remote sensing

MDPI

Article

A Novel Hybrid Attention-Driven Multistream Hierarchical Graph Embedding Network for Remote Sensing Object Detection

Shu Tian [1], Lin Cao [1,2,*], Lihong Kang [3], Xiangwei Xing [3], Jing Tian [3], Kangning Du [1], Ke Sun [4], Chunzhuo Fan [3], Yuzhe Fu [3] and Ye Zhang [5]

[1] Key Laboratory of Information and Communication Systems, Ministry of Information Industry, Beijing Information Science and Technology University, Beijing 100101, China
[2] Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100101, China
[3] Beijing Remote Sensing Information Institute, Beijing 100094, China
[4] Software College, Shenyang Normal University, Shenyang 110034, China
[5] School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China
* Correspondence: charlin@bistu.edu.cn; Tel.: +86-138-1045-0940

Abstract: Multiclass geospatial object detection in high-spatial-resolution remote-sensing images (HSRIs) has recently attracted considerable attention in many remote-sensing applications as a fundamental task. However, the complexity and uncertainty of spatial distribution among multiclass geospatial objects are still huge challenges for object detection in HSRIs. Most current remote-sensing object-detection approaches fall back on deep convolutional neural networks (CNNs). Nevertheless, most existing methods only focus on mining visual characteristics and lose sight of spatial or semantic relation discriminations, eventually degrading object-detection performance in HSRIs. To tackle these challenges, we propose a novel hybrid attention-driven multistream hierarchical graph embedding network (HA-MHGEN) to explore complementary spatial and semantic patterns for improving remote-sensing object-detection performance. Specifically, we first constructed hierarchical spatial graphs for multiscale spatial relation representation. Then, semantic graphs were also constructed by integrating them with the word embedding of object category labels on graph nodes. Afterwards, we developed a self-attention-aware multiscale graph convolutional network (GCN) to derive stronger for intra- and interobject hierarchical spatial relations and contextual semantic relations, respectively. These two relation networks were followed by a novel cross-attention-driven spatial- and semantic-feature fusion module that utilizes a multihead attention mechanism to learn associations between diverse spatial and semantic correlations, and guide them to endowing a more powerful discrimination ability. With the collaborative learning of the three relation networks, the proposed HA-MHGEN enables grasping explicit and implicit relations from spatial and semantic patterns, and boosts multiclass object-detection performance in HRSIs. Comprehensive and extensive experimental evaluation results on three benchmarks, namely, DOTA, DIOR, and NWPU VHR-10, demonstrate the effectiveness and superiority of our proposed method compared with that of other advanced remote-sensing object-detection methods.

Keywords: graph convolutional networks (GCNs); object detection; remote-sensing images; graph learning; attention mechanism

Citation: Shu, T.; Lin, C.; Kang, L.; Xing, X.; Tian, J.; Du, K.; Su, K.; Fan, C.; Fu, Y.; Zhang, Y. A Novel Hybrid Attention-Driven Multistream Hierarchical Graph Embedding Network for Remote Sensing Object Detection. *Remote Sens.* **2022**, *14*, 4951. https://doi.org/10.3390/rs14194951

Academic Editor: Leyuan Fang, Chunhui Zhao, Jinchang Ren, Weiwei Sun, Shou Feng, Nan Su and Yiming Yan

Received: 2 September 2022
Accepted: 29 September 2022
Published: 4 October 2022

1. Introduction

With the rapid development of remote-sensing sensors and space technology, many high-spatial-resolution remote-sensing images (HSRIs) are available for satisfying specific needs. A significant enhancement in the quality and quantity of remote-sensing imagery has further broadened their wide range of applications, such as image segmentation [1],

image classification [2], data fusion [3], and object detection [4]. As a fundamental and crucial task for most remote-sensing-imagery-based applications, object detection has drawn increasing attention, especially for practical applications based on optical HSRIs. The main purpose of object detection is to assign each corresponding object to a category and regress the correct location for each predicted object. Over the past few years, a large number of object detection approaches have been explored, demonstrating their reasonable performance [5,6]. Existing object detection approaches in optical remote-sensing imagery can be divided into traditional and deep-learning-based object-detection approaches. Traditional object detection methods usually begin with the generation of regions of interest (ROIs) accompanied by implementations of imagery-pixel-based clustering and object detection using traditional handcrafted image features.

With the rise of deep convolutional neural networks (CNNs), a series of deep-learning-based approaches have emerged to address the problems of object detection in HSRIs that also achieved remarkable achievements. These existing deep-learning-based object detection methods can mainly be categorized into one- and two-stage object-detection methods. Two-stage object-detection approaches usually start with decomposing the input imagery into a set of object region proposals and then iteratively and optimally selecting the most contributing region proposals as the preudoinstance-level labels. Lastly, the selected object region proposals with the most contributions are fed into training the detectors for object category prediction and location estimation.

Among two-stage detection approaches, faster R-CNN [7] can achieve outstanding performance in HSRI object detection, taking advantage of the strength of CNNs to learn powerful features with high discriminative ability from remote-sensing imagery. The major innovation of faster R-CNN is introducing region proposal networks (RPNs) to extract ROIs, which ensures the highly efficient operation of a uniform deep-learning-based object detection framework. Although the above approach has achieved impressive detection results, its execution is always slow for some practical HSRI-based engineering applications. Different from two-stage approaches, the implementation of a one-stage approach can be regarded as a single regression process to analyze the relations of the pixel values of the input imagery with the set of neighborhoods and their variations, including object locations and the corresponding category information. A one-stage object detection method generally treats detection tasks as end-to-end regression problems of object location and category information, and realizes bounding-box regression and category prediction by learning the network models. This kind of detection approaches can achieve higher detection speed, but they always have lower detection accuracy than that of two-stage object detection methods.

So far, most object detection explorations in remote-sensing images have been inspired by object detection in natural images [8,9]. Compared with natural images, HSRIs generally have higher appearance ambiguity among the same or different categories of geospatial objects, and more complexity and diversity on background information and spatial distribution, as shown in Figure 1. In addition, HSRIs have multiscale implicitly structured characteristics that consist of pixel clustering within the corresponding neighborhoods into the object parts; then these series of pixel clusters are further grouped into the objects and spatial distributions of different categories of geospatial objects in HSRIs. Therefore, object detection is still a challenging task for several applications in the field of remote-sensing imagery.

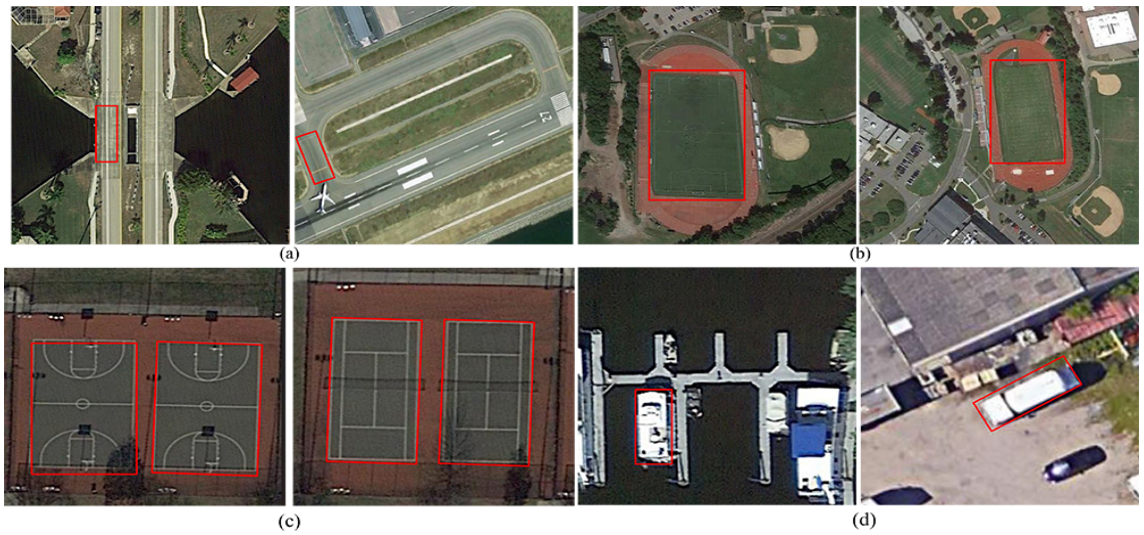

Figure 1. Appearance ambiguities among multiclass remote-sensing objects. Ambiguity between the object classes of (**a**) bridges and roads, (**b**) ground track fields and soccer fields, (**c**) basketball and tennis courts, and (**d**) ships and large vehicles.

To tackle these challenges, in this paper, we present a novel hybrid attention-driven multistream hierarchical graph embedding network (HA-MHGEN) that investigates by utilizing hierarchical spatial relations with contextual semantic relations to improve remote-sensing object-detection performance. Specifically, we first utilized the powerful generalization capacity of a region proposal network (RPN) to generate a set of region proposals; this operation was equipped with a similar region feature encoding network to that of the advanced faster R-CNN framework. Then, the hierarchical spatial graph construction module is presented to model inherent inter- and intra-object spatial relations across multiscale feature maps. A semantic graph construction module is also employed to characterize implicit and diverse semantic relations among object category labels in which every graph node could be regarded as an object category. The connectivity of each pair of semantic graph nodes relies on the discrimination of correlations between object category labels. A self-attention-driven hierarchical graph convolutional network was designed for capturing rich inter- and intra-object contextual spatial relations among across multiscale feature maps that are achieved from the input images via a pretrained backbone model. Furthermore, a semantic GCN operation is exploited to derive implicit global semantic relations among the category labels. Subsequently, a novel cross-attention-driven semantic and spatial fusion module is investigated to explore the complementarity between the learned spatial and semantic features to utilize cross-attention to guide them between the spatial and semantic modalities; it also enhances the discrimination ability of the fused features to further improve object-detection performance. The main contributions of our work are summarized as follows.

(1) We developed a hierarchical graph construction module to effectively model multiscale spatial relation graphs that were implemented on the pre-extracted region proposals of multiscale feature maps. Moreover, a category semantic graph construction model is presented to characterize semantic relations among object category labels. These two kinds of graph representations are then fed into the proposed HA-MHGEN, which further enhances their representative and discriminative capacity.

(2) We propose a self-attention-driven hierarchical graph embedding network (SA-HGEN) that enables capturing comprehensive intra- and interobject spatial and semantic relations. The introduction of self-attention can promote SA-HGEN to adaptively focus on more effective and important inter- and intra-object correlations across spatial and semantic features, respectively. This series of derived spatial and semantic relation features are

guided by the self-attention mechanism, providing higher discrimination for distinguishing subtle differences between location-variable and complicated remote-sensing objects.

(3) To explore the complementarity between learned spatial and semantic features to further enhance their discriminative ability for object detection, a novel cross-attention-driven fusion module was designed for spatial and semantic graph feature fusion. This mainly adopts the multihead attention mechanism to guide one graph modality to another, which enables learning the optimal associations between semantic and spatial graph features, and endows the fused features with more powerful discriminative ability, improving object-detection performance.

Extensive experimental results on three benchmark datasets demonstrate the effectiveness and superiority of our proposed HA-MHGEN framework. The remainder of this paper is organized as follows. In Section 2, we briefly review several related works regarding multiclass geospatial object detection in HRSIs, and introduce the attention mechanism and graph convolutional networks (GCNs). Then, we present the details of HA-MHGEN in Section 3. In Section 4, we compare HA-MHGEN with several state-of-the-art approaches on three public benchmarks to indicate the effectiveness of the proposed method. In Section 5, we provide a brief conclusion.

2. Related Work

As a fundamental and crucial part of HSRI analysis and processing, object detection has gained increasing attention, and many object-detection approaches have been explored in recent years. According to the manner of feature extraction and representation, these existing methods can be categorized into traditional handcrafted-based and deep-learning-based methods. In this section, we briefly review previous object-detection approaches in HSRIs that have had two stages of developments, namely, low-level handcrafted-feature-based and high-level deep-learning-based detection frameworks. In addition, we introduce several recent advanced-learning models, including the attention mechanism and graph convolutional networks (GCNs), and their achievements in the field of object detection in HSRIs.

2.1. Traditional Handcrafted-Feature-Based Object-Detection Approaches

Many previous works regarding the object-detection problem mainly relied on the low-level handcrafted features, such as SIFT, HOG, and BOW. For example, Li et al. [10] proposed scale-orientation SIFT-based restriction criteria to solve feature-matching problems between different remote-sensing images. Sirmacek et al. [11] combined SIFT features and graph theory for building detection in remote-sensing imagery. SIFT was first adapted for detecting possible building regions under different imaging conditions, and each detected keypoint of the building regions could be regarded as the vertex of a graph. Then, building detection results could be achieved by analyzing the connection relations of keypoint-based graphs. Tao et al. [12] represented an airport by using a series of scale-invariant feature transform (SIFT) keypoints, detecting it through an improved SIFT-based matching strategy; object-detection results were achieved by using prior knowledge to select the most possible candidate object regions. The HOG feature descriptor that enables to effectively extract edge or local shape information has also been widely investigated for object detection. For example, Dalal et al. [13] first proposed a HOG that represented the objects by using the distributions of gradient intensities and orientations in spatially distributed object regions. Xiao et al. [14] fused orientation normalization, feature space mapping, and an elliptic Fourier transformation to obtain the rotational invariance of HOG; then, derived combination-based HOG features were fed into the detector for remote-sensing object detection. Cheng et al. [15] presented a detection framework called the collection of part detectors (COPD) that adopted the feature pyramid strategy for extracting multilevel HOG features for detection tasks. The BoW model is another famous handcrafted-based feature descriptor for object detection. The BoW model regards every scene as the aggregation of a series of unsorted regions that contain parts of the information of each category [16].

The BoW descriptor and several of its variants have also been widely explored for remote-sensing object detection. For example, Meynberg et al. [17] utilized the BoW model with two alternative local features encoded as improved Fisher vectors for object detection in remote-sensing imagery. Sun et al. [18] utilized a sliding window for searching for object locations, and extracted each object feature representation by using the spare spatial coding BoW (SSCBoW). Then, the set of SSCBOW-based features were incorporated with a linear support vector machine (SVM) for object detection in remote-sensing imagery. Moreover, several other types of handcrafted-based texture features were proposed for detection tasks [19]. Although these existing handcrafted-based features have achieved impressive performance for several specific object-detection tasks, there still exist certain limitations for object detection in HSRIs, as they lack the capability to capture high-level semantic information that is required to distinguish multiple categories of objects, especially in situations when visual recognition tasks are more challenging.

2.2. Deep-Learning-Based Object-Detection Approaches

With the rapid development and remarkable achievements of deep-learning technology, deep-learning-based object-detection approaches are receiving increasing attention. In recent years, many advanced deep-learning-based object-detection approaches have been developed in the field of HSRIs, and the most famous deep learning technology is the convolutional neural network (CNN) model. Compared traditional handcrafted features that are dependent on abundant human ingenuity, features of CNN models are directly extracted from the pixel values of images by using a neural network. Furthermore, the deep architecture of a CNN enables extracting more powerful features and capturing more abstract characteristics, which significantly improves object detection in HSRIs [20,21]. Essentially, these state-of-the-art object-detection approaches are mainly divided into two categories, two- and one-stage detection approaches. The former detection approaches always divide their executions into two stages, the region proposal module and classification. The latter type of detection methods exploit the implementation of object detection as an end-to-end procedure that aims to simultaneously predict the bounding box, object confidence, and probability of the corresponding object category.

For object-detection algorithms based on the two-stage strategy, Han et al. [22] proposed a transfer-learning detection framework based on faster-RCNN for detecting multiple categories of geospatial objects in HSRIs that mainly utilized a pretrained mechanism to boost the efficiency of different traits by transfer learning from the natural-imagery domain to the HSRI domain. Cheng et al. [23] presented an effective object-detection framework that imposed a a rotation-invariant regularizer and a Fisher discrimination regularizer to train CNN-based models. By combining these two regularizers to optimally learn the whole network, learning features both had rotation-invariant characteristics and constrained the CNN-based features to have a small within-class scatter, but large between-class separation. Li et al. [24] developed a uniform CNN model-based detection method that was incorporated with a region proposal network (RPN) and a local-contextual feature network. Then, a dual-channel feature fusion subnetwork was proposed for tackling the problem of the appearance ambiguity of multiple categories of geospatial objects. Deng et al. [25] derived a novel CNN-based feature descriptor by adjusting the concatenated rectified linear unit (ReLU) and the inception subnetwork. Then, a multiscale object region proposal network was first adopted for possible object region generation from some intermediate layers, and an object-detection layer followed by using a set of aggregated learning features. For object-detection algorithms based on the one-stage strategy, Chen et al. [26] utilized the strategy of transfer learning, a single deep CNN-based model, and limited labeled training images for carrying out tasks of airplane detection in an end-to-end trainable way. Wang et al. [27] developed a novel one-stage detection framework called the full-scale object-detection network (FSoD-Net) whose architecture was composed of a multiscale enhancement network backbone and scale-invariant regression layers (SIRLs) in a cascade way. Zhang et al. [28] presented a semantic-context-aware network (SCANet) model that

could execute multiscale geospatial object detection by fusing learned multiscale feature representations and completely semantic context information. Kang et al. [29] developed a detector on the basis of widely used proposal-free detection framework YOLOv2 with a reweighted module that reassigned weights for the learning series of deep features with the use of corresponding support images. Wang et al. [30] presented a uniform detection method called the feature-merged single-shot detection (FMSSD) framework that exploited the context information aggregation of multiscale feature maps to extract complete features with high discriminative ability to improve object detection. Chen et al. [31] proposed a refined single-stage detector that adopted an enhanced feature extraction network by combining it with a bidirectional residual feature pyramid network and a multiscale feature fusion module in order to further boost remote-sensing object-detection performance. Although these detection approaches are widely applied in the HSRIs and can achieve impressive results, there are still several problems that are induced by shooting angles, objection distribution, and imaging range, and they also hardly satisfy the increasing demands of object-detection tasks in HSRIs.

2.3. Attention-Mechanism-Based Object-Detection Approaches

Attention mechanism-based approaches are gaining increasing attention to further explore object detection in HSRIs. The focus was originally on machine translation to automatically select the most relevant portions of a source sentence and predict corresponding targeted words. For some computer visual tasks, such as classification [32], change detection [33], and object detection [34], an attention mechanism is generally designed to select important weights in the learned feature maps and strengthen feature description by focusing on more important information. Introducing the attention mechanism into a detection framework significantly boosts the accuracy and efficiency of object detection in HSRIs. In the field of remote sensing, widely applied attention models are mainly the spatial-attention, channel-attention, and spatial–channel hybrid attention models. For example, Chen et al. [35] proposed a novel detector by integrating multiscale information with the spatial- and channel-attention models, fusing features. The two former attention mechanisms enable effectively improving the detection performance of HSRIs. Yang et al. [36] proposed the SCRDet detector for small-, cluttered-, and rotated-object detection. In its implementation, SCRDet jointly learnt two attention models to suppress noise and focus on extracting more object region features, which resulted in great detection performance for small and densely arranged objects in remote-sensing imagery. Wang et al. [37] developed an end-to-end multiscale visual attention network (MS-VAN) that adopted a skip-connected encoder–decoder model to learn attention weights at different scales and then extract multiscale features for object detection. Lu et al. [38] proposed an novel one-stage attention-based detector, attention and feature fusion SSD, which first designed a multilayer fusion module in order to explore complete semantic information and extract multiscale features. Then, a dual-path attention model was introduced into the learning process of the whole detection framework to overcome problems induced by background noise and ensure the extraction of key features for object detection.

2.4. Graph Convolutional Networks in Remote-Sensing Vision Applications

The operations of CNN are generally executed on regular structured data regions that lack the ability to explore the implicitly structural characteristics underlying the data. To address this drawback, graph convolutional networks (GCNs) have been proposed that enable directly carrying out graph convolutions on graph nodes and their spatial neighbors. As an extension of traditional CNNs, GCNs have been extensively investigated in various visual object tasks. For instance, Yang et al. [39] proposed the DSGCN method that first utilized a remarkable two-stage mask-RCNN [40] detection framework to iteratively detect and select high-quality proposals, and construct subgraphs. Then, a GCN-D network was exploited for cluster detection. Shi et al. [41] developed an adaptive GCN that could learn the multiple topological relations of graph-structured data across different GCN layers in an end-to-end

manner; then, the set of learned graph-based features could effectively enhance the accuracy of active recognition. He et al. [42] proposed a novel object-detection framework consisting of a GCN-based relation inferring module and a self-adopted attention module that could enhance the discriminative ability of object features by aggregating the geometric and visual relationships, and efficiently improve object-detection performance. Due to the powerful feature representation capacity of GCNs, they are also widely applied in the field of remote sensing. Compared with natural images, remote-sensing imagery has more complex content, especially HSRIs, for which it is hard to mine the intrinsic relationships hidden in diverse imagery information. In this situation, thanks to the ability of GCNs to effectively capture spatial relations from complex contextual patterns from HSRIs and the performance for some visual tasks in HSRIs was improved. For instance, Chaudhuri et al. [43] presented a Siamese graph convolutional network (SGCN) for remote-sensing image retrieval tasks that utilized a graph correlation-based metric to measure the similarity between a pair of remote images, producing satisfactory results. Khan et al. [44] proposed a multilabel remote-sensing scene-recognition method based on GCNs that first mainly utilized a region adjacency graph (RAG) to obtain a series of discriminative graph-based features from remote-sensing images, explore more complete semantic information for remote-sensing scenes, and enhance performance for multicategory scene recognition. Although GCNs have had remarkable success in some remote-sensing applications, there still exist potential limitations for HSRI-based visual tasks, such as object detection. The main reason is that these approaches generally ignore exploration for diverse-type and multiple-scale graph relations in HSRIs, which cannot further improve object-detection performance.

3. Proposed Method

The goal of the proposed HA-MHGEN is to enhance the performance of multiclass geospatial object detection by exploring abundant semantic and spatial information. Figure 2 depicts the whole scheme of the HA-MHGEN framework. A hierarchical spatial graph model and a semantic graph model were constructed. The former graph characterizes multiscale spatial relation features on the basis of proposed regions on multiscale feature maps that are acquired from a backbone network in the faster R-CNN architecture. The latter graph was adopted for global semantic correlation representation among object category labels. After that, the series of constructed spatial graphs are imported into self-attention-driven multistream embedding GCNs for intra- and interobject spatial contextual relation discrimination and reasoning. The constructed semantic graphs are also fed into another GCN for implicitly intrinsic semantic relation learning. Then, the two kinds of deduced relation features follow a novel cross-attention module for learning the association between spatial and semantic information that provides complementarity to them in order to further enhance the discrimination ability of the fused features. Lastly, the learned features are fed into the regression and classification layers for multiclass remote-sensing object detection. The design and implementation details of the subnetworks of our proposed HA-MHGEN are introduced in the following subsections.

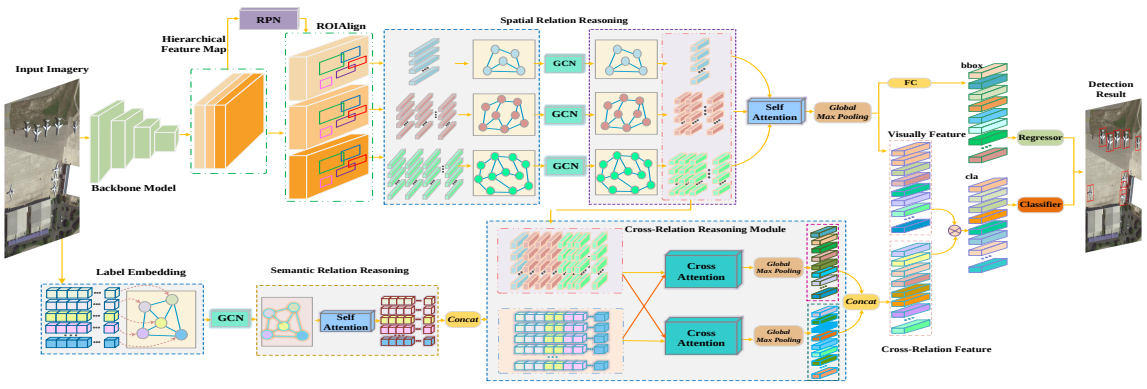

Figure 2. The whole architecture of the proposed HA-MHGEN for remote-sensing object detection.

3.1. Hierarchical Spatial Graph and Semantic Graph Construction

(1) Hierarchical Spatial Graph Construction

Given an arbitrary input image $I = \{I_1, I_2, \ldots, I_N\}$, we first employed a region proposal network to generate a series of candidate object proposal regions $R = \{r_i^{(c)}, i = 1, 2, 3, \ldots, N, c = 1, 2, 3.\}$ with various sizes. i denotes the number of proposals in each channel feature map, and c is the channel number of feature maps. Next, we adopted RoIAlign to extract the visual features of the generated object proposal on multiscale feature maps, which aims to project the acquired varying-size visual features into fixed-size features. On the basis of these postprocessed visual features crossing multiscale feature maps with fixed-dimensional representations, we first constructed a hierarchical graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ to represent the contextual intra- and interobject spatial and appearance relations. The layer number of our proposed hierarchical graph was empirically set to $c = 3$. Then, we could build the hierarchical graph from the two following perspectives:

(a) **Nodes**: Let $X_{(c)} = (x_c^1, x_c^2 \cdots x_{n_c}^p)$ denote the set of post-processed proposal features by the RoIAlign algorithm, where $x_c^p \in \mathbb{R}^N$, p is the number of object proposal feature in the c-th feature map, N is the feature dimension. Specifically, each proposal region feature corresponded to a fixed-form feature vector $\boldsymbol{x}_{n_c}^p = [o_x, o_y, w, h]$, where (o_x, o_y) denotes the spatial coordinate of the top-left point of each proposal region, h/w is the aspect ratio of each proposal region. The nodes of the three-layer hierarchical graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ can thus be defined as $\mathcal{V} = \{V^{(c)}, c = 1, 2, 3.\}$, where $V^{(c)}$ denotes the collection of graph nodes at the c-th feature map, and each node $v_c^{(p)} \in V^{(c)}$ corresponds to a pre-extracted object proposal that is assigned postprocessed CNN-based features x_c^p.

(b) **Edges**: The spatially connected relations of each pair of graph nodes among different feature maps are controlled by edge term $\mathcal{E} = \{E^{(w)}, E^{(b)}\}$, where $E^{(w)}$ represents the edges within each feature channel, and $E^{(w)}$ denotes the edges that cross different channels of feature maps. In each feature layer, each edge $e_{ij}^w \in E^w$ connects nodes v_i^w and v_j^w if they either are neighbors or have the same adjacency nodes. Among different feature layers, each edge $e_{ij}^a \in E^a$ denotes the spatially connected relation between nodes v_i^a and v_j^a that are acquired from different layers of feature maps. The connections of the pairs of graph nodes are governed by edge weights that indicate the spatial relations of object region proposals. The adjacency matrix that consists of the calculated inter- and intrafeature layer edge weights can be constructed with the following formulation:

$$A_{ij}^{spa} = \exp\left(-\frac{\left\|x_i - x_j\right\|_2}{\delta}\right), \tag{1}$$

where A_{ij}^{spa} denotes the adjacency matrix of hierarchical graph $\mathcal{G}$, and δ is a hyperparameter that was empirically set to $\delta = \{0.2, 0.8, 0.4\}$ in the following experimental verification on three datasets.

(2) Semantic Graph Construction The exploitation of semantic graphs provides an effective way of capturing statistical correlations among pairs of object category labels.

(a) **Nodes**: To construct semantic graph $\mathcal{G}_{sem}(\mathcal{V}_{sem}, \mathcal{E}_{sem})$, we first adopted Word2Vec [45] to transform each object label into the corresponding semantic feature vector. The collection of these word embeddings of the object category labels can be denoted as $S^{(i)} = (s^1, s^2 \cdots s^k, i = 1, 2, \ldots, k)$, where $s^i \in \mathbb{R}^N$. We regarded each feature vector of the semantic embedding as a graph node, and the node collection is represented as $v_{sem}^i = (s^1, s^2 \cdots s^p, i = 1, 2, \ldots, k)$, where $v_{sem}^i \in \mathcal{V}^k$.

(b) **Edges**: The collected semantic relations between pairs of object category labels rely on their co-occurence relations that are leveraged to depict the likelihood of each pair of object categories appearing simultaneously in the same image. In this situation, a semantic edge is denoted as $\mathcal{E}_{sem}\{v_{sem}^i, v_{sem}^j\} = e_{sem}\{P(v_{sem}^i, v_{sem}^j)\}$, which is acquired by calculating the conditional probability $P(v_{sem}^i, v_{sem}^j)$ that labels i and j would occur together in the same

image. For acquiring adjacency matrix A_{ij}^{sem}, we first computed the number of occurrences of object category label pairs in the dataset in order to obtain occurrence matrix $M \in \mathbb{R}^{N \times N}$, where N is the number of object categories, and M_{ij} denotes the number of times that labels i and j occur together. Then, we calculated the number of times that a single label i occurs in dataset Q_i. After that, the formulation of conditional probability of label pairs is denoted by $P_{i \to j} = M_{ij}/Q_i$, where $P_{i \to j}$ and indicates the probability that label i occurs when label j does, and $P_{i \to j}$ is not equal to $P_{j \to i}$. On the basis of the conditional probability of the label pairs, we could obtain semantic adjacency matrix A_{ij}^{sem} as follows:

$$A_{ij}^{sem} = \begin{cases} 1, & \text{if } P_{i \to j} \geq \eta \\ 0, & \text{otherwise} \end{cases}; \tag{2}$$

where η is a threshold that is used to filter out noisy edges if a pair of category labels never co-occur in the dataset. Threshold η was empirically set to $\delta = \{0.4, 0.6, 0.4\}$ in the following experiments on the three datasets.

3.2. Hierarchical Spatial and Semantic Relation Learning

With hierarchical spatial matrix A_{ij}^{spa} and semantic adjacency matrix A_{ij}^{sem}, constructed graph-structured feature matrices are deeply investigated by the self-attention driven GCN network to completely learn semantic and spatial relations. For capturing comprehensive spatial and semantic relations, we first constructed the GCN on the constructed spatial and semantic graphs in order to acquire the set of basic spatial and semantic relation graph units.

(1) Spatial Relation Learning Module

Generally, the operation of a GCN can be regarded as the generation of the GCN to the graph domain, which could effectively enhance relation representation among derived graph-structured features. In this paper, we applied a spectral GCN to relation learning. For the given graph-structured feature vector $x_c^p \in \mathbb{R}^N$, the spectral graph convolution was equivalent to the product of x_c and filter kernel g_θ, $\theta \in \mathbb{R}^N$ [46]. The matrix-vector form of GCN operation can thus be formulated as follows:

$$x_c * g_\theta = \boldsymbol{G}_\theta(\mathcal{L}) x_c; \tag{3}$$

where G_θ denotes a diagonal matrix whose elements are the parameterized eigenvalues via function θ, $\mathcal{L}$ is the normalized Laplacian that can be denoted as $\mathcal{L} = I_N - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, A is the corresponding adjacency matrix, and D is the corresponding degree matrix. According to the Chebyshev expansion of the graph Laplacian [47], the first-order approximation of the above formulation can be defined as follows:

$$x_c * g_\theta \approx \theta\left(I_N + \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\right) x_c; \tag{4}$$

Then, the generalized operation of the matrix form on each graph convolutional layer can be formulated as follows.

$$\boldsymbol{X_c}^{l+1} = \boldsymbol{A}\boldsymbol{X_c}^{l}\boldsymbol{W}^{l}; \tag{5}$$

where ${X_c}^{l+1}$ and ${X_c}^{l}$ denote the input in the l-th layer of GCN and the output in the $l+1$-th layer of GCN, respectively, and $\boldsymbol{W}^l$ represents the learned weight matrix. After that, the operation of each GCN layer can be expressed as a nonlinear function:

$$\boldsymbol{X_c}^{l+1} = \text{LeakyReLU}\left(\boldsymbol{A}\boldsymbol{X_c}^{l}\boldsymbol{W}^{l}\right); \tag{6}$$

where LeakyReLU is the nonlinear operation with the negative input slope being equal to 0.01, which was adopted to normalize the final hierarchical spatial relation. For the constructed spatial adjacency matrix in each channel of feature map $A_{ij(c)}^{spa}$, where c is the

channel number of feature map with $c = 1, 2, 3$, graph propagation in the multiple-layer GCN for the deduction of corresponding spatial relations can be derived as follows:

$$\boldsymbol{X_{c(spa)}}^{l+1} = \text{LeakyReLU}\left(\boldsymbol{A_{ij(c)}^{spa}} \boldsymbol{X_{c(spa)}}^{l} \boldsymbol{W_{c(spa)}}^{l}\right); \tag{7}$$

where ${X_{c(spa)}}^{l}$ and ${X_{c(spa)}}^{l+1}$ denote the input and output of each spatial graph convolutional layer, respectively, and $\boldsymbol{W_{c(spa)}}^{l}$ is the corresponding learned spatial relation weight matrix.

Similar to spatial graph relation propagation by using a GCN, we performed the GCN operation on the constructed semantic graph, and took semantic adjacency matrix A_{ij}^{sem} as the input; the derived semantic graph relation could be achieved with the following multilayer GCN propagation:

$$\boldsymbol{X_{c(sem)}}^{l+1} = \text{LeakyReLU}\left(\boldsymbol{A_{ij(c)}^{sem}} \boldsymbol{X_{c(sem)}}^{l} \boldsymbol{W_{c(sem)}}^{l}\right); \tag{8}$$

where ${X_{c(sem)}}^{l}$ and ${X_{c(sem)}}^{l+1}$ denote the input and output of each spatial graph convolutional layer, respectively, and $\boldsymbol{W_{c(sem)}}^{l}$ is the corresponding learned spatial relation weight matrix.

(2) Self-Attention-Driven Spatial and Semantic Relation Module

To acquire the hierarchical spatial relationships across different feature scales (feather channels) and the contextual semantic relations, we first utilized the self-attention model [48] to capture the explicit relations between multiscale spatial graph feature maps. Then, this self-attention mechanism was also applied in order to deduce contextual semantic graph features. Such a self-attention model not only effectively captures explicit spatial relationships across multiple spatial scales, but also exploits implicitly contextual semantic relations between different kinds of remote-sensing objects. The internal operation of the self-attention mechanism is shown in Figure 3.

(a) Self-Attention-Driven Hierarchical Spatial Relation Learning

According to the obtained spatial relation features from the GCN, the multiscale spatial feature concatenation can be denoted as $X_{c(spa)}^{s} \in \mathbb{R}^{N \times M \times s}$. In the operation of the self-attention module, we first squeezed the spatial feature into a graph node descriptor $H_{spa}^{s} \in \mathbb{R}^{N \times M \times 1}$ along the node feature dimension, which was implemented by performing a standard 1D convolutional operation and the sigmoid activation function; this propagation could be formulated as follows:

$$\mathbf{H}_{spa}^{s} = \text{Sigmoid}\left(\text{Conv}^{1\times1}\left(X_{c(spa)}^{s}\right)\right) \tag{9}$$

where H_{spa}^{s} denotes the squeezed node features. Then, spatial attention weight matrix $\mathbf{A}_{spa}^{s} \in \mathbb{R}^{N \times M \times 1}$ could be acquired by performing normalization along the feature channel dimension to ensure that the attention weights sum up to 1; this operation can be defined as follows:

$$\mathbf{A}_{spa}^{s} = \frac{\mathbf{H}_{spa}^{s}}{\sum_{s=1}^{c} \mathbf{H}_{spa}^{s}} \tag{10}$$

where $s \in [1, C]$ are the indices of feature channel dimension; we empirically set $C = 3$ as the number of the spatial feature channels (or feature spatial scale). We then adopted element-wise multiplication between the derived attention weight matrix and the spatial feature vectors in order to obtain attention weighted hierarchical spatial feature vectors $H_{spa}^{s'}$, which are defined as follows:

$$\mathbf{H}_{spa}^{s'} = \sum_{s=1}^{3} \mathbf{H}_{spa}^{s} \otimes \mathbf{A}_{spa}^{s} \tag{11}$$

where ⊗ denotes the Hadamard product. By using the self-attention mechanism on the multiscale spatial graph learning process, we allowed the self-attention-driven GCN to deduce the complete intra- and interobject channel-wise hierarchical spatial relations.

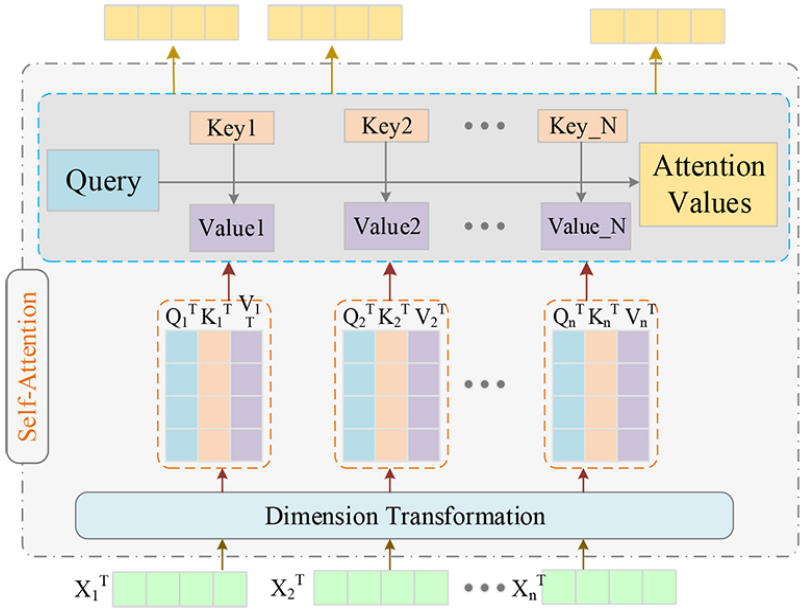

Figure 3. The internal operation of the self-attention mechanism for hierarchical spatial and contextual semantic relation learning.

(b) Self-Attention-Driven Contextual Semantic Relation Learning

Similar to the hierarchical spatial graph learning process, we could also capture the contextual semantic relations between different kinds of objects. Specifically, for semantic concatenation $X^{s}_{c(sem)} \in \mathbb{R}^{N\times M\times s}$, we also squeezed the semantic features into a graph node descriptor, and then executed a standard 1D convolutional operation and the sigmoid activation function in order to obtain the semantic graph node descriptor as follows:

$$\mathbf{H}^{s}_{sem} = \text{Sigmoid}\left(\text{Conv}^{1\times 1}\left(\mathbf{X}^{s}_{c(sem)}\right)\right) \tag{12}$$

where H^{s}_{sem} denotes the squeezed node features. Afterwards, the semantic attention weight matrix $\mathbf{A}^{s}_{sem} \in \mathbb{R}^{N\times M\times 1}$ could be acquired by performing normalization along the feature channel dimension to ensure that the attention weights sum up to 1; this operation can be defined as follows:

$$\mathbf{A}^{s}_{sem} = \frac{\mathbf{H}^{s}_{sem}}{\sum_{s=1}^{c}\mathbf{H}^{s}_{sem}} \tag{13}$$

Then, we could also obtain attention weighted contextual vectors with the following operation:

$$\mathbf{H}^{s'}_{sem} = \sum_{s=1}^{3}\mathbf{H}^{s}_{sem}\otimes\mathbf{A}^{s}_{sem} \tag{14}$$

where $H^{s'}_{sem}$ are the contextual semantic feature vectors.

(3) Cross-Attention-Driven Spatial and Semantic Relation Learning

To further explore the interactions of inter- and intraobject spatial and semantic relations, we adopted the multihead attention module to learn the cross-relation feature representations between the spatial and semantic graph relation branches, as shown in Figure 4.

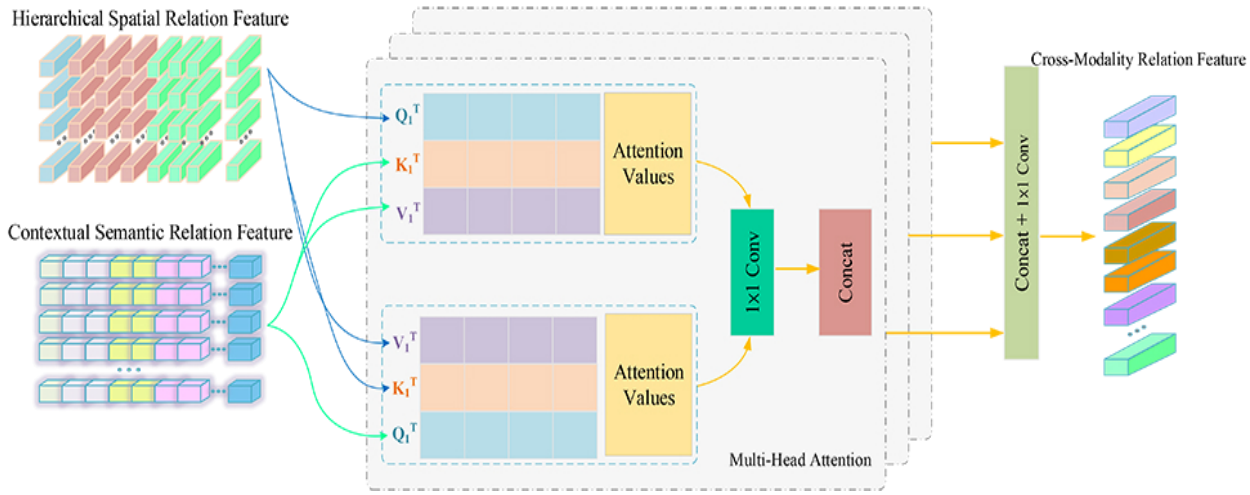

Figure 4. Architecture of the multihead attention mechanism for cross-modality relation fusion.

(a) Multihead attention module

The operation of the multihead attention mechanism [49] is mapping a series of queries and key-value pairs to the weighted sum form of inputted values. Specifically, for given features E, we first adopted the linear projection for computing three embedded feature terms in the i-th head; this operation can be defined as the following formulation:

$$\mathbf{Q}_i = \mathbf{E}\mathbf{W}_i^Q, \mathbf{K}_i = \mathbf{E}\mathbf{W}_i^K, \mathrm{V}_i = \mathbf{E}\mathbf{W}_i^V; \tag{15}$$

where queries Q_i, keys K_i and value Q_i denote the three input terms with dimensions d_k, d_k, d_v, respectively. $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \in \mathbb{R}^{d\times d}$ are trainable parameter matrices. After that, we calculated the i-th attention weight matrix $\mathbf{A}_i \in \mathbb{R}^{(N+1)\times(N+1)}$ of the corresponding input features as follows:

$$\mathbf{A}_i = \mathrm{softmax}\left(\frac{\mathbf{Q}_i\mathbf{K}_i^\top}{\sqrt{d_k}}\right)\mathbf{V}_i; \tag{16}$$

where $\top$ denotes the matrix transpose operation. On this basis, the multihead attention matrix could be acquired by first performing an attention function on multiple linear projections of the queries, keys, and values in a parallel way. Then, these values are concatenated and once again projected. The calculated operation of the multihead attention can be defined as follows:

$$\mathrm{MultiHead}\ (\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \ \mathrm{Concat}\ (\mathbf{A}_1, \ldots, \mathbf{A}_h)\mathbf{W}^o; \tag{17}$$

where $\mathbf{W}^o$ denotes the weight matrix in linear output function of the multihead attention module, and h is the number of attention heads.

(b) Spatial and Semantic Relation Fusion with the Multihead Attention Module

Due to the advantage of the multihead attention model for mining interactions between cross-modal information, we adopted it for spatial and semantic relation fusion. Specifically, for the derived spatial relation feature concatenation X^{Spa} and semantic relation feature concatenation X^{Sem}, we first defined the parallel attention layers as l. Then, we utilized linear projections to transform each spatial and semantic feature into the corresponding multihead query, key, and value as follows:

$$\mathbf{Q}_i^{spa} = \mathbf{X}_i^{spa}\mathbf{W}_i^Q, \mathbf{K}_i^{sem} = \mathbf{X}_i^{sem}\mathbf{W}_i^K, \mathbf{V}_i^{sem} = \mathbf{X}_i^{sem}\mathbf{W}_i^V; \tag{18}$$

where $X^{Spa} \in \mathbb{R}^{M\times\frac{N}{h}}$, $X^{Sem} \in \mathbb{R}^{M\times\frac{N}{h}}$ are divided by spatial feature X^{Spa} and semantic feature X^{Sem} in the feature channel direction, respectively. $\mathbf{W}_i^Q \in \mathbb{R}\frac{d}{l}\times\frac{d_k}{h}$, $\mathbf{W}_i^K \in \mathbb{R}\frac{d}{l}\times\frac{d_k}{h}$,

$\mathbf{W}_i^V \in \mathbb{R}^{\frac{d}{l}} \times \frac{d_v}{h}$ denote the projection weight matrix. We empirically set $d_k = d_v = d$ for each of these weight matrices, and $i \in \{1, \cdots, l\}$ is the number of the attention layers. Then, the output of the i-th attention layer can be defined as follows:

$$\begin{aligned} \text{Head}_i(\mathbf{Q}_i^{spa}, \mathbf{K}_i^{sem}, \mathbf{V}_i^{sem}) &= \text{softmax}\left(\frac{Q_i^{Spa}\left(K_i^{Sem}\right)^\top}{\sqrt{d_k/l}}\right) V_i^{Sem} \\ &= \text{softmax}\left(\frac{\mathbf{X}_i^{Spa}\mathbf{W}_i^Q\left(\mathbf{X}_i^{Sem}\mathbf{W}_i^K\right)^\top}{\sqrt{d_k/h}}\right)\mathbf{X}_i^{Sem}W_i^V; \end{aligned} \tag{19}$$

After that, intramodality information from spatial and semantic relations can be obtained with the following formulation:

$$\mathbf{H}_c^{Spa\rightarrow Sem} = \text{ Concat }(\text{ Head }_1, \ldots, \text{ Head }_l)\mathbf{W}^O \tag{20}$$

where $\mathbf{W}^O \in \mathbb{R}^{d_v \times d}$ denotes the weight matrix in a linear function, and $\mathbf{H}_c^{Spa\rightarrow Sem}$ is the final intermodality feature representation of the spatial and semantic relations. Similar to the above cross-modality fusion process from spatial to semantic relation, the transformation from a semantic into a spatial relation can also be defined as in the following formula:

$$\mathbf{Q}_i^{spa} = \mathbf{X}_i^{sem}\mathbf{W}_i^Q, \mathbf{K}_i^{spa} = \mathbf{X}_i^{spa}\mathbf{W}_i^K, \mathbf{V}_i^{sem} = \mathbf{X}_i^{spa}\mathbf{W}_i^V; \tag{21}$$

$$\begin{aligned} \text{Head}_i(\mathbf{Q}_i^{sem}, \mathbf{K}_i^{spa}, \mathbf{V}_i^{spa}) &= \text{softmax}\left(\frac{Q_i^{Sem}\left(K_i^{Spa}\right)^\top}{\sqrt{d_k/l}}\right) V_i^{Spa} \\ &= \text{softmax}\left(\frac{\mathbf{X}_i^{Sem}\mathbf{W}_i^Q\left(\mathbf{X}_i^{Spa}\mathbf{W}_i^K\right)^\top}{\sqrt{d_k/h}}\right)\mathbf{X}_i^{Spa}W_i^V; \end{aligned} \tag{22}$$

The intramodality information from semantic and spatial relations could thus also be acquired with the following formulation:

$$\mathbf{H}_c^{Sem\rightarrow Spa} = \text{ Concat }(\text{ Head }_1, \ldots, \text{ Head }_l)\mathbf{W}^P; \tag{23}$$

where $\mathbf{W}^P$ is the corresponding weight matrix in a linear function, and $\mathbf{H}_c^{Sem\rightarrow Spa}$ denotes the final intermodality feature representation of the semantic and spatial relations. Lastly, the joint feature representation $\mathbf{H}_c^{cross}$ of the spatial and semantic relations can be achieved with the following formula:

$$\mathbf{H}_c^{cross} = \text{Conv}\left(\text{Concat}\left(\mathbf{H}_c^{Spa\rightarrow Sem}, \mathbf{H}_c^{Sem\rightarrow Spa}\right)\right) \tag{24}$$

where Concat denotes the concatenation operation in the feature direction, and Conv is a standard 1D convolutional operation. Now, we summarize the self-attention-driven hierarchical spatial relation features, contextual semantic features, and their cross-modality relation features using a global max-pooling layer that can be defined as follows:

$$\mathbf{H}^f = \text{Concat}\left(\mathbf{H}_{spa}^{s'}, \mathbf{H}_{sem}^{s'}, \mathbf{H}_c^{cross}\right) \tag{25}$$

where H^f denotes the fused relation features of the spatial and semantic relations, and their intramodality relation features.

3.3. Objective Function of HA-MHGEN

Intrinsically, the proposed HA-MHGEN is a two-stage object-detection approach. We optimized a novel objective loss function that was specifically designed for multiclass

geospatial object detection. The objective loss function of HA-MHGEN is composed of classification loss, localization loss, and margin-based ranking loss. The overall multitask loss function can be defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{Cls}} + \alpha\mathcal{L}_{\text{Loc}} + \beta\mathcal{L}_{\text{MR}} \tag{26}$$

where $\mathcal{L}_{\text{Cls}}$ denotes classification loss, $\mathcal{L}_{\text{Reg}}$ is object localization loss, and $\mathcal{L}_{\text{MR}}$ is RPN loss. α and β denote two hyperparameters that can be regarded as the two-weight decaying coefficients of regularization terms. The main role of these two hyperparameters is to control the trade-offs of the three regularization terms in the overall multitask loss function during the learning process.

For the task of object classification, we adopted the cross-entropy loss function to force the predicted object categories to be aligned with the ground-truth categories; the classification loss formula is defined as follows:

$$\mathcal{L}_{\text{cls}} = \frac{1}{N_{\text{cls}}}\sum_{n} -y_n \log \mathcal{H}_n; \tag{27}$$

where y_n denotes the ground-truth object classification label, N is the number of object categories, and $\mathcal{H}_n$ represents the final classification score for each object category that can be calculated with a softmax function:

$$\mathcal{H}_n = \frac{e^{H_f^n}}{\sum_{n=1}^{N} e^{H_f^n}} \tag{28}$$

where $\mathcal{H}_n$ is the final classification possibility of object category n, and $\sum_{n=1}^{N} \mathcal{H}_n = 1$.

For the task of object localization, we utilized smooth L_1 loss to penalize misalignments between the predicted object proposal and ground-truth regions. Smooth L_1 loss [50] is defined as follows:

$$\text{Smooth } L_1(x) = \begin{cases} 0.5x^2, & \text{if x} < 0 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{29}$$

where x denotes the difference between the predicted IoU score and the true IoU value between the predicted object proposal and ground-truth regions. The general process of object proposal region localization can be regarded as bounding-box regression from a predicted object box to a nearby ground-truth one, whose formulation can be defined as follows:

$$\begin{aligned} b_x &= \left(o'_x - o_x\right)/w, \quad b_y = \left(o'_y - o_y\right)/h \\ b_w &= \log\left(w'/w\right), \quad b_h = \log\left(h'/h\right) \end{aligned} \tag{30}$$

where b_x, b_y, b_w and b_h are the set of box regression parameters, where (o'_x, o'_y) denotes the spatial coordinate of the top-left point of each predicted proposal region, and h' and w' are the corresponding width and height of the predicted object proposal region, respectively. (o_x, o_y) denotes the spatial coordinate of the top-left point of each object ground-truth region, and w and y are the corresponding width and height of the ground-truth one, respectively.

In addition, we introduce margin-based ranking loss $\mathcal{L}_{\text{MR}}$ for predicting the similarity of the pre-extracted region proposals to the ground-truth ones whose formulation is defined as [51]:

$$\begin{aligned} \mathcal{L}_{\text{MR}} &= \sum_{i=1}^{P} \left\{ \begin{array}{l} x_{n_c}^i \times \max\{m^+ - s_i, 0\} \\ +(1 - x_{n_c}^i) \times \max\{s_i - m^-, 0\} + \Delta_i \end{array} \right\} \\ \Delta_i &= \sum_{j=i+1}^{P} \left\{ \begin{array}{l} \left[x_{n_c}^i = x_{n_c}^j\right] \times \max\{|s_i - s_j| - m^-, 0\} \\ +\left[x_{n_c}^i \neq x_{n_c}^j\right] \times \max\{m^+ - |s_i - s_j|, 0\} \end{array} \right\} \end{aligned} \tag{31}$$

where $x_{n_c}^i$ represents the feature vector of the i-th object proposal region, and P is the number of pre-extracted object proposal regions in the RPN. s_i is a classification score that denotes the foreground probability with regard to the i-th predicted object proposal region. m^+ and m^- indicate the upper and lower limits of the desired probabilities, respectively.

4. Experiments

In this section, we first introduce the datasets that were used for the experimental verifications: DIOR, NWPU VHR-10, and DOTA. Then, we evaluate the performance of the proposed method on the three public benchmark datasets above and compare our proposed object-detection method with 11 state-of-the-art object-detection methods to demonstrate its effectiveness and superiority.

4.1. Datasets and Evaluation Metrics

(1) Dataset Description

DOTA: A multiscale optical remote-sensing dataset that was released by the researchers of Wuhan University (WHU) [52]. This dataset contains 2806 images that were collected from various remote-sensing sensors and platforms, including the Google Earth service, and the GF-2 and JL-1 satellites; the dimensions of each image range from 800 × 800 to 4000 × 4000 pixels. The DOTA dataset comprises 188,282 object instances that exhibit various characteristics regarding scale, orientation, and shape among 15 categories of geospatial objects. These fully annotated object categories are airplanes, baseball diamonds (BD), bridges, ships, ground field tracks (GFTs), small vehicles (SVs), large vehicles (LVs), tennis courts (TCs), basketball courts (BCs), storage tanks (STs), soccer ball fields (SBFs), roundabouts (RAs), swimming pools (SPs), harbors, and helicopters (HCs). This dataset was split into three parts, the training, validation, and testing sets; it randomly selected half of the total images as the training set, one-sixth as the validation set, and one-third as the testing set.

DIOR: A large-scale publicly available benchmark dataset for remote-sensing object detection that was published by Northwestern Polytechnic University in 2020 [53] . DIOR contains 23,463 optical remote-sensing images and 192,472 instances covering 20 object categories. All images in the dataset were collected from Google Earth; the spatial resolutions ranged from 0.5 to 30 m, and each image was resized to 800 × 800 pixels. The sizes of object instances in DIOR have a wide range of variations regarding spatial resolutions and large variability among inter- and intracategory object instances. Its object categories are airplanes, airports, baseball fields (BFs), basketball courts (BCs), bridges, chimneys, dams, expressway service areas (ESA), expressway toll stations (ETSs), harbors (HBs), golf courses (GCs), ground track fields (GTFs), overpasses (OPs), ships, stadiums (STMs), storage tanks (STs), tennis courts (TCs), train stations (TSs), vehicles, and windmills (WMs). DIOR was divided into three subsets, namely, the training, validation, and testing sets. To ensure that the three subsets had similar distributions, we randomly selected 11,725 images as the trainval set; the training set contained 5862 images, the validation set had 5863 images, and the remaining 11,738 images were the testing set.

NWPU VHR-10: Another optical remote-sensing object-detection dataset that is a ten-category geospatial object-detection dataset that was labeled by Northwestern Polytechnic University in 2014. Its ten object categories are airplanes, ships, storage tanks (STs), baseball diamonds (BDs), tennis courts (TCs), basketball courts (BCs), ground track field (GTFs), harbors, bridges, and vehicles. This dataset is consists of multisource and multiresolution remote-sensing images. The total number of the whole dataset's images is 800, of which 715 were collected from Google Maps with a spatial resolution range from 0.5 to 2 m, and the remaining 85 were pan-sharpened color-infrared images with a spatial resolution of 0.08 m that were collected from Vaihingen. Among the 800 images, 650 included at least one object to be detected. The remaining 150 images were mainly composed of background information with no object to be detected.

(2) Evaluation Metrics

To evaluate our proposed detection approach, the average precision (AP) score [54] was adopted to assess the multiple categories' object detection in optical remote-sensing imagery. AP calculation is usually dependent on an assessment indicator called the intersection over union (IoU), which is used for estimating whether a detected object is correct. The IoU is the overlap ratio between the predicted bounding boxes and the corresponding ground truths. The IoU evaluates the degree of coincidence between the bounding box S_{gt} of the ground truth and the predicted bounding box S_p on the basis of the Jaccard index. The formulation of the IoU is defined as follows:

$$\mathrm{IoU} = \frac{\mathrm{area}\left(S_p \cap S_{gt}\right)}{\mathrm{area}\left(S_p \cup S_{gt}\right)}; \tag{32}$$

where $\mathrm{area}\left(S_p \cap S_{gt}\right)$ denotes the intersectional area between predicted bounding box S_p and ground truth S_{gt}, $\mathrm{area}\left(S_p \cup S_{gt}\right)$ is the union area of S_p and S_{gt}. During the implementation of the training procedure, the selection of correctly detected results relies on the given threshold of the IoU. If the IoU value is larger than 0.5, the predicted bounding boxes are true positives (TPs), while predicted bounding boxes with an IoU value smaller than 0.5 are false positives (FPs). Detection precision is thus defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}; \tag{33}$$

where TP indicates the number of correctly detected objects, and FP represents the number of objects in which the predicted results were not matched with any bounding boxes of the ground truths. To assess the performance of the detection model, the percentage of the total true objects correctly detected by the detection model is also needed. Therefore, recall is defined as follows:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}; \tag{34}$$

where FN represents the number of objects that were not detected.

4.2. Implementation Details and Parameter Analysis

In the experiment, we selected the classical faster R-CNN as the baseline model, which was integrated with ResNet-101 [55], pretrained on ImageNet [56] for a series of basic feature extraction. To construct the spatial relation graph, the input number of object region proposals that were generated on each image was set to 256, and the IoU threshold of the nonmaximal suppression (NMS) operation was set to 0.7 in order to remove redundant bounding boxes. Then, the ROI align operation was utilized to boost region-proposal-based feature processing. For multicategory object label representation in the branch of semantic relation network, Word2Vec with 300 dimensions was adopted for semantic label feature embedding. During the training process, we utilized the stochastic gradient descent (SGD) algorithm to optimize the parameters of the proposed HA-MHGEN, and after each decay step, it was divided by 10. For the DOTA, DIOR, and NWPU VHR-10 datasets, the overall iterations were all set to 150k. The momentum of SGD was set to 0.9 with weight decay of 0.0005, and the batch size was fixed to 16. The dropout rate was set with a probability of 0.3, which was employed to prevent the overfitting problem. In the stage of cross-attention-driven spatial and semantic fusion, the number of heads in the multihead attention module was set to 4. The upper and lower limits of margin probabilities m^+ and m^- in the loss function were lastly set to 0.7 and 0.3, respectively. The proposed HA-MHGEN was implemented on the Pytorch 1.4.0 framework with two GeForce GTX1080ti GPUs.

In the training process, learning rate η, and the two hyperparameters of regularization α and β played a more important role in optimizing HA-MHGEN. The combinations of the above three parameters can effectively result in the final detection performance.

During the training process of our proposed MA-MHGEN model, learning rate η and the two hype-parameters α and β in the overall loss function were three important parameters that could affect object-detection performance. For different datasets, we set different optimal parameter combinations. Then, we first investigated how object detection is affected by different parameter combinations on the validation sets of the three datasets. In our work, the initial learning rates were set to $\eta = \{0.0005, 0.001\}$, which were reduced by multiplying them by 0.1 every 50 epochs. Training stopped after the final epochs or once there was no improvement on the corresponding validation sets. The two hyperparameters were set to $\alpha = \{0.01, 0.02, 0.03, 0.04, 0.05\}$ and $\beta = \{0.01, 0.02, 0.03, 0.04, 0.05\}$. Figure 5 shows the corresponding training and validation loss curves of the HA-MHGEN on the aforementioned three datasets, and Figure 6 indicates the object results that are measured in terms of mean AP (mAP) under different parameter settings on the three datasets. Figure 6 shows that the highest detection performance on the DOTA validation set was 78.32%, which was obtained with $\eta = 0.001$, $\alpha = 0.05$, and $\beta = 0.01$. Then, the best detection result on the DIOR validation set was 74.72%, which was acquired from the parameter combination of $\eta = 0.0005$, $\alpha = 0.04$, and $\beta = 0.02$. The greatest detection performance on NWPU VHR-10 was 93.39%, which was achieved with the combination of $\eta = 0.001$, $\alpha = 0.02$, and $\beta = 0.04$. On the basis of the above experimental verifications, we empirically adopted the above set of optimal parameter combinations for the following experiments on the three datasets.

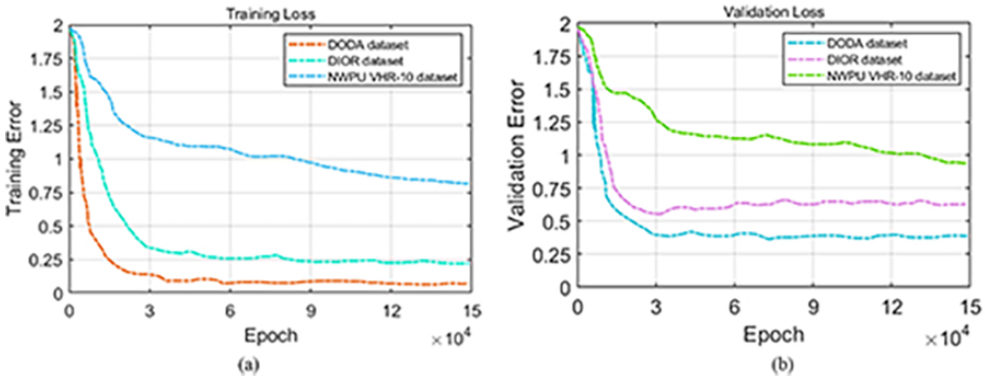

Figure 5. Training errors and validation errors using the proposed HA-MHGEN on three public datasets. (**a**) Training errors on three public datasets. (**b**) Validation errors on three public datasets.

In our work, relation learning and discrimination were dependent on the spatial and semantic graph relation matrices, and the connection between each pair of graph nodes was governed by an edge weight that reflected their relationship. In a relation learning process, stronger relationships of spatial or semantic graph nodes can deduce better relation feature representation, and can further improve object-detection performance. Then, we also investigated how object detection was affected by spatial and semantic graph edge parameters δ and μ. We also verified their relation on the validation set of the three datasets above. Figure 7 reports the object results based on the mAP under different parameter settings on the three datasets. Figure 7 shows that these two hyperparameters moderately affected the object-detection results, and the best results were obtained with $\delta = 0.2, \mu = 0.4$ on DOTA, $\delta = 0.8, \mu = 0.6$ on DIOR, and $\delta = 0.4, \mu = 0.4$ on NWPU VHR-10. Consequently, we empirically set the three parameter combinations above for the following experimental verifications.

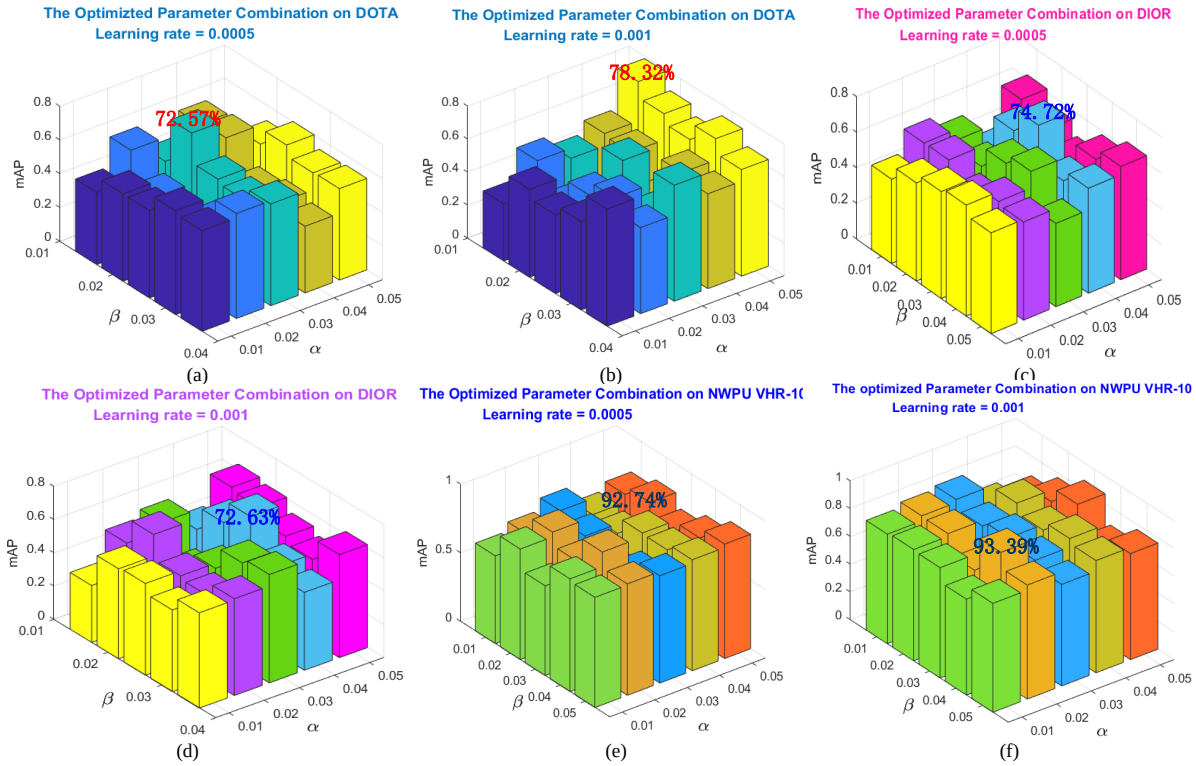

Figure 6. Multiclass object-detection results in terms of mAP values over all object categories on three public datasets under different parameter combinations. (**a**) Learning rate $\eta = 0.0005$ on DOTA. (**b**) Learning rate $\eta = 0.001$ on DOTA. (**c**) Learning rate $\eta = 0.0005$ on DIOR. (**d**) Learning rate $\eta = 0.001$ on DIOR. (**e**) Learning rate $\eta = 0.0005$ on NWPU VHR-10. (**f**) Learning rate $\eta = 0.001$ on NWPU VHR-10.

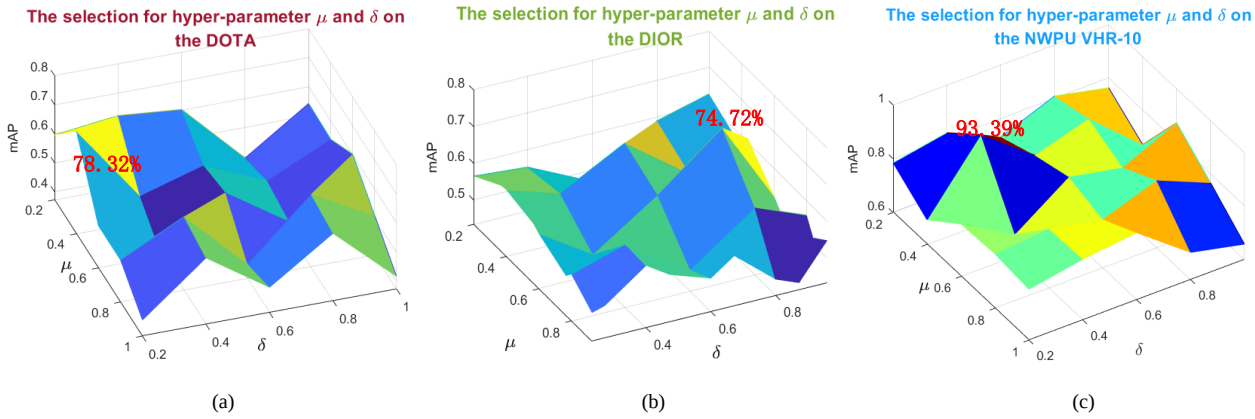

Figure 7. Selection of parameters δ and μ based on object-detection performance in terms of mAP values over all object categories on three public datasetS under different parameter combinations. (**a**) Selection of parameters δ and μ on DOTA. (**b**) Selection of parameters δ and μ on DIOR. (**c**) Selection of parameters δ and μ on NWPU VHR-10.

During the implementation of MA-MHGEN, the layer number of the hierarchical graph was also a critical parameter to be considered. Therefore, we also analyzed its influence on object-detection performance using the validation sets of the above datasets. Keeping other settings unchanged, the number of hierarchical graph layers varied in the set of $\{1, 8\}$; the overall mAP values that were obtained from the three datasets are shown in Figure 8. mAP was decreased if the number of hierarchical graph layers was higher than 3. The best performance of MA-MGGEN on the three datasets was also under the

condition that the graph layer number was set to 3. Therefore, we empirically set $c = 3$ in the following experimental verifications.

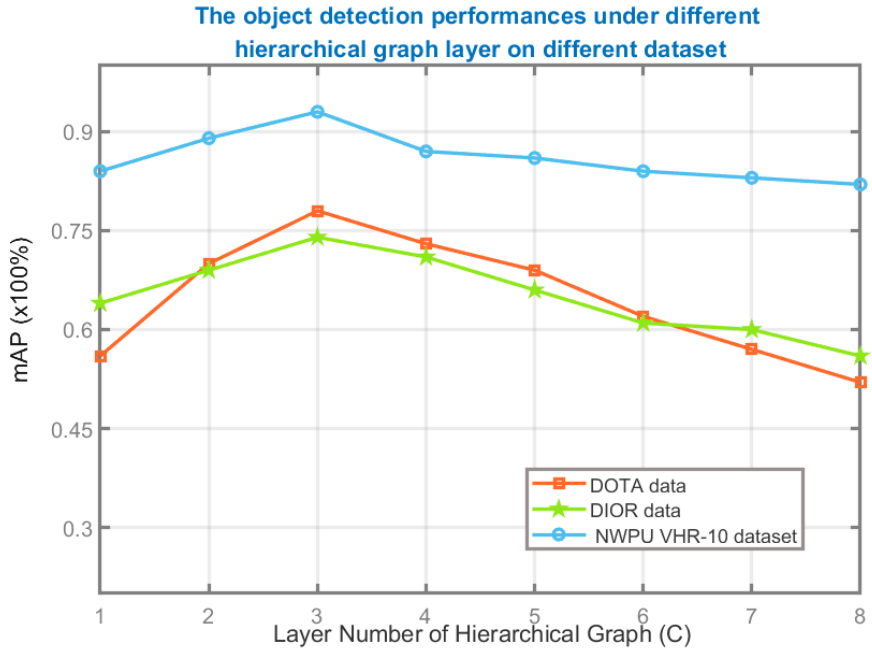

Figure 8. Object-detection performance using different hierarchical graph layers on three datasets.

4.3. Comparisons and Analysis Using Different Network Configurations

To investigate the impact of different relation modules explored by HA-MHGEM, we adjusted various network configurations to examine their detection performance. Five types of configurations were adopted for experimental comparison and analysis: the spatial relation module (Spa-Ra), semantic relation module (Sem-Ra), self-attention-driven spatial relation module (SA-Spa-Ra), self-attention-driven semantic relation module (SA-Sem-Ra), and the multihead attention driven spatial and semantic relation module (MH-SS-Ra). Table 1 indicates the detailed quantitative evaluation results in terms of mAP values, which were achieved by using the above network configurations. For the DOTA dataset, the mAP values of Spa-Ra and Sem-Ra were boosted by 4.68% and 2.26% compared to the mAP values of SA-Spa-Ra and SA-Sem-Ra, respectively. This demonstrates the effectiveness of the self-attention model in improving spatial- and semantic-relation reasoning.

Table 1. Performance of multiclass object detection on the DOTA, DIOR, and NWPU VHR-10 datasets that was achieved with different configurations of HA-MHGEN in terms of mAP values (%).

	Configuration					DOTA	DIOR	NWPU VHR-10
	Spa-Ra	**Sem-Ra**	**SA-Spa-Ra**	**SA-Sem-Ra**	**MH-SS-Ra**	**mAP**	**mAP**	**mAP**
HA -MHGEN	✓	-	-	-	-	69.34	67.92	86.35
	✓	✓	-	-	-	72.13	69.15	89.78
	✓	✓	✓	-	-	74.02	71.42	90.47
	✓	✓	✓	✓	-	74.39	72.35	91.96
	✓	✓	✓	✓	✓	78.32	74.72	93.39

The multihead attention module (MH-SS-Ra) for cross-modality relation fusion and discrimination achieved 4.3% and 3.93% better mAP values compared to the configurations of SA-Spa-Ra and SA-Sem-R, respectively. This indicates the superiority of the cross-relation reasoning using the multihead attention mechanism. For the more challenging DIOR dataset, MH-SS-Ra also achieved the best overall detection accuracy (74.72%), which showed improvements of 6.8%, 5.57%, 3.3% and 2.37% compared to the configurations of

Spa-Ra, Sem-Ra, SA-Spa-Ra, and SA-Sem-Ra, respectively. For NWPU VHR-10, detection performance using MH-SS-Ra in terms of mAP value was further improved, achieving 7.04%, 3.61%, 2.92%, and 1.41% compared with the configurations of Spa-Ra, Sem-Ra, SA-Spa-Ra, and SA-Sem-Ra, respectively. These series of comparison results effectively demonstrate the advantage of hybrid attention for diverse relation discrimination among hierarchical spatial graphs and contextual semantic graphs, which also further boosts multiclass object-detection performance in HRSIs.

4.4. Quantitative and Qualitative Comparison and Analyses

To verify that our proposed HA-MHGEN was superior to other state-of-the-art approaches, we selected several advantageous CNN- and GCN-based object-detection approaches for quantitative and qualitative comparison and analysis. These comparative approaches consisted of two-stage object detectors faster R-CNN [7] and FPN [57], one-stage object detectors Yolo-v3 [58], FMSSD [30], and RetinaNet [59], anchor-free detectors FCOS [60] and SRAF-Net [61], typical keypoint-based detection methods CornerNet [62] and CenterNet [63], and representative GCN-based detection methods MGCN [64] and STGCN [65]. Several experiments were carried out on the DOTA, DIOR, and NWPU VHR-10 datasets. Then, we individually analyzed these comparative results on the three public benchmark datasets above.

4.4.1. Comparison and Analysis on the DOTA Dataset

The DOTA dataset consists of 15 categories of object instances with complicated visual diversity and high appearance ambiguity; different kinds of geospatial remote-sensing objects also have complex spatial distributions and underlying spatial relations. Thus, the dataset is very suitable for the detection performance verification of remote-sensing objects with explicit and implicit relation characteristics. On the DOTA dataset, we employed the proposed HA-MHGEN and several state-of-the-art detection approaches, and obtained a series of quantitative experimental results that are reported in Table 2. Our proposed HA-MHGEN achieved higher detection indices of 78.32% than those of other state-of-the-art detection approaches, either classical two- or one-stage object detectors. Compared to classical two-stage methods faster R-CNN and FPN, the overall detection performance of mAP using HA-MHGEN achieved an improvement of 22.54% (78.32% versus 55.78%) and 6.32% (78.32% versus 72%), respectively. The main reason is that faster R-CNN only adopts the last feature layer of the backbone for object proposal region extraction, which failed to accurately predict and localize for some small objects with dense spatial distribution. Although the FPN adopts multiscale feature layers for region proposal extraction, it also ignores implicit spatial relations for object detection, which also results in some unsatisfying detection results. Comparing it with famous one-stage detection approaches in terms of mAP, namely, with RetinaNet, Yolo-v3, and FMSSD, the proposed HA-MHGEN achieved an improvement of 15.05% (78.32% versus 63.27%), 12.65% (78.32% versus 65.67%) and 5.89% (78.32% versus 72.43%), respectively. In general, the regression and prediction layer of the above existing one-stage detectors cannot always afford an abundant receipt field for distinguishing multiclass objects with a large-scale spatial span. Thus, these methods also cannot provide satisfactory object-detection results. We also ran some recent anchor-free-based one-stage detectors on DOTA for quantitative comparisons, namely, CornerNet, FCOS, SRAF-Net, and CenterNet. Compared with these anchor-free-based detection approaches, our HA-MHGEN outperformed CornerNet,FCOS, SRAF-Net, and CenterNet by 22.54% (78.32% versus 55.78%), 15.63% (78.32% versus 62.69%), 12.59% (78.32% versus 65.73%) and 4.38% (78.32% versus 73.94%) in terms of mAP, respectively. Technically, it is always hard for the pre-extracted object proposal regions of one-stage detectors to overcome the overfitting problem of continuous object scale representation within a large-scale spatial span. In addition, due to complex background interference and diverse object appearance ambiguity discrimination challenges in optical remote-sensing scenes, these anchor-free-based approaches cannot achieve the desirable detection performance.

To further indicate the superiority of our method, we also compared it with recent GCN-based detection approaches MGCN and STGCN. Table 2 Lines 11 to 13 indicate that HA-MHGEN improved the overall detection performance by 1.61% (78.32% versus 76.61%) and 1.48% (78.32% versus 76.84%) compared with MGCN and STGCN, respectively. Table 2 shows that MGCN achieved more better detection results for airplanes and ships, mainly because MGCN considers the background information around the objects for graph construction and spatial relation discrimination. Then, it can overcome the visual homogeneity between object and background information, and achieve better detection accuracy for the object categories of airplanes and ships. Nevertheless, the overall detection performance was still worse than that of our method. In addition, Figure 9 shows the ROC curves of all comparative approaches for object detection, which also demonstrates the better detection performance of our proposed method. The effectiveness and robustness of the proposed HA-MHGEN framework is conclusively indicated by the comprehensive and convincing experimental results.

Table 2. Comparisons with state-of-the-art object-detection approaches in terms of AP (%) and mAP (%) on the DOTA dataset.

Methods	Airplane	BD	Bridge	Ship	GTF	BC	SV	LV	TC	ST	SBF	RA	SP	Harbor	HC	mAP
CornerNet	67.85	78.94	53.59	27.08	68.05	63.75	31.39	46.52	87.96	53.57	62.46	69.79	43.06	58.79	23.94	55.78
Faster R-CNN	76.15	63.70	29.60	67.70	54.86	50.10	67.70	62.59	86.89	67.33	55.84	40.91	43.64	65.71	48.34	56.13
FCOS	88.81	71.63	56.35	68.92	40.86	67.31	49.37	74.58	89.56	70.77	44.63	70.97	42.71	66.90	36.97	62.69
RetinaNet	72.99	68.17	65.96	68.59	76.22	62.33	25.51	62.78	84.20	51.31	57.78	80.87	57.81	65.96	48.50	63.27
Yolo-v3	93.91	68.78	45.93	85.56	51.92	66.82	50.12	60.67	93.88	83.47	52.45	45.01	55.85	74.03	56.68	65.67
SRAF-Net	88.93	72.76	50.10	83.77	45.93	70.32	59.51	75.69	93.00	67.08	55.63	62.69	47.36	71.45	41.80	65.73
FPN	88.70	75.10	52.60	84.50	59.20	81.30	69.40	78.80	90.60	82.60	52.50	62.10	66.30	76.60	60.10	72.00
FMSSD	89.11	81.51	48.22	76.87	67.94	82.67	69.23	73.56	90.71	73.33	52.65	67.52	80.57	72.37	60.15	72.43
CenterNet	97.37	78.56	49.39	90.30	53.39	66.11	62.16	80.24	94.58	85.75	64.86	69.02	75.63	78.86	66.82	73.94
MGCN	98.13	82.74	56.15	90.46	57.14	67.98	66.85	83.76	96.17	86.98	65.78	72.54	78.19	80.62	67.26	76.71
STGCN	90.42	79.87	63.39	86.42	76.54	80.08	77.46	87.87	86.83	82.45	68.19	69.43	65.08	81.41	57.17	76.84
Our Method	94.57	81.07	61.76	88.67	78.16	81.98	79.15	88.62	88.79	82.13	69.87	70/17	67.67	83.15	58.98	78.32

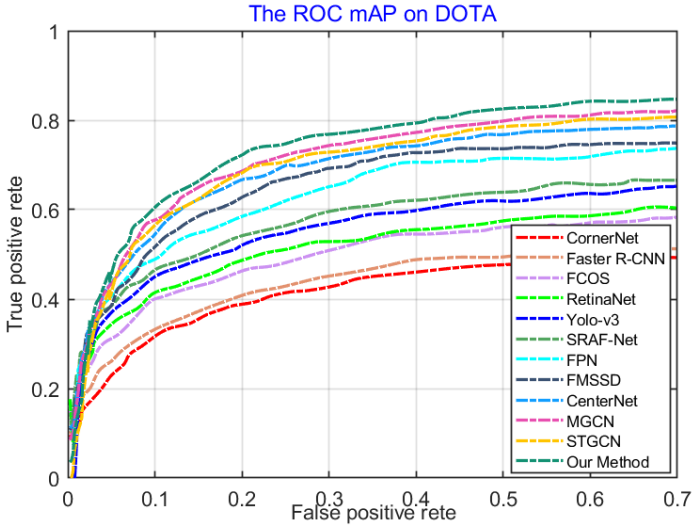

Figure 9. Comparisons of ROC curves using different detection methods on the DOTA dataset.

On the basis of a series of quantitative comparison results, our proposed HA-MHGEN demonstrated superiority in multiclass geospatial object detection. Figure 10 demonstrates visual object-detection results using the proposed HA-MHGEN framework on the DOTA dataset. Figure 10 shows that the proposed method allowed for successfully detecting most of the objects in all the unseen categories of objects in the DOTA dataset. There were also still some false detections and missing detections that are marked with red and blue dashed bounding boxes, respectively. For example, as shown in Figure 10d, the SP object category was leaked, and Figure 10h shows a false detection in which a rectangular rooftop was predicted as a large vehicle (LV). We will continue to address these deficiencies in future work.

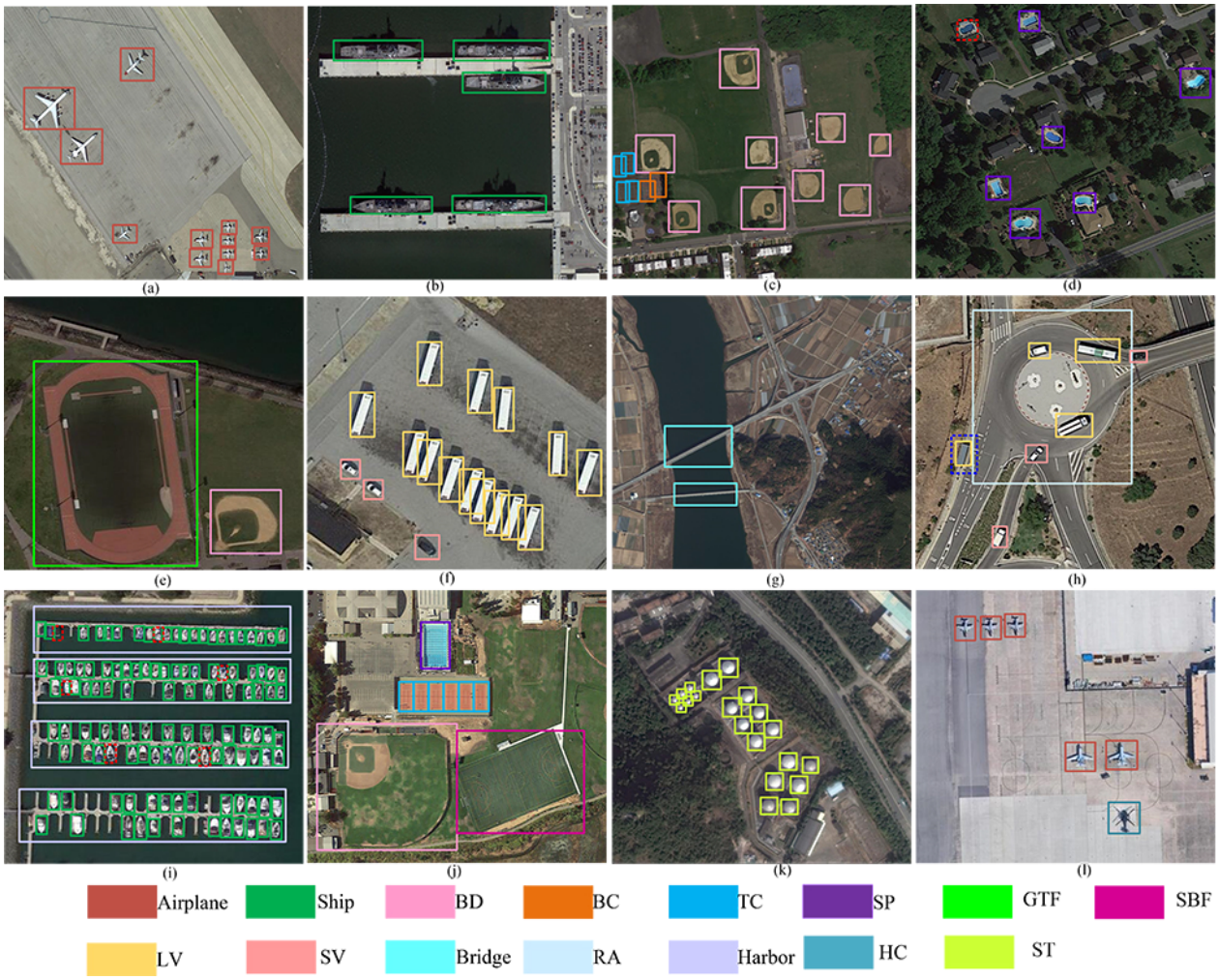

Figure 10. Visualization of multiclass object-detection results by using the proposed HA-MHGEN on the DOTA dataset; red and blue dashed lines denote missing and false detections, respectively. (**a**) Airplane. (**b**) Ship. (**c**) TC, BD, and BC. (**d**) SP. (**e**) BD and GTF. (**f**) LV and SV. (**g**) Bridge. (**h**) SV, LV, and RA. (**i**) Harbor and ship. (**j**) SBF, SP, BD, and TC. (**k**) ST. (**l**) Airplane.

4.4.2. Comparisons and Analysis on the DIOR Dataset

The DIOR dataset consists of more object categories and object instances than those of the DOTA dataset; thus, different intra- or interobject classifications have greater interference for appearance discriminants. Underlying spatial or contextual semantic relations between remote-sensing objects are also hard to decouple and distinguish. To further verify the effectiveness of our proposed HA-MHGEN, we adopted a more challenging optical remote-sensing object-detection dataset, DIOR, for experimental comparisons and analysis. Table 3 shows that our proposed HA-MHGEN achieved better object-detection performance than that of some classical two-stage approaches including faster R-CNN and the FPN, and boosted famous one-stage detectors such as Yolo-v3, FCOS, and FMSSD.

The table also shows that our HA-MHGEN achieved relatively high performance for some objects with diverse spatial distribution, such as airplanes with 88.93% mAP and ships with 86.14% mAP, and also on some large-scale objects with high appearance ambiguity with the complex background, such as bridges with 52.34% mAP, TCs with 89.53% mAP, BCs wth 92.08% mAP, and GCs with 86.78% mAP. In this situation, FMSSD leveraged contextual information in multiscale and the same feature maps, which enabled to guide the network to pay more attention to some objects with different multiscales, and then achieved relatively good object-detection performance. Nevertheless, this method only considers multiscale visually contextual information and ignores multiscale spatial

information between objects, which also results in some unsatisfactory detection results. In addition, compared with some recent anchor-free-based approaches, such as recent keypoint-based methods CornerNet, CenterNet, and SRAF-Net, there were improvements of 13.12% (74.72% versus 61.60%), 10.62% (74.72% versus 64.10%) and 5.02% (75.55% versus 69.70%) in terms of mAP, respectively.

Among the above methods, by introducing scene-relevant information into the feature discriminative process, SRAF-Net achieved relatively better detection performance. However, it also produced some unsatisfactory results for some objects whose appearances had more homogeneity with the background, such as GTF and ST. We also compared the proposed HA-MHGEN with graph-based methods MGCN and STGCN. As illustrated in Lines 10 to 12 of Table 3, the proposed HA-HMGEN achieved improvements of 1.15% (74.72% versus 73.57%) and 0.99% (74.72% versus 73.57%) compared with MGCN and STGCN, respectively. It is obvious that our method achieved comparable and even superior detection performance. The effectiveness and superiority of our method for object detection can be attributed to HA-MHGEN having self-attention-based spatial- and semantic-relation-driven hybrid attention modules that allow for learning powerful spatial and semantic feature representation, and hierarchical spatial and semantic relation discrimination enhancement to effectively boost detection performance for multiclass remote-sensing objects with high appearance ambiguity and complex spatial distribution. Figure 11 also shows the ROC curves of all the compared approaches for object detection, which also demonstrates the superiority of our proposed HA-MHGEN model.

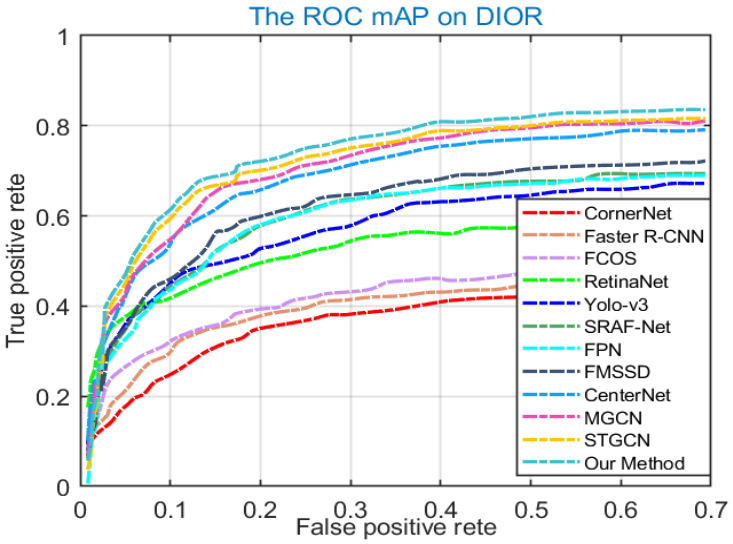

Figure 11. Comparisons of ROC curves using different detection methods on the DIOR dataset.

We further show visual detection on the challenging DIOR dataset to qualitatively demonstrate the effectiveness of the proposed HA-MHGEN. Figure 8 shows that most of the multiclass remote-sensing objects could be accurately and tightly localized by the predicted bounding boxes. Objects with high appearance ambiguity and adjacent locations were accurately detected, which further indicates the effectiveness of our proposed HA-MHGEN for remote-sensing object detection. In addition, there were some unsatisfactory detection results. For example, as shown in Figure 12e,h, the object category of VE was missing, and Figure 12n shows a false detection in which a road was predicted as bridge. In future work, we will continue to overcome these drawbacks.

Table 3. Comparisons with state-of-the-art object-detection approaches in terms of AP (%) and mAP (%) on the DIOR dataset.

Methods	Airplane	BF	Bridge	GTF	Ship	STM	TC	BC	ST	Harbor	Airport	ESA	Chimney	Dam	VE	GC	TS	OP	ETS	WM	mAP
Faster R-CNN	51.40	62.20	27.00	61.80	56.10	41.80	73.90	80.70	39.60	43.70	61.60	53.40	74.20	37.30	34.30	69.60	44.70	49.00	45.10	65.30	53.60
Yolo-v3	67.50	65.80	34.20	68.90	86.80	40.30	83.90	86.80	67.80	54.30	54.70	55.70	73.50	34.30	49.10	67.30	32.30	51.70	49.60	73.60	59.90
FCOS	73.60	84.30	32.10	17.10	51.10	71.40	77.40	46.70	63.10	73.20	62.00	76.60	52.40	39.70	71.90	80.80	37.20	46.10	58.40	82.70	60.00
Center -Net	64.00	65.70	34.80	66.00	81.30	53.50	80.90	86.30	63.70	45.30	66.30	60.80	73.10	41.10	46.30	73.00	44.10	53.30	54.20	78.80	61.60
Retina -Net	63.40	83.30	48.20	59.10	72.00	82.40	90.10	78.40	80.70	47.60	47.80	53.20	67.90	49.40	47.70	66.30	55.00	45.70	73.60	92.00	63.40
Corner -Net	68.50	85.20	46.90	16.80	34.50	89.10	84.70	78.40	40.00	68.60	77.10	73.90	76.90	60.20	45.00	79.10	52.30	58.90	74.80	70.10	64.10
FPN	54.00	63.30	44.80	76.80	71.80	68.30	81.10	80.70	53.80	46.40	74.50	76.50	72.50	60.00	43.10	76.00	59.50	57.20	62.30	81.20	65.10
SRAF -Net	88.40	92.60	83.80	16.20	59.40	80.90	87.90	90.60	55.60	76.40	76.50	86.80	83.80	58.60	53.20	82.80	90.60	58.00	66.80	91.00	69.70
FMSSD	85.60	75.80	40.70	78.60	84.90	76.70	87.90	89.50	65.30	62.00	82.40	67.10	77.60	64.70	44.50	80.80	62.40	58.00	61.70	76.30	71.10
MGCN	87.19	73.97	52.34	80.13	86.14	79.15	89.24	91.09	68.13	68.65	85.92	72.09	79.23	66.16	47.33	83.16	66.19	53.89	67.45	73.98	73.57
STGCN	88.13	72.54	49.76	85.32	86.76	77.91	88.15	90.12	70.16	69.73	83.17	74.11	79.57	67.98	44.54	85.90	67.23	54.17	68.12	71.25	73.73
Our Method	88.93	77.13	52.34	81.51	87.24	78.09	89.53	92.08	72.23	71.42	85.17	74.17	75.32	71.23	46.87	86.78	69.18	52.46	71.86	70.98	74.72

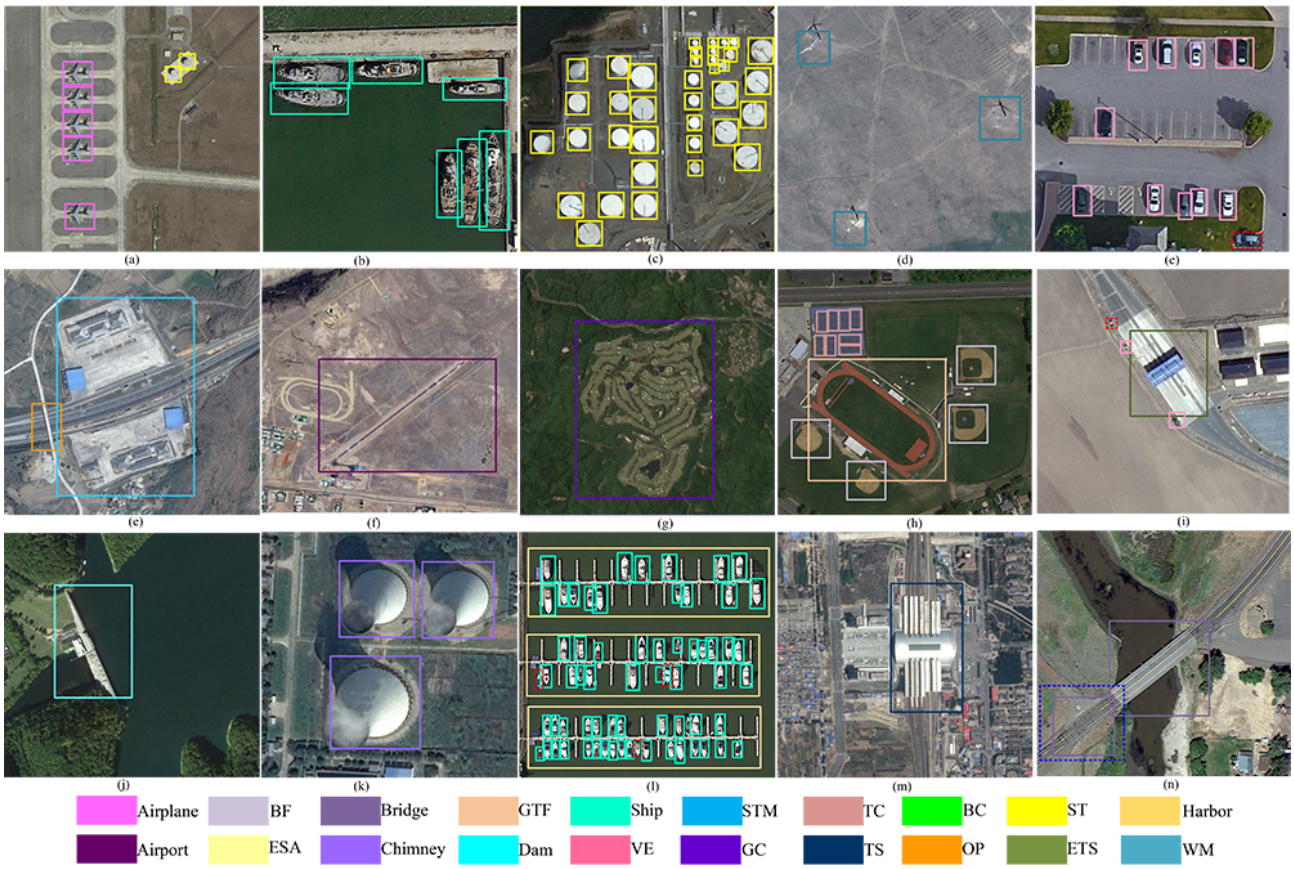

Figure 12. Visualization of multiclass object-detection results by using the proposed HA-MHGEN on the DIOR dataset; red and blue dashed lines denote missing and false detections, respectively. (**a**) Airplane and ST. (**b**) Ship. (**c**) ST. (**d**) WM. (**e**) VE. (**f**) Airport. (**g**) GC. (**h**) BF, TC and GTF. (**i**) VE and ETS. (**j**) Dam. (**k**) Chimney. (**l**) Harbor. (**m**) TS. (**n**) Bridge.

4.4.3. Comparisons and Analysis on the NWPU VHR-10 Dataset

In order to verify the generalization of the proposed HA-MHGEN, we also constructed another group of comparative experiments on the NWPU-VHR-10 dataset. The series of quantitative results that were produced by our method and the compared methods are reported in Table 4. Among all methods, the proposed HA-MHGEN was the only detection approach that surpassed 87%, and it obtained the best AP value for some large-scale objects, such as bridges (98.68%) and harbors (97.65%). Our proposed spatial and semantic hybrid attention could thus effectively focus on large-scale objects and suppress interference information of complex backgrounds.

Although Yolo-v3, RetinaNet, FMSSD, and FCOS utilize a multiscale regression layer and multiscale anchor-based initialization, our method also achieved better detection accuracy. Specifically, the overall detection accuracy of HA-MHGEN in terms of mAP was 22.07% higher than that of Yolo-v3, 15.26% higher than that of RetinaNet, 2.99% higher than that of FMSSD, and 1.25% than that of FCOS. The main reason is that it is hard for these one-stage methods to address the problem of anchor initialization with diverse visual and spatial scale changes, and then they always results in some leak detection cases for densely spatial arranged remote-sensing objects. In addition, compared with two-stage approaches faster R-CNN and FPN, our method achieved improvements of 12.45% (93.39% versus 80.94%) and 3.31% (93.39% versus 90.08%) in terms of mAP, respectively. Furthermore, the overall performance of our method was 4.89% higher than that of STGCN and 2.31% more than that of MGCN. Figure 13 shows a comparison of ROC curves by using the object-detection approaches, also indicating the superiority of our proposed HA-MHGEN.

Table 4. Comparisons with state-of-the-art object-detection approaches in terms of AP (%) and mAP (%) on the NWPU VHR-10 dataset.

Methods	Airplane	Ship	BD	BC	TC	ST	Vehicle	Bridge	Harbor	GTF	mAP
CornerNet	72.10	53.67	44.13	83.16	67.79	56.51	77.25	92.67	49.17	99.96	69.64
CenterNet	73.09	71.97	87.41	73.37	65.12	59.91	55.75	53.76	75.48	95.27	71.11
Yolo-v3	92.50	62.90	59.48	47.99	64.08	56.12	72.59	59.48	70.06	92.17	71.32
RetinaNet	99.56	78.16	99.55	65.18	83.37	82.29	71.91	40.25	65.66	95.38	78.13
Faster R-CNN	97.83	78.66	89.99	58.80	80.85	90.68	73.09	63.33	80.68	95.47	80.94
SRAF-Net	94.59	83.80	53.99	92.38	88.39	72.84	89.21	96.95	63.53	98.95	83.45
STGCN	95.76	94.82	93.45	86.92	85.83	95.03	87.39	73.41	84.86	87.62	88.50
FMSSD	99.70	89.90	98.20	96.80	86.00	90.30	88.20	80.10	75.60	99.60	90.40
FPN	100	90.86	96.84	95.05	90.67	99.99	90.19	50.86	93.67	100	90.80
MGCN	98.36	92.15	99.16	97.24	86.87	91.02	89.86	81.34	77.19	97.67	91.08
FCOS	99.99	85.21	97.75	80.34	95.80	96.94	88.92	88.92	95.04	99.67	92.14
Our Method	97.19	88.86	98.68	83.09	94.17	98.97	90.54	87.64	97.65	97.13	93.39

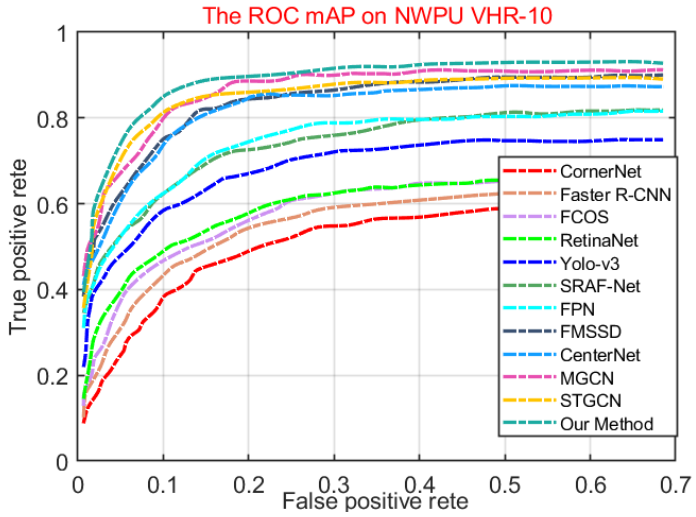

Figure 13. Comparisons of ROC curves using different detection methods on the NWPU VHR-10 dataset.

The superior detection performance by the proposed HA-MHGEN regarding mAP was mainly assisted by considering the intramodality relations using the cross-attention module, which can help the network in learning the association between spatial and semantic relations. HA-MHGEN can thereby complement spatial and semantic relation feature extraction, and further enhances the discrimination ability of fused features to boost multiclass object-detection performance. To indicate the validity of the proposed HA-HMGEN, a series of visual multiclass geospatial object-detection results are shown in Figure 14, which further demonstrate the generalization and effectiveness of our proposed method for multiclass remote-sensing object detection. However, there were also some unsatisfactory detection results. For example, as shown in Figure 14e,h, the object categories of BCs and vehicles were missing, and Figure 14n shows a false detection in which a road was also predicted as bridge. In addition, for some objects with high appearance homogeneity with the background, such as airplanes and ships, HA-MHGEN achieved worse detection performance than that of MGCN, as listed in Table 4. The main reason is that MGCN adopts spatial contextual information in superpixel segmentation for graph construction, which takes into account more complete object information and their background information for relation discrimination. In future work, we will continue to amend these drawbacks to further improve the detection performance of our method.

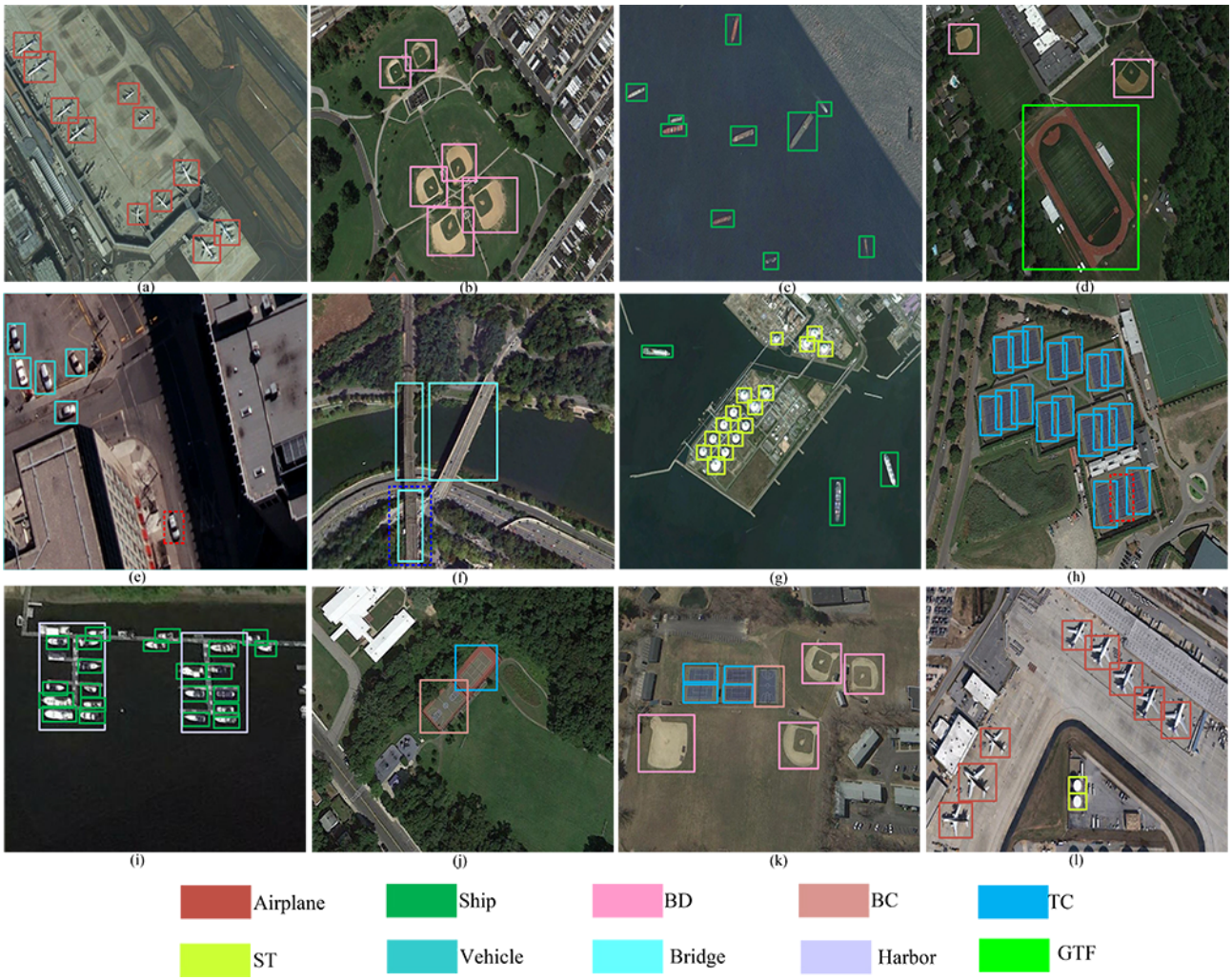

Figure 14. Visualization of multiclass object-detection results using the proposed HA-MHGEN on the NWPU VHR-10 dataset; red and blue dashed lines denote missing and false detections, respectively. (**a**) Airplane. (**b**) BD. (**c**) Ship. (**d**) GTF. (**e**) Vehicle. (**f**) Bridge. (**h**) TC. (**i**) Harbor. (**j**) TC and BC. (**k**) BD, BC, and TC. (**l**) Airplane and ST.

4.4.4. Computational Cost Analysis

Aside from the detection accuracy comparison and analysis between the proposed HA-MHGEN and other famous detection methods, the computational speed and complexity of detectors are also important. Therefore, to further illustrate the effectiveness of the proposed HA-MHGEN in practical cases, we also compare the computational speed and complexity of our method with that of 11 other famous detection methods with different backbones. Specifically, we adopted the mean inference time for computational speed comparison and analysis, and parameters and giga floating-point operations (GFLOPs) were utilized for complexity analysis. For a fair comparison, we randomly selected 10 remote-sensing images from each dataset and resized them into 512 × 512 pixels for experimental verification. Table 5 reports the series of comparisons regarding mean inference time, parameters, and GFLOPs: CornetNet needed the most inference time and computational consumption. Faster R-CNN and FPN needed almost the same inference time, but FPN had more parameters and GFLOPs for calculation, and then improved complexity. FCOS, MGCN, and STGCN almost had the same inference time, and the parameters and GFLOPs of MGCN and STGCN were higher than those of FCOS, which indicates that MGCN and STGCN had high computation complexity. CenterNet, SRAF-Net, and FMSSD were almost the same in terms of inference time, and SARF-Net and FMSSD needed a higher computational cost. YOLO-v3 and the proposed HA-MHGEN had similar inference times, and

HA-MHGEN had fewer parameters and GFLOPs than those of YOLO-v3, which indicates the relatively low computational complexity of HA-MHGEN. These comparisons show that HA-MHGEN both achieved the best performance of object detection in the datasets, and had acceptable computational speed and complexity.

Table 5. Computational cost comparison and analysis of different approaches.

Methods	Backbone	map@DOTA	map@DIOR	map@NWPU VHR-10	Params (M)	GFLOPs	Inference Times (ms)
Faster R-CNN	ResNet-101	56.17	55.24	82.37	60.7	81.6	67
Yolo-v3	DarkNet-53	66.92	58.97	72.16	60.04	82.4	28
FCOS	ResNet-101	61.73	61.37	92.39	51.2	70.65	55
CenterNet	ResNet-101	74.08	62.51	72.07	52.7	75.2	44
RetinaNet	ResNet-101	64.19	63.27	79.6	56.9	81.3	91
CornerNet	Hourglass-54	55.82	64.59	71.84	112.7	287.6	127
FPN	ResNet-101	72.43	66.18	91.34	50.7	112.3	69
SRAF-Net	ResNet-101	66.37	70.39	84.55	62.9	87.2	46
FMSSD	VGG-16	74.76	72.37	91.17	61.3	84.2	42
MGCN	ResNet-101	77.92	72.94	91.39	62.4	87.2	52
STGCN	ResNet-101	76.18	73.15	92.07	64.6	90.1	56
Our Method	ResNet-101	78.79	74.96	93.27	51.4	67.9	33

5. Conclusions

In this paper, we presented a novel hybrid attention-driven multistream hierarchical graph embedding network (HA-MHGEN) for multiclass geospatial object detection in high-spatial-resolution remote-sensing images (HSRIs). First, to explore intra- and interobject hierarchical spatial relations and contextual semantic relations, a self-attention-aware multiscale GCN model was designed to more effectively extract spatial and semantic relations features, respectively. Second, we proposed a cross-attention-driven intramodality relation fusion model that takes advantage of the multihead attention mechanism to learn associated joint feature representations between diverse spatial and complex semantic correlations, and adopted the cross-attention mechanism to further improve the discrimination ability of the fused features. Lastly, the derived spatial-relation, semantic-relation, and their intramodality relation features were embedded into the uniform learning framework to improve detection performance. Comprehensive experiments were conducted on three benchmark datasets, DOTA, DIOR, and NWPU VHR-10, and the effectiveness and superiority of the proposed HA-HMGEN was demonstrated both quantitatively and qualitatively. In the future, we will pay more attention to novel advantage feature fusion models and further enhance the performance of our proposed detection framework.

Author Contributions: S.T., L.C. and L.K. conceived the research; S.T. and L.C. wrote the code, performed the analysis, and wrote the article; X.X., J.T. and K.D. analyzed the results; K.S., C.F. and Y.F. collected the dataset; and Y.Z. revised the manuscript. All authors have read and agreed to published version of the manuscript.

Funding: this work was supported by the National Natural Science Foundation of China under grant 62001032, the National Science Foundation of China under grant U20A20163, and the Scientific Research Project Beijing Municipal Education Commission under grants KZ202111232049 and KM202011232021.

Institutional Review Board Statement: not applicable.

Informed Consent Statement: not applicable.

Data Availability Statement: not applicable.

Conflicts of Interest: the authors declare no conflict of interest.

References

1. Wang, Y.; Li, Y.; Chen, W.; Li, Y.; Dang, B. DNAS: Decoupling Neural Architecture Search for High-Resolution Remote Sensing Image Semantic Segmentation. *Remote Sens.* **2022**, *14*, 3864. [CrossRef]
2. Ji, X.; Huang, L.; Tang, B.-H.; Chen, G.; Cheng, F. A Superpixel Spatial Intuitionistic Fuzzy C-Means Clustering Algorithm for Unsupervised Classification of High Spatial Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3490. [CrossRef]
3. Cheng, F.; Fu, Z.; Tang, B.; Huang, L.; Huang, K.; Ji, X. STF-EGFA: A Remote Sensing Spatiotemporal Fusion Network with Edge-Guided Feature Attention. *Remote Sens.* **2022**, *14*, 3057. [CrossRef]
4. Qin, H.; Li, Y.; Lei, J.; Xie, W.; Wang, Z. A specially optimized one-stage network for object detection in remote sensing images. *IEEE Geosci. Remote. Sens. Lett.* **2021**, *18*, 401–405. [CrossRef]
5. Ma, W.; Guo, Q.; Wu, Y.; Zhao, W.; Zhan, X.; Ji, L. A novel multi-model decision fusion network for object detection in remote sensing images. *Remote Sens.* **2019**, *11*, 737. [CrossRef]
6. Qin, H.; Li, Y.; Lei, J.; Xie, W.; Wang, Z. Cross-scale feature fusion for object detection in optical remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 431–435.
7. Ren, S.; He, K.; Gishick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
8. Cheng, G.; Yan, B.; Shi, P.; Li, K.; Yao, X.; Guo, L.; Han, J. Prototype-CNN for few-shot object detection in remote sensing images. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *60*, 1–10. [CrossRef]
9. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
10. Li, Q.; Wang, G.; Liu, J.; Chen, S. Robust scale-invariant feature matching for remote sensing image registration. *IEEE Geosci. Remote Sens. Lett.* **2009**, *6*, 287–291.
11. Sirmacek, B.; Unsalan, C. Urban-area and building detection using SIFT keypoints and graph theory. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1156–1167. [CrossRef]
12. Tao, C.; Tan, Y.; Cai, H.; Tian, J. Airport detection from large IKONOS images using clustered SIFT keypoints and region information. *IEEE Geosci. Remote Sens. Lett.* **2010**, *8*, 128–132. [CrossRef]
13. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
14. Xiao, Z.; Liu, Q.; Tang, G.; Zhai, X. Elliptic Fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. *Int. J. Remote Sens.* **2015**, *36*, 618–644. [CrossRef]
15. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
16. Zhang, Y.; Yuan, Y.; Feng, Y.; Lu, X. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5535–5548. [CrossRef]
17. Meynberg, O.; Cui, S.; Reinartz, P. Detection of high-density crowds in aerial images using texture classification. *Remote Sens.* **2016**, *8*, 470. [CrossRef]
18. Sun, H.; Sun, X.; Wang, H.; Li, Y.; Li, X. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geosci. Remote Sens. Lett.* **2011**, *9*, 109–113. [CrossRef]
19. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
20. Liu, J.; Yang, D.; Hu, F. Multiscale Object Detection in Remote Sensing Images Combined with Multi-Receptive-Field Features and Relation-Connected Attention. *Remote Sens.* **2022**, *14*, 427. [CrossRef]
21. Zhang, K.; Shen, H. Multi-Stage Feature Enhancement Pyramid Network for Detecting Objects in Optical Remote Sensing Images. *Remote Sens.* **2022**, *14*, 579. [CrossRef]
22. Han, X.; Zhou, Y.; Zhang, L. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
23. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Trans. Image Process.* **2018**, *28*, 265–278. [CrossRef] [PubMed]
24. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 2337–2348.
25. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
26. Chen, Z.; Zhang, T.; Ouyang, C. End-to-end airplane detection using transfer learning in remote sensing images. *Remote Sens.* **2018**, *10*, 139. [CrossRef]
27. Wang, G.; Zhuang, Y.; Chen, H.; Liu, X.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FSoD-Net: Full-scale object detection from optical remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–18. [CrossRef]
28. Zhang, K.; Wu, Y.; Wang, J.; Wang, Y.; Wang, Q. Semantic Context-Aware Network for Multiscale Object Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
29. Zhang, K.; Wu, Y.; Wang, J.; Wang, Y.; Wang, Q. Few-shot object detection via feature reweighting. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8420–8429.

30. Wang, P.; Sun, X.; Diao, W.; Fu, K.; Zhang, T.; Li, L.; Dong, S.; Sang, Q. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 3377–3390. [CrossRef]
31. Chen, S.; Dai, B.; Tang, J.; Luo, B.; Wang, W.; Lv, K.; Dong, S.; Sang, Q. A refined single-stage detector with feature enhancement and alignment for oriented object. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8898–8908. [CrossRef]
32. Li, X.; Drng, J.; Fang, Y. Enhanced TabNet: Attentive Interpretable Tabular Learning for Hyperspectral Image Classification. *Remote Sens.* **2022**, *14*, 716.
33. Pan, F.; Wu, Z.; Liu, Q.; Xu, Y.; Wei, Z.DCFF-Net: A Densely Connected Feature Fusion Network for Change Detection in High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11974–11985. [CrossRef]
34. Li, X.; Deng, J.; Fang, Y. Few-shot object detection on remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
35. Chen, J.; Wan, J.; Zhu, G.; Xu, G.; Deng, M. Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **2019**, *17*, 681–685. [CrossRef]
36. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 8232–8241.
37. Wang, C.; Bai, X.; Wang, S.; Zhou, J.; Ren, P. Multiscale visual attention networks for object detection in VHR remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2018**, *16*, 310–314. [CrossRef]
38. Lu, X.; Ji, J.; Xing, Z.; Miao, Q. Attention and feature fusion SSD for remote sensing object detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *70*, 1–9. [CrossRef]
39. Yang, L.; Zhan, X.; Chen, D.; Yan, J.; Lov, C.; Lin, D. Learning to cluster faces on an affinity graph. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2298–2306.
40. Yang, L.; Zhan, X.; Chen, D.; Yan, J.; Lov, C.; Lin, D. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
41. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 12026–12035.
42. He, C.; Lai, S.; Lam, K. Improving object detection with relation graph inference. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2537–2541.
43. Chaudhuri, U.; Banerjee, B.; Bhattacharya, A. Siamese graph convolutional network for content based remote sensing image retrieval. *Comput. Vis. Image Underst.* **2019**, *184*, 22–30. [CrossRef]
44. Khan, N.; Chaudhuri, U.; Banerjee, B.; Chaudhuri, S. Graph convolutional network for multi-label VHR remote sensing scene recognition. *Neurocomputing* **2019**, *357*, 36–46. [CrossRef]
45. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.-S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*.
46. Hammond, D.K.; Vandergheynst, P.; Gribonval, R. Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.***2011**, *30*, 129–150. [CrossRef]
47. Kopf, T.N.; Welling, X. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.-N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
49. Xiao, L.; Wu, X.; Wu, W.; Yang, J.; He, L. Multi-Channel Attentive Graph Convolutional Network with Sentiment Fusion for Multimodal Sentiment Analysis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 4578–4582.
50. Girshick, R. Fast r-cnn. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
51. Hsieh, T.-I.; Lo, Y.-C.; Chen, H.-T.; Liu, J.T.-L. One-shot object detection with co-attention and co-excitation. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
52. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp.3974–3983.
53. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
54. Dong, Z.; Wang, M.; Wang, Y.; Zhu, Y.; Zhang, Z. Object detection in high resolution remote sensing imagery based on convolutional neural networks with suitable object scale features. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 2104–2114. [CrossRef]
55. He, C.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
56. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
57. Lin, C.-Y.; Piotr, D.; Ross, G.; He, K.; Bharah, H.; Serge, B. Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

58. Remon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2017**, arXiv:1804.02767.
59. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Piotr, D. Focal Loss for Dense Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2980–2988.
60. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9627–9636.
61. Liu, J.; Li, S.; Zhou, C.; Cao, X.; Gao, Y.; Wang, B. SRAF-Net: A Scene-Relevant Anchor-Free Object Detection Network in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
62. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
63. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
64. Jiang, B.; Jiang, X.; Tang, J.; Luo, B.; Huang, S. Multiple graph convolutional networks for co-saliency detection. In Proceedings of the International Conference on Multimedia and Expo (ICME), Shanghai, China, 8-12 July 2019; pp. 332–337.
65. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018.