IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 2, 2021, accepted July 7, 2021, date of publication July 12, 2021, date of current version July 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096553

Gaze-Aware Graph Convolutional Network for Social Relation Recognition

XINGMING YANG[1,2,3], FEI XU[3], KEWEI WU[1,2,3], ZHAO XIE[1,2,3], (Member, IEEE), AND YONGXUAN SUN[3]

[1] Key Laboratory of Knowledge Engineering With Big Data, Hefei University of Technology, Ministry of Education, Hefei 230601, China
[2] Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology, Hefei 230601, China
[3] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China

Corresponding author: Kewei Wu (wu_kewei1984@163.com)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB1002203, in part by the Fundamental Research Funds for the Central Universities of China under Grant PA2020GDSK0059, in part by the National Nature Science Foundation of China under Grant 61503111 and Grant 61876058, and in part by the Anhui Natural Science Foundation under Grant 1808085MF168.

ABSTRACT Social relation, as the basic relation in our daily life, is vital for social action analysis. However, how to learn the social feature between people is still not tackled. In this work, we propose a gaze-aware graph convolutional network (GA-GCN) for social relation recognition, which targets discovering the context-aware social relation inference with gaze-aware attention. To predict the gaze direction, we apply a convolutional network trained with gaze direction loss. Then, we build a graph convolutional inference module, which is a two-stream graph inference with both gaze-aware attention and distance-aware attention. The attention can pick up relevant context objects for context-aware representation. We further introduce additional scene features and construct a multiple feature fusion module, which can adaptively learn social relation representation from both scene feature and context-aware feature. Extensive experiments on the PISC and the PIPA datasets demonstrate that our GA-GCN can find interesting contextual objects and achieves state-of-the-art performances.

INDEX TERMS Social relation recognition, graph convolutional network, gaze direction, gaze-aware attention, graph inference.

I. INTRODUCTION

Social relation recognition is to aware the intimate/ non-intimate relation between people, which is the basic social structure in our daily life. The social relation is vital for intelligent machines to help machines act appropriately. The relations can apply to social media platforms to alarm privacy risks and to analyze human actions. However, visual social relation recognition is still challenging because the social relation feature is potential and complex. The features include not only the attributes of people but also the cues of surrounding objects in the scene.

Early attempts describe the relation as the interactions between people. Lu *et al.* [1] use a visual module to describe the object and use the predicates model to learn the relation. Sun *et al.* [2] consider both body and head region to extract features of gender, age, clothes, and even expression. Wang *et al.* [3] design a selection module to exact features of the body and head. Li *et al.* [4] consider the basic attention as a first glance at the interesting object and design a dual-glance model to exploit social features. However, the above methods mainly consider the relation of person pair and do not explicitly consider the object graph in the scene.

Recent works extend the relation of person pair to that of person-object pair because objects can suggest the role of the individuals. For intimate relations, context information may be the beer for friends, the TV for family, and the flowers for couples. For non-intimate relations, the context cues include the goods shelf for commercial and the document for professional. Goel *et al.* [5] provide a Social Relationship Graph Generation Network as an explicit knowledge graph with both person and object. Zhang *et al.* [6] introduce pose joint to Person-Pose Graph and design a Multi-Granularity Reasoning method. They can learn the context feature from

The associate editor coordinating the review of this manuscript and approving it for publication was Felix Albu.

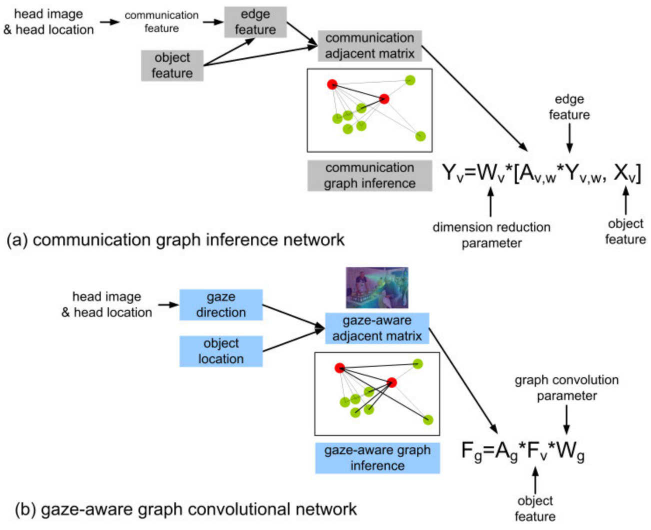
head image
& head location
communication feature
edge feature
object feature
communication adjacent matrix
communication graph inference
edge feature
$Y_v=W_v*[A_{v,w}*Y_{v,w}, X_v]$
dimension reduction parameter
object feature
(a) communication graph inference network
head image
& head location
gaze direction
object location
gaze-aware adjacent matrix
gaze-aware graph inference
graph convolution parameter
$F_g=A_g*F_v*W_g$
object feature
(b) gaze-aware graph convolutional network

FIGURE 1. Different graph structures of gaze-aware attention. In our work (b), we focus on a graph with gaze direction and object location. We optimize the graph convolution parameters to learn the gaze-aware feature.

surrounding objects with graph inference but still fail on the image when the features are noisy for social relation. This implies us to find attention to select objects for the graph inference.

Existing graph methods consider the interesting object selection with the gaze communication. As shown in Figure 1(a), gaze communication has been predicted from the object and edge features [25]. This communication encodes the location feature implicitly and limits the candidate objects. In our work, we are inspired by the selection in the gaze receptive field. Figure 1(b) shows our flexible gaze-aware attention, which first designs a Gaussian-based spatial distribution to capture the receptive field of gaze communication, and then calculates the attention with each object location.

To this end, we exploit a gaze-aware graph convolutional network (GA-GCN) for social relation recognition. Given the person-object graph, gaze attention is extracted using a Gaussian-based estimation. The attention learns the weights of the link between the nodes in the person-object graph, and the strong link indicates the attentive person-object pair. The attention composes the gaze-aware graph structure, which is vital to learn the dynamic of social relations in interactions.

The core modules of the GA-GCN are in Figure 2. (1) Inspired by the social graph structure in real life, the GA-GCN translates the objects into an object graph to learn the contextual social relation feature. (2) We employ gaze attention to explore the person-object selection in the graph structure. Different from the existing GCN with appearance-based self-attention, we propose a Gaussian-based model to describe the attention of gaze direction. To predict the gaze direction, we design a gaze direction loss to train a convolutional network fed with head image and head position. (3) We generate the graph structure with gaze-aware attention to learn the dynamics in the person-object pair. Unlike previous attention suffers from the weak link between person-object pair with long-distance, our gaze direction aware graph can find reliable candidate objects. (4) As complementary features, our GA-GCN joint considers the gaze-direction aware feature, distance-aware feature, and a scene feature. We design a multiple feature fusion module to adaptively learn social relation representation.

Our contributions are summarized as follows:

(1) We propose a gaze-aware graph convolutional network (GA-GCN) for social relation recognition, which targets discovering the context-aware social relation inference with gaze-aware attention.

(2) We build a graph convolutional inference module, which joins the gaze-aware attentive inference and the distance-aware attentive inference. We design the graph convolutional inference with residual connection to exploit the gradient for feature learning.

(3) We design a GazeNet to predict the gaze direction from the head image and head position. The gaze direction estimates the gaze-aware attention to pick up context objects for graph structure generation.

(4) We construct a multiple feature fusion module, which can adaptively learn social relation representation joint with both scene-level and object-level context features. Extensive experiments are conducted on the PISC and the PIPA datasets, and our GA-GCN achieves new state-of-the-art performances.

II. RELATED WORK

A. SOCIAL RELATION RECOGNITION

The social relation is the interaction between person pair. The key to the recognition is to learn the person features to describe social relations. Sun *et al.* [2] extract features from both body and head to represent the clothes and appearance. Guo *et al.* [7] consider face and scene features to predict social relation. Yan *et al.* [8] introduce segment to enhance the scene feature with semantics. Wang *et al.* [3] design a feature selection module to adaptively learn the social feature from head and body. Aimar *et al.* [9] apply social relation to a user wearing a photo-camera system. Some other works study the social relation from a video. Lv *et al.* [10] learn high-level semantic information of spatial, temporal, and audio for social interactions in videos. They further present an Attentive Sequences Recurrent Network model to fuse multiple visual features [11]. Fan textitet al. [25] exploit social relation as the atomic-level gaze communication with a spatial-temporal graph neural network. The above methods only consider the feature of person pair without the feature from contextual objects.

B. GRAPH CONVOLUTIONAL NETWORK

The person-object graph can provide the contextual feature for social relations. Wang *et al.* [13] consider the message propagation between objects can represent the

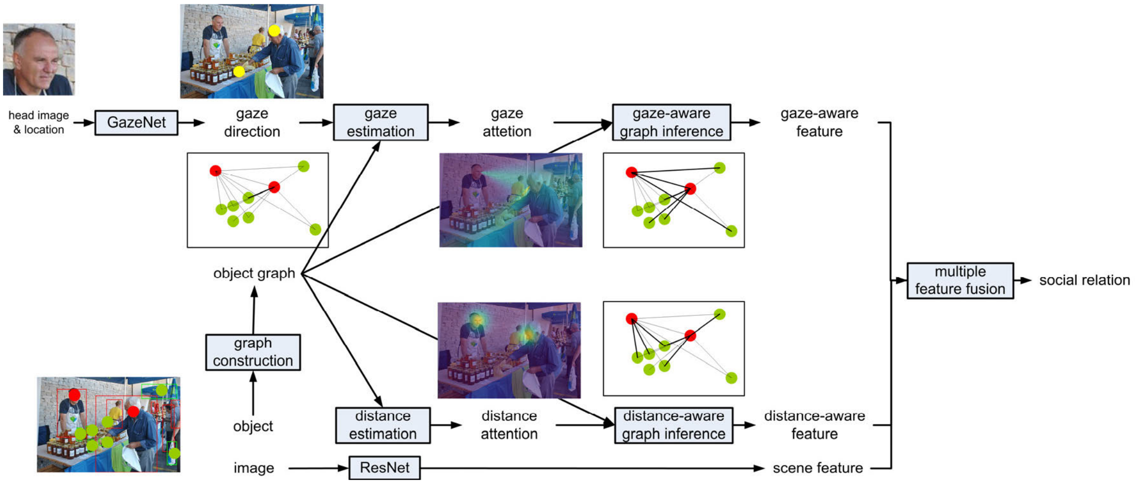

FIGURE 2. The overview of gaze-aware graph convolutional network for social relation recognition.

interaction between them. Zhang *et al.* [6] introduce pose joint into a person-pose graph to learn the features with Multi-Granularity Reasoning. Liu *et al.* [14] propose a Multi-scale Spatial-Temporal Reasoning (MSTR) framework to recognize social relations from videos. Li *et al.* [15] design a dual-glance model, which considers the first glance as the attention of region proposals and the second glace use enhanced feature for social relation recognition. They further use an adaptive focal loss to reduce the ambiguity in social relationship labels [4].

The object graph has also been studied for image captions. Lu *et al.* [1] provide a predicates model to describe the relationship between objects. Li *et al.* [16] design a multi-level scene graph to jointly predict object, phrase, and caption regions. Xu *et al.* [17] predict the scene graph with message passing Recurrent Neural Network. Zellers *et al.* [18] learn the subgraphs for object and scene. Herzig *et al.* [19] study the graph permutation invariance in a scene graph. Huang *et al.* [20] use social context graph for person attribute recognition. Vicol *et al.* [21] learn a movie graph to understand human relationships and interactions. Liao *et al.* [22] use social relation-based graph to recommend the group event.

C. GAZE ESTIMATION

Gaze direction can find interesting contextual objects for social relations. Gaze direction can be detected by head posture. Zhang *et al.* [23] estimate gaze direction with head image and group the social relation with it. Lian *et al.* [24] use head image and head position to predict gaze direction. Fan *et al.* [25] design a deep CNN to classify the gaze location with shifted grids labels. Besides the head posture, the scene image can predict the salience object for gaze direction. Recasens *et al.* [26] detect the gaze point with both head image and scene image. Varadarajan *et al.* [27] learn the direction with both head and body features. Fan *et al.* [28] use shared attentive object to predict gaze direction. Zhuang *et al.* [29] uses a recurrent structure to fuse individual gazes.

Recent works study gaze direction with eye features. Cheng *et al.* [30] build a network with two eyes asymmetry. Yu *et al.* [31] annotate gaze label with a gaze redirection network. Zhang *et al.* [32] predict gaze angle with face model and camera parameters. Zhang *et al.* [33] use ResNet to predict gaze direction Under Extreme Head Pose and Gaze Variation. Martinikorena *et al.* [34] estimate gaze with camera parameter and a geometrical compound model.

However, the above methods have not applied gaze direction to learn the context feature for social relation. And our work target at discovering the context-aware social relation inference with gaze-aware attention.

III. METHODOLOGY

Our work focuses on how to learn social relation representation in a spatial graph with the context-aware feature. Figure 2 shows our GA-GCN contains three branches. The first branch uses GazeNet to predict gaze direction and design a gaze-aware graph to learn the gaze-aware feature. The second branch learns distance-aware features with a distance-aware graph. The third branch learns a scene feature with ResNet. Finally, we design a multiple feature fusion module to adaptively learns social relation representation from both context-aware feature and scene feature.

A. GAZE DIRECTION NETWORK

The gaze direction can select the attentive objects for the social relation. The direction is the key to gaze-aware attention. Therefore, we design a gaze direction network (GazeNet). Given the head position h_i and its head image x_{h_i}, The GazeNet detect the gaze point of each person as

$$g_i = GazeNet(x_{h_i}, h_i) \tag{1}$$

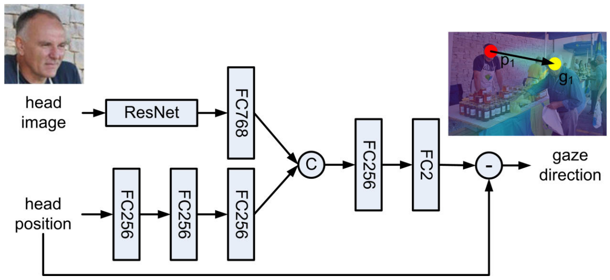

FIGURE 3. The architecture of the GazeNet to detect the gaze direction of a person.

We apply the network of [24]. As shown in figure 3, we first resize the head image to 224 × 224 and feed it into a ResNet-50 for feature extraction. Then transform it into an FC layer with a unit of 768 outputs. We use the coordinate of head position when the original image size is normalized to 1 × 1. We feed the head position into three FC layers to get the head position feature. Then we concatenate the head appearance feature with the head position feature and then transform it with two FC layers to predict the gaze point.

We do not directly use the gaze point, because the head feature can only show the gaze direction and not indicate the distance of the object. Therefore, we estimate the gaze direction as a normalized vector from the head position to the gaze point. To train the GazeNet, we use the gaze direction loss as

$$Loss_g(g_i, g_i^{gt}) = 1 - \frac{\overrightarrow{h_ig_i} \cdot \overrightarrow{h_ig_i^{gt}}}{\left|\overrightarrow{h_ig_i}\right| \cdot \left|\overrightarrow{h_ig_i^{gt}}\right|} \tag{2}$$

where g_i and g_i^{gt} are the predicted and ground truth gaze point, respectively. The $\overrightarrow{h_ig_i}$ and $\overrightarrow{h_ig_i^{gt}}$ are the predicted and ground-truth gaze direction, respectively.

B. GAZE AWARE GRAPH CONVOLUTIONAL NETWORK

1) GRAPH CONSTRUCTION

The graph convolutional network organizes various objects to learn the context-aware feature of social relation. Let G = {V, E} be the graph and its nodes and edges. The nodes are the detected objects, and the edges are the links between two objects. Because the gaze direction points from a person to an attentive object, we label the nodes with the person and other objects in the graph and divide the nodes set into person set and object set.

We use the RCNN [35] to extract the nodes in the graph. We consider RCNN can describe not only the feature of a person, such as the clothes, gender, but also the feature of objects to indicate the office scene or commercial scene. As shown in figure 4, we model two types of graphs with distance-aware links and gaze-aware links, respectively. The value of distance edge is the attention estimated with distance. The value of gaze edge is the attention estimated with gaze direction.

2) GAZE AWARE ATTENTION

Gaze direction can select the relevant objects with similar directions, and organize these objects to learn social relation representation. We use gaze-aware attention to describe the link between person and objects. Given a person with head position h_i and its gaze point g_i, we can get the gaze direction $\overrightarrow{h_ig_i}$. With additional object position v_j, we can get the object direction $\overrightarrow{h_iv_j}$. Then the angle between these two directions is

$$\theta(h_i, g_i, v_j) = \arccos \frac{\overrightarrow{h_ig_i} \cdot \overrightarrow{h_iv_j}}{\left|\overrightarrow{h_ig_i}\right| \cdot \left|\overrightarrow{h_iv_j}\right|} \tag{3}$$

We use a Gaussian distribution to estimate the gaze-aware attention from the person as

$$\begin{aligned} \alpha_g(p_i, v_j) &= \mathcal{N}(\theta(h_i, g_i, v_j), 0, \sigma_g) \\ &= \frac{1}{\sqrt{2\pi}\sigma_g} \exp\{-\frac{\theta^2(h_i, g_i, v_j)}{2\sigma_g^2}\} \end{aligned} \tag{4}$$

were p_i is the person i, v_j is the node j, $\mathcal{N}(.)$ is the Gaussian distribution, σ_g is the parameter of gaze-aware Gaussian distribution. This suggests that the object with the large angle has low attention. We further normalize the attention with various objects except the person. We define the attention of the diagonal element of the person as 1.

To estimate the attention from an object o_i, we consider only the diagonal element of the object is 1 and the rest is 0. This is because an object does not have gaze.

$$\alpha_g(o_i, v_j) = \begin{cases} 1 & v_j = o_i \\ 0 & v_j \neq o_i \end{cases} \tag{5}$$

3) JOINT GAZE-AWARE AND DISTANCE-AWARE ATTENTION

Because the social distance can describe the social relation, we further introduce distance-aware attention. Given two object position v_i and v_j, we estimate the distance-aware attention as

$$\begin{aligned} \alpha_d(v_i, v_j) &= \mathcal{N}(d(v_i, v_j), 0, \sigma_d) \\ &= \frac{1}{\sqrt{2\pi}\sigma_d} \exp\{-\frac{\left|v_iv_j\right|^2}{2\sigma_d^2}\} \end{aligned} \tag{6}$$

where σ_d is the parameter of Gaussian distribution. This suggests that the object with a large distance has low attention. Then, the joint attention is the joint probability of the distance-aware attention and gaze-aware attention.

$$\alpha_{dg}(v_i, v_j) = \alpha_d(v_i, v_j) \cdot \alpha_g(v_i, v_j) \tag{7}$$

The attention selects interesting nodes and constructs the adjacent matrix for graph inference. The matrix is the size of C × C, where C is the number of nodes. figure 5 shows the adjacent matrix with three types of attention. Each attention finds different attentive objects. The distance-aware attention can select near objects. The gaze-aware attention can select

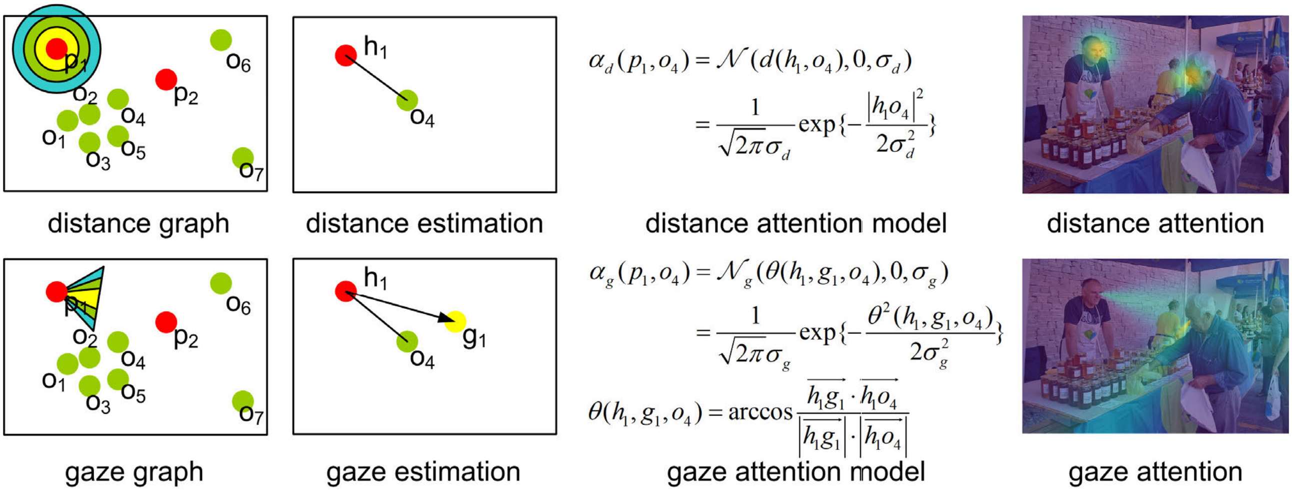
p1
o2
o1
o3
o4
o5
p2
o6
o7
distance graph
h1
o4
distance estimation
$\alpha_d(p_1,o_4)=\mathcal{N}(d(h_1,o_4),0,\sigma_d)$
$=\frac{1}{\sqrt{2\pi}\sigma_d}\exp\{-\frac{|h_1o_4|^2}{2\sigma_d^2}\}$
distance attention model
distance attention
p1
o2
o1
o3
o4
o5
p2
o6
o7
gaze graph
h1
g1
o4
gaze estimation
$\alpha_g(p_1,o_4)=\mathcal{N}_g(\theta(h_1,g_1,o_4),0,\sigma_g)$
$=\frac{1}{\sqrt{2\pi}\sigma_g}\exp\{-\frac{\theta^2(h_1,g_1,o_4)}{2\sigma_g^2}\}$
$\theta(h_1,g_1,o_4)=\arccos\frac{\overrightarrow{h_1g_1}\cdot\overrightarrow{h_1o_4}}{|\overrightarrow{h_1g_1}|\cdot|\overrightarrow{h_1o_4}|}$
gaze attention model
gaze attention

FIGURE 4. Distance attention estimation (top) and gaze attention estimation (bottom). We use the position of the head and object to estimate distance attention. We use the position of the head, the gaze, and the object to estimate the gaze position.

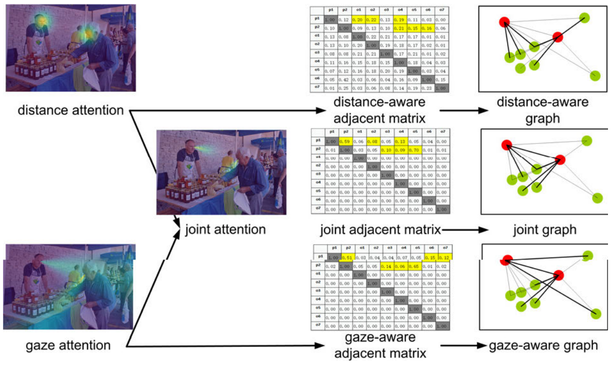

FIGURE 5. Distance-aware graph model and gaze-aware graph model. We show the attention distribution, the adjacent matrix with attention for graph inference, and the graph labeled with strong attention.

an object with a similar direction to the gaze direction. The joint attention reduces the attention from p_1 to o_6, because the object o_6 is far away. The adjacent matrix is used to group a person with relevant objects and infer person-object features with them.

C. GAZE STRUCTURAL GRAPH INFERENCE

Given the deep feature of the detected object from the RCNN F_v, the spatial graph convolution feature with gaze-aware attention is computed as:

$$F_g^{GCN}=A_gF_vW_g \tag{8}$$

where $A_g = \{\alpha_g(v_i,v_j)\}_{i,j}$ is the adjacent matrix with gaze-aware attention, W_g is the parameter of graph convolution. The inference fuses the attentive context feature to describe the person in the scene. Similarly, we also provide graph convolution feature with distance-aware attention and joint attention as $F_d^{GCN}=A_dF_vW_d$ and $F_{dg}^{GCN}=A_{dg}F_vW_{dg}$, where A_d and A_{dg} is the adjacent matrix with distance-aware attention and joint attention, respectively.

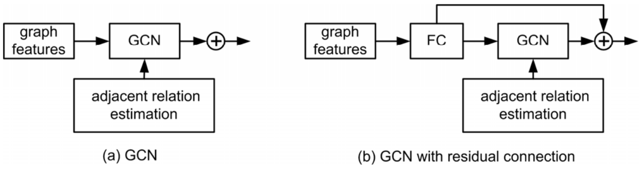
graph features
GCN
adjacent relation estimation
(a) GCN
graph features
FC
GCN
adjacent relation estimation
(b) GCN with residual connection

FIGURE 6. The module of GCN with residual connection.

As the complex social image generates many marginal samples in the feature space. This leads to degradation of training GCN. Therefore, we further exploit a GCN with residual connection to overcome the problem, as shown in figure 6. In the GCN with residual connection, we feed the graph features into an FC layer, which learns hidden features without feature shape transformed. Then we add a residual connection across the GCN and sum the GCN output and the GCN input. This benefits to exploit the gradients of the hidden features.

D. MULTIPLE FEATURE FUSION

Besides the above context-aware feature provided by graph convolution, our network adds a scene feature branch to extract the global feature for scene understanding, which is also important for social relations. This branch uses Resnet-50 as the backbone and takes the FC layer feature as the scene feature.

Each feature, including the gaze-aware feature, distance-aware feature, and scene feature, provides an independent interpretation for social relation. Inspired by the inception network, which can learn the local sparse structure of the convolutional network to approximate the dense inference, we further disclose a three-branch network with multiple feature fusion. We concatenate three features and use an FC layer with a unit of 1024 outputs to fuse them, and another

FC layer to predict the social relation.

$$\begin{cases} z = FC(F_{fusion}) \\ F_{fusion} = FC([F_g^{GCN}, F_d^{GCN}, F_s]) \end{cases} \tag{9}$$

where [.] is the concatenate function, F_g^{GCN} is the 4096-dimensional feature with gaze-aware attention, F_d^{GCN} is the 4096-dimensional feature with distance-aware attention, and F_s is the 2048-dimensional scene feature. Our network constructs different spatial graphs for each social relation of two people.

E. LOSS FUNCTION

As the majority of the loss is generated from the classified negatives and dominates the gradient, inspired by the focal loss, we use a modulating factor with a focusing parameter to balance the importance of positive/negative examples. Given the prediction z and the ground truth z^{gt}, the social relation loss is

$$Loss_z(z, z^{gt}) = \begin{cases} -(1-z)^{\gamma} \log z & z^{gt} = 1 \\ -z^{\gamma} \log(1-z) & z^{gt} = 0 \end{cases} \tag{10}$$

where γ is the focusing parameter. When z is approximate 1, the factor goes to 0 and this loss for well-classified example is down-weighted. As the parameter increases, the loss of correct examples gets reduce. In turn, this increases the importance of misclassified examples.

IV. EXPERIMENTS

A. DATASET

We conduct experiments on two widely used social relation recognition datasets: the PISC and the PIPA. The PISC dataset is a large-scale People in Social Context (PISC) dataset [15]. It has 22,670 images where the person pairs are annotated for domain recognition (i.e. Intimate, Not-Intimate, and No Relation) and relationship recognition (i.e. Friends, Family, Couple, Professional, Commercial, and No Relation). As in [4] for domain recognition, we randomly select 4000 images (15,497 samples) as the test set, 4000 images (14,536 samples) as the validation set and ~~use~~the remaining images (49,017 samples) as the training set. For relationship recognition, we sampled the test and validation split to have a balanced classification. Specially, we select 1250 images (250 per relation) with 3961 samples as the test set and 500 images (100 per relation) with 1505 samples as the validation set. The remaining images (55,400 samples) are used as the training set.

The PIPA dataset [2] has 16 fine-grained relationship categories. As in [5], it is divided into 6289 images (13,672 relationships) for training, 270 images (706 relationships) for validation, 2649 images (5075 relationships) for testing.

To train the gaze direction network, we annotate the head bounding box without the pre-detected model. We annotate the gaze point on the interesting object in the gaze direction. We use the same train/val/test split for gaze direction prediction as that for social relation recognition.

TABLE 1. The angular error of the GazeNet on two datasets.

Method	PISC	PIPA
GazeNet without head position	26.2°	29.3°
GazeNet with head position	22.3°	27.8°

B. IMPLEMENTATION DETAILS

We use PyTorch to implement the proposed method. Our network has a two-stage training. The first stage is for gaze direction prediction. We employ ResNet-50 to extract the head image feature. The network is initialized with the model pre-trained by ImageNet. Then, it is fed into the GazeNet and is trained with the gaze direction loss by Adam optimization. We set the learning rate as 0.001, while the fine-tuning model has a lower learning rate of 0.0001. We use a batch size of 32 and a momentum of 0.9 during training.

The second stage is for social relation recognition. We employ the pre-train RCNN [35] to initial the nodes of the graph and extract their features with the FC7 layer in RCNN. We estimate the distance-aware attention with objects' location and the gaze-aware attention with additional head and gaze direction location. We organize the attention between person-object pairs to form the adjacent matrix for graph inference. Then we get the context-aware feature for social relations. Besides, we employ ResNet-50 to extract the scene feature. In the multiple feature fusion module, we use two FC layers. The first layer has 1024 output units and the second layer predict the social relation. We train the GA-GCN network with the focal loss and the same optimization.

C. EVALUATION METRIC

We use angular error (Angular) between predicted gaze direction and ground truth direction to evaluate the gaze direction prediction. We use mean average precision (mAP) to evaluate the social relation recognition.

D. ABLATION STUDY OF GAZE DIRECTION PREDICTION

As the GazeNet takes the head image feature and the head position feature as input, we do the ablation study with/without the head position feature. Table 1. shows the Angular error of GazeNet on two datasets. As gaze direction is restrained with the head posture, given the available head image, the GazeNet without head position feature can indicate the gaze direction. When we introduce the head position, the GazeNet with head position features further reduces the angular error. This suggests that the head position can localize the field of the view with the assumption that the interesting objects are around the center of the scene. The PISC dataset shows a lower error than the PIPA dataset, which is mainly because it has more training images than the PIPA dataset.

E. ABLATION STUDY OF SOCIAL RELATION RECOGNITION

1) EFFECT OF GAUSSIAN PARAMETER IN THE GCN

We do the ablation study of social relation with the domain recognition on the PISC dataset. Our adjacent matrix of GCN

TABLE 2. The mAP of the domain recognition with different joint gaze/distance graph inference.

Method	$\sigma_d = w_{img}/2$	$\sigma_d = w_{img}/4$	$\sigma_d = w_{img}/8$
$\sigma_g = 15°$	77.2	78.1	77.6
$\sigma_g = 30°$	78.8	80.1	79.9
$\sigma_g = 60°$	78.3	79.7	79.1

has two main Gaussian parameters, which are the angular standard deviation σ_g and distance standard deviation σ_d. We initial the angular standard deviation with 15, 30, 60-degree angles, and the distance standard deviation with 1/2, 1/4, 1/8 of the image width w_{img}. Table 2 shows the mAP of the domain recognition on the PISC dataset with different joint gaze/distance graph inference. This model has a single graph with joint probability by early fusing the two attentions

We notice that (1) two-stream model with $\sigma_g = 30°$ and $\sigma_d = w_{img}/4$ gives the best performance in Table 2. (2) The model with small distance $\sigma_d = w_{img}/8$ has similar performance to that with $\sigma_d = w_{img}/4$, which suggest interesting objects mainly exist within a small distance. Although objects still exist within a large distance, the attention to the interesting object reduces in the model with a larger distance $\sigma_d = w_{img}/2$, which results in low performance. (3) Compared to the distance parameter, the model is more sensitive to the direction parameter. The model with a small direction angle is lower than that with a large direction angle. This suggests the attention within the small-angle field may fail to catch the interesting objects. Although the model with a larger angle field still contains interesting objects, it weakens the attention to them, and reduces the performance.

2) EFFECT OF THE VARIOUS GRAPH LINKS

Table 3 shows the mAP of the domain recognition with various graph links. The Gaussian parameter is $\sigma_g = 30°$ and $\sigma_d = w_{img}/4$. (1) We consider the distance-aware graph as the baseline model because the distance implies reachable objects. (2) The gaze-aware graph outperforms the distance aware graph because gaze direction can further select the interesting objects. (3) The model with joint gaze/distance select object considers both two attentions and outperforms that with single attention, which suggests the two attentions are complementary. The increment of the joint model to the distance-aware model is larger than that of the gaze-aware model, which suggests the gaze-aware attention provides more additional evidence than the distance-aware attention. (4) The best model in Table 3 is the two-stream model. It outperforms the joint model because it has two graphs to learn the local sparsity structure, while the joint model with a single graph cannot fully optimize the graph structure.

3) EFFECT OF THE MULTIPLE STREAM FEATURES

Table 4 shows the mAP of the domain recognition with multiple stream social relation features. We notice that (1) the model with scene feature performs lower than that with distance-aware graph, which suggests the feature without contextual information cannot effectively infer the social relation between persons. (2) When we add the scene feature, the performance is improved because the scene suggests a social event, such as a public place, office, market. The scene also implies the role of the actor in the event. The scene feature is a complement to the feature of gaze communication.

TABLE 3. The mAP of the domain recognition with various graph links.

Method	GCN	GCN with residual
distance	76.9	78.0
gaze	80.1	81.2
joint gaze/distance	80.9	81.9
two stream gaze/distance	83.2	84.4

TABLE 4. The mAP of the domain recognition with multiple stream social relation features.

Method	GCN	GCN with residual
scene feature	75.5	NA
distance+scene	79.7	80.8
gaze+secen	82.9	83.9
joint gaze/distance+scene	83.3	84.5
two stream gaze/distance+scene	86.6	87.7

4) EFFECT OF THE GCN WITH RESIDUAL CONNECTION

We do the ablative study with the GCN module and GCN with residual connection in Table 4 and Table 5. We notice that (1) the residual connection increases the performance on each graph structure. (2) The two-stream gaze+distance gives the highest improvement because of the complementary structure of the two graphs. (3) When we introduce scene feature, the performance rises because the scene feature improves the total loss, which also benefits graph-based social feature learning.

F. COMPARISON WITH STATE-OF-THE-ART METHODS

Our comparison methods contain two groups: (1) the model without graph inference, including Union-CNN [1], Two stream CNN [2], DSFS (Deep supervised feature selection) [3]. (2) the model with graph inference. including SRG-GN (Social Relationship Graph Generation Network) [5], MGR (Multi-Granularity Reasoning) [6], Dual-glance [4]. Table 5 shows the comparison of deep architectures of social relation methods. STGR [25] is the spatio-temporal graph reasoning model on video sequence, which cannot perform with image feature on the PISC and the PIPA dataset.

Table 6 shows the comparison with the state-of-the-art methods for the domain recognition on the PISC dataset. The data are cited from each method, and the data with * are cited

TABLE 5. Comparison of deep architectures of social relation methods.

Method	Human feature	Contextual feature	Feature extraction	Feature fusion	Loss function
annotated object					
Two stream CNN* [2]	attributive feature of head and body	NA	CaffeNet	SVM	attribute fine-tuned loss
DSFS [3]	attributive feature of head and body	NA	CaffeNet	feature selection from normalization	cross-entropy loss of relation, normalization term
SRG-GN [5]	attributive feature of body	scene feature, activity feature	attribute ConvNet, VGG-16	message passing in GRU	multi-task loss
object proposal with R-CNN					
Union-CNN [1]	visual feature of body, words vector	NA	VGG-16, word2vec	multiplication of probability	relation loss, variance loss of prediction, rank loss of occurrence frequency
MGR [6]	body feature, joint feature	contextual obejct feature with GCN	ResNet-101, deconvolution pose estimation network	weighted fusion	relation loss
Dual-glance [4]	location, body feature, person pair feature	contextual obejct feature	ResNet-101, VGG-16	attentive contextual fusion	adaptive focal loss of relation
STGR [25]	head feature, location	contextual obejct feature, scene feature, gaze-aware GCN, LSTM	ResNet-50	NA	loss of relation
Ours	body feature	contextual obejct feature with gaze-aware GCN and distance-aware GCN, scene feature, object selection	Faster R-CNN, ResNet-50, GazeNet	FC-based weighted fusion	focal loss of relation

TABLE 6. Comparison with the state-of-the-art methods for the domain recognition on the PISC dataset.

Method	mAP	Intimate	Non-intimate	No-relation
Union-CNN* [1]	75.2	81.5	75.3	NA
Two stream CNN* [2]	76.9	82.1	76.5	NA
Dual-glance [4]	85.8	85.8	83.1	NA
GA-GCN [Ours]	86.6	86.1	84.9	88.9
GA-GCN with residual [Ours]	87.7	87.3	86.0	89.8

TABLE 7. Comparison with the state-of-the-art methods for the relationship recognition on the PISC dataset.

Method	mAP	Friends	Family	Couple	Professional	Commercial	No-relation
Union-CNN* [1]	49.3	42.7	52.5	45	70.2	49.4	NA
Two stream CNN* [2]	54.9	58.1	58.5	47.3	72.7	52.3	NA
MGR [6]	64.4	64.6	67.8	60.5	76.8	34.7	70.4
Dual-glance [4]	65.2	60.6	64.9	54.7	82.2	58.0	NA
SRG-GN [5]	71.6	25.2	80	100	78.4	83.3	62.5
GA-GCN [Ours]	72.7	62.7	73.3	75.3	80.7	76.6	71.4
GA-GCN with residual [Ours]	73.6	63.1	73.5	78.3	82.7	76.8	71.8

from [4]. The Union-CNN [1] gets low performance because it only uses the feature of the object to analyze the interactions between pairs of objects. And Two stream CNN [2] apply both body and head region feature to describe gender, age, clothing. However, the above methods only consider object pair features without context information. Dual-glance [4] design a two-stage method. The first glance fixates at the person of interest and the second glance deploys an attention mechanism to exploit contextual cues. The graph in the dual-glance model is built on region proposals, which may be noisy for social relations. Our GA-GCN outperforms above methods because we design an explicit graph for person pair and their contextual objects. And our GA-GCN with residual connection further increases the performance because we exploit the gradients to learn effect hidden features.

Table 7 shows the comparison with the state-of-the-art methods for relationship recognition on the PISC dataset. We further discuss the model with explicit graph inference. The SRG-GN introduce context objects in Social Relationship Graph Generation Network [5], which use an explicit knowledge graph to represent human relation and attributes. The MGR introduces pose joint into a graph with Multi-Granularity Reasoning [6]. Our GA-GCN outperforms them because we apply social attention to estimate the adjacent matrix of the graph. Specially, we introduce both gaze-aware attention and distance-aware attention to learn context-aware features for social relations. By introducing an FC layer to learn hidden features, Our GA-GCN with residual connection increases mAP by 2.0 %, compared with the Dual-glance model [4].

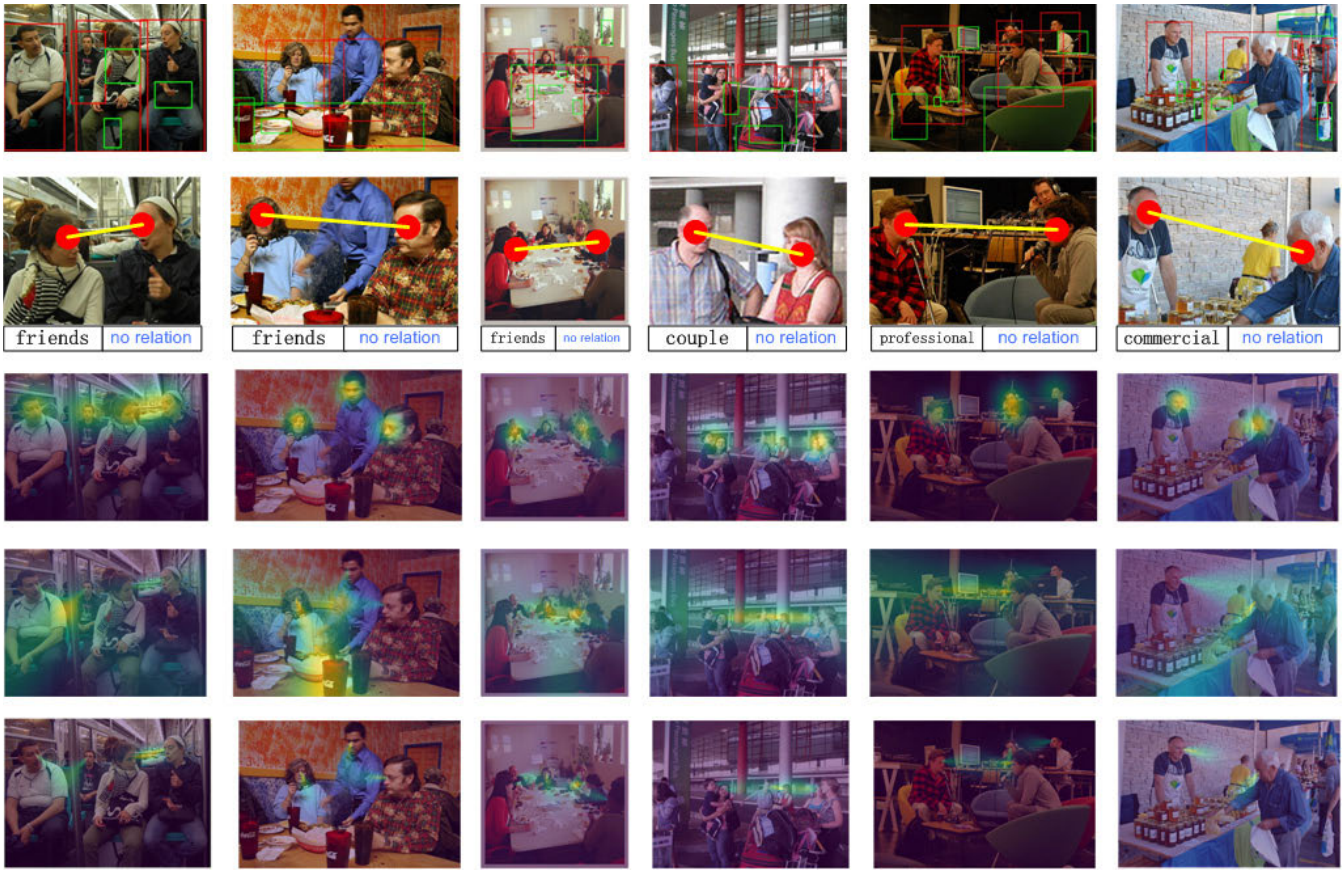

FIGURE 7. Correct prediction from our GA-GCN model (black label) while the distance-aware model fails (blue label). The first row shows the objects predicted by the RCNN. The second row shows the correct prediction of the person pair. The third/fourth/fifth row shows the distance-aware attention, gaze-aware attention, and joint attention, respectively.

TABLE 8. Comparison with the state-of-the-art methods on the PIPA dataset.

Method	mAP
SRG-GN [5]	53.6
Two stream CNN [2]	57.2
Dual-glance [4]	59.6
DSFS [3]	61.5
MGR [6]	64.4
GA-GCN [Ours]	65.5
GA-GCN with residual [Ours]	66.6

Table 8 shows the comparison with the state-of-the-art method on the PIPA dataset. Our GA-GCN outperforms the DSFS [3] because the DSFS only discusses the feature selection from body and head regions and does not use the context. Our method still outperforms other graph inference methods because of the adjacent matrix with attention estimation. We notice the SRG-GN [5] gets low performance, which probably because PIPA has 16 relationship categories and limited training images. Our gaze-aware attention can select the interesting objects and robust to 16 relationship categories on the PIPA dataset. Finally, Our GA-GCN with residual connection increases mAP by 2.2 %, compared with the MGR model [6] on the PIPA dataset.

G. VISUALIZATION

We further visualize the social prediction for relation recognition on the PISC dataset. We also show the contextual objects, distance-aware attention, gaze-aware attention, joint attention in figure 7. We notice that our GA-GCN can correct the prediction (black label) while the distance-aware model fails (blue label). Person pair with large distance may predict as no relation in the 2nd 3rd 5th and 6th column. In the 1st and 4th column, the person pair with small distance but they were different clothes can be fails predict as no-relation. Our gaze-aware attention can find the interesting person and imply the relation between them. Further, our GA-GCN combines scene features to distinguish the intimate (friends and couple) and non-intimate (professional and commercial) relation.

V. CONCLUSION

In this paper, we aim to discover the context-aware social relation inference with gaze direction and exploit a gaze-aware graph convolutional network (GA-GCN) for social relation recognition. Our GA-GCN is a two-stream graph, which joint both gaze direction and distance. We design a GazeNet fed with the head image and head position to predict the gaze direction. The direction can estimate the gaze-aware attention for graph inference. As the scene is also an important attribute for social relation, we add a global branch and design a multiple feature fusion module to adaptively learn social relation representation from both scene feature and context-aware feature. Extensive experiments are conducted on the PISC and the PIPA datasets. We do the ablation study of GazeNet for gaze direction prediction, and that of the GA-GCN for social relation recognition. We further visualize the gaze-aware attention to show that our attention can find the interesting objects and correct the prediction while the distance-aware model fails. Our method is limited to dynamic gaze communication. Therefore, we intend to embed our gaze-aware graph into temporal inference in future work.

REFERENCES

[1] C. Lu, R. Krishna, M. S. Bernstein, and F. Li, ''Visual relationship detection with language priors,'' in *Proc. Eur. Conf. Comput. Vis.*, vol. 1, 2016, pp. 852–869.

[2] Q. Sun, B. Schiele, and M. Fritz, ''A domain based approach to social relation recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 435–444.

[3] M. Wang, X. Du, X. Shu, X. Wang, and J. Tang, ''Deep supervised feature selection for social relationship recognition,'' *Pattern Recognit. Lett.*, vol. 138, pp. 410–416, Oct. 2020.

[4] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, ''Visual social relationship recognition,'' *Int. J. Comput. Vis.*, vol. 128, no. 6, pp. 1750–1764, Jun. 2020.

[5] A. Goel, K. T. Ma, and C. Tan, ''An end-to-end network for generating social relationship graphs,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11186–11195.

[6] M. Zhang, X. Liu, W. Liu, A. Zhou, H. Ma, and T. Mei, ''Multi-granularity reasoning for social relation recognition from images,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1618–1623.

[7] X. Guo, L. F. Polania, J. Garcia-Frias, and K. E. Barner, ''Social relationship recognition based on a hybrid deep neural network,'' in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–5.

[8] H. Yan and C. Song, ''Semantic three-stream network for social relation recognition,'' *Pattern Recognit. Lett.*, vol. 128, pp. 78–84, Dec. 2019.

[9] E. S. Aimar, P. Radeva, and M. Dimiccoli, ''Social relation recognition in egocentric photostreams,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3227–3231.

[10] J. Lv, W. Liu, L. Zhou, B. Wu, and H. Ma, ''Multi-stream fusion model for social relation recognition from videos,'' in *Proc. ICCV*, 2018, pp. 355–368.

[11] J. Lv, B. Wu, Y. Zhang, and Y. Xiao, ''Attentive sequences recurrent network for social relation recognition from video,'' *IEICE Trans. Inf. Syst.*, vol. E102.D, no. 12, pp. 2568–2576, Dec. 2019.

[12] A. K. A. Recasens, C. Vondrick, and A. Torralba, ''Where are they looking?'' in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 199–207.

[13] Z. Wang, T. Chen, J. Ren, W. Yu, H. Cheng, and L. Lin, ''Deep reasoning with knowledge graph for social relationship understanding,'' in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1021–1028.

[14] X. Liu, W. Liu, M. Zhang, J. Chen, L. Gao, C. Yan, and T. Mei, ''Social relation recognition from videos via multi-scale spatial-temporal reasoning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3566–3574.

[15] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, ''Dual-glance model for deciphering social relationships,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2669–2678.

[16] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, ''Scene graph generation from objects, phrases and region captions,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1270–1279.

[17] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, ''Scene graph generation by iterative message passing,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3097–3106.

[18] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, ''Neural motifs: Scene graph parsing with global context,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.

[19] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson, ''Mapping images to scene graphs with permutation-invariant structured prediction,'' in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7211–7221.

[20] Q. Huang, Y. Xiong, and D. Lin, ''Unifying identification and context learning for person recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2217–2225.

[21] P. Vicol, M. Tapaswi, L. Castrejón, and S. Fidler, ''MovieGraphs: Towards understanding human-centric situations from videos,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8581–8590.

[22] G. Liao and X. Deng, ''Leveraging social relationship-based graph attention model for group event recommendation,'' *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 8834450-1–14-8834450, Oct. 2020.

[23] L. Zhang and H. Hung, ''Beyond F-formations: Determining social involvement in free standing conversing groups from static images,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1086–1095.

[24] D. Lian, Z. Yu, and S. Gao, ''Believe it or not, we know what you are looking at!'' in *Proc. Asian Conf. Comput. Vis.*, vol. 3, 2018, pp. 35–50.

[25] L. Fan, W. Wang, S.-C. Zhu, X. Tang, and S. Huang, ''Understanding human gaze communication by spatio-temporal graph reasoning,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5723–5732.

[26] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba, ''Following gaze in video,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1444–1452.

[27] J. Varadarajan, R. Subramanian, S. R. Buló, N. Ahuja, O. Lanz, and E. Ricci, ''Joint estimation of human pose and conversational groups from social scenes,'' *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 410–429, Apr. 2018.

[28] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu, ''Inferring shared attention in social scene videos,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6460–6468.

[29] N. Zhuang, B. Ni, Y. Xu, X. Yang, W. Zhang, Z. Li, and W. Gao, ''MUGGLE: MUlti-stream group gaze learning and estimation,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3637–3650, Oct. 2020.

[30] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, ''Gaze estimation by exploring two-eye asymmetry,'' *IEEE Trans. Image Process.*, vol. 29, pp. 5259–5272, 2020.

[31] Y. Yu and J.-M. Odobez, ''Unsupervised representation learning for gaze estimation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7312–7322.

[32] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, ''MPIIGaze: Real-world dataset and deep appearance-based gaze estimation,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 1, pp. 162–175, Jan. 2019.

[33] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, ''ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation,'' in *Proc. Eur. Conf. Comput. Vis.*, vol. 5, 2020, pp. 365–381.

[34] I. Martinikorena, A. Larumbe-Bergera, M. Ariz, S. Porta, R. Cabeza, and A. Villanueva, ''Low cost gaze estimation: Knowledge-based solutions,'' *IEEE Trans. Image Process.*, vol. 29, pp. 2328–2343, 2020.

[35] S. Ren, K. He, R. Girshick, and J. Sun, ''Faster R-CNN: Towards real-time object detection with region proposal networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

XINGMING YANG received the Ph.D. degree from the University of Science and Technology of China, in 2005. He is currently an Associate Professor with the Hefei University of Technology. His main research interests include computer vision, intelligent multimedia systems, media processing, content analysis, and visual information systems.

FEI XU is currently pursuing the master's degree with the Hefei University of Technology. His research interests include computer vision, pattern recognition, and machine learning.

ZHAO XIE (Member, IEEE) received the Ph.D. degree in computer science from the Hefei University of Technology, in 2007. He is currently an Associate Research Fellow with the Hefei University of Technology. His research interests include computer vision, pattern recognition, and machine learning.

KEWEI WU received the Ph.D. degree in computer science from the Hefei University of Technology, in 2013. He is currently an Associate Research Fellow with the Hefei University of Technology. His research interests include computer vision, pattern recognition, and machine learning.

YONGXUAN SUN received the Ph.D. degree in computer science from the Hefei University of Technology, in 2013. He is currently a Lecturer with the Hefei University of Technology. His research interests include computer vision, pattern recognition, and machine learning.