

STAR: A First-Ever Dataset and a Large-Scale Benchmark for Scene Graph Generation in Large-Size Satellite Imagery

Yansheng Li^{1b}, Senior Member, IEEE, Linlin Wang^{1b}, Tingzhu Wang^{1b}, Xue Yang^{1b}, Junwei Luo^{1b}, Qi Wang^{1b}, Senior Member, IEEE, Youming Deng, Wenbin Wang, Xian Sun^{1b}, Senior Member, IEEE, Haifeng Li^{1b}, Member, IEEE, Bo Dang^{1b}, Yongjun Zhang^{1b}, Member, IEEE, Yi Yu^{1b}, Member, IEEE, and Junchi Yan^{1b}, Senior Member, IEEE

Abstract—Scene graph generation (SGG) in satellite imagery (SAI) benefits promoting understanding of geospatial scenarios from perception to cognition. In SAI, objects exhibit great variations in scales and aspect ratios, and there exist rich relationships between objects (even between spatially disjoint objects), which makes it attractive to holistically conduct SGG in large-size very-high-resolution (VHR) SAI. However, there lack such SGG datasets. Due to the complexity of large-size SAI, mining triplets $\langle \text{subject}, \text{relationship}, \text{object} \rangle$ heavily relies on long-range contextual reasoning. Consequently, SGG models designed for small-size natural imagery are not directly applicable to large-size SAI. This paper constructs a large-scale dataset for SGG in large-size VHR SAI with image sizes ranging from 512×768 to 27860×31096 pixels, named STAR (Scene graph generaTion in Large-size satellite imageRy), encompassing over 210K objects and over 400K triplets. To realize SGG in large-size SAI, we propose a context-aware cascade cognition (CAC) framework to understand SAI regarding

object detection (OBD), pair pruning and relationship prediction for SGG. We also release a SAI-oriented SGG toolkit with about 30 OBD and 10 SGG methods which need further adaptation by our devised modules on our challenging STAR dataset.

Index Terms—Large-size satellite imagery, object detection, relationship prediction, scene graph generation benchmark.

I. INTRODUCTION

SCENE graph generation (SGG) [1], [2], [3] in satellite imagery (SAI) aims to detect objects and predict relationships between objects, where scene graph is a high-level structured representation of SAI via extensive triplets $\langle \text{subject}, \text{relationship}, \text{object} \rangle$. Compared with traditional perceptual understanding tasks in SAI, such as object detection [4], [5], [6], scene classification [7], [8], [9] and semantic segmentation [10], [11], [12], SGG in SAI works towards cognitively mining the high-value knowledge contained in SAI (e.g., in Fig. 1, $\langle \text{ship1}, \text{away from}, \text{dock1} \rangle$ can be inferred by the spatial layout between ship1 and dock1 and the wake trailing of ship1). SGG in SAI benefits cognitive interpretation tasks in SAI, like image retrieval [13], image captioning [14] and image visual question answering [15], and further serves various applications, e.g., traffic planning [16], [17], circuit layout [18]. As illustrated in Fig. 1(a) and (b), large-size VHR SAI allows for the retention of detailed information on small objects without compromising the integrity on large objects, which provides mandatory support for predicting rich relationships between multi-scale objects. However, to the best of our knowledge, there is a significant lack of both datasets and approaches for SGG in large-size VHR SAI. Therefore, there is an urgent need to explore datasets and approaches around SGG in large-size VHR SAI to pursue the cognitive understanding of SAI.

Table I shows existing datasets for the SGG task in SAI [19], [20], which face limitations concerning the number of objects, number of triplets, and image size. Moreover, they tend to provide horizontal bounding box (HBB) annotations with too much background disturbance, hurting the recognition of relationships between man-made objects with distinct oriented boundaries in SAI. For the object detection (OBD) task in SAI, while the datasets for OBD in SAI [5], [21], [22] are continually

Received 2 July 2024; revised 26 October 2024; accepted 24 November 2024. Date of publication 28 November 2024; date of current version 5 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42371321 and Grant 42030102 and in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102. Recommended for acceptance by Y.-Y. Chuang. (Corresponding authors: Linlin Wang; Tingzhu Wang; Xue Yang.)

Yansheng Li, Linlin Wang, Tingzhu Wang, Junwei Luo, Bo Dang, and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; wangll@whu.edu.cn; tingzhu.wang@whu.edu.cn; luojunwei@whu.edu.cn; bodang@whu.edu.cn; zhangyj@whu.edu.cn).

Xue Yang is with the Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yangxue-2019-sjtu@sjtu.edu.cn).

Qi Wang is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com).

Youming Deng is with the Department of Computer Science, Cornell University, Ithaca, NY 14853 USA (e-mail: ymdeng@cs.cornell.edu).

Wenbin Wang is with the College of Computer and Information Technology, China Three Gorges University, Yichang 443002, China (e-mail: wangwenbin@ctgu.edu.cn).

Xian Sun is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunxian@mail.ie.ac.cn).

Haifeng Li is with the School of Geosciences and Info-Physics, Central South University, Changsha 410083, China (e-mail: lihaifeng@csu.edu.cn).

Yi Yu is with the School of Automation, Southeast University, Nanjing 210096, China (e-mail: yuyi@seu.edu.cn).

Junchi Yan is with the School of Artificial Intelligence, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: yanjunchi@sjtu.edu.cn).

The dataset and toolkit are available at: <https://linlin-dev.github.io/project/STAR>.

Digital Object Identifier 10.1109/TPAMI.2024.3508072

multi-class objects with diverse object scales (e.g., same category: bridges, docks; different categories: "airplanes and aprons", "ships and docks"), extreme aspect ratios (e.g., bridges, runways and gravitydams), and arbitrary orientations (e.g., trucks, ships, airplanes, bridges, runways). These factors make it challenging to holistically detect these objects while preserving the complete details necessary for accurate relationship prediction. **(ii) Pair pruning.** Large-size VHR SAI usually contains a large number of objects and a plethora of object pairs (the number of which grows exponentially as the number of objects), which inevitably results in excessive computation for exhaustive methods based on all object pairs. Thus, it is necessary to prune object pairs without losing high-value pairs. Notably, the triplets are usually partially annotated in reality (i.e., many non-annotated object pairs still carry meaningful relationships), which suggests that some approaches treating pair sparsity as a binary classification problem (i.e., annotated pairs are treated as positive samples and non-annotated pairs are treated as negative samples) are not reasonable. **(iii) Relationship prediction.** Different from small-size natural imagery, there are a large number of object pairs containing rich knowledge, which are even spatially disjoint in large-size SAI. A few relationships in large-size SAI can be directly predicted by isolated pairs, while most high-value relationships need to be inferred with the help of context. As illustrated in Fig. 1(c) and (d), only simple triplets (e.g., $\langle crane1, over, dock1 \rangle$) in large-size SAI can be predicted when relying only on isolated pairs. However, more complex triplets can be inferred when the context of the pairs is introduced (e.g., $\langle ship2, parallelly docked at, dock1 \rangle$ and $\langle ship5, parallelly docked at, dock2 \rangle$ can be inferred as $\langle ship2, docking at the different dock with, ship5 \rangle$).

In this paper, we propose a context-aware cascade cognition (CAC) framework to understand SAI at three levels: OBD, pair pruning and relationship prediction. **First**, a holistic multi-class object detection network (HOD-Net) incorporating multi-scale context is adopted to detect multi-scale objects in large-size VHR SAI. **Then**, to reduce the complexity of predicting pairs without losing high-value pairs, a pair proposal generation (PPG) network via adversarial reconstruction is designed to obtain contextual interaction pairs containing rich knowledge. **Finally**, considering the dependence of relationship prediction in large-size SAI on the contexts of object pairs, a relationship prediction network with context-aware messaging (RPCM) is proposed to comprehensively predict the relationship types.

In summary, the STAR dataset and CAC framework proposed in this paper aim to establish a large-scale benchmark for SGG in large-size VHR SAI. Under this benchmark, new algorithms can be further developed. The contributions are:

- We propose STAR, a first-ever dataset for SGG in large-size VHR SAI. Containing more than 210K objects and over 400K triplets across 1,273 complex scenarios globally, it provides more challenging and practical testbed for the community.
- To address SGG in large-size VHR SAI, a context-aware cascade cognition (CAC) framework is established where the HOD-Net is devised for holistic detection of multi-class multi-scale objects in large-size SAI, and then the

pair proposal generation (PPG) network and relationship prediction network with context-aware messaging (RPCM) network are designed to retain high-value pairs as well as to predict relationship types of these pairs.

- We release an open-source toolkit, encompassing about 30 OBD methods and 10 SGG methods, designed to accommodate OBD and SGG in large-size VHR SAI. The versatile toolkit unifies the SGG process in natural imagery and SAI, offering flexibility and ease of use. Our toolkit will provide valuable contributions to the development of the SGG community.
- Rigorous experiments on the STAR dataset, employing state-of-the-art algorithms, underscore the indispensability of this dataset and the efficacy of the algorithms. The experimental results showcased in this paper serve as a large-scale benchmark for future research in this direction.

II. RELATED WORK

A. Scene Graph Generation Datasets

Scene graph has attracted extensive attention from many researchers as a powerful tool for the understanding and reasoning analysis of images [14], [15], [27], [28]. Various datasets have been proposed for SGG in recent years.

1) *Scene Graph Generation Datasets in Natural Images:* Visual Phrases [29] is an early dataset for visual relationship detection, which is a task similar to SGG, and it has 8 object categories and 13 relationship categories. RW-SGD [13] may be the first dataset for SGG, is constructed by gathering 5,000 images from YFCC100m [30] and Microsoft COCO datasets [31], containing 93,832 object instances and 112,707 relationship instances. VRD [32] is constructed for the task of visual relationship detection, which has 100 object categories from 5,000 images and contains 37,993 relationship instances. Visual Genome (VG) [33] is a large-scale scene graph dataset, which consists of 108,077 images with tens of thousands of object and relationship categories. Later, VG variants [34], [35] are gradually introduced by researchers to address its key shortcomings. Specifically, VG-150 [34] retains only the most common 150 object categories and 50 predicate categories to reflect a more realistic scenario. VrR-VG [35] argues that many predicates in VG-150 could be easily estimated by statistics, and then re-filters the original VG categories. However, there are still cases of redundancy and ambiguity of predicates in VrR-VG. Similar shortcomings are also found in the popular datasets GQA [36], Open Images V4 and V6 [37]. In addition, the co-occurrence of multiple relationships between subject-object pairs is common in the real world, but most previous datasets treat edge prediction as single-label categorization.

2) *Scene Graph Generation Datasets in SAI:* Compared to the natural imagery field, the development of SGG in SAI field has been relatively slow. SGG datasets in SAI [19], [20], [38], [39] with detailed annotation and considering the characteristics of SAI are rare. GRTRD [19] may be the first SGG dataset in SAI, which focuses on 12 kinds of objects and contains 19,904 object instances and 18,602 relationship instances. RSSGD [20] is a SAI-oriented SGG dataset constructed based on the remote

sensing caption dataset RSICD [40], which has about 39 object categories and 16 relationship categories. Unfortunately, the image size of GRTRD dataset is 600×600 pixels, and the image size of RSSGD dataset is only 224×224 pixels. Since the large-size VHR SAI scenarios are cropped into small image blocks, the objects and relationships contained in each image are very sparse, and many object pairs containing high-value relationships are corrupted, which makes it difficult to support high-level cognitive understanding of SAI. In addition, the objects in the above dataset are annotated by HBB with too much background disturbance, which does not provide accurate information for recognizing relationships between regular man-made objects with distinct oriented boundaries in large-size VHR SAI. To meet the need for cognitive understanding of SAI, a large-scale SGG dataset named STAR for large-size VHR SAI is constructed, which contains multiple complex scenarios and meaningful relationships. In Table I, we give the statistics of popular OBD datasets and SGG datasets in SAI. Compared to small-scale datasets for the SGG task in SAI, such as RSSGD dataset and GRTRD dataset, our STAR outperforms them by at least an order of magnitude in both the number of objects and relationships on all/single images. For OBD datasets in SAI [22], [23], [24] lacking relationship annotations, our STAR dataset provides effective improvements in both object category diversity and scene complexity. The STAR dataset, a large-scale and comprehensive SGG dataset, supports both the OBD task and the SGG task based on the HBB-based and OBB-based detectors in large-size VHR SAI, serving as an invaluable resource for cognitive understanding of SAI.

B. Scene Graph Generation Methods

1) *Scene Graph Generation Methods in Natural Imagery:* Existing SGG methods for natural imagery have been dominated by the two-stage process consisting of OBD and relationship prediction between objects. Given the need for compatibility between OBD models and relationship prediction models within the SGG framework, the majority of existing SGG models, in alignment with the literature [41], [42], adopt Faster R-CNN as the standard for the OBD task. To generate high-quality scene graphs from images, a series of works have been explored in several directions, such as task-specific [34], [43], [44], [45], [46] and dataset-specific [41], [47], [48]. Notably, aggregating contextual information between objects has been shown to be effective for SGG. Xu et al. [34] first combined contextual information to refine the features of objects and relationships by using GRU to pass messages between graphs. Based on the discovery that there are some inherent relationship patterns in subgraphs, Motifs [43] constructed a new global context computing mechanism. A series of context-based studies [49], [50], [51] have consecutively explored the messaging mechanisms of objects and relationships. Most recently, HetSGG [46] attempted to capture the relationships between objects in more detail by proposing a heterogeneous graph and proved its effectiveness. However, the above methods of exhaustive enumeration based on all pairs tend to result in excessive computation, which makes it difficult to run under common computational resources (e.g.,

24GB of memory in a single GPU) for large-size VHR SAI with a large number of objects and more object pairs.

2) *Scene Graph Generation Methods in SAI:* Limited by a few semantic understanding datasets in SAI, there are only a few approaches related to SGG in SAI. Shi and Zou [52] proposed a framework for SAI captioning by exploring whether machines can automatically generate human-like linguistic descriptions of SAI, and this work provides an initial exploration of high-level understanding of SAI. Chen et al. [19] proposed a geospatial relationship triplet representation dataset (GRTRD) based on the characteristics of high-resolution SAI and then adopted the "object-relationship" message-passing mechanism to realize the prediction of geographic objects and geospatial relationships. Li et al. [20] proposed a multi-scale semantic fusion network (MSFN) for SGG in SAI based on their dataset RSSGD, which integrates global information through a graph convolution network to improve the accuracy of SGG. However, the above methods are mainly designed for small-size SAI with HBB and do not take into account the characteristics of large-size VHR SAI, such as the extreme aspect ratio of the objects, the great scale variation among the objects, and the random orientation of the objects. Furthermore, due to the limitation of the dataset, these methods do not have the ability to predict relationships between objects separated at long-range in large-size VHR SAI.

III. DETAILS OF THE STAR DATASET

A. Image Collection and Annotation

1) *Image Collection:* To meet the needs of practical applications, images in the STAR dataset are collected from Google Earth, with a spatial resolution ranging from 0.15m to 1m. Considering the value of cognitive understanding of large-size VHR SAI scenarios in practical applications closely related to human activities, such as transportation, energy, and life, we collected scenarios from more than 1,200 airports, ports, wind power stations, nuclear power stations, thermal power stations, construction sites, sports, service areas, toll stations, traffic bridges, and dams from all over the world. Unlike objects in natural imagery, objects in SAI are oriented in a variety of directions. Therefore, HBB-based annotation cannot provide accurate spatial information for oriented objects. To wrap objects more accurately and develop more suitable algorithms for oriented objects in SAI, all object instances in the STAR dataset are annotated with OBB. Samples of annotated instances in the STAR dataset can be seen in Fig. 2.

2) *Annotation Criteria:* The construction of the SGG dataset for large-size VHR SAI mainly contains two tasks: object annotation and relationship annotation. Referring to the category system of mainstream OBD datasets in SAI, and combining with the actual requirements of the SGG task and some downstream tasks such as visual question answer and image caption, we carry out detailed interpretation and value analysis of the object categories appearing in the above scenarios, and finally choose 48 objects categories as shown in Fig. 3(a). Specifically, we select to annotate "smoke" and "vapor" because they can be used as a basis for judging the working status of power plants. In addition, the collection and annotation of objects like "containmentvessel",

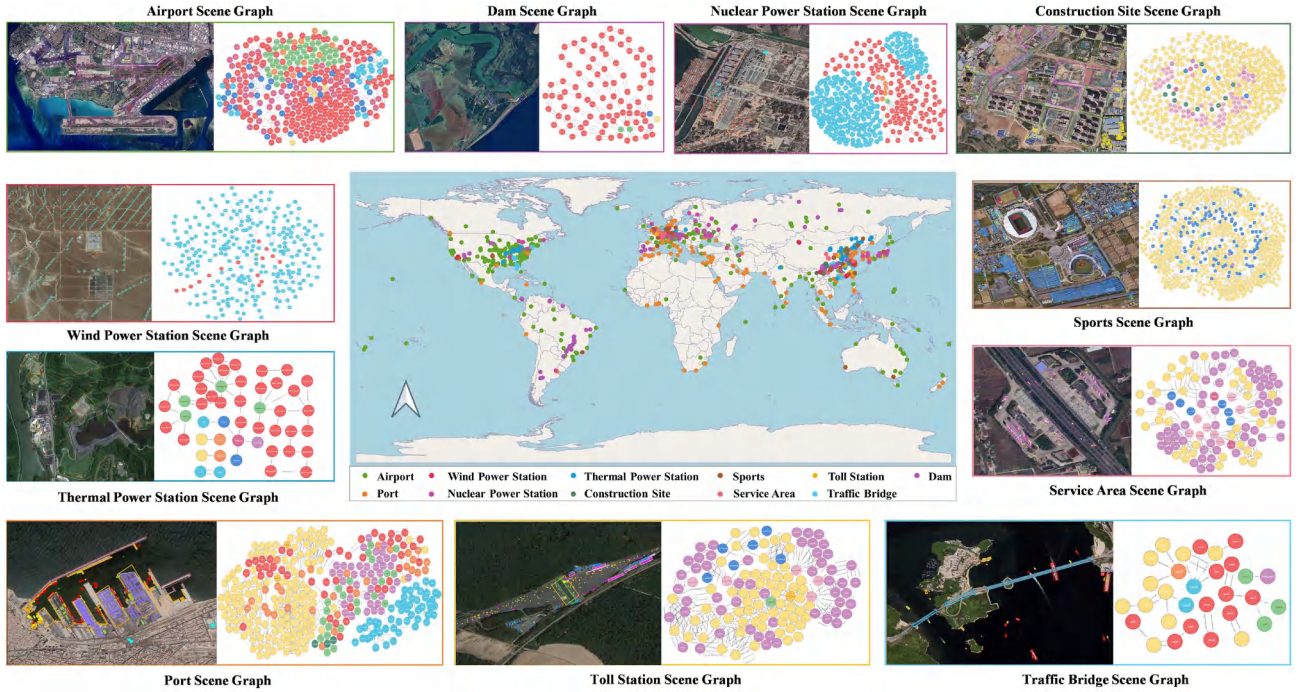


Fig. 2. The geographical distribution map of the sampled images from the proposed STAR dataset.

”latticetower”, ”substation” and ”gravitydam” are expected to play an important role in SAI-based electricity research [18].

For the relationship description, unlike the “human-eye view” perspective of natural imagery, the “birds-eye view” perspective in SAI makes the semantic relationships between objects absolute, so we discard the orientation words such as “east”, “south”, “west”, and “north”. Moreover, we have learned a lesson from the ambiguous labeling of SGG datasets in natural imagery and removed relationship descriptions whose semantics are obviously ambiguous. After careful screening, the semantic relationships are defined into 8 major categories: “distance warning”, “spatial topology”, “functional description”, “circuit layout”, “movement status”, “emission status”, “construction status” and “parking status”, which contain 58 subcategories as in Fig. 3(b).

Fig. 4 gives the interaction mapping between objects and relationships. There are eight colors, representing the eight major categories of relationships consistent with Fig. 3(b). It can be seen that almost every relationship category is associated with multiple object categories, which reflects the complex interactions between objects in large-size VHR SAI.

3) *Annotation Management*: STAR dataset has a standardized dataset construction pipeline: pre-annotation and rules refinement stage, large-scale detailed annotation stage, and quality-checking stage. In the pre-annotation stage, 6 experts in the SAI field are first invited to formulate annotation rules, and then the main authors pre-annotate the images with 11 types of scenarios, which are examined and evaluated by these experts, thereby forming a detailed object and relationship annotation document. During the large-scale annotation stage, the main authors provide comprehensive training to 9 professionals with rich interpretation experience in SAI, and each image is annotated by

2 professionals and checked by 1 professional at the same time. During the quality-checking stage, cross-sampling and revisions were performed by main authors.

Specifically, we use RoLabelImg4 to manually generate fine-grained oriented labeled boxes for the objects. Each object is represented by four clockwise-aligned angular coordinates $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ and category labels, and each triplet is represented by $\langle \text{subject}, \text{relationship}, \text{object} \rangle$. For a more meaningful annotation, all pre-annotated pairs (both long-range and short-range) are selected based on the interaction graph in Fig. 4 during the relationship annotation process.

B. Data Characteristics and Analysis

1) *Data Characteristics*: Compared with the existing OBD dataset in SAI, the STAR dataset has unique properties in task diversity, category diversity, and scale diversity. Sparse high-value pairs, high intra-class variation and inter-class similarity are also important characteristics of the STAR dataset.

Task Diversity: STAR supports both the OBD and SGG tasks in large-size VHR SAI. After processing the annotated images, it can further satisfy the various task requirements like OBD based on HBB/OBB in large-size VHR SAI, SGG based on HBB/OBB in large-size VHR SAI.

Category Diversity: STAR has various scene categories, various object categories, and various relationship categories. Specifically, it contains 11 categories of VHR SAI scenarios over the world, 48 categories of important objects, and 58 categories of high-value relationships, which will bring new opportunities for cognitive understanding of SAI.

Scale Diversity: In the STAR dataset, there are many typical objects with extreme aspect ratios (e.g., runways, bridges,

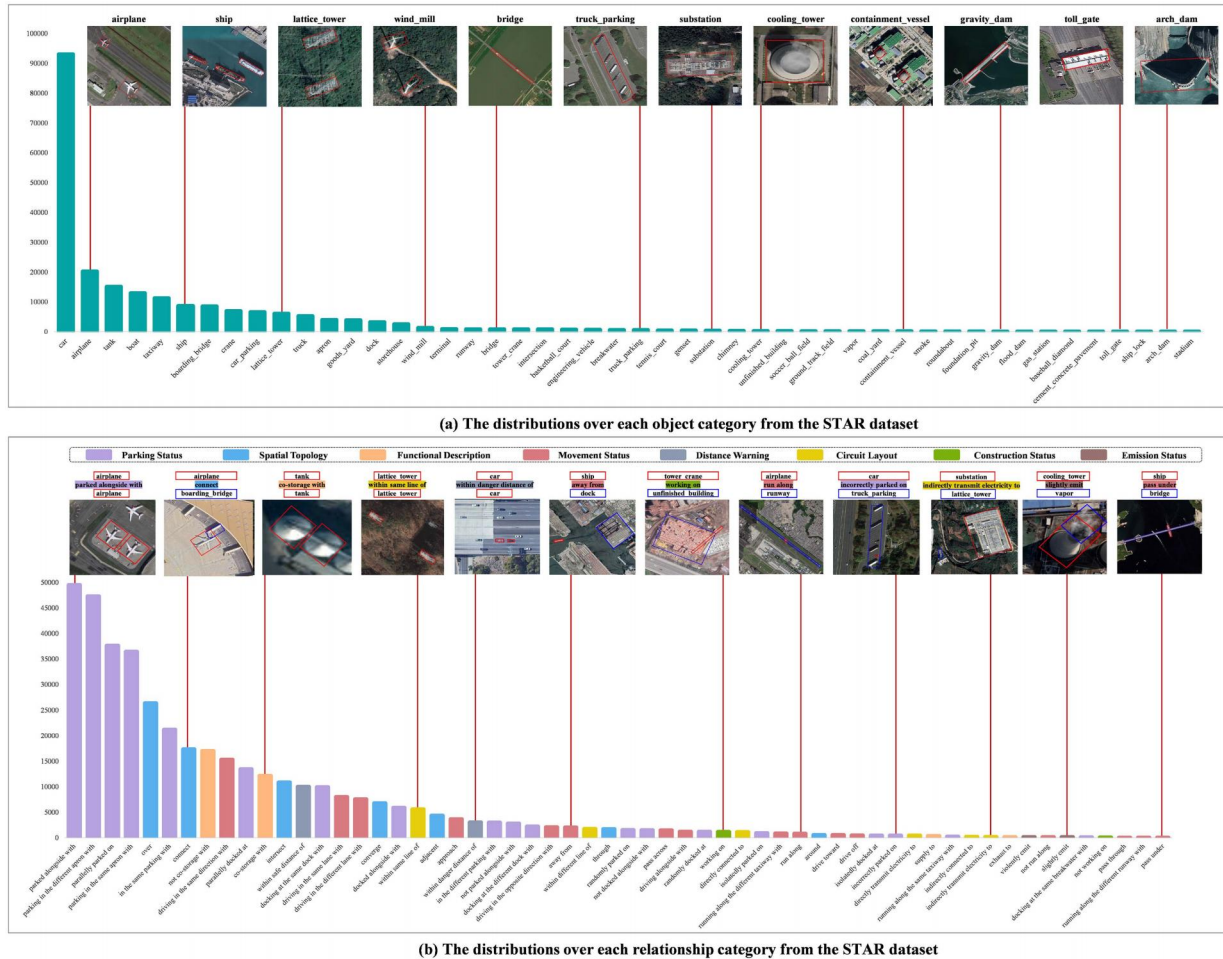


Fig. 3. Statistics and visualization of objects (a) and relationships (b) from the STAR dataset. The relationships are color-coded to show parking status, spatial topology and functional description, movement status, distance warning, circuit layout, construction status and emission status. Some typical objects and interesting triplets are visualized.

gravitydams), and there are large scale variations between objects of the same category (e.g., airplanes, ships) and different categories (e.g., “boats and docks”, “cars and carparkings”) that contain rich knowledge, which will bring great challenges to the OBD and SGG tasks.

Sparse High-Value Pairs: There are an enormous number of objects in a large-size SAI. If relationship prediction is performed for all object pairs, more than $N(N-1)$ triplets will appear for N objects in the case of one-to-many relationships (e.g., in the STAR dataset where the number of objects in one image is up to 6,800, the possible triplets will be more than 40,000,000), which will cause memory overflow under common computational resources. In fact, not all object pairs are of interest to us, and only a few of them contain high-value relationships. Therefore, it will be a new challenge to select high-value pairs from the numerous pairs for the SGG task in large-size VHR SAI.

High Intra-Class Variation and Inter-Class Similarity: Intra-class variation of relationships is mainly due to the differential appearance and diverse subject-object pairs. The inter-class similarity of relationships arises from similar appearance representations, but manifests itself differently for different relationship

categories. As shown in Fig. 5, for the object pairs of the same type, the relationship labels between them are non-unique (e.g., $\langle \text{ship}, \text{away from}, \text{dock} \rangle$, $\langle \text{ship}, \text{approach}, \text{dock} \rangle$), and for the object pairs of different types, their relationship labels may be the same (e.g., $\langle \text{ship}, \text{away from}, \text{dock} \rangle$, $\langle \text{truck}, \text{away from}, \text{toll gate} \rangle$).

2) Data Statistics and Visualization: The statistical results of object and relationship annotations from the STAR dataset are shown in Fig. 3(a) and (b), respectively. It can be seen that there is an obvious long-tail effect for both object and relationship categories in the STAR dataset, which is also consistent with reality. The highest-frequency objects are common small objects such as “cars”, “airplanes”, “tanks” and “boats”. The highest-frequency relationships are mainly related to the parking states, such as “parked alongside with”, “parking in the different apron with”, “parallelly parked on” and “parking in the same apron with”.

IV. THE PROPOSED METHOD

Fig. 6 presents the overall architecture of the proposed CAC framework, which mainly contains three parts: OBD, pair pruning and relationship prediction. Considering the drastic changes

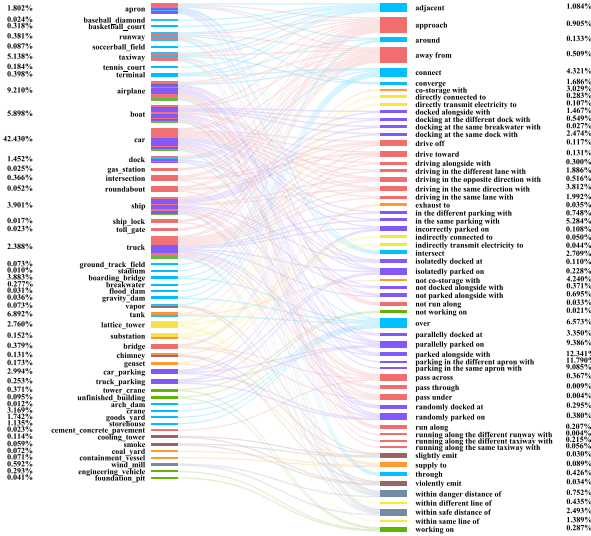


Fig. 4. Interaction mapping between objects and relationships. There are eight colors in the figure, which represent the types of eight relationships: parking status (in purple), spatial topology (in blue) and functional description (in orange), movement status (in red), distance warning (in grey), circuit layout (in gold), construction status (in green) and emission status (in brown). The values on either side indicate the proportion of each object category and relationship category, respectively.



Fig. 5. Examples of intra-class variations and inter-class similarities in relationship on the STAR dataset.

in the scales and aspect ratios of objects in large-size VHR SAI, a HOD-Net that flexibly integrates multi-scale contexts is proposed for the holistic detection of multi-class objects. Facing the problem of memory overflow caused by pair redundancy in large-size VHR SAI under common computational resources, we propose the PPG network via adversarial reconstruction to obtain object pair ranking scores, which can efficiently sift out contextual interaction pairs containing rich knowledge under the condition of no-negative samples. Considering the dependence of relationship prediction in large-size SAI on the contexts of object pairs, the RPCM network for predicting relationship types of these pairs is proposed, which enhances the cognitive ability of the model by fusing the contexts of objects and relationships in triplets. Based on the three levels in the CAC framework, the definition of the SGG task in large-size VHR SAI is presented here. Specifically, viewing the object entities as nodes of the scene graph, and the relationships between the entities as directed edges connecting different entities, the SGG task in

large-size VHR SAI can be described as:

$$P(G_{SA} | I) = P(E | I)P(E_S, E_O | E, I)P(R | E_S, E_O, I), \quad (1)$$

where I denotes large-size SAI, G_{SA} denotes the scene graph in large-size SAI, E is object entities, and R is the relationships between subjects E_S and objects E_O .

A. Oriented Object Detection in Large-Size VHR SAI

OBD in large-size SAI is an indispensable basis for SGG in large-size SAI, which directly affects the prediction accuracy of the triplets. Compared with small-size or low-resolution imagery, large-size VHR SAI can not only accommodate more complete large objects (e.g., runways, docks), but also greatly retain important details of small objects (such as the head and tail of a car, the head and tail of an airplane), which can help to mine more high-value knowledge in SAI. In processing large-size VHR SAI, mainstream deep learning-based OBD methods, whether supervised [53], [54], [55], [56] or weakly supervised [57], [58], [59], [60], adopt a cropping strategy [4], [61], which results in large objects, such as runways, being cut off. In addition to the cropping strategy, some OBD methods tackle the original large-size imagery with a fixed-window down-sampling strategy [62], resulting in the loss of massive detail information of the image. Recently, HBD-Net for single-class bridge detection in large-size VHR SAI is proposed [26], and the SDFF architecture in HBD-Net effectively solves the large scale variations in single-class bridge detection by fusing multi-scale contexts via the dynamic image pyramid (DIP) of the large-size image. However, this single-class OBD only considers the intra-class variations of the objects, which is not applicable in the multi-class OBD task with both intra-class and inter-class variations.

To address the aforementioned problem, it extends HBD-Net to the holistic multi-class OBD in large-size VHR SAI, named HOD-Net whose total loss of the HBB-based and OBB-based detectors in large-size VHR SAI is defined as:

$$\mathcal{L}_O = \sum_{m=1}^M \left(\frac{1}{\Gamma^m} \sum_{i \in \Delta^m} \mathcal{L}_i^{cls} + \frac{1}{\Gamma^{m+}} \sum_{j \in \Delta^{m+}} w_j^{reg} \mathcal{L}_j^{reg} \right), \quad (2)$$

$$\mathcal{L}_H = \sum_{m=1}^M \left(\frac{1}{\Gamma^m} \sum_{i \in \Delta^m} \mathcal{L}_i^{cls} + \frac{1}{\Gamma^{m+}} \sum_{j \in \Delta^{m+}} \mathcal{L}_j^{reg} \right), \quad (3)$$

where m means the layer of the DIP, $m \in [1, M]$. Δ^m and Δ^{m+} are the set of all samples and the set of positive samples, respectively. Γ^m and Γ^{m+} indicate the total number of all samples and positive samples in the m -th layer, respectively. w_j^{reg} denotes the regression weight obtained using the SSRW strategy [26]. Regression loss \mathcal{L}_j^{reg} is set with reference to the smoothed L1 loss defined in [63].

To enable the model to adaptively focus on different object categories in different layer, a hierarchical adaptive weighted strategy is designed in classification loss $\mathcal{L}_i^{cls} = -t_c \log(e^{\Psi_i^c} / \sum_{c=1}^C e^{\Psi_i^c})$. Specifically, t_c represents the one-hot vector, and $\Psi_{wei}^m = \mathbf{w}^m \odot \Phi^m$ denotes the class-weighted

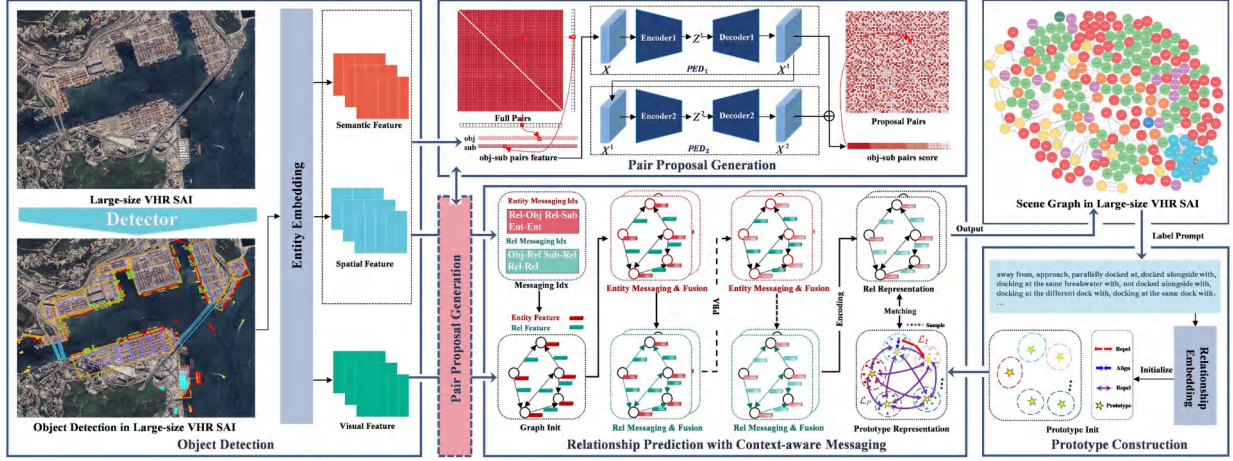


Fig. 6. The overall architecture of the CAC framework. First, multi-scale objects in large-size VHR SAI are holistically detected by HOD-Net. Then, contextual interaction pairs containing rich knowledge are selected by the PPG network via adversarial reconstruction. Finally, high-value relationships between subject and object are predicted by the RPCM network.

score. $\mathbf{w}^m = [w_1^m, \dots, w_C^m]$ is the learnable weight, and $\Phi^m = [\phi_1^m, \dots, \phi_C^m]$ denotes the confidence of the network. \mathcal{C} is the number of object categories, and \odot is the element-wise product.

B. Pair Proposal Generation via Adversarial Reconstruction

Given N objects in the image, the number of subject-object pairs is $N(N-1)$, which exhibits exponential growth as N increases. There are a huge amount of subject-object pairs in a large-size VHR SAI, and the pair redundancy problem will cause the model to be ineffective under common computational resources due to memory overflow. Some researchers regard pair pruning as a binary classification problem for the SGG task in natural imagery, i.e., annotated object pairs are regarded as positive samples, and non-annotated object pairs are regarded as negative samples. It is worth noting that it is difficult to exhaust the annotation of triplets for large-size VHR SAI, and most of the non-annotated object pairs still have high-value semantic relationships, so it is unreasonable to directly regard the non-annotated object pairs as negative samples. To address this problem, this paper proposes a pair proposal generation (PPG) network via adversarial reconstruction, which can realize the ranking of triplets according to their score based on their containment of high-value knowledge without negative samples, and effectively addresses the issue of selecting meaningful pairs with contextual associations in the case of only positive samples for the SGG task in large-size VHR SAI.

To enable the model to autonomously learn the reconstruction ability for high-value pairs without negative samples, the PPG network is divided into two-stage adversarial training with pair encoder-decoder (PED), because the auto-encoder can achieve stability and reconstruct the inputs well during adversarial training. The PPG network contains two encoder networks: encoder1 (E_1), and encoder2 (E_2) as well as two decoder networks: decoder1 (D_1), and decoder (D_2), which constitute the two auto-encoders PED_1 and PED_2 , as shown in Fig. 6. Two PED are trained to reconstruct the input feature X , which represents the union of spatial and semantic features for object pairs.

Similar to the gaming strategy in generative adversarial network (GAN) [64], PED_1 and PED_2 are trained in an adversarial manner, in which PED_1 tries to deceive PED_2 , while PED_2 tries to distinguish whether the data is the input X or X^1 reconstructed by PED_1 . During training stage, the feature X is squeezed into the latent space Z by E_1 , and then reconstructed by D_1 , which can be represented as $D_1(E_1(X))$. PED_2 is trained by $D_2(E_2(X^1))$ to distinguish whether the data is the original input X or from PED_1 , during which PED_1 maintains learning to deceive PED_2 . Overall, PED_1 and PED_2 similar to minimal-maximal games:

$$\zeta = \min_{PED_1} \max_{PED_2} G \|X - PED_2(PED_1(X))\|_2. \quad (4)$$

The training objective of PED_1 is to minimize the reconstruction error $\|X_i - X_i^1\|_2$ to learn the latent representation of the data, and the training objective of PED_2 is to reduce the reconstruction error $\|X_i - X_i^2\|_2$ to a minimum. As the iterations n are constantly changing, the loss is:

$$\mathcal{L}_{PED} = \frac{1}{n} \|X - X^1\|_2 + \left(1 - \frac{1}{n}\right) \|X - X^2\|_2. \quad (5)$$

In the inference stage, n in (5) is set to 2, and then the top k_1 proposal pairs are selected as input indices of subject-object pairs for relationship prediction.

C. Relation Prediction With Context-Aware Messaging

For large-size VHR SAI, any two objects can often be directly associated through significant relationships or indirectly associated with other medium objects, so the introduction of context-aware messaging from objects and relationships is essential to enhance the cognitive ability of the model. In addition, numerous possible subject-object pairs present different visual appearances, resulting in large intra-class variation among the same relationships, which pose critical challenges for SGG in large-size VHR SAI. This paper proposes a relationship prediction network with context-aware messaging (RPCM), which mainly consists of two components: a progressive bi-context

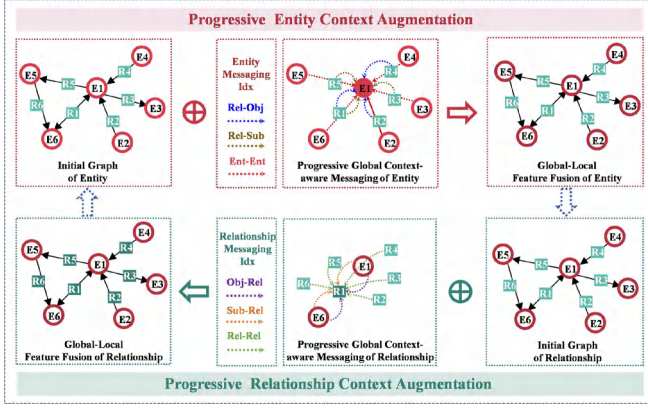


Fig. 7. The proposed PBA module: the details of messaging and fusion for one progressive context augmentation of entity-relationship.

augmentation component and a prototype-guided relationship learning component. The former uses context messaging from nodes and edges to augment the local representations of objects and relationships. Meanwhile, the latter continuously optimizes and updates the semantic prototype under instance-level and prototype-level constraints. Ultimately, accurate relationship prediction is achieved by matching the predicted relationship representation with the prototype representation in the same semantic space.

1) *Progressive Bi-Context Augmentation Network*: In the sparse graph formulated based on the proposed PPG network, the semantic representations of both objects and relationships are influenced by neighboring objects and relationships. However, the interplay between object-object, object-relationship, and relationship-relationship interactions may exhibit substantial diversity in different contexts. Attention graph network [65], which leads the model to focus on reliable context information by assigning the corresponding weights to different regions of an image, has shown its effectiveness for SGG in natural imagery [46], [50]. Nonetheless, due to the complexity of large-size VHR SAI, conventional SGG models often exhibit deficiencies in relationship context inference capabilities when processing SAI. Moreover, although the integration of global context information assists in enhanced relationship recognition, more fine-grained discriminative features are primarily extracted from the triplets themselves. An overabundance of global context information may induce significant confusion. In our work, a novel progressive bi-context augmentation (PBA) component is proposed. As shown in Fig. 7, PBA mainly consists of entity-relationship progressive global context-aware messaging and global-local feature fusion. The progressive global context-aware messaging module for entity-relationship adaptively learns global information from neighboring entities and relationships using a progressive strategy. The global-local feature fusion module fuses the collected global and local features to obtain augmented features. After several iters of nodes and edges in the sparse graph, the dual optimization and updating of entity and relationship representations are accomplished.

Progressive Global Context-Aware Messaging: Considering that it is easy to confuse when introducing too much unnecessary

context information, we adopt a progressive training strategy to solve this problem. In this strategy, all entities and relationships first collect reliable information from their direct neighboring nodes and edges via the attention graph network, and then learn their indirectly related contexts with the help of neighboring nodes and edges, so that each entity and relationship can learn more reliable contextual information without introducing extra noise.

Sparse scene graphs constructed based on the proposed PPG network can provide multiple types of messaging indexes for entities and relationships. For $i + 1$ iters, different types of messaging (i.e., entity-entity, relationship-subject, and relationship-object) are performed on entities to learn context-aware information of entities:

$$\hat{F}_{i+1}^e = \sigma(w^{ee} F_i^e \alpha^{ee} + w^{rs} F_i^r \alpha^{rs} + w^{ro} F_i^r \alpha^{ro}), \quad (6)$$

where $\sigma(\cdot)$ denotes the sigmoid activation function [66]. α^{ee} , α^{rs} and α^{ro} represent the importance of messaging from entity to entity, relationship to subject, and relationship to object, respectively. By learning to adjust α [65], each iteration leads to a change in attention and affects subsequent iters. w^{ee} , w^{rs} and w^{ro} denote the weight matrix of the directed messaging from entity to entity, relationship to subject, and relationship to object, respectively. F_i^e and F_i^r are entity and relationship representation, respectively.

Similarly, different types of messaging (i.e., relationship-relationship, subject-relationship, object-relationship) on relationships are used to learn the global contextual information of relationships:

$$\hat{F}_{i+1}^r = \sigma(w^{rr} F_i^r \alpha^{rr} + w^{sr} F_i^e \alpha^{sr} + w^{or} F_i^e \alpha^{or}). \quad (7)$$

Global-Local Feature Fusion: Introducing reliable global information helps to better recognize relationships, but more fine-grained discriminative features mainly come from local. Therefore, we adopt a global-local feature fusion strategy to capture reliable global information and maintain the local more discriminative features. Specifically, the context augmented feature of entity F_{i+1}^e can be obtained by fusing the context-aware feature of entity \hat{F}_{i+1}^e from global and the entity feature F_0^e from local. Similarly, the context augmentation feature of relationship F_{i+1}^r can be obtained. After L iters, the predicted relationship representations Rel with context-aware augmentation are updated. Note that the predicted relationship representations in Section IV-C1 are optimized and updated simultaneously with the prototype representations in Section IV-C2.

2) *Prototype-Guided Relationship Learning*: Prototype learning allows SGG models to focus on all categories equally in training, and is effective in alleviating the category imbalance problem of datasets [45]. However, previous linear classifier-based methods often lack the ability to model data at different semantic levels and suffer from a lag in representational capabilities (e.g., $\langle ship, away from, dock \rangle$ and $\langle ship, approach, dock \rangle$ correspond to different relationship labels, but it is clear that $\langle ship, away from, dock \rangle$ and $\langle ship, approach, dock \rangle$ are closer than $\langle truck, away from, gas station \rangle$ in the visual representation). We

propose a prototype-guided relationship learning method that utilizes category labels as prompts to help generate learnable relationship prototypes, and prototype representations updated dynamically by instance-level constraints and prototype-level constraints to help the SGG model enhance the discriminability between prototypes.

Prototype Construction: To construct more discriminative prototype representations, word embeddings (Glove) of different category labels $[label_1, label_2, \dots, label_C]$ are utilized to construct initial class-specific semantic prototypes $\mathbf{T} = [t_1, t_2, \dots, t_C]$. After obtaining the initial prototype representations, they are fed into a learnable prototype encoder Enc_p , during which the prototype representations $Pro = Enc_p(\mathbf{T})$ are allowed to be updated and matched with the predicted relationship representations Rel in the same semantic space by $r = MAP(Rel)$ and $p = MAP(Rro)$, where $MAP(\cdot)$ is two-layer MLPs.

Instance-Wise Matching Loss: To enable the model to learn more fine-grained category variation, the matching constraints between samples and prototypes are injected into the learning objective during instance-level matching to achieve positive samples closer to the prototypes and negative samples further away from the prototypes. Specifically, we obtain instance-level loss \mathcal{L}_{IC} and \mathcal{L}_{ID} based on cosine distance and euclidean distance, respectively:

$$\mathcal{L}_{IC} = -\log \frac{\exp(\langle \bar{r}, \bar{p} \rangle / \tau)}{\sum_{i=0}^C \exp(\langle \bar{r}, \bar{p}_i \rangle / \tau)}, \quad (8)$$

$$\mathcal{L}_{ID} = \max(0, q^+ - q^- + \gamma_1), \quad (9)$$

where $q^- = \|r^{Neg} - \bar{p}_i\|_2^2$, $q^+ = \|r^{true} - \bar{p}_i\|_2^2$. r^{true} denotes positive samples with label annotations, and r^{Neg} is the negative samples sampled in order of distance with the exclusion of positive samples.

Prototypical Matching Loss: The purpose of prototype contrast learning is to make samples of the same category tightly clustered around the corresponding prototype to form tight clusters, with clusters corresponding to different prototypes pulling away from each other. To achieve this goal, we utilize the semantic structure captured by relationship prototypes to realize prototype contrast learning loss \mathcal{L}_{PC} and loss \mathcal{L}_{PD} .

$$\mathcal{L}_{PC} = \|\bar{p} \cdot \bar{p}^T\|_{2,1}, \quad (10)$$

$$\mathcal{L}_{PD} = \max\left(0, -\mathcal{N}_{select}^k\left(\|\bar{p}_i - \bar{p}_j\|_2^2\right) + \gamma_2\right), \quad (11)$$

where $\mathcal{N}_{select}^k(D_M)$ denotes the selection of the top k smallest values in the distance matrix D_M .

The overall training loss \mathcal{L} of our RPCM is as follows, without any bias on the weight of each term:

$$\mathcal{L} = \mathcal{L}_{IC} + \mathcal{L}_{ID} + \mathcal{L}_{PC} + \mathcal{L}_{PD}. \quad (12)$$

In the inference stage, the prototype type with the highest cosine similarity s_i is used for relationship prediction:

$$R_{class} = \underset{i}{\operatorname{argmax}} (s_i \mid s_i = \langle \bar{r}, \bar{p}_i \rangle / \tau), \quad (13)$$

where \bar{r} and \bar{p} are both regularization operations.

V. EXPERIMENTS AND ANALYSIS

A. Benchmark

Task Settings: For the OBD task, we establish two types of OBD benchmarks: HBB-based and OBB-based detectors in large-size VHR SAI. For the SGG task, we evaluate SGG in large-size VHR SAI following three conventional sub-tasks of SGG in natural imagery: Predicate Classification (PredCls), Scene Graph Classification (SGCls) and Scene Graph Detection (SGDet).

Evaluation Metrics: For the OBD task, mean average precision (mAP) is adopted as the evaluation metric. For the SGG task, different from the one-to-one relationship prediction in natural imagery, the relationship prediction for the SGG task in large-size VHR SAI is one-to-many. Based on the common evaluation metrics recall@K (R@K) and mean recall@K (mR@K) of the SGG task in natural imagery [41], we propose multi-label recall@K (MR@K) and mean multi-label recall@K (mMR@K) as the multi-label evaluation metrics for SGG in large-size VHR SAI.

To evaluate the performance of the SGG model more comprehensively, the harmonized mean (HMR@K) of both MR@K and mMR@K is taken as the comprehensive evaluation metric for the SGG task in large-size VHR SAI. For all tasks, the prediction can be considered as a true positive (TP) result only when it matches the ground-truth triplet $\langle \text{subject}, \text{relationship}, \text{object} \rangle$ and has at least 0.5 IoU between the bounding boxes of subject-object and the bounding boxes of ground-truth. As for the K, instead of the setting (K= 50/100), which is commonly used by the SGG task in natural imagery, we adopt the setting (K= 1,500/2,000) due to the average number of triplets per large-size VHR SAI far exceeds that in natural imagery.

B. Implementation Details

Object Detector: Taking the number of instances and the type of scenarios as the basis of division, we split the dataset into train set (60%), val set (20%), and test set (20%). The detectors are trained on STAR train set using AdamW [95] as a default optimizer, and the performance is reported on STAR test set. All the listed models are trained on a single NVIDIA RTX 3090 GPU. We set the batch size to 2 and the initial learning rate to 1×10^{-4} . The mean average precision (mAP) is calculated following PASCAL VOC 07 [96].

Scene Graph Generation: On top of the frozen detectors, we train all SGG models on STAR train set using SGD as an optimizer on a single GPU and evaluate them on STAR test set. Batch size and initial learning rate are 4 and 1×10^{-3} for the PredCls and SGCls tasks, and 2 and 1×10^{-3} for the SGDet task, respectively. For the SGDet task, object detection is performed using a Per-Class NMS [41], [43] with 0.5 IoU. Considering the computational resource consumption caused by numerous invalid object pairs in large-size VHR SAI, we use the proposed PPG network in the overall framework, which enables the SGG task in large-size VHR SAI to be realized in common computational conditions (e.g., one single GPU).

TABLE II

BASELINE RESULTS (%) OF HBB-BASED AND OBB-BASED DETECTORS ON STAR TEST SET. A_P DENOTES AIRPLANE, B_B IS BOARDING_BRIDGE, L_T IS LATTICE_TOWER, B_D IS BRIDGE, R_W IS RUNWAY, S_L IS SHIP_LOCK, AND G_D REPRESENTS GRAVITY_DAM. ALL EXPERIMENTS ARE BASED ON THE STANDARD '1x' (12 EPOCHS) TRAINING SCHEDULE. † MEANS THAT SWIN-L [67] IS USED AND OTHERS INDICATES RESNET50 [68] AS THE BACKBONE

OBD	Detectors	Venue	ship	boat	truck	car	a_p	crane	b_b	tank	l_t	b_d	r_w	s_l	dock	...	g_d	mAP
HBB	Faster R-CNN [69]	NeurIPS'15	30.7	7.2	26.2	38.1	<u>54.3</u>	23.6	24.7	<u>21.1</u>	21.9	5.3	0.0	0.0	4.2	...	0.0	32.2
	RetinaNet [70]	ICCV'17	26.1	4.6	25.6	35.5	52.6	20.1	21.0	18.4	21.9	3.3	0.0	0.0	2.8	...	0.0	24.6
	Cascade R-CNN [71]	TPAMI'19	32.0	<u>7.8</u>	29.3	39.1	55.0	<u>26.4</u>	<u>25.6</u>	20.9	22.1	6.0	0.0	0.0	5.6	...	0.0	32.4
	FCOS [72]	TPAMI'20	20.0	2.3	23.7	34.6	48.0	14.8	17.4	17.6	11.5	1.2	0.0	0.0	2.0	...	0.0	20.8
	TOOD [73]	ICCV'21	29.3	5.5	<u>30.8</u>	<u>38.4</u>	53.6	23.5	24.8	19.8	20.6	4.0	0.0	<u>3.3</u>	3.8	...	1.2	30.1
	GCL [62]	KBS'23	30.6	7.2	26.6	38.1	<u>54.3</u>	23.5	24.5	<u>21.1</u>	21.8	5.4	6.6	0.0	5.0	...	2.2	34.4
	HOD-Net (Ours)	-	<u>32.5</u>	7.6	28.5	37.5	53.3	25.5	24.5	21.8	<u>22.8</u>	<u>9.7</u>	<u>9.3</u>	0.9	<u>12.3</u>	...	<u>37.7</u>	<u>45.2</u>
	HOD-Net† (Ours)	-	35.4	11.8	31.9	36.7	54.0	33.5	27.1	20.8	23.1	15.7	20.9	4.7	15.4	...	40.3	53.7
OBB	Deformable DETR [74]	ICLR'21	18.1	5.5	15.8	42.5	85.1	12.6	32.7	51.1	21.2	0.5	0.0	0.0	2.1	...	0.0	17.1
	ARS-DETR [75]	TGRS'24	44.6	16.4	36.1	60.8	88.6	40.5	59.0	54.7	41.6	8.2	0.0	0.0	7.8	...	0.0	28.1
	RetinaNet [70]	ICCV'17	39.3	14.3	36.7	65.3	88.5	42.9	43.1	50.8	44.8	2.1	0.0	0.0	5.2	...	0.0	21.8
	ATSS [76]	CVPR'20	38.4	15.5	31.4	67.6	89.0	35.6	39.2	52.6	43.2	5.1	0.0	0.0	5.1	...	0.0	20.4
	KLD [77]	NeurIPS'21	43.1	12.7	39.1	65.7	88.7	47.2	50.0	51.2	48.0	3.9	0.0	0.0	12.7	...	0.0	25.0
	GWD [78]	ICML'23	42.5	14.5	38.4	65.8	89.1	52.2	54.0	51.0	49.2	4.4	0.0	0.0	12.6	...	0.0	25.3
	KFIoU [79]	ICLR'23	41.6	14.1	39.5	67.1	89.4	47.7	54.6	51.4	47.0	8.3	0.0	0.0	9.2	...	4.5	25.5
	DCFL [80]	CVPR'23	47.3	10.2	45.6	69.7	89.2	56.8	52.9	55.2	49.2	13.2	0.0	0.0	11.3	...	2.3	29.0
	R³Det [81]	AAAI'21	44.0	15.3	45.3	68.6	89.6	52.5	43.5	53.6	47.5	4.3	0.0	0.0	11.2	...	0.0	23.7
	S²A-Net [82]	TGRS'21	46.6	13.9	45.1	69.3	89.7	52.2	52.4	57.9	48.9	13.5	0.0	0.0	11.7	...	0.0	27.3
	RepPoints [83]	ICCV'19	26.3	5.6	27.1	62.5	89.0	40.3	35.3	54.2	38.6	9.1	0.0	0.0	4.9	...	0.0	19.7
	CFA [84]	CVPR'21	45.6	15.1	44.3	66.8	88.8	57.3	48.6	56.1	46.4	2.9	0.0	0.0	13.7	...	0.0	25.1
	Oriented RepPoints [85]	CVPR'22	48.0	16.9	43.8	67.9	89.3	59.0	49.7	55.1	46.5	9.1	0.0	0.0	11.4	...	0.0	27.0
	G-Rep [56]	RS'23	43.3	17.2	47.2	68.5	89.2	53.3	54.1	54.7	40.5	12.5	0.0	0.0	4.6	...	0.0	26.9
	SASM [86]	AAAI'22	40.6	10.4	42.6	68.1	89.0	56.3	41.8	51.0	44.1	14.5	0.0	0.0	14.3	...	0.0	28.2
	FCOS [87]	ICCV'19	43.3	12.7	44.9	68.2	89.5	56.9	53.6	52.9	44.4	12.7	0.0	0.0	12.9	...	0.0	28.1
	CSL [88]	ECCV'20	43.8	12.6	45.5	67.3	89.1	55.5	52.1	49.9	44.3	10.3	0.0	0.0	8.9	...	0.0	27.4
	PSC [89]	TPAMI'24	44.6	18.0	44.3	66.2	89.6	56.4	55.4	53.1	44.8	16.1	0.0	0.0	13.9	...	0.0	30.5
	H2RBox-v2 [58]	NeurIPS'23	43.7	12.2	44.0	67.7	89.3	50.7	51.4	49.2	45.4	15.7	0.0	0.0	12.2	...	0.0	27.3
	Faster R-CNN [69]	NeurIPS'15	50.4	20.9	48.1	68.0	89.6	20.9	54.4	55.0	48.6	18.8	0.0	0.0	14.0	...	0.0	32.6
	Gliding Vertex [90]	TPAMI'20	50.6	23.8	38.6	68.1	89.4	56.9	54.7	54.2	45.9	17.8	0.0	0.0	14.9	...	0.0	30.7
	RoI Transformer [54]	CVPR'19	56.6	26.2	50.3	69.1	89.9	59.7	<u>63.4</u>	57.3	51.0	15.8	0.0	0.0	13.6	...	0.0	35.7
	ReDet [91]	CVPR'21	<u>60.1</u>	25.5	50.0	<u>69.9</u>	90.1	68.0	63.7	60.1	54.4	25.2	0.0	0.0	18.6	...	0.0	39.1
	Oriented R-CNN [92]	ICCV'21	57.6	<u>26.6</u>	51.6	67.8	89.5	65.0	62.8	54.1	49.1	21.2	0.0	0.0	14.8	...	0.1	33.2
	GCL [62]	KBS'23	57.1	<u>26.6</u>	51.5	67.8	89.4	64.2	62.2	54.1	49.1	21.1	0.0	0.0	15.1	...	0.1	33.7
	LSKNet-S [93]	ICCV'23	59.0	25.5	<u>53.2</u>	70.9	89.7	<u>65.5</u>	62.4	53.6	55.0	24.7	0.0	0.0	19.5	...	11.4	37.8
	PKINet-S [94]	CVPR'24	55.2	22.4	48.0	67.5	<u>89.9</u>	59.1	56.3	50.7	47.4	21.1	0.0	0.0	19.0	...	0.2	32.8
	HOD-Net (Ours)	-	57.2	23.5	49.4	64.9	89.5	65.5	56.6	56.5	<u>56.2</u>	<u>25.9</u>	<u>11.3</u>	<u>3.7</u>	<u>20.7</u>	...	<u>48.6</u>	<u>43.6</u>
	HOD-Net† (Ours)	-	65.2	32.6	53.4	66.4	89.7	69.3	63.1	<u>59.3</u>	57.6	38.7	35.7	46.9	29.4	...	78.6	55.9

Bold indicates the best result, and underline indicates second-best result.

TABLE III

RESULTS (%) OF DIFFERENT TRAINING STRATEGIES ON STAR TEST SET. A_P DENOTES AIRPLANE, B_B DENOTES BOARDING_BRIDGE, L_T DENOTES LATTICE_TOWER, B_D IS BRIDGE, R_W IS RUNWAY, S_L IS SHIP_LOCK, AND G_D REPRESENTS GRAVITY_DAM. ALL EXPERIMENTS ARE BASED ON THE STANDARD '1x' (12 EPOCHS) TRAINING SCHEDULE. SWIN-L [67] IS ADOPTED AS THE BACKBONE

OBD	Detectors	ship	boat	truck	car	a_p	crane	b_b	tank	l_t	b_d	r_w	s_l	dock	...	g_d	mAP
HBB	Faster R-CNN† [69] (Resizing)	7.5	0.0	12.6	1.3	1.8	3.4	0.3	0.6	0.4	6.9	29.5	4.7	10.1	...	41.6	33.7
	Faster R-CNN† [69] (Cropping)	35.6	<u>9.5</u>	<u>29.8</u>	37.1	53.1	<u>31.6</u>	27.2	20.9	<u>22.3</u>	<u>14.0</u>	0.0	0.0	<u>10.8</u>	...	10.2	<u>42.6</u>
	HOD-Net† (Ours)	35.4	11.7	31.9	<u>36.6</u>	<u>53.0</u>	32.7	<u>26.6</u>	<u>20.8</u>	23.1	16.6	<u>20.9</u>	4.7	15.4	...	<u>36.3</u>	53.2
OBB	Oriented R-CNN† [92] (Resizing)	22.4	0.3	32.4	8.8	15.6	11.9	4.5	9.1	4.5	14.5	<u>0.8</u>	<u>3.0</u>	<u>23.0</u>	...	<u>68.4</u>	30.9
	Oriented R-CNN† [92] (Cropping)	<u>64.8</u>	<u>31.2</u>	54.9	69.7	<u>89.6</u>	<u>69.1</u>	<u>63.0</u>	<u>57.7</u>	<u>34.3</u>	0.0	0.0	22.2	...	4.5	<u>41.9</u>	
	HOD-Net† (Ours)	65.2	32.6	<u>53.4</u>	<u>66.4</u>	89.7	69.3	63.1	59.3	57.6	38.7	35.7	46.9	29.4	...	78.6	55.9

Bold indicates the best result, and underline indicates second-best result.

C. Results and Analysis for OBD

Comparison With Baselines: To explore the properties of STAR and provide guidelines for future OBD task in large-size VHR SAI, we conduct a comprehensive evaluation of about 30 OBD methods and analyze the results. Table II provides a comprehensive benchmark on STAR, including single/two-stage, CNN/Transformer, anchor-based/anchor-free and supervised/weakly-supervised methods. We adopt MMDetection¹ [97] and MMRotate² [98] as the toolkits and all experiments are based on the standard '1x' (12 epochs) training

schedule. It can be observed that our HOD-Net significantly outperforms other methods under the same backbone (ResNet50 [68]) for both HBB-based and OBB-based detectors, and our HOD-Net† (Swin-L) achieves even better performance. Additionally, we evaluate three OBD strategies (see Table III). Our HOD-Net achieves 53.2% and 55.9% mAP on the HBB-based and OBB-based detectors, respectively (using a 0.5 IoU threshold). Specifically, under the same backbone (Swin-L [67]), it achieves a 10.6%/19.5% improvement in mAP compared to the baseline Faster R-CNN detector (cropping/resizing) in the HBB detection task, and a 12.9%/23.9% improvement in mAP metric compared to the baseline Oriented R-CNN detector (cropping/resizing) in the OBB detection task. Notably, Oriented

¹<https://github.com/ZhuZi24/STAR-MMDetection>

²<https://github.com/yangxue0827/STAR-MMRotate>

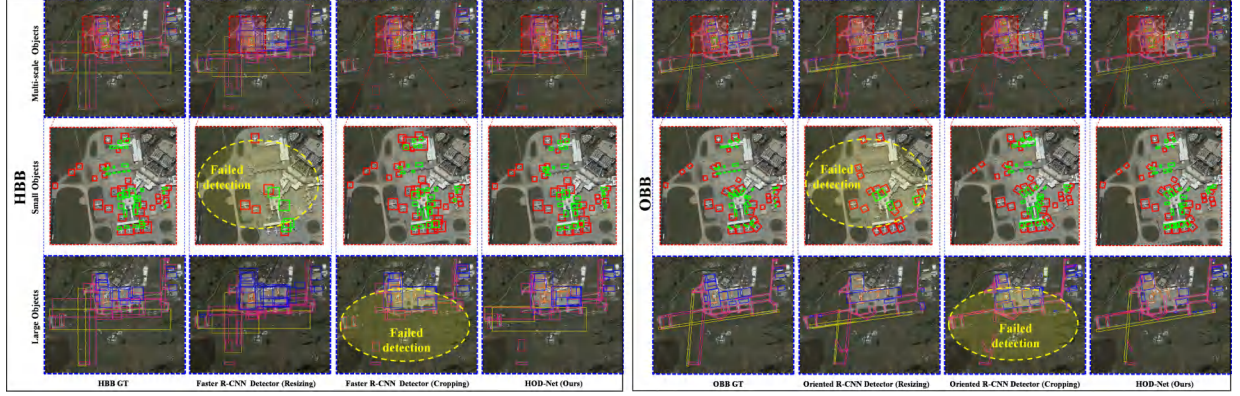


Fig. 8. The visualization results of HBB-based and OBB-based detectors on STAR test set.

TABLE IV
BASELINE RESULTS (%) OF SGG ON STAR TEST SET

OBD Type	Relationship Prediction	Venue	PredCls			SGCls			SGDet		
			MR@1500/2000	mMR@1500/2000	HMR@1500/2000	MR@1500/2000	mMR@1500/2000	HMR@1500/2000	MR@1500/2000	mMR@1500/2000	HMR@1500/2000
HBB	IMP [34]	CVPR'17	49.90/51.62	18.51/19.40	27.00/28.20	42.48/43.63	16.46/17.02	23.73/24.49	23.01/23.72	8.25/8.60	12.15/12.62
	Motif [43]	CVPR'18	61.81/63.78	31.33/32.66	41.58/43.20	45.13/46.19	21.46/22.57	29.09/30.32	28.36/29.32	11.49/11.95	16.35/16.98
	GPS-Net [50]	CVPR'20	65.68/67.10	34.92/36.11	45.60/46.95	31.99/33.16	14.61/15.40	20.06/21.03	30.83/31.58	12.71/13.10	18.00/18.52
	SHA [42]	CVPR'22	64.03/65.77	31.98/33.16	42.66/44.10	48.08/49.54	<u>21.60/22.46</u>	<u>29.81/30.91</u>	30.21/31.04	<u>13.24/13.65</u>	<u>18.41/18.96</u>
	HETSGG [46]	AAAI'23	63.83/65.26	28.70/29.72	39.60/40.84	30.03/31.08	10.50/11.05	15.56/16.30	25.19/25.82	8.17/8.50	12.34/12.79
	PE-Net [45]	CVPR'23	63.21/65.11	<u>35.94/37.10</u>	<u>45.82/47.27</u>	38.84/40.05	19.46/20.20	25.93/26.86	28.11/29.22	12.80/13.40	17.59/18.37
	RPCM (Ours)	-	<u>64.27/65.67</u>	38.73/39.70	48.33/49.48	<u>46.28/47.55</u>	25.32/26.12	32.73/33.72	<u>30.33/31.36</u>	14.09/14.78	19.24/20.09
OBB	IMP [34]	CVPR'17	51.52/53.19	21.47/22.31	30.31/31.43	46.08/47.50	18.57/19.38	26.47/27.53	18.13/19.03	5.60/5.99	8.56/9.11
	Motif [43]	CVPR'18	62.02/64.00	29.40/30.64	39.89/41.44	50.44/52.29	<u>24.21/25.36</u>	<u>32.72/34.16</u>	20.63/21.60	7.78/8.22	11.30/11.91
	GPS-Net [50]	CVPR'20	65.38/66.87	32.52/34.08	44.80/46.22	29.17/30.36	14.28/14.90	19.17/19.99	24.18/25.04	9.54/10.09	13.68/14.38
	SHA [42]	CVPR'22	<u>65.00/67.04</u>	33.26/34.78	44.00/45.80	<u>51.01/52.63</u>	23.30/24.15	31.99/33.11	28.01/29.06	<u>10.18/10.79</u>	<u>14.93/15.74</u>
	HETSGG [46]	AAAI'23	63.81/65.37	29.74/30.98	40.57/42.04	27.64/28.96	12.18/12.78	16.91/17.73	19.45/20.32	5.44/5.73	8.50/8.94
	PE-Net [45]	CVPR'23	62.87/64.98	<u>36.99/38.29</u>	<u>46.58/48.19</u>	41.79/43.45	20.64/21.60	27.63/28.86	21.26/22.50	8.75/9.30	12.40/13.16
	RPCM (Ours)	-	64.23/65.86	41.24/42.30	50.23/51.51	51.29/52.72	30.04/30.85	37.89/38.92	<u>27.23/28.50</u>	11.53/12.07	16.20/16.96

Bold indicates the best result, and underline indicates second-best result.

R-CNN[†] (HOD-Net's baseline) has 267.08M parameters and 929.13 GFLOPS, and HOD-Net[†] has 534.16M parameters and 929.13 GFLOPS. Importantly, for objects with extreme aspect ratios (e.g., runways, shiplocks), the proposed HOD-Net outperforms other methods in OBB detection by a great margin of (34.9%/31.2%).

Qualitative Results and Visualization: In Fig. 8, we further give the visualizations of different methods on OBB-based detectors. It can be seen that many large objects (e.g., docks, aprons) cannot be detected holistically due to the integrity being destroyed in the detection based on cropping strategy, and a lot of small objects (e.g., airplanes, boats, cranes, boardingbridges) cannot be detected correctly due to the loss of information in the detection based on resizing strategy. The proposed HOD-Net shows superiority in the detection results of multi-scale objects and achieves almost the same results as the annotation for objects with extreme aspect ratios (e.g., runways). The above results show that HOD-Net can provide mandatory support for the SGG task in large-size VHR SAI.

D. Results and Analysis for SGG

Comparison With the State-of-the-Art Methods: As shown in Section V-C, HOD-Net[†] (with Swin-L [67] adopted as the

backbone) exhibits superior performance for the OBD task, which is a prerequisite for achieving accurate relationship prediction in the SGG task. For a fair comparison, all experiments utilized the same toolkit³ and employed HOD-Net[†] as the detector for the SGG task. We use MR@1500/2000, mMR@1500/2000, and HMR@1500/2000 as evaluation metrics of the SGG task. In Table IV, we conduct a comprehensive evaluation of 14 experiments and validated the effectiveness of RPCM by comparing it with six state-of-the-art SGG methods on both HBB-based and OBB-based detectors. For the SGCIs task based on the OBB-based detector, our RPCM achieves 37.89%/38.92% on HMR@1500/2000, outperforming the SOTA method by a large margin of 5.17%/4.76%. For the SGDet task, our RPCM also outperforms the previous method with a 1.27%/1.82% improvement on HMR@1500/2000.

Qualitative Results and Visualization: The SGG visualization results of different models on the STAR dataset are shown in Fig. 9. HETSGG has a limited ability to learn discriminative details (e.g., HETSGG fails to discriminate specific relationships $\langle \text{ship}, \text{away from/approach}, \text{dock} \rangle$ by the wake trailing of the ship). PE-Net does not have the ability to constrain relationship prediction by relying on contextual reasoning (e.g.,

³<https://github.com/ZhuZi24/SGG-Toolkit>

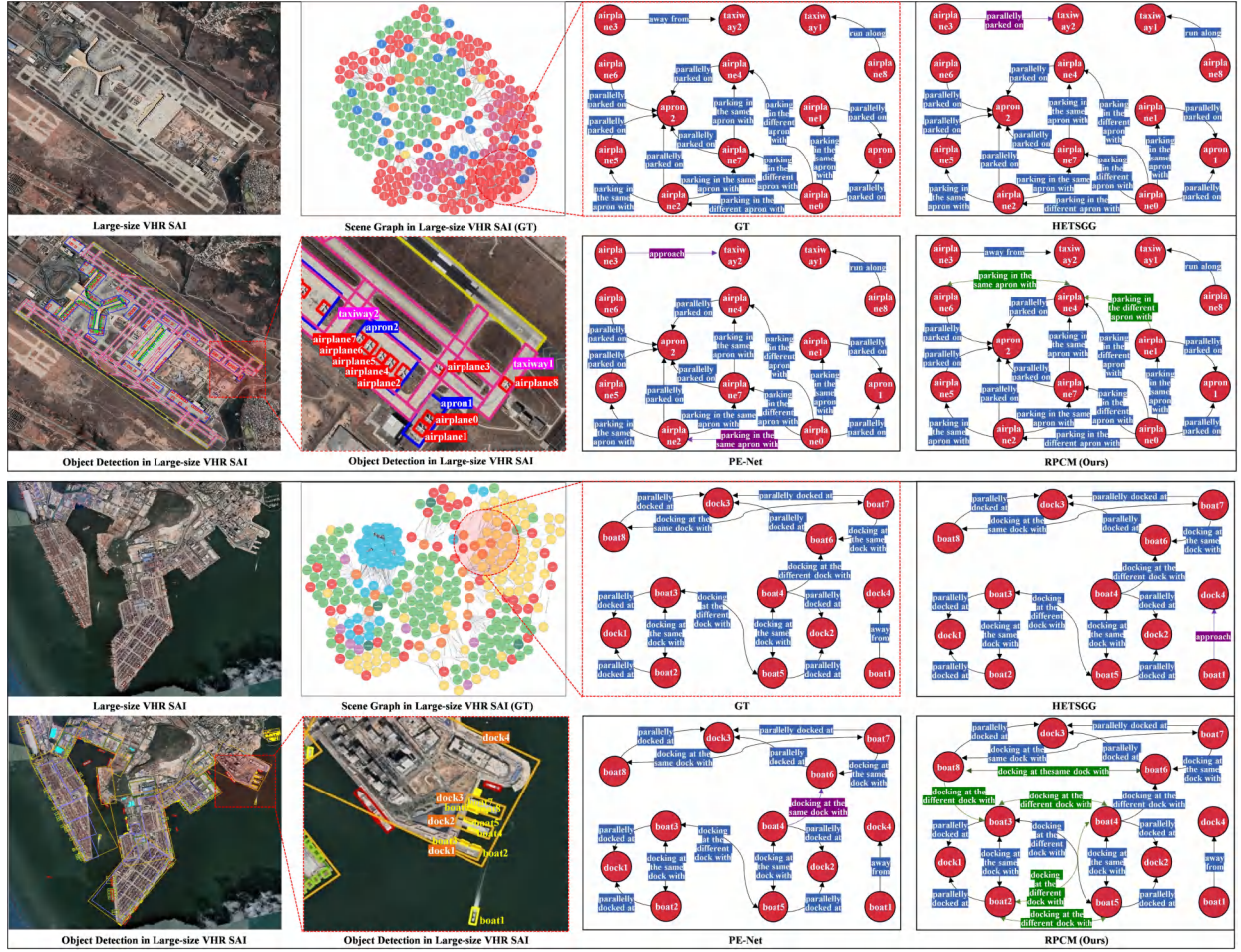


Fig. 9. The visualization results of the SGG task. Blue edges represent the correctly predicted relationships, purple edges indicate the ground-truth (GT) relationships that failed to be detected, and green edges are reasonable relationships predicted by the model but not annotated in the GT.

TABLE V
COMPARISON OF OUR PPG AND OTHER PAIR PRUNING METHODS

Models	PredCls			SGCIs			SGDet		
	MR@1500/2000	mMR@1500/2000	HMR@1500/2000	MR@1500/2000	mMR@1500/2000	HMR@1500/2000	MR@1500/2000	mMR@1500/2000	HMR@1500/2000
Randomization+RPCM	51.49/52.62	31.96/32.46	39.44/40.15	43.12/43.54	26.02/26.28	32.46/32.78	16.31/16.57	6.66/6.76	9.46/9.60
ABS-PRD [39]+RPCM	<u>53.32/53.92</u>	<u>33.45/33.75</u>	<u>41.11/41.51</u>	<u>44.04/44.58</u>	<u>26.16/26.41</u>	<u>32.82/33.17</u>	<u>17.02/17.49</u>	<u>7.02/7.68</u>	<u>9.94/10.67</u>
PPG+RPCM (Ours)	64.23/65.86	41.24/42.30	50.23/51.51	51.29/52.72	30.04/30.85	37.89/38.92	27.23/28.50	11.53/12.07	16.20/16.96

Bold indicates the best result, and underline indicates second-best result.

inferring the unreasonable triplet $\langle \text{airplane0}, \text{parking in the same apron with}, \text{airplane2} \rangle$ via triplet $\langle \text{airplane2}, \text{parallelly parked on}, \text{apron2} \rangle$ and triplet $\langle \text{airplane0}, \text{parallelly parked on}, \text{apron1} \rangle$. In summary, our RPCM network can effectively guide the model in learning the discriminative details of different relationships.

E. Ablation Study on SGG

1) *Effect of PPG Network*: Table V gives the quantitative comparison of the proposed PPG network with other pair pruning methods on the STAR dataset when the sorted top 10,000 pairs are selected. The ABS-PRD method based on statistics shows only a slight improvement over the randomized method across all metrics, indicating that the complex semantic

and spatial combinations of object pairs for SGG in large-size VHR SAI make it difficult to identify the combination patterns of objects through statistics. The score matrix, as depicted in the first row of Fig. 10, becomes sparse after the PPG network learning, effectively retaining the annotated pairs. This highlights the robust learning capacity of the proposed PPG network when dealing with pairs containing rich knowledge. The local visualization shows that the PPG network can retain the annotated pairs with high scores, as well as some pairs that are non-annotated but may contain high-value relationships, this is because the annotation of triplets in the real annotation is hardly exhaustive, so achieving the same sparsity as ground-truth is not necessary.

2) *Effect of PBA Iterations*: In Table VI, we compare the accuracy and time complexity of PredCls under different numbers

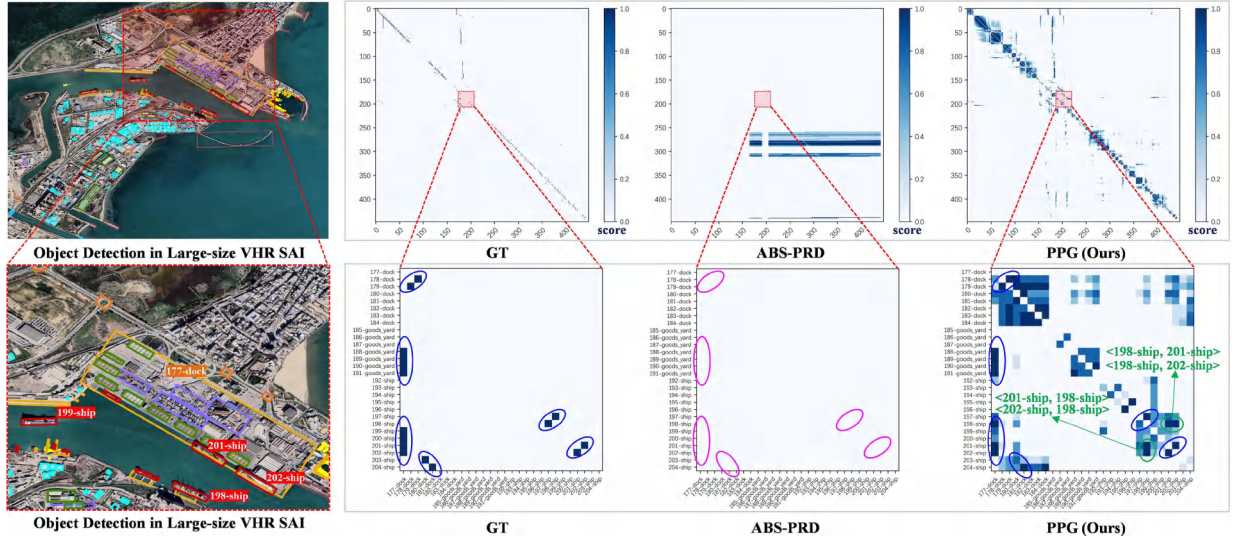


Fig. 10. The visualization results of pair score matrix. Blue edges represent the ground-truth (GT) pairs correctly predicted, purple edges indicate the GT pairs that failed to be detected, and green edges are reasonable pairs predicted by the model but not annotated in the GT.

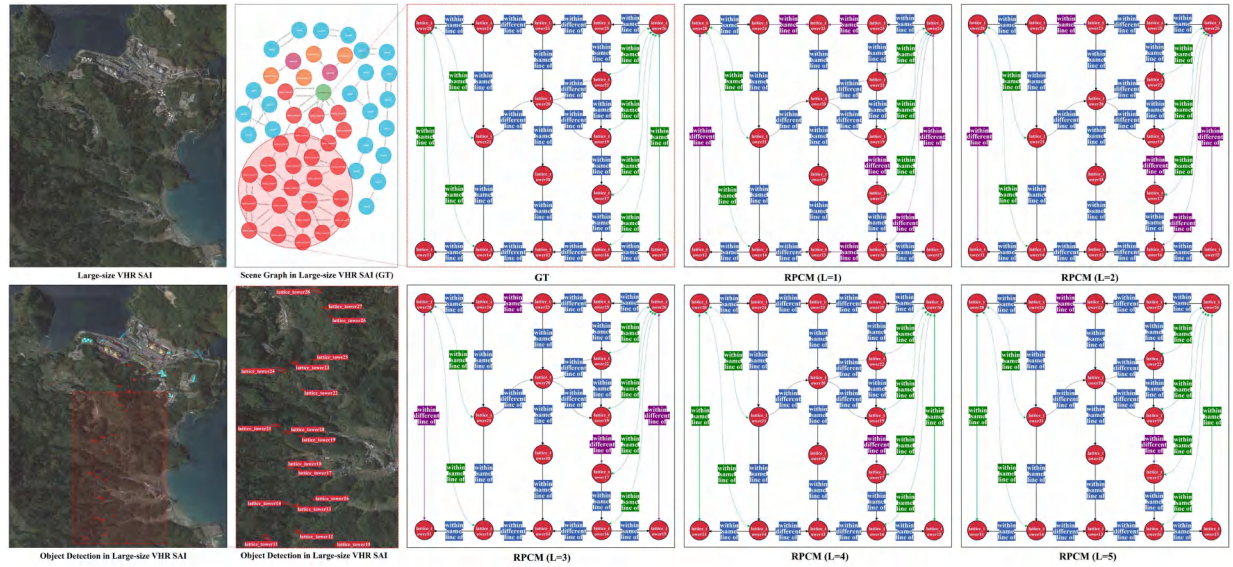


Fig. 11. Visualization of different PBA iters L. Blue edges represent the correctly predicted relationships, purple edges indicate incorrectly predicted relationships, and green edges are reasonable relationships predicted by the model but not annotated in the ground-truth (GT).

TABLE VI
COMPARISON WITH DIFFERENT PBA ITES L. TIME DENOTES THE INFERENCE TIME IN SECONDS PER IMAGE PER DEVICE, MEASURED ON 1 DEVICE

Iters	Time	PredCls		
		MR@1500/2000	mMR@1500/2000	HMR@1500/2000
L=1	1.51	63.56/65.44	39.12/40.54	48.43/50.06
L=2	1.59	64.14/ 65.79	40.10/41.18	49.35/50.65
L=3	1.67	63.90/65.54	<u>40.79/ 41.88</u>	<u>49.79/ 51.10</u>
L=4	1.74	64.23/65.86	41.24/42.30	50.23/51.51
L=5	1.82	62.14/64.09	39.92/41.09	48.61/50.08

Bold indicates the best result, and underline indicates second-best result.

of PBA iters L (set to 1, 2, 3, 4 and 5). It can be observed that iters L has a minor effect on the time complexity of our model.

Notably, the HMR@K keeps improving with the increase in iters when iters are less than 4. This suggests that proper contextual information can effectively improve the model performance, but excessive introduction of contextual information may cause disturbance. In this paper, iters L is set to 4.

To explore the intrinsic impact of context introduction on the model, Fig. 11 gives examples of visualizations obtained from different PBA iters. It can be seen that latticetowers that are far apart (e.g., "latticetower11 and latticetower28", "latticetower15 and latticetower26", and "latticetower16 and latticetower26") have incorrect relationship predictions due to insufficient contextual information when L is set to 1. As L increases, the introductions of more contextual information cause the wrong relationships to be gradually corrected. When L is set to 4,

TABLE VII

ABLATION STUDY ON EACH COMPONENT OF THE RPCM. OCA AND RCA DENOTE OBJECT CONTEXT AUGMENTATION AND RELATIONSHIP CONTEXT AUGMENTATION, RESPECTIVELY. CLA (CLASSIFIER) INDICATES THE METHOD (PROTOTYPE MATCHING (PM)/ LINEAR CLASSIFICATION) FOR RELATIONSHIP PREDICTION

Component			PredCls		
OCA	RCA	Cla	MR@1500/2000	mMR@1500/2000	HMR@1500/2000
×	×	PM	63.10/65.28	36.53/38.05	46.27/48.08
✓	×	PM	63.11/65.09	<u>37.86/39.07</u>	<u>47.87/48.83</u>
✓	✓	Linear	66.55/68.18	34.58/35.80	45.51/46.95
✓	✓	PM	<u>64.23/65.86</u>	41.24/42.30	50.23/51.51

Bold indicates the best result, and underline indicates second-best result.

the relationships between latticetowers at different distances are predicted with better results. It is worth noting that the previous incorrect relationship prediction occurred when L is 5, which may be caused by excessive contextual interference. Overall, the introductions of proper contexts help relationship inference and excessive context brings disruption, consistent with the quantitative metrics in Table VI.

3) *Effect of Different Components*: We perform component ablation by gradually removing them from the RPCM on the PredCls task, as shown in Table VII. Not surprisingly, the experimental results show that all three components, object context augmentation, relationship context augmentation, and prototype matching, are important for the performance of SGG. Specifically, prototype matching enables it to capture the discriminative features of different relationship categories in the semantic space, which in turn effectively distinguishes different categories with strong similarities. The object augmentation and relationship augmentation modules assist the model in learning contextual reasoning constraints for relationship prediction, as shown by the visualization in Fig. 9.

VI. CONCLUSION

We have proposed a first-ever dataset named STAR for SGG in large-size VHR SAI. It covers 1,273 complex scenarios worldwide, with image sizes ranging from 512×768 to $27,860 \times 31,096$ pixels, including more than 210,000 objects (with both OBB and HBB annotations) and more than 400,000 triplet annotations. The large image size, the extensive sample volume, and the diversity of object scale and relationship semantics make STAR a valuable dataset, providing indispensable data support for advancing a new challenging and meaningful task: SGG in large-size VHR SAI. Furthermore, we propose the CAC framework, for the SGG task in large-size VHR SAI, to achieve a cascading understanding of SAI from three levels: object detection, pair pruning and relationship prediction. In future work, we will continue to enrich the STAR dataset and improve the SGG models in SAI, maximizing the potential of SGG in various downstream tasks, such as image retrieval, image captioning and visual question answering. Furthermore, considering the growing availability of satellite video and the importance of integrating spatio-temporal information for accurately modeling dynamic interaction scenes, we plan to explore spatio-temporal SGG in satellite video, thereby opening up new avenues for

real-time monitoring and analysis in applications such as traffic planning, disaster response and maritime situational awareness.

REFERENCES

- [1] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, Jan. 2023.
- [2] Y. Cong, M. Y. Yang, and B. Rosenhahn, "RelTR: Relation transformer for scene graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11 169–11 183, Sep. 2023.
- [3] S. Sun, S. Zhi, Q. Liao, J. Heikkilä, and L. Liu, "Unbiased scene graph generation via two-stage causal modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12 562–12 580, Oct. 2023.
- [4] G.-S. Xia et al., "Dota: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
- [5] X. Sun et al., "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.
- [6] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2023.
- [7] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," in *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
- [8] K. Nogueira, O. A. Penatti, and J. A. Dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognit.*, vol. 61, pp. 539–556, 2017.
- [9] H. Zhu et al., "A spatial-channel progressive fusion RESNet for remote sensing classification," *Inf. Fusion*, vol. 70, pp. 72–87, 2021.
- [10] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (OBIA): A review of algorithms and challenges from remote sensing perspective," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 115–134, 2019.
- [11] M. Wurm, T. Stark, X. X. Zhu, M. Weigand, and H. Taubenböck, "Semantic segmentation of slums in satellite images using transfer learning on fully convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 150, pp. 59–69, 2019.
- [12] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 175, pp. 20–33, 2021.
- [13] J. Johnson et al., "Image retrieval using scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3668–3678.
- [14] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 685–10 694.
- [15] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10313–10322.
- [16] P. Reinartz, M. Lachaise, E. Schmeer, T. Krauss, and H. Runge, "Traffic monitoring with serial images from airborne cameras," *ISPRS J. Photogrammetry Remote Sens.*, vol. 61, no. 3–4, pp. 149–158, 2006.
- [17] M. Smigaj et al., "Monitoring riverine traffic from space: The untapped potential of remote sensing for measuring human footprint on inland waterways," *Sci. Total Environ.*, vol. 860, 2023, Art. no. 160363.
- [18] L. Matikainen et al., "Remote sensing methods for power line corridor surveys," *ISPRS J. Photogrammetry Remote Sens.*, vol. 119, pp. 10–31, 2016.
- [19] J. Chen, X. Zhou, Y. Zhang, G. Sun, M. Deng, and H. Li, "Message-passing-driven triplet representation for geo-object relational inference in HRSI," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8003205.
- [20] P. Li, D. Zhang, A. Wulamu, X. Liu, and P. Chen, "Semantic relation model and dataset for remote sensing scene understanding," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 7, 2021, Art. no. 488.
- [21] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [22] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.

- [23] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.
- [24] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
- [25] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411.
- [26] Y. Li, J. Luo, Y. Zhang, Y. Tan, J.-G. Yu, and S. Bai, "Learning to holistically detect bridges from large-size VHR remote sensing imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Dec. 2024.
- [27] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proc. 4th Workshop Vis. Lang.*, 2015, pp. 70–80.
- [28] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Dense relational captioning: Triple-stream networks for relationship-based captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6271–6280.
- [29] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1745–1752.
- [30] B. Thomee et al., "The new data and new challenges in multimedia research," 2015, *arXiv:1503.01817*.
- [31] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 740–755.
- [32] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 852–869.
- [33] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, 2017.
- [34] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5410–5419.
- [35] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, and T. Mei, "VRR-VG: Refocusing visually-relevant relationships," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10 403–10 412.
- [36] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6700–6709.
- [37] A. Kuznetsova et al., "The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, 2020.
- [38] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707411.
- [39] Y. Li, K. Li, Y. Zhang, L. Wang, and D. Zhang, "Aug: A new dataset and an efficient model for aerial image urban scene graph generation," 2024, *arXiv:2404.07788*.
- [40] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.
- [41] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3716–3725.
- [42] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19 427–19 436.
- [43] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5831–5840.
- [44] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6619–6628.
- [45] C. Zheng, X. Lyu, L. Gao, B. Dai, and J. Song, "Prototype-based embedding network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22 783–22 792.
- [46] K. Yoon, K. Kim, J. Moon, and C. Park, "Unbiased heterogeneous scene graph generation with relation-aware message passing neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 3285–3294.
- [47] C. Chen, Y. Zhan, B. Yu, L. Liu, Y. Luo, and B. Du, "Resistance training using prior bias: Toward unbiased scene graph generation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 212–220.
- [48] Y. Deng et al., "Hierarchical memory learning for fine-grained scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 266–283.
- [49] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 670–685.
- [50] X. Lin, C. Ding, J. Zeng, and D. Tao, "GPS-Net: Graph property sensing network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3746–3753.
- [51] R. Li, S. Zhang, B. Wan, and X. He, "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11 109–11 119.
- [52] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 6, pp. 3623–3634, Jun. 2017.
- [53] X. Yang et al., "Automatic ship detection in remote sensing images from Google Earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, 2018, Art. no. 132.
- [54] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.
- [55] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8232–8241.
- [56] L. Hou, K. Lu, X. Yang, Y. Li, and J. Xue, "G-Rep: Gaussian representation for arbitrary-oriented object detection," *Remote Sens.*, vol. 15, no. 3, 2023, Art. no. 757.
- [57] X. Yang, G. Zhang, W. Li, X. Wang, Y. Zhou, and J. Yan, "H2rbox: Horizontal box annotation is all you need for oriented object detection," in *Proc. Int. Conf. Learn. Representations*, 2023.
- [58] Y. Yu et al., "H2rbox-v2: Incorporating symmetry for boosting horizontal box supervised oriented object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 59137–59150.
- [59] J. Luo, X. Yang, Y. Yu, Q. Li, J. Yan, and Y. Li, "PointOBB: Learning oriented object detection via single point supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16730–16740.
- [60] Y. Yu et al., "Point2RBox: Combine knowledge from synthetic visual patterns for end-to-end oriented object detection with single point supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16783–16793.
- [61] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 966–970.
- [62] X. Chen et al., "Coupled global-local object detection for large VHR aerial images," *Knowl.-Based Syst.*, vol. 260, 2023, Art. no. 110097.
- [63] R. Girshick, "Fast R-CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [64] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [65] P. Velickovic et al., "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [66] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
- [67] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 012–10 022.
- [68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [69] 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.
- [70] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
- [71] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.
- [72] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.

- [73] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *Proc. 2021 IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3490–3499.
- [74] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [75] Y. Zeng, Y. Chen, X. Yang, Q. Li, and J. Yan, "ARS-DETR: Aspect ratio-sensitive detection transformer for aerial oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5610315.
- [76] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.
- [77] X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18 381–18 394.
- [78] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 11 830–11 841.
- [79] X. Yang et al., "The KFIoU loss for rotated object detection," in *Proc. Int. Conf. Learn. Representations*, 2023.
- [80] C. Xu et al., "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7318–7328.
- [81] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3163–3171.
- [82] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.
- [83] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.
- [84] Z. Guo, C. Liu, X. Zhang, J. Jiao, X. Ji, and Q. Ye, "Beyond bounding-box: Convex-hull feature adaptation for oriented and densely packed object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8792–8801.
- [85] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1829–1838.
- [86] L. Hou, K. Lu, J. Xue, and Y. Li, "Shape-adaptive selection and measurement for oriented object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 923–932.
- [87] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
- [88] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 677–694.
- [89] Y. Yu and F. Da, "On boundary discontinuity in angle regression based arbitrary oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6494–6508, Oct. 2024.
- [90] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.
- [91] J. Han, J. Ding, N. Xue, and G.-S. Xia, "ReDet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2786–2795.
- [92] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.
- [93] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 16 794–16 805.
- [94] X. Cai, Q. Lai, Y. Wang, W. Wang, Z. Sun, and Y. Yao, "Poly kernel inception network for remote sensing detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27 706–27 716.
- [95] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [96] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.
- [97] K. Chen et al., "Mmdetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv: 1906.07155*.
- [98] Y. Zhou et al., "MMRotate: A rotated object detection benchmark using pytorch," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 7331–7334.



Yansheng Li (Senior Member, IEEE) received the BS degree in information and computing science from Shandong University, Weihai, China, in 2010, and the PhD degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015. He is currently a full professor and vice dean with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. He has authored more than 100 peer-reviewed papers, such as *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *CVPR*, *ECCV*, and *AAAI*. His research interests include knowledge graph, deep learning and their applications in remote sensing Big Data mining. He is an associate editor of *IEEE Transactions on Geoscience and Remote Sensing* and a junior editorial member of *The Innovation*.



Linlin Wang received the BS degree in geography and information science from Henan University, Kaifeng, China in 2019 and the ME degree in cartography and geographical information engineering from the China University of Geosciences, Wuhan, China in 2022. She is currently working toward the PhD degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. Her main research interests include remote sensing scene graph generation.



Tingzhu Wang received the BE degree from the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, China in 2023. He is currently working toward the MS degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. He has published one paper in *ECCV*. His research interests include image scene graph generation and video scene graph generation.



projects for oriented object detection.

Xue Yang received the BE degree in automation from Central South University, Hunan, China, in 2016, the MS degree from the Chinese Academy of Sciences University, Beijing, China, in 2019 and the PhD degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2023. His research interest is computer vision. He published first-authored papers in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IJCV*, *CVPR*, *ECCV*, *ICCV*, *NeurIPS*, and *ICML*. He is also the leading contributor to the MMRotate and AlphaRotate open-source



Junwei Luo received the BS degree in remote sensing science and technology from Wuhan University, Wuhan, China in 2023. He is currently working toward the MS degree with the School of Remote Sensing and Information Engineering, Wuhan University. He has published a first-authored paper in *CVPR*. His research interests include remote sensing object detection and remote sensing vision-language foundation model.



Qi Wang (Senior Member, IEEE) received the BE degree in automation and the PhD degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition, and remote sensing.



Bo Dang received the BS degree in remote sensing science and technology from Wuhan University, Wuhan, China in 2022. He is currently working toward the PhD degree with the School of Remote Sensing and Information Engineering, Wuhan University. He has published several papers in CVPR, AAAI, *ISPRS Journal of Photogrammetry and Remote Sensing*, etc. His research interests include remote sensing semantic segmentation and remote sensing foundation model.



Youming Deng received the BS degree in remote sensing science and technology from Wuhan University, Wuhan, China in 2022. He is currently working toward the PhD degree with the Department of Computer Science, Cornell University, New York, United States. He has published a first-authored paper in ECCV. His research interests include scene graph generation, computer vision, and 3D vision.



Yongjun Zhang (Member, IEEE) received the BS degree in geodesy, the MS degree in geodesy and surveying engineering, and the PhD degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively. He is currently a full professor and dean with the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. His research interests include aerospace and low-altitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence-driven remote sensing image interpretation, and 3-D city reconstruction.



Wenbin Wang received the BS degree in computer science and technology from Nankai University, Tianjin, China, in 2017, and the PhD degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, in 2022. He is now an assistant professor with the College of Computer and Information Technology, China Three Gorges University, China. His current research interests mainly include computer vision, scene understanding, and scene graph generation.



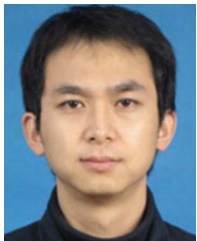
Yi Yu (Member, IEEE) received the BE and PhD degrees from the School of Automation, Southeast University, in 2016 and 2022. He is currently an assistant researcher with Southeast University, Nanjing, China. His research interest is computer vision. His first-authored papers have been published in high-quality journals and premiere scientific conferences including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Information Forensics and Security*, CVPR, NeurIPS, and etc.



Xian Sun (Senior Member, IEEE) received the BS degree from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004, and the MS and PhD degrees from the Institute of Electronics, Chinese Academy of Sciences, China, in 2009. He is currently a professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.



Junchi Yan (Senior Member, IEEE) received the PhD degree in electrical engineering, Shanghai Jiao Tong University in 2015 when he was with IBM. He is a professor and associate director with the School of Artificial Intelligence and Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. Before that, he was a senior research staff member with IBM Research where he started his career since April 2011. His research interest includes machine learning and computer vision. He served area chair for conferences including ICLR, ICML, NeurIPS, SIGKDD, CVPR, AAAI, IJCAI, etc., and associate editor for journals including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *ACM Transactions on Probabilistic Machine Learning* and *Pattern Recognition*. He is a fellow of IET and IAPR. He is the recipient of CVPR 2024 best paper candidate award.



Haifeng Li (Member, IEEE) received the PhD degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2009, and the MS degree in transportation engineering from the South China University of Technology, Guangzhou, China, in 2005. He is currently a professor with the School of Geosciences and Info-Physics, Central South University, Changsha, China. His research interests include geo/remote sensing Big Data and artificial intelligence.