Relational Part-Aware Learning for Complex Composite Object Detection in High-Resolution Remote Sensing Images

Shuai Yuan, Lixian Zhang, Runmin Dong, Jie Xiong, Juepeng Zheng, *Member, IEEE*, Haohuan Fu, *Senior Member, IEEE*, and Peng Gong

***Abstract*—In high-resolution remote sensing images (RSIs), complex composite object detection (e.g., coal-fired power plant detection and harbor detection) is challenging due to multiple discrete parts with variable layouts leading to complex weak inter-relationship and blurred boundaries, instead of a clearly defined single object. To address this issue, this article proposes an end-to-end framework, i.e., relational part-aware network (REPAN), to explore the semantic correlation and extract discriminative features among multiple parts. Specifically, we first design a part region proposal network (P-RPN) to locate discriminative yet subtle regions. With butterfly units (BFUs) embedded, feature-scale confusion problems stemming from aliasing effects can be largely alleviated. Second, a feature relation Transformer (FRT) plumbs the depths of the spatial relationships by part-and-global joint learning, exploring correlations between various parts to enhance significant part representation. Finally, a contextual detector (CD) classifies and detects parts and the whole composite object through multirelation-aware features, where part information guides to locate the whole object. We collect three remote sensing object detection datasets with four categories to evaluate our method. Consistently surpassing the performance of state-of-the-art methods, the results of extensive experiments underscore the effectiveness and superiority of our proposed method.**

***Index Terms*—Complex composite object detection, high-resolution remote sensing images (RSIs), inter-relationship, Transformer.**

I. Introduction

OBJECT detection in remote sensing areas, one of the most interesting yet formidable issues, laying the groundwork for interpreting and understanding remote sensing images (RSIs) [1]. Owing to the achievements in high-resolution RSIs datasets and deep learning algorithms, tremendous progress in the accuracy and efficiency of object detection in remote sensing has been witnessed [1], [2], [3], [4], [5], [6], [7], [8]. However, most of the existing algorithms are designed for clearly defined single-object detection like vehicle detection [9], [10], yet overlooking many complex composite objects in optical RSIs (e.g., coal-fired power plant and airport) which we should think of as a whole. These complex composite objects provide essential support for society (e.g., power plants for electricity generation and airports for transportation), so monitoring them in RSIs is equally important. With a target to identify these combined complexes with multiple parts and nonrigid layouts, and the difficulties arising from the complicated background and blurred boundaries, it is a challenging research problem.

Compared with single-object detection, complex composite object detection in RSIs is difficult for two reasons and Fig. 1 shows the comparison between complex composite objects and single objects. First, these objects are characterized by intricate parts with various layouts. For example, a coal-fired power plant contains chimneys and condensing towers, and such complex detection target involves problems including complex spatial relationships between parts and nonrigid boundaries. Nonrigid boundaries can enlarge the sizes of bounding boxes and decrease the precision. The complex composite manner indicates the parts are discrete, and other textures between parts make the composite spatial relationships weak and disturbed, leading to difficulties in detecting a composite object as a whole precisely. Second, complex composite objects are frequently situated amidst surroundings with similar textures, further complicating detection. For instance, coal-fired power

Manuscript received 3 February 2024; revised 7 April 2024; accepted 15 April 2024. This work was supported by the National Natural Science Foundation of China under Grant T2125006. This article was recommended by Associate Editor L. Jiao. *(Corresponding authors: Juepeng Zheng; Haohuan Fu.)*

Shuai Yuan and Peng Gong are with the Department of Geography, The University of Hong Kong, Hong Kong, China (e-mail: shuai914@connect.hku.hk; penggong@hku.hk).

Lixian Zhang is with the High Performance Computing Department, National Supercomputing Center in Shenzhen, Shenzhen 518055, China (e-mail: zhanglx18@tsinghua.org.cn).

Runmin Dong is with the Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, and the Xi'an Institute of Surveying and Mapping Joint Research Center for Next-Generation Smart Mapping, Tsinghua University, Beijing 100190, China (e-mail: drm@mail.tsinghua.edu.cn).

Jie Xiong is with the Department of Strategy, Entrepreneurship and International Business, ESSCA School of Management, 49000 Angers, France (e-mail: jie.xiong@essca.fr).

Juepeng Zheng is with the School of Artificial Intelligence, Sun Yat-sen University (Zhuhai), Zhuhai 510275, China (e-mail: zhengjp8@mail.sysu.edu.cn).

Haohuan Fu is with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, and also with the Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, and the Xi'an Institute of Surveying and Mapping Joint Research Center for Next-Generation Smart Mapping, Tsinghua University, Beijing 100190, China (e-mail: haohuan@tsinghua.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2024.3392474.

Digital Object Identifier 10.1109/TCYB.2024.3392474

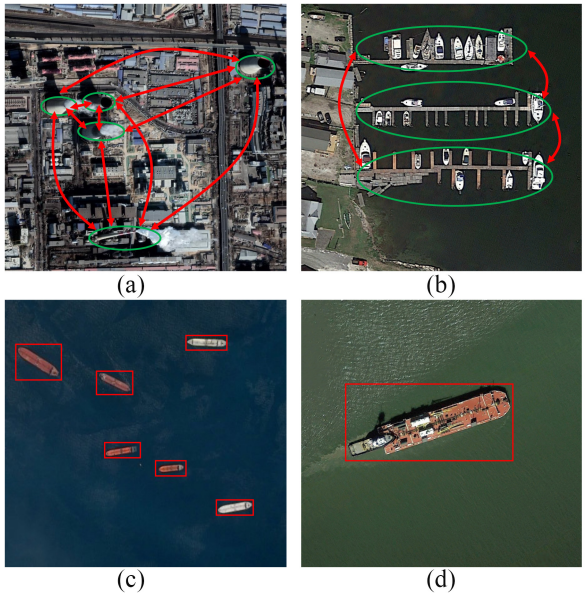


Fig. 1. Differences between complex composite objects (i.e., (a) coal-fired power plants and (b) harbors) and single objects (i.e., (c) and (d) ships) in optical RSIs. The green ovals in (a) and (b) denote the distinct parts of coal-fired power plants and harbors. The red lines in (a) and (b) represent the complex inter-relationships between parts in composite objects. Compared with composite objects, single objects have clear-defined boundaries and simpler backgrounds around.

plants are often located in industrial areas where other similar industrial infrastructures may hamper coal-fired power plant detection performance. Similar surroundings contribute to the blurred boundaries and puzzle the bounding box localization. Unlike single objects, such as cars or ships, which own a unified structure and a unified semantic meaning without significant internal complexity, composite objects, such as a coal-fired power plant, are characterized by a more complex structure composed of multiple semantic meanings with internal complexity. As the red lines and green ovals in Fig. 1(a) and (b), complex yet weak spatial inter-relationships and blurred boundaries caused by multiple components with various layouts make composite objects harder to detect than single objects.

Nevertheless, commonly used CNN-based object detection methods rely on feature extraction from local regions and use these features to generate bounding boxes. For composite object detection, these algorithms may fail to handle the semantic gap between low-level features and high-level understanding of objects caused by complex and diverse spatial inter-relationships between parts [11], [12]. Additionally, the highly variable appearance of parts makes it difficult to generalize across different instances of the same object [6]. Consequently, the direct application of existing algorithms to composite object detection is ill-advised, and part-based methods are better for discovering discriminative and subtle components.

Part-based methods are used in fine-grained visual classification tasks [13], [14], [15], aiming to generate rich feature representations [16], [17] or localize parts for feature enhancement [14], [18], [19]. By modeling a complex structure as an assemblage of distinct parts that can be localized and recognized individually, part-based methods offer heightened efficacy for composite object detection. Recently, a few efforts [6], [20], [21] have been made on part-based methods for composite object detection in RSIs. For example, Sun et al. [6] proposed a unified part-based CNN-based network consisting of a part localization module and a context refinement module to localize the most representative part features. Although previous work has reached promising results, the attention to constraints on local feature learning and simple concatenation of part features lead to the regardless of discriminative parts and the potential in long-range spatial inter-relationships. We argue that investigating the potential correlation between parts and constructing a global semantic understanding of objects can significantly benefit composite object detection in RSIs.

To this end, we propose a relational part-aware network (REPAN) to explore the inter-relationships and extract discriminative features among multiple parts. The effectiveness of REPAN is based on three main modules, i.e., a part region proposal network (P-RPN) to discover discriminative regions, and a feature relation Transformer (FRT) to construct the correlation, and a contextual detector (CD) to detect parts and the whole composite object. During the part region proposal, multiscale features are fused by butterfly units (BFUs), reducing aliasing and confusion between different scale layers. The peak responses in multiscale-aware features are then selected to generate part proposal regions. After that, FRT discovers global relationships and inter-relationships by part-and-global joint learning with multiattention heads. To fully utilize the contextual correlation, CD is proposed to combine part and global features for final classification and detection.

In summary, our main contribution lies in the following.

1) We introduce the REPAN, a comprehensive solution tailored for the complex composite object detection RSIs, which is conceived as an end-to-end model, establishing both local and global correlations through the integration of the weakly supervised part proposal and a novel Transformer-based part-and-global joint learning strategy.
2) To enhance the precision and robustness of discriminative part feature discovery, we devise a novel P-RPN featuring BFUs. The incorporation of BFUs mitigates feature confusion, enabling the network to identify distinctive part features with greater accuracy.
3) We introduce an FRT to facilitate the learning of relationships and foster high-level correlations between global and part-specific features. This module empowers the network to develop a holistic understanding of objects, transcending the limitations of focusing solely on single parts.

The remainder of this article is organized as follows. First, we briefly introduce the related work in Section II. Then, we elaborate our proposed method REPAN in Section III. We illustrate experiment details, and comparative studies in Section IV, and conduct ablation studies and discussion in Section V. Finally, the conclusion of this article and future work is presented in Section VI.

II. Related Work

A. Composite Object Detection in RSIs

Composite objects hold equal significance alongside single objects in terms of both natural and social attributes [6]. However, the lack of targeted methods for detecting composite objects in RSIs poses challenges in effectively monitoring them, thereby significantly impacting local development, urban planning, and ecological preservation. This gap has recently garnered attention, leading to the emergence of specialized techniques aimed at detecting composite objects in RSI. Works [22], [23], [24], [25], [26], [27], [28] utilized traditional CNN-based methods to detect composite objects, such as airports, schools, etc. For example, Cheng et al. [23] proposed an anchor-free-oriented proposal generator to detect oriented coal-fired power plants in the DIOR-R dataset. Yao et al. [25] proposed a two-stage network to detect airports and expressway service centers. Cheng et al. [29] designed a two-stage method to address the misalignments of spatial and feature in oriented airport detection. Cheng et al. [30] proposed a spatial and channel Transformer to capture the deep correlations for oriented coal-fired power plant detection and so on. Cai et al. [31] used hard example learning and a weight-balanced strategy in airport detection to improve performance within an overwhelming number of easy examples and a few hard examples. Fu et al. [32] proposed a CNN-based one-stage detector with a feature-enhanced module to detect schools. However, traditional CNN-based methods cannot handle the complex and various component layouts and distribution and the detection performance is limited. Works [6], [20], [21] then developed part-based methods to generate rich feature representation and get discriminative part features. For instance, Yin et al. [20] designed a multiattention part-based network for coal-fired power plant detection. The context attention module and part-based attention module strengthen the component features. Qian et al. [21] developed a part-based topology distillation network for composite object detection in RSIs by locating the discriminative part features. However, the aforementioned works all focused on local feature representation [green ovals in Fig. 1(a) and (b)], regardless of the semantic relationships between local features [red lines in Fig. 1(a) and (b)]. Thus, in this article, we aim to not only discover discriminative part information but also reveal the spatial relationships in composite objects via our proposed method.

B. Part-Based Methods

Compared with standard object detection methods (i.e., SSD [33], Faster R-CNN [34], Cascade R-CNN [35], Libra R-CNN [36], etc.), part-based methods have emerged as a powerful approach in tackling the challenges of fine-grained visual classification tasks by dissecting objects into distinct components or parts, focusing on local regions that contain discriminative cues. For example, Fu et al. [26] developed the RA-CNN, which used a recurrent learning approach to locate the most discriminative region in an image and improve classification accuracy. Zheng et al. [37] proposed MA-CNN where part generation and feature learning can reinforce each other via multiattention. Zheng et al. [38] designed PA-CNN to realize fine-grained object detection via progressive-attention learning step by step. Ji et al. [39] proposed ACNet, which contained an attention convolutional binary neural tree architecture for weakly supervised fine-grained classification. Ding et al. [40] proposed AP-CNN which enhanced the performance by learning both high-level semantic and low-level detailed feature representation. Part-based methods are appropriate for composite object detection in remote sensing because composite objects in RSIs usually consist of multiple parts with different textures, shapes, and scales. CNN-based detectors that treat the object as a whole may not capture the subtle but crucial differences between the parts of the object. Instead of holistic understanding, part-based methods are capable of localizing and extracting important part regions of the object in the presence of background interference. However, existing work on composite object detection in remote sensing which utilizes part-based methods all focus on local feature representation, regardless of the potential relationships between local features, which we think can develop a semantic understanding of the detector. Therefore, in this work, we propose to detect composite objects via two steps in one unified framework, i.e., discovering the discriminative parts and exploring the potential spatial relationships.

C. Transformers in Vision

Transformers are a type of neural network architecture that has gained significant popularity in natural language processing (NLP) tasks due to their strong capability to construct long-range dependencies in text [41], [42]. They have recently been introduced to computer vision tasks and achieved promising results [43], [44], [45]. Unlike traditional CNN, which relies on local spatial relationships, Transformers capture global contextual information by attending to all image regions simultaneously. For example, the very first work, ViT [43], utilized a fully Transformer architecture for image classification. The multihead self-attention mechanism helped the network understand global-range relationships. However, the strong ability also brings high computation costs. Works [44], [46] attempted to combine CNN and Transformers, exploiting the advantages from both sides: reduced computation costs, local feature understanding and global-range understanding. For example, Liu et al. [44] designed shifted windows to reduce the computation complexity in high-resolution images. Due to the holistic relationship understanding ability, in this article, we introduce an FRT to encourage global-and-part joint learning for exploring relationships between the globe and parts in composite objects. With the combination of CNN and Transformers, the network can learn local feature representation and build holistic semantic correlation in the meantime.

III. Approach

A. Preliminary and Overview

Given an input image $\mathcal{I}$ and the respective ground-truth label y, our ultimate goal is to find a suitable detection function

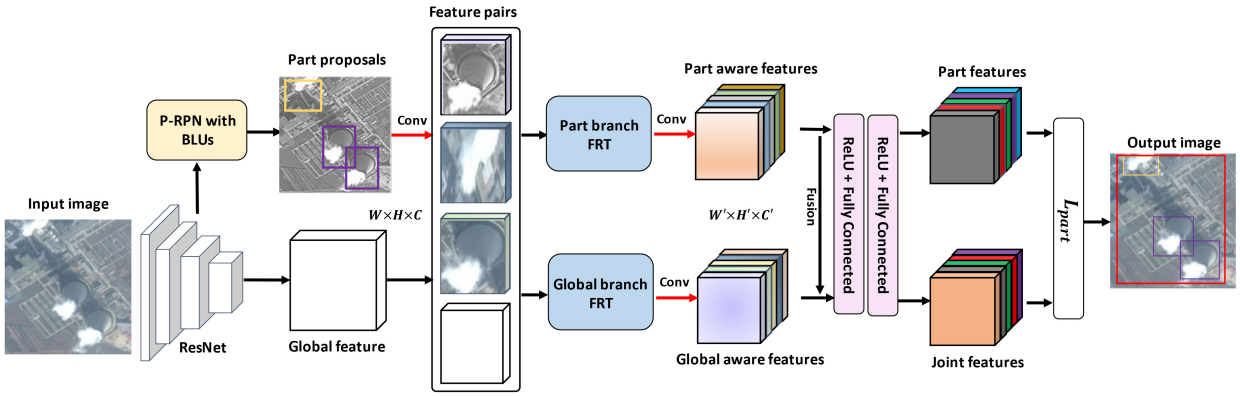


Fig. 2. Architecture of our proposed REPAN. P-RPN with BFUs alleviates feature confusion problems and generates multiscale-aware pyramid features with discriminative part proposals (three proposals as an example). Resized part proposal features and global features form multiple feature pairs as the input of FRTs with the part branch and the global branch. This global-and-part joint learning constructs holistic-and-inter relationships. The CD then utilizes relation-aware features for final detection.

$\mathcal{F}$ that calculates $\hat{y} = \mathcal{F}(\mathcal{I})$, which means we need to optimize the loss $\mathcal{L}$ between predicted values and labels

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\mathcal{F}(\mathcal{I} \mid \mathcal{R}, \mathcal{P}), y) \tag{1}$$

where θ^* is the optimized learned parameters generated by the optimized $\mathcal{L}$, and θ represents the learned parameters. The detection results rely on clear spatial relationships $\mathcal{R}$ and strong part feature representation $\mathcal{P}$. Assuming the object is in one image, because REPAN will reveal both part and global relationships through Transformers $\mathcal{A}$, we could divide the optimization body into global optimization and part optimization

$$\theta^* = \arg\min_{\theta} \Big\{ \mathcal{L}(\mathcal{F}\big((\mathcal{I} \mid \mathcal{A}(\mathcal{I}; \theta_g)), y\big) + \sum_{p} \lambda_p \big(\mathcal{L}(\mathcal{F}(\mathcal{I} \mid \mathcal{A}(\mathcal{I}; \theta_p), p), y)\big)\Big\} \tag{2}$$

where $\theta_{\{g,p\}}$ represent the global and part learned parameters through Transformers, and λ_p represents the individual part weight. Therefore, we divide the optimization problem in this work into three aspects: 1) strong part feature representation; 2) clear holistic correlation construction; and 3) clear intercorrelation construction. Thus, our proposed network is designed to model these factors via accurate part proposal discovery and a clear correlation understanding of our strategy is illustrated in Fig. 2. First, we localize the discriminative parts through P-RPN to get strong part feature representation, and then construct the inter-and-holistic correlation between localized parts and the globe through FRT. Finally, we combine part and global features through a CD to fully utilize the contextual correlation for final classification and detection. We will explain the details in the following sections.

B. Part Region Proposal Network

In traditional object detection tasks, generating region proposals from the discriminative information of the object features is vital before detection, where region proposal network (RPN) is most widely used [34]. Vanilla RPN utilizes feature maps from Faster R-CNN to generate proposals. Recently some works [12], [47], [48], [49] designed multiscale feature fusion methods to fully utilize pyramid features. For example, work [48] proposed a multiscale convolutional feature fusion strategy to use the highest-level feature map to supervise multiscale feature maps, making full use of semantic information in high-level features. By applying a feature pyramid, small-scaled components can be obtained in shallow-layer features and large-scaled components can be found in deep-layer features. Despite the promising results of these feature fusion strategies, there remains a critical gap in addressing the feature-scale confusion problem. Investigating this issue is paramount for mitigating the inherent scale variation challenges observed in part regions within composite object detection. Feature-scale confusion problem refers to the challenge in distinguishing features related to small-scale components from those associated with large-scale components in pyramid features, particularly in the context of composite object detection. Shallow-layer features contain rich detailed information beneficial for detecting small-scale components. However, they may also include characteristics of large-scale components, leading to interference in accurately localizing small components. Conversely, deep-layer features emphasize information about large-scale components, causing a decay in the representation of detailed and location information reserved in shallow layers. This feature-scale confusion hinders precise detection, especially in scenarios involving composite objects with diverse scales. As a result, kinds of multilayer fusion by convolution layers can neither address this feature-scale confusion problem nor extract the inter-relationship between certain parts, let alone lead to additional computation costs from convolutions. Hence, we propose a P-RPN to address the feature-scale confusion caused by aliasing effects, exploring the relationship between different scale feature representations containing different parts without extra convolution computation costs.

As shown in Figs. 2 and 3(a), we assume that input pyramid features have four levels (i.e., $\{F_1, F_2, F_3, F_4\}$). With the number increasing, the feature transfers from the shallow to deep layers. To extract the relationship between different

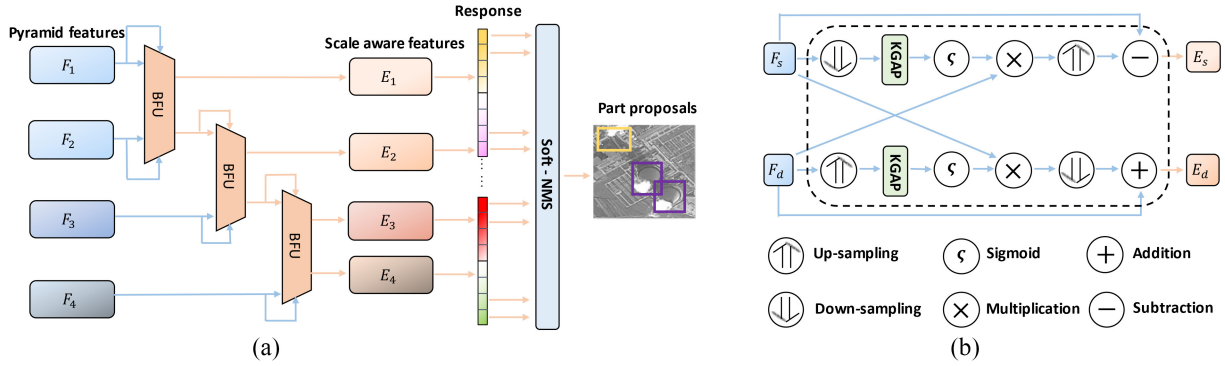


Fig. 3. Details of (a) P-RPN and (b) BFU. The pyramid features are fused progressively to alleviate the feature confusion problems with various scales. Top-P discriminative part proposals are then discovered in the generated multiscale-aware pyramid features.

feature representations in shallow layers F_s and deep layers F_d, we embed the BFUs in the P-RPN. Instead of using convolution layers, BFUs follow an unsupervised manner. As Fig. 3(b) shows, first F_d and F_s are transformed to F_d^s and F_s^d by upsampling $\Uparrow$ and downsampling $\Downarrow$, to keep the spatial resolution the same as F_s and F_d. To obtain more representative features and filter out less relevant information, a K-global average pooling (KGAP) along the channel dimension with a Sigmoid activation layer is placed after the up/down sampling to get the corresponding attention maps A_d^s and A_s^d of the transferred features F_d^s and F_s^d. The attention maps can be described as follows:

$$F_d^s = F_d{}^{\Uparrow}, \ F_s^d = F_s{}^{\Downarrow} \tag{3}$$

$$\mathrm{KGAP}_{(i,j)} = \frac{1}{K}\sum_{l=1}^{K} F_{(i,j)}^{l}(x) \tag{4}$$

$$A_d^s = \varsigma\big(\mathrm{KGAP}(F_d^s)\big), \ A_s^d = \varsigma\Big(\mathrm{KGAP}(F_s^d)\Big) \tag{5}$$

where $F_{(i,j)}^{l}(x)$ means the value at the position (i,j) in the lth feature map F^l, and ς means the Sigmoid function. $\mathrm{KGAP}_{(i,j)}$ means conduction of global average pooling on K representative values at the position (i,j) along the channel dimension. KGAP calculates the average value across a set of representative values (K values) at each position along the channel dimension. By aggregating information across feature maps, KGAP can obtain more representative features and selectively emphasize important spatial positions. Note that the selection of K along the channel dimension has a threshold T to filter unrelated pixels simply, and the values that are larger than T can be inputs of KGAP

$$T = \frac{1}{C}\sum_{l=1}^{C} F_{(i,j)}^{l}(x). \tag{6}$$

Thus, attention maps indicate the different discriminative parts from the channel aspects. The inter-relationship between F_s and F_d is obtained by the pixel-wise multiplication of the attention maps with the corresponding features. Then, F_s is further fused with the inter-relationships by pixel-wise subtraction to focus on the small parts. F_d is further fused with the inter-relationships by pixel-wise addition to gain more detailed information about significant components. These can be denoted by

$$E_s = F_s \ominus \left(F_d \odot A_s^d\right)^{\Uparrow}, \ E_d = F_d \oplus \left(F_s \odot A_d^s\right)^{\Downarrow}. \tag{7}$$

Therefore, for the original pyramid features with four levels with three BFUs embedded hierarchically as Fig. 3(a) shows, we can obtain four multiscale-aware pyramid features (i.e., $\{E_1, E_2, E_3, E_4\}$), which own the specific feature representations in various levels concerning the relationship between parts with different scales.

We note the peak response coordinates in each channel of each multiscale-aware pyramid feature. Then, we flatten multiscale-aware pyramid features and mark the peak responses R. Since the peak responses indicate the accuracy and confidence of prediction for corresponding regions, we generate part proposals according to these peak responses. With the initial anchor settings $\mathbf{A}$, we select the P most scored parts with different scales to locate the discriminative parts. We also apply soft nonmaximum suppression (Soft-NMS) [50] to eliminate the overlapping anchors. As a result, our P-RPN is able to find the most discriminative parts with improved precision and less computation cost. The main idea of the P-RPN is shown in Algorithm 1. Note that P-RPN is in a weakly supervised manner. Given image-level annotations, we can follow this coarse-grained label to extract discriminative part proposals, and then construct the discriminative part feature representation.

C. Feature Relation Transformer

After obtaining the most discriminative parts $\mathcal{P}$, we have addressed one of the dependencies in composite object detection mentioned in (1), the strong part feature representation. Then, the main problem is building inter-and-holistic semantic relationships between parts and the globe. The vanilla Transformers in vision can extract the long-term dependencies between pixels, yet with high computation costs. It is efficient to embed Transformers into CNNs, with larger receptive fields, lower computation costs, and an end-to-end training manner. Thus, we design an FRT combined with CNN architectures to investigate the correlations. The detailed architecture is shown in Fig. 4. Ideally, building correlations between every two parts can promote contextual understanding

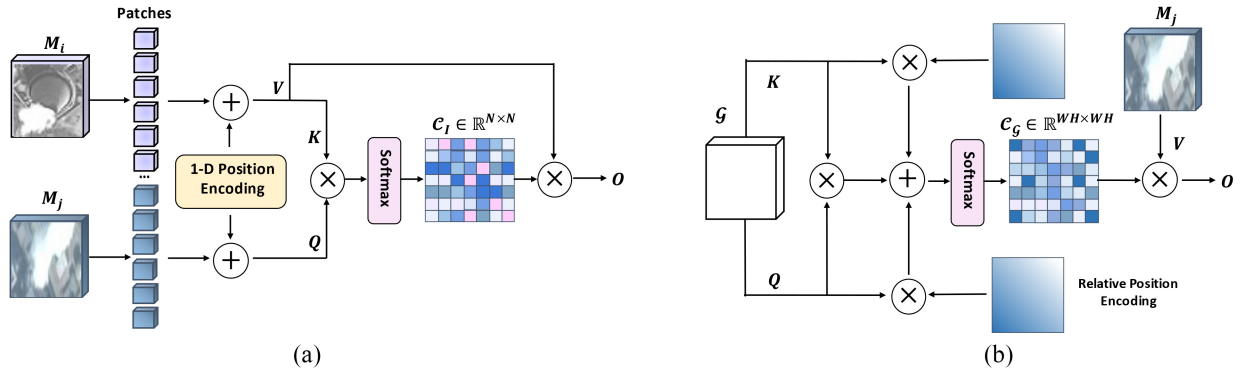


Fig. 4. Detailed structure of the FRT. (a) Interunderstanding branch constructs inter-relationships between parts. (b) Global understanding branch constructs holistic relationships around the globe. We set one feature pair as an example. Multiple feature pairs with a multihead attention mechanism build effectiveness.

Algorithm 1 P-RPN

Input: Input image $\mathcal{I}$, Hyperparameters: Number of pyramid features $\mathbf{X}$, Number of parts P, Initial anchor settings $\mathbf{A}$

Output: Discovery part set: $\mathcal{P} = \{p_1, p_2, \ldots, p_P\}$

Initialize discovery part set $\mathcal{P} = \emptyset$, and multi-scale aware pyramid features $\mathbf{E} = \emptyset$
Generate vanilla pyramid features from $\mathcal{I}$ through backbone $\mathbf{F} = \{F_1, F_2, \ldots, F_\mathbf{X}\}$
while $i < \mathbf{X} - 1$ **do**
  $(E_i, E_{i+1}) = BFU(F_i, F_{i+1})$
  $F_{i+1} = E_{i+1}$
  $\mathbf{E}$ append $(\mathbf{E}, E_i)$
  $i = i + 1$
  else
  $\mathbf{E}$ append $(\mathbf{E}, E_i, E_{i+1})$
end while
Rearrange $\mathbf{E}$ and mark peak responses R
Sort R and select top P peak responses
Generate anchors $\mathbf{A}(R_i)_{i=1}^P$
$\mathcal{P}$ append $P_i = SoftNMS(\mathbf{A}(R_i)_{i=1}^P)$
return $\mathcal{P} = \{p_1, p_2, \ldots, p_P\}$

maximally. However, the surge in computation costs will also influence performance. In addition, different part features with different confidence should not share the same weight. Part features with higher confidence should have more chances to interact with other parts, and global features can interact with all part features to build holistic correlation. Therefore, the first step is rescaling part and global features for later joint learning in Transformers. We resize part proposals to the same size with the input image (i.e., 512×512) and then resend the part proposal set $\mathcal{P}$ back to the backbone to extract the part features $\mathcal{M} = \{M_1, M_2, \ldots, M_P\} \in \mathbb{R}^{W\times\ H\times C}$, where the confidence decreases progressively since M_1. To build the correlation, we divide the construction into two branches. For the global understanding, we let the global feature $\mathcal{G} \in \mathbb{R}^{W\times\ H\times C}$ and every single part feature $(M_i)_{i=1}^P$ form an input pair $[(\mathcal{G}, M_1), (\mathcal{G}, M_2), \ldots, (\mathcal{G}, M_P)]$. For the inter-relationship construction, every part feature M_k combines with every other part features with lower confidence $(M_i)_{i=k+1}^P$ to form an input pair $[(M_1, M_2), \ldots, (M_1, M_P), (M_2, M_3), \ldots, (M_{P-1}, M_P)]$. In total, we will get a number of $\mathcal{N}_{IN} = P + P(P-1)/2$ input pairs and output $\mathcal{N}_{\text{OUT}} = P/2 + P(P-1)/4$ transformed features.

Patch Partition: As Fig. 4 shows, the patch partition has two branches. For interunderstanding, the input feature pair $\{[(M_i, M_j)], i > j, \{i, j\} \leq P\}$ is first cropped and flattened into a sequence with a size of 16×16. The sliding window keeps an overlapping size of 8 to maintain the consistency of the cropped features. Therefore, the feature dimension of each patch is $16 \times 16 \times C$, and the number of patches is $N = 2 \times \lfloor (H-8)/8 \rfloor \times \lfloor (W-8)/8 \rfloor$. A linear embedding layer is set after the patch partition to gain the linear projection with a shape of $16 \times 16 \times C'$. For global understanding, the input feature pair $\{[(\mathcal{G}, M_i)], i \leq P\}$ remains unchanged due to our 2-D relative position encoding strategy.

Position Encoding: Position encoding can preserve the original structural information when transforming the feature maps into multiple vectors [41]. This spatial location information in the patch level provides contextual understanding and promotes prediction performance. Similar to the patch partition process, position encoding also has two branches for local and global understanding, respectively. For interunderstanding, we want to build the intercorrelation map $\mathcal{C}_I \in \mathbb{R}^{N\times N}$ which shows the relation level of every single patch in one part to other patches in another part in the feature map. We have two sequences of patches through the linear embedding layer. We follow [43] to build 1-D absolute positional information for these two sequences in the raster order. Then, we formulate the process as

$$Z_i = \left[M_i^1, M_i^2, \ldots, M_i^N\right] \oplus \left[\mathcal{E}_1^1, \mathcal{E}_1^2, \ldots, \mathcal{E}_1^N\right] \tag{8}$$

$$\mathcal{C}_I = Z_i \mathbf{W}_Q \left(Z_j \mathbf{W}_K\right)^T, \ i < j \leq P \tag{9}$$

where $Z_{i,j}$ denotes the output after position encoding, and M_i^j represents the jth patch in the original feature M_i, and $\mathcal{E}_1^i$ means the 1-D position encoder corresponding to the ith patch. $\mathbf{W}_{\{\mathbf{Q},\mathbf{K}\}}$ is the learnable matrix of query and key projection. For global understanding, we want to build a global correlation matrix $\mathcal{C}_G \in \mathbb{R}^{WH\times WH}$. This matrix shows the relation level of each single pixel to all the other pixels in the feature map. To

receive the long dependencies between pixels in grid, we here follow [51] to employ relative position encoding in contextual mode. The process can be formulated

$$\mathcal{C}_G = \mathcal{G}_{i,:}\mathbf{W}_Q\mathbf{W}_K^T\mathcal{G}_{j,:}^T + \mathcal{G}_{i,:}\mathbf{W}_Q\left(\mathbf{r}_{i,:;j,:}^K\right)^T + \mathcal{G}_{j,:}\mathbf{W}_K\left(\mathbf{r}_{i,:;j,:}^Q\right)^T \quad (10)$$

where $\{Q, K\}$ denotes the query and key, and R denotes the encoding matrix. $\mathbf{W}_{\{\mathbf{Q},\mathbf{K}\}}$ is the learnable matrix of query and key projection, and $\mathbf{r}$ denotes a trained vector.

Multihead Attention: After patch and position embedding, the generated correlation maps can represent the holistic-and-inter relationships between pixels. Further, we multiply the correlation maps with the value projections $V_{\{G,I\}}$ of original features to get multiple Transformer heads $H_{\{G,I\}}$

$$H_{\{G,I\}} = \frac{\exp\{\mathcal{C}_{\{G,I\}}\}/\sqrt{C}}{\sum_{j=1}^{WH}\exp\{\mathcal{C}_{\{G,I\},:,j}\}/\sqrt{C}}V_{\{G,I\}}. \quad (11)$$

For each Transformer, we generate $\mathcal{H}$ Transformer heads with different projection weights of query, key and value, which could largely promote the construction of relation learning. Then, the attention heads are concatenated and further sent to a convolution layer to obtain relation-aware feature maps. To be specific, for global relationship between $(\mathcal{G}, M_j)$ and inter-relationship between (M_i, M_j), we have two kinds of relation-aware feature maps

$$\mathcal{O}_G^j = \mathrm{Conv}\left(\mathrm{Concat}(H_1^{\{G,j\}}, \ldots, H_{\mathcal{H}}^{\{G,j\}})\right) \quad (12)$$

$$\mathcal{O}_i^j = \mathrm{Conv}\left(\mathrm{Concat}(H_1^{\{i,j\}}, \ldots, H_{\mathcal{H}}^{\{i,j\}})\right). \quad (13)$$

D. Contextual Detector

To fully utilize the relation-aware feature maps from global and part aspects, we design a two-branch detector to obtain both global and part classification (see Fig. 2). The first branch takes the part-aware feature maps as inputs and classifies the parts by part annotations. The second branch fuses the part and global aware feature maps to form a more powerful representative feature for whole complex detection. With two fully connected layers and two ReLU activation layers subsequently, the features are reformed to a confidence matrix. In order to classify the parts and locate the bounding boxes, with training the detector in an end-to-end manner, we design a multitask loss function L_{part}

$$\begin{aligned} L_{\text{part}} &= L_{wcls} + L_{wloc} + L_{ploc} + L_{pcls} \\ &= \lambda_1(L_{cls}(S, C^{gt} + \lambda_2\left[C^{gt} > 0\right]L_{\text{loc}}(B^w, B^{gt})) \\ &\quad + \lambda_3(\sum_{s=1}^{4}(L_{cls}(s, c^{gt}) + \lambda_4\left[c^{gt} > 0\right]L_{\text{loc}}(b^w, b^{gt})) \end{aligned} \quad (14)$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters to balance different losses; L_{loc} is the bounding box loss (i.e., GIoU Loss [52]) for both global and part detection, where B^w is the regressed bounding box for the whole object, and B^{gt} is the ground-truth bounding box for the whole object, and b^w is the regressed bounding box for parts, and b^{gt} is the ground-truth bounding box for parts; L_{cls} is the classification loss for the global and part classification (i.e., softmax cross entropy loss), in which S means the prediction of the whole classification, and s means the prediction of part classification, and C^{gt} means the ground truth of whole object labels, and c^{gt} means the ground truth of part labels.

IV. Experiments and Evaluation

A. Dataset

Because there is a lack of open-source composite object detection datasets in RSIs, we collect a remote sensing complex composite object detection dataset to validate the performance of our proposed method based on three different open-source remote sensing datasets, i.e., DIOR [53], BUAA-FFPP60 [54], and DOTA [55], [56]. The DIOR dataset is one of the largest high-resolution remote sensing object detection datasets, containing 20 categories and over 20 000 annotated images. The BUAA-FFPP60 dataset contains 1-m spatial resolution RSIs of over 60 coal-fired power plants in the Beijing–Tianjin–Hebei region in North China. The DOTA dataset contains 15 common categories with a wide variety of scales, orientations, and shapes. Three datasets share the same source (i.e., Google Earth) and the same resolution range (i.e., $\leq$ 1 m), guaranteeing consistency in our collected dataset and the training process. As Table II illustrates, we collect coal-fired power plants from BUAA-FFPP60, harbors from DIOR and DOTA, airports from DIOR and DOTA, and expressway service areas from DIOR. For the coal-fired power plant, parts (i.e., chimney and condensing tower) form the whole functionality and structure. For one harbor, multiple shipyards are distinguishing parts from the whole harbor. For one expressway service area, the areas located on both sides of the expressway are different parts. For the airport, the airport runway and the terminals as parts consist of the whole airport.

B. Parameter Settings

We conduct our experiment on PyTorch deep learning framework [59], with 4 NVIDIA GeForce RTX 2080 Ti GPUs and 50 training epochs. The batch size is set as 4. The learning rate starts at 0.05 and decreases by a decreasing factor of 0.1 after every 10 epochs. We use mini-batch stochastic gradient descent (SGD) [60] as the optimizer for classifier training, and set a momentum of 0.9 and a weight decay of 0.0005. We set anchors with ratios of {0.2, 0.3, 0.5, 1, 2, 3, 5} and scales of {100, 150, 200, 400}. Besides, we also use multiscale training with the long edge set to 2000 and the short edge randomly sampled from [400, 1400], and online hard example mining (OHEM) [61] to handle hard example learning. Additionally, Soft-NMS [50] is also used to eliminate overlapping proposals.

C. Comparison With State of the Arts

We conduct a comparative study between our proposed REPAN and other state-of-the-art models on our collected dataset, including five standard object detection methods (i.e., Faster R-CNN [34], SSD [33], Cascade R-CNN [35], Dynamic R-CNN [57], Libra Faster R-CNN [36]), five part-based object detection methods (i.e., RA-CNN [58], MA-CNN [37],

TABLE I
COMPARISON RESULTS ON OUR COLLECTED DATASET, INCLUDING COAL-FIRED POWER PLANT, HARBOR, EXPRESSWAY SERVICE AREA, AND AIRPORT (%)

Method	Coal-fired power plant		Harbor		Expressway service area		Airport		**Average**	
	Precision	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision	Accuracy
SSD [33]	69.08	68.41	53.02	49.53	72.56	72.02	71.38	73.12	66.51	65.77
Faster R-CNN [34]	69.95	72.98	52.29	52.76	73.89	72.08	72.11	71.53	67.06	67.34
Cascade R-CNN [35]	74.27	74.15	58.34	58.92	75.76	74.33	74.16	75.54	70.63	70.74
Dynamic R-CNN [57]	74.14	75.80	56.72	58.45	75.24	77.07	76.60	72.84	70.68	71.04
Libra Faster R-CNN [36]	80.89	81.43	66.23	69.16	77.13	79.92	80.27	81.48	76.13	78.00
PBNet [6]	85.22	84.71	73.42	76.83	84.58	83.29	84.79	85.10	86.57	83.78
PCAN [20]	84.19	86.24	71.80	73.99	85.09	86.27	85.61	85.36	83.15	84.12
RA-CNN [58]	84.74	84.70	71.52	72.85	83.67	85.73	87.11	85.19	81.76	82.12
MA-CNN [37]	86.43	84.59	73.31	73.48	83.34	85.51	87.54	88.40	82.66	83.00
PA-CNN [38]	88.86	88.11	76.45	77.32	86.83	85.91	88.36	89.12	85.13	85.12
ACNet [39]	87.32	86.24	78.94	80.89	88.79	90.25	89.64	89.83	86.17	86.80
AP-CNN [40]	88.32	**90.41**	81.67	83.02	**89.28**	91.46	89.48	90.21	87.19	88.78
REPAN (Ours)	**90.32**	90.37	**82.87**	**84.49**	**89.28**	**92.74**	**90.19**	**90.36**	**88.17**	**89.49**

TABLE II
DETAILED INFORMATION OF THREE OPEN-SOURCE DATASETS USED IN OUR WORK

Dataset	DIOR [53]	BUAA-FFPP60 [54]	DOTA [55], [56]
Source	Google Earth	Google Earth	Google Earth
Resolution	0.1-1m	1m	0.1-1m
Category	Harbor,Airport, Expressway Service Area	Coal-fired power plants	Harbor,Airport
Training	1,873	800	643
Testing	2,036	92	285
Total	3,909	892	928

PA-CNN [38], ACNet [39], and AP-CNN [40]) and two composite object detection methods for RSIs (i.e., PBNet [6] and PCAN [20]). Table I lists the accuracy and precision of the aforementioned methods on our collected dataset. Because some part-based methods are in a weakly supervised manner, to be fair, we use image-level annotations in airport detection, harbor detection, and expressway service area detection. We employ complete humancrafted annotations, including part- and image-level labels only in the comparison of coal-fired power plant detection. Based on the same experimental settings for all methods, our REPAN achieves leading performance over other methods.

Comparison on Coal-Fired Power Plant Detection: Our REPAN achieves the best performance concerning both accuracy and precision at 90.37% and 90.32%, respectively. We can find that Libra Faster R-CNN [36] achieves the best performance in the standard method category, but is outperformed by REPAN by 9.43% in precision and 8.94% in accuracy. AP-CNN [40] outstrips all other part-based methods with a precision of 88.32% and an accuracy of 90.41%. REPAN still reaches a gain of 2.0% in precision, with a 0.04% loss on accuracy.

Comparison on Harbor Detection: REPAN still has the leading position with an accuracy of 82.87% and a precision of 84.49%, outperforming the second-best method (AP-CNN [40]) by a margin of 1.2% in precision and 1.47% in accuracy.

Comparison on Expressway Service Area Detection: Expressway service areas have more heterogeneous surroundings, making detecting them easier. Thus, recent part-based methods, i.e., AP-CNN [40] and ACNet [39], both achieve good results of 89.28% and 88.79% in precision. Though similarly good performance is achieved by other methods, our method can reach higher performance with an accuracy of 92.74% and reach equal performance in precision.

Comparison on Airport Detection: Even though the ACNet [39] achieves high scores with a precision of 89.64% and an accuracy of 89.83%, our REPAN still shows better performance, reaching a precision of 90.19% and an accuracy of 90.36%. We can see that exploring inter-relationships and building holistic understanding provide promising potential and performance on different kinds of composite object detection.

Visualization Comparison: Fig. 5 shows the visualization comparison results. We visualize the detection results of our REPAN and other state-of-the-art methods on four composite object categories. The results are divided into comparisons with part-based methods and comparisons with standard object detection methods for better observation. In the coal-fired power plant detection (the 1st row and the 5th row), we can find that our REPAN clearly classifies all parts of coal-fired power plants and detects the whole coal-fired power plant properly, achieving the best performance against all other methods. Part-based methods basically detect the whole coal-fired power plant, but the location and the range are not accurate. Besides, the fine-grained detection of parts is confusing, affected by similar surroundings and the smoke. The standard object detection methods generate multiple redundant bounding boxes which exceedingly decrease the accuracy. In the harbor detection (the 2nd row and the 6th row), there are similar situations to that in the coal-fired power plant detection. AP-CNN also generates redundant bounding boxes influenced by the ships. In two simpler tasks, the expressway service area detection (the 3rd row and the 7th row) and the airport detection (the 4th row

Fig. 5. Visualization results of the comparison study. (a) Input images; (b) ground truth; (c) REPAN (ours); (d) AP-CNN [40]; (e) ACNet [39]; (f) PA-CNN [38]; (g) MA-CNN [37]; (h) RA-CNN [58]; (i) Dynamic R-CNN [57]; (j) Libra Faster R-CNN [36]; (k) Faster R-CNN [34]; (l) SSD [33]; and (m) Cascade R-CNN [35].

and the 8th row), most models achieve better performance compared with their own performance in the tasks before. However, our REPAN still has the best performance for accurate bounding boxes which are closest to the ground truth.

In summary, the comprehensive experimental results show the robustness and ability of our REPAN across diverse composite object detection scenarios. Compared with existing methods, including part-based methods, standard methods, and composite object detection methods in remote sensing, REPAN outstands by two pivotal factors: 1) extracting strong part feature representations via P-RPN and 2) building clear spatial relationships via FRT.

V. Discussion

In this section, we conduct ablation studies for REPAN, discussing the effectiveness of each component in our work, including the part feature representations by part region proposal, the relationship awareness by FRT, and the joint correlation-aware features. We will also analyze the advantages and disadvantages of our methods compared with other latest methods, regarding the feature fusion ways, computation efficiency, etc. Through these deliberations, we provide insights into the effectiveness of REPAN while offering a comprehensive understanding of its operational mechanisms.

Effectiveness of Part Region Proposal: Fig. 6 shows the effectiveness of part region proposal from the perspectives of

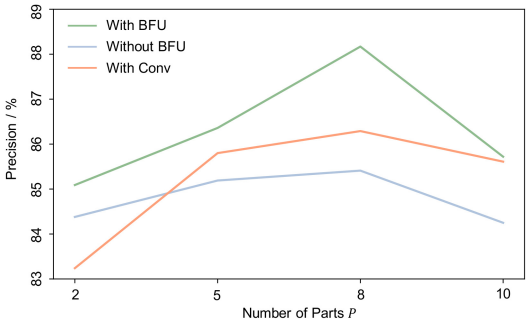


Fig. 6. Ablation study on P-RPN. We examine the effects of BFU and the number of parts P. The green line denotes the model is with BFU. The blue line denotes that we remote BFU directly. The orange line denotes that we replace BFU with convolution layers.

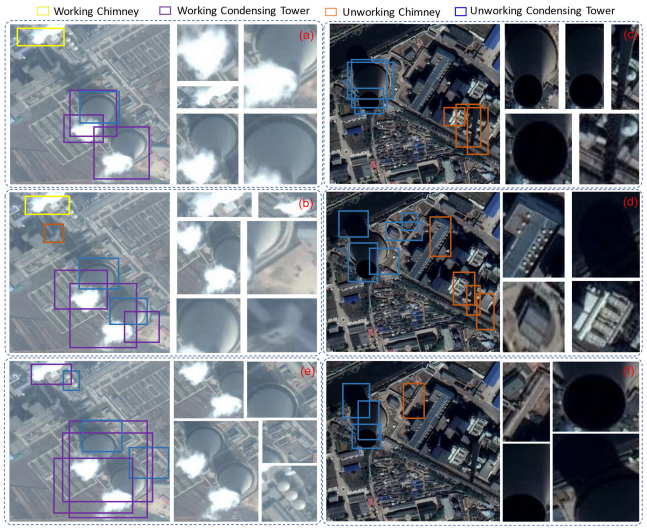


Fig. 7. Effectiveness visualization of BFU. We here take the coal-fired power plant detection task as an example. (a) and (c) are the part proposals generated by P-RPN with BFUs and the zoom-in images of parts. (b) and (d) are the part proposals generated by P-RPN without BFUs and the zoom-in images of parts. (e) and (f) are the part proposals generated by vanilla RPN and the zoom-in images of parts. P-RPN with BFUs can effectively discover the most discriminative parts in the image. More unrelated regions will be selected by P-RPN without BFUs. Vanilla RPN cannot extract the most useful proposals and the most harmful is that vanilla RPN can only recognize each part as a single object, not a part, resulting in a non-end-to-end composite object detection.

the number of parts P and the embedded BFUs. Considering the multiparts in one image, the value of P should be at least 2. We can see that with the value increasing, the precision increases at the beginning and then decreases when P is larger than 8. When the value of P is comparatively small, the real number of parts in one image exceeds the value, so it is hard to represent all the part features by an insufficient value of P. When the value is larger than 8, part proposal redundancy and confusion hinder the learning of discriminative features, leading to a drop in precision. As for the effectiveness of BFU, we conduct ablation experiments by simply removing BFU and replacing BFU with convolution layers. We can see that BFU still outperforms convolution layers no matter what P is. For example, when the P equals 8, if we remove the BFU and directly concatenate pyramid feature maps, the precision drops from 88.17% to 85.41%. If we replace BFU with convolution layers, the precision drops to 86.29% with about two times more model parameters and computation cost. Our final model holds about 28.8 M parameters for learning compared to the model with convolution layers of 45 M.

To be direct, we visualize the comparisons in Fig. 7. We can see the part region proposal with BFUs can locate significant parts more precisely by identifying the clear semantic information of the particular features. The part region proposal without BFUs may select some unrelated regions, influencing the later detector decision and increasing the possibility of false detection. With BFUs, the selected parts are more discriminative, which indicates the network pays more attention to the important parts and eventually contributes to better detection performance.

Also, we compare the effectiveness of BFU with other feature fusion methods. Table III shows the comparison of different feature fusion methods. To be fair, we embed these feature fusion methods into our network to make comparisons. P-RPN not only outperforms other methods regarding precision and accuracy but also owns less computation complexity. It can be illustrated from two aspects. First, different feature fusion methods have different orientations. Even though other feature fusion methods have promising performance on public datasets, these cannot address the feature-scale confusion problem. Designed for single-object detection, these methods just aim to make full use of semantic information. However, P-RPN is designed for composite object detection, which can erase large object features in shallow layers and add detailed information to deep layers. Addressing feature scale confusion is vital for accurate composite object detection, ensuring precise localization of diverse object components and facilitating semantic understanding of their relationships. Second, these methods are implemented with convolution layers, which increase the computation complexity and model parameters. BFUs in P-RPN only use pooling, multiplication, addition, and subtraction to complete scale-aware feature generation. This indicates that BFUs designed for feature-scale confusion problems are proper for composite object detection, which may not perform on single-object detection as well as on composite object detection.

TABLE III
COMPARISON OF FEATURE FUSION METHODS

Method	Precision	Accuracy	Params
Aug-FPN [12]	85.13	86.73	32.4M
AF-FPN [47]	81.94	79.98	31.8M
Qu et al. [48]	84.27	85.36	30.9M
MF2CNet [49]	83.09	84.25	34.7M
BFU (Ours)	**88.17**	**89.49**	**28.8M**

Effectiveness of FRTs: To evaluate the effectiveness of the FRTs, here we conduct the ablation studies on global–part branches, position embedding, and the number of attention heads $\mathcal{H}$. As Table IV shows, when we only use the global branch or part branch, the precision drops to 86.96% and 87.42%, respectively. It indicates that building local

TABLE IV
ABLATION STUDY ON FRT (%)

Branch	Number of attention head $\mathcal{H}$	Position encoding	Precision
N/A	N/A	✗	84.43 (baseline)
Part	2	✓	86.78
		✗	86.61
	6	✓	87.42
		✗	87.37
	10	✓	87.47
		✗	87.41
Global	2	✓	86.80
		✗	86.75
	6	✓	86.96
		✗	86.94
	10	✓	86.81
		✗	86.75
Both	2	✓	87.54
		✗	87.49
	6	✓	**88.17**
		✗	87.89
	10	✓	87.74
		✗	87.61

TABLE V
ABLATION STUDY ON CD (%)

Features	Original features	Global-aware features	Part-aware features	Global-aware with part-aware features
Precision	82.18 (baseline)	85.13	84.97	**88.17**

TABLE VI
COMPARISON OF COMPUTATION EFFICIENCY

Model		GFlops ↓	Params ↓
RA-CNN [58]		37.8	33.0M
MA-CNN [37]		35.5	30.5M
ACNet [39]		33.1	30.0M
AP-CNN [40]		31.9	**28.0M**
REPAN	$Full$	**31.8**	28.8M
	$-P\text{-}RPN$	31.0	27.9M
	$-FRT$	25.6	20.8M
	$-CD$	29.7	27.7M

TABLE VII
SELECTION OF HYPERPARAMETERS IN MULTITASK LOSS FUNCTION

λ_1	λ_2	λ_3	λ_4	Precision
1	1	1	1	85.93
2	2	1	1	85.09
1	1	2	2	**88.17**
1	1	2	1	86.74
1	1	1	2	86.50

or global relationships alone can cause a loss of attention and information from the other side. Besides, the difference between the two precision results shows part inter-relationship exploration may play a more important role in the learning process. The position embedding also contributes to a little improvement in precision. There are some decreases when we remove the position embedding, no matter whether we use both two branches or not. However, the differences are comparatively small compared with other components. Table IV also demonstrates the effectiveness of the right number of attention heads $\mathcal{H}$. The detection performance will decrease if $\mathcal{H}$ is lower or higher than 6. Specifically, the precision with 2 or 10 heads are 87.91% and 87.74%, decreasing by 0.26% and 0.43%, respectively. Too few attention heads bring less learnable weights and limit the learning ability of the model. Too many attention heads will lead to insufficient learning and a larger computation cost.

Effectiveness of CD: We verify the contributions of the fusing features (global-aware feature maps with part-aware feature maps) in boosting detection performance. We compare the performance of whole composite object detection by involving the part-aware features and only using the global-aware features. Table V shows the combination of part-aware features with the global-aware features can improve the detection performance by 3.04% in precision. The global-aware features or part-aware features alone could also improve the performance compared with the original global features by 2.95% or 2.79% in precision. It proves that the fusion of discriminative part information does not only help to classify and localize the parts, but also be beneficial for whole object detection.

Computation Efficiency and Hyperparameters: To make a full comparison and evaluate effectiveness from all aspects, here we compare our method's efficiency with existing approaches and compare different modules' computation costs to offer a comprehensive perspective on computational performance. As Table VI shows, REPAN has the least GFlops compared with the existing SOTA part-based methods, with the second least parameters. Besides, we compare different modules' computation costs, which shows the FRT contributes the most computation costs compared with the other two modules. We also conduct an ablation study on the hyper-parameters in the multitask loss to analyze their influence on the model performance. As Table VII shows, we conduct five experiments with different value combinations. When we set all parameters as 1, all tasks share the same importance. When we pay more attention to the whole object detection, the precision drops a little. The precision reaches the highest when we pay attention to the part detection. It may indicate in composite object detection tasks, the classification and localization of parts are more important and challenging than those of the whole object, which is also intuitive because the more clearly the model understands and detects parts, the better the detection of the whole object can be achieved.

In summary, our ablation studies on part region proposal, FRT, and CD confirm the following.

1) Both strong part feature representation and clear semantic correlation contribute to the improvement of

composite object detection performance. Existing works just focus on strong part feature representation, but it is also the relationships between parts that lead to the distinct characteristics of every single composite object. Thus, exploring the potential relationships beside strong part feature representation brings improvements.

2) Effective fusion of global-aware features and part-aware features is good for final detection. Simply utilizing global-aware or part-aware features can also improve the performance because of the reveal of correlation. Feature fusion improves the performance further, which can be assumed that it combines global correlation and part correlation, where joint awareness of the correlation between parts and the globe is realized.

VI. Conclusion

In this article, we focus on the complex composite object detection in RSIs and propose a REPAN. We are the first to point out that composite object detection is based on part feature representation and spatial relationships. For part feature representation, we design a P-RPN to discover discriminative parts robustly and precisely by alleviating the feature confusion. For spatial relationships, it is the first time to propose an FRT to build the holistic and intersemantic correlation by global-and-part joint learning. The potential correlation between the globe and parts addresses the problems of complex weak spatial relationships and similar texture disturbance. With the relation-aware features generated by Transformers, the CD conducts final detection with promising performance. Evaluations of our collected dataset with four categories and the comprehensive ablation studies demonstrate the superiority of our proposed REPAN. In the future, we will continue to explore the large-scale application ability of our REPAN.

References

[1] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016.

[2] J. Zheng et al., "Growing status observation for oil palm trees using unmanned aerial vehicle (UAV) images," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 95–121, Mar. 2021.

[3] G. Li, Z. Liu, D. Zeng, W. Lin, and H. Ling, "Adjacent context coordination network for salient object detection in optical remote sensing images," *IEEE Trans. Cybern.*, vol. 53, no. 1, pp. 526–538, Jan. 2023.

[4] W. Li, R. Dong, H. Fu, and L. Yu, "Large-scale oil palm tree detection from high-resolution satellite images using two-stage convolutional neural networks," *Remote Sens.*, vol. 11, no. 1, p. 11, 2019.

[5] Z. Chen, D. Chen, Y. Zhang, X. Cheng, M. Zhang, and C. Wu, "Deep learning for autonomous ship-oriented small ship detection," *Safety Sci.*, vol. 130, Oct. 2020, Art. no. 104812.

[6] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 173, pp. 50–65, Mar. 2021.

[7] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 3735–3739.

[8] J. Zheng et al., "Surveying coconut trees using high-resolution satellite imagery in remote atolls of the pacific ocean," *Remote Sens. Environ.*, vol. 287, Mar. 2023, Art. no. 113485.

[9] D. Yu and S. Ji, "A new spatial-oriented object detection framework for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 10, Nov. 2021, Art. no. 4407416.

[10] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[11] Y. Pang, Y. Li, J. Shen, and L. Shao, "Towards bridging semantic gap to improve semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4230–4239.

[12] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "AugFPN: Improving multi-scale feature learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12595–12604.

[13] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," 2014, *arXiv:1406.2952*.

[14] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1173–1182.

[15] S. Cao, D. Joshi, L. Gui, and Y.-X. Wang, "HASSOD: Hierarchical adaptive self-supervised object detection," in *Proc. 37th Adv. Neural Inf. Process. Syst.*, 2024, pp. 1–23.

[16] L. Zhang, S. Huang, W. Liu, and D. Tao, "Learning a mixture of granularity-specific experts for fine-grained categorization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8331–8340.

[17] Y. Gao, X. Han, X. Wang, W. Huang, and M. Scott, "Channel interaction networks for fine-grained image categorization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10818–10825.

[18] Z. Huang and Y. Li, "Interpretable and accurate fine-grained recognition via region grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8662–8672.

[19] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 27, pp. 1487–1500, 2018.

[20] W. Yin, W. Diao, P. Wang, X. Gao, Y. Li, and X. Sun, "PCAN—Part-based context attention network for thermal power plant detection in remote sensing imagery," *Remote Sens.*, vol. 13, no. 7, p. 1243, 2021.

[21] W. Qian, Z. Yan, Z. Zhu, and W. Yin, "Weakly supervised part-based method for combined object detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5024–5036, Jun. 2022, doi: 10.1109/JSTARS.2022.3179026.

[22] S. Li, Y. Xu, M. Zhu, S. Ma, and H. Tang, "Remote sensing airport detection based on end-to-end deep transferable convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1640–1644, Oct. 2019.

[23] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Jun. 2022, Art. no. 5625411.

[24] Y. Xu, M. Zhu, S. Li, H. Feng, S. Ma, and J. Che, "End-to-end airport detection in remote sensing images combining cascade region proposal networks and multi-threshold detection networks," *Remote Sens.*, vol. 10, no. 10, p. 1516, 2018.

[25] Y. Yao et al., "On improving bounding box representations for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 21, Dec. 2022, Art. no. 5600111.

[26] H. Fu, X. Fan, Z. Yan, and X. Du, "Detection of schools in remote sensing images based on attention-guided dense network," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 11, p. 736, 2021.

[27] S. Yuan, J. Zheng, Y. Huang, J. Liu, H. Fu, and R. C. Cheung, "CO-detector: Towards complex object detection with cross-part feature learning in remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 1941–1944.

[28] S. Yuan, J. Zheng, L. Zhang, R. Dong, R. C. C. Cheung, and H. Fu, "MUREN: MUltistage recursive enhanced network for coal-fired power plant detection," *Remote Sens.*, vol. 15, no. 8, p. 2200, 2023.

[29] G. Cheng et al., "Dual-aligned oriented detector," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5618111.

[30] G. Cheng, Q. Li, G. Wang, X. Xie, L. Min, and J. Han, "SFRNet: Fine-grained oriented object recognition via separate feature refinement," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, May 2023, Art. no. 5610510.

[31] B. Cai, Z. Jiang, H. Zhang, D. Zhao, and Y. Yao, "Airport detection using end-to-end convolutional neural network with hard example mining," *Remote Sens.*, vol. 9, no. 11, p. 1198, 2017.

[32] H. Fu, X. Fan, Z. Yan, X. Du, H. Jian, and C. Xu, "Feature enhanced anchor-free network for school detection in high spatial resolution remote sensing images," *Appl. Sci.*, vol. 12, no. 6, p. 3114, 2022.

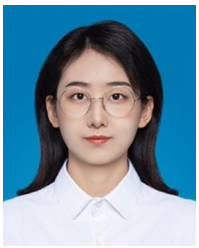
[33] W. Liu et al., "SSD: Single shot MultiBox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[35] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
[36] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.
[37] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5209–5217.
[38] H. Zheng, J. Fu, Z.-J. Zha, J. Luo, and T. Mei, "Learning rich part hierarchies with progressive attention networks for fine-grained image recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 476–488, 2020.
[39] R. Ji et al., "Attention convolutional binary neural tree for fine-grained visual categorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10468–10477.
[40] Y. Ding et al., "AP-CNN: Weakly supervised attention pyramid convolutional neural network for fine-grained visual classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2826–2836, 2021.
[41] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[42] K. Han et al., "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023.
[43] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.
[44] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.
[45] Y. Li et al., "Global transformer and dual local attention network via deep-shallow hierarchical feature fusion for retinal vessel segmentation," *IEEE Trans. Cybern.*, vol. 53, no. 9, pp. 5826–5839, Sep. 2023.
[46] L. Yuan et al., "Tokens-to-token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 558–567.
[47] Z. Zuo et al., "AFFPN: Attention fusion feature pyramid network for small infrared target detection," *Remote Sens.*, vol. 14, no. 14, p. 3412, 2022.
[48] Z. Qu, C. Cao, L. Liu, and D.-Y. Zhou, "A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4890–4899, Sep. 2022.
[49] L. Bai, Q. Liu, C. Li, Z. Ye, M. Hui, and X. Jia, "Remote sensing image scene classification using multiscale feature fusion covariance network with octave convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Mar. 2022, Art. no. 5620214.
[50] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS–Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5561–5569.
[51] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10033–10041.
[52] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 658–666.
[53] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.
[54] H. Zhang and Q. Deng, "Deep learning based fossil-fuel power plant monitoring in high resolution remote sensing images: A comparative study," *Remote Sens.*, vol. 11, no. 9, p. 1117, 2019.
[55] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
[56] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.
[57] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: Towards high quality object detection via dynamic training," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 260–275.
[58] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4438–4446.
[59] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. 33rd Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[60] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2012, pp. 421–436.
[61] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.

Shuai Yuan received the bachelor's degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2021, and the M.Phil. degree from the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, in 2023. He is currently pursuing the Ph.D. degree in global sustainability with the Department of Geography, The University of Hong Kong, Hong Kong, China.

His research interests include artificial intelligence, object detection, land cover change detection, and wetland mapping.

Lixian Zhang received the Ph.D. degree in ecology from the Department of Earth System Science, Tsinghua University, Beijing, China, in 2024.

He is a Postdoctoral Researcher affiliated with the National Supercomputing Center in Shenzhen, Shenzhen, China. His research interests include building extraction from remote sensing images, deep learning, and remote sensing image super-resolution reconstruction.



Runmin Dong received the Ph.D. degree in ecology from the Department of Earth System Science, Tsinghua University, Beijing, China, in 2022.

She is a Postdoctoral Researcher affiliated with the Department of Earth System Science, Tsinghua University. Her research interests include remote sensing, artificial intelligence, land cover mapping, super-resolution, image fusion, image synthesis, and self-supervised learning.

Jie Xiong received the Ph.D. degree in management from EMLYON Business School, Écully, France, in 2013.

He is an Associate Professor with the ESSCA School of Management, Angers, France. His research has been published in journals, such as IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, *Regional Studies*, and *Journal of Environmental Management*. His research interests include catching up of latecomer industries and firms, sustainable development and innovation, and interface research between artificial intelligence and management science, with emerging markets focus.

Juepeng Zheng (Member, IEEE) received the Ph.D. degree from the Department of Earth System Science, Tsinghua University, Beijing, China, in 2023.

He is currently an Assistant Professor with the School of Artificial Intelligence, Sun Yat-sen University (Zhuhai), Zhuhai, China. He is also a Researcher with the National Supercomputing Center in Shenzhen, Shenzhen, China. His research interests include remote sensing image understanding, high-performance computing, deep learning, transfer learning, and parallel computing for remote sensing applications.

Haohuan Fu (Senior Member, IEEE) received the Ph.D. degree in computing from Imperial College London, London, U.K., in 2009.

He is currently a Professor with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, and the Ministry of Education Key Laboratory for Earth System Modeling and the Department of Earth System Science, Tsinghua University, Beijing, China. He is also the Deputy Director of the National Supercomputing Center in Wuxi, Wuxi, China. His research interests include high-performance computing in Earth and environmental sciences, computer architectures, performance optimizations, and programming tools in parallel computing.

Peng Gong received the Ph.D. degree in geography from the University of Waterloo, Waterloo, ON, Canada, in 1990.

He is currently the Chair Professor of Global Sustainability with the University of Hong Kong, Hong Kong. He built the Department of Earth System Science and was the Dean for the School of Sciences, Tsinghua University, Beijing, China. He was also the Founding Director for the Tsinghua Urban Institute, Beijing. He had previously taught at the University of Calgary, Calgary, AB, Canada, and the University of California at Berkeley, Berkeley, CA, USA. He has authored or coauthored more than 600 articles and ten books. His research interests include mapping, monitoring, and modeling of global environmental change, and modeling of environmentally related infectious diseases, such as schistosomiasis, avian influenza, dengue, and COVID-19, and healthy and sustainable cities.