inductive learning of graph (Hamilton, Ying, & Leskovec, 2017). By leveraging the self-attention layers, GNNs acquire a more powerful representation ability in dynamics of graph data (Liu, Wang, & Ji, 2021; Veličković, Cucurull, Casanova, Romero, Liò, & Bengio, 2018; Yang, Zhang, & Cai, 2020).

In social relation inference, GNNs have been introduced for representation learning of human interactions (Li et al., 2020; Wang et al., 2018; Zhang et al., 2019). For instance, Wang et al. (2018) applied gated GNN with a graph-attention mechanism on knowledge graphs to facilitate social relationship recognition. Zhang et al. (2019) designed a person-object graph and a person-pose graph, and conducted social relation reasoning on these two graphs by GNN. Li et al. (2020) proposed a graph relation reasoning network to infer social relations by building a graph for each image, where the nodes represent the persons and the edges represent the relations. In this paper, inspired by the design of GNN in Kipf and Welling (2017), we propose a graph-based CQM to extract relation queries with logical constraints among different types of social relations from an image.

2.3. Transformer

Transformer was first introduced by Vaswani et al. (2017) as a new building block with self-attention mechanism in natural language processing for machine translation. The attention mechanism in transformers usually refers to a neural network that aggregates information from the entire input, *e.g.*, the whole sentence in machine translation.

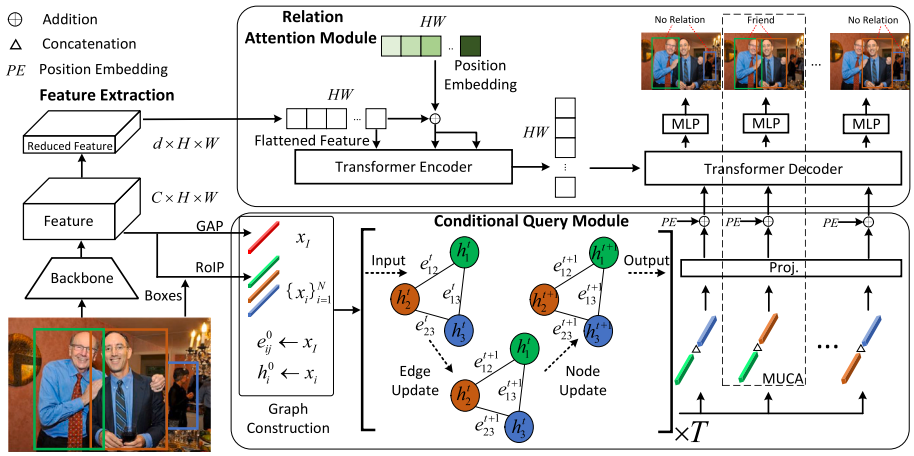


Fig. 4. The proposed end-to-end network (*i.e.*, GRIT) for social relation inference. Given an image with person bounding boxes as input, the step of feature extraction generates a set of features, including a dimension-reduced image feature map, a global image feature x_I (red) and RoI-pooled box features $\{x_i\}_{i=1}^N$. In the graph-based conditional query module, a complete graph is constructed with the node and edge features initialized from x_i and x_I, respectively. The edge and node features are updated alternatively in each layer of the conditional query module. After T rounds of updates, we concatenate two node features to generate a query feature for each person pair. The relation query and the reduced image feature map are fed into the relation attention module to produce relation prediction for the person pair. In particular, a relational query for each person pair in the conditional query module and a global transformer decoder in the relation attention module achieve MUlti-level Conditional Attention (MUCA). *RoIP* and *GAP* refer to RoI pooling and global average pooling, respectively. The "Proj". means a linear layer for dimension reduction.

Thus, the structure of transformer in Vaswani et al. (2017) naturally is capable of representation learning in a global view.

Transformer was originally designed for the sequence-to-sequence tasks (Sutskever, Vinyals, & Le, 2014), and then transferred to other domains (Chu et al., 2021; Guo et al., 2021; Han et al., 2022; Wang, Chakraborty, & Stella, 2021; Wang et al., 2021). In Vaswani et al. (2017), stacked encoders and decoders were designed to capture global dependencies regardless of token distance in a sequence. In Dosovitskiy et al. (2020), vision transformer was developed by stacking encoders and splitting images into patches. The challenge for vision transformer is its intensive computation complexity due to image size (Han et al., 2022). Recently, in Liu et al. (2021), Swin Transformer was proposed for solving the problem of computation complexity, as well as generalizing the structure of transformer.

We note that there are several studies on human object interactions with transformer (Kim, Lee, Kang, Kim, & Kim, 2021; Roy & Fernando, 2021; Zou et al., 2021). But those methods cannot be directly adopted to solve the problem of social relation inference because social relations characterize the connections between persons and also social relation inference requires more high-level understanding on scenes, objects and human interactions. Thus, in this work, we design a transformer-style network to attack the challenge of multi-level conditional attention in social relation inference.

3. Methodology

The overall architecture is presented in Fig. 4. The framework of GRIT technically consists of a CQM and an RAM. In GRIT, an image is first passed to extract RoI features of all persons and the global image feature. Then the person features and global image feature are fed into CQM. Specifically, the CQM is designed to fuse local person features and global context to generate informative relation queries for RAM. The relation queries may help to capture logical constraints among different types of social relations. Inside CQM, a stack of graph convolutional layers updates edges and nodes in an alternative manner. Based on relation queries and the flattened image feature, RAM infers the relations of all person pairs in an image within a single pass. The combination of CQM and RAM can adaptively attend to important clues for social relation recognition. GRIT is trained in an end-to-end manner. In the following, details of the modules will be introduced separately.

3.1. Feature extraction

Given an image with its person bounding boxes, we use a backbone model to extract features for GRIT. The backbone model in this study includes two options: transformer base model and Resnet101. We extracts two types of features, *i.e.*, the features of persons and the global image feature.

Starting from the initial image $I \in \mathbb{R}^{3\times H_0\times W_0}$, a backbone generates a feature map $f_I \in \mathbb{R}^{C\times H\times W}$. Typically, $H = H_0/32$, $W = W_0/32$, and the value of C depends on the backbone model. For instance, $C = 1024$ if the backbone is Swin Transformer Base model (Liu et al., 2021) and $C = 2048$ if the backbone is Resnet101 (He, Zhang, Ren, & Sun, 2016). We then extract the feature representations of each person and the whole image from the feature map f_I using RoI pooling (RoIP) and global average pooling (GAP), respectively. Specifically, given feature map f_I with N bounding boxes $b_1, b_2, \ldots, b_N$ for N persons in image I, we obtain the feature representations of the ith person in the image, denoted as x_i as follows:

$$x_i = RoIP(f_I, b_i) \in \mathbb{R}^C, \tag{1}$$

and the feature representation of the whole image x_I is extracted by applying a GAP layer on the f_I:

$$x_I = GAP(f_I) \in \mathbb{R}^C. \tag{2}$$

3.2. Conditional query module

The CQM is designed to fuse local person features x_i and global image feature x_I to generate relation queries for all person pairs in the image, which are used as inputs to the transformer's decoder in RAM. In essence, the design of CQM is inspired by Li et al. (2020), Wang et al. (2018), Zhang et al. (2019). In the following, we introduce the details of graph construction and iteration.

3.2.1. Graph construction

We build a graph for each image to generate relation queries. In the graph, each person in an image is modeled as a node, and each pair of persons in the image has a message-passing edge, *i.e.*, there is a complete graph for each image. Denote $\mathcal{G} = (\mathcal{V}, \xi)$ as the complete graph with node set $\mathcal{V}$ and edge set ξ in an image. For the ith node $v_i \in \mathcal{V}$ in $\mathcal{G}$, we set its initial node feature representation as $h_i^0 = x_i \in \mathbb{R}^C$. Correspondingly, each edge has a feature representation, and we set the initial edge feature representation between node i and node j as $e_{ij}^0 = x_I \in \mathbb{R}^C$.

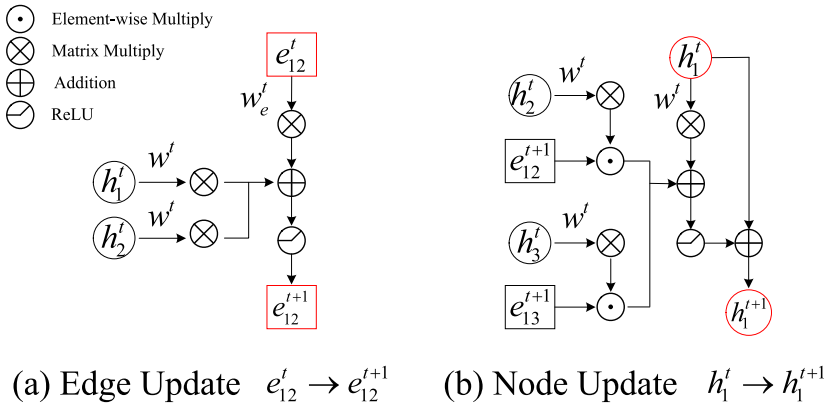


Fig. 5. Illustration of edge and node updates in CQM. The red boxes emphasize the units in $e_{12}^t \rightarrow e_{12}^{t+1}$ and $h_1^t \rightarrow h_1^{t+1}$.

3.2.2. Graph iteration

The CQM is a stack of graph convolutional layers, and there is an edge update and a node update in each layer, as illustrated in Fig. 5. In total, the edge and node feature representations are updated iteratively for T times in CQM. Specifically, at $(t+1)$-th layer the edge and node representations are updated as follows:

$$e_{ij}^{t+1} = \sigma(W^t h_i^t + W^t h_j^t + W_e^t e_{ij}^t), \tag{3}$$

$$h_i^{t+1} = h_i^t + \sigma(W^t h_i^t + \frac{1}{\bar{N}_i} \sum_{j \in \mathcal{N}_i} e_{ij}^{t+1} \odot W^t h_j^t), \tag{4}$$

where $\mathcal{N}_i$ is the set of neighbors for node i, $\bar{N}_i$ is the size of the set $\mathcal{N}_i$, $W^t, W_e^t \in \mathbb{R}^{C \times C}$ with $t = 0, 1, \ldots, T-1$ are the learnable parameters at each layer, σ is the ReLU function, and $\odot$ is the Hadamard point-wise multiplication operator. Intuitively, through iterative updates on edge and node representations, the local personal feature and global image feature are fused to form better feature representations.

Finally, we obtain relation query for each person pair by concatenating the node feature of the two persons from the last layer. Namely, we have

$$q_{ij} = \langle h_i^T, h_j^T \rangle,$$

where $\langle\rangle$ means concatenating feature vectors of two persons. Note that one can also directly use the edge features, or concatenate the edge features with node features from person pairs, to form the relation query. However, we empirically found that it did not lead to better performance.

The design of CQM is inspired by GatedGCN (Dwivedi, Joshi, Laurent, Bengio, & Bresson, 2020). We note that CQM in Eqs. (3)–(4) is an anisotropic variant of GCN (Dwivedi et al., 2020), while most of GNNs (such as GCN (Kipf & Welling, 2017) and GraphSAGE (Hamilton et al., 2017)) are isotropic. Typically in an anisotropic GCN, different neighbors have different weights in the node update equation. CQM employs learnable edge gates to learn the importance of different neighbors in Eq. (4). For comparisons of anisotropic and isotropic GNNs, one can refer to Dwivedi et al. (2020).

3.3. Relation attention module

The RAM is designed based on conventional encoder–decoder transformer (Vaswani et al., 2017). Intuitively, attention layers inside the transformer allow for automatically identifying important clues in images to help social relation inference.

3.3.1. Transformer encoder

We first apply an 1×1 convolution layer to reduce the dimension of the feature map f_I from C to d, and then flatten f_I into a sequence of visual tokens $\{f_i \in \mathbb{R}^d, i = 1, 2, \ldots, HW\}$. Each encoder layer has a standard architecture and consists of a multi-head self-attention module and a feed-forward network. Following traditional settings in other transformer models, we add a learnable positional embedding to the input of the encoder.

3.3.2. Transformer decoder

The inputs to the decoder are the relation queries among all person pairs from CQM, *i.e.*, q_{ij}, and a learnable positional embedding. As different images may have different number of persons, we pad the relation queries to the maximal number of person pairs in the dataset. Denote the maximal number of persons in the dataset as P, the relation queries are pad to have length of P^2. In addition, we add learnable positional embedding to the relation queries. By design, the positional embedding helps the decoder to be aware of different person pairs.

As RAM receives P^2 relation queries as input, correspondingly, the number of outputs from RAM is P^2. These outputs are passed through an MLP with shared weights for relation classification. We denote the final outputs as $\{x_{ij}\}_{1 \le i,j \le P}$. For a relation inference problem with K classes, we have $x_{ij} \in \mathbb{R}^K$.

3.4. Graph-based Relation Inference Transformer (GRIT)

The overall framework of GRIT, which consists of CQM and RAM, is jointly optimized end-to-end with cross-entropy as the objective function. Specifically, the loss function in GRIT can be expressed as follows:

$$\mathcal{L} = \sum_{i,j}^{P} \sum_{k=1}^{K} - \left[y_{ij}^k \log(x_{ij}^k) + (1 - y_{ij}^k) \log(1 - x_{ij}^k) \right], \tag{5}$$

where $y_{ij} \in \mathbb{R}^K$ is the one hot encoding vector of the ground true relation between the ith and the jth persons in the image and $\log(\cdot)$ is the logarithmic operator.

Although there are some pioneering studies that apply GCN in social relation inference (Wang et al., 2018; Zhang et al., 2019), most of them treat the problem as pairwise social relation recognition, *i.e.*, they predict social relations of person pairs independently. This clearly limits the power of GCN. Most recently, Li et al. (2020) proposed a GCN-based model called GR^{2}N to restrict logical connections of persons, which can simultaneously reason all relations for each image and achieved the state-of-the-art results in social relation inference. Although CQM and GR^{2}N (Li et al., 2020) both are variants of GCN, our CQM is more scalable. Specifically, as we expand the range of social relation

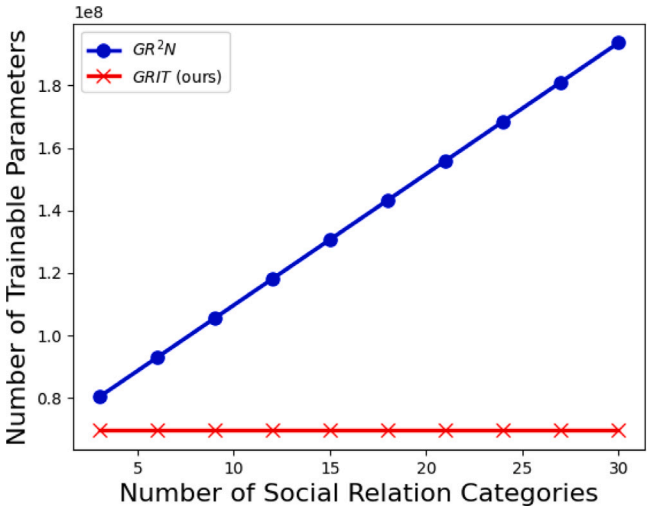


Fig. 6. Comparison on the number of trainable parameters between GR^{2}N and GRIT with varying numbers of social relation categories.

categories, both GRIT and GR^{2}N encounter an increase in the number of trainable parameters. However, it is worth noting that the growth in GRIT's parameters is negligible when compared to GR^{2}N. On one hand, we incorporate a Multi-Layer Perceptron (MLP) at the final stage to map relation features to the respective relation categories in GRIT. Consequently, the number of relation categories only affects the dimension of trainable parameters in the MLP layer, which constitutes only a small fraction of GRIT's overall parameters. On the other hand, GR^{2}N takes a different approach by constructing multiple virtual relation graphs to explicitly capture the robust logical constraints among various types of social relations. For a social relation inference problem involving K categories, GR^{2}N introduces K virtual relation graphs to model each of these K social relationships independently. This results in a linear dependency between the number of trainable parameters in GR^{2}N and the number of social relation categories, a more pronounced effect compared to GRIT. To illustrate this, we show the number of trainable parameters between GR^{2}N and GRIT when varying the number of social relation categories in Fig. 6. The number of trainable parameters of GR^{2}N is linearly dependent on the number of social relation categories. In contrast, the number of parameters in GRIT has negligible increase, which is around 69M under the default setting specified in Section 4.2.

In addition, compared to previous attention-based methods such as Dual-Glance (Li et al., 2017) or GRM (Wang et al., 2018), the attention in RAM is conditional to the graph-based query from CQM, and is multi-level and flexible. Clearly, RAM removes the step of object detection and is able to attend to scenes, objects or human interactions in an image conditional on each person pair.

4. Experiments

In this section, we conduct extensive experiments based on two image-based datasets (*i.e.*, PIPA and PISC datasets) and one video-based dataset (*i.e.*, VISR). We first present the description of datasets and the implementation details. Then we evaluate the performance of GRIT through comparisons with benchmarks and ablation study. Besides, we visualize the attention map of sample images from GRIT and prior works to show the effectiveness of MUCA. Finally, we present some discussions on the societal impact. The codes and experimental results are publicly available on github.[2]

[2] https://github.com/IFBigData/GRIT

4.1. Datasets

We conducted experiments on two image-based social relation datasets, namely the PIPA dataset (Zhang, Paluri, et al., 2015) and the PISC dataset (Li et al., 2017). Additionally, we conducted experiments on a video-based social relation dataset named VISR (Liu et al., 2019). The PIPA dataset partitions social life into five social domains (coarse-grained) and sixteen social relations (fine-grained). There are 13,729 person pairs for training, 709 for validation, and 5,106 for testing. The *accuracy* over all classes is typically used to evaluate all methods on PIPA dataset. The PISC dataset has a hierarchy of three coarse-grained relations (intimate, non-intimate, no relation) and six fine-grained relations (friend, family, couple, professional, commercial, and no relation). We follow the standard train/val/test split in Li et al. (2017). Specifically, for the coarse-grained relationship, we divide the dataset into a training set of 13,142 images and 49,017 relationship instances, a validation set of 4000 images and 14,536 instances and a test set of 4000 images and 15,497 instances. For the fine-grained relationship, the train/val/test set consist of 16,828 images and 55,400 instances, 500 images and 1505 instances, 1250 images and 3961 instances, respectively. For more details and statistics of the datasets, please refer to Appendix. VISR comprises 8240 video clips, each with a duration ranging from 10 to 30 s and eight types of social relation are defined in this dataset (*i.e.*, Leader-subordinate, Colleague, Service, Parent-offspring, Sibling, Couple, Friend and Opponent). While GRIT is designed to recognize social relationships between pairs of individuals, the VISR dataset labels are defined at the video level. Some videos within VISR contain more than two people, leading to label ambiguity because each video is assigned only one label. To adapt this dataset to our model, we implemented a uniform sampling approach, selecting 20 frames from each input video and only retaining those frames featuring precisely two individuals. This preprocessing step reduced the dataset from 8240 videos to 6982 videos, resulting in a total of 131,940 images. During our experiments, we divided the dataset randomly into training, validation, and testing subsets using a ratio of 7:1:2.

4.2. Implementation details

We train GRIT with Adam Optimizer in an end-to-end manner and set the learning rate of backbone and the rest of network to 10^{-5} and 10^{-4}, respectively. The GRIT network is trained for 10 epochs with a batch size of 16. The number of layers in CQM is set to be 2, *i.e.*, $T = 2$. The transformer has 3 encoder and 3 decoder layers, each with 8 self-attention heads. The dropout in RAM is set as 0.2.

We adopt different backbones in GRIT. We report the results with ImageNet pretrained Resnet101 model as backbone and call the model as GRIT-R101. Following Li et al. (2020), the input images are resized to 448 × 448 for GRIT-R101. In addition, we also report results with ImageNet pretrained Swin Transformer Base model (*i.e.*, Swin-B in Liu et al., 2021) as backbone, denoted as SW224. The input image are resized to 224 × 224 and the corresponding models are called GRIT-SW224.

We conduct all our experiments on a server with Intel Xeon(R) E5-2683 CPU, 512 GB memory and 4 GeForce RTX 2080Ti GPUs. Using the Distributed Data Parallel mode in pytorch, we can run each experiment in parallel on 4 GPUs. In this case, it costs around 1467 s to train GRIT-SW224 on PISC dataset.

4.3. Evaluation metrics

The mean average precision (mAP) is a popular evaluation metric for object detection task (localization and classification). Following the definition in Everingham, Van Gool, Williams, Winn, and Zisserman (2010), for a given task and class, the precision/recall curve is computed by ranked output from a method. Precision is defined as the proportion of all examples above that rank which are from the positive

class. The average precision (AP) is a way to summarize the precision–recall curve into a single value representing the average of all precision. It is defined as the mean precision at a series spaced recall level [0, $\frac{1}{M}$, $\frac{2}{M}$, ...,1] where M is the number of positive samples.

$$AP = \frac{1}{M} \sum_{r \in \left\{0, \frac{1}{M}, \ldots, 1\right\}} P_{interp}(r). \tag{6}$$

At each recall level of r, the precision is the maximum precision for which the corresponding recall exceeds r:

$$P_{interp}(r) = \max_{\hat{r}:\hat{r} \geq r} p(\hat{r}). \tag{7}$$

The mAP is computed by taking mean of AP of each class.

4.4. Comparisons with benchmarks

In this subsection, we compare GRIT with the following existing methods. For fair comparisons, we report the best results in experiments for Tables 3 and 4 following the routine in this research field.

4.4.1. Discussions on benchmarks

We present the details of prior methods as follows.

Pair-CNN (Li et al., 2017): A backbone with RoI pooling is used to extract features of two persons, which are concatenated and fed into an MLP for classification.

Dual-Glance (Li et al., 2017): The first glance focuses on the pair of persons. The second glance extracts the information of objects in the context and apply attention mechanism to refine the prediction.

SRG-GN (Goel et al., 2019): Scene and human attribute context features are extracted by five CNNs. These Features are passed through a graph inference network using GRU and message passing scheme.

GRM (Wang et al., 2018): A weighted graph is constructed to represent the persons and objects existing in an image, and then a gated graph attention network is applied to predict social relations.

MGR (Zhang et al., 2019): Two GNNs are applied to learn features in the person-pose graph and the person-object graph.

GR^{2}N (Li et al., 2020): A GNN is designed to model all relationships in a graph which can provide strong logical constraints among different types of social relations.

HF-SRGR (Li et al., 2021): A framework for extraction and fusion of the hybrid features in social relation classification is proposed, and is named as social relation graph reasoning model driven by hybrid-features (HF-SRGR).

For fair comparisons, we also report the results in PairCNN, Dual-Glance, GRM and GR^{2}N by substituting their backbone with SW224. For ease of presentation, the corresponding models are called Pair-SW224, DG-SW224, GRM-SW224 and GR^{2}N-SW224, respectively. In particular, Dual-Glance uses both Resnet101 and VGG as its backbone, in DG-SW224 we replace both of them with SW224. We note that the source codes of SRG-GN and MGR are not available and their results cannot be reproduced. Besides, as their performance are inferior to GR^{2}N, we do not include the paper results of SRG-GN and MGR in Tables 3 and 4. Interested readers can refer to their papers (Goel et al., 2019; Zhang et al., 2019) for performance comparison when using resnet as backbone. We note that the result of GRM for coarse relation recognition on PIPA dataset in Table 2 is empty because the adjacency matrix in this case was not released by its authors.

All of the prior methods, except GR^{2}N, treat the problem as pairwise social relation recognition, *i.e.*, they predict social relations of person pairs separately. In contrast, GR^{2}N and our GRIT consider the social relations among all persons in one image jointly. Besides, GR^{2}N lacks attention mechanism by design. Dual-Glance and GRM use object attention to assist in social relation inference. Our GRIT adopts a graph-based transformer-style network to achieve multi-level conditional attention.

Table 2
Comparisons between GRIT and other object attention methods on PIPA and PISC datasets with backbones *Resnet101*. We report accuracy on PIPA and mAP on PISC (in %).

Dataset	Method	Coarse (Domain)	Fine (Relation)
PIPA	Dual-Glance	65.2	59.6
	GRM	–	62.3
	GRIT-R101 (ours)	**73.4** (+8.2)	**66.7** (+4.4)
PISC	Dual-Glance	79.7	63.2
	GRM	82.8	68.7
	GRIT-R101 (ours)	**84.6** (+1.8)	**75.0** (+6.3)

Table 3
Comparisons of the accuracy (in %) between GRIT and other state-of-the-art methods on PIPA dataset.

Backbone	Method	Coarse (Domain)	Fine (Relation)
R101	Pair-CNN	65.9	58.0
	Dual-Glance	65.2	59.6
	GRM	–	62.3
	GR^{2}N	72.3	64.3
	GRIT-R101 (ours)	**73.4** (+1.1)	**66.7** (+2.4)
SW224	Pair-SW224	77.6	70.7
	DG-SW224	79.1	70.5
	GRM-SW224	–	69.9
	GR^{2}N-SW224	78.8	69.2
	GRIT-SW224 (ours)	**80.4** (+1.3)	**71.5** (+0.8)

4.4.2. Experimental comparison

In the comparisons, we address the following questions with experimental results shown in Tables 2, 3 and 4.

Q1: How does GRIT perform by comparing all the attention-based methods?

Q2: By using Resnet101 as backbones, how do GRIT and other methods perform?

Q3: By using SW224 as backbone, how do GRIT and other methods perform?

For Q1, we note that Dual-Glance, GRM and GRIT all use attention mechanism. The results of these methods are presented in Table 2. We observe that GRIT significantly outperforms other two object attention methods. For instance, GRIT-R101 achieves 6.3% absolute improvement in fine relation recognition on PISC dataset compared to GRM. Recall that both Dual-Glance and GRM require object detection as a preprocessing step. Their performance heavily rely on the object detection algorithm, i.e., the number of object classes to be detected and the detection accuracy. In contrast, our proposed GRIT is an end-to-end framework that can automatically attend to scenes, objects or human interactions in the images. These results validate the effectiveness of MUCA in GRIT.

For Q2, by following the backbone setting of GR^{2}N, we show the results of GRIT with Resnet101 in Tables 3 and 4. We observe that GRIT-R101 outperforms all other methods on PIPA dataset and improves the current state-of-the-art method, *i.e.*, GR^{2}N, by 1.1% absolute improvement for coarse relation recognition and 2.4% absolute improvement for fine relation recognition, respectively. Consistent observations can be found from Table 4. We observe that GRIT-R101 achieves an mAP of 84.6% for the coarse-grained recognition and 75.0% for the fine-grained recognition, outperforming GR^{2}N by 1.5% and 2.3% in absolute improvements, respectively. This result demonstrates the superiority of GRIT with Resnet101 as backbone.

For Q3, we observe that GRIT-SW224 consistently outperforms all previous methods and establishes new state-of-the-art results on both

Table 4
Comparisons of the per-class recall for each relation and the mAP over all relations (in %) between our GRIT and other state-of-the-art methods on PISC dataset. "Back.": Backbone, Int: Intimate, Non: Non-Intimate, NoR: No Relation, Fri: Friend, Fam: Family, Cou: Couple, Pro: Professional, Com: Commercial.

Back.	Method	Coarse relationships				Fine relationships						
		Int	Non	NoR	mAP	Fri	Fam	Cou	Pro	Com	NoR	mAP
R101	Pair-CNN	70.3	80.5	38.8	65.1	30.2	59.1	69.4	57.5	41.9	34.2	48.2
	Dual-Glance	73.1	84.2	59.6	79.7	34.4	68.1	76.3	70.3	57.6	60.9	63.2
	GRM	81.7	73.4	65.5	82.8	59.6	64.4	58.6	76.6	39.5	67.7	68.7
	GR^2N	81.6	74.3	70.8	83.1	60.8	65.9	84.8	73.0	51.7	70.4	72.7
	GRIT-R101 (ours)	83.6	74.0	70.1	**84.6** (+1.5)	66.3	65.4	52.0	79.4	33.3	79.6	**75.0** (+2.3)
SW224	Pair-SW224	82.2	74.2	49.0	75.9	71.8	69.0	53.1	84.7	29.1	51.9	69.6
	DG-SW224	84.4	76.8	53.1	78.9	69.5	68.4	59.8	87.9	33.6	56.1	70.3
	GRM-SW224	81.0	75.4	56.0	78.2	68.8	69.2	60.5	84.6	34.5	56.0	70.9
	GR^2N-SW224	85.8	78.9	67.5	84.2	70.0	67.5	62.9	88.6	25.1	82.4	75.3
	GRIT-SW224 (ours)	84.7	76.8	70.8	**85.6** (+1.4)	70.6	71.9	55.5	87.7	25.7	78.3	**78.3** (+3.0)

Table 5
Comparisons between GRIT and other methods on PIPA and PISC datasets with different backbones. We report accuracy on PIPA and mAP on PISC (in %).

Backbone	Method	PIPA		PISC	
		Coarse (Domain)	Fine (Relation)	Coarse	Fine
R101	Pair-CNN	65.9	58.0	65.1	48.2
	Dual-Glance	65.2	59.6	79.7	63.2
	SRG-GN	–	53.6	–	71.6
	GRM	–	62.3	82.8	68.7
	MGR	–	64.4	–	70.0
	GR^2N	72.3	64.3	83.1	72.7
	HF-SRGR	–	65.9	84.6	73.3
	GRIT-R101 (ours)	73.4	66.7	84.6	75.0
SW224	GRIT-SW224 (ours)	80.4	71.5	85.6	78.3
SW384	GRIT-SW384 (ours)	**81.2** (+8.9)	**73.7** (+7.8)	**88.5** (+3.9)	**82.9** (+9.6)

datasets. In particular, GRIT-SW224 achieves 1.4% and 3.0% absolute improvements in PISC coarse and fine relationships recognition respectively as compared to previous best method. In contrast, GR^2N-SW224 outperforms other previous methods on PISC dataset, but its performance is inferior to other methods on PIPA dataset. The results in Tables 3 and 4 based on SW224 as backbone further validate the advantage of our proposed network architecture.

Note that from Table 4, we can observe that in GRIT, the relations COU and COM have the greatest decreases in per-class recall as compared to GR^2N. In fact, we verify that COU and COM have the smallest sample numbers in PISC dataset, *i.e.*, their sample numbers are only 3.1% and 1.7% of all the samples, respectively. Thus we conjecture that the performance decreases in these two relations are mainly due to the sample imbalance. We note that in GR^2N, the authors introduce sample weights in the loss to deal with the sample imbalance issue. In GRIT we do not apply any additional techniques to tackle this issue, which is not the focus of this paper and we will leave it as future work.

4.4.3. Comparison with different backbones

In this subsection, we report the experimental results with GRIT with different backbone settings. By combining the results from Q2 and Q3, we observe that most of the methods have improvements if we replace their backbone with SW224. For instance, GR^2N-SW224 achieves an mAP of 75.3% in PISC fine relationships recognition, which is 2.6% higher in absolute value as compared to GR^2N. It suggests that Swin Transformer can extract more representative features than Resnet model. In addition to Resnet101 and SW224, here we also report results with more transformer-based backbones, *i.e.*, SW384 (Liu et al., 2021). Specifically, SW384 is the Swin Transformer Base backbone pre-trained on ImageNet with input images resized into 384×384.

From Table 5, we can observe that among all the backbones, GRIT with SW384 as backbone achieves the best performance in this task on both datasets. In particular, there are around 9% absolute improvement in terms of accuracy on PIPA dataset and more than 10% absolute improvement in terms of mAP in fine relation recognition on PISC dataset. This further suggests that Swin Transformer Base is a powerful backbone that can generate representative features.

4.4.4. Unified mAP

We note that all the previous methods report accuracy on PIPA dataset while report mAP on PISC dataset. The inconsistency is likely to prejudice comparison of performance among different methods. To address this issue, we propose to report mAP on both datasets. Beside the best mAP, we report the mean and standard deviation of mAP among 10 random runs, which can reduce the influence of randomness.

In our experiments, we find that GRIT-SW384 consistently achieves superior performance as compared to all previous methods. The results are shown in Table 6. In particular, there is at least an absolute improvement of 4% on both datasets if we compare GRIT-SW384 with previous methods. In contrast, we can observe that none of the previous methods consistently outperforms others in terms of mAP on both datasets. For example, GR^2N-SW224 performs well on PISC dataset, but its performance is inferior to the simple Pair-SW224 on PIPA dataset. Besides, Table 6 also shows the experiment results on VISR dataset. It is worth highlighting that our proposed GRIT demonstrates superior performance compared to all other methods in VISR. Notably, GRIT-SW384 achieves the highest performance, boasting a remarkable mAP of 51.3.

Table 6
Comparisons of our GRIT and other methods using Swin Transformer as backbone on various datasets. We report the mean and standard deviation of mAP (in%) among 10 random runs. "Abs. Imp." is short for "Absolute Improvement".

Method	PIPA		PISC		VISR
	Coarse (Domain)	Fine (Relation)	Coarse	Fine	
Pair-SW224	60.2 ± 0.3	38.3 ± 0.1	75.6 ± 0.2	69.2 ± 0.3	44.2 ± 0.5
DG-SW224	60.8 ± 0.2	38.3 ± 0.4	78.3 ± 0.2	69.7 ± 0.5	43.9 ± 0.5
GRM-SW224	–	36.4 ± 0.5	77.8 ± 0.1	70.4 ± 0.5	–
GR^2N-SW224	58.3 ± 0.5	34.1 ± 0.6	83.9 ± 0.2	73.0 ± 2.8	47.0 ± 0.6
GRIT-SW224	64.0 ± 0.2	39.2 ± 1.4	85.4 ± 0.1	77.6 ± 0.3	49.2 ± 0.6
GRIT-SW384	**65.4 ± 0.6**	**44.8 ± 1.1**	**88.3 ± 0.1**	**82.5 ± 0.2**	**51.3 ± 0.5**
Abs. Imp.	4.6	6.5	4.4	9.5	4.3

Table 7
Comparison of GRIT and previous methods in terms of inference speed and computational cost on PIPA-Fine test dataset with batch size being 1.

Methods	Images/s	GPU memory	Number parameters	FLOPs	Acc
Pair-CNN	21	920MB	59M	31.3G	58.0
Dual-Glance	16	2194MB	266M	66.6G	59.6
GR^2N	27	1208MB	134M	17.5G	64.3
GRIT-R101 (ours)	29	984MB	69M	16.9G	66.7

Table 8
Ablation study of GRIT's module on PISC dataset. We report the mean and standard deviation of mAP (in %) among 10 random runs on PISC dataset.

Settings	Coarse	Fine
Pair-SW224	75.6 ± 0.2	69.2 ± 0.3
GRIT-SW224 w/o CQM	84.6 ± 0.2	75.9 ± 0.6
GRIT-SW224 w/o RAM	84.4 ± 0.1	76.4 ± 0.4
GRIT-SW224	**85.4 ± 0.1**	**77.6 ± 0.3**

Table 9
Ablation study of transformer module with different number of encoder and decoder on PISC dataset. We report the mean and standard deviation of mAP (in %) among 10 random runs on PISC dataset.

Settings	Coarse	Fine
Enc1 & Dec1	85.0 ± 0.2	78.0 ± 0.3
Enc3 & Dec3 (Default)	85.4 ± 0.1	77.6 ± 0.3
Enc6 & Dec6	85.0 ± 0.1	77.2 ± 0.1

4.4.5. Inference efficiency and computational cost

We show the inference efficiency and the computational cost of various methods in Table 7. The table highlights a few important findings. First, Dual-Glance utilizes Faster R-CNN for object detection within images and employs both ResNet-101 and VGG-16 for feature extraction. Although this method is quite competitive in accuracy, it significantly inflates the number of parameters, as well as the GPU memory and computational requirements during inference, as shown in Table 7. Second, despite Pair-CNN having the least number of parameters, it does not achieve the fastest inference speed nor the lowest computational expense. This could be due to two key issues. The initial factor is that Pair-CNN analyzes cropped patches of images for each individual, which adds to the processing time and computational load for feature extraction via ResNet. In contrast, our method extracts feature from the whole image and uses ROI pooling to get feature for each person. The second issue is that Pair-CNN is designed to predict relationships only between pairs of individuals, meaning that images with more than two people necessitate multiple iterations of the model to ascertain all pairwise relations. Lastly, we observe that GRIT delivers not only the highest inference speed and the lowest computational demand but also exemplifies superior performance with reasonable GPU memory usage. Besides, similar findings are also observed in other datasets in this study.

4.5. Ablation study

4.5.1. Module Ablation

The modules of RAM and CQM are the core of our GRIT model, which work together to perform MUCA. On one hand, CQM is customized for the social relation inference task to generate effective relational queries. On the other hand, RAM can generate multi-level attention on the image, which automatically identify important clues to assist in social relation inference. To verify the effectiveness of these two modules, we conduct a module ablation experiment. By default, we use SW224 as the backbone. We remove CQM (GRIT-SW224 w/o CQM) and RAM (GRIT-SW224 w/o RAM) from GRIT-SW224 respectively. We also compare Pair-SW224, which can be treated as removing CQM and RAM simultaneously and performing relation inference of person pairs separately. The experiment results are shown in Table 8. Instead of reporting only the best mAP, we report the mean and standard deviation of mAP among 10 random runs, which can reduce the influence of randomness. We observe that the mAP in both levels of PISC dataset drops significantly after removing CQM or RAM. The performance degradation demonstrates that the proposed GRIT needs RAM and CQM to work together, and thus can effectively perform social relation inference.

4.5.2. Structure Ablation in RAM

To further explore the influence of structure in GRIT, we design an experiment that uses different numbers of encoder and decoder layers in RAM. As shown in Table 9, we observe that the changes in mAP on both coarse and fine grained on PISC dataset are very minor. These results suggest that GRIT-SW224 is insensitive to the depth of RAM. By default, we set the number of layers in both encoder and decoder to be 3.

4.6. Qualitative evaluation

In this subsection, we show the sample images to illustrate the effectiveness of MUCA and discuss the limitation of GRIT. Besides, we visualize the attention results from GRIT and GRM to demonstrate the difference between MUCA and object attention. For GRIT, we use the multi-head cross-attention weights from the last decoder layer in RAM. We first average across multiple heads to get the attention scores on the feature map, and then rescale and resize them to the original image size to get the attention heatmap on the original image. Recall that

(a) GRIT attends to the human interaction between the child and the adult.

(b) GRIT attends to the chair and accordions of the players.

(c) GRIT attends to the background, food, body interaction, etc.

(d) GRIT attends to body contact.

(e) GRIT attends to human interactions.

(f) GRIT attends to background objects.

Fig. 7. Visualization of sample images with attention heatmap where GRIT makes correct predictions.

GRM first detects objects in the image and then learns attention weights on those detected objects. To visualize the attention results from GRM, we show the detected objects with colors that are proportional to the attention weights.

We show the sample images demonstrating the benefits of MUCA in Fig. 7. In particular, the proposed GRIT successfully attends to scenes, objects and human interactions in different scenarios for social relation inference. For example, the body contact is attended in Fig. 7(d), and a plenty of objects including the food and the table are attended in Fig. 7(c). The results in Fig. 7 clearly show the effectiveness of MUCA in GRIT.

Sample images with attention heatmap are shown in Fig. 8, which compare the attention results between our GRIT and the prior object attention method GRM. In Fig. 8(a), one customer is buying juice in front of a booth. GRIT understands the context by paying attentions on the juice, the oranges and the signboard in the booth. However, limited by the power of object detection algorithms in the pre-processing step, GRM only has attention weights on the oranges and the customer. In Fig. 8(b), a couple were rowing a boat on the lake. Although both GRIT and GRM notice about the boat, GRIT has multi-level conditional attention on extra information such as the persons and the water. In Fig. 8(c), two friends are looking at the computer. GRIT infers their relation by attending on the computer, while GRM ignores this vital information. In Fig. 8(d), two sport team members are playing soccer on a field. GRM attends to the soccer only, while GRIT has multi-level attention on the soccer, the field and human interactions.

We also present failure cases of GRIT in Fig. 9. In Fig. 9(a), we observe that our method fails to make correct predictions when the persons in the images are too small or being occluded by other objects. When there exists overlapping of persons, it is also difficult for GRIT to attend and make correct predictions, as shown in Fig. 9(b). Besides, GRIT may still lack the ability to attend to age. As illustrated in Figs. 9(c)–9(d), GRIT makes wrong predictions on scenario where age is crucial for social relation inference. Moreover, some hard samples (in Figs. 9(e)–9(f)) are difficult to predict, as they typically require the model to attend to some subtle but important clues.

5. Discussions

In this section, we discuss the rational of transformer for social relation inference. We also disclose the societal impacts.

5.1. Rational of GRIT for social relation inference

Our GRIT contains RAM (as shown in Fig. 4) to adaptively achieve the goal of multi-level attentions to scenes, objects and human interactions for social relation inference. Besides, GRIT is equipped with CQM to characterize the relationships of multiple person pairs in one image. GRIT significantly outperforms the state-of-the-art methods on datasets of PIPA and PISC.

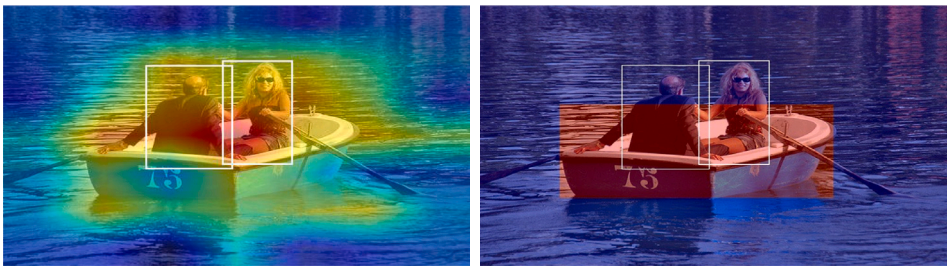
The side product of GRIT is the explainable results of predictions in social relation inference. One may easily justify the effectiveness of GRIT by directly visualize the attention heatmap over an image.

5.2. Societal impacts

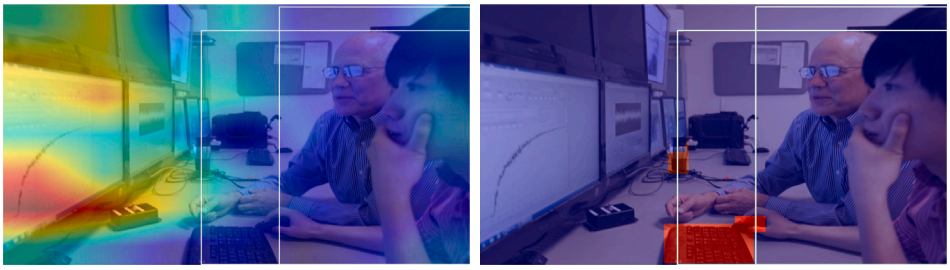
There are potentially many interesting real-world applications of our proposed model. For instance, GRIT can be used in social media platform as a friend recommendation engine based on still images. Or it can be used as a module to assist in the task of image caption generation. In general, GRIT can be used as a plus-and-play module for applications that involved understanding the social relations among people from still images.

As GRIT deals with social relation inference, one may raise concerns about the potential relation bias, *i.e.*, there may be severe biases against recognizing social relations among minority or non-traditional family groups. This is a common issue for most deep-learning based methods when there exists biases in the dataset that involved people. One of the effective ways to address this issue is to collect as many as data samples with diversity from various data sources to eliminate the biases in the training dataset. Note that potential negative societal impacts exist in most studies of deep learning, and they can be eventually solved with the advancement of research.

(a) Commercial.



(b) Couple.



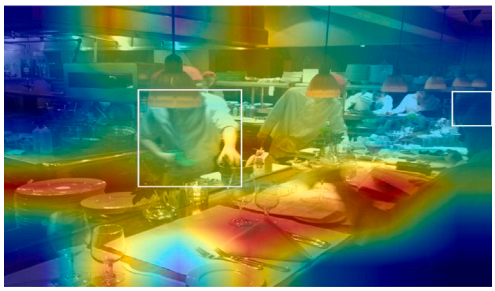
(c) Friend.

(d) Sport team members.

Fig. 8. Visualization of sample images with attention heatmap. Images in (a) and (b) are from PISC dataset. Images in (c) and (d) are from PIPA dataset. Images in the first and second columns are results from GRIT and GRM, respectively.

6. Conclusion

In this paper, we propose a novel model of GRIT for social relation inference from the viewpoint of multi-level conditional attention. GRIT consists of a graph-based Conditional Query Module (CQM) and a Relation Attention Module (RAM). We design a transformer-based network to achieve multi-level attention of images in a global view for classifying social relations. The graph-based CQM in GRIT is concise in learning the interaction of all person pairs in an image within a single pass. We develop an iterative update strategy for a graph to generate informative relation queries in CQM. We design RAM with an encoder–decoder transformer for classifying social relations, which contains powerful self-attention units. Extensive experiments clearly demonstrate that the proposed GRIT is superior than the prior methods. The proposed GRIT significantly outperforms previous attention-based methods and establishes new state-of-the-art results on PIPA and PISC datasets comparing to all existing methods, *e.g.*, with absolute improvements of 7.8% for fine relation on PIPA and 9.6% for fine relation on PISC. Ablation study and qualitative evaluation show the effectiveness of the proposed modules. The query design in this study might inspire new directions of investigations on new transformer for computer vision and natural language processing. Future directions lie in multi-level conditional attention networks for image caption generation, behavior analysis, etc, and also the adoption of pre-trained models for social relation inference.

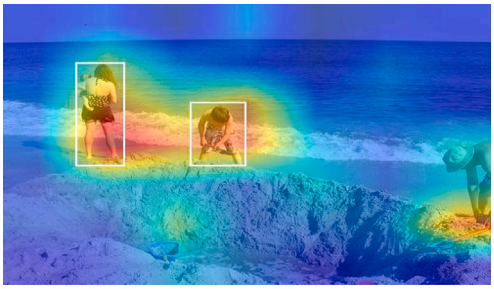


(a) True label is no relation, and predicted label is friend. Persons in the image are occluded by other objects.

(b) True label is family, and predicted label is friend. Bounding boxes of the two persons overlap too much.

(c) True label is gramma-grandchild, and predicted label is mother-child. GRIT may fail to attend to age.

(d) True label is mother-child, and predicted label is friends. GRIT may fail to attend to age.



(e) True label is family, and predicted label is friend. GRIT fails to attend to some important clues, such as the baby.

(f) True label is couple, and predicted label is family. As there is no intimate action, it is difficult to infer their relation.

Fig. 9. Visualization of sample images with attention heatmap where GRIT makes wrong predictions.

CRediT authorship contribution statement

Xiaotian Yu: Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Hanling Yi:** Methodology, Software, Writing – original draft, Writing – review & editing. **Qie Tang:** Conceptualization, Data curation, Methodology, Software, Visualization, Writing – review & editing. **Kun Huang:** Data curation, Visualization, Writing – review & editing. **Wenze Hu:** Writing – review & editing. **Shiliang Zhang:** Conceptualization, Supervision. **Xiaoyu Wang:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix. Statistics of the datasets

The statistics of PIPA and PISC in coarse-grained and fine-grained relations are shown in Tables A.10–A.13. Table A.14 provides a breakdown of the number of videos for various social relations within the VISR dataset. There are in total 8240 videos. We randomly select 20 frames from each video and only retaining those frames featuring precisely two individuals. Note that the train/val/test splits for PISC coarse and fine are different. We note that both PIPA and PISC have the problem of label imbalance. In particular, in the PISC fine dataset, the relations Couple and Commercial have the smallest numbers of samples among all the relations, as shown in Table A.12. This label imbalance issue may introduce bias in training data and thus pose additional challenge in model learning.

Table A.10
Statistics of PIPA fine dataset. We show the number of social relations in train/val/test set.

	Train.	Valid.	Test
Father-child	332	32	168
Mother–child	448	45	190
Grandpa–grandchild	46	3	11
Grandma–grandchild	37	0	15
Friends	3,054	187	1,833
Siblings	608	32	231
Classmates	128	71	13
Lovers/Spouse	503	49	313
Presenters–audience	194	12	91
Teacher–student	23	15	33
Trainer–trainee	83	1	54
Leader–subordinate	10	1	14
Band members	520	25	211
Dance team members	17	5	326
Sport team members	863	5	294
colleagues	6,863	226	1,309

Table A.11
Statistics of PIPA coarse dataset. We show the number of social relations in train/val/test set.

	Attach.	Reci.	Mating	Hierarch.r	Coal.	All
Train	863	3,790	503	310	8,263	13,729
Val	80	290	49	29	261	709
Test	384	2,077	313	192	2,140	5,106

Table A.12
Statistics of PISC fine dataset. We show the number of social relations in train/val/test set.

	Friend	Family	Couple	Professional	Commercial	No relation
Train	12,686	7,818	1,552	20,842	523	11,979
Val	332	249	102	311	164	347
Test	790	677	256	858	354	1,026

Table A.13
Statistics of PISC coarse dataset. We show the number of social relations in train/val/test set.

	Intimate	Non-intimate	No relation	All
Train	19,131	15,334	14,552	49,017
Val	5,205	4,942	4,389	14,536
Test	5,803	4,893	4,801	15,497

Table A.14
Statistics of VISR dataset. We show the number of social relations in train/val/test set.

	Train.	Valid.	Test
Leader-subordinate	718	102	205
Colleague	954	136	273
Service	393	56	112
Parent-offspring	716	102	205
Sibling	322	46	92
Couple	604	86	173
Friend	1,275	182	364
Opponent	786	112	225

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *IEEE conference on computer vision and pattern recognition* (pp. 961–971).

Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., et al. (2016). SALSA: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(8), 1707–1720. http://dx.doi.org/10.1109/TPAMI.2015.2496269.

Cai, L., Li, J., Wang, J., & Ji, S. (2021). Line graph neural networks for link prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, X., & Lawrence Zitnick, C. (2015). Mind's eye: A recurrent visual representation for image caption generation. In *IEEE conference on computer vision and pattern recognition* (pp. 2422–2431).

Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., et al. (2021). Twins: Revisiting the design of spatial attention in vision transformers. In *Annual conference on neural information processing systems*: *vol. 1*, (no. 2), (p. 3).

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*: *vol. 29*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *International conference on learning representations*.

Dwivedi, V. P., Joshi, C. K., Laurent, T., Bengio, Y., & Bresson, X. (2020). Benchmarking graph neural networks. *TSP*, *12*, 50–500.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, *88*(2), 303–338.

Fan, H., Zhang, F., Wei, Y., Li, Z., Zou, C., Gao, Y., et al. (2021). Heterogeneous hypergraph variational autoencoder for link prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/TPAMI.2021.3059313.

Goel, A., Ma, K. T., & Tan, C. (2019). An end-to-end network for generating social relationship graphs. In *IEEE conference on computer vision and pattern recognition* (pp. 11186–11195).

Gori, M., Monfardini, G., & Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings. 2005 IEEE international joint conference on neural networks*: *vol. 2*, (no. 2005), (pp. 729–734).

Gottman, J., Levenson, R., & Woodin, E. (2001). Facial expressions during marital conflict. *Journal of Family Communication*, *1*(1), 37–57.

Guo, Y., Zheng, Y., Tan, M., Chen, Q., Li, Z., Chen, J., et al. (2021). Towards accurate and compact architectures via neural architecture transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hamilton, W., Ying, Z., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Annual conference on neural information processing systems*: *vol. 30*.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/TPAMI.2022.3152247.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hess, U., Blairy, S., & Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal behavior*, *24*(4), 265–283.

Hoai, M., & Zisserman, A. (2014). Talking heads: Detecting humans and recognizing their interactions. In *IEEE conference on computer vision and pattern recognition* (pp. 875–882).

Isufi, E., Gama, F., & Ribeiro, A. (2021). EdgeNets: Edge varying graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jiang, X., Zhu, R., Li, S., & Ji, P. (2020). Co-embedding of nodes and edges with graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kim, B., Lee, J., Kang, J., Kim, E. S., & Kim, H. J. (2021). HOTR: End-to-end human-object interaction detection with transformers. In *IEEE conference on computer vision and pattern recognition* (pp. 74–83).

Kim, D. J., Oh, T. H., Choi, J., & Kweon, I. S. (2021). Dense relational image captioning via multi-task triple-stream networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International conference on learning representations*.

Kitayama, S., & Markus, H. R. (2000). The pursuit of happiness and the realization of sympathy: Cultural patterns of self, social relations, and well-being. *Culture and Subjective Well-being*, *1*, 113–161.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems* (pp. 253–256).

Levie, R., Monti, F., Bresson, X., & Bronstein, M. M. (2018). Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, *67*(1), 97–109.

Li, W., Duan, Y., Lu, J., Feng, J., & Zhou, J. (2020). Graph-based social relation reasoning. In *European conference on computer vision* (pp. 18–34). Springer.

Li, L., Qing, L., Wang, Y., Su, J., Cheng, Y., & Peng, Y. (2021). HF-SRGR: a new hybrid feature-driven social relation graph reasoning model. *The Visual Computer*, 1–14.

Li, J., Wong, Y., Zhao, Q., & Kankanhalli, M. S. (2017). Dual-glance model for deciphering social relationships. In *IEEE international conference on computer vision* (pp. 2650–2659).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE international conference on computer vision*.

Liu, X., Liu, W., Zhang, M., Chen, J., Gao, L., Yan, C., et al. (2019). Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *2019 IEEE/CVF conference on computer vision and pattern recognition* (pp. 3561–3569). http://dx.doi.org/10.1109/CVPR.2019.00368.

Liu, M., Wang, Z., & Ji, S. (2021). Non-local graph neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Marchetti, F., Becattini, F., Seidenari, L., & Del Bimbo, A. (2020). Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Priebe, C. E., Shen, C., Huang, N., & Chen, T. (2021). A simple spectral failure mode for graph convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Roy, D., & Fernando, B. (2021). Action anticipation using pairwise human-object interactions and transformers. *IEEE Transactions on Image Processing*, *30*, 8116–8129.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.

Sperduti, A., & Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, *8*(3), 714–735.

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/TPAMI.2022.3148210.

Sun, Q., Schiele, B., & Fritz, M. (2017). A domain based approach to social relation recognition. In *IEEE conference on computer vision and pattern recognition* (pp. 3481–3490).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Annual conference on neural information processing systems* (pp. 3104–3112).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Annual conference on neural information processing systems* (pp. 5998–6008).

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *International conference on learning representations*.

Wan, H., Zhang, M., Du, J., Huang, Z., Yang, Y., & Pan, J. Z. (2021). FL-MSRE: A few-shot learning based approach to multimodal social relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*: *vol. 35*, (no. 15), (pp. 13916–13923).

Wang, J., Chakraborty, R., & Stella, X. Y. (2021). Spatial transformer for 3D point clouds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Z., Chen, T., Ren, J., Yu, W., Cheng, H., & Lin, L. (2018). Deep reasoning with knowledge graph for social relationship understanding. In *International joint conferences on artificial intelligence* (pp. 1021–1028).

Wang, G., Gallagher, A., Luo, J., & Forsyth, D. (2010). Seeing people in social context: Recognizing people and social relationships. In *European conference on computer vision* (pp. 169–182). Springer.

Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., et al. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE international conference on computer vision*.

Wang, R., Yan, J., & Yang, X. (2020). Combinatorial learning of robust deep graph matching: an embedding based approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yan, R., Xie, L., Tang, J., Shu, X., & Tian, Q. (2020). HiGCIN: Hierarchical graph-based cross inference network for group activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, X., Zhang, H., & Cai, J. (2020). Auto-encoding and distilling scene graphs for image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, M., Liu, X., Liu, W., Zhou, A., Ma, H., & Mei, T. (2019). Multi-granularity reasoning for social relation recognition from images. In *International conference on multimedia and expo* (pp. 1618–1623).

Zhang, Z., Luo, P., Loy, C. C., & Tang, X. (2015). Learning social relation traits from face images. In *IEEE international conference on computer vision* (pp. 3631–3639).

Zhang, N., Paluri, M., Taigman, Y., Fergus, R., & Bourdev, L. (2015). Beyond frontal faces: Improving person recognition using multiple cues. In *IEEE conference on computer vision and pattern recognition* (pp. 4804–4813).

Zhang, P., Xue, J., Zhang, P., Zheng, N., & Ouyang, W. (2020). Social-aware pedestrian trajectory prediction via states refinement LSTM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. http://dx.doi.org/10.1109/TPAMI.2020.3038217.

Zhang, J., Yang, Y., Zhuo, L., Tian, Q., & Liang, X. (2019). Personalized recommendation of social images by constructing a user interest tree with deep features and tag trees. *IEEE Transactions on Multimedia*, *21*(11), 2762–2775.

Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., et al. (2021). End-to-end human object interaction detection with hoi transformer. In *IEEE conference on computer vision and pattern recognition* (pp. 11825–11834).