Group Visual Relation Detection

Fan Yu, *Student Member, IEEE*, Beibei Zhang, *Student Member, IEEE*, Tongwei Ren, *Member, IEEE*, Jiale Liu, Gangshan Wu, *Member, IEEE*, and Jinhui Tang, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a novel visual relation detection task, named Group Visual Relation Detection (GVRD), for detecting visual relations whose subjects and/or objects are groups (GVRs), inspired by the observation that groups are common in image semantic representation. GVRD can be deemed as an evolution over the existing visual relation detection task that limits both subjects and objects of visual relations as individuals. We propose a Simultaneous Group Relation Prediction (SGRP) method that can simultaneously predict groups and predicates to address GVRD. SGRP contains an Entity Construction (EC) module, a Feature Extraction (FE) module, and a Group Relation Prediction (GRP) module. Specifically, the EC module constructs instances, group candidates, and phrase candidates; the FE module extracts visual, location and semantic features for these entities; and the GRP module simultaneously predicts groups and predicates, and generates the GVRs. Moreover, we construct a new dataset, named *COCO-GVR*, to facilitate solutions to GVRD task, which consists of 9,570 images from *COCO* dataset and 31,855 manually labeled GVRs. We test and validate the performance of SGRP by extensive experiments on *COCO-GVR* dataset. It shows that SGRP outperforms the baselines generated from the state-of-the-art visual relation detection and scene graph generation methods.**

***Index Terms*—Visual relation detection, group visual relation, entity construction, group prediction, predicate prediction.**

I. Introduction

Visual relation detection (VRD) aims to explore interactions between instances and represent them with triplets in the form of <subject, predicate, object> [1], which acts as the foundation of many visual understanding tasks, such as visual captioning [2], visual dialogue [3] and visual referring [4]. Current VRD methods [1], [5], [6] can only handle the visual relations whose subjects and objects are both individuals, *i.e.*, a single instance, which limits the effectiveness of VRD in representing the images with complex content. As shown in Fig. 1(a), VRD requires repetitive visual relations to represent "persons stand at dining table", "persons cut cake" and "cups on dining table". However, these repetitive visual relations can be represented in a more compact way by merging the persons and cups into groups (Fig. 1(b)). To simplify relation triplet representation, we represent a group of instances in a specific category with a star label, *e.g.*, person* denotes a group of person. Furthermore, to represent the content of Fig. 1(c), VRD can only predict the intuitive predicate "beside" between chairs and the dinging table, but it cannot provide more inductive predicate "surround" by treating all the chairs as a group (Fig. 1(d)). In this paper, we name a set of instances (no less than two instances) with similar characteristics and same visual relation(s) to other instance(s) as a "*group*" in VRD, and propose a new VRD task named *Group Visual Relation Detection* (GVRD), which aims to detect the visual relations whose subjects and/or objects are groups (GVRs). To distinguish GVRD from the existing VRD task, we refer to the relations whose subjects and objects are both individuals as *individual visual relations* (IVRs) and the VRD task only detecting IVRs as *individual VRD* (IVRD). GVRD addresses the limitations of IVRD by providing a more comprehensive representation of image content. The key advantage of GVRD lies in its ability to capture contextual and holistic relationships among multiple entities and their collective structure. This allows for a deeper understanding of complex scenes where meaning arises from the interactions and arrangements of multiple objects, rather than just pairwise relationships. On one hand, GVRs encompass relationships that cannot be represented in IVRs. For instance, the relationship <person*, surround, dining table> cannot be substituted with multiple IVR triplets like <person, surround, dining table> because an individual person cannot "surround" an object on their own. On the other hand, GVRs are not simply a clustering of IVRs. For example, <person*, compete with, person*> cannot be reduced to multiple <person, compete with, person> triplets, as not every individual in the subject group is necessarily

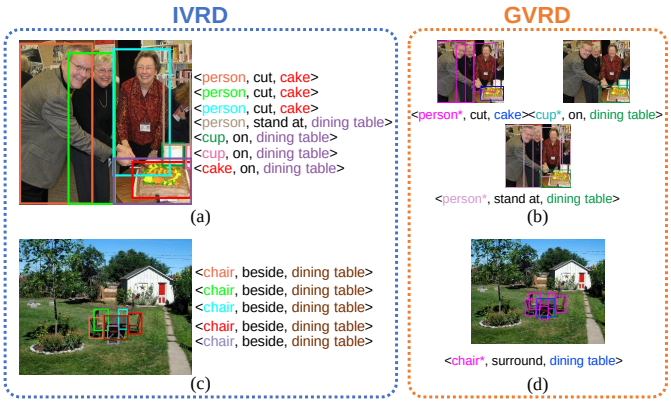


Fig. 1. Comparison between IVRD and GVRD, and some GVRD examples. GVRD can represent image content with more compact and inductive visual relations.

This work is supported by National Natural Science Foundation of China (62072232), the Fundamental Research Funds for the Central Universities (021714380026) and the Collaborative Innovation Center of Novel Software Technology and Industrialization. (Corresponding author: Tongwei Ren.)

Fan Yu, Beibei Zhang, Tongwei Ren, Jiale Liu and Gangshan Wu are with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China (e-mail: {yf, zhangbb}@smail.nju.edu.cn; rentw@nju.edu.cn; licsber@gmail.com; gswu@nju.edu.cn).

Jinhui Tang is with the Nanjing University of Science and Technology (jinhuitang@njust.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes a comprehensive list of symbols, along with extended details pertaining to the dataset and experiments. Contact yf@smail.nju.edu.cn or rentw@nju.edu.cn for further questions about this work.

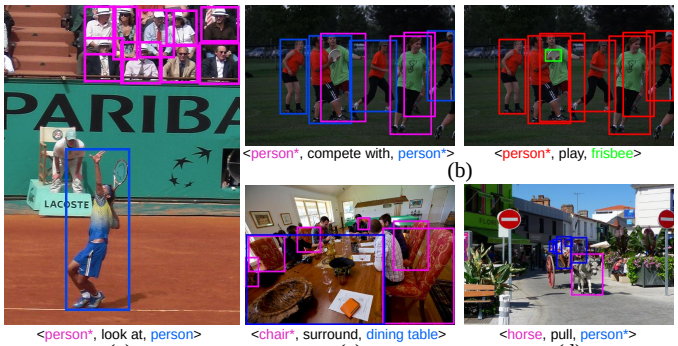


Fig. 2. Some examples of GVRs with group subjects and/or group objects.

competing with every individual in the object group.

Groups of human have already been studied in group activity recognition (GAR) [7], which is evolved from single-person action recognition and two persons' interaction recognition. A group activity (also called collective activity) is performed by multiple persons simultaneously. It has to be noted that, GAR only focuses on human groups and targets at the intra-group activities in videos, while GVRD requires to recognize groups of all categories and the relations between instances and groups or different groups, *i.e.*, inter-group relations. Conceptually, the core of group activity recognition lies in understanding contextual interaction among individuals within a scene. It is also necessary to reason upon the interactions of different individuals. However, it is used to learn a consistent group activity among the individuals, not the relationship between different groups. First of all, GVRD not only focuses on human groups, but considers different kinds of instances. For example, <person*, surround, dining table> is the relationship between a human group and a dining table, and <chair*, surround, dining table> is the relationship between a chair group and a dining table. In addition, group division is a part in GVRD, and GVRD needs to detect many different inter-group relationships, while existing works on GAR terms to recognize one group activity in one time.

Following the definition of VRD task, GVRD requires to generate relation triplets and localize the subjects and the objects of the detected relation triplets. In subject/object localization, there are two reasonable choices: one is to use a bounding box containing all the instances in the group, and the other is to use a set of bounding boxes of all the instances in the group. To provide more accurate localization, We use bounding box set instead of a single bounding box to represent the locations of group subjects and group objects in GVRD. Figure 2 shows some examples of GVRD, which can be categorized into three types, only group subject <subject*, predicate, object> (Fig. 2(a)-(c)), only group object <subject, predicate, object*> (Fig. 2(d)), and both group subject and group object <subject*, predicate, object*> (Fig. 2(b)). In particular, the relations between/among the instances within a group are not included in GVRD because they are more similar to the attributes of individuals in IVRD.

Compared with IVRD, GVRD has extremely more subject/object candidates due to explosive combination possibility in group generation. Supposing that the number of the detected instances belonging to the k^{th} category is n_k, there are $\sum_k n_k$ subject/object candidates in IVRD and $\sum_k 2^{n_k}$ subject/object candidates in GVRD. Moreover, the composition of groups may be changeable in different GVRs, *e.g.*, the players are considered as a group in the GVR <person*, play, frisbee> but they are divided into two groups in the GVR <person*, compete with, person*> in Fig. 2(b). Therefore, a demand is posed for GVRD that group prediction and predicate prediction should be simultaneously processed.

A straightforward solution of GVRD is to cluster the detected relation triplets by IVRD methods, *e.g.*, combining the subjects/objects in the same category whose predicate is identical to that of an object/subject into a group subject/object. However, the combination of IVRD and clustering cannot effectively handle GVRD. Firstly, the similarity measurement for determining the instances belonging to a group is neither simple nor explicit. For example, the referee and the audience both look at the player but they cannot be treated as a group in Fig. 2(a) because they have different identities and positions, while the players in the same team in Fig. 2(b) should be treated as a group despite their separate positions. Secondly, the relations among groups may not be detected between all the instances within the two groups in IVRD. For example, two teams with four players (red clothes) and three players (green clothes) compete with each other in Fig. 2(b), but it is hard to detect all the 12 visual relations in the form of <person, compete with, person> in IVRD because many players in two teams are standing apart and they do not have direct interactions. Thirdly, some predicates are intrinsically related to groups and only exist in GVRD. As shown in Fig. 2(c), <chairs, surround, dining table> is detected in GVRD, but only <chair, beside, dining table> can be detected in IVRD.

In this paper, we propose a novel GVRD method named *Simultaneous Group Relation Prediction* (SGRP), which simultaneously predicts groups and predicates in GVRs. SGRP is proposed on the basis of an observation: the range of a group is not absolute but related to another instance or group and the relationships between them. Thus, we consider that a group can be confirmed only when a GVR related to it is confirmed. Also, generating several group candidates first and confirming their components later significantly reduce the number of candidates. SGRP contains three modules: an Entity Construction (EC) module, a Feature Extraction (FE) module and a Group Relation Prediction (GRP) module. Specifically, EC module constructs all the entities for GVRD, *i.e.*, instances, group candidates from instances, and phrase candidates, by pairing group candidates. FE module extracts features for entities, *i.e.*, the visual and location features of instances, visual, location and semantic features of group candidates, and visual features of phrase candidates. GRP module simultaneously predicts groups and predicates of GVRs using entity features, and generates the final GVRs. The Group Candidate Construction (GCC) and Phrase Candidate Construction (PCC) sub-module in the FE module and the Group Prediction (GP) branch and Predicate Prediction (PP) branch in the GRP module work together to distinguish GVRs from IVRs.

To the best of our knowledge, GVRD is a new task

without existing dataset. Hence, we construct a new dataset *COCO-GVR* on the basis of the *COCO* dataset [8]. We first filter the images that do not contain GVRs according to the instance number of each class in each image. Every GVR in a candidate image is manually annotated by selecting instances in subject and/or object groups and choosing or typing predicate between subject and object. The annotation data are then post-processed for correction and completion. The *COCO-GVR* dataset contains 9,570 images with 31,855 GVRs in total and it is splited into a training set with 8,056 images and a test set with 1,514 images.

In summary, our contributions are threefold: 1) We propose a new GVRD task aiming to detect visual relations related to group in a given image, capturing more complex aspects of group-level content understanding. 2) We propose a novel SGRP method to simultaneously predict group and group related visual relations, *i.e.*, group confirmation and predicate prediction influence each other, which strengthens the robustness of our SGRP method and makes a more accurate prediction of group visual relations. 3) We construct the first dataset *COCO-GVR* for GVRD evaluation, which consists of 9,570 images from *COCO* dataset and 31,855 manually labelled GVRs.

II. Related Works

Visual Relation Detection and Scene Graph Generation. VRD is first formally introduced by Lu *et al.* [1] to explore interactions between pairs of objects in a given image, along with a method leveraging language priors. Subsequent research [5], [6], [9] has built upon this idea, further incorporating linguistic information. Xu *et al.* [10] propose a related task named SGG, which represents image content as a graph, where nodes correspond to objects with attributes, and edges denote relationships between them. Alongside the introduction of SGG, a message-passing-based method was proposed. Building on this concept, several approaches [11]–[14] have employed message passing to refine node and edge predictions by extracting visual features from object proposals and improving inference quality through graph-based reasoning.

Different structures are exploited to encode the context for VRD and SGG, such as graph [13], [15], tree [16], Transformer [17]–[19] and attention mechanism [20]. As many prototypes regularly appear in visual relations, which comply with the rules of human natural language, internal knowledge, *i.e.*, common sense, is applied to assisting visual relations prediction [11], [21], [22]. However, in recent years the opinion that methods for VRD and SGG should focus on "learning" to "infer" visual relations rather than fit the statistical bias has received much attention. Thus, an increasing number of methods [13], [23]–[27] pay attention to unbiased inference and one-shot learning. Further, informative and fine-grained predicates are explored to detect implicit visual relationships [28], [29]. Unseen relation detection is also explored to address the difficulty in obtaining visual relation annotations [12], [30]. Moreover, some recent research starts to focus on semantically important visual relations [31], [32], uncertainty visual relations [33], and incremental visual relations [34]. With the development of fundamental models, a unified framework [35] is proposed to predict scene graphs and a similar task, human-object interactions [36]–[41]. In addition, visual relationship detection is extending to video understanding [42]. Furthermore, VRD and SGG can be used in many applications. In remote sensing area, scene graph generation is explored for [43]. As an extension of VRD and SGG, social relation detection focuses on the abstract relationship between people [44], [45].

There are several datasets are constructed for VRD and SGG. The first dataset for VRD and SGG is the *VRD* dataset [1], which contains 5,000 images with 100 object categories and 70 predicates. The largest dataset for VRD and SGG is the *Visual Genome (VG)* dataset [46], the latest version (1.4) of which contains 108,077 images with 82,827 object categories and 37,342 predicates. Since the objects and relations in *VG* are noisy, *VG150* [10] and *VG200* [47] are constructed to clean up the annotations in VG. *VG150* contains 108,077 images with 150 object categories and 50 predicates, while *VG200* contains 99,658 images with 200 object categories and 100 predicates. On the basis of *VG*, *GQA* [48], *UnRel* [30] and *VrR-VG* [49] are constructed for further visual relationship reasoning. The *ViROI* [31] dataset and the *VG-KR* [32] dataset are constructed for semantically important relation prediction, on the basis of *COCO* and *VG*, respectively. Additionally, PSG [50] is constructed for panoptic scene graph generation.

Compared to VRD and SGG, which are limited to handling individual subjects and objects, GVRD focuses on detecting visual relations involving subjects and/or objects in group formations, offering more compact representation of visual relations and predicting intuitive predicates that elude inference from individual visual relations. Existing VRD and SGG datasets do not support GVRD evaluation, making COCO-GVR the first dataset specifically designed for this task.

Group Activity Recognition. GAR aims to detect activities performed by multiple persons and some work on group activity also needs to distinguish different groups. GAR evolves from action recogition and focuses on "multiperson action" recognition [51], which does not explicitly recognize relations between person groups. Choi *et al.* [52] introduce a paradigm where actions are recognized with the context, *i.e.,* what other humans are doing in the scene, and construct a new dataset. Amer *et al.* [53] address a new problem to detect and localize a wide range of activities, and Ibrahim *et al.* [54] build another volleyball dataset with preson detections, person action labels and group activity labels. The traditional methods are based on hand-crafted features with probabilistic graphical models or AND-OR grammar methods [55]. With the development of deep learning, convolutional neural networks [54], recurrent neural networks [56] and graph convolution networks [57]–[59] are proposed to encode context features for GAR. Since the annotation for GAR requires expensive cost, weakly-supervised learning [60]–[62] and self-supervised learning [63] are also explored.

Several datasets have been developed to evaluate the performance of GAR methods. Collective Activity Dataset (CAD) [52] comprises 44 short video sequences

(approximately 2,500 frames) capturing five group activities—crossing, waiting, queueing, walking, and talking—and six individual actions (NA, crossing, waiting, queueing, walking, and talking). The group activity label for each frame is determined by the activity in which the majority of individuals are engaged. Volleyball Dataset [54] includes 4,830 clips extracted from 55 volleyball games, with 3,493 clips for training and 1,337 for testing. Each clip is annotated with one of eight group activity labels: right set, right spike, right pass, right winpoint, left set, left spike, left pass, and left winpoint. Social-CAD [64] enhances the original CAD by providing annotations for subgroup activity recognition. Instead of assigning a single group activity label to an entire scene, Social-CAD identifies distinct social groups and assigns corresponding activity labels to each group.

Unlike GAR, which focuses exclusively on human groups and aims to identify intra-group activities in videos, GVRD takes a more expansive approach by recognizing groups across all categories and identifying relationships both between individual instances and groups as well as among different groups. From a methodological perspective, while GAR does involve reasoning about individual interactions, this is done with the aim of deriving a consistent group activity, rather than establishing relationships between distinct groups. In contrast, GVRD is broader in scope. It goes beyond human groups to encompass various types of instances, and group identification is only one aspect of GVRD. It also requires detecting a diverse range of inter-group relationships. GAR datasets are typically focused on recognizing a single group activity at a time.Though more recent datasets have introduced annotations for subgroup activities in videos, these datasets remain centered on activities within human groups. The dataset for GVRD aims to capture a broader range of relationships, encompassing various types of groups and instances. Thus, datasets for GAR are not suitable for the GVRD task.

III. Dataset

To support the evaluation of the GVRD task, we construct the first dataset, named *COCO-GVR* on the basis of the *COCO* dataset [8], a large-scale dataset for object detection, segmentation, and captioning. *COCO-GVR* is constructed by labeling group relation triplets implied in image. First of all, we filter the images and keep those containing at least a group and another instance, which may form a candidate GVR. For convenient annotation, we provide some possible predicates from the VG [46] dataset. We select the overlapping images of the *VG* dataset and the *COCO* dataset and map the instances in the *COCO* dataset to those in *VG* by matching classes and calculating the interaction of union (IoU) of the bounding boxes provided by the two datasets. Then we extract some triplets in the form of <*COCO* instance category, *VG* predicate, *COCO* instance category> and show possible predicates after annotators confirm the subject and object in a GVR.

We develop an annotation tool with an interface shown in Fig. 3. The topmost area shows all the instance categories in the current image. Below the category words, the left area is

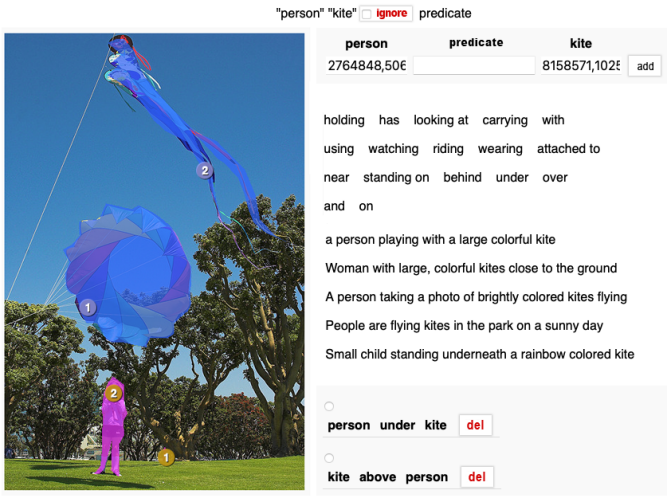


Fig. 3. An example of our annotation tool for GVR annotation.

an image with corresponding instance segmentations, and the right area includes possible predicates, captions for reference, and the group relation triplets that have been annotated. All the instances of one category are highlighted in the image after the category word in the topmost area is chosen, and annotators can further choose instances to construct the group by clicking the corresponding segmentations. When the annotation for subject and object is completed, the tool shows possible predicates from the VG according to categories of subject and object. Annotators can also choose words from the captions or type in other possible predicates. For additional context, 23 annotators were involved in the annotation process, which spanned five days. The images were evenly distributed among the annotators, and three of the authors also contributed to the annotation. To establish clear standards for data annotation, the authors pre-annotated hundreds of examples as references prior to the start of the process. These examples served as training materials for all annotators, who were instructed to label as many group relationships as possible to minimize ambiguities in the annotations. Throughout the annotation process, a portion of the annotated data was randomly sampled each day for review. Feedback was promptly provided to annotators to address any identified issues and ensure consistent quality.

Afterwards, we perform the following four processes for data cleaning: 1) Manually correct the misspellings, lemmatize predicates and merge synonyms. 2) Add new relations by reversing the position of subject and object with opposite predicates according to existing relations (*e.g.*, "kite above person" and "person under kite"). 3) Remove the relations that do not satisfy the definition of GVRs, *i.e.,* filter those whose subject and object are individual instances. 4) Remove the relations whose subject and object contain the same instances.

The *COCO-GVR* dataset consists of 9,570 images with 31,855 annotated GVRs. Each instance is represented with its category, bounding box and segmentation, and each GVR is represented with its subject instances, predicate and object instances. The *COCO-GVR* dataset contains 80 categories of instances, 96 categories of predicates, including 39 verbs, 3

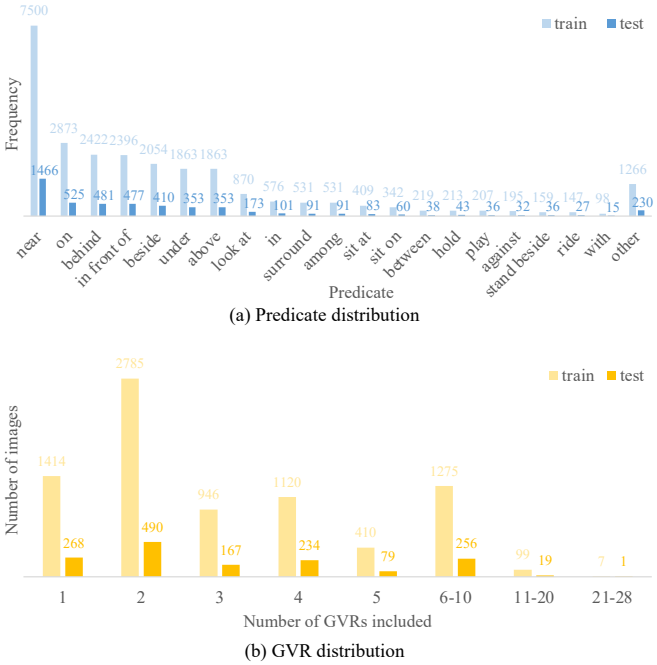


Fig. 4. Data distribution in the *COCO-GVR* dataset.

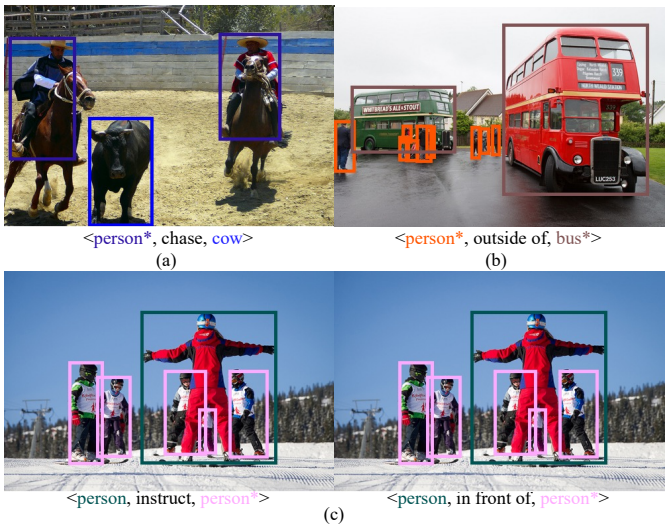


Fig. 5. Examples of noise handling in the *COCO-GVR* dataset.

prepositions, 17 spatial and 37 preposition phrases, and 4,610 unique GVRs in total. The *COCO-GVR* dataset is divided into a training set and a test set with similar distribution of predicates and GVRs (shown in Fig. 4), containing 8,056 images with 26,734 GVRs (3.32 GVRs per image) and 1,514 images with 5,121 GVRs (3.38 GVRs per image), respectively.

Relation annotation in GVRD, like IVRD, is susceptible to supervisory noise, which may arise from uncertainty, incorrect label assignments, or a coarse characterization of labels. As illustrated in Fig. 5(a), the relationship between the two people and the cow is labeled as "chase", but it could easily be misidentified as "drive out". This ambiguity arises from contextual uncertainties or occlusions. To mitigate such errors, we incorporate contextual understanding and common knowledge during data annotation. The hierarchical composition of groups and relations can also introduce noise. For instance, in Fig. 5(b), the relation <person*, outside of, bus*> might be split into two distinct annotations: <(some) person*, towards, (green) bus> and <(some) person*, towards, (red) bus>. Sometimes, even with contextual cues and common knowledge, it is difficult to determine which individuals are heading towards the green bus and which are heading towards the red bus. In such cases, we opt for a "not-wrong" choice, where <person*, outside of, bus*> remains a valid and unambiguous annotation. Additionally, we account for scenarios in which a single pair of entities may exhibit multiple predicates simultaneously. To address this, we provide multi-label annotations for subject-object pairs with multiple possible relationships, as shown in Fig. 5(c). The COCO-GVR dataset is released at https://magus.ink/resource, and it will be continuously refined according to use feedbacks.

IV. Method

Figure 6 shows the framework of our SGRP method. In EC module, we detect the instances and further construct the group candidates by combining instances and the phrase candidates by pairing of groups and instances/groups. In FE module, we extract features for the instances, the group candidates and the phrase candidates, including visual features, location features and semantic features. In GRP module, we simultaneously predict the final groups and the predicates with two branches, GP and PP, and generate the GVRD results. SGRP predicts groups and relationships interdependently instead of detecting groups first and then recognizing relationships between them. The GCC sub-module constructs group candidates and the final groups are confirmed in GP branch. The GP branch predicts the confidence of each instance, which represents the probability whether an instances belongs to a group candidate, and the PP branch exploits the instance confidence in predicate prediction. Hence, loss for predicate prediction can also used in optimizing weights for group prediction. Moreover, predicate prediction is not limited by final group confirmation but adapts the weights of different instance features according to instance confidence, which strengthens the robustness.

A. Entity Construction

Given an image, we firstly use a traditional object detector consisting of ResNet101, FPN and RPN to detect instances and generate instance raw visual features. The key challenge in group detection is managing the vast number of potential group candidates generated when dividing instances into groups. To tackle this, we generate group candidates using instance-similarity thresholds, which effectively reduce the number of candidates. The process of GCC and PCC is illustrated in Fig. 7. GCC relies on the similarity measurements between instances. Specifically, the similarity between instance

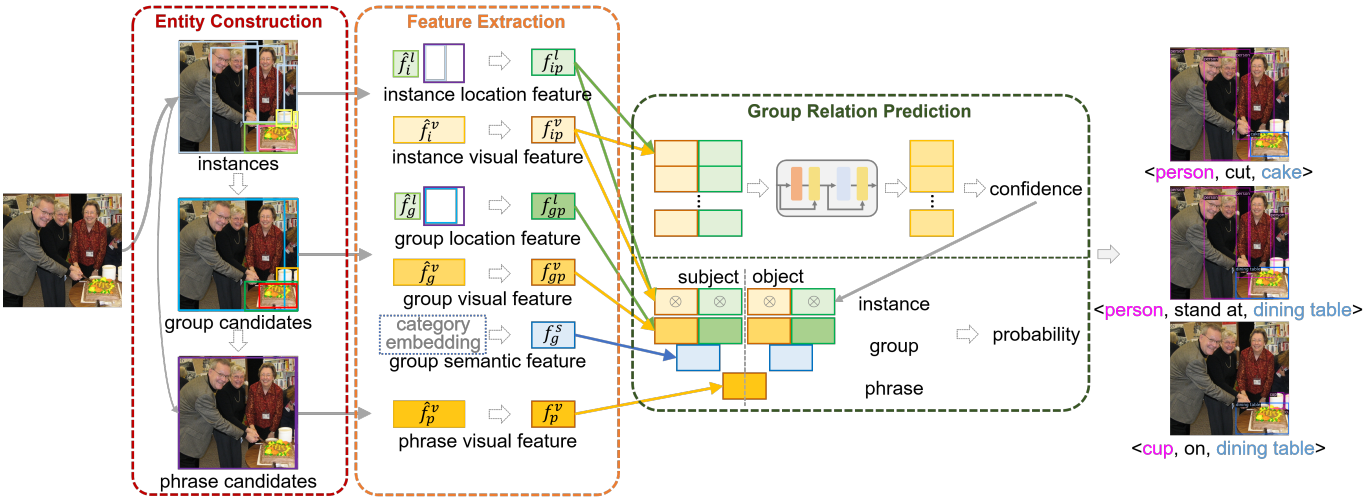


Fig. 6. An overview of our SGRP method. It contains three modules, Entity Construction module, Feature Extraction module, and Group Relation Prediction module.

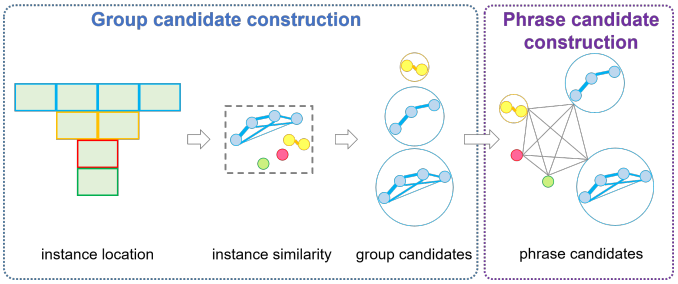


Fig. 7. The detailed process for constructing group candidates and phrase candidates in EC module.

pairs is computed using the normalized cosine values of their global location features:

$$\mathbf{S}_{ij} = \frac{\bar{f}_i^l \cdot \bar{f}_j^l}{\| \bar{f}_i^l \|_2 \cdot \| \bar{f}_j^l \|_2}, \qquad \bar{\mathbf{S}}_{ij} = \frac{\mathbf{S}_{ij} - \mathbf{S}_{\min}}{\mathbf{S}_{\max} - \mathbf{S}_{\min}}, \tag{1}$$

where $\bar{f}_i^l$ and $\bar{f}_j^l$ are normalized global location features of instance i and j, *i.e.*, concatenated coordinates of top-left corner and bottom-right corner of the bounding box after being normalized with image width and height, $\mathbf{S}_{ij}$ represents the instance cosine-similarity between i and j, $\bar{\mathbf{S}}_{ij}$ is the similarity between i and j after normalization, $\mathbf{S}_{\max}$ and $\mathbf{S}_{\min}$ represent the maximum and minimum values in the similarity matrix, respectively. Since the similarity matrix also computes the similarity between identical instances, the value of $\mathbf{S}_{\max}$ is always 1. It is worth noting that, group candidates are constructed in GCC based on relative location similarity instead of vision similarity, because the composition of groups are usually related to spatial position while instances in different groups could be quite similar in vision.

Since instance similarity is inherently symmetric, we construct a undirected graph by treating each instance as a node and weight the edge between a pair of nodes according to the similarity between the corresponding instances. To generate various group candidates, we binarize the edge weights in the undirected graph with different thresholds, which are 0.5 and 0 in our experiments, and remove the edges with zero weights. In each binarized undirected graph, we treat each maximal connected subgraphs as a group candidate. Thus, instances belonging to the same category can only be divided into no more than two group candidates, thereby eliminating the risk of combinatorial explosion. Finally, in PCC submodule, we construct phrase candidates by combining two group candidates or an instance and a group candidate.

Training Details. Generation of training samples in GVRD are more complicated than that in IVRD. In IVRD, training samples can be generated according to instance pairs. The pairs with relationships are positive samples and those without relationships are negative. However, there exists combination explosion in GVRD, and there must be a large amount of negative samples generated during training. Too many negative samples cause the quantity imbalance between positive samples and negative samples and also result in the dramatic increase of computation costs. In the follow-up modules, losses will be measured in phrase, thus phrase candidates are matched with groundtruth phrases to generate training samples. Though we have constructed group candidates according to instance similarity to limit the number, there still could be a lot of phrase candidates. During the training period, we use the groundtruth instances instead of the detected instances, which reduces the complexity and ambiguity in training sample generation. Assume $\tilde{i}$ is an instance in groundtruth, g is a constructed group candidate, p is a constructed phrase candidate, $\tilde{i}|g$ denotes $\tilde{i}$ belongs to g, and $g|p$ denotes that g belongs to p, *i.e.*, g is the group subject/object of p. $\widehat{P}$ is the set of positive phrase samples, which contains all phrase candidates whose subject and object both respectively cover no less than 60% of the subject and the object instances of at least one phrase in groundtruth. Instance $\tilde{i}|g|\widehat{p}$ in group candidate $g|\widehat{p}$ of positive phrase $\widehat{p}$ is assigned with label 1 if this instance truly belongs to $g|\widehat{p}$, and assigned with label 0 otherwise. We name the sub-group, which is composed by

instances with labels 1, groundtruth group $\tilde{g}$. The other phrases are negative samples and all instances in group candidates of negative phrases are assigned with label 0.

B. Feature Extraction

The features used in the subsequent module are categorized into three levels: object level, group level, and phrase level, and three types: visual, location, and semantic. The selection of these feature levels aligns with the structural components of GVRs. At the core, instances form the basic elements of GVRs, which are then grouped to create the subject and object of a relational phrase. This hierarchy makes it natural to define features at the instance, group, and phrase levels. Our choice of feature categories follows widely accepted practices in VRD and SGG research [6], [20], [47], [65]. Visual features are traditional and foundational in VRD [1] and SGG [10]. Semantic features were introduced when VRD is proposed [1] and have been extensively applied [23]. Location features implying the relative spatial information between subjects and objects are also a critical form of hidden knowledge and has been studied in many works [47], [66]. While adhering to these principles, we made specific adjustments based on the characteristics of each feature level. At group level, we use location, visual, and semantic features. This provides a comprehensive representation of the group. At object level, we rely on location and visual features but exclude semantic features. A group consists of objects from the same category, leading to a lack of distinctiveness in semantic features due to the uniformity within the group. At phrase level, only semantic features are applied. The location of the entire phrase, represented as the union box of all instances, is irrelevant for relationship prediction. The known semantics are derived from the subject group and object group categories, while the predicate semantics remain to be predicted. This structured approach ensures that the features are tailored to the specific requirements of each level, enhancing the effectiveness of GVRD.

Instance feature extraction. For each instance, we use *RoIAlign* [67] to extract its raw visual feature $\hat{f}_i^v$ from backbone feature maps generated by ResNet101, whose dimension is 1,024 in our experiments, and extract its raw location feature $\hat{f}_i^l$, *i.e.*, concatenated coordinates of the top-left corner and the bottom-right corner of the bounding box. Instance visual feature of instance i in phrase p, *i.e.*, f_{ip}^v, is generated by transforming $\hat{f}_i^v$ to the dimension of 512 linearly. Instance location feature of instance i in phrase p, *i.e.*, f_{ip}^l, is generated by normalizing $\hat{f}_i^l$ with width and height of phrase p and transforming the vector composed by the normalized coordinates to the same dimension of f_{ip}^v.

Group candidate feature extraction. We treat the minimal bounding box of all the instances belonging to a group candidate as the bounding box of the group candidate. Similar to instance feature extraction, we extract the visual feature and the location feature of a group candidate in a phrase by *RoIAlign* [67] and coordinate normalization, respectively, along with linear transformation. Moreover, we assign the category of the instances belonging to a group candidate to the group candidate, and extract the semantic feature of the group candidate by applying word embedding on its category with GloVe model [68]. To balance the effects of different features, we also transform the semantic features to the dimension of 512.

Phrase candidate feature extraction. We treat the minimal bounding box of the subject and the object of a phrase candidate as the bounding box of the phrase. Similar to instance and group feature extraction, we extract the visual feature of a phrase candidate by *RoIAlign* [67] with linear transformation to dimension of 512.

C. Group Relation Prediction

As shown in Fig. 8, we simultaneously predict groups and predicates with two branches GP and PP within each phrase candidate, and these two branches are interdependent in prediction.

In GP branch, the concatenated visual features and location features of all the instances belonging to the subject and the object of a phrase candidate are encoded for group prediction. Since instances in a group do not spontaneously form a sequence, and the range of a group is related to another instance/group in the relation, we encode the features of instances in groups by a cross-attention Transformer encoder. The traditional Transformer [69] encoder uses the self-attention mechanism, which calculates attention by matrix multiplication of input features. We calculate the cross-attention by matrix multiplication of subject instance features and object instance features in our cross-attention Transformer encoder:

$$A_{g^o \to g^s} = \text{softmax}\Big(\frac{F_{\mathcal{I}|g^s} \ F_{\mathcal{I}|g^o}^T}{\sqrt{d}}\Big), \tag{2}$$

where $A_{g^o \to g^s}$ is the attention matrix of subject group g^s from object group g^o, $F_{\mathcal{I}|g^s}$ is the combined visual and location feature matrix of all instances in group g^s, $F_{\mathcal{I}|g^o}^T$ is the combined visual and location feature matrix transpose of all instances in group g^o, and d is the instance number of each group after padding. We first calculate the intermediate feature matrix $F_{\mathcal{I}|g^s}^{'}$ of instances in the subject group:

$$F_{\mathcal{I}|g^s}^{'} = \eta(F_{\mathcal{I}|g^s} + A_{g^o \to g^s} \times F_{\mathcal{I}|g^o}), \tag{3}$$

where the feature matrix of instances in subject group before encoding $F_{\mathcal{I}|g^s}$ is updated by the product of feature matrix of instances in object group before encoding $F_{\mathcal{I}|g^o}$ and the attention of subject group g^s from object group g^o; η represents normalization. The final encoded feature matrix $F_{\mathcal{I}|g^s}^{''}$ of instances in the subject group is generated as follows:

$$F_{\mathcal{I}|g^s}^{''} = \eta\Big(F_{\mathcal{I}|g^s}^{'} + \tau\big(\iota(F_{\mathcal{I}|g^s}^{'})\big)\Big), \tag{4}$$

where τ is an activation function and ι is a linear transformation. The encoded feature matrix $F_{\mathcal{I}|g^o}^{''}$ of instances in the object group is generated similarly. We further predict the confidence of the instance belonging to the group candidate:

$$\gamma_{i|g} = \text{sigmoid}\big(\mu(f_{i|g}^{''})\big), \tag{5}$$

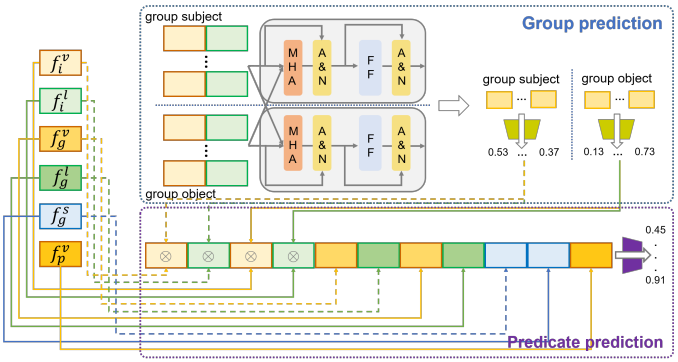


Fig. 8. The details of the GP branch and the PP branch in GRP module.

where $f''_{i|g}$ denotes the encoded feature of instance i which belongs to group candidate g after the cross-attention Transformer encoding; $\gamma_{i|g}$ is the confidence that instance i belongs to group candidate g; μ is a linear transformation to reduce the dimension to 1. If $\gamma_{i|g}$ is lower than a threshold, which equals 0.5 in our experiments, instance i will be removed from group candidate g. In this way, we can update the group candidates, and generate the final groups.

In PP branch, we use the phrase feature and all the features of the instances and the group candidates belonging to the phrase. Specifically, to adapt the influence of different instances in group candidates, the visual features and the location features of all the instances belonging to the subject and the object are weighted averaged by their confidences:

$$f^{\#}_{\mathcal{I}|g} = \frac{1}{\mid \mathcal{I}|g \mid} \sum_{i \in \mathcal{I}|g} \gamma_{i|g} \cdot f^{\#}_{i|g}, \tag{6}$$

where $\mathcal{I}|g$ denotes the set of instances belonging to group g; $f^{\#}$ denotes visual feature f^v or location feature f^l generated by FE module; $\mid . \mid$ denotes the set cardinality; the weight $\gamma_{i|g}$ is calculated according to Eq. (5).

We concatenate instance visual features ($f^v_{\mathcal{I}|g^s}$ and $f^v_{\mathcal{I}|g^o}$), instance location features ($f^l_{\mathcal{I}|g^s}$ and $f^l_{\mathcal{I}|g^o}$), group visual features ($f^v_{g^s}$ and $f^v_{g^o}$), group location features ($f^l_{g^s}$ and $f^l_{g^o}$), group semantic features ($f^s_{g^s}$ and $f^s_{g^o}$), and the phrase visual feature (f^v_p), and predict a probability vector β by linear transformation and the sigmoid activation function. β represents the probabilities of all the predicate to the phrase, and it contains a no-relation element to represent that there is no suitable predicates between the subject and the object. To handle the data imbalance in GVRD, we adjust the β according to dataset statistics [70] during inference. Finally, some generated GVRs are merged if they have the same predicate and their subject/object instance sets are overlapped.

Training Details. In GP branch training, we only calculate the loss of group prediction for groups of phrases in $\widehat{P}$. As mentioned in Section IV-A, we set the groundtruth confidence $\tilde{\gamma}_{\tilde{i}|\tilde{g}}$ to 1 if groundtruth instance $\tilde{i}$ belongs to groundtruth group $\tilde{g}$ and 0 otherwise. The loss function of GP branch is as follows:

$$\mathcal{L}_{GP} = \frac{1}{\mid \widehat{P} \mid} \sum_{p \in \widehat{P}} \sum_{g|p} \frac{1}{\mid \mathcal{I}|g \mid} \Big(\sum_{\tilde{i} \in I|g} \xi(\gamma_{\tilde{i}|g}, \tilde{\gamma}_{\tilde{i}|\tilde{g}}) \Big), \tag{7}$$

where $\mathcal{I}|g$ is the set of all the instances belonging to g; ξ is the binary cross entropy function; $\mid . \mid$ denotes the set cardinality. Moreover, we use features of groundtruth instances in the groundtruth object group $F_{\widetilde{\mathcal{I}}|\tilde{g^o}}$ to generate the encoded features of instances in the subject group $F''_{\mathcal{I}|g^s}$, and use features of groundtruth instances in the groundtruth subject group $F_{\widetilde{\mathcal{I}}|\tilde{g^s}}$ to generate the encoded features of instances in the object group $F''_{\mathcal{I}|g^o}$ in training.

In PP branch training, we generate a groundtruth predicate probability vector $\tilde{\beta}_p$ for each positive phrase candidate sample p in $\widehat{P}$ by setting the vector elements, which represent predicate categories, to 1 if there are the corresponding GVRs in groundtruth and 0 otherwise. For a negative phrase candidate sample not in $\widehat{P}$, we also generate its groundtruth predicate probability vector $\tilde{\beta}_p$ by setting all the elements to 0 except the no-relation element, which is set to 1.

The loss function of predicate prediction is as follows [31]:

$$\mathcal{L}_{PP} = \frac{1}{|P|} \sum_{p \in P} \mathcal{L}_p, \tag{8}$$

where P is the set of all phrase candidate samples; $\mathcal{L}_p$ denotes the loss of of phrase p, which is calculated as follows:

$$\begin{aligned} \mathcal{L}_p = &- \frac{\sum_k \tilde{\beta}_k (1-\beta_e)^2 \log(\beta_e)}{\sum_k \tilde{\beta}_k} \\ &- \frac{\sum_k (1-\tilde{\beta}_k) \beta_e \log(1-\beta_e)}{\sum_k (1-\tilde{\beta}_k)}, \end{aligned} \tag{9}$$

where β and $\tilde{\beta}$ denote the predicted predicate probability vector and groundtruth predicate probability vector of a phrase candidate, respectively; β_e is the e^{th} element of β. Here, we emphasize the loss penalty on the vector elements when their groundtruth values are 1 by considering such vector elements are more sparse in the training data [31].

The final loss is computed as the sum of the losses from the GP branch and the PP branch:

$$\mathcal{L} = \mathcal{L}_{GP} + \mathcal{L}_{PP}. \tag{10}$$

V. Experiments

A. Experimental Settings

Following the experimental setting of IVRD, we use recall and mean recall of top t results to evaluate the performance of SGRP for GVRD. Our method predicts a probability score for each triplet, the triplets are sorted according to the scores. All the experiments are conducted on the *COCO-GVR* dataset. Since there could be several relationships between the same subject and object, we use no graph constrains recall (R@t) and mean recall (mR@t) in our experiments. We set t to 10, 20 and 30 in evaluation, because 98.68% images have no more than 10 GVRs, 99.92% images have no more than 20 GVRs, and the maximal number of GVRs in an image in the *COCO-GVR* dataset is 28.

Similar to IVRD evaluation, we consider that a GVRD result is correct when the predicted relation triplet is the same as the groundtruth and the subject and object IoU between the GVRD result and the groundtruth are both larger than a threshold,

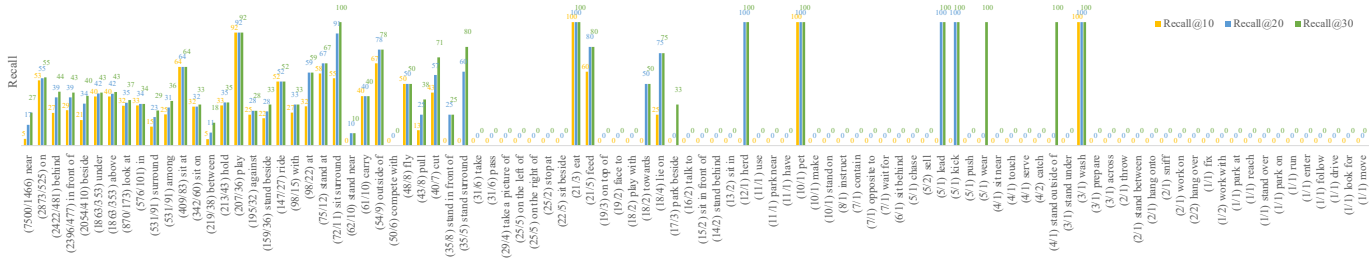


Fig. 9. Recall@10, Recall@20, and Recall@50 for 96 predicates on the *COCO-GVR* dataset. The occurrence times of each predicate in the training set and the test set is indicated alongside the predicate name (e.g., "near" appears 7,500 times in the training set and 1,466 times in the test set).

which equals 0.5 in our experiments. In particular, we treat all the bounding boxes of the instances belonging to a group as a mask, and calculate the IoU between the two masks, *i.e.*, $IoU = \frac{\bigcup_{ij}(\bigcap(b_i^p, b_j^g))}{\bigcup_{ij}(\bigcup(b_i^p, b_j^g))}$. Here, b_i^p and b_j^g denote the bounding boxes of the i^{th} instance in the GVRD result and the j^{th} instance in the groundtruth, respectively; $\bigcup(.)$ and $\bigcap(.)$ denote the union area and the interaction area of two bounding boxes, respectively.

B. Implementation Details

Our method is built on the Detectron2 framework [71], utilizing the pre-trained model "R101-FPN-3x.pkl", which was trained on the COCO dataset [8]. The object detector in Detectron2 is based on the Mask R-CNN architecture, consisting of ResNet101, FPN and RPN .

All the experiments are conducted on a server with CPU E5-2680, GPU 3090 and 64GB memory. The average processing time of our SGPR method is 0.26 seconds per image.

C. Component Analysis

We evaluate the effectiveness of different components: the EC module (Table I), the GP branch in the GRP module (Table III) and the PP branch (Table IV) in the GRP module. We also evaluate the effectiveness and the inference performance of the hyperparameters that used to generate group candidates, the results are shown in Table II. Finally, considering the long-tail distribution of different predicates, the performance of each predicate is demonstrated in Fig. 9.

TABLE I
ABLATION STUDY RESULTS OF THE EC MODULE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	R@10	mR@10	R@20	mR@20	R@30	mR@30
VIS	25.48	12.7	33.37	20.05	**38.53**	22.73
GPP	23.12	11.15	30.23	16.95	35.21	21.21
PPP	25.60	13.74	33.12	19.11	37.75	23.06
SGRP	**25.70**	**14.71**	**33.65**	**22.19**	**38.53**	**25.78**

We first compare our SGRP method with three variants that change the EC module in the original method, and all these variants are also able to generate group and phrase candidates with different instances of one category. The results are shown in Table I. 1) The first variant, denoted as "VIS", uses instances' visual features to calculate their similarity and generate group candidates. 2) The second variant, denoted as "GPP", pretrains another detector to directly generate bounding boxes of group candidates, *i.e.*, group proposal, and matches group bounding boxes with instance bounding boxes to generate group candidates. The phrase candidates are also generated by combining each two group candidates. 3) The third variant, denoted as "PPP", pretrains another detector to directly generate bounding boxes of phrase candidates, *i.e.*, phrase proposal, and matches phrase bounding boxes with instance bounding boxes to generate phrase candidates with groups. The performance of "VIS" is slightly worse than that of our SGRP method, proving that dividing instances into groups by their global location is more effective. The metrics of "GPP" and "PPP" are both lower than ours, and the performance of "GPP" is worse than "PPP". We suppose that "GPP" and "PPP" may generate many invalid proposals and it is harder for the network to predict correct GVRs. Since the area of phrase proposals is usually larger than that of group proposals and the phrase proposals of different relations could be similar, proposal generation for "GPP" is harder than "PPP". These three variants show that candidate group and phrase generation in the bottom-top style, *i.e.*, according to similarity of instances is effective, and calculating similarity by instances' global location is better than their visual features.

We evaluate the effectiveness of the thresholds used for dividing instances into group candidates, the results are shown in Table II. Since our method confirm the components of a group at the last module along with the prediction of the relationship predicates, we actually do not need the sufficiently accurate group candidates at the GCC module. In constract, since the GCC module cannot predict sufficiently accurate group candidates, overly fine-grained group candidate divisions may lead to a large number of incorrect group candidates, negatively impacting the final relationship prediction results. However, we still need a certain level of diversity in group candidates to reduce the difficulty for subsequent modules in determining the group components. Meanwhile, we test the running time for inference, the results are also shown in Table II. The results show that while FPS decreases as the number of instance similarity thresholds increases, the inference time remains within a practical and manageable range.

To evaluate the effectiveness of the GP branch, we design four variants, and the results are shown in Table III. 1) The first variant, denoted as "w/o CA", uses self-attention like

TABLE II
ABLATION STUDY RESULTS OF SIMILARITY THRESHOLDS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Thresholds	R@10	mR@10	R@20	mR@20	R@30	mR@30	FPS
0	25.48	12.32	**33.68**	21.07	38.51	23.47	**3.85**
0,0.5	**25.70**	**14.71**	33.65	**22.19**	**38.53**	**25.78**	3.80
0,0.3,0.6,0.9	25.19	11.27	32.51	19.17	37.57	22.73	3.39
0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9	25.01	12.04	32.47	19.42	37.45	23.66	3.15

TABLE III
ABLATION STUDY RESULTS OF THE GP BRANCH IN THE GRP MODULE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	R@10	mR@10	R@20	mR@20	R@30	mR@30
w/o CA	24.06	11.66	31.67	15.6	36.89	20.37
w/o GP	**25.91**	12.63	33.61	19.97	38.57	23.2
BiLSTM	25.27	13.3	33.31	19.22	38.18	23.31
GCN	25.72	11.77	33.53	18.16	**38.92**	24.35
SGRP	25.70	**14.71**	**33.65**	**22.19**	38.53	**25.78**

TABLE IV
ABLATION STUDY RESULTS OF THE PP BRANCH IN THE GRP MODULE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Methods	R@10	mR@10	R@20	mR@20	R@30	mR@30
w/o G&P	24.17	12.64	31.58	16.96	36.32	21.75
w/o P	25.41	11.95	33.27	18.86	38.14	22.45
Mean	25.19	12.79	33.41	20.26	38.51	25.24
SGRP	**25.70**	**14.71**	**33.65**	**22.19**	**38.53**	**25.78**

the traditional Transformer instead of the cross-attention in the GP branch. 2) The second variant, denoted as "w/o GP", considers all instances in the group candidates actually belong to the group and predicts confidence 1.0 for each instance. 3) The third variant, denoted as "BiLSTM", uses the bi-direction LSTM [72] to encode the visual and location features of instances in groups. 4) The forth variant, denoted as "GCN", uses the GCN [73] to encode the visual and location features of instances in groups. Our method outperforms the "w/o CA" at all metrics, showing that cross-attention is more effective than self-attention in group prediction, which also validates the observation that the range of the subject/object group in a GVR is relative to the range of the corresponding object/subject group. Also, all metrics of "BiLSTM" are worse than ours, proving the effectiveness of our cross-attention Transformer encoder. Compared with "w/o GP" and "GCN", our method achieves better performance at all mean recalls, but is slightly lower at R@10 and R@30. The main reason is that a lot of images in the *COCO-GVR* dataset are not complex in content and the groups in these images do contain all instances of same category, especially in the cases with high frequency predicates. Though R@10 of SGRP is 0.81% and 0.08% lower than that of "w/o GP" and "GCN" repectively, mR@10 of SGRP is 16.47% and 24.98% higher than that of "w/o GP" and "GCN" repectively. Similarly, though R@30 of SGRP is 0.10% and 1.00% lower than that of "w/o GP" and "GCN" repectively, mR@30 of SGRP is 11.12% and 5.87% higher than that of "w/o GP" and "GCN" repectively. Thus, the GP branch and the cross-attention Transformer encoder used in the GP branch are effective.

To evaluate the effectiveness of the PP branch, we design three variants. The results are shown in Table IV. 1) The first variant, denoted as "w/o G&P", only uses instance features in the PP branch. 2) The second variant, denoted as "w/o P", uses instance features and group features in the PP branch. 3) The third variant, denoted as "Mean", uses average instance features to generate features for predicate prediction instead of using the weighted average instance features, whose weights are the instance confidences predicted by the GP branch. The performance of "w/o G&P" is obversely worse than ours, and when group features are added, *i.e.,* the "w/o P" variant, the performance is relatively improved. Comparing the performance of "w/o P" with ours, we find that additional phrase features greatly help to enhance the mean recall but slightly contribute to recall. We suppose that the instances belonging to a group may be dispersed from each other and a phrase area may contain many different GVRs, but phrase features can still provide more context information to assist for GVRD especially in the cases with low frequency predicates. Furthermore, SGRP outperforms "Mean" at all metrics, which demonstrates the effectiveness of using instance confidence as "weight" for predicate prediction.

Considering the long-tail problem, we show the performance of each predicate in Fig. 9. The predicates are ordered based on their frequency in the training set. The results reveal that the performance of different predicates is not solely determined by their frequency. While head predicates are more likely to be predicted correctly, and tail predicates tend to be harder to predict, higher frequency predicates do not necessarily achieve higher recall. For instance, the most frequent predicate, "near", does not perform the best across all three metrics, whereas "sit surround" demonstrates strong performance in all metrics. Additionally, tail predicates like "eat", "pet", and "wash" also show high performance. The reason for this could be that predicates with clear and explicit patterns are easier to predict, while those with broader and more vague definitions are harder to learn.

D. Comparison with the SOTA baselines

As GVRD is a novel task with no existing methods explicitly designed for it, we compare our approach with IVRD methods retrained on the *COCO-GVR* dataset, adjusted for IVRD. Following this, a clustering strategy was applied to generalize their IVRD outputs into GVRD results. This approach provides a consistent and systematic way to adapt IVRD methods to the GVRD task, enabling meaningful comparisons despite the differences in their original design

TABLE V
COMPARISON RESULTS OF SGRP *vs.* IVRD METHODS ON *COCO-GVR*. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. THE RESULTS OF THE IVRD METHODS ON *VG150* ARE ALSO LISTED FOR REFERENCE.

Methods	GVRD (*COCO-GVR*)						IVRD (*VG150*)			
	R@10	mR@10	R@20	mR@20	R@30	mR@30	i-R@50	i-mR@50	i-R@100	i-mR@100
IMP [10]	9.16	1.62	13.79	2.02	18.24	2.75	3.44	-	4.24	-
KERN [74]	9.20	0.62	13.92	0.96	16.74	1.22	27.10	6.40	29.80	7.30
MFURLN [6]	12.42	1.40	17.59	2.28	21.68	2.87	14.40	-	16.50	-
EGTR [19]	3.53	0.68	5.64	1.13	6.91	1.39	30.20	7.90	34.30	10.10
Graph R-CNN [15]	14.76	3.39	20.04	4.36	24.19	5.17	11.40	-	13.70	-
BGNN [13]	16.91	7.19	23.59	9.32	28.9	10.44	31.00	10.70	35.80	12.60
HL-Net [75]	18.41	7.01	25.21	9.84	29.72	11.73	33.70	-	38.10	-
PENET [76]	17.91	8.82	24.76	11.75	30.21	14.67	30.70	12.40	35.20	14.50
Motifs [11], [23]	17.44	5.51	23.69	8.46	28.31	11.83	32.10	5.50	36.90	6.80
VCTree [16], [23]	17.67	4.82	24.37	7.34	29.19	8.62	31.50	5.70	36.20	6.90
Transformer [23]	17.36	4.92	23.43	6.90	27.92	8.07	33.04	8.13	37.40	8.66
TDE-Motifs [23]	16.29	2.93	22.55	3.91	27.4	4.94	16.90	8.20	20.30	9.80
TDE-VCTree [23]	16.23	2.54	22.67	3.67	27.59	4.43	19.40	9.30	23.20	11.10
BA-SGG-Motif [77]	14.33	12.44	20.52	15.51	24.51	16.81	23.00	13.50	26.90	15.60
BA-SGG-VCTree [77]	12.85	8.30	18.45	13.15	22.12	15.06	21.70	13.50	25.50	15.70
BA-SGG-Transformer [77]	12.79	12.88	17.83	16.20	21.66	21.07	-	14.80	-	17.10
DLFE-Motifs [78]	18.3	7.38	24.92	11.32	29.68	15.11	-	11.70	-	13.80
DLFE-VCTree [78]	18.26	7.56	25.17	9.92	30.25	13.68	-	11.80	-	13.80
EBM-Motifs [79]	18.00	7.08	24.66	9.98	29.27	11.63	31.74	7.72	36.29	9.27
EBM-VCTree [79]	19.41	6.84	26.28	9.71	30.97	11.26	31.36	7.71	35.87	9.10
EBM-Transformer [79]	18.02	6.41	24.51	8.96	29.08	12.24	-	-	-	-
FGPL-Motifs [25]	9.67	3.36	14.04	5.10	17.20	6.30	-	15.40	-	18.20
FGPL-VCTree [25]	12.36	9.02	17.01	13.05	20.07	15.81	-	16.20	-	19.10
FGPL-Transformer [25]	13.61	10.13	18.59	15.60	21.83	18.46	-	17.40	-	20.30
RTPB-Motifs [24]	17.28	14.20	23.28	19.63	27.24	21.71	19.00	13.10	22.50	15.50
RTPB-VCTree [24]	17.32	12.58	23.14	16.37	27.71	20.64	18.10	12.80	21.30	15.10
RTPB-DualTrans [24]	16.42	12.82	21.79	16.47	25.89	21.90	19.70	16.50	23.40	19.00
DHL-Motifs [65]	18.2	11.12	24.33	15.42	28.72	16.62	24.70	17.80	28.80	20.70
DHL-VCTree [65]	17.2	12.03	22.65	14.91	26.52	16.96	23.30	17.40	27.10	20.00
DHL-Transformer [65]	17.03	11.52	23.61	16.39	27.65	20.20	23.20	18.20	27.30	21.00
SGRP	**25.70**	**14.71**	**33.65**	**22.19**	**38.53**	**25.78**	-	-	-	-

purposes. Table V shows the comparison result. For fair comparison, we use the same instance detector in all the two-stage baselines to our method, retrain their predicate classification modules with the IVRs departed from the GVRs in groundtruth. Since "EGTR" [19] is a one-stage method, we adhere to its original pipeline by employing Deformable DETR for embedded object detection, using pretrained weights from COCO. The predicted IVRs are then clustered to GVRs in two steps: 1) Iterate over the IVR list, merge the subject/object instances which have the same category and the same predicate with the same object/subject into a group object, and construct a relation list containing the "individual-to-group" relation triplets; 2) Iterate over the "individual-to-group" list, and merge the subject/object similar to the Step 1) to construct a relation list containing the "group-to-group" relation triplets. We collect the GVRs generated in the Step 2), order the generated GVRs with the average confidence of its clustered IVRs, and treat them as the GVRD result of each baseline. Since there are no absolute mapping between IVRs and GVRs, *e.g.*, when some individuals are "beside" another individual, they do not necessarily "surround" the individual, we do not further map IVRs to GVRs.

To provide a better reference for the IVRD performance of these methods, we have listed the results evaluated on *VG150* [10]. It is important to note that the metrics used are the commonly adopted recall and mean recall under graph constraints, simplified as i-R@t and i-mR@t. For IVRD, the value of t is typically set to 50 and 100. We can see that, our SGRP method outperforms all the baselines at all metrics, although some recent methods, *e.g.*, "PENET", "BA-SGG", "RTPB", "DLFE", "DHL" and "EBM", also have fairly good performance. The results show that SGRP can handle the GVRD task better by predicting groups and predicates in GVRs interdependently than clustering GVRs from IVRs. For one thing, the IVRD baselines cannot perfectly predict the correct predicate for all instances in a group between another instance or each instance in another group; for another, even if all these IVRs are predicted correctly, the clustering strategy may divide instances that have the same relation with another instance/group but quite different characteristics into a group. In comparison, group candidates are constructed on the basis of diverse instance similarity thresholds in our method and it can handle such a case more effectively. Moreover, the features generated by IVRD methods are hard to support the predicates only for GVRs, *e.g.*, all the instances sharing the "surround" relation with another instance/group may have diverse features generated by IVRD methods. We retrain these baselines with IVRs departed from the GVRs in groundtruth in our experiments, which actually leads these IVRD methods to detect IVRs that are likely to be clustered as GVRs, however, it is harder for these IVRD methods to detects correct IVRs for GVR clustering if they are only trained with IVR groundtruth. Combining features at instance, group and phrase levels together, our method is not confined to instance features

Fig. 10. Qualitative comparison of the GVRD results predicted by SGRP and baseline methods with the groundtruth.

and is more robust for detecting GVRs.

E. Qualitative Analysis

To provide a clearer demonstration of SGRP's effectiveness, Fig. 10 presents some visualized results. For each image, a groundtruth example is displayed alongside predicted examples selected from the top 10 results based on their scores. If any predicted GVR matches the groundtruth in all three aspects: subject category, predicate category, and object category, the highest-scoring matching GVR is shown in Fig. 10 for qualitative analysis. If no predicted GVR fully matches the groundtruth, the highest-scoring GVR that matches the groundtruth in both subject and object categories is displayed. For comparison, we include clustering results from "EBM-VCTree", "RTPB-DualTrans", and "RTPB-Motifs", as these methods achieve the second-highest performance at the six evaluation metrics respectively.

From the qualitative comparison, it is evident that SGRP outperforms the baseline methods relying on clustering IVRs when predicting GVRs. For example, in the first row, SGRP successfully predicts the relationship <person*, look at, cake>, while none of the other methods predict any GVR between "person" and "cake". In the second row, SGRP accurately predicts <person*, outside of, bus*>, whereas the clustering IVRs approach of RTPB-DualTrans and RTPB-Motifs predicts the correct predicate but fails to group the "person" instances correctly. Similarly, EBM-VCTree predicts only a person near the buses, failing to generalize a group of people outside of the buses. In the third row, SGRP predicts the "wash" relationship between "person" and "elephant", while all three baseline methods incorrectly predict the relationship as "ride", a common collocation that does not match the actual scenario. This highlights that GVRs are less influenced by fixed collocations compared to IVRs, as groups are composed of multiple instances. GVRs with the same subject, object, and predicate categories can demonstrate more diverse patterns, reflecting real-world variability. Furthermore, GVRD requires relationship generalization, which is challenging when specific IVRs cannot be easily clustered into a generalized GVR. Additionally, the clustering-based approaches calculate a GVR's score as the average confidence of its clustered IVRs. This scoring method is sensitive to low-confidence IVRs, potentially lowering the overall score of the GVR and causing it to rank lower. In contrast, a simple GVR formed by high-confidence IVRs may rank higher, despite being less accurate. However, SGRP does have limitations, as seen in the third row, where it misidentifies the precise range of the subject and object. In this case, distinguishing that only three people are washing the lying elephant remains a significant challenge. This demonstrates areas for further refinement in SGRP's performance.

VI. Conclusion and Future Work

We proposed the GVRD task, which aims to detect the visual relations whose subjects and/or objects are groups in images and output the bounding boxes of the subjects and objects. GVRD allows for a deeper understanding of complex scenes where meaning arises from the interactions and arrangements of multiple objects, rather than just pairwise relationships. To solve GVRD, we proposed a SGRP method that can simultaneously predict groups and the group related visual relations. Our SGRP method consists of an entity construction module, a feature extraction module, and a group relation prediction module. To evaluate the performance of

solutions for GVRD, we constructed the first dataset *COCO-GVR*, containing 9,570 images with 31,855 annotated GVRs. We conducted experiments on the *COCO-GVR* dataset and proved that SGRP outperforms IVRD methods. We also examined the effectiveness of different components in the SGRP method, which is validated by our experimental results.

GVRD provides valuable contextual information that can significantly enhance visual representations for image understanding and improve a variety of downstream tasks. For instance, in the field of image captioning, GVRD can enrich scene descriptions by incorporating group-level interactions, which is exemplified by captions such as "A group of people are gathered around a dining table''. Besides, in visual question answering, questions involving group behaviors, such as "What is the group of people doing?" can be answered more accurately with the semantic insights provided by GVRD. Moreover, the rich and general semantic information derived from GVRD significantly enhances scene understanding, offering a higher-level abstraction for tasks like image retrieval and social relation analysis.

Despite its potential benefits, GVRD still faces many challenges. Although the proposed SGRP method performs well on some common predicates, it struggles with predicting infrequent ones. Its performance may further deteriorate when detecting GVRs in open-world scenarios. Meanwhile, the efficiency of current GVRD methods needs to be improved to facilitate their applications in large-scale scene understanding.

REFERENCES

[1] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European Conference on Computer Vision*, 2016, pp. 852–869.

[2] L. Li, X. Gao, J. Deng, Y. Tu, Z.-J. Zha, and Q. Huang, "Long short-term relation transformer with global gating for video captioning," *IEEE Transactions on Image Processing*, vol. 31, pp. 2726–2738, 2022.

[3] J. Yu, X. Jiang, Z. Qin, W. Zhang, Y. Hu, and Q. Wu, "Learning dual encoding model for adaptive visual understanding in visual dialogue," *IEEE Transactions on Image Processing*, vol. 30, pp. 220–233, 2020.

[4] H. Wang, Y. Du, Y. Zhang, S. Li, and L. Zhang, "One-stage visual relationship referring with transformers and adaptive message passing," *IEEE Transactions on Image Processing*, vol. 32, pp. 190–202, 2022.

[5] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in *IEEE International Conference on Computer Vision*, 2017, pp. 1974–1982.

[6] Y. Zhan, J. Yu, T. Yu, and D. Tao, "On exploring undetermined relationships for visual relationship detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5128–5137.

[7] B. Ni, S. Yan, and A. Kassim, "Recognizing human group activities with localized causalities," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1470–1477.

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[9] S. Sharifzadeh, S. M. Baharlou, M. Schmitt, H. Schütze, and V. Tresp, "Improving scene graph classification by exploiting knowledge from texts," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 2189–2197.

[10] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5410–5419.

[11] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.

[12] S. Jae Hwang, S. N. Ravi, Z. Tao, H. J. Kim, M. D. Collins, and V. Singh, "Tensorize, factorize and regularize: Robust visual relationship learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1014–1023.

[13] R. Li, S. Zhang, B. Wan, and X. He, "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 109–11 119.

[14] S. Khandelwal and L. Sigal, "Iterative scene graph generation," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 295–24 308.

[15] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *European Conference on Computer Vision*, 2018, pp. 670–685.

[16] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6619–6628.

[17] Y. Lu, H. Rai, J. Chang, B. Knyazev, G. Yu, S. Shekhar, G. W. Taylor, and M. Volkovs, "Context-aware scene graph generation with seq2seq transformers," in *IEEE International Conference on Computer Vision*, 2021, pp. 15 931–15 941.

[18] Y. Cong, M. Y. Yang, and B. Rosenhahn, "Reltr: Relation transformer for scene graph generation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11 169–11 183, 2023.

[19] J. Im, J. Nam, N. Park, H. Lee, and S. Park, "Egtr: Extracting graph from transformer for scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 229–24 238.

[20] X. Dong, T. Gan, X. Song, J. Wu, Y. Cheng, and L. Nie, "Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 427–19 436.

[21] M. Diomataris, N. Gkanatsios, V. Pitsikalis, and P. Maragos, "Grounding consistency: Distilling spatial common sense for precise visual relationship detection," in *IEEE International Conference on Computer Vision*, 2021, pp. 15 911–15 920.

[22] J. Lu, L. Chen, Y. Song, S. Lin, C. Wang, and G. He, "Prior knowledge-driven dynamic scene graph generation with causal inference," in *ACM International Conference on Multimedia*, 2023, pp. 4877–4885.

[23] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3716–3725.

[24] C. Chen, Y. Zhan, B. Yu, L. Liu, Y. Luo, and B. Du, "Resistance training using prior bias: toward unbiased scene graph generation," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 212–220.

[25] X. Lyu, L. Gao, Y. Guo, Z. Zhao, H. Huang, H. T. Shen, and J. Song, "Fine-grained predicates learning for scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 467–19 475.

[26] T. He, L. Gao, J. Song, and Y.-F. Li, "State-aware compositional learning toward unbiased training for scene graph generation," *IEEE Transactions on Image Processing*, vol. 32, pp. 43–56, 2022.

[27] J. Yang, C. Wang, L. Yang, Y. Jiang, and A. Cao, "Adaptive feature learning for unbiased scene graph generation," *IEEE Transactions on Image Processing*, vol. 33, pp. 2252–2265, 2024.

[28] L. Li, L. Chen, Y. Huang, Z. Zhang, S. Zhang, and J. Xiao, "The devil is in the labels: Noisy label correction for robust scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 869–18 878.

[29] X. Li, P. Miao, S. Li, and X. Li, "Mlmg-sgg: Multi-label scene graph generation with multi-grained features," *IEEE Transactions on Image Processing*, vol. 33, pp. 1549–1559, 2022.

[30] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in *IEEE International Conference on Computer Vision*, 2017, pp. 5179–5188.

[31] F. Yu, H. Wang, T. Ren, J. Tang, and G. Wu, "Visual relation of interest detection," in *ACM International Conference on Multimedia*, 2020, p. 1386–1394.

[32] W. Wang, R. Wang, S. Shan, and X. Chen, "Sketching image gist: Human-mimetic hierarchical scene graph generation," in *European Conference on Computer Vision*, 2020, pp. 222–239.

[33] X. Li, T. Wu, G. Zheng, Y. Yu, and X. Li, "Uncertainty-aware scene graph generation," *Pattern Recognition Letters*, vol. 167, pp. 30–37, 2023.

[34] X. Li, G. Zheng, Y. Yu, N. Ji, and X. Li, "Relationship-incremental scene graph generation by a divide-and-conquer pipeline with feature adapter," *IEEE Transactions on Image Processing*, 2024.

[35] T. He, L. Gao, J. Song, and Y.-F. Li, "Toward a unified transformer-based framework for scene graph generation and human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 6274–6288, 2023.

[36] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Acp++: Action co-occurrence priors for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 9150–9163, 2021.

[37] Y. Gao, Z. Kuang, G. Li, W. Zhang, and L. Lin, "Hierarchical reasoning network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8306–8317, 2021.

[38] M. Gu, Z. Zhao, W. Jin, R. Hong, and F. Wu, "Graph-based multi-interaction network for video question answering," *IEEE Transactions on Image Processing*, vol. 30, pp. 2758–2770, 2021.

[39] H. Wang, L. Jiao, F. Liu, L. Li, X. Liu, D. Ji, and W. Gan, "Ipgn: Interactiveness proposal graph network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 6583–6593, 2021.

[40] D. Yang, Y. Zou, Z. Li, and G. Li, "Learning human-object interaction via interactive semantic reasoning," *IEEE Transactions on Image Processing*, vol. 30, pp. 9294–9305, 2021.

[41] Z. Zeng, P. Dai, X. Zhang, L. Zhang, and X. Cao, "Cognition guided human-object relationship detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 2468–2480, 2023.

[42] T.-T. Nguyen, P. Nguyen, and K. Luu, "Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 384–18 394.

[43] Z. Lin, F. Zhu, Y. Kong, Q. Wang, and J. Wang, "Srsg and s2sg: A model and a dataset for scene graph generation of remote sensing images from segmentation results," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[44] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Dual-glance model for deciphering social relationships," in *IEEE International Conference on Computer Vision*, 2017, pp. 2650–2659.

[45] Y. Guo, F. Yin, W. Feng, X. Yan, T. Xue, S. Mei, and C.-L. Liu, "Social relation reasoning based on triangular constraints," in *AAAI Conference on Artificial Intelligence*, 2023, pp. 737–745.

[46] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[47] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5532–5540.

[48] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6700–6709.

[49] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, and T. Mei, "Vrr-vg: Refocusing visually-relevant relationships," in *IEEE International Conference on Computer Vision*, 2019, pp. 10 403–10 412.

[50] J. Yang, Y. Z. Ang, Z. Guo, K. Zhou, W. Zhang, and Z. Liu, "Panoptic scene graph generation," in *European Conference on Computer Vision*, 2022, pp. 178–196.

[51] S. S. Intille and A. F. Bobick, "Recognizing planned, multiperson action," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 414–445, 2001.

[52] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *IEEE International Conference on Computer Vision Workshops*, 2009, pp. 1282–1289.

[53] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *European Conference on Computer Vision*, 2012, pp. 187–200.

[54] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1971–1980.

[55] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1354–1361.

[56] J. Tang, X. Shu, R. Yan, and L. Zhang, "Coherence constrained graph lstm for group activity recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 636–647, 2019.

[57] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "Stagnet: An attentive semantic rnn for group activity recognition," in *European Conference on Computer Vision*, 2018, pp. 101–117.

[58] H. Yuan, D. Ni, and M. Wang, "Spatio-temporal dynamic inference network for group activity recognition," in *IEEE International Conference on Computer Vision*, 2021, pp. 7476–7485.

[59] Z. Xie, C. Jiao, K. Wu, D. Guo, and R. Hong, "Active factor graph network for group activity recognition," *IEEE Transactions on Image Processing*, vol. 33, pp. 1574–1587, 2024.

[60] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, "Social adaptive module for weakly-supervised group activity recognition," in *European Conference on Computer Vision*, 2020, pp. 208–224.

[61] D. Kim, J. Lee, M. Cho, and S. Kwak, "Detector-free weakly supervised group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 083–20 093.

[62] L. Wu, M. Tian, Y. Xiang, K. Gu, and G. Shi, "Learning label semantics for weakly supervised group activity recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 6386–6397, 2024.

[63] W. Zhou, L. Kong, Y. Han, J. Qin, and Z. Sun, "Contextualized relation predictive model for self-supervised group activity representation learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 353–366, 2023.

[64] M. Ehsanpour, A. Abedin, F. Saleh, J. Shi, I. Reid, and H. Rezatofighi, "Joint learning of social groups, individuals action and sub-group activities in videos," in *European Conference on Computer Vision*, 2020, pp. 177–195.

[65] C. Zheng, L. Gao, X. Lyu, P. Zeng, A. El Saddik, and H. T. Shen, "Dual-branch hybrid learning network for unbiased scene graph generation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1743–1756, 2024.

[66] H. Zhou, C. Zhang, and C. Hu, "Visual relationship detection with relative location mining," in *ACM International Conference on Multimedia*, 2019, pp. 30–38.

[67] K. He, G. Gkioxari, P. Doll
'ar, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.

[68] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[70] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *International Conference on Learning Representations*, 2021.

[71] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[73] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[74] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6163–6171.

[75] X. Lin, C. Ding, Y. Zhan, Z. Li, and D. Tao, "Hl-net: Heterophily learning network for scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 476–19 485.

[76] C. Zheng, X. Lyu, L. Gao, B. Dai, and J. Song, "Prototype-based embedding network for scene graph generation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 783–22 792.

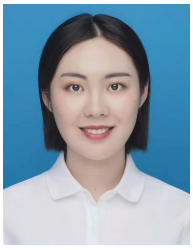
[77] Y. Guo, L. Gao, X. Wang, Y. Hu, X. Xu, X. Lu, H. T. Shen, and J. Song, "From general to specific: Informative scene graph generation via balance adjustment," in *IEEE International Conference on Computer Vision*, 2021, pp. 16 383–16 392.

[78] M.-J. Chiou, H. Ding, H. Yan, C. Wang, R. Zimmermann, and J. Feng, "Recovering the unbiased scene graphs from the biased ones," in *ACM International Conference on Multimedia*, 2021, pp. 1581–1590.

[79] M. Suhail, A. Mittal, B. Siddiquie, C. Broaddus, J. Eledath, G. Medioni, and L. Sigal, "Energy-based learning for scene sraph generation," in *IEEE Conference on Computer Vision and Pattern Recogition*, 2021, pp. 13 936–13 945.

VII. Biography Section

Fan Yu received the B.E. degree from Nanjing University, Nanjing, China in 2018. She is currently pursuing the Ph.D. degree in Nanjing University, Nanjing, China. Her research interests include visual relationship detection and its applications. She was in the champion teams of ECCV 2018 PIC challenge, MM 2020 DVU challenge and MM 2023 DVU challenge.



Beibei Zhang received the B.E. degree from Nanjing University, Nanjing, China in 2020. She is currently pursuing the Ph.D. degree in Nanjing University, Nanjing, China. Her research interests include video analysis and understanding. She was in the champion teams of MM 2020 DVU challenge and MM 2022 DVU challenge.

Tongwei Ren (Member, IEEE) received the B.S., M.E., and Ph.D. degrees from Nanjing University, Nanjing, China, in 2004, 2006, and 2010, respectively. He joined Nanjing University in 2010, and at present he is a professor. His research interest mainly includes multimedia computing and its real-world applications. He has published more than 40 papers in top-tier journals and conferences. He was a recipient of the best paper candidate awards of ICIMCS 2014, PCM 2015, and MMAsia 2020, and he was in the champion teams of ECCV 2018 PIC challenge, MM 2019 VRU challenge, MM 2020 DVU challenge, MM 2022 DVU challenge and MM 2023 DVU challenge.

Jiale Liu received the B.E. degree from Nanjing Institute of Technology, Nanjing, China in 2022. He worked as an intern at Nanjing University from 2021 to 2022. His research interests include visual relationship detection.

Gangshan Wu (Member, IEEE) received the B.Sc., M.S., and Ph.D. degrees from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 1988, 1991, and 2000, respectively. He is currently a Professor with the School of Computer Science, Nanjing University. His current research interests include computer vision, multimedia content analysis, multimedia information retrieval, digital museum, and large-scale volumetric data processing.

Jinhui Tang (Senior Member, IEEE) received the B.E. and Ph.D. degrees from the Univerisity of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Nanjing University of Science and Technology, Nanjing, China. Hehas authored more than 200 articles in toptier journals and conferences. His research interests include multimedia analysis and computer vision. He was a recipient of the Best Paper Awards in ACM MM 2007 and ACM MM Asia 2020, the Best Paper Runner-Up in ACM MM 2015. He has served as an Associate Editor for IEEE TKDE, IEEE TMM, IEEE TNNLS and IEEE TCSVT. He is a Fellow of IAPR.