Learning to Aggregate Multi-Scale Context for Instance Segmentation in Remote Sensing Images
Ye Liu, Huifang Li, Member, IEEE, Chao Hu, Shuang Luo, Yan Luo, and Chang Wen Chen, Fellow, IEEE

Abstract—The task of instance segmentation in remote sensing images, aiming at performing per-pixel labeling of objects at the instance level, is of great importance for various civil applications. Despite previous successes, most existing instance segmentation methods designed for natural images encounter sharp performance degradations when they are directly applied to top-view remote sensing images. Through careful analysis, we observe that the challenges mainly come from the lack of discriminative object features due to severe scale variations, low contrasts, and clustered distributions. In order to address these problems, a novel context aggregation network (CATNet) is proposed to improve the feature extraction process. The proposed model exploits three lightweight plug-and-play modules, namely, dense feature pyramid network (DenseFPN), spatial context pyramid (SCP), and hierarchical region of interest extractor (HRoIE), to aggregate global visual context at feature, spatial, and instance domains, respectively. DenseFPN is a multi-scale feature propagation module that establishes more flexible information flows by adopting interlevel residual connections, cross-level dense connections, and feature reweighting strategy. Leveraging the attention mechanism, SCP further augments the features by aggregating global spatial context into local regions. For each instance, HRoIE adaptively generates RoI features for different downstream tasks. Extensive evaluations of the proposed scheme on iSAID, DIOR, NWPU VHR-10, and HRSID datasets demonstrate that the proposed approach outperforms state-of-the-arts under similar computational costs. Source code and pretrained models are available at https://github.com/yeliudev/CATNet.

Index Terms—Feature pyramid networks, global context aggregation, instance segmentation, object detection.

Manuscript received 22 March 2022; revised 19 June 2023 and 24 August 2023; accepted 20 November 2023. Date of publication 23 January 2024; date of current version 8 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 42371366, in part by the Hubei Provincial Key Research and Development Program under Grant 2023BAB066, and in part by the Fundamental Research Funds for the Central Universities under Grant 2042023kfyq04. (Corresponding author: Huifang Li.)

Ye Liu is with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China, and also with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR, China (e-mail: ye-liu@whu.edu.cn; coco.ye.liu@connect.polyu.hk).
Huifang Li and Chao Hu are with the School of Resource and Environmental Sciences, Wuhan University, Wuhan 430079, China (e-mail: huifangli@whu.edu.cn; chaohu@whu.edu.cn).
Shuang Luo is with Changjiang Spatial Information Technology Engineering Company Ltd., Wuhan 430074, China (e-mail: sluo@whu.edu.cn).
Yan Luo is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR, China (e-mail: silver.luo@connect.polyu.hk).
Chang Wen Chen is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, SAR, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: changwen.chen@polyu.edu.hk).
Digital Object Identifier 10.1109/TNNLS.2023.3336563

I. INTRODUCTION

Recent advances in satellites and remote sensing techniques have generated a large variety of high-resolution remote sensing images, bringing great challenges to manual manipulation and processing. Therefore, automatic analysis and understanding of these images are becoming increasingly essential for various civil applications [1], [2], [3], [4], [5], [6]. As a fundamental yet challenging task in computer vision, instance segmentation, which is a combination of object detection and semantic segmentation aiming at predicting binary masks of objects at the instance level, has been widely used to extract fine-grained object information from both optical images.
remote sensing images and synthetic aperture radar (SAR) images. It has attracted considerable attention in recent years.

Most existing works on object detection and instance segmentation [4], [7], [8], [9], [10], [11], [12] have been successful in conventional front-view scenes. However, when directly applied to remote sensing images, a large number of these methods inevitably encounter performance degradations [13], [14], [15], [16], [17]. Compared with natural images, remote sensing images are typically viewed from the top, capturing large areas with limited object discrepancies, shown in Fig. 1. We analyze the peculiarities of these scenes and divide the challenges into five distinct aspects, i.e., scale variation, arbitrary orientation, clustered distribution, low contrast, and cluttered background, as illustrated in Fig. 2. The first three aspects lead to complicated object patterns while the remaining two bring interfering information from the background. These phenomena are scarce in natural scenes, thus only a few works have considered these aspects. We argue that all the challenges above are due to the lack of discriminative object features in remote-sensing images. That is, the visual appearances of individual objects in remote sensing images are not informative enough for directly adopting existing schemes to perform instance segmentation. Such a view is also supported by previous works [18], [19], [20], [21], [22].

A natural question will be: *How to enhance the inadequate features to achieve better instance segmentation results in remote sensing images?* Considering that, in a general instance segmentation pipeline [9], [10], [11], object representations are directly cropped from either of the feature maps from the backbone or neck, containing only local features with irreversible information loss. In this work, we mitigate this problem by introducing context aggregation network (CATNet), a novel framework for global context aggregation. The key idea is that context information of images, coming from different feature pyramid levels, spatial positions, or receptive fields, shall provide extra prior for segmenting indistinguishable objects. Note that existing works [23], [24] only regard context as spatial correlations. We expand and explicitly disentangle the concept of context into three domains, i.e., *feature, space, and instance*. The implication is that when detecting and segmenting objects, this model may augment the visual information by: 1) balancing the heterogeneous features; 2) fusing information from the background or other correlate objects; and 3) adaptively refining intermediate representations for each instance and task. These three different domains are capable of modeling global visual context from coarse to fine at different granularities, capturing more discriminative object information.

The proposed framework intends to leverage three plug-and-play modules and construct the aforementioned context aggregation pipeline. In the feature domain, we argue that, in the feature pyramid built up by a backbone, flexible information flows may reduce information confounding and handle multi-scale features more effectively. Based on this analysis, a dense feature pyramid network (DenseFPN) is proposed to enable adaptive feature propagation. This module has a pyramid structure with stackable basic blocks consisting of top-down and bottom-up paths. We adopt interlevel residual connections [25], cross-level dense connections [26], and feature reweighting strategy to enable the module to learn its optimal feature propagation manner. In the spatial domain, long-range spatial dependencies in remote sensing images bring more complementary information to blurry objects than in natural scenes. So the spatial context pyramid (SCP) is adopted to capture global spatial context in each feature pyramid level. This module learns to aggregate features from the whole feature map and combines them into each pixel using adaptive weights. Such a strategy guarantees that only useful global information is fused into local regions, without decreasing the discrepancies among objects. As for the instance domain, we argue that object representations should be adaptively refined for each instance and downstream task. For example, performing object classification needs an overall view, while segmentation requires more zoomed-in view details. The demand for different sizes of receptive fields also varies among instances. Hence, we introduce a hierarchical region of interest extractor (HRoIE) to generate RoI features per instance and task. After cropping the instance feature maps from all levels, this module starts from the highest or lowest scale and fuses the features level-by-level in a hierarchical manner. The pixel-wise attention mechanism is exploited to combine neighboring feature maps. These modules are lightweight while having the flexibility for scalable model design. Overall, the main contributions of this article are threefold as follows.

1) The expansion and explicit disentangling of the concept of context into *feature, spatial*, and *instance* domains have resulted in superior performance in both optical remote sensing image and SAR image segmentation. To the best of our knowledge, this is the first work that considers global visual context beyond spatial dependencies.

2) The proposed CATNet is capable of utilizing DenseFPN, SCP, and HRoIE to learn and aggregate the global visual context from different domains for object detection and instance segmentation in remote sensing images.

3) The proposed scheme has been tested on a wide variety of datasets, including iSAID, DIOR, NWPU VHR-10, and HRSID, and new state-of-the-art performance has been obtained with similar computational costs.

The rest of this article is organized as follows. Related works and comparisons are discussed in Section II. Detailed formulations of DenseFPN, SCP, and HRoIE are introduced in Section III. Section IV presents extensive experimental results.
Fig. 3. Overall architecture of the proposed framework. The process of global context aggregation is realized by three modules, namely, (a) DenseFPN, (b) SCP, and (c) HRoIE. These modules are designed to aggregate global context information from different feature pyramid levels, spatial positions, and receptive fields at feature, spatial, and instance domains, respectively.

II. RELATED WORK

A. Instance Segmentation in Remote Sensing Images

Instance segmentation is a challenging and broadly studied problem in computer vision. Similar to object detection [7], [8], the majority of instance segmentation approaches can be divided into two schemes, namely, one-stage methods and two-stage methods. As a straightforward design, one-stage methods [12], [27] adopt the bottom-up strategy that performs semantic segmentation at the image level, and further separates individual objects using clustering or metric learning. These methods often possess considerable efficiencies but are largely restricted by their localization accuracy. Compared to this paradigm, two-stage methods [9], [10], [11], [28] separate the segmentation pipeline into two phases, i.e., region proposal generation and task-specific postprocessing, resulting in a top-down style. Benefiting from two-time bounding box regression, these methods usually achieve better results on object localization and mask prediction. Some recent works [18], [19], [20], [21], [29], [30], [31], [32], [33], [34] try to tackle the problem of scene classification and object detection in remote sensing images, but they do not pay partial attention to instance segmentation. Our proposed context aggregation strategy can be integrated into both one-stage and two-stage methods, while HRoIE is not used in one-stage methods since it’s not necessary to crop the feature maps. Further experimental results demonstrate that our modules can steadily boost performances.

B. Multi-Scale Feature Propagation

Scale variation is a long-standing challenge in most visual recognition tasks, and it is a common solution to leverage a pyramid structure to represent the visual features under different resolutions. In the area of dense prediction, FPN [35] is the first work that builds up a feature pyramid, and propagates the information among different levels. Such a design rapidly became a standard for most instance segmentation models. However, FPN only propagates the features using a top-down path, which is suboptimal for multi-level feature fusion. Thus, PAFPN [28] was proposed to incorporate an extra bottom-up propagation path, and FPG [36] further introduced a multipathway feature pyramid that can better capture cross-level information. MHN [37] tackles the semantic gap problem by leveraging semantic feature maps. A similar multibranch structure is also used in TridentNet [38] and NETNet [39] for scale-aware training and generating scale-aware features. From the perspective of basic operators, scale-equalizing pyramid convolution (SEPC) [40] tends to fuse the feature maps using 3-D convolutions. More recently, recursive feature pyramid (RFP) [41] reuses the backbone to capture deeper semantics via feedback connections from FPN. In order to reduce the computational costs, NAS-FPN [42] is obtained by introducing neural architecture search (NAS) [43] to optimize the architecture of FPN, and BiFPN [44] is constructed by carefully designing the feature fusion blocks. We argue that compared with hard-coded feature aggregation paths, flexible information flows may reduce confounding and handle multi-scale features more effectively so DenseFPN is proposed to learn the optimal feature aggregation strategy during training.

C. Global Context Modeling

One of the most representative properties of convolutional neural networks (CNNs) is local dependency modeling. The
segmentation would be a collection of tuples whichever the predefined categories. The output of instance detect and segment all the objects in the dependencies in feature and instance domains.

been achieved by these works, all these methods merely for global context modeling. Although promising results have been produced by these works, all these methods merely for global context modeling. Although promising results have been achieved by these works, all these methods merely consider context as long-range spatial correlations, ignoring the dependencies in feature and instance domains.

III. PROPOSED APPROACH

In this section, we introduce our approach to global context aggregation. As shown in Fig. 3, the entire framework can be divided into three submodules, namely, DenseFPN, SCP, and HRoIE. These modules aim to aggregate global context information in order from feature, spatial, and instance domains, respectively.

A. Overview

Given an image $x$ and a set of object categories of interest $S = \{1, \ldots, N\}$, the task of instance segmentation aims to detect and segment all the objects in $x$, where they belong to whichever the predefined categories. The output of instance segmentation would be a collection of tuples $T = \{(b, m, s)\}$, where $b \in \mathbb{R}^d$ denotes the bounding box of the object, $m$ represents a binary mask in which $m_{i,j} \in \{0, 1\}$ indicates whether the pixel $(i, j)$ belongs to this object, and $s \in S$ is a one-hot vector describing the object category. Note that a single object may be presented by separate masks.

We adopt Mask R-CNN [9], a common two-stage instance segmentation framework, as our baseline. The whole pipeline is constructed by extracting visual features, generating region proposals, and performing bounding box regression, object classification as well as mask prediction on each proposal. A heterogeneous feature pyramid is first built by extracting visual features from each stage of the backbone. In order to make the features more discriminative, we exploit DenseFPN and SCP to propagate object information among different levels and regions. After enhancing the feature pyramid, task-specific RoI features are generated by HRoIE for each proposal. Details of these modules are introduced in Sections III-B–III-D.

B. Dense Feature Pyramid Network

Multi-scale feature propagation aims to aggregate visual features from different backbone stages, that is given an input feature pyramid $C = \{C_1, C_2, \ldots\}$, where $C_i$ denotes the feature map from the stage $i$, the goal is to propagate the features among different levels to produce an enhanced feature pyramid $P = \{P_1, P_2, \ldots\}$, in which the features are more informative for downstream tasks. Formally, the resolution of feature map $C_i$ or $P_l$ is $1/2^l$ of the input image.

The basic architecture of DenseFPN is shown in Fig. 3(a), where each node represents a feature map and the lines stand for information flows. This module takes $C_2 \sim C_5$ as inputs and first down-samples them to 256 channels using $1 \times 1$ convolutions, producing $C'_2 \sim C'_5$. An extra $3 \times 3$ convolution with stride $= 2$ is applied to $C_5$ to generate $C'_5$. So that $C'_2 \sim C'_5$ are with the same number of channels but different resolutions. Subsequently, these features are passed through several stacked basic blocks for feature-level context aggregation. In each block, the input feature pyramid is processed by a top-down and a bottom-up aggregation path, in which interlevel residual connections [25], cross-level dense connections [26], and feature reweighting strategy are adopted.

Fig. 4 illustrates the detailed feature propagation strategy in basic blocks. In the top-down path, output features $C_{l,4}$ of each feature pyramid level are generated by fusing the features from the current level and all upper levels, then performing a
parameterized transform upon the fused features

\[ C_{i+1} = \text{Bottleneck} \left( C_i + \sum_{j=i+1}^{i_{\text{max}}} \left[ \text{Resize}(C_j) \cdot w_{i,j} \right] \right). \]  

(1)

Here, Bottleneck(·) denotes a ReLU activation layer followed by a 3 × 3 bottleneck [25] without activations. We observe that adopting only one nonlinearity before the bottleneck structure brings better performance. Resize(·) represents a max pooling layer, and \( w_{i,j} \) is a learnable reweighting term for aggregating features from level \( j \) to level \( i \). The weights \( w_{i,j} \) are vectors with lengths corresponding to their levels, the values are normalized from raw values using softmax by

\[ w_{i,j} = \frac{\exp(v_{i,j})}{\sum_{k=1}^{N_i} \exp(v_{i,k})}. \]  

(2)

where \( v_{i,j} \) denotes the raw weight vector and \( j \) is the index of each element. Using the normalization above can stabilize the learning process. Similar to the top-down path, bottom-up features \( C_{2\uparrow} \sim C_{6\uparrow} \) are computed by

\[ C_{i} = \text{Bottleneck} \left( C_i + \sum_{j=i+1}^{i-1} \left[ \text{Resize}(C_{j}) \cdot w_{i,j} \right] \right). \]  

(3)

where \text{Resize}(·) represents a bilinear interpolation layer and other notations are consistent with (1). We adopt residual connections to preserve the original features and prevent gradient vanishing. Leveraging the flexible architecture and feature reweighting strategy, DenseFPN has the capability to optimize the information flow of context aggregation in the feature domain during training.

### C. Spatial Context Pyramid

After aggregating feature maps across different levels, the feature pyramid remains to contain spatially local information, thus we introduce an SCP to further augment the features by...
learning the global spatial context within each level. Former attempts in this area [23], [45], [46], [47], [48] normally integrate several spatial or channel attention blocks into the backbone to enable global receptive fields. Some architectures of these blocks are presented in Fig. 5. Among these methods, NLSNet [23] is a general solution that computes pixel-wise correlations via embedded Gaussian for every spatial position, while squeeze-and-excitation network (SENet) [45] tackles the problem from the perspective of channel attentions. To combine the spatial and channel modulation abilities, a global context network (GCNet) [47] is proposed to learn a single attention map for an image. However, we observe that in remote sensing images with objects only covering small areas, this design may bring too much useless background information to objects. To tackle this problem, we propose adding an extra path on top of this structure to learn the informativeness of each pixel. Our core idea is that if the features of a pixel are informative enough, there’s not much need to aggregate features from other spatial positions. Such a soft reweighting strategy can effectively fuse global features while reducing information confounding.

The architecture of SCP is shown in Fig. 3(b). This module also has a pyramid structure and thus can be easily inserted after the backbone or neck. Each layer consists of a context aggregation block (CABlock) with a residual connection. The detailed design of this block is presented in Fig. 5(d). In each block, pixel-wise spatial context is aggregated by

$$Q_i^j = P_i^j + a_i^j \cdot \sum_{m=1}^{N_i} \frac{\exp(w_k P_i^{jm})}{\sum_{j=1}^{N_i} \exp(w_k P_i^{jm})} \cdot w_k P_i^{jm}$$  (4)

where $P_i$ and $Q_i$ denote the input and output feature maps of level $i$ in the feature pyramid, each contains $N_i$ pixels. $j, m \in \{1, N_i\}$ indicate the indices of each pixel. $w_k$ and $w_v$ are linear projection matrices for projecting the feature maps. In practice, we use $1 \times 1$ convolutions to perform the mapping. The formula above simplifies the widely used self-attention mechanism [67] by replacing the matrix multiplication between query and key with a linear projection, largely reducing the parameters and computational costs. Beyond GCNet, we apply $a_i$, a reweighting matrix with the same shape as $P_i$ and $Q_i$, to balance the extent of aggregating global spatial context for each pixel. This matrix can also be generated as simply as a linear projection from $P_i$ with sigmoid activation

$$a_i^j = \frac{1}{1 + \exp\left(-w_a P_i^j\right)}.$$  (5)

Similarly, $j \in \{1, N_i\}$ is the matrix index. Here, the output of the sigmoid function shall be regarded as the ratio of information that should be aggregated from a global context. We conducted extensive visualizations and experiments on the effectiveness of SCP. The results show that SCP tends to aggregate global context into areas with similar semantics, e.g., from small vehicle to large vehicle. Furthermore, we also observe that placing SCP after DenseFPN rather than incorporating CABlock in the backbone leads to better performance while preserving a smaller model size. This due to the sequential basis of multidomain global context aggregation. Detailed discussions will be conducted in Section IV.

### D. Hierarchical Region of Interest Extractor

Most two-stage object detection and instance segmentation methods lack sufficient attention on the RoI extractor, which may cause severe information loss since only a single scale is considered. The original purpose of this design is to enable large proposals to benefit from low-level features that capture higher localization accuracy, while small proposals can obtain more contextual information due to the larger receptive field of high-level features. We argue that such a hard assignment strategy may not be suitable for all proposals. Recent works [28], [68] also proved that simply computing the sum of RoI features cropped from all layers achieves slightly better performance, indicating that leveraging multi-scale features might alleviate the information loss.

In this work, we further deal with this problem by proposing an HRoIE to perform task-specific hierarchical feature fusion for each instance. This module is inserted after SCP as displayed in Fig. 3(c). Our hypothesis is that humans can easily perform object detection and segmentation because they focus
their attention on objects in a hierarchical manner. For example, when a person tries to classify an object, he or she would look at the object itself at first. If the object's appearance is not discriminative, the person would gradually look at the surrounding things to gain better information. On the opposite, when segmenting an object at pixel level, a human would look at the whole object to have a comprehensive understanding of its shape, and then zoomed-in view to obtain detailed boundary information for accurate segmentation. We implement the idea above by cropping the features of proposals $R_i$ from all feature pyramid levels in $Q_i$ using RoIAlign [9], and utilize several attention blocks to fuse the features hierarchically and adaptively per instance and task. As shown in Fig. 6, for each task, the RoI features are initialized from an empty matrix and combined with features from different levels in a hierarchical manner by

$$F_{b/m}^j = F_{b/m}^i + R_{b/m}^i \cdot \text{Sigmoid} \left( \left[ F_{b/m}^i \parallel R_{b/m}^i \right] \cdot w_i \right).$$

Here, $R_i$ denotes the cropped features at level $i$, $F_{b/m}^i$ and $R_{b/m}^i$ represent the aggregated RoI features at different levels, $w_i$ is the linear projection weight, and $\parallel$ means matrix concatenation at channel dimension. The procedure above computes the pixel-wise attention weights for feature aggregation, thus RoI features can be generated adaptively per instance and task. In practice, we adopt a bottom-up path for the object detection head and a top-down path for the mask prediction head.
In this section, we extensively evaluate the proposed method on iSAID, DIOR, NWPU VHR-10, and HRSID datasets, in which the first three datasets are based on optical remote sensing images, while the last one is for SAR images. Note that DIOR only provides bounding box annotations, thus we only evaluate the object detection performance on this dataset.

### A. Datasets and Evaluation Metrics

1) **iSAID** [15] is a large-scale dataset for instance segmentation in aerial images. The images in iSAID are inherited from DOTA [14], which is popular for oriented object detection. It contains 15 classes of 655,451 instances in 2806 images, with all the objects independently annotated from scratch. The spatial resolutions of images are in a large range between 800 and 13,000. Following previous works [22], [51], [53], we split them into 800 × 800 patches with a stride of 200 for fair benchmarking with existing methods. When conducting detailed comparisons and ablation studies, we adopt a smaller patch size of 512 × 512 with a 128 stride to reduce the training cost. The abbreviations of classes are airplane (AI), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track (GT) field, harbor (HA), bridge (BR), and vehicle (VE).

2) **DIOR** [16] is a complex aerial images dataset labeled by horizontal and oriented bounding boxes. It contains 23,463 images with 190,288 instances, covering 20 object classes. Object sizes in DIOR have severe interclass and intraclass variabilities. The complexity of this dataset is also reflected in different imaging qualities, weather, and seasons. The abbreviations of classes are airplane (AL), airport (AR), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DA), expressway service (ES) area, expressway toll (ET) station, golf field (GC), ground track (GT) field, harbor (HA), overpass (OV), ship (SH), stadium (ST), storage tank (SA), tennis court (TC), train station (TS), vehicle (VE), and wind mill (WM).

3) **NWPU VHR-10** [13] was originally designed for object detection in aerial images and has been enriched with instance-level mask annotations [64]. It contains ten object categories in 800 high-resolution images, among which 650 are positive and 150 are negative without any objects of interest. Since this dataset has no official train/test split available, we randomly select 70% of the images for training while the others for testing. Our splits will be released for a fair comparison with the following methods. The abbreviations of classes are airplane (AI), ship (SH), storage tank (ST), baseball diamond (BD), tennis court (TC), basketball court (BC), ground track (GT) field, harbor (HA), bridge (BR), and vehicle (VE).

4) **HRSID** [17] is a recently introduced dataset for ship detection and segmentation in SAR images. This dataset contains a total of 5604 high-resolution SAR images with 16,951 ship instances. All the instances in this dataset are annotated with pixel-level masks. The spatial resolutions of the images are 0.5, 1, and 3 m.

We follow the standard evaluation metric that utilizes mean average precision (mAP) to measure the detection and segmentation performances. A prediction is considered a true positive (TP) when the bounding box or mask of the object has an intersection over union (IoU) with its corresponding ground truth greater than a threshold \( \theta_{iou} \), and the predicted class label is correct. For iSAID, NWPU VHR-10, and HRSID datasets, we compute the mean of mAPs under \( \theta_{iou} \) ranging from 0.05 to 0.95. For the DIOR dataset, only the mAPs under \( \theta_{iou} = 0.5 \) are considered following the original paper. Aside from the accuracy metrics, we also evaluate the efficiencies of the models using the number of parameters (Params) and floating-point operations (FLOPs), measuring the storage and computational efficiencies, respectively. We consider these two metrics since they are both hardware-irrelevant and have been widely adopted in previous studies.

### B. Implementation Details

We choose Mask R-CNN [9], Faster R-CNN [7], and RetinaNet [8] with ResNet-50 [25] backbone as our baselines for different scenarios. The backbone is pretrained on ImageNet [73] and fine-tuned with the detectors. All the parameters in the first stage are frozen after pretraining. If not specified, five basic blocks are included in all DenseFPN modules. In order to stabilize the training process, synchronized batch normalizations (SyncBN) [74] are used among intermediate layers. When testing, we adopt Soft-NMS [75] to suppress the duplicate results with IoU larger than 0.5, and a maximum of 1000 predictions would be made for each image, since most objects are heavily overlapped in remote sensing images.

We use a stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, momentum 0.9, and weight decay 0.0001 to learn the parameters for all models. Each training batch contains 8 images. On the iSAID dataset, we follow the standard 1× training schedule that drops the learning rate by 0.1 at epoch 8 and 11 and stop training at epoch 12 for efficient parameter tuning and ablation studies on the validation set. When benchmarking on the test set, we train our model on both training and validation sets independently annotated from scratch. The spatial resolutions of images are in a large range between 800 and 13,000.

### IV. EXPERIMENTS

In this section, we extensively evaluate the proposed method on iSAID, DIOR, NWPU VHR-10, and HRSID datasets, in which the first three datasets are based on optical remote sensing images, while the last one is for SAR images. Note that DIOR only provides bounding box annotations, thus we only evaluate the object detection performance on this dataset.

#### TABLE V

| Method          | Backbone       | AP\(_0\) | AP\(_m\) |
|-----------------|----------------|---------|---------|
| RetinaNet [8]   | ResNet-101     | 59.8    | –       |
| Faster R-CNN [7]| ResNet-101     | 63.9    | –       |
| Cascade R-CNN [52]| ResNet-101 | 66.8    | –       |
| HRDSNet [17]   | HRFPN-W40      | 69.4    | –       |
| Mask R-CNN [9] | ResNet-101     | 65.4    | 54.3    |
| MS R-CNN [54]  | ResNet-101     | 64.9    | 54.4    |
| Cascade R-CNN [52]| ResNet-101 | 67.6    | 54.7    |
| HTC [10]       | ResNet-101     | 68.4    | 55.4    |
| HQ ISNet [64]  | HRFPN-W40      | 66.7    | 54.2    |
| GCBA [66]      | ResNet-101     | 69.4    | 57.3    |
| CATNet         | ResNet-50      | 71.7    | 58.2    |
| CATNet + Aug.  | ResNet-50      | 73.3    | 59.6    |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
To justify the effectiveness of our method, class-wise instance segmentation results are also provided in Table II. Our model obtains better performances in most categories, including “storage tank,” “baseball diamond,” and “tennis court.” In other categories, our method is still comparable with most existing works.

D. Object Detection on DIOR Dataset

We then evaluate our approach to the DIOR dataset. Note that this dataset only provides bounding box annotations for objects, we only validate the object detection performance on it. Such a comparison can also demonstrate the generalization ability of the modules—they could potentially benefit all the dense prediction tasks including object detection and instance segmentation. Table III presents the class-wise performance comparison with existing methods. Our methods are built upon two meta algorithms for one-stage and two-stage object detection, which are RetinaNet [8] and Faster R-CNN [7]. Note that one-stage models do not need RoI extractors so CATNet* only contains DenseFPN and SCP. The comparison shows that our proposed method outperforms all the previous methods, with or without data augmentations. Moreover, the performances of our model with ResNet-50 backbones are even better than the previous state-of-the-art with ResNet-101 by a noticeable margin. This demonstrates the effectiveness and generalization ability of the proposed modules.

E. Instance Segmentation on NWPU VHR-10 Dataset

To validate the significance of our method, we also conduct instance segmentation experiments on an extra NWPU VHR-10 Dataset, and the performance comparisons are shown in Table IV. The first group includes one-stage methods while the second group contains two-stage methods. Following previous works, we only report the average precisions for mask predictions to focus on the instance segmentation task. The proposed CATNet can still perform better than all the previous approaches with heavier backbones.

F. Instance Segmentation on HRSID Dataset

Beyond optical remote sensing images, we also evaluate our model on the more challenging SAR images on the HRSID dataset. The results are presented in Table V, in which the first group contains object detection models while the second group includes two-stage methods. Following previous works, we only report the average precisions for mask predictions to focus on the instance segmentation task. The proposed CATNet still performs better than all the previous approaches with heavier backbones.

G. Visualizations

We also provide some visualizations to conduct an in-depth study on the significance and effectiveness of our method.

1) Context Aggregation Weights: To demonstrate the effectiveness and significance of SCP, we visualize the context aggregation weights in GCNet [47] and SCP in Fig. 7. Each row represents the weights of all classes that are aggregated.
TABLE VII
Detailed Comparisons on iSAID val. set. The Baseline Model Is Mask R-CNN With ResNet-50-FPN as Backbone and Neck. (a) Types of Multi-Scale Feature Propagation Modules. (b) Types of Spatial Context Modules. (c) Orders of the Proposed Modules. (d) Types of Region of Interest Extractors. (e) Ablation Study and Efficiency Comparisons

(a)

| Method       | Chn. | Depth | APb, APm_1 | PLOPs |
|--------------|------|-------|-------------|-------|
| Baseline     | 256  | -     | 43.8 36.5   | 44.05M | 114.96G |
| PAFPN [28]   | 256  | -     | 44.4 37.0   | 47.59M | 121.31G |
| HRFPN [69]   | 256  | -     | 44.1 36.8   | 44.64M | 129.09G |
| CARAFE [70]  | 256  | -     | 44.1 36.9   | 49.65M | 115.71G |
| BFP [71]     | 256  | -     | 44.2 37.0   | 44.31M | 115.23G |
| AugFPN [72]  | 256  | -     | 44.3 36.8   | 45.82M | 115.03G |
| TridenNet [38] | 256  | -     | 42.6 32.9   | 35.28M | 838.99G |
| RFP [41]     | 256  | -     | 45.3 37.4   | 68.80M | 153.88G |
| BiFPN [44]   | 256  | -     | 44.1 36.8   | 45.89M | 112.49G |

(b)

| Method                  | Position | Reduced | APb, APm_1 | PLOPs | FLOPs |
|-------------------------|----------|----------|-------------|-------|-------|
| Baseline                | -        | -        | 43.8 36.5   | 44.05M | 114.96G |
| GABlock [48]            | C3 → C5  | -        | 44.5 37.0   | 43.25M | 156.52G |
| CCBlock [46]            | C3 → C5  | -        | 44.1 36.7   | 68.97M | 132.43G |
| NLBlock [23]            | C3 → C5  | -        | 44.2 37.0   | 83.93M | 142.92G |
| GCBlock [47]            | C3 → C5  | -        | 44.3 37.1   | 54.03M | 121.97G |
| CBlock (ours)           | C3 → C5  | -        | 44.5 37.6   | 49.04M | 118.47G |
| GABlock [48]            | P2 → P6  | -        | 44.9 36.9   | 45.60M | 119.20G |
| CCBlock [46]            | P2 → P6  | -        | 43.6 36.2   | 44.46M | 116.76G |
| NLBlock [23]            | P2 → P6  | 1        | 44.0 36.9   | 45.36M | 120.71G |
| GCBlock [47]            | P2 → P6  | -        | 44.2 37.2   | 47.41M | 114.97G |
| SCP (ours)              | P2 → P6  | 1        | 44.6 37.7   | 44.71M | 116.41G |

(c)

| Module Order                      | APb, APm_1 | #Params | FLOPs |
|-----------------------------------|-------------|---------|-------|
| Backbone (w. CBlock)              | 45.4 37.5   | 72.95M  | 155.67G |
| Backbone → SCP → DFPN             | 45.5 37.6   | 53.64M  | 150.09G |
| Backbone → DFPN → SCP             | 45.9 37.8   | 53.64M  | 150.09G |

(d)

| Method                     | Direction | APb, APm_1 | #Params | FLOPs |
|----------------------------|-----------|-------------|---------|-------|
| Baseline                   | -         | 43.8 36.5   | 44.05M  | 114.96G |
| GroIoE [68]                | -         | 44.0 36.9   | 47.54M  | 157.92G |
| SUM                        | -         | 43.9 36.7   | 44.05M  | 114.96G |
| CONCAT                     | -         | 44.0 36.7   | 48.34M  | 188.18G |
| AFFINITY                   | -         | 43.9 36.8   | 46.15M  | 186.97G |
| HRoIE (ours)              | + & +     | 44.2 37.1   | 45.10M  | 151.00G |
| HRoIE (ours)              | + & +     | 44.1 37.0   | 45.10M  | 151.00G |
| HRoIE (ours)              | + & +     | 44.3 36.9   | 45.10M  | 151.00G |
| HRoIE (ours)              | + & +     | 44.4 37.2   | 45.10M  | 151.00G |

2) Training Process: In order to study the model training process, we also visualize the overall training losses in Fig. 8. All the models are trained under the 1× schedule as stated above, and the training losses of the last 8 epochs are plotted. The solid curves indicate the smoothed losses of different models, while the light-colored areas denote the standard deviations. We start from our baseline (i.e., Mask R-CNN [9]) and gradually incorporate the three modules to study the overall converge processes. The comparison indicates that all the proposed modules, especially DenseFPN, contribute to faster learning and higher performance, as the loss values decline faster and can reach lower minimums. Notably, the full version of CATNet converges about 30% faster than the baseline while reaching a 10% lower final loss.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
3) Model Predictions and Attention Maps: Fig. 9 shows some qualitative results on iSAID, DIOR, and HRSID datasets. Each image patch is visualized by its feature aggregation weights in SCP (top) and the final object detection or instance segmentation results (bottom). The feature aggregation weights indicate that our method focuses more on areas that contain objects rather than pure background, and aggregates them into regions with poor features. Furthermore, we also visualize the detailed comparisons on mask prediction for hard cases in Fig. 10. Our model can better handle small, blurry, and clustered objects, which are common in both optical and other kinds of remote-sensing images. These results demonstrate that our method can effectively detect and segment objects accurately in multiple scenes.

4) Discriminativeness of Features: To verify that the proposed method can effectively enhance the inadequate features and make them more discriminative, we conducted two studies on the hidden states. First, we visualize the RoI features for some common objects in the iSAID dataset in Fig. 1. The feature maps are obtained by randomly selecting a channel from the RoI features. By aggregating context gradually from different domains, clearer object boundaries can be seen from the feature maps. Second, we randomly sampled 1000 objects from each category, extracted their average pooled RoI features, and plotted them on a low dimensional manifold using t-SNE [77]. The results are shown in Fig. 11. In our baseline method, features of some categories are heavily confounded, since the top-view and low-quality images cannot provide enough information for perception. By introducing multi-scale context aggregation, our method makes the features much more discriminative, as the clusters are more isolated, while features for semantically related categories (e.g., large vehicle & small vehicle and helicopter & plane) remain close. We also compared the averaged class-wise variances of features, proving that the enhanced feature clusters are more cohesive and, thus more discriminative.

H. Detailed Comparisons and Ablation Study

To study the effectiveness of the proposed modules individually, we conduct extensive experiments to compare different settings with some representative methods. All the experiments are performed on iSAID val set using the standard 1x training recipe with 512 × 512 patches inputs. To verify the lightweight characteristics of our method, which means the modules merely introduce negligible extra parameters and computations to the baselines, Params, and FLOPs are reported to compare the space and time complexities of models.

1) Plug-and-Play Abilities: We first verify the plug-and-play abilities of the proposed modules by incorporating them into different instance segmentation models. Aside from Mask R-CNN, we choose Cascade R-CNN [52], Mask Scoring R-CNN [54], and PointRend [11] as base models, and sequentially incorporate DenseFPN, SCP, and HRoIE into these frameworks. The class-wise instance segmentation performances are reported in Table VI. We observe that for both object detection and instance segmentation tasks, our methods can steadily boost performances. Among the three proposed modules, DenseFPN provides the most significant improvement on mAPs, while SCP and HRoIE also bring considerable gains with few extra parameters. We claim that the proposed modules shall be compatible with any CNN-based instance segmentation pipelines.

2) Types of Multi-Scale Feature Propagation Modules:

We then study the accuracies and efficiencies of different multi-scale feature propagation modules. Table VII(a) presents a detailed comparison among different feature pyramid networks. Compared with existing representative methods, DenseFPN works distinctly better on both object detection and instance segmentation tasks with fewer computational costs. It is worth noting that a single-layer DenseFPN can already obtain considerable gains from the baseline with fewer or comparable computational costs. Simply stacking more basic blocks in DenseFPN can further boost the performance, indicating its capability and flexibility of model scaling.

3) Types of Spatial Context Modules: Table VII(b) shows the comparison among multiple spatial context modules. Compared with the baseline, NLNet [23] can effectively bring higher performance with a high computational cost. GCNet [47] solves the problem of computational complexity, but leads to an extra information confounding problem. With the help of reweighting context, our proposed CABlock steadily outperforms GCNet using different channel reduction rates. In previous studies, these spatial attention modules are used as plugins of backbones to enhance their global context modeling abilities for image classification. However, we observe that such a strategy is suboptimal for some modules, due to the large computational consumption with limited performance gains, as can be seen in the table. In this work, we propose to move these modules from the backbone to the layers after FPN, so that the input channels can be largely reduced, while the semantics in the feature maps are more consistent. The comparison of different designs of CABlocks (i.e., used as backbone plugins or incorporated into SCP)
Fig. 9. Qualitative results on multiple datasets. We visualize the spatial context attention maps in SCP (odd rows) and the final outputs (even rows) on iSAID, DIOR, and HRSID datasets. The results demonstrate that our model has the ability to perform object detection and instance segmentation accurately in both optical remote sensing images and SAR images. (a) iSAID (instance segmentation in aerial images). (b) DIOR (object detection in aerial images). (c) HRSID (instance segmentation in SAR images).

Fig. 10. Detailed comparisons on mask predictions for small, blurry, and clustered objects. The image patches are from the iSAID val set.

Fig. 11. Visualization of the sampled RoI features from different models using t-SNE. The averaged class-wise variances are also reported.

shows that the architecture of SCP brings more performance gains with negligible computational costs.

4) Order of DenseFPN and SCP: Since both DenseFPN and SCP are pyramid structures for multi-scale features, the order of these two modules could be changed. We compare the different orders of these modules in Table VII(c). The first row means using CABlocks are backbone plugins and adopting DenseFPN after backbone. This design is the de-facto standard for most existing works. However, it can only lead to suboptimal performances. The second row places SCP before DenseFPN so that spatial context is aggregated before feature context. This order still cannot perform well enough. The last row shows our final design, which places SCP after DenseFPN, achieving the best performance on both object...
be a better solution to handle the global context in multiple domains. We foresee the potential of incorporating multi-scale context aggregation in transformer-based models.

REFERENCES

[1] L. Liu, Z. Yang, G. Li, K. Wang, T. Chen, and L. Lin, “Aerial images meet crowdsourced trajectories: A new approach to robust road extraction,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 7, pp. 3308–3322, 2022.
[2] G. Ding, D. Yang, T. Wang, S. Wang, and Y. Zhang, “Crowd counting via unsupervised cross-domain feature adaptation,” IEEE Trans. Multimedia, vol. 25, pp. 4665–4678, 2022.
[3] G. Ding, M. Cui, D. Yang, T. Wang, S. Wang, and Y. Zhang, “Object counting for remote-sensing images via adaptive density map-assisted learning,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022.

V. CONCLUSION AND FUTURE WORK

In this work, we provided an in-depth analysis of global context modeling in remote sensing images and proposed CATNet, a novel framework that leverages three lightweight plug-and-play modules, i.e., DenseFPN, SCP, and HRoIE, to aggregate the global visual context in feature, spatial, and instance domains, respectively. It has been demonstrated that the collaboration among these modules can effectively enhance the discriminative object features for promoting both object detection and instance segmentation performances. We expect that the new understanding of the global context and the design of the proposed modules will benefit future research in this area. Below we discuss the limitations and future work in this direction.

The motivation of this work is to design lightweight plug-and-play modules for existing instance segmentation pipelines. Each proposed module aggregate global context in one domain only. Such a strategy may be suboptimal since cross-domain contexts could be correlated with each other. Therefore, a more unified framework shall be designed from scratch to mitigate the limitations of existing instance segmentation models. The recently introduced transformers [67] might...
Huifang Li (Member, IEEE) received the B.S. degree in geographical information science from the China University of Mining and Technology, Xuzhou, China, in 2008, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2013.

She is currently a Professor with the School of Resources and Environmental Science, Wuhan University. She focuses on the study of urban remote sensing, including complex radiometric correction, image segmentation, scene classification, and urban environment analysis.

Chao Hu received the B.S. degree in geographic information science and the M.S. degree in geomatics engineering from Wuhan University, Wuhan, China, in 2019 and 2021, respectively.

His research interests include remote sensing, deep learning, and object detection.

Ye Liu received the B.S. degree in geographical information science and the B.E. degree in computer science from Wuhan University, Wuhan, China, in 2020. He is currently pursuing the Ph.D. degree with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China.

He was a Research Assistant at the State University of New York, Buffalo, NY, USA, and the Chinese University of Hong Kong, Shenzhen, China, in 2019. He is serving as an Associate at the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS), Boston, MA, USA, and a Research Intern at Tencent ARC Lab, Shenzhen, China. His research interests include computer vision, multimedia, and urban computing, with a particular focus on foundation models and vision-language understanding.

Shuang Luo received the B.S. degree in geographic information systems from the China University of Petroleum, Qingdao, China, in 2015, and the Ph.D. degree in cartography and geographic information engineering from Wuhan University, Wuhan, China, in 2021.

He is currently an Engineer with Changjiang Spatial Information Technology Engineering Company Ltd., Wuhan. His research interests include deep learning and remote sensing image processing.

Yan Luo received the bachelor’s degree in remote sensing from Wuhan University, Wuhan, China, in 2019. She is currently pursuing the Ph.D. degree with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China.

She has undertaken internships at Microsoft Research Asia (MSRA), Beijing, China, and Didi, Beijing. Concurrently, she serves as a Research Fellow at the Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Her research focuses on the convergence of technology and social impact, with a particular emphasis on leveraging big data to drive societal innovation. In addition to her academic pursuits, she has also taken on the role of founder, establishing a startup in alignment with her passion for advancing impactful technological solutions.

Chang Wen Chen (Fellow, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 1983, the M.S.E.E. degree from the University of Southern California, Los Angeles, CA, USA, in 1986, and the Ph.D. degree from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1992.

He was with the Faculty of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA, from 1992 to 1996, and the Faculty of Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA, from 1996 to 2003. He was the Allen Henry Endow Chair Professor at the Florida Institute of Technology, Melbourne, FL, USA, from 2003 to 2007. He served as an Empire Innovation Professor of computer science and engineering at the University at Buffalo, State University of New York, Buffalo, New York, USA, from 2008 to 2021. From 2017 to 2020, he served as the Dean of the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. He is currently the Chair Professor of visual computing at The Hong Kong Polytechnic University, Hong Kong SAR, China. His research expands a broad range of topics in multimedia communication, the Internet of Video Things, multimedia systems, image/video processing, machine learning, and multimedia signal processing.

Dr. Chen has been an SPIE Fellow since 2007 and a member of Academia Europaea since 2021. He and his students have received nine best paper awards or best student paper awards. He has also received several research and professional achievement awards, including the Sigma Xi Excellence in Graduate Research Mentoring Award in 2003, the Alexander von Humboldt Research Award in 2009, the University at Buffalo Exceptional Scholar for Sustained Achievement Award in 2012, the State University of New York System Chancellor’s Award for Excellence in Scholarship and Creative Activities in 2016, and the Distinguished ECE Alumni Award from University of Illinois at Urbana–Champaign in 2019. He served as the Conference Chair for several major IEEE, ACM, and SPIE conferences related to multimedia, video communications, and signal processing, including ACM MM2020 and ACM MM2018 in recent years. He was the Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2006 to 2009 and IEEE TRANSACTIONS ON MULTIMEDIA from 2014 to 2016. He has been an Editor of several other major IEEE TRANSACTIONS and journals, including PROCEEDINGS OF THE IEEE and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS.