Research Article manuscript No.
(will be inserted by the editor)

CAGNet: A Context-aware Graph Neural Network for Detecting Social Relationships in Videos

Fan Yu · Yaqun Fang · Zhixiang Zhao · Jia Bei · Tongwei Ren · Gangshan Wu

Received: date / Accepted: date

Abstract Social relationships, such as parent-offspring and friends, are crucial and stable connections between individuals, especially at the person level, and are essential for accurately describing the semantics of videos. In this paper, we analogize such a task to scene graph generation, which we call video social relationship graph generation (VSRGG). It involves generating a social relationship graph for each video based on person-level relationships. We propose a context-aware graph neural network (CAGNet) for VSRGG, which effectively generates social relationship graphs through message passing, capturing the context of the video. Specifically, CAGNet detects persons in the video, generates an initial graph via relationship proposal, and extracts facial and body features to describe the detected individuals, as well as temporal features to describe their interactions. Then, CAGNet predicts pairwise relationships between individuals using graph message passing. Additionally, we construct a new dataset, VidSoR, to evaluate VSRGG, which contains 72 hours of video with 6276 person instances and 5313 relationship instances of eight relationship types. Extensive experiments show that CAGNet can make accurate predictions with a comparatively high mean recall (mRecall) when using only visual features.

Keywords Video analysis · Social relationship detection · Scene graph generation · Message passing

All the authors are with Nanjing University, State Key Laboratory for Novel Software Technology, Nanjing, China. (Email: yuf@smail.nju.edu.cn, fangyq@smail.nju.edu.cn, zhaozx@smail.nju.edu.cn, beijia@nju.edu.cn, rentw@nju.edu.cn, gswu@nju.edu.cn). Corresponding author: Jia Bei.

1 Introduction

Recent advances in the study of the synergy between vision and language have led to the understanding of video content, where the analysis of the relationships between two persons/objects has received much attention [1–3]. Existing research has mostly focused on the detection of visual relationships between objects [1,2], i.e., the interactions and spatial relationships represented visually in one or more video frames, while less attention has been paid to the more stable and essential relationships, such as social relationships [3–5]. As an indispensable part of people's daily lives, social relationship refers to the association between different people, such as colleagues, couples and friends [6].

Social relationship analysis lays the groundwork for high-level visual reasoning tasks. This analysis enhances the generation of descriptive visual captions [7,8], such as "the father gives an apple to his son" instead of "a man gives an apple to a child". It also improves the accuracy of answers to visual questions [9,10]. For instance, it is able to answer "his father" to the question "who is giveing the child an apple?" instead of just "a man". In addition, social relationship analysis enables the extraction of individuals' attributes in social media networks and facilitates the provision of personalized recommendations and services [11].

Traditional research on social relationship analysis focuses primarily on recognizing social relationships in images [12–15]. However, image-based social relationship analysis methods are not suitable for video analysis tasks because they lack the ability to aggregate temporal information and recognize related pairs of people that do not appear in the same frame. Previously, video-based methods only annotated one or more social relationships in a given video [3, 4] without

providing person-level social relationship analysis. The effectiveness of understanding video content is limited because, especially when there are multiple people in the videos, existing methods are unable to determine which pair of people or relationship is being described or which relationship is being mentioned.

Person-level social understanding [16–18] requires predicting the social relationship between two persons. We draw an analogy between this task and scene graph generation [19–21], where the goal is to construct a graph of social relationships between individuals in a video. We refer to this task as video social relationship graph generation (VSRGG). Given a video V (Fig. 1 (a)), VSRGG requires detecting the set of people $P = p_k$ appearing in V, and determining the social relationship of each pair of people $< p_i, p_j >$ or no social relationship between them. VSRGG presents the social relationship analysis result of a video in the form of a graph, named social relationship graph (SRG), where each vertex denotes a detected person in P and the edge between two vertexes denotes the social relationship of the corresponding pairs of people (Fig. 1 (b)).

Compared to existing tasks for analyzing social relationships in images and videos, VSRGG faces several technical challenges. The richness of pairs of people with social relationships in videos exceeds that of images. In video, two individuals can have a social relationship without appearing in the same frame. This increases the computational complexity and costs of VSRGG compared to the tasks analyzing social relationships in images. Furthermore, detecting social relationships at the person-pair level proves to be more challenging than detecting social relationships at the video level. Because the performance of the methods in VSRGG relies heavily on the accuracy of person detection, we introduce another task called weak video social relationship graph generation (VSRGG*). VSRGG* takes video, along with a collection of character faces and bodies, as input to generate the social relationship graph (SRG). In other words, VSRGG* uses annotated faces and bodies instead of detected faces and bodies. Compared to VSRGG, VSRGG* focuses specifically on the ability of the method to perform classification.

Most methods for VSRGG use multimodal features extracted from video, audio, and transcripts [5, 16–18, 23–25]. However, visual features are not completely exploited, especially in the fusion of character features and temporal context features. The methods in Refs. [5, 16, 18, 23, 25] encode visual features sequentially, while the methods in Refs. [17, 24] encode graphs to update visual features. Sequential encoding methods often combine temporal information with character features using a simple concatenation strategy. Wu et al. [24]

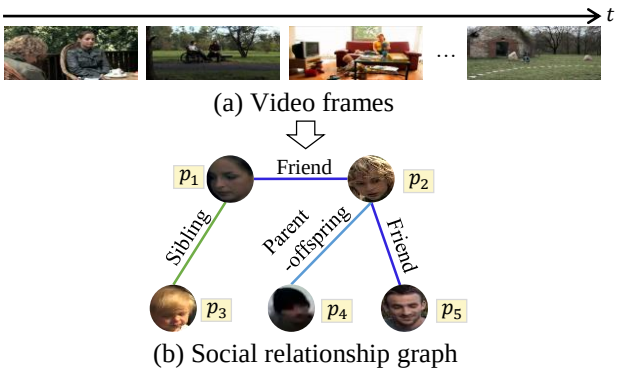

Fig. 1 An example of the generation of a social relationship graph on a given video, where p represents person and t represents time (source from the DVU challenge [22]).

proposed combining frame-level graph convolutional network (GCN) and clip-level GCN to generate the final social graph. Frame-level GCN updates character representations and character-pair representations with multi-modal and multi-view information on each frame. Long short-term memory (LSTM) is used to temporally accumulate these representation features and generate the input node of the clip-level graph. Hu et al. [17] used an overall GCN and a distinctive-level GCN. The overall level GCN with intra-edges and inter-edges propagates the representation of each character. The distinctive-level GCN focuses on the interaction between two characters. LSTM is also used to encode visual features temporally. The two methods both encode temporal context with LSTM, which is still a sequential-style encoder with limited whole-time encoding capabilities. In this paper, we introduce a novel VSRGG method, named the context-aware graph neural network (CAGNet), which focuses on visual information encoding for social relationship analysis in video. Figure 2 shows an overview of CAGNet. CAGNet exploits the message passing mechanism to capture visual context while taking temporal information into account. To address the limited presence of social relationships in video, we propose a graph proposal module to construct a sparse SRG. Each edge in the graph represents a potential social relationship between two individuals. Then, facial, body and interaction features are extracted. Next, we propose a cross-vertex message passing mechanism that incorporates temporal context information into each edge representation during prediction, which can encode visual features in a global view. We employ both temporal context edge representation and discriminative vertex representation during prediction to identify the social relationship of each edge and remove edges without social relationships.

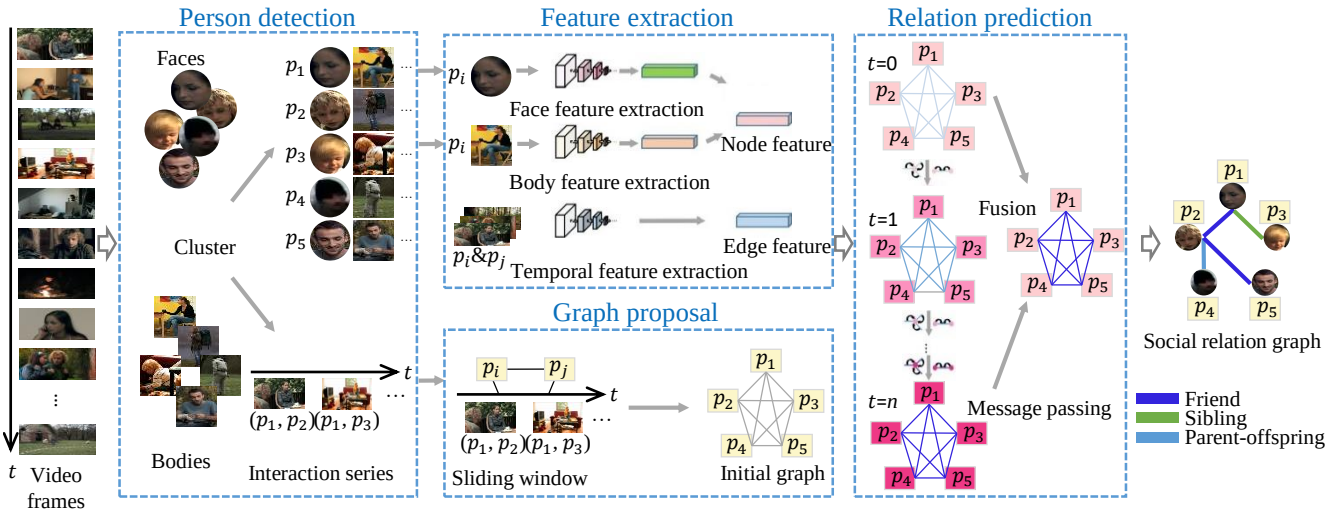

Fig. 2 An overview of the proposed context-aware graph neural network (CAGNet) method (source from the DVU challenge [22]).

The early video social relationship analysis datasets [3, 4] are annotated with video-level labels. Currently, MovieGraphs [26] is the most frequently used dataset for VSRGG, which was collected from 51 movies and focused on human-centered situations such as people's interactions, relationships, emotions, and motivations. BiliBili [16] videos have been collected for VSRGG; however, they have not been published yet. Hence, we construct another video social relationship dataset named VidSoR for VSRGG. The dataset consists of 1798 video clips with 8 relationships. In comparison to MovieGraphs, VidSoR has more valid video clips, the average duration of the clips is longer, and there are more average relationship instances in each clip. Thus, our dataset presents significant challenges in social relationship detection. We evaluate the performance of the proposed CAGNet method using this dataset. The results demonstrate that GACNet outperforms state-of-the-art baselines.

In summary, the contributions of this study are as follows.

1) We create VidSoR, a more challenging dataset for evaluating VSRGG. This dataset was collected from a larger number of data sources, and consists of more valid clips with a greater number of relationship instances and a longer average duration.
2) We propose a novel method called CAGNet for VSRGG. This method incorporates a temporal message passing mechanism to effectively utilize visual information in a global view. In addition, it achieves superior performance to methods that rely on multi-modal features.

2 Related Work

2.1 Social Relationship Recognition

As an important component of social network analysis, researchers have devoted significant attention to image social relationship recognition. Initially, researchers focused primarily on kinship recognition or verification, as demonstrated by the studies in Refs. [27–31]. Inspired by Ref. [32], Zhang et al. [6] utilized relationship traits to represent diverse relationships and employed human facial expressions to explore relationship traits between individuals. Based on domain theory [33], Sun et al. [12] introduced a dataset named "people in photo album", which included pair-wise relationships such as father-son, friends, and colleagues. A double-stream CaffeNet was utilized to perform classification in their works. Li et al. [13] defined social relationships according to prototype theory [34] and proposed a dataset named "people in social context". In their work, they introduced a dual glance net that can capture individual visual information and extract information of background objects. However, since static images are unable to capture the temporal dynamics between characters, this task may differ considerably from our task. Wang et al. [14] linked background objects with relationships between people and constructed a knowledge graph to assist in social relationship classification and reasoning. Zhang et al. [35] extracted global knowledge and intermediate-level information from the relative position of the image scene and objects to infer social relationships between people. Fine-grained information on key points of the human body was used to establish a person-pose graph. Goel et al. [15]

presented the initial social relationship graph generation task. Different from our social relationship graph, they utilized age and gender to distinguish different person instances rather than faces. They employed a message passing mechanism to merge the vertex feature with the edge feature and used the merged feature for pair-wise relationship prediction. In contrast, we pass messages for bi-direction, which enables each edge representation to capture the whole context of the social relationship graph.

2.2 Video Social Relationship Analysis

In recent years, researchers have begun to pay attention to video social relationship recognition. Lv et al. [3] defined video social relationship recognition as a multi-label video classification task and proposed the first video social relationship recognition dataset named social relationship in videos (SRIV), which contains approximately 3000 video clips with multi-label annotation. They designed a multi-stream model to extract and fuse features of RGB images, audio series and optical flow and used a late fusion strategy to merge different features for final classification. Liu et al. [4] introduced the first single label video social relationship dataset, the video social relationship (ViSR) dataset. In their work, three types of graphs were constructed that were designed to capture the behavior of each person, the interaction between different people, and the person-object relationship separately. Moreover, a novel pyramid graph convolutional network was designed to extract features from three graphs, which were recently merged with the global feature to perform final classification. MovieGraphs [26] and HLVU [36] were constructed for person-level social relationship analysis in videos. Kukleva et al. [5] presented a joint framework to predict both interactions and relationships between characters utilizing visual and textual features. Cao et al. [23] proposed fusing spatio-temporal and multi-modal semantic knowledge in videos. Xu et al. [16] designed a multi-stream architecture for jointly embedding visual and textual information after character pair searching. Multi-modal cues in a hierarchical-cumulative GCN structure was integrated by Wu et al. [24] to generate the social graphs for characters. Teng et al. [25] proposed a self-supervised multi-modal feature learning framework based on the Transformer model. Hu et al. [17] introduced a hierarchical-cumulative graph convolutional network to integrate the short-term multi-modal cues to generate frame-level graphs and aggregate all frame-level subgraphs along the temporal trajectory to construct a global video-level social graph with various social relationships among multiple characters. Hu et al. [18] integrated automatic speech recognition, natural language understanding, face recognition and face clustering, and extracted multi-modal video relationships.

2.3 Scene Graph Generation

Another related topic is the generation of scene graphs, which is widely studied in the computer vision field to describe the spatial and structural relationships between objects. In fact, the concept of using graph-based context to improve scene understanding has been explored by many studies in recent decades [37, 38]. For example, Johnson et al. [39] were the first to introduce the problem of modeling objects and their relationships using scene graphs, which aimed to simultaneously detect objects and their pairwise relationships. Zellers et al. [40] proposed capturing higher-order repeated structures of scene graphs for better performance. Similarly, Yang et al. [41] developed an attention-aware GCN framework to update node and relationship representations by propagating context between nodes in candidate scene graphs, and RNNs were used by Xu et al. [42] to jointly refine object and relationship features in an iterative way to construct the scene graphs. Wu et al. [24] introduced a GCN to generate the social relationship graph for multiple characters in videos.

Inspired by the structural representation of scene graphs, we approach the video social relationship recognition task by generating a social graph.

3 METHOD

To investigate visual information for analyzing social relationships in videos, we propose a context-aware graph neural network composed of four modules: a person detection module, a graph proposal module, a feature extraction module and a relationship prediction module. The person detection module identifies faces and bodies in a given video and groups them together as characters. The graph proposal module builds a character graph that includes features of characters and their interactions. The feature extraction module produces features for faces, bodies and frames that capture person interactions. The relationship prediction module utilizes message passing to combine entity features and context features, thereby enhancing the representation capability of graph features for predicting social relationships.

3.1 Person Detection

We first need to find the main characters in the videos to detect relationships between people. Inspired by Ref. [12], we detect both faces and bodies for further feature extraction. One keyframe per second is selected for each clip and faces in keyframes are detected using RetinaFace [43]. The faces are clustered via consensus-driven propagation [44] and each cluster is treated as a detected person. To filter out unimportant people, only the clusters with required number of faces (at least six faces in our experiments) are retained. Moreover, similar to Ref. [13], bodies appearing in keyframes are detected by Faster-RCNN [45], which is pre-trained on the MSCOCO dataset [46]. Although many effective pedestrian detection models have been proposed, these pre-trained models perform worse than the pre-trained Faster-RCNN on the MSCOCO dataset. For example, the state-of-the-art Cascade Mask-RCNN [47] pre-trained on the CrowdHuman [48] dataset achieves only 58.01% coverage rate of the pre-annotated faces, while Faster-RCNN pre-trained on the MSCOCO achieves 99.24% coverage rate. A possible explanation is that the pedestrian detetion models are usually pre-trained on the datasets with entire small bodies, but people appearing in the VidSoR dataset are mostly cut off and noticeable, which is more similar to the case in the MSCOCO dataset. Considering that the performance of the pre-trained Faster-RCNN is sufficient, we do not re-train pedestrian detection models but use Faster-RCNN for body detection.

Cross validation between bodies and faces detected is utilized for body detection filtering. If the bounding box of a detected body can cover more than 95% of a detected face, the detected body is retained and assigned to the face. Moreover, if all the faces of a person do not have assigned body, that person is omitted. If more than one body is assigned to the same face, we randomly select one of the assigned bodies. Based on face clustering and face-body cross validation, the person set in a video is extracted as $P = \{p_i | i = 1, 2, 3, \cdots\}$. Each person is represented by a series of faces and bodies. Each person p_i is treated as a vertex in the SRG. Moreover the interaction series $C = \{c_t | t = 1, 2, 3, \cdots\}$ is constructed, where each c_t represents the collection of people that appear at time t.

3.2 Graph Proposal

Each SRG of a video consists of the vertices representing the people appearing in the video and the edges representing their social relationships. Considering that an SRG is sparser than a complete graph, we propose the potential edges between detected people to reduce the computational cost in relationship prediction.

The potential edges between the vertices in the SRG are proposed based on whether the corresponding two people have potential social relationship. We assume that two people with a social relationship should have interactions at some points in the video. Interactions are determined via the following method: co-occurrence of the two people within a sliding window of several adjacent keyframes. In our experiments, the size of the sliding window is set to 2, i.e., two people are considered to be interacting if they appear in the same keyframe or two adjacent keyframes. As illustrated in Fig. 2, in the graph proposal module, we add an edge between the corresponding two vertices for each pair of person appearing in the same sliding window.

Following the procedures mentioned above, an initial graph $G_0 = \{\{v_i\}, \{e_{mn}\}\}$ is generated, where v_i denotes the vertex representing person p_i and e_{mn} denotes an edge between vertex v_m and v_n in G_0.

3.3 Feature Extraction

Social relationship detection is related to both the personal attributes of people, such as age, sexuality and clothing, and the interactions between people. Thus, we extract face features and body features for VSRGG. There are many face instances and body instances in the keyframes for each person p_i in the video. If we extract features for all the face instances and body instances, the features would be large-scale and redundant. Hence, only the most representative instances for each person are selected for feature extraction. Since clear and front face can provide more effective features, face instance $\hat{f}_i$ with the largest area is selected. Bodies in videos, especially movies and dramas, are often incomplete due to occlusions or close-ups, thus body instance $\hat{b}_i$ with the maximum height-width ratio is selected to capture the whole body of each person.

Features are extracted based on $\hat{f}_i$ and $\hat{b}_i$, using a VGG16 model [49] pre-trained on the UTKFace dataset [50] and a ResNet50 model [51] pre-trained on the Market1501 dataset [52,53]. We represent the vertex v_i by fusing the features of $\hat{f}_i$ and $\hat{b}_i$, with a two-layer multi-layer perceptron

$$\boldsymbol{v}_i^0 = \phi([\boldsymbol{f}_i^f \cdot \boldsymbol{f}_i^b]), \tag{1}$$

where $\boldsymbol{v}_i^0$ denotes the fused feature of vertex v_i in G_0, $\boldsymbol{f}_i^f$ and $\boldsymbol{f}_i^b$ denote the features of $\hat{f}_i$ and $\hat{b}_i$, respectively, $[\cdot]$ denotes vector concatenation, and ϕ is implemented with a 4096×2048 dimension fully connected layer followed by the ReLU Layer.

Features that describe the interactions between people are also extracted. When two people are represented by adjacent vertices in G_0, they co-occur in one or more sliding windows of adjacent keyframes. The keyframes within these sliding windows are collected as

$$\Theta_{ij} = \cup_t \{K_t, \cdots, K_{t+\Delta t} \mid p_i, p_j \in c_t \cup \cdots \cup c_{t+\Delta t}\}, \quad (2)$$

where Θ_{ij} denotes the set of keyframes of sliding windows containing person pairs p_i and p_j, K_t denotes a keyframe at time t, Δt denotes the width of sliding window, which is 2 in our experiments, $p_i, p_j \in c_t \cup \cdots \cup c_{t+\Delta t}$ denotes p_i and p_j appearing in the sliding window beginning at t. The number of keyframes in Θ_{ij} is adjusted to a constant, which is 16 in the experiments, by uniformly sampling or oversampling existing frames. Based on the adjusted Θ_{ij}, a 3D-ResNet50 [54] pretrained on the activity network [55] is applied to extract the feature $\boldsymbol{e}_{ij}^0$ representing the edge e_{ij} in G_0.

3.4 Relationship Prediction

Context information is leveraged for relationship prediction. Inspired by iterative message passing for scene graph generation [42], a complete graph message passing mechanism is introduced, in which the vertex feature and edge feature are iteratively updated with messages from adjacent vertices and edges.

Let v_i denote vertex i and its corresponding feature is $\boldsymbol{v}_i$. e_{ij} denotes the edge between v_i and v_j, and its corresponding feature is $\boldsymbol{\varepsilon}^{ij}$. $E_i = \{e_{ik_1}, \cdots, e_{ik_n}\}$ is the set of all edges connected to v_i, and $V_{ij} = \{v_i, v_j\}$ is the set of the two vertices of the edge e_{ij}. We calculate the vertex message $\boldsymbol{\nu}_i^t$ and edge message $\boldsymbol{\xi}_{ij}^t$ as follows:

$$\boldsymbol{\nu}_i^t = \sum_{e_{mn} \in E_i} \sigma(\boldsymbol{\omega}_e[\boldsymbol{v}_i^t \cdot \boldsymbol{\varepsilon}_{mn}^t]) \cdot \boldsymbol{\varepsilon}_{mn}^t, \quad (3)$$

$$\boldsymbol{\xi}_{ij}^t = \sum_{v_k \in V_{ij}} \sigma(\boldsymbol{\omega}_v[\boldsymbol{v}_k^t \cdot \boldsymbol{\varepsilon}_{ij}^t]) \cdot \boldsymbol{v}_k^t, \quad (4)$$

where $\boldsymbol{\nu}_i^t$ and $\boldsymbol{\xi}_{ij}^t$ denote message for vertex v_i and edge e_{ij} at iteration t, respectively; e_{mn} denote an edge in E_i; v_k denote a vertex in V_{ij}; $\boldsymbol{\omega}_e$ and $\boldsymbol{\omega}_v$ are two weighted vectors to calculate the attention weights of the collected vertices or edges for the message, respectively; $\boldsymbol{v}_i^t$ and $\boldsymbol{v}_k^t$ denote features of vertex v_i and v_k and at iteration t, respectively; $\boldsymbol{\varepsilon}_{mn}^t$ and $\boldsymbol{\varepsilon}_{ij}^t$ denote features of edge e_{mn} and e_{ij} and at iteration t, respectively; $\sigma(\cdot)$ denotes sigmoid function for converting attention weights to range (0,1); t denotes the $t-$th iteration in message passing, which is initially set to 0.

We upgrade G_0 to the final SRG by updating the representation of its vertices and edges iteratively through two gate recurrent units [56], in which we use the message as the input and the representation as the hidden state. The representation features of the graph's vertices and edges are updated by Eqs. (**??**) and (**??**):

$$\boldsymbol{v}_i^{t+1} = \Psi_v(\boldsymbol{\nu}_i^t, \boldsymbol{v}_i^t), \quad (5)$$

$$\boldsymbol{\varepsilon}_{ij}^{t+1} = \Psi_e(\boldsymbol{\xi}_{ij}^t, \boldsymbol{\varepsilon}_{ij}^t), \quad (6)$$

where Ψ_v and Ψ_e are two gate recurrent units for updating edge representation and vertex representation, respectively.

Unlike in scene graph generation [42], the category label of each vertex cannot be used because all the vertices in a SRG represent people, leading to vertex representation confusion, i.e. all of the vertex representations in the graph are similar and the prediction result can easily overfit the most common undirected relationships such as friends and colleagues. One primary solution is to treat each person as an independent category; however, there are uncertain person numbers in each video and people in different videos are not the same. Another solution proposed in Ref. [4] is to only aggregate messages from vertices to edges and use edge representation for social relationship prediction. However, such a solution hampers the propagation of context information. Therefore, instead of constraining the representation of vertices, we fuse the final representation of each edge e_{ij} with the initial representations of its two vertices v_i and v_j, and conduct context-aware social relationship prediction on a fully connected layer using Eq.(7):

$$\boldsymbol{r}_{ij} = \phi([\boldsymbol{v}_i^0 \cdot \boldsymbol{\varepsilon}_{ij}^n \cdot \boldsymbol{v}_j^0]), \quad (7)$$

where $\boldsymbol{r}_{ij}$ denotes the predicted social relationship on edge e_{ij}, i.e., the social relationship between person p_i and p_j; $\boldsymbol{v}_i^0$ and $\boldsymbol{v}_j^0$ denote the initial representations of v_i and v_j, respectively; $\boldsymbol{\varepsilon}_{ij}^n$ denotes the final representation of e_{ij} and ϕ implemented with a fully connected layer.

4 Dataset

Early video social relationship analysis datasets such as SRIV [3] and ViSR [4], only provide social relationship tags for video clips. These datasets lack social relationship annotations at the person level. MovieGraphs [26] and the extension of the HLVU [36] in the deep video understanding challenge are datasets depicting human-centered situation, containing interactions between characters, their relationships and various visible and inferred properties such as the reasons behind certain interactions. The annotations in the MovieGraphs and HLVU datasets can support VSRGG evaluation, but are noisy and sparse. The scale of the HLVU is much smaller than that of the MovieGraphs. Xu et al. [16]

constructed a BiliBili dataset especially for VSRGG, but this dataset has not yet been published.

We construct a new VidSoR dataset for the VSRGG task. The dataset consists of videos collected from over 750 episodes of more than 300 different TV dramas, encompassing four categories: situation comedy, legal/medical, modern life and romance. We exclude categories such as cartoon and fiction, which are less relevant to people's daily lives. We identify and define eight relationship types based on domain theory [33]. Table 1 presents a detailed description of the relationship types.

The main process of constructing the VidSoR dataset contains four steps. Fig. 3 displays the key steps of annotating the relationships:

1) Each episode is cut into multiple clips ranging from 1 min to 4 min 51 s.
2) A filtering process is conducted to ensure that at least two individuals are included in each video clip and clips with opening and ending are discarded.
3) Faces are then detected by RetinaFace [43] with a sample rate of one frame per second, and two annotators are asked to manually categorize recognizable faces into different individuals as the pivot face set of each character. Unimportant characters in the video clips such as pedestrians in the background are discarded.
4) Two annotators are asked to annotate the results separately and compare them. A third annotator is asked to vote in case of conflicting opinions. If the third annotator is unsure of the annotation results, the disputed video clip is discarded.

The VidSoR dataset comprises 1798 valid video clips totaling 72 h in duration, with an average duration of 144 s. Additionally, there is an average of 3.50 person instances and 2.95 relationship instances in each video. Some annotation examples are shown in Fig. 4.

There are a total of 5313 relationship instances in the VidSoR dataset, and their distribution between the training set and test set is depicted in Fig. 5. The dataset has a long-tail distribution, with certain relationship types, such as "friends" and "colleagues", appearing more frequently than the others, posing a challenge to our task. Two constraints are ensured for the dataset split. First, the video clips from different TV drama categories are similarly distributed between the training and test sets. In addition, video clips from the same TV drama are restricted to appear in only one split, preventing the method from enhancing its performance by memorizing repeated characters within the same TV drama. Initially, we split our dataset into a training set and a test set at a ratio of 3:1 in accordance with the split ratio of ViSR. To meet these constraints, we make minor adjustments and end up with a training set size of 1347 episodes and the test set size of 451 episodes.

Table 1 Detailed relationship types.

Relationship	Description
Colleague	Co-workers, classmates
Couple	Husband-wife, boyfriend or girlfriend
Friend	Friends
Leader-subordinate	Boss-employee
Opponent	Enemy
Parent-offspring	Grandparent-grandchild, parent-child
Service	Waiter-customer
Sibling	Brothers, sisters

Table 2 compares the VidSoR and MovieGraphs datasets. VidSoR is derived from 750 episodes of more than 300 diverse TV dramas, while MovieGraphs is derived from 51 films. Although a greater number of video clips are included in MovieGraphs, VidSoR contains a greater number of valid clips that include social relationship annotations. Additionally, the average number of relationship instances per clip is higher in VidSoR. In terms of clip duration, the average clip length in MovieGraphs is 44 s, while in VidSoR it is 144 s. As a result, VidSoR is more complex and challenging than MovieGraphs. The social relationship annotation of VidSoR follows domain theory [33] and categorizes all descriptions into 8 classes. Although MovieGraphs encompass 107 types of social relationships, current methods reduce them to 5 [16], 8 [23, 25, 57] and 15 [5, 23, 57] types of social relationships..

Table 2 Comparison between the MovieGraphs [26] dataset and our VidSoR dataset.

Dataset	MovieGraphs	VidSoR
Source	51 movies	300+ TV dramas
Valid clips	1551	1798
Avg. duration	44 s	144 s
Relationship instances	2329	5313
Avg. relationship instances	1.50	2.95
Relationship types	106	8

5 Experiments

5.1 Evaluation Metrics

Following Refs. [2, 12, 13], we employ the mean average precision (mAP) and mean recall over all classes (mRecall) to evaluate different methods. In particular, the mAP is more important in the generation of social relationship graphs, where false predictions can severely affect scene understanding.

During testing, for each person, we select the face image with the feature of minimum distance to the center of its face collection. If the selected face has more than

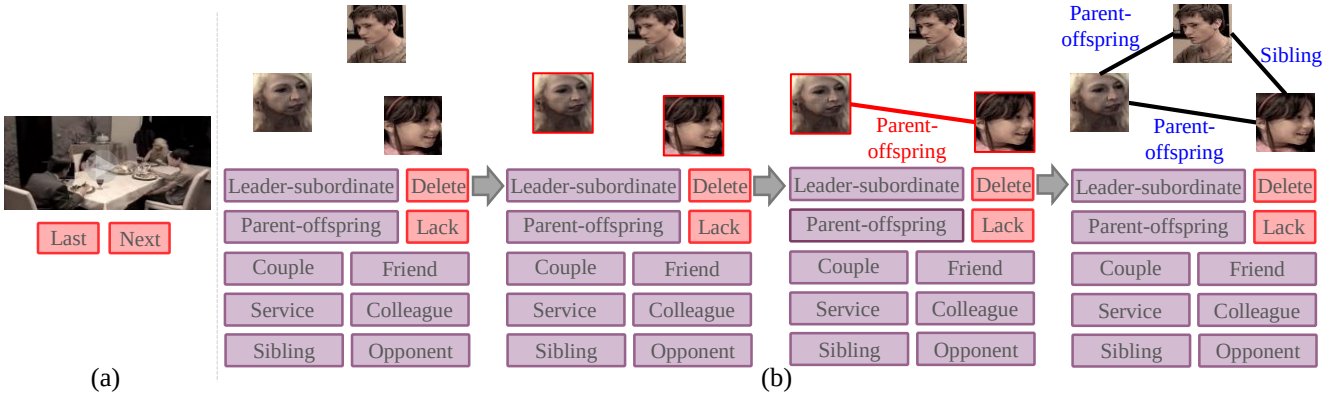

Fig. 3 An example of the annotation process (source from the DVU challenge [22]). (a) Watch the video clip to be annotated. (b) Confirm characters, choose a pair of characters, choose the relationship label of the pair and finally annotate all relationships.

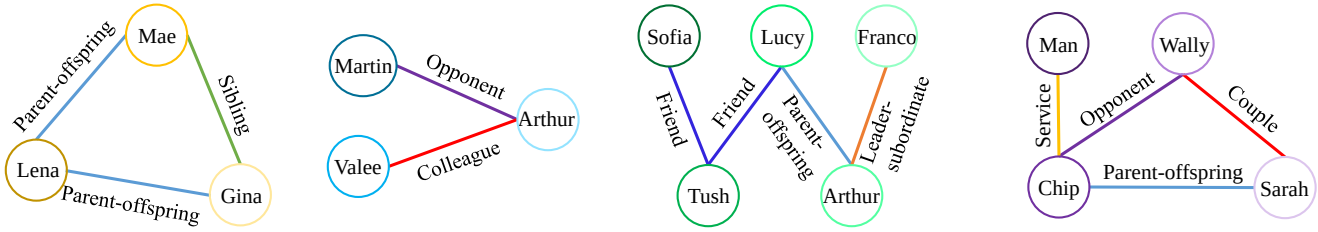

Fig. 4 Examples of annotations in the VidSoR dataset that we create.

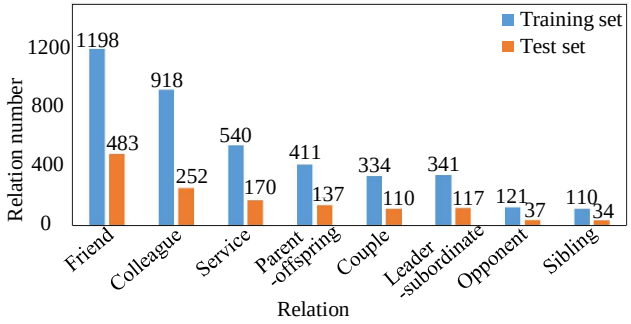

Fig. 5 Distribution of different relationship types in the training set and test set.

0.5 mean intersection-over-union (mIoU) with one of the annotated person's pivot faces, or has a cosine similarity to the pivot face of its nearest annotated person greater than 0.6, we consider this face the targeted person. If duplicated persons are detected for one person in the ground truth, we choose the face collection that has more faces of this person as the person to evaluate its relationships.

5.2 Implementation Details

In the VSRGG task, faces and bodies of persons need to be detected by RetinaFace and Faster-RCNN, respectively. Due to incorrect clusters or unfiltered pedestrians, there are large numbers of negative relationship samples in proposals generated by the graph proposal module. To address this issue, we retain the negative and positive samples at a ratio of 1:1 during training and only during training.

In the VSRGG* task, the input faces and bodies are those after manual annotation. Although in dataset construction, the original faces are also detected by RetinaFace, they are filtered and categorized by annotators. Thus, the bodies that are aligned to the faces after manual annotation are also manually filtered. Moreover, no sample strategy is used for VSRGG*.

In the CAGNet method, all the modules must be trained except the person detection module. The feature extraction submodels initialized with pre-trained parameters also need to be fine-tuned. In both tasks, our model is trained using the stochastic gradient descent optimizer, with the learning rate initialized to 1.0×10^{-5} and a batch size of 8. Our experiments are conducted using an RTX-3090 GPU with 24 GB of memory.

5.3 Comparison with State-of-the-Arts

We compare our CAGNet with different methods related to VSRGG, including methods for social relationship recognition in images: UnionCNN [1], PairCNN [13], First-glance [13], Dual-glance [13], graph reasoning model (GRM) [14] and social relationship graph generation network (SRG-GN) [15]; social relationship recognition at video level: multi-scale spatial-temporal

reasoning (MSTR) [4]; video social relationship detection at person level: learning interactions and relationships between characters (LIREC) [5]; methods for video scene graph generation: gated spatio-temporal fully-connected energy graph (GSTEG) [19], spatial-temporal transformer (STTran) [20] and target adaptive context aggregation network (TRACE) [21]. UnionCNN [1] is an image-based visual relationship detection method that combines visual cues with a single CNN and language priors from a word embedding model. PairCNN [13] contains two CNNs with shared weights to encode two persons separately. First-glance [13] only looks at pairs of people and makes a rough prediction directly. Dual-glance [13] takes another glance at region proposals, and aggregates the region-level predictions to refine the results. GRM [14] constructs a graph of persons and objects, and propagates node messages through the graph to fully explore the interaction of the persons with the contextual objects. SRG-GN [15] builds a social relationship graph and uses memory cells to update social relationship states using scenes and attribute contexts. MSRT [4] proposes a multi-scale spatial-temporal reasoning framework containing triple graphs to predict video-level social relationships. LIREC [5] proposes learning interaction and relationship between movie characters jointly with fused visual, audio and text features. GSTEG [19] constructs a fully-connected spatio-temporal graph to jointly learn spatial and temporal information about visual relationships in videos. STTran [20] proposes a spatial-temporal transformer framework to encode the spatial context within single frames and decode visual relationship representations with temporal dependencies across frames. TRACE [21] proposes a target adaptive context aggregation network following the detect-to-track paradigm with hierarchical relationship tree. Although there are several new methods for VSRGG recently [16–18, 23–25, 57], their codes have not been published and they are evaluated on the pre-processed MovieGraphs and the self-constructed datasets, which are not suitable for comparison.

We adapt all of these baselines to VSRGG. (1) Image-based methods: Because image social relationship recognition methods require related people to appear in the same image, all key frames with more than one person are selected and each person pair appearing in the same key frame is treated as the input. During the training process, each pair of people is taken and the ground truth is the annotation of that pair. The VSRGG task uses sampling strategies to our method to handle large numbers of negative samples. During the test, pair-wise relationship prediction is performed and predictions of the same person-pair are collected. After that, a voting mechanism is employed to aggregate predictions from different key frames on each person-pair. (2) Video scene graph generation methods: We transfer social relationship annotations to scene graph annotations and train the methods with person-level social relationship as the groundtruth for supervised learning. (3) Video-level Social relationship analysis methods: We adjust person-level social relationship annotations to video-level annotations as the ground truth for supervised learning by cutting video clips into shorter clips according to person co-occurrence, and each clip contains one person-pair with one social relationship. (4) Person-level Social relationship analysis methods: Since LIREC predicts interaction and social relationship together and VidSoR does not provide interaction annotations, we adapt the interaction prediction branch to relationship prediction with social relationship supervision.

As summarized in Table 3, CAGNet achieves balanced performance on both mAP and mRecall over two tasks. Despite significant exploration of person representation, context encoding and feature fusion in image-based methods, these methods do not excel in video social relationship analysis. Image-based social relationship detection methods only detect relationships when people are visible in the same frame. In contrast, video-based detection methods can identify relationships even when people appear in different video frames. Consequently, the performances of image-based methods are slightly inferior to ours.

Methods for scene graph generation primarily concentrate on detecting visual relationships between common objects in videos. However, these methods have been limited in their ability to analyze social relationships that require the reasoning about implicit information. Consequently, the metric values for most of these methods are relatively low, with the exception of the mAP of STTran.

The MSRT method, designed for video-level social relationship analysis, shows poor performance in person-level social relationship analysis. The LIREC method, which leverages multi-modal features for person-level video social relationship detection, obtains the highest mRecalls in both VSRGG and VSRGG*. However, our method using only visual features achieves 88.35% (7.05 vs. 7.98) and 93.61% (12.74 vs. 13.61) of LIREC's mRecalls in VSRGG and VSRGG*, respectively, while yielding significantly higher mAP values (10.04 vs. 2.07, 17.79 vs. 10.49).

In summary, our method can make accurate predictions with a comparatively high mRecall when using only visual features. The proposed graph module helps to detect more complete relationship graphs and thus improve the mRecall. The feature extraction module can

help to incorporate both static personal information and information about interactions. The message passing module can help to aggregate context information.

Table 3 Comparison of our method with state-of-the-art baselines in video social relationship graph generation (VSRGG) and weak video social relationship graph generation (VSRGG*). mAP and mRecall denote mean average precision and mean recall over all classes, respectively. The bold numbers refer to the best results and the underlined numbers refer to the second to the best results.

Method	VSRGG		VSRGG*	
	mAP	mRecall	mAP	mRecall
UnionCNN [1]	7.09	1.10	17.01	11.54
PairCNN [13]	7.59	1.08	16.98	11.10
First-glance [13]	7.28	1.49	16.93	11.45
Dual-glance [13]	8.22	1.45	15.92	11.68
GRM [14]	5.88	2.23	16.54	10.99
SRG-GN [15]	9.68	2.14	<u>17.49</u>	11.12
GSTEG [19]	8.40	2.22	9.56	4.15
STTran [20]	**10.76**	3.28	12.75	9.45
TRACE [21]	8.82	3.73	10.31	10.30
MSRT [4]	3.68	2.46	4.22	6.47
LIREC [5]	2.07	**7.98**	10.49	**13.61**
CAGNet(Ours)	<u>10.04</u>	<u>7.05</u>	**17.73**	<u>12.74</u>

5.4 Ablation Study

Backbone. VGG16 [49], ResNet50 [51]/ and 3D-ResNet50 [54] are used as feature extraction submodels in our method. We also evaluate the performance of a variant of CAGNet, which is denoted as "CAGNet-ViT" and exploits ViT-B [58] for face and body feature extraction and ViViT-B [59] for temporal feature extraction. As shown in Table 4, the performance of CAGNet is worse than that of CAGNet-ViT. It is assumed that the scale of the dataset is not large enough to train a method using the transformer as the backbone. Additionally, it is observed that GPU memory is sometimes exceeded during training of the variant. Thus, CAGNet based on VGG and ResNet can achieve better performance.

Table 4 Evaluation of our method using different backbones in VSRGG*. CAGNet represents the context-aware graph neural network using VGG [49] and ResNet [51]/ as the backbone and CAGNet-ViT represents the context-aware graph neural network using transformer as the backbone.

Backbone	mAP	mRecall
CAGNet-ViT	13.14	12.49
CAGNet	**17.73**	**12.74**

Body and facial features. In our method, both body and facial features are extracted. To validate the influence of body and facial features to proposed CAGNet, the performance of CAGNet trained with only facial feature (Face), only body feature (Body) and both (Body+Face) is presented. As shown in Table 5, CAGNet with both facial and body features has the best performance on all metrics. The results show that both facial and body features are important. Furthermore, the results also illustrate that face features are more effective. On one hand, face implies a lot of information, such as gender, age and emotion of the character, while body does not focus on such fine-grained information; on the other hand, bodies in videos such as movies and dramas are often incomplete because of occlusion or close-ups, which hardly hurts the expression ability of body features.

Table 5 Evaluation of our method using different features in VSRGG*.

Feature	mAP	mRecall
Face	14.05	11.58
Body	13.14	6.48
Body + Face	**17.73**	**12.74**

Sliding window width. The sliding window width is a key factor for identifying potential SRGs. An increase in the sliding window width has not only the potential to improve the completeness of the relationship graph, but also results in more false positive relationship proposals. Hence, we calculate the F-score to evaluate the performance:

$$F\text{-score} = \frac{(1+\beta^2) \times \text{mAP} \times \text{mRecall}}{\beta^2 \times \text{mAP} + \text{mRecall}}, \tag{8}$$

where the β^2 takes a value of 0.3 to increase the weight of the mAP. It is worth noting that the increase in completeness is limited and will not continue to increase as the sliding window width increases. We evaluate our method on VSRGG* with width the ranging from 1 to 5, and the results are shown in Table 6. When the sliding window width is 2, the mAP and F-score reach the maximum values.

Table 6 Evaluation of our method with different sliding window widths in VSRGG*. "1-5" indicate the width of a sliding window. F-score is used to evaluate the balance between mAP and mRecall.

Metric	1	2	3	4	5
mAP	12.76	**17.73**	15.14	12.88	14.43
mRecall	7.94	12.74	15.07	13.76	**17.46**
F-score	11.19	**16.26**	15.12	13.07	15.03

Combinations of representations. In our method, the edge representations are used after message passing, and the vertex representations are used before

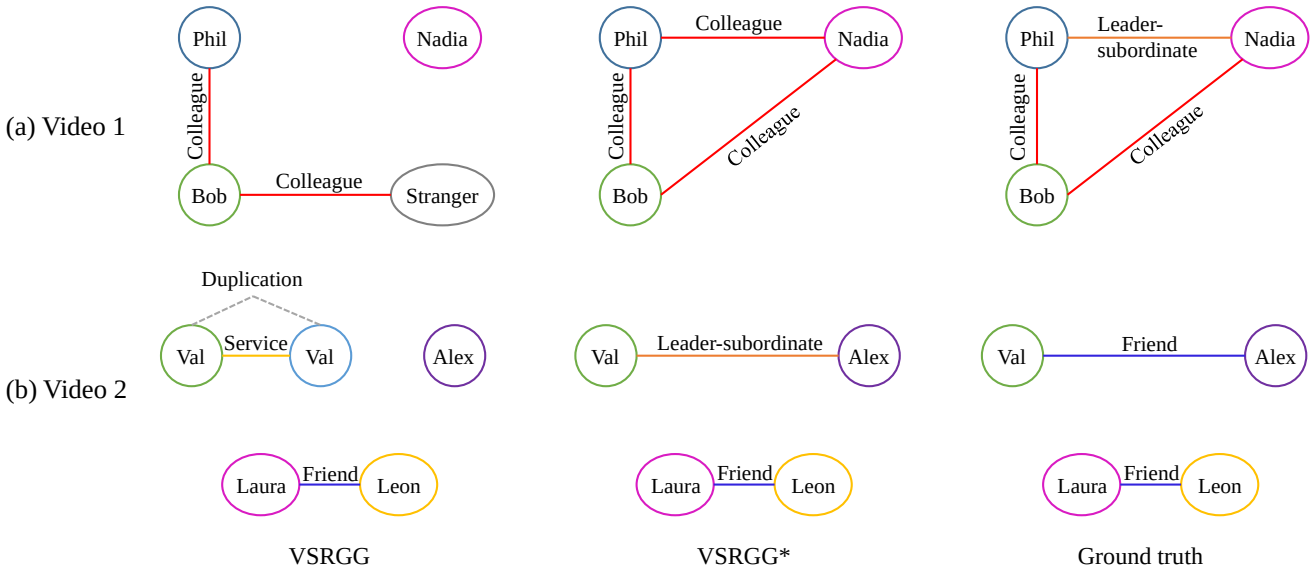

Fig. 6 Qualitative results of our method. (a) and (b) represent two different videos, respectively. Three columns are the results on video social relationship graph generation (VSRGG), weak video social relationship graph generation (VSRGG*) and ground truth of the two videos.

message passing to capture context information and incorporate distinguishable personal features. To study the influence of different combinations of vertex and edge representations, a set of experiments that include $E_0, E_n, V_0+E_0, V_n+E_n, V_n+E_0, V_n+E_0$ are conducted, where E_0 and E_n mean edge representation before and after message passing, and V_0 and V_n mean vertex representation before and after message passing. We also calculate the F-score to evaluate the performance.

As displayed in Table 7, the proposed method outperforms $V_0 + E_0$ on mAP and F-score, demonstrating the effectiveness of message passing for incorporating context information. Moreover, it outperforms $V_n + E_n$ on mAP and F-score, which shows that distinguishable vertex representation is helpful for making balanced prediction for all categories of relationships.

Table 7 Evaluation of our method using different representation combinations in VSRGG*. V_0 and V_n represent the use of vetex features before and after message passing, respectively. E_0 and E_n represent the use of edge features before and after message passing, respectively.

Combination	mAP	mRecall	F-score
E_0	16.55	15.11	16.19
E_n	15.13	14.67	15.02
$V_0 + E_0$	16.13	**15.51**	15.98
$V_n + E_n$	15.76	13.90	15.28
$V_n + E_0$	13.93	15.27	14.21
$V_0 + E_n$	**17.73**	12.74	**16.26**

5.5 Qualitative Results

The qualitative results are shown in Fig. 6, and the performance of our method on VSRGG is worse than that on VSRGG*. In particular, the performance of mRecall on VSRGG is even less than half that of the mRecall on VSRGG*. There are two possible reasons and we use qualitative examples to better illustrate the performance gap. One reason is that in the VSRGG task, there are false positives. As presented in Fig. 6, there are duplicate detections of the same person or some people as missing. The other reason is that even for the correctly detected person, there are some incorrect faces in the face cluster of each person, which leads to false predictions between correct person pairs. As shown in Fig. 6, the prediction of VSRGG* is better than that of VSRGG for correctly detected person pairs.

6 Conclusions

In this paper, we propose a novel CAGNet method for the person-level video social relationship analysis task VSRGG, which requires the construction of a social relationship graph containing social relationships between people appearing in a given video. CAGNet consists of person detection, graph proposal, feature extraction and relationship prediction to detect relationships in videos. Furthermore, a more complex and challenging dataset VidSoR is constructed for VSRGG evaluation. The dataset consists of 6276 person instances and 5313 relationship instances. The CAGNet method is compared with several state-of-the-art baselines on the VidSoR dataset, and achieves comparatively satisfactory performance.

Acknowledgements
Ao Zhang and Lusha Chen provided preliminary work and technical support for this project.

Funding
This work was supported by the National Natural Science Foundation of China (No. 62072232), the Fundamental Research Funds for the Central Universities (No. 021714380026) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Availability of data and materials
The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Abbreviations
CAGNet, context-aware graph neural network; SRG, social relationship graph; SRIV, social relationship in video; ViSR, video social relationship; VSRGG, video social relationship graph generation.

Declarations

Competing interests
The authors have no relevant financial or non-financial interests to disclose.

Author contributions
Jia Bei is the leader of the project and has set up the entire framework for the review. All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Fan Yu, Yaqun Fang, Zhixiang Zhao and Jia Bei. The first draft of the manuscript was written by Fan Yu, Yaqun Fang and Jia Bei. Zhixiang Zhao, Tongwei Ren and Gangshan Wu commented on previous versions of the manuscript. All the authors have read and approved the final version of the manuscript.

References

1. Cewu Lu, Ranjay Krishna, Michael Bernstein, & FeiFei Li. Visual relationship detection with language priors. In *B. Leibe, J. Matas, N. Sebe, et al. (Eds.), Proceedings of the 14th European Conference on Computer Vision*, pages 852–869, Cham: Springer, 2016.
2. Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, & Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, Piscataway: IEEE, 2020.
3. Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, & Huadong Ma. Multi-stream fusion model for social relation recognition from videos. In *K. Schoeffmann, T. H. Chalidabhongse, C.-W. Ngo, et al. (Eds.), Proceedings of the 24th International Conference on Multimedia Modeling*, pages 355–368, Cham: Springer, 2018.
4. Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, et al. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3566–3574, Piscataway: IEEE, 2019.
5. Anna Kukleva, Makarand Tapaswi, & Ivan Laptev. Learning interactions and relationships between movie characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Piscataway: IEEE, 2020.
6. Zhanpeng Zhang, Ping Luo, Chen-Change Loy, & Xiaoou Tang. Learning social relation traits from face images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3631–3639, Piscataway: IEEE, 2015.
7. Wentian Zhao, Xinxiao Wu, & Xiaoxun Zhang. MemCap: Memorizing style knowledge for image captioning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, number 07, pages 12984–12992, Palo Alto: AAAI Press, 2020.
8. Shaoxiang Chen & Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, number 01, pages 8191–8198, Palo Alto: AAAI Press, 2019.
9. Sanket Shah, Anand Mishra, Naganand Yadati, & Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, number 01, pages 8876–8884, Palo Alto: AAAI Press, 2019.
10. Wenya Guo, Ying Zhang, Jufeng Yang, & Xiaojie Yuan. Re-attention for visual question answering. 30:6730–6743, 2021.
11. Le Wu, Peijie Sun, Yanjie Fu, Richang Hong, Xiting Wang, & Meng Wang. A neural influence diffusion model for social recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 235–244, New York: ACM, 2019.
12. Qianru Sun, Bernt Schiele, & Mario Fritz. A domain based approach to social relation recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3481–3490, Piscataway: IEEE, 2017.
13. Junnan Li, Yongkang Wong, Qi Zhao, & Mohan S Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2650–2659, Piscataway: IEEE, 2017.
14. Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, & Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *J. Lang (Ed), Proceedings of the 27th International Joint Conference on Artificial Intelligence*, Palo Alto: AAAI Press, 2018.

15. Arushi Goel, Keng Teck Ma, & Cheston Tan. An end-to-end network for generating social relationship graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11195, Piscataway: IEEE, 2019.
16. Tong Xu, Peilun Zhou, Linkang Hu, Xiangnan He, Yao Hu, & Enhong Chen. Socializing the videos: A multimodal approach for social relation recognition. *ACM Transactions on Multimedia Computing, Communications, & Applications*, 17(1):1–23, 2021.
17. Yibo Hu, Chenyu Cao, Fangtao Li, Chenghao Yan, Jinsheng Qi, & Bin Wu. Overall-distinctive gcn for social relation recognition on videos. In *DT. Dang-Nguyen, et al. (Eds.), Proceedings of the 29th Proceedings of the International Conference on Multimedia Modeling*, pages 57–68, Cham: Springer, 2023.
18. Yibo Hu, Chenghao Yan, Chenyu Cao, Haorui Wang, & Bin Wu. Social relation graph generation on untrimmed video. In *DT. Dang-Nguyen, et al. (Eds.), Proceedings of the 29th International Conference on Multimedia Modeling*, pages 739–744, Cham: Springer, 2023.
19. Yao-Hung Hubert Tsai, Santosh Divvala, Louis-Philippe Morency, Ruslan Salakhutdinov, & Ali Farhadi. Video relationship reasoning using gated spatio-temporal energy graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Piscataway: IEEE, 2019.
20. Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, & Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16372–16382, Piscataway: IEEE, 2021.
21. Yao Teng, Limin Wang, Zhifeng Li, & Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13688–13697, Piscataway: IEEE, 2021.
22. Keith Curtis, George Awad, Shahzad Rajput, & Ian Soboroff. Hlvu: A new challenge to test deep understanding of movies the way humans do, Retrieved June 19, 2024, from https://www-nlpir.nist.gov/projects/trecvid/dvu/dvu.development.dataset/.
23. Chenyu Cao, Chenghao Yan, Fangtao Li, Zihe Liu, Zheng Wang, & Bin Wu. Recognizing characters and relationships from videos via spatial-temporal and multimodal cues. In *Proceedings of the IEEE International Conference on Big Knowledge*, pages 174–181, Piscataway: IEEE, 2021.
24. Shiwei Wu, Joya Chen, Tong Xu, Liyi Chen, Lingfei Wu, Yao Hu, et al. Linking the characters: Video-oriented social graph generation via hierarchical-cumulative gcn. In *Proceedings of the ACM International Conference on Multimedia*, pages 4716–4724, New York: ACM, 2021.
25. Yiyang Teng, Chenguang Song, & Bin Wu. Learning social relationship from videos via pre-trained multimodal transformer. *IEEE Signal Processing Letters*, 29:1377–1381, 2022.
26. Paul Vicol, Makarand Tapaswi, Lluis Castrejon, & Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8581–8590, Piscataway: IEEE, 2018.
27. Gang Wang, &rew Gallagher, Jiebo Luo, & David Forsyth. Seeing people in social context: Recognizing people and social relationships. In *K. Daniilidis, P. Maragos, & N. Paragios (Eds.), Proceedings of the 11th European Conference on Computer Vision*, pages 169–182, Piscataway: IEEE, 2010.
28. Hamdi Dibeklioglu, Albert Ali Salah, & Theo Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1497–1504, Piscataway: IEEE, 2013.
29. Ruogu Fang, Kevin D Tang, Noah Snavely, & Tsuhan Chen. Towards computational models of kinship verification. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1577–1580, Piscataway: IEEE, 2010.
30. Siyu Xia, Ming Shao, Jiebo Luo, & Yun Fu. Understanding kin relationships in a photo. *IEEE Transactions on Multimedia*, 14(4):1046–1056, 2012.
31. Yuanhao Guo, Hamdi Dibeklioglu, & Laurens Van der Maaten. Graph-based kinship recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 4287–4292, Piscataway: IEEE, 2014.
32. Donald J Kiesler. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90(3):185, 1983.
33. Daphne Blunt Bugental. Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, 126(2):187, 2000.
34. Alan P Fiske. The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99(4):689, 1992.
35. Meng Zhang, Xinchen Liu, Wu Liu, Anfu Zhou, Huadong Ma, & Tao Mei. Multi-granularity reasoning for social relation recognition from images. In *IEEE International Conference on Multimedia and Expo*, pages 1618–1623, Piscataway: IEEE, 2019.
36. Keith Curtis, George Awad, Shahzad Rajput, & Ian Soboroff. Hlvu: A new challenge to test deep understanding of movies the way humans do. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 355–361, New York: ACM, 2020.
37. Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, & Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4613–4623, Piscataway: IEEE, 2019.
38. An-An Liu, Yu-Ting Su, Wei-Zhi Nie, & Mohan Kankanhalli. Hierarchical clustering multi-task learning for joint human action grouping and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):102–114, 2016.
39. Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, et al. Image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3668–3678, Piscataway: IEEE, 2015.
40. Rowan Zellers, Mark Yatskar, Sam Thomson, & Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, Piscataway: IEEE, 2018.
41. Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, & Devi Parikh. Graph R-CNN for scene graph generation. In *V. Ferrari, M. Hebert, C. Sminchisescu, et al. (Eds.), Proceedings of the 15th European Conference on Computer Vision*, pages 670–685, 2018.
42. Danfei Xu, Yuke Zhu, Christopher B Choy, & FeiFei Li. Scene graph generation by iterative message passing. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 5410–5419, Piscataway: IEEE, 2017.
43. Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, & Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, Piscataway: IEEE, 2020.
44. Xiaohang Zhan, Ziwei Liu, Junjie Yan, Dahua Lin, & Chen Change Loy. Consensus-driven propagation in massive unlabeled data for face recognition. In *V. Ferrari, M. Hebert, C. Sminchisescu, et al. (Eds.), Proceedings of the 15th European Conference on Computer Vision*, pages 568–583, Piscataway: IEEE, 2018.
45. Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Proceedings of the 29th International Conference on Neural Information Processing Systems*, pages 91–99, 2015.
46. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, et al. Microsoft coco: Common objects in context. In *D. Fleet, T. Pajdla, B. Schiele, et al. (Eds.), Proceedings of the 13th European Conference on Computer Vision*, pages 740–755, Cham: Springer, 2014.
47. Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, & Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337, Piscataway: IEEE, 2021.
48. Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, et al. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint. arXiv:1805.00123*, 2018.
49. Karen Simonyan & Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint. arXiv:1409.1556*, 2014.
50. Zhifei Zhang, Yang Song, & Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, Piscataway: IEEE, 2017.
51. Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, Piscataway: IEEE, 2016.
52. Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, & Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1116–1124, Piscataway: IEEE, 2015.
53. Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, & Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1487–1495, Piscataway: IEEE, 2019.
54. Kensho Hara, Hirokatsu Kataoka, & Yutaka Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, Piscataway: IEEE, 2018.
55. Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, & Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 961–970, Piscataway: IEEE, 2015.
56. Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, & Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *Proceedings of the Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.
57. Chenghao Yan, Zihe Liu, Fangtao Li, Chenyu Cao, Zheng Wang, & Bin Wu. Social relation analysis from videos via multi-entity reasoning. In *Proceedings of the International Conference on Multimedia Retrieval*, pages 358–366, New York: ACM, 2021.
58. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint. arXiv:2010.11929*, 2020.
59. Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, & Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, Piscataway: IEEE, 2021.