Improving Visual Relationship Detection With Two-Stage Correlation Exploitation

Hao Zhou, Chongyang Zhang, *Member, IEEE*, Muming Zhao, Yan Luo, *Graduate Student Member, IEEE*, and Chuanping Hu

***Abstract*— Visual relationship detection, as a challenging task used to find and distinguish interactions between object-pairs in one image, has received much attention recently. In this work, we devise a unified visual relationship detection framework with two types of correlation exploitation to address the combination explosion problem in the object-pairs proposing stage and the non-exclusive label problem in the predicate recognition stage. In the object-pairs proposing stage, with the exploitation of relative location correlation between two objects in one pair, one location-embedded rating module (LRM) is developed to effectively select plausible proposals. In the predicate recognition stage, one label-correlation graph module (LGM) is introduced to measure the implicit semantic correlation among predicates; and then assign discrete distributed labels to predicates to improve the precision of top-n recall. Experiments on the two widely used VRD and VG datasets show that our proposed method outperforms current state-of-the-art methods.**

***Index Terms*— Visual relationship detection, graph neural network, label distribution.**

I. Introduction

THANKS to the development of deep learning, many deep CNN models have achieved good performances in computer vision tasks, including object detection [2], [3], classification [4], [5] and salient detection [6], [7]. However, one of the further goals in computer vision is image understanding: capturing the semantic information in one image. Generally, visual relationships can be briefly expressed as triplets $\langle sub - pred - ob \rangle$, where *sub*, *pred* and *ob* mean *subject*, *predicate* and *object*, respectively. Visual relationship detection attempts to distinguish interactions of object-pairs. By capturing semantic relation information in the real-world, it plays an essential role in a wide range of industrial application scenarios [8]–[10]. To be specific, Liu *et al*. [8] leverage instance-level relationships to assist in object detection. Peng *et al*. [9] exploit the relationship detection to learn more discriminative features for unsupervised cross-media retrieval. Qi *et al*. [10] devise a group relationship module to model players' interaction for better sports video captioning. Thus, visual relationship detection has received increasing attention recently.

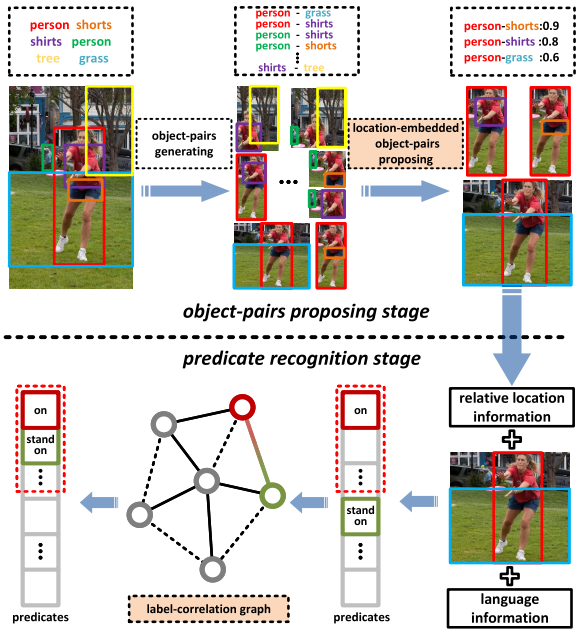

Fig. 1. The illustration of our proposed pipeline modified from [1]. In the object-pairs proposing stage, we use relative location to select plausible proposals effectively. In the predicate recognition stage, we use one graph to increase probabilities of all plausible *predicates*.

While there is much ongoing research in visual relationship detection works, most of them use a two-stage pipeline: the object-pairs proposing stage and the predicate recognition stage, as illustrated in Figure 1. Although quick strides have been made in this field, there still exist two main challenges: the combination explosion problem in the object-pairs proposing stage and the non-exclusive label problem in the predicate

Manuscript received April 21, 2020; revised September 3, 2020; accepted October 11, 2020. Date of publication October 21, 2020; date of current version July 2, 2021. This work was supported in part by the National Key Research and Development Program under Grant 2017YFB1002401, in part by NSFC under Grant 61971281, and in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 18DZ2270700 and Grant 18DZ1112300. This article was recommended by Associate Editor Y. Yang. *(Corresponding author: Chongyang Zhang.)*

Hao Zhou and Yan Luo are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China (e-mail: zhouhao_0039@sjtu.edu.cn; luoyan_bb@sjtu.edu.cn).

Chongyang Zhang is with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China, and also with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: sunny_zhang@sjtu.edu.cn).

Muming Zhao is with Robotics and Autonomous Systems, Data61 CSIRO, Brisbane, QLD 4069, Australia (e-mail: muming.zhao@data61.csiro.au).

Chuanping Hu is with the School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: cphu@vip.sina.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2020.3032650

recognition stage. In the object-pairs proposing stage, prior works [11], [12] follow the naive proposing method: $N(N-1)$ object-pairs are combined based on N detected objects. Thus, performances are heavily dependent on N. To cover more visual relationship triplets in one image, we tend to reserve as many as possible objects. As a result, the combinations will grow exponentially along with N increasing. To handle the combination explosion problem, one feasible proposal method is to select M ($M \ll N(N-1)$) plausible proposals for the following stage. However, this method faces one substantial challenge: how to select M plausible object-pairs from $N(N-1)$ combinations? Some works [13], [14] attempt to reserve specific proposals based on objectiveness scores from the pre-trained detection model. Intuitively, object confidences can not indicate whether objects are related to one another. For example, despite the fact that both "person(green box)" and "shirt(purple box)" have high object confidences (seen in Figure 1), there is a low probability for them as relationships due to distant relative location information.

Another long-standing issue that exists in the subsequent predicate recognition stage is the non-exclusive label problem. That is, as a one-hot recognition task, one basic assumption is that categories are disjoint in the label space. Thus, each visual sample belongs only to a single category. However, some $predicates$ in the label space do not satisfy this assumption and have very similar semantic meanings, which result in the blurred visual border among these predicates. In other words, one visual object-pair can be associated with a set of labels, not only a one-hot category. For example, as shown in Figure 1, "person-grass" can be assigned with both "on" and "stand on" as relationships. Another example is "beside" and "near", which are different categories yet with similar semantic information. Thus, the non-exclusive label problem may cause the network to deteriorate if we model the visual relationship detection task as one-hot recognition learning.

Considering the fact that most visual relationships are semantic concepts defined by humans, there are latent correlations or human knowledge hidden in them that has not been fully exploited by existing methods. Firstly, relative location correlation can be exploited to handle the combination explosion problem. Intuitively, relative location correlation implies location bias in human-annotated relationships, which can assist in selecting plausible proposals. For example, for object-pairs of "person - shirt", only those with spatial overlap may be annotated as the predicate of "wear"; for those without any spatial overlap, the probability to be one plausible proposal can be very low. That is, plausible triplets may satisfy one or more specific relative location bias: above or under, overlap or separate, etc. Thus, relative location correlation can be modeled to indicate whether objects are related to one another. Secondly, implicit semantic correlation can be exploited to handle the non-exclusive label problem. Specifically, implicit semantic correlation refers to ambiguity among $predicates$. Gao *et al.* [15] propose deep label distribution learning (DLDL) to address the ambiguity in age estimation and head pose estimation in which networks are supervised with discrete distribution labels. They find label ambiguity will improve recognition performance if it can be reasonably exploited. Thus, when predicate recognition is formulated as DLDL, we can leverage implicit semantic correlation to further improve performances.

Based on the above motivation, we devise a unified visual relationship detection framework with **T**wo-stage **C**orrelation **E**xploitation, termed **TCE**. In the object-pairs proposing stage, we propose one location-embedded rating module (**LRM**) based on relative location correlation to reduce the number of object-pair proposals. Through relative location encoding, **LRM** produces rating scores indicating the probabilities of objects being related to one another, so that plausible proposals are reserved. In the predicate recognition stage, to further reveal predicate-level semantic correlations, we construct a label-correlation graph module (**LGM**), which is supervised by discrete distribution labels. Due to expensive distribution annotations, a progressive training scheme is deployed to update a predicate-level similarity matrix that converts one-hot labels to distribution labels automatically. **LGM** encourages higher probabilities for all possible $predicates$, so that higher top-n recall can be achieved. The contributions of our work are summarized below.

(1) We revisit important issues in visual relationship detection and generalize the main challenges into two types of problems: the combination explosion problem and the non-exclusive label problem. For the first time, we address the two problems with two-stage correlation exploitation and devise a unified framework, termed **TCE**.

(2) We devise two novel modules, named **LRM** and **LGM**, for the two-stage correlation exploitation. **LRM** is incorporated into the object-pairs proposing stage to select plausible proposals; **LGM** is incorporated into the predicate recognition stage to increase probabilities of all plausible $predicates$.

(3) We achieve state-of-the-art results on two widely used datasets, namely, Visual Relationship Detection (VRD) [11] and Visual Genome (VG) [16], [17], and the improvement is significant, especially in relationship detection metric. In terms of R@100,k=1, we obtain 3.37% gains on VRD and 4.35% gains on VG.

The remainder of the paper is organized as follows. In Section II, we provide a brief review of related work. The proposed **TCE** is introduced in Section III. In Section IV, we present qualitative and quantitative evaluations of **TCE** on two datasets. Finally, we draw conclusions in Section V.

II. Related Work

In the early years, visual relationship detection was regarded as a phrase classification task [18], [19] that yielded poor generalizations because of the large scale of relationships. Recently, most methods [11], [14], [17], [20] have detected relationships expressed as triplets. There are also many works focusing on a specific type of relationship. For example, Ga *et al.* [21] learned a small set of prototypical spatial relationships by modeling the relative location between objects. Hu *et al.* [22] explicitly have defined a set of spatial pose-object interactions exemplars and achieve human-object relationship recognition by measuring the matching scores.

Wang *et al*. [23] designed a conditional random field to model human-human relationships in the spatio-temporal space. In our work, we focus on more general visual relationship detection, which is generally divided into two stages.

In the object-pairs proposing stage, Li *et al*. [13] proposed a triplet NMS based on the product of objectiveness scores to remove redundant object-pairs. However, there exists a gap between higher objectiveness scores and more plausible object-pairs. Recently, Zhan *et al*. [24] explored undetermined relationships and achieved significant improvements. In contrast to [24] that directly incorporates determinate confidences into final predictions, we utilize rating scores that indicate probabilities of objects being related to one another to select plausible proposals. Thus, our method can reduce computational complexity and achieve better performances. Despite that Zhang *et al*. [25] proposed "relationship proposal networks" based on a similar motivation, there are many differences between the **LRM** and their approach. In contrast to predicting visual and spatial scores separately, the **LRM** combines visual and relative location information and produces final rating scores directly. Moreover, their work ignores objectiveness scores; we integrate the outputs of object detection and the **LRM** to select and rank object-pair proposals.

In the predicate recognition stage, the basic method [11] takes union regions of object-pairs as inputs and adds language priors to preserve alignments with human perception. In the work of Xu *et al*. [26], visual context within an image was discovered and integrated for scene graph inference. Through propagating a message between region-centric context and object visual features, they were utilized for context modeling in [26] to reinforce each other. Furthermore, many works focus on multi-modal feature fusion. Other than visual information, different modal information has provided auxiliary features that play an important role in many tasks. By incorporating textual phrases into visual feature maps, Zhao *et al*. [27] achieved specific objects localization. By concatenating temporal position embeddings and video representations, Mun *et al*. [28] obtained significant improvements in action grounding. Recently, language and location information have also proven to be beneficial for predicate recognition. For example, SIL [29] is shown to capture semantic dependencies from language information and learn the adaptive parameters of the predicate classification model. Yu *et al*. [14] encoded location information into a one-dim vector as auxiliary features. Zhou *et al*. [30] introduced a spatial-channel attention mechanism guided by language and position. Following a similar motivation, we deploy a multi-modal fusion module to integrate the three-modal features. Moreover, to the best of our knowledge, we are the first to formulate predicate recognition into label distribution learning and devise an **LGM** to capture semantic correlation among *predicates*.

Deep label distribution learning (DLDL) [15], [31] was proposed to focus on label ambiguity and convert one-hot labels into discrete distribution labels. Rather than one-hot labels as in Figure 2.a, the network in Figure 2.b is supervised with discrete distribution labels to capture more precise supervision information. For age estimation and head pose estimation, Gao *et al*. [15] assigned distribution labels to each instance according to their ground truth. By updating label distributions, Yi *et al*. [32] maintained one network to obtain correct labels from noisy labels. Recently, Cho *et al*. [33] extracted distribution labels from teacher networks to guide student networks and thus achieved knowledge transfer as "distillation". Considering the fact that label distribution can naturally describe ambiguity among all possible labels, we introduce it to handle the relationship detection task with predicate ambiguity. Generally, label distributions are generated from independent networks [33] or are hand-crafted. For example, the Gaussian kernel is chosen to transform one-hot annotations into discrete distribution labels. As illustrated in Figure 2.c, we devise a learnable similarity matrix that can directly convert one-hot annotations to distribution labels. Although Qi *et al*. [34] also construct an affinity matrix to generate cross-camera soft labels for persons in the semi-supervision Re-ID task, our proposed similarity matrix aims at capturing the semantic correlation among *predicates*. Similar to multi-prototype learning of face media [35], each row of the similarity matrix represents the distribution prototype for a specific *predicate*. During the training procedure, the constructed matrix is updated progressively using the previous average predictions.

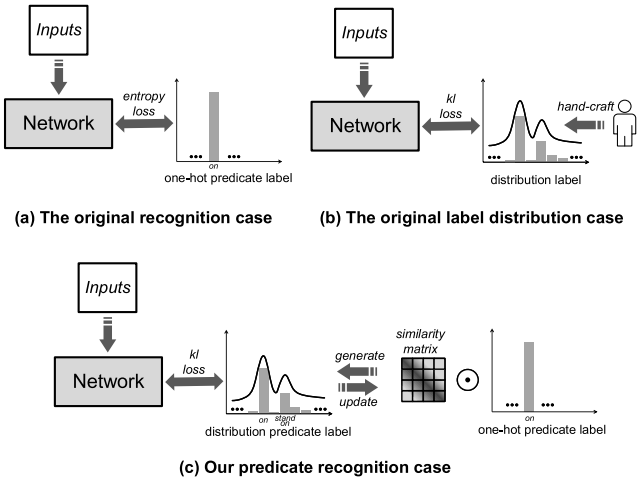

Fig. 2. The case of our predicate recognition model compared with two original cases.

A preliminary version of this article was published in [1]. This version (1) delves into two ignored problems: the combination explosion problem and the non-exclusive label problem; (2) re-formulates the visual relationship recognition as multiple label distribution learning (seen in Figure 2.c) instead of one-hot learning (seen in Figure 2.a); (3) devises a unified network with two-stage correlation exploitation **TCE**; and (4) performs additional ablation analyses and visualizations to clearly illustrate the advantages of our approach.

III. Two-Stage Visual Relationship Detection

In this section, we first describe an overall architecture of our proposed visual relationship detection framework **TCE**,

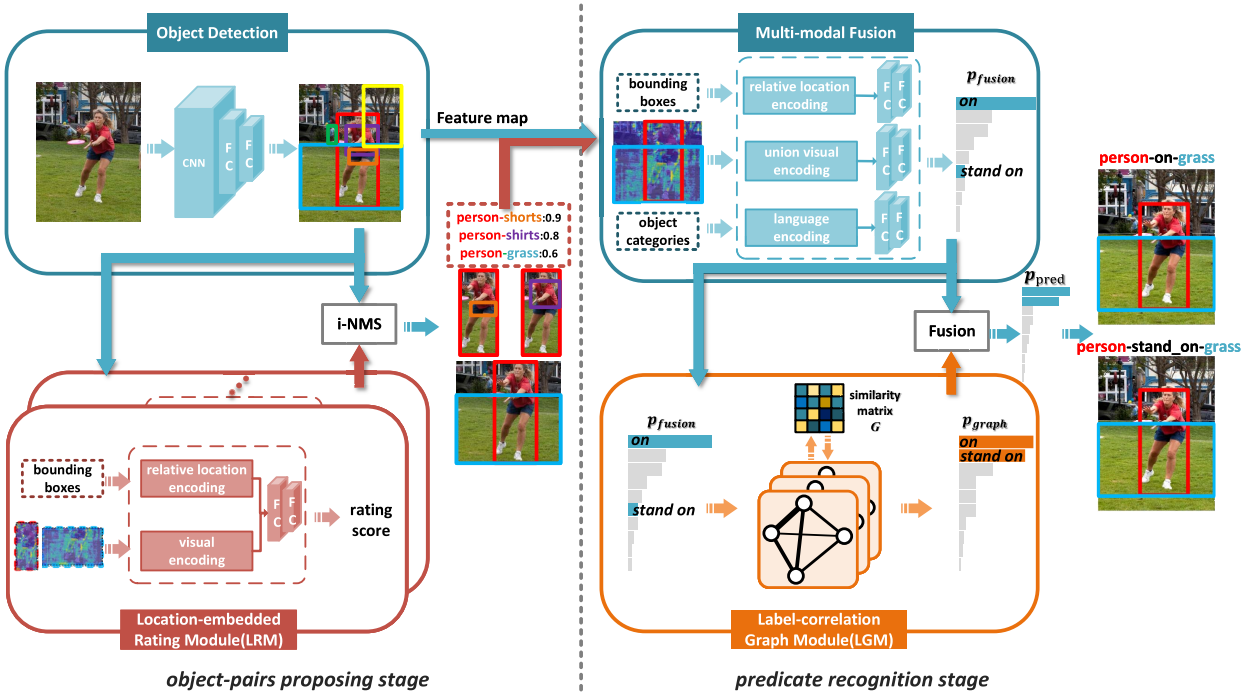

Fig. 3. The structure of our proposed framework,termed **TCE**. Red and orange boxes are the location-embedded rating module (termed **LRM**) and label-correlation graph module (termed **LGM**), respectively.

and then give a detailed introduction of each module in **TCE**. Training and inference procedures are presented in the end.

A. Overview

The overall framework of our proposed **TCE** is illustrated in Figure 3, which can be divided into two stages. The first stage is the object-pairs proposing stage, including the object detection model and the **LRM**(red box). The second stage is the predicate recognition stage, including multi-modal fusion and the **LGM**(orange box). First, given one image, a pretrained object detection model predicts object bounding boxes, objectiveness scores and object categories. Then, visual and relative location encoding are fed into the **LRM** to output rating scores. Based on the outputs of both the object detection and **LRM**, plausible proposals are reserved and ranked by our improved proposing scheme, termed i-NMS. Finally, the union visual, language, and relative location encoding is fed into the multi-modal fusion module and then into the **LGM** to recognize $predicates$.

B. Location-Embedded Object-Pair Rating

Our proposed **LRM** is built on the object detection model. As mentioned in the motivation, both visual and relative location assist in determining whether two objects are related to one another. Based on the pretrained object detection backbone, two FC layers that follow ROI-Align pooling [36] are set to extract visual representations. The visual representations are denoted as $\mathcal{R}_{vis}(sub) \in \mathbb{R}^d$ and $\mathcal{R}_{vis}(ob) \in \mathbb{R}^d$, where d is the visual representation dimension.

Relative location information is encoded with bounding boxes of two objects, defined as $b_{sub} = [x_{sub}, y_{sub}, w_{sub}, h_{sub}]$ and $b_{ob} = [x_{ob}, y_{ob}, w_{ob}, h_{ob}]$. sub (ob) denotes $subject$ ($object$), (x, y) is the upper left corner, and w, h are the width and height, respectively. Given a pair of bounding boxes, let W_u, H_u and S_u denote the width, height and area of the union region, respectively. In our work, we define the individual position as $[\frac{x}{W_u}, \frac{y}{H_u}, \frac{x+w}{W_u}, \frac{y+h}{H_u}, \frac{S}{S_u}]$ and the mutual position as $[\frac{x_{sub}-x_{ob}}{w_{ob}}, \frac{y_{sub}-y_{ob}}{h_{ob}}, \log\frac{w_{sub}}{w_{ob}}, \log\frac{h_{sub}}{h_{ob}}]$. Two individual position encodings and one mutual position encoding are concatenated and then L2-normalization is applied to obtain the final relative location encoding $\mathcal{R}_{loc}(sub, ob)$.

In **LRM**, we concatenate $\mathcal{R}_{vis}$ and $\mathcal{R}_{loc}$ and get the representation of object-pairs as input

$$\mathcal{R}_{lrm}(sub, ob) = [\mathcal{R}_{vis}(sub), \mathcal{R}_{vis}(ob), \mathcal{R}_{loc}(sub, ob)]. \quad (1)$$

LRM outputs rating scores s, which denote the probabilities of object-pairs being related to one another. Therefore, s is in the range [0,1], and defined as

$$\begin{aligned} h_{lrm} &= \mathcal{F}_o(\mathcal{R}_{lrm}; \boldsymbol{\Theta}_o), \\ s &= \frac{1}{1 + e^{-h_{lrm}}}, \end{aligned} \quad (2)$$

where $\mathcal{F}_o(\cdot; \boldsymbol{\Theta}_o)$ is an output network implemented by two FC-layers, and $\boldsymbol{\Theta}_o$ are the parameters.

Considering the fact that annotations for the **LRM** are not available, we generate binary rating labels (denoted as l) indicating whether object-pairs are related as relationships. In our work, binary rating labels are generated automatically by existing relationship annotations.

Let $\mathcal{B}^* = \{\langle b^*_{sub,1}, b^*_{ob,1}\rangle, \langle b^*_{sub,2}, b^*_{ob,2}\rangle, \cdots, \langle b^*_{sub,M}, b^*_{ob,M}\rangle\}$ denote a set of relationship annotations in one image, where M is the number of relationship annotations, and $\langle b^*_{sub,i}, b^*_{ob,i}\rangle$ is the bounding-box for one object-pair. $IoU(b_i, b_j)$ denotes the area of intersection between bounding-boxes b_i and b_j divided by the area of their union. We define triplet IoU scores ($tIoU$) to measure the maximum overlap between each object-pair $\langle b_{sub}, b_{ob}\rangle$ and human-annotated relationships $\mathcal{B}^*$:

$$\begin{aligned} u^m_{\langle b_{sub},b_{ob}\rangle} &= IoU(b_{sub}, b^*_{sub,m}) \cdot IoU(b_{ob}, b^*_{ob,m}), \\ tIoU_{\langle b_{sub},b_{ob}\rangle} &= \max\{u^1_{\langle b_{sub},b_{ob}\rangle}, \cdots, u^M_{\langle b_{sub},b_{ob}\rangle}\}, \end{aligned} \tag{3}$$

where $u^m_{\langle b_{sub},b_{ob}\rangle}$ is the overlap between $\langle b_{sub}, b_{ob}\rangle$ and the m-th human-annotated relationship. The larger the $tIoU$ is, the closer the corresponding object-pair is to one human-annotated relationship.

Based on $tIoU$, binary rating labels are set as 1 (up to $thresh\_high$) or 0 (below to $thresh\_low$), where the two thresholds are set as 0.45 and 0.25 in our experiment. Proposals with $tIoU$ lying between the two thresholds are removed in the training process. The loss of LRM is

$$\mathcal{L}_{lrm} = \frac{1}{N}\sum_{n=1}^{N}[l_n \cdot \log s_n + (1 - l_n) \cdot \log(1 - s_n)], \tag{4}$$

where N is the batch size.

C. Improved Object-Pairs Proposing Scheme

Our method produces plausible proposals from the outputs of object detection and the **LRM**. First, given each object-pair $\langle b_i, b_j\rangle$ for one image, we define proposal scores as

$$\tilde{s}_{\langle b_i,b_j\rangle} = s \cdot p_{obn}(b_i) \cdot p_{obn}(b_j), \tag{5}$$

where p_{obn} is the objectiveness score from object detection network. Then, based on the greedy NMS [37], our improved NMS, termed i-NMS, is implemented using proposal scores $\tilde{s}$. Similar to NMS, we remove the object-pair with the lower $\tilde{s}$ if the IoUs of two object-pairs reach the threshold(N_t) in i-NMS. Finally, plausible proposals along with $\tilde{s}$ are reserved for the second stage.

To further illustrate the details, we compare the differences between our i-NMS and the similar triplet NMS [13]: **Different inputs**. Given each object-pair $\langle b_i, b_j\rangle$, triplet NMS is only based on the product of two objectiveness scores $p_{obn}(b_i) \cdot p_{obn}(b_j)$; i-NMS is based on the proposal score $\tilde{s}$ in Eq.5; **Different outputs**. In triplet NMS, the outputs are only object-pair proposals. In i-NMS, the outputs contain both the proposal set and proposal score set, which means reserved proposals are ranked by proposal scores $\tilde{s}$.

D. Multi-Modal Features Fusion

In the predicate recognition stage, models recognize interaction given the union region of each object-pair proposal. However, visual information is too limited to capture subtle relation representation. As auxiliary features, language provides semantic information and location reflects specific spatial priors for different predicates. Thus, inspired by previous works [13], [30], we integrate visual, language and relative location information together to achieve more robust and discriminative feature representation in the multi-modal fusion module.

Analogous to [13], [17], we encode language information through concatenating the corresponding word2vec [38] embeddings. The word2vec mapping is trained offline with the whole English Wikipedia using GloVe. In this task, most object categories(as language inputs) are predefined common words that can be directly mapped to a 300-dimensional vector. For some phrase categories, we average the word vectors of each word. As a result, language representations $\mathcal{R}_{lan}(sub, ob)$ are encoded as

$$\mathcal{R}_{lan}(sub, ob) = [word2vec(sub), word2vec(ob)]. \tag{6}$$

In contrast to [26] in which region-centric context is integrated explicitly, we argue the union region of object-pairs already contains some visual context information for capturing predicates. Thus, similar to the above subsection, the flatten operation following ROI-Align is set to encode union visual area as $\mathcal{R}'_{vis}(sub, ob)$; the location information adopts the same relative location encoding $\mathcal{R}_{loc}(sub, ob)$. Then, three FC layer branches are implemented to map them into the same space as the fused relationship representation $\mathcal{R}_{fusion} \in \mathbb{R}^{d'}$:

$$\mathcal{R}_{fusion} = \mathcal{F}_1(\mathcal{R}'_{vis}; \boldsymbol{\Theta}_1) \odot \mathcal{F}_2(\mathcal{R}_{lan}; \boldsymbol{\Theta}_2) \odot \mathcal{F}_3(\mathcal{R}_{loc}; \boldsymbol{\Theta}_3), \tag{7}$$

where $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3$ are learnable weights and $\odot$ denotes the dot product operation. Finally, a linear operation, followed by a softmax layer, is applied to convert $\mathcal{R}_{fusion}$ into predicate probabilities $\boldsymbol{p}_{fusion}$.

$$\boldsymbol{p}_{fusion} = \text{Softmax}(\theta \cdot \mathcal{R}_{fusion} + b_o). \tag{8}$$

E. Predicate Recognition Using the Label-Correlation Graph

To reveal the implicit label-correlation, we construct the **LGM** to model the similarities among $predicates$. **LGM** mainly contains a label-correlation graph $\mathcal{G}$ and a predicate-level similarity matrix $\boldsymbol{G}$. Each node of $\mathcal{G}$ represents a specific $predicate$ and each edge between two nodes represents the implicit label-correlation. Each row of $\boldsymbol{G}$ represents the distribution prototype for a specific $predicate$ and is used to convert one-hot annotations to discrete distribution labels. Rather than a hand crafted definition, $\boldsymbol{G}$ is learned in a data-driven manner. According to the observation in [39], secondary predictions of deep networks are prone to be reasonable predictions. Thus, we update $\boldsymbol{G}$ progressively using previous average predictions to enhance the secondary predictions.

In our work, we follow the structure of GGNN [40] and define the label-correlation graph as $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{A})$, where $\boldsymbol{V}$ denotes a node set and $\boldsymbol{A}$ denotes an edge set (also determines the message passing in the graph). The detailed propagation process is defined as

$$\begin{aligned} \boldsymbol{h}_v^{(1)} &= [\boldsymbol{x}_v^\top, \boldsymbol{0}]^\top, \\ \boldsymbol{a}_v^{(t')} &= \boldsymbol{A}_v^\top [\boldsymbol{h}_1^{(t'-1)\top} \cdots \boldsymbol{h}_{|V|}^{(t'-1)\top}]^\top + \boldsymbol{b}, \\ \boldsymbol{h}_v^{(t')} &= GRU(\boldsymbol{h}_v^{(t'-1)}, \boldsymbol{a}_v^{(t')}), \end{aligned} \tag{9}$$

where GRU is the Gated Recurrent Unit [33], and $\boldsymbol{h}_v^{(t')}$ denotes the v-th node features in step t'. Each node v of $\mathcal{G}$ denotes a type of $predicate$, thus $|\boldsymbol{V}|$ is equal to the number of defined $predicates$. Each bit of predicate probability $\boldsymbol{p}_{fusion}$ is initially fed into the corresponding node. Furthermore, a transform network is deployed between two nodes (such as h_i and h_j) to compute the coefficient c_{ij} that indicates the correlation of node i to node j:

$$c_{ij} = \mathcal{F}_4([\boldsymbol{h}_i, \boldsymbol{h}_j]; \boldsymbol{\Theta}_4), \tag{10}$$

where $\mathcal{F}_4(\cdot; \boldsymbol{\Theta}_4)$ is two FC-layers, and $\boldsymbol{\Theta}_4$ are the parameters. Then, the message passing structure $\boldsymbol{A} \in \mathbb{R}^{|\boldsymbol{V}|\times|\boldsymbol{V}|}$ is determined by normalizing each row of node coefficients:

$$A_{ij} = \text{Softmax}_j(c_{ij}) = \frac{e^{c_{ij}}}{\sum_{k=1}^{|\boldsymbol{V}|} e^{c_{ik}}}. \tag{11}$$

Finally, the output of **LGM** is

$$\begin{aligned} o_v &= O(\boldsymbol{h}_v^{(e)}\top, p_{fusion,v}), v = 1, 2, \cdots, |\boldsymbol{V}|, \\ \boldsymbol{p}_{graph} &= \text{Softmax}([o_1, \cdots, o_{|\boldsymbol{V}|}]), \end{aligned} \tag{12}$$

where O is an FC-layer network, $p_{fusion,v}$ is the v-th value of $\boldsymbol{p}_{fusion}$ and $\boldsymbol{h}_v^{(e)}$ denotes the v-th node feature in the final step.

Due to the unavailability for distribution annotations, we use $\boldsymbol{G}$ to convert one-hot annotations $\boldsymbol{y}$ into discrete distribution labels $\tilde{\boldsymbol{y}} = \boldsymbol{G}^\top \boldsymbol{y}$, where the value $\boldsymbol{G}_{ij}$ denotes the diffusion probability of category i to j. $\boldsymbol{G}$ is updated iteratively using previous average predictions.

Specifically, let $\boldsymbol{G}^{(t)}$ denote the similarity matrix after the t-th epoch and $\boldsymbol{G}^{(0)} = \boldsymbol{I}$ initially is equivalent to utilizing one-hot annotations as supervision information. After each epoch t, statistical distributions $\bar{\boldsymbol{g}}_v^{(t)}$ are accumulated by averaging predictions $\boldsymbol{p}_{fusion}$ across all samples belonging to the v-th $predicate$ in the training set. To enhance the secondary predictions [39], softer predicate distributions are produced by temperature T,

$$\boldsymbol{g}_v = \frac{\bar{\boldsymbol{g}}_v^{\frac{1}{T}}}{\sum_{j=1}^{|\boldsymbol{V}|} \bar{g}_{vj}^{\frac{1}{T}}}, \tag{13}$$

where higher temperature T encourages the production of softer distributions. Then, $\boldsymbol{G}$ is updated by the previous statistical distributions $\boldsymbol{g}$ as

$$\boldsymbol{G}_v^{(t)} = \boldsymbol{G}_v^{(t-1)} + \alpha \boldsymbol{g}_v^{(t-1)}, v = 1, 2, \cdots, |\boldsymbol{V}|, \tag{14}$$

where α controls the update operation. Finally, **LGM** is optimized with KL-divergence between $\boldsymbol{p}_{graph}$ and the label distribution $\tilde{\boldsymbol{y}} = \boldsymbol{G}^\top \boldsymbol{y}$,

$$\begin{aligned} \mathcal{L}_{kl} &= \frac{1}{N}\sum_{n=1}^{N} KL(\tilde{\boldsymbol{y}}_n || \boldsymbol{p}_{graph,n}), \\ KL(\tilde{\boldsymbol{y}} || \boldsymbol{p}_{graph}) &= \sum_{j=1}^{|\boldsymbol{V}|} \tilde{y_j} \log(\frac{\tilde{y_j}}{p_{graph,j}}). \end{aligned} \tag{15}$$

Through mimicking distribution labels, the **LGM** is forced to capture the correlations among predicates progressively. During the inference procedure, our method only runs a single propagation process without $\boldsymbol{G}$. In the predicate recognition stage, the final predictions aggregate the outputs of the multi-modal fusion module and the **LGM** as

$$\boldsymbol{p}_{pred} = \mu \boldsymbol{p}_{fusion} + (1 - \mu)\boldsymbol{p}_{graph}, \tag{16}$$

where μ balances the outputs of the above two modules. To prevent a trivial solution, the network is still supervised with one-hot annotations as

$$\mathcal{L}_{pred} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{cel}(\boldsymbol{p}_{pred,n}, \boldsymbol{y}_n) + \mathcal{L}_{cel}(\boldsymbol{p}_{fusion,n}, \boldsymbol{y}_n), \tag{17}$$

where $\mathcal{L}_{cel}$ is the cross-entropy loss.

F. Training and Inference Procedures

During the training procedure, our framework is trained in two stages. First, the object detection model is optimized along with the **LRM** in the object-pairs proposing stage. Thus, the training loss in the first stage is defined as:

$$\mathcal{L}_1 = \mathcal{L}_{odm} + \lambda_1 \mathcal{L}_{lrm}, \tag{18}$$

where $\mathcal{L}_{odm}$ is for the object detection model, which just follows [2]. Then, we freeze the network of the first stage, and optimize the multi-modal fusion along with the **LGM** in the predicate recognition stage. The training loss is defined as:

$$\mathcal{L}_2 = \mathcal{L}_{pred} + \lambda_2 \mathcal{L}_{kl}, \tag{19}$$

During the inference procedure, the final visual relationship predictions integrate the outputs of the two stages modules. Specifically, the visual relationship probability for object-pair $\langle b_i, b_j \rangle$ is

$$\boldsymbol{P}_{\langle b_i, b_j \rangle} = p_{obn}(b_i) \cdot p_{obn}(b_j) \cdot \boldsymbol{p}_{pred} \cdot \tilde{s}. \tag{20}$$

IV. Experiments

In this section, we compare **TCE** with state-of-the-art methods, and investigate the gains with various components and hyper-parameters. To present an intuitive insight into how our model works, we also visualize the update process of the similarity matrix as well as some detection results.

A. Datasets

Visual Relationship Detection dataset(VRD) [11] contains 70 predicates and 100 objects. In total, it contains 37,993 relationships with 6,672 relations and 24.25 predicates per object category. Moreover, 1,877 relationships only exist in the testing set for the zero-shot evaluation. The whole dataset is split into 4,000 images for training and 1,000 images for testing.

Visual Genome (VG) [16] is a large scale relationship dataset. Unlike VRD that is constructed by computer vision experts, VG is annotated by crowd workers and thus $predicates$ and objects are noisy. Specifically, many $predicates$ are uncommon and objects are in different forms, which causes a more serious long-tail distribution problem. Thus, we adopt a clean subset [17] in which an official

TABLE I
COMPARISON OF TCE WITH STATE-OF-THE-ART METHODS ON VRD, WHERE BOLD FONT INDICATES BEST RESULTS, UNDERLINED SECOND-BEST

Methods	Predicate Detection			Phrase Detection				Relationship Detection			
	R@100/50 k=1	R@100 k=70	R@50 k=70	R@100 k=1	R@50 k=1	R@100 k=70	R@50 k=70	R@100 k=1	R@50 k=1	R@100 k=70	R@50 k=70
VTransE [17]	44.76	-	-	22.42	19.42	-	-	15.20	14.07	-	-
Language-Pri [11]	47.87	84.34	70.97	17.03	16.17	24.90	20.04	14.70	13.86	21.51	17.35
TCIR [20]	53.59	-	-	25.26	23.88	-	-	23.39	20.14	-	-
CDD-Net [41]	-	93.76	87.57	-	-	-	-	-	-	26.14	21.46
VIP [13]	-	-	-	-	-	27.91	22.78	-	-	20.01	17.31
DR-Net [42]	-	81.90	80.78	-	-	23.45	19.93	-	-	20.88	17.73
VRL [43]	-	-	-	22.60	21.37	-	-	20.79	18.19	-	-
Factorizable Net [44]	-	-	-	-	-	30.77	26.03	-	-	21.20	18.32
LKD [14]	55.16	94.65	85.64	24.03	23.14	29.43	26.32	21.34	19.17	31.89	22.68
STL [45]	-	-	-	33.48	28.92	-	-	26.01	22.90	-	-
Zoom-Net [46]	55.98	94.56	89.03	28.09	24.82	37.34	29.05	21.41	18.92	27.30	21.37
SIL [29]	56.56	-	-	24.50	20.82	-	-	16.01	13.81	-	-
BLOCK [47]	-	92.58	86.58	-	-	28.96	26.32	-	-	20.96	19.06
MF-URLN [24]	58.20	-	-	36.10	31.50	-	-	26.80	23.90	-	-
HGAT [48]	**59.54**	**97.02**	**90.91**	-	-	-	-	24.63	22.52	27.73	22.90
MCN [49]	58.10	-	-	37.10	31.80	-	-	28.00	24.50	-	-
TCE(ours)	57.93	96.05	90.25	**40.01**	**33.46**	**45.69**	**36.69**	**31.37**	**26.76**	**37.13**	**30.19**

pruning is applied. In the clean subset, *predicates* with less than 5 samples are filtered out; objects in different forms are merged. For example, "young woman" and "lady" are merged to "woman". As a clean subset, there are 99,658 images with 200 object categories and 100 predicates. We follow the same train/test split in [17], where 73,801 images are for training and 25,857 images are for testing. For a fair comparison, all methods below use the same subset on VG.

B. Experimental Setup

Our experiments are based on PyTorch, and the object detection follows the official released Faster R-CNN [2]. Considering the good performance of ResNet50-FPN [50], we adopt the same structure as our backbone network and share it in all components. In the object-pairs proposing stage, the initial learning rate is set as 0.02, and the momentum as 1e-4. Stochastic gradient descent (SGD) is used to optimize our model for 20 epochs. In the experiments, λ_1 is set as 1, the number of reserved object-pair proposals (N_o) is 110, and the threshold for NMS (N_t) is 0.25. In the predicate recognition stage, we set the initial learning rate as 0.01 divided by 10 at epoch 20 and 30 (40 epochs in total), and the momentum as 1e-4. To accelerate the convergence of training, Adam [51] is applied to optimize the second stage. During the training phase, λ_2 is set as 1 empirically. In regard to the fusion coefficient μ, it controls the importance of the two components in the recognition stage and is set as 0.5. In addition, the similarity matrix is updated after 10 epochs, and the update coefficient is set as 0.1.

C. Evaluation Metrics

Because annotations of visual relationships are not exhaustive, mAP evaluation metrics will penalize positive predictions absent in annotations. In our work, we follow [11] to apply Recall@50 (R@50) and Recall@100 (R@100), where top-n predictions are retrieved in one image. In the same manner as [14], a hyper-parameter k is set to take the top k *predicates'* predictions into consideration per object-pair. Specifically, R@n,$k = 1$ is equivalent to R@n in [11], and R@n,$k = 70$ in VRD/R@n,$k = 100$ in VG is equivalent to taking all *predicates* into consideration. Furthermore, we adopt three evaluation metrics, including predicate detection, phrase detection and relationship detection. Predicate detection (which provides bounding boxes and object categories) evaluates performances of distinguishing *predicates*. Moreover, phrase detection (which needs to detect union regions of object-pairs) and relationship detection (which needs to detect objects) can evaluate performances of detecting plausible object-pairs.

D. Comparison to State-of-the-Art Methods

To validate the effectiveness of our proposed method, we evaluate **TCE** with current state-of-the-art methods on VRD and VG.

1) Experiments on Visual Relationship Detection: We compare our proposed method that is denoted as "**TCE**(ours)" with some related methods [11], [13], [14], [17], [20], [24], [29], [41]–[49] in three metrics in Table I. Compared to these prior works, our method "**TCE**(ours)" achieves competitive or better performances in different metrics. In predicate detection, **HGAT** [48] constructs an object-level attention graph and a triplet-level attention graph and is the current state-of-the-art method. Compared to **HGAT**, "**TCE**(ours)" still achieves comparable performances, especially 90.25% vs 90.91% for R@50,$k = 70$. It means the gains of our method mainly depend on improvements of accuracy for all plausible predicates. Furthermore, as more challenging metrics, phrase detection and relationship detection can better reflect practice scenarios. "**TCE**(ours)" outperforms the current state-of-the-art method **MCN** [49], e.g., 31.37% vs 28.00% for R@100,$k = 1$ in relationship detection. It means our method

TABLE II

COMPARISON WITH PREVIOUS METHODS ON THE VRD ZERO-SHOT SET; BOLD FONT INDICATES BEST RESULTS, UNDERLINED SECOND-BEST

Methods	R@100 k=1	R@50 k=1	R@100 k=70	R@50 k=70
Language-Pri [11]	8.45	8.45	50.04	29.77
TCIR [20]	16.42	16.42	-	-
Weakly-sup [53]	23.6	23.6	-	-
LKD [14]	16.98	16.98	74.65	54.20
MCN [49]	**26.7**	**26.7**	-	-
TCE(ours)	26.52	26.52	**86.66**	**72.97**

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS ON VG; BOLD FONT INDICATES BEST RESULTS, UNDERLINED SECOND-BEST

k	Methods	Predicate Detection R@100	Predicate Detection R@50	Phrase Detection R@100	Phrase Detection R@50	Relationship Detection R@100	Relationship Detection R@50
1	VTransE [17]	62.87	62.63	10.45	9.46	6.04	5.52
	Shuffle [54]	62.94	62.71	-	-	-	-
	VSA-Net [55]	64.53	64.41	9.97	9.72	6.28	6.02
	PPR-FCN [12]	64.86	64.17	11.08	10.62	6.91	6.02
	DSL [56]	-	-	15.61	13.07	8.00	6.82
	STL [45]	-	-	18.13	14.62	9.41	7.93
	SIL [29]	68.91	68.63	12.05	10.60	6.64	5.96
	MF-URLN [24]	**72.20**	**71.90**	32.10	26.60	16.50	14.40
	HGAT [48]	68.32	68.11	-	-	-	-
	MCN [49]	71.90	71.60	32.70	26.90	17.10	14.90
	TCE(ours)	71.25	70.95	**34.31**	**26.90**	**21.45**	**17.22**
100	DR-Net [42]	91.26	88.26	25.74	20.28	22.23	17.51
	CDD-Net [41]	74.92	70.42	-	-	-	-
	HGAT [48]	**96.65**	90.05	-	-	-	-
	TCE(ours)	96.23	**91.19**	**35.04**	**27.75**	**22.82**	**18.47**

TABLE IV

COMPARISON WITH OUR CONFERENCE VERSION IN R@N,$k = 1$; BOLD FONT INDICATES THE BEST RESULTS

Datasets	Methods	Predicate Detection R@100	Predicate Detection R@50	Phrase Detection R@100	Phrase Detection R@50	Relationship Detection R@100	Relationship Detection R@50
VRD	RLM[1]	57.19	57.19	39.74	33.20	31.15	26.55
VRD	**TCE**	**57.93**	**57.93**	**40.01**	**33.46**	**31.37**	**26.76**
VG	RLM[1]	69.87	69.57	33.92	26.60	21.17	16.96
VG	**TCE**	**71.25**	**70.95**	**34.31**	**26.90**	**21.45**	**17.22**

can produce more plausible proposals and better adapt to real applications.

In Table II, we also compare our proposed method with some existing methods on the zero-shot set, where relationship triplets that belong to the training set are removed. Thus, it can reflect the generalization ability of our proposed recognition method. Compared to current state-of-the-art methods, our method "**TCE**(ours)" also achieves competitive performances. While **LKD** [14] introduces external linguistic knowledge, our method outperforms it considerably, e.g., 72.97% vs 54.20% for R@50,$k = 70$. Thus, by capturing the semantic label-correlation, it also helps models to improve the performances in some unseen triplets.

2) Experiments on Visual Genome: To further validate the effectiveness of our proposed method "**TCE**(ours)" on a large-scale dataset, we compare our method with some existing methods on VG. Considering the fact that most existing methods are only evaluated in R@n,$k = 1$ on VG, we follow the same metric. As shown in Table III, "**TCE**(ours)" achieves comparable or improved performances. In particular, it outperforms the current state-of-the-art methods in phrase detection and relationship detection. For example, compared to **MCN** [24], "**TCE**(ours)" yields 4.35% and 2.32% gains for R@100 and R@50 in relationship detection, respectively. Furthermore, several methods experimented in R@n,$k = 100$ are also listed in the second row of Table III, where all the of predictions are evaluated. As shown in Table III, "**TCE**(ours)" still achieves the current state-of-the-art in almost all metrics.

3) Comparison With the Preliminary Version: Our proposed **TCE** is based on the preliminary version **RLM** [1] and made many improvements. The different performances are listed in Table II and we clarify the benefits and necessity of the improvements from two aspects. On the one hand, **TCE** obtains more gains. We find the obtained gains are consistent on both the VRD and VG datasets. On the other hand, the training procedure for **TCE** is more flexible. **RLM** has to compute the relative location bias from training sets. Then a static graph is built based on explicit location similarities among predicate categories. Once new training samples are introduced, a corresponding graph needs to be constructed again. **TCE** is updated progressively based on a learnable graph in the **LGM**. Thus, it can learn the implicit semantic correlation in a data-driven manner.

E. Component Analysis

In our work, we follow the prior works to construct our object detection model and keep the same settings.

1) Impact of Reserved Object-Pair Proposals: To explore the performance of our proposed **LRM**, we experiment on different numbers of object-pair proposals in the object-pairs proposing stage. We compare our method with two other popular methods, termed "**simple top**" and "**simple product**". "**simple top**" (adopted by [11], [12]) first filters out objects based on objectiveness scores and then combines them into object-pairs. "**simple product**" (adopted by [13], [14]) directly selects object-pairs based on the product of $\langle sub - ob \rangle$ objectiveness scores. With the outputs of the **LRM**, "**TCE**(ours)" introduces them into both i-NMS and final predictions, while "**TCE**†(ours)" only feeds them into i-NMS. Specifically, plausible object-pair proposals are reserved but not ranked in "**TCE**†(ours)". In the interests of fairness, all methods apply the same predicate recognition branch [1].

The results are shown in Figure 4. The left one is the results in phrase detection and the right one is the results in relationship detection. We can observe our method achieves significant improvements compared with the two other methods. With reserved proposals increasing, both "**TCE**†(ours)" and "**TCE**(ours)" obtain more gains rapidly, which validates that the **LRM** can assist in reserving plausible object-pairs.

TABLE V
PERFORMANCES OF MULTI-MODAL FUSION MODULE IN DIFFERENT SETTINGS IN PREDICATE DETECTION

Settings						Entire set			Zero-shot set		
visual	language	location	concate	average	product	R@100/50,k=1	R@100,k=70	R@50,k=70	R@100/50,k=1	R@100,k=70	R@50,k=70
✓						33.74	86.06	72.74	12.83	73.57	53.29
✓	✓				✓	53.78	94.68	86.96	19.33	80.92	61.76
✓		✓			✓	48.83	92.71	83.71	24.55	84.75	68.86
✓	✓	✓	✓			56.39	94.44	88.35	23.70	85.29	71.34
✓	✓	✓		✓		52.28	94.93	88.79	23.44	84.94	71.77
✓	✓	✓			✓	**56.81**	**95.97**	**89.33**	**24.72**	**87.17**	**71.34**

TABLE VI
PERFORMANCES OF LGM IN DIFFERENT SETTINGS IN PREDICATE DETECTION

Settings	Entire set		Zero-shot set	
	R@50,k=1	R@50,k=70	R@50,k=1	R@50,k=70
complete LGM	**57.93**	**90.25**	**26.52**	**72.97**
without graph	56.81	89.33	24.72	71.34
without learnable graph	57.61	89.93	26.26	71.69
without updated similarity matrix	56.60	89.60	25.75	72.11

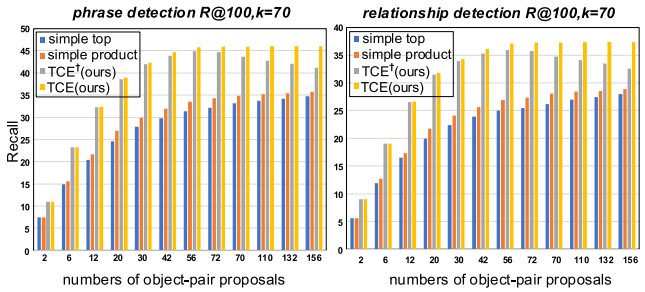

Fig. 4. Impact of reserved object-pair proposals between our methods (**TCE**(ours) and **TCE**†(ours)) and two other methods (**simple top** and **simple product**). The X-axis denotes the number of reserved object-pair proposals, and the Y-axis denotes recall values.

Generally, when only a few proposals are reserved (approximately less than 100 proposals), our methods still outperform others. Interestingly, when the reserved numbers further increase, the performances of "**TCE**†(ours)" drop slightly. We surmise that too many reserved proposals will inevitably cause redundancy. To alleviate this issue, we further rank the reserved object-pairs, termed "**TCE**(ours)".

2) Discussion of Multiple Modal Feature Fusion: To investigate the influences of different components in the multi-modal fusion module, we evaluate different settings without the **LGM** on VRD. All of the experiments are evaluated in predicate detection listed in Table V. Two kinds of settings are analyzed below.

The first one is different modal feature fusion. The inputs contain language encoding, relative location encoding and visual features, termed "language", "location" and "visual" briefly. In our experiments, both "visual+location" and "visual+language" outperform "visual". It proves that solely implementing visual features is limited in capturing the subtle interactions between objects. Three-modal feature fusion achieves the best performances. Thus, we consider that language, location and visual features provide different inferences for predicate recognition.

The second one is different ways of combining multi-modal features; "concate", "average" and "product" denote concatenation, average and element wise product operation, respectively. In our experiment, "product" outperforms others in all evaluation metrics. We consider that the "product" operation makes the gradient easily flow backward into the three-modal branches compared to the other two operations.

*3) Benefits of Settings in the **LGM**:* In this subsection, we investigate the gains of our **LGM** in different settings (seen in Table VI). We compare our method termed "complete LGM" with methods in three kinds of settings that include the following: without graph, without learnable graph, and without updated similarity matrix. Specifically, the entire **LGM** is removed in "without graph". "without learnable graph" means that edge weights in the graph are not learnable and are set as the average of all nodes initially. Finally, "without updated similarity matrix" means the similarity matrix is an identity matrix and is not updated, which is also equivalent to degenerating distribution learning into traditional one-hot recognition learning. In our experiments, "complete LGM" achieves the best performances in all evaluation metrics, which means that all settings in the **LGM** contribute to improving performances.

4) Hyper Parameter Sensitivity: The performances of **TCE** might be influenced by hyper parameters. We conducted extended experiments on the sensitivity of **TCE** to the hyper-parameter on VRD. In the object-pairs proposing stage, there are mainly three hyper parameters, including N_o, λ_1 and N_t. N_o is the number of reserved object-pair proposals that are already explored above. λ_1 balances the object detection and **LRM**, and N_t is the threshold for NMS. As listed in Table VII, λ_1 is traversed in a range of [0.1, 2], and performances are best at 1. With N_t increasing in the range of

TABLE VII

PERFORMANCES OF DIFFERENT SETTINGS IN THE OBJECT-PAIRS PROPOSING STAGE ON THE ENTIRE SET

Settings		Phrase Detection R@100 k=1	Phrase Detection R@50 k=1	Phrase Detection R@50 k=70	Relationship Detection R@100 k=1	Relationship Detection R@50 k=1	Relationship Detection R@50 k=70
λ_1	0.1	34.29	29.18	30.57	26.64	23.38	24.97
	0.5	38.13	31.95	35.00	29.77	25.50	28.88
	1	**40.01**	**33.46**	**36.69**	**31.37**	**26.76**	**30.19**
	2	39.15	32.99	36.36	30.68	26.30	29.98
N_t	0.1	39.63	33.25	36.59	31.00	26.71	30.18
	0.25	**40.01**	**33.46**	**36.69**	31.37	26.76	30.19
	0.4	39.81	33.35	36.57	31.78	27.17	30.40
	0.5	39.36	32.82	36.19	**32.04**	**27.11**	**30.45**

TABLE VIII

PERFORMANCES OF DIFFERENT SETTINGS IN THE PREDICATE RECOGNITION STAGE IN PREDICATE DETECTION

Settings		Entire set R@100/50 k=1	Entire set R@50 k=70	Zero-shot set R@100/50 k=1	Zero-shot set R@50 k=70
λ_2	0.1	56.99	90.05	25.32	72.54
	0.5	57.50	89.96	26.01	71.17
	1	**57.93**	**90.25**	**26.52**	**72.97**
	2	57.20	89.40	24.64	70.06
T	1	56.93	89.95	25.24	71.17
	1.5	57.25	90.09	25.58	71.00
	2	**57.93**	**90.25**	**26.52**	**72.97**
	2.5	57.41	89.93	24.98	71.86

TABLE IX

TESTING TIME PER IMAGE FOR DIFFERENT NUMBERS OF OBJECT-PAIRS AND OBJECT DETECTORS

Detectors \ Object-pairs	60	110	200	w/o constraints
Faster R-CNN	144 ms	182 ms	250 ms	1,348 ms
YOLOv3	72 ms	110 ms	178 ms	-

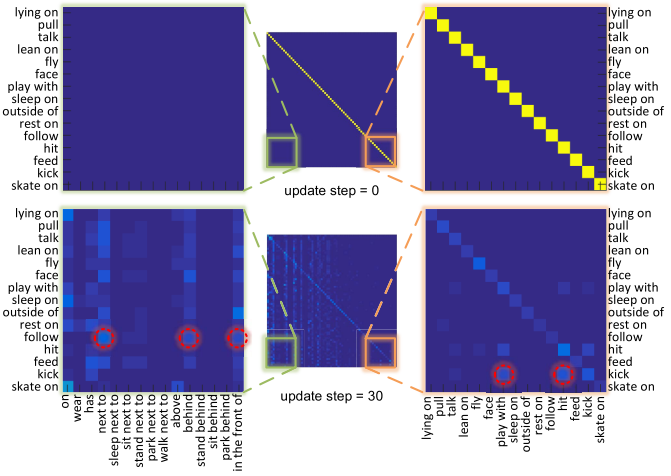

Fig. 5. Visualization of the predicate-level similarity matrix after several update steps on VRD. Red dot circles are the semantically similar labels.

[0.1, 0.5], performances drop slightly in phrase detection and the opposite occurs in relationship detection. In the predicate recognition stage, there are mainly two hyper parameters, including λ_2 and T. As listed in Table VIII, λ_2 balances the one-hot annotations and discrete distribution labels, and T controls the smooth degree of predicate distributions. The performances of **TCE** are robust with respect to choices of λ_2 and T, and they slightly better when λ_2 is 1 and T is 2.

F. Model Complexity and Inference Speed

Based on the ResNet50-FPN, the computational complexity of **TCE** is 31.55 GMACs. When 110 object-pairs are reserved within one image, the testing time is 182 ms per image run on a 1080Ti GPU. As listed in Table IX, the bottlenecks of the inference speed are the number of reserved object-pairs and the type of object detectors. "w/o constraints" denotes the naive proposing method: $N(N-1)$ object-pairs are reserved based on N detected objects. Thus, it's possible to further increase inference speed by maintaining fewer object-pairs and deploying a lightweight object detection model. For example, based on 60 reserved object-pairs and YOLOv3 [56], the testing time is substantially reduced to 72 ms per image.

G. Model Visualization

To investigate how **TCE** works, we visualize the similarity matrix after several update steps in VRD, as shown in Figure 5. The initial similarity matrix is an identity matrix. After 30 update steps, we find the matrix captures the implicit label-correlation among *predicates*. For example, if the ground truth is "follow", then "next to", "behind" and "in the front of" are assigned higher scores (seen in the lower left matrix of Figure 5); therefore, it means our **LGM** captures spatial correlation information among them. Moreover, if the ground truth is "kick", then "hit" and "play with" are assigned with higher scores (seen in the lower right matrix of Figure 5); therefore, it means our **LGM** captures semantic correlation information among them. Thus, our method learns an effective and plausible predicate distribution progressively to improve the performances of ambiguous predicate recognition.

We also visualize some detection results (seen in Figure 6) on VRD. The top-5 object-pair proposals are provided and each proposal only reserves the top-1 predicate in relationship detection. To focus on the gains of our **LRM**, we compare predicted relationships between models with and without the **LRM**. "w/o LRM" removes the **LRM** and selects object-pairs based on the product of objectiveness scores. From the top row in Figure 6, we can see that relationships in "w LRM" are more diverse and closer to the ground truth than those in "w/o LRM". The top-4 predicate predictions are provided in predicate detection in the bottom row of Figure 6. To focus on the gains of our **LGM**, we compare differences between models with and without the **LGM**. With the help of the **LGM** (termed "w LGM"), the correct and plausible *predicates* are assigned higher scores and rankings than those without the **LGM** (termed "w/o LGM").

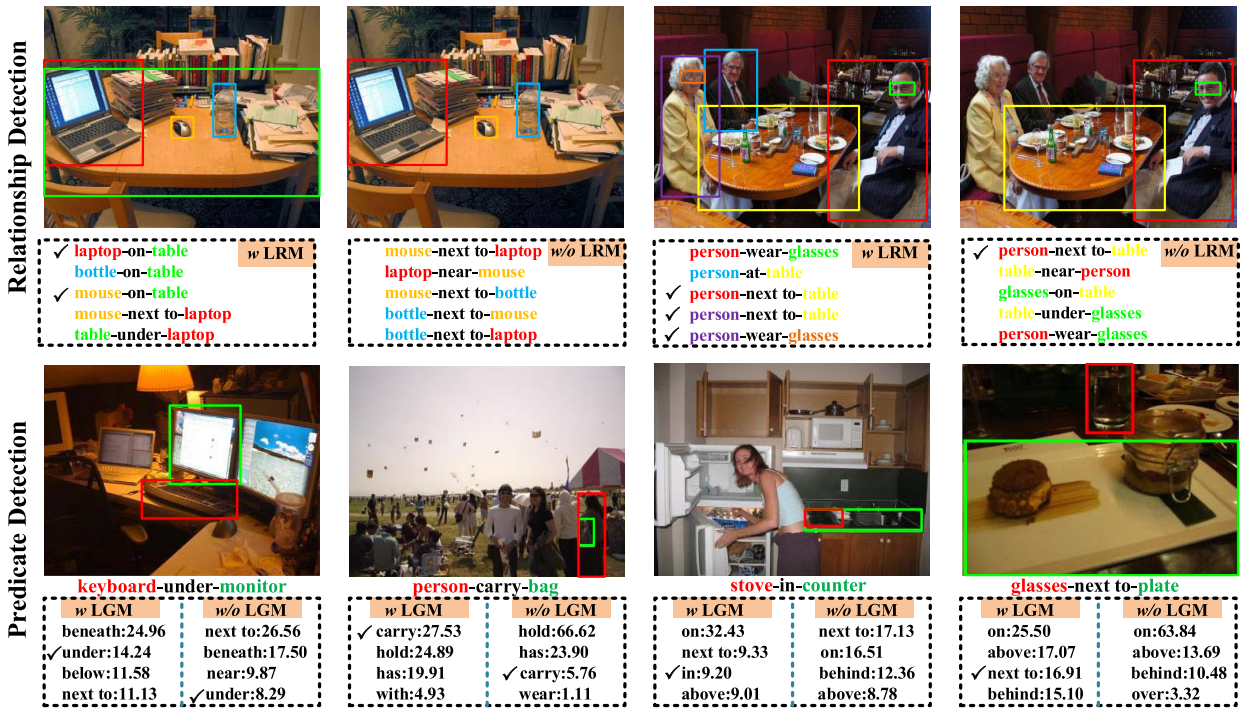

Fig. 6. Qualitative examples of relationship detection (top row) and predicate detection (bottom row) on VRD. The ✓denotes the ground-truths.

V. Conclusion

In this paper, we explore two challenging issues (redundancy and ambiguity) neglected by prior works and devise a unified network **TCE**. We propose the **LRM** to reserve plausible proposals in the object-pairs proposing stage, and the **LGM** to increase the probabilities of all plausible *predicates* in the predicate recognition stage. The experiments on the VRD and VG datasets validate the effectiveness of our method. In future work, we will attempt to extend this method to grasp relationships in videos or 3D real-world scenarios, hoping that relationship detection can inspire solving action recognition and video understanding in the real world.

References

[1] H. Zhou, C. Zhang, and C. Hu, "Visual relationship detection with relative location mining," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 30–38.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[3] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Spatio-temporal closed-loop object detection," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1253–1263, Mar. 2017.

[4] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang, "Multi-modal curriculum learning for semi-supervised image classification," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3249–3260, Jul. 2016.

[5] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2017.

[6] C. Wang, S. Dong, X. Zhao, G. Papanastasiou, H. Zhang, and G. Yang, "SaliencyGAN: Deep learning semisupervised salient object detection in the fog of IoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2667–2676, Apr. 2020.

[7] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, May 2017.

[8] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6985–6994.

[9] Y. Peng and J. Chi, "Unsupervised cross-media retrieval using domain adaptation with scene graph," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Nov. 15, 2020, doi: 10.1109/TCSVT.2019.2953692.

[10] M. Qi, Y. Wang, A. Li, and J. Luo, "Sports video captioning via attentive motion representation and group relationship modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2617–2633, Aug. 2020.

[11] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. ECCV*, Dec. 2016, pp. 852–869.

[12] H. Zhang, Z. Kyaw, J. Yu, and S.-F. Chang, "PPR-FCN: Weakly supervised visual relation detection via parallel pairwise R-FCN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4233–4241.

[13] Y. Li, W. Ouyang, X. Wang, and X. Tang, "ViP-CNN: Visual phrase guided convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7244–7253.

[14] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1974–1982.

[15] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017.

[16] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[17] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, p. 5.

[18] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. CVPR*, Jun. 2011, pp. 1745–1752.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[20] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards context-aware interaction recognition for visual relationship detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 589–598.

[21] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[22] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang, "Exemplar-based recognition of human–object interactions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 647–660, Apr. 2016.

[23] Z. Wang, S. Liu, J. Zhang, S. Chen, and Q. Guan, "A spatio-temporal CRF for human interaction understanding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1647–1660, Aug. 2017.

[24] Y. Zhan, J. Yu, T. Yu, and D. Tao, "On exploring undetermined relationships for visual relationship detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5128–5137.

[25] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, "Relationship proposal networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5678–5686.

[26] N. Xu, A.-A. Liu, Y. Wong, W. Nie, Y. Su, and M. Kankanhalli, "Scene graph inference via multi-scale context modeling," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Apr. 28, 2020, doi: 10.1109/TCSVT.2020.2990989.

[27] F. Zhao, J. Li, J. Zhao, and J. Feng, "Weakly supervised phrase localization with multi-scale anchored transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5696–5705.

[28] J. Mun, M. Cho, and B. Han, "Local-global video-text interactions for temporal grounding," in *Proc. CVPR*, Jun. 2020, pp. 10810–10819.

[29] L. Zhou, J. Zhao, J. Li, L. Yuan, and J. Feng, "Object relation detection based on one-shot learning," 2018, *arXiv:1807.05857*. [Online]. Available: http://arxiv.org/abs/1807.05857

[30] H. Zhou, C. Hu, C. Zhang, and S. Shen, "Visual relationship recognition via language and position guided attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2097–2101.

[31] X. Geng, "Label distribution learning," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1734–1748, Jul. 2016.

[32] K. Yi and J. Wu, "Probabilistic end-to-end noise correction for learning with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7017–7025.

[33] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

[34] L. Qi, L. Wang, J. Huo, Y. Shi, and Y. Gao, "Progressive cross-camera soft-label learning for semi-supervised person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2815–2829, Sep. 2020.

[35] J. Zhao *et al.*, "Multi-prototype networks for unconstrained set-based face recognition," 2019, *arXiv:1902.04755*. [Online]. Available: http://arxiv.org/abs/1902.04755

[36] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[37] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[38] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[40] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," 2015, *arXiv:1511.05493*. [Online]. Available: http://arxiv.org/abs/1511.05493

[41] Z. Cui, C. Xu, W. Zheng, and J. Yang, "Context-dependent diffusion network for visual relationship detection," in *Proc. ACM Multimedia Conf. Multimedia Conf. - MM*, 2018, pp. 1475–1482.

[42] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3298–3308.

[43] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 848–857.

[44] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable Net: An efficient subgraph-based framework for scene graph generation," in *Proc. ECCV*, Sep. 2018, pp. 335–351.

[45] D. Chen, X. Liang, Y. Wang, and W. Gao, "Soft transfer learning via gradient diagnosis for visual relationship detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1118–1126.

[46] G. Yin *et al.*, "Zoom-Net: Mining deep feature interactions for visual relationship recognition," in *Proc. ECCV*, Sep. 2018, pp. 322–338.

[47] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proc. AAAI*, 2019, pp. 8102–8109.

[48] L. Mi and Z. Chen, "Hierarchical graph attention network for visual relationship detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13886–13895.

[49] Y. Zhan, J. Yu, T. Yu, and D. Tao, "Multi-task compositional network for visual relationship detection," *Int. J. Comput. Vis.*, vol. 128, no. 8, pp. 2146–2165, 2020.

[50] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[52] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Weakly-supervised learning of visual relations," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5179–5188.

[53] X. Yang, H. Zhang, and J. Cai, "Shuffle-then-assemble: Learning object-agnostic visual relationship features," in *Proc. ECCV*, Sep. 2018, pp. 36–52.

[54] C. Han, F. Shen, L. Liu, Y. Yang, and H. T. Shen, "Visual spatial attention network for relationship detection," in *Proc. ACM Multimedia Conf. Multimedia Conf. - MM*, 2018, pp. 510–518.

[55] Y. Zhu and S. Jiang, "Deep structured learning for visual relationship detection," in *Proc. AAAI*, 2018, pp. 7623–7630.

[56] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

Hao Zhou received the B.S. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2017. He is currently pursuing the Ph.D. degree with Shanghai Jiao Tong University, Shanghai, China. His research interests include visual relationship detection, video action localization, and visual understanding.

Chongyang Zhang (Member, IEEE) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2008. He is currently a Professor with the Institute of Image Communication and Network Engineering, Department of Electronic Engineering, Shanghai Jiao Tong University. His research interests are in the area of machine learning and computer vision, especially on object detection, crowd counting, action recognition, and event detection. He has published over 50 international journals or conference papers on these topics.

Muming Zhao received the B.E. degree from Xidian University, Xi'an, China, in 2013, and the dual Ph.D. degree from Shanghai Jiao Tong University (SJTU), Shanghai, China, and the University of Technology Sydney (UTS), Sydney, Australia, in 2020. She is currently a Postdoctoral Fellow with the Robotics and Autonomous Systems Group, CSIRO Data61. Her research interests include machine learning and computer vision.

Yan Luo (Graduate Student Member, IEEE) received the B.S. degree in information engineering from Southeast University, China. She is currently pursuing the Ph.D. degree in information and communication Engineering with Shanghai Jiao Tong University, China. Her research lines are focused on pedestrian detection, pattern recognition, machine learning, computer vision, and intelligent transportation systems.

Chuanping Hu received the Ph.D. degree from Tong Ji University, Shanghai, China, in 2007. He was a Research Fellow and the Director of the Third Research Institute of the Ministry of Public Security, China. He is currently working with Zhengzhou University and is also a specially appointed Professor in Shanghai Jiao Tong University, Shanghai, China. He has authored more than 20 articles, has edited five books, and is the holder of more than 30 authorized patents. His research interests include machine learning, computer vision, and intelligent transportation systems.