CATNet: Context AggregaTion Network for Instance Segmentation in Remote Sensing Images

Ye Liu, Huifang Li, Member, IEEE, Chao Hu, Shuang Luo, Huanfeng Shen, Senior Member, IEEE, 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 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 directly applied to top-view remote sensing images. Through careful analysis, we observe that the challenges mainly come from 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 inter-level residual connections, cross-level dense connections, and feature re-weighting 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. We carry out extensive evaluation of the proposed scheme on the challenging iSAID, DIOR, NWPU VHR-10, and HRSID datasets. The evaluation results demonstrate that the proposed approach outperforms state-of-the-arts with similar computational costs. Code is available at https://github.com/yeliudev/CATNet.

Index Terms—Instance Segmentation, Object Detection, Global Context Aggregation, Self Attention

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, including urban planning, resource detection, and environmental monitoring. 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 instance level, has been widely used to extract fine-grained object information from both optical 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 seg-
The challenges of performing instance segmentation in remote sensing images lead to complicated object patterns while low contrast and cluttered background bring interfering information from the background. These challenges are due to the lack of discriminative object features in remote sensing images. Here, scale variation, arbitrary orientation, and clustered distribution lead to complicated object patterns while low contrast and cluttered background 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 are due to the lack of discriminative object features in remote sensing images. That is, visual appearances of individual objects in remote sensing images are not informative enough for directly adopting existing schemes to perform instance segmentation.

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 [3–5], 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 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 [14–17] only regard context as spatial correlations. We expand and explicitly disentangle the concept of context into three domains, i.e. feature, spatial, 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 correlative objects, and 3) adaptively refining intermediate representations for each instance and task. These three different domains is 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. Fig. 1 shows a glimpse of object features processed by these modules. In 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. It is based on this analysis that 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 inter-level residual connections [18], cross-level dense connections [19], and feature re-weighting strategy to enable the module to learn its optimal feature propagation manner. In 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 details. The demand for different sizes of receptive fields also varies among instances. Hence, we introduce 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 fuse the features level-by-level in a hierarchical manner. 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 paper are three-fold:

- The expansion and explicit disentangling the concept of context into feature, spatial, and instance domains have resulted in superior performance in remote sensing image segmentation. To the best of our knowledge, this is the first work that considers global visual context beyond spatial dependencies.
- 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.
- The proposed scheme has been tested on broad variety of datasets, including iSAID, DIOR, NWPU VHR-10, and HRSID, and new state-of-the-art performance has been obtained with similar computational cost.
The rest of this paper 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 and in-depth analysis on public datasets. Finally, concluding remarks are summarized in Section V.

II. RELATED WORKS

A. Instance Segmentation in Remote Sensing Images

Instance Segmentation is a challenging and broadly studied problem in computer vision. Similar to object detection [1], [2], the majority of instance segmentation approaches can be divided into two schemes, namely single-stage methods and two-stage methods. As a straightforward design, single-stage methods [6]–[8], [20], [21] adopt the bottom-up strategy that performs semantic segmentation at image level, and further separates individual objects using clustering or metric learning. These methods often possess considerable efficiency, but are largely restricted by their localization accuracy. Compared to this paradigm, two-stage methods [3]–[5], [22] separate the segmentation pipeline into two phases, i.e. region proposal generation and task-specific post-processing, resulting 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 [23]–[29] try to tackle the problem of object detection in remote sensing images, but they do not pay particular attention to instance segmentation. Our proposed context aggregation strategy can be integrated into both single-stage and two-stage methods, while HRoIE is not used in single-stage methods since it’s not necessary to crop the feature maps. Further experimental results demonstrate that our modules have the capability to steadily boost the performances of existing methods.

B. Global Context Modeling

One of the most representative properties of convolutional neural networks (CNNs) is local dependency modeling. Multiple convolution layers can be stacked together to increase the receptive fields, learning high resolution localization features at low levels and low resolution semantic features at high levels. The key to handle the heterogeneity is to propagate multi-level features from the backbone appropriately. A common practice in earlier works [22], [30]–[32] is to adopt the strategy of multi-scale feature propagation, but these methods only propagate the features along fixed paths. We claim that flexible information flows may reduce the information confounding and better aggregate multi-scale features. Besides, long-range spatial dependency modeling has also been proved to be effective for dense prediction tasks [16], [33]. As a pathbreaking work, non-local neural network (NLNet) [16] shows that global spatial context can be aggregated by computing pixel-level pairwise correlations, but it also suffers from the problem of high computational cost. Some extensions of NLNet [34]–[36] tend to address this problem by simplifying the correlation estimation. Promising results have been achieved by these works, however, all these methods only consider the concept of context as long-range spatial correlations, ignoring the global dependencies in feature and instance domains.

Fig. 3. Overall architecture of the proposed framework. The process of global context aggregation is realized by three modules, namely a) dense feature pyramid network, b) spatial context pyramid, and c) hierarchical region of interest extractor. 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.

Fig. 4. Detailed procedure of multi-scale feature propagation in dense feature pyramid network. ⊕ and ⊖ denote element-wise addition and ReLU activation respectively. The cross-level feature maps are adaptively combined using feature re-weighting strategy.
III. The Proposed Approach

In this section, we introduce our approach on global context aggregation. As shown in Fig. 3, the entire framework can be divided into three sub-modules, namely DenseFPN, SCP, and HRoIE. These modules are between the backbone and prediction heads, aiming to aggregate global context information from different domains.

A. Overview

Given an image \(x\) and a set of object categories of interest \(S = \{1, ..., N\}\), the task of instance segmentation aims to detect and segment all the objects in \(x\), where they belong to whichever the pre-defined categories. The output of instance segmentation would be a collection of tuples \(T = \{(b, m, s)\}\), where \(b \in \mathbb{R}^{3}\) 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 [3], 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 the following sections.

B. Dense Feature Pyramid Network

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

The basic architecture of our proposed DenseFPN is presented in Fig. 3 (a), where each node represents a feature map and black lines stand for information flows. This module takes \(C_2 \sim C_5\) as inputs and firstly 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 \(C'_6\). So that \(C'_2 \sim C'_6\) are with the same number of channels but different resolutions. Subsequently, these features are passed through a number of 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 inter-level residual connections [18], cross-level dense connections [19], and feature re-weighting strategy are adopted.

Fig. 4 illustrates the detailed feature propagation strategy in basic blocks. In the top-down path, output features \(C_{i+4}\) 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.

\[
C_{i+4} = \text{Transform}(C_i + \sum_{j=1}^{\max \ell_{max}} [\text{Resize}(C'_j) \cdot w_{ij}]) \quad (1)
\]

Here, \(\text{Transform}(\cdot)\) denotes a ReLU activation layer followed by a \(3 \times 3\) bottleneck [18] without activations. We observe that adopting only one nonlinearity before the bottleneck structure brings better performance. \(\text{Resize}(\cdot)\) represents a max pooling layer used to unify the resolutions of feature maps, and \(w_{ij}\) is a learnable re-weighting term for aggregating features from level \(j\) to level \(i\). The weights \(w_{ij}\) are vectors with lengths correspond to their levels, the values are normal-
design may bring too much useless background information to remote sensing images that objects only cover small areas, this a single lightweight module. However, we observe that in approach that combines NLNet [16] and SENet [33] into backbone to enable global receptive fields. Some architectures these blocks are presented in Fig. 5. Among these methods, further augment the features by learning global spatial context within each level. Former attempts in this area [16], [33]–

C. Spatial Context Pyramid

After aggregating feature maps across different levels, the feature pyramid remains containing spatially local information, thus we introduce spatial context pyramid (SCP) to further augment the features by learning global spatial context within each level. Former attempts in this area [16], [33]–[36] normally integrate several visual attention blocks into the backbone to enable global receptive fields. Some architectures of these blocks are presented in Fig. 5. Among these methods, global context network (GCNet) [36] is a simple but effective approach that combines NLNet [16] and SENet [33] into a single lightweight module. However, we observe that in remote sensing images that objects only cover small areas, this design may bring too much useless background information to objects. To tackle this problem, we propose to add 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 re-weighting strategy can effectively fuse local and global features while reducing information confounding.

The architecture of SCP is shown in Fig. 3(b). This module also has a pyramid structure, 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

\[
Q_i^l = P_i^l + a_i^l \cdot \sum_{j=1}^{N_l} \left[ \sum_{m=1}^{N_m} \frac{\exp(w_k P_i^m)}{\sum_{n=1}^{N_n} \exp(w_n P_i^n)} \cdot w_v P_i^l \right]
\]

where \(P_i\) and \(Q_i\) denote the input and output feature maps of level \(i\) in the feature pyramid, each contains \(N_l\) pixels. \(j, m \in \{1, N_l\}\) indicate the indices of each pixel. \(w_k\) and \(w_v\) are linear transform 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 [39] by replacing the matrix multiplication between query and key with a linear transform, largely reducing the parameters and computational costs. Beyond GCNet, we apply \(a_i\), a re-weighting 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 simple as a linear transform from \(P_i\) with softmax normalization.

\[
a_i^l = \frac{\exp(w_a P_i^j)}{\sum_{n=1}^{N_n} \exp(w_a P_i^n)}
\]

Similarly, \(j, n \in \{1, N_l\}\) are the matrix indices. We conduct extensive experiments on the effectiveness of \(a_i\). Fig. 7 visualizes the comparison of feature aggregation weights between GCNet and SCP. The results show that our model tends to
aggregate object features from the same category. Please refer to Section LV-H for detailed discussions.

D. Hierarchical Region of Interest Extractor

Most two-stage object detection and instance segmentation methods lack of sufficient attention on the RoI extractor, which may cause severe information loss since only a single scale is considered. The original design of this 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 this hard assignment strategy may not be suitable for all proposals. Recent works [22], [40] also prove that simply computing the sum of RoI features cropped from all layers achieves slightly better performance.

In this work, we further deal with this problem by proposing hierarchical region of interest extractor (HRoIE) to perform task-specific RoI 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 focalize 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 zoom in repeatedly to obtain more 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 [3], and utilize several attention blocks to fuse the features 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

\[
F_{b/m}^i = F_{b/m}^i + R_i^j \cdot \text{Sigmoid}([F_{b/m}^i \parallel R_i^j] \cdot w_i) \quad (6)
\]

Here, \( R_i \) denotes the cropped features at level \( i \), \( F_{b/m}^i \) and \( F_{b/m}^j \) represent the aggregated RoI features at different levels, \( w_i \) is the linear transform 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 object detection head and a top-down path for mask prediction head.

IV. EXPERIMENTS

In this section, we extensively evaluate the proposed method on iSAID, DIOR, NWPU VHR-10, and HRSID datasets. The modules are firstly evaluated under instance segmentation task on iSAID dataset, followed by object detection task on DIOR and NWPU VHR-10 datasets, to demonstrate the effectiveness in optical remote sensing images. HRSID dataset is also used to verify the capability of generalization in SAR images.

A. Datasets and Evaluation Metrics

**iSAID** [11]: iSAID is a large-scale dataset for instance segmentation in aerial images. All the images in iSAID are inherited from DOTA [10], which is popular for oriented object detection. It contains 15 classes of 655,451 instances in 2,806 images, with all the objects independently annotated from scratch. The spatial resolutions of images are in a large range between 800 and 13,000. We split them into 512 × 512 patches during training and testing. The abbreviations of classes are SH - ship, ST - storage tank, BD - baseball diamond, TC - tennis court, BC - basketball court, GT - ground track field, BR - bridge, LV - large vehicle, SV - small vehicle, HE - helicopter, SP - swimming pool, RO - roundabout, SB - soccer ball field, PL - plane, and HA - harbor.

**DIOR** [12]: DIOR is a complex aerial images dataset labeled by only horizontal bounding boxes. It contains 23,463 images with 190,288 instances, covering 20 object classes. Object sizes in DIOR have severe inter-class and intra-class variabilities. The complexity of this dataset is also

| Method                 | AP_b | AP_m | SH | ST | BD | TC | BR | LV | SV | HE | SP | RO | SB | PL | HA |
|------------------------|------|------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Mask R-CNN [3]         | 43.8 | 36.5 | 47.2 | 36.8 | 54.0 | 76.7 | 32.3 | 32.7 | 22.9 | 39.4 | 15.3 | 4.9 | 30.1 | 37.3 | 46.3 | 48.1 | 26.8 |
| + DenseFPN             | 45.5 | 37.7 | 48.8 | 37.7 | 55.6 | 77.5 | 35.8 | 33.4 | 23.8 | 39.6 | 16.3 | 5.7 | 29.9 | 39.8 | 46.7 | 49.2 | 28.4 |
| + SCP                  | 45.9 | 37.8 | 48.5 | 38.2 | 55.8 | 77.8 | 38.6 | 32.5 | 25.0 | 39.8 | 16.2 | 5.9 | 28.9 | 39.3 | 46.4 | 49.6 | 27.9 |
| + HRoIE                | 46.2 | 38.5 | 48.4 | 38.4 | 55.4 | 77.8 | 40.4 | 35.3 | 25.8 | 39.5 | 16.4 | 6.3 | 31.0 | 38.3 | 47.8 | 50.2 | 29.0 |

**Cascade R-CNN** [37]

| Method                 | AP_b | AP_m | SH | ST | BD | TC | BR | LV | SV | HE | SP | RO | SB | PL | HA |
|------------------------|------|------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| + DenseFPN             | 46.8 | 38.2 | 49.6 | 38.9 | 55.1 | 78.8 | 34.4 | 34.4 | 25.1 | 41.4 | 16.8 | 4.8 | 31.2 | 41.0 | 45.9 | 49.9 | 28.4 |
| + SCP                  | 47.4 | 38.4 | 49.5 | 39.0 | 55.9 | 77.9 | 35.6 | 34.8 | 25.3 | 40.7 | 16.7 | 5.7 | 31.1 | 41.2 | 47.5 | 49.7 | 28.8 |
| + HRoIE                | 47.8 | 38.9 | 49.1 | 39.2 | 55.7 | 78.3 | 37.4 | 35.5 | 25.4 | 41.2 | 16.8 | 8.7 | 31.2 | 41.8 | 47.1 | 50.0 | 29.9 |

**MS R-CNN** [38]

| Method                 | AP_b | AP_m | SH | ST | BD | TC | BR | LV | SV | HE | SP | RO | SB | PL | HA |
|------------------------|------|------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| + DenseFPN             | 44.0 | 37.8 | 48.6 | 36.6 | 54.9 | 77.0 | 39.6 | 35.2 | 22.5 | 39.9 | 15.5 | 9.7 | 30.1 | 39.2 | 46.2 | 49.9 | 29.7 |
| + SCP                  | 45.9 | 38.9 | 49.2 | 37.9 | 56.2 | 77.8 | 40.8 | 35.4 | 24.5 | 40.9 | 16.2 | 8.6 | 31.8 | 40.5 | 46.6 | 50.3 | 29.7 |
| + HRoIE                | 46.1 | 39.3 | 49.5 | 37.8 | 56.6 | 77.9 | 37.5 | 37.7 | 25.0 | 41.2 | 16.4 | 9.8 | 32.0 | 40.6 | 50.1 | 51.0 | 30.2 |

**PointRend** [5]

| Method                 | AP_b | AP_m | SH | ST | BD | TC | BR | LV | SV | HE | SP | RO | SB | PL | HA |
|------------------------|------|------|----|----|----|----|----|----|----|----|----|----|----|----|----|
| + DenseFPN             | 44.6 | 38.3 | 49.4 | 39.2 | 54.7 | 77.6 | 35.7 | 33.4 | 23.8 | 41.7 | 17.0 | 5.5 | 30.9 | 37.7 | 47.6 | 52.0 | 31.7 |
| + SCP                  | 45.2 | 38.7 | 50.9 | 39.4 | 55.7 | 78.4 | 36.6 | 33.7 | 25.0 | 43.1 | 16.8 | 5.0 | 30.1 | 39.2 | 46.2 | 52.7 | 31.4 |
| + HRoIE                | 45.7 | 39.3 | 51.6 | 39.7 | 55.8 | 78.6 | 37.7 | 36.0 | 25.6 | 42.8 | 16.8 | 5.7 | 29.8 | 39.9 | 47.7 | 53.1 | 31.6 |

\[
F_{b/m}^i = F_{b/m}^i + R_i^j \cdot \text{Sigmoid}([F_{b/m}^i \parallel R_i^j] \cdot w_i) \quad (6)
\]
reflected in different imaging qualities, weather, and seasons. The abbreviations of classes are AL - airplane, AR - airport, BF - baseball field, BC - basketball court, BR - bridge, CH - chimney, DA - dam, ES - expressway service area, ET - expressway toll station, GC - golf field, GT - ground track field, HA - harbor, OV - overpass, SH - ship, ST - stadium, SA - storage tank, TC - tennis court, TS - train station, VE - vehicle, and WM - wind mill.

NWPU VHR-10 [9]: NWPU VHR-10 is another widely used dataset for object detection in aerial images. It has 800 high resolution images, among which 650 are positive and 150 are negative without any objects of interests. This dataset contains annotations of 10 object categories. All the objects are annotated using horizontal bounding boxes that are publicly accessible. The abbreviations of classes are AL - airplane, SH - ship, ST - storage tank, BD - baseball diamond, TC - tennis court, BC - basketball court, GT - ground track field, HA - harbor, BR - bridge, and VE - vehicle.

HRSID [13]: HRSID is a recently introduced dataset for ship detection and segmentation in SAR images. This dataset contains a total of 5,604 high-resolution SAR images with 16,951 ship instances. All the instances in this dataset are

### TABLE II
OBJECT DETECTION RESULTS ON DIOR DATASET.

| Method          | Backbone | AP  |
|-----------------|----------|-----|
| R-CNN           | VGG-16   | 37.7 |
| ResNet-50       | 56.1     |
| ResNet-101      | 65.1     |
| Faster R-CNN    | VGG-16   | 54.1 |
| Mask R-CNN      | ResNet-50| 63.5 |
| RetinaNet       | ResNet-50| 65.7 |
| PANet           | ResNet-50| 66.1 |
| CBD-E           | ResNet-101| 67.8 |
| CSFF            | ResNet-101| 68.0 |
| FSoD-Net        | MNE-Net  | 71.8 |
| HawkNet         | ResNet-50| 72.0 |
| SCRDet++        | ResNet-50| 69.4 |
| CATNet†         | ResNet-50| 76.3 |
| CATNet‡         | ResNet-50| 77.7 |
| CATNet† + Aug.  | ResNet-50| 78.6 |
| CATNet‡ + Aug.  | ResNet-50| 81.9 |

### TABLE III
OBJECT DETECTION RESULTS ON NWPU-VHR-10 DATASET.

| Method          | Backbone | AP  |
|-----------------|----------|-----|
| R-CNN           | VGG-16   | 73.1 |
| ResNet-50       | 88.7     |
| ResNet-101      | 87.7     |
| Faster R-CNN    | VGG-16   | 84.5 |
| HRBM            | ZFNet    | 87.1 |
| ASBL            | ResNet-50| 89.3 |
| CAD-Net         | ResNet-101| 91.5 |
| DCL-Net         | ResNet-101| 94.6 |
| CBD-E           | ResNet-50| 95.0 |
| CATNet†         | ResNet-50| 95.8 |
| CATNet‡         | ResNet-50| 96.4 |
| CATNet† + Aug.  | ResNet-50| 97.4 |
| CATNet‡ + Aug.  | ResNet-50| 97.7 |

### TABLE IV
EXPERIMENTAL RESULTS ON HRSID DATASET.

| Method          | Backbone | AP  |
|-----------------|----------|-----|
| R-CNN           | ResNet-50| 59.8 |
| Faster R-CNN    | ResNet-101| 63.9 |
| Cascade R-CNN   | ResNet-101| 66.8 |
| HRDSID          | HRFPN-W40| 69.4 |
| Mask R-CNN      | ResNet-101| 65.4 |
| Cascade R-CNN   | ResNet-101| 67.0 |
| MS R-CNN        | ResNet-101| 65.9 |
| HTC             | ResNet-101| 68.4 |
| CATNet†         | ResNet-50| 70.5 |
| CATNet‡         | ResNet-50| 71.7 |
| CATNet† + Aug.  | ResNet-50| 72.8 |
| CATNet‡ + Aug.  | ResNet-50| 73.3 |

1 RetinaNet based, 2 Faster R-CNN based, 3 Aug. - Multi-Scale Training & Testing

Aug. - Multi-Scale Training & Testing
annotated with pixel-level masks. Spatial resolutions of the images are 0.5\(m\), 1\(m\), and 3\(m\).

We follow the standard evaluation metric that utilizes mean average precision (mAP) to measure the detection and segmentation performances. A result is considered as a true positive when the bounding box or mask of the object has intersection over union (IoU) with its ground truth greater than a threshold \(\theta_{IoU}\), and the predicted class label is correct. For iSAID dataset, we compute the mean of mAPs under \(\theta_{IoU}\) ranging from 0.05 to 0.95. For other datasets, only the mAPs under \(\theta_{IoU} = 0.5\) are considered according to the original papers.

**B. Implementation Details**

We choose Mask R-CNN [3], Faster R-CNN [1], and RetinaNet [2] with ResNet-50 [18] backbone as our baselines. The backbone is pre-trained on ImageNet [53] and finetuned when training the detector. All the parameters in the first stage are frozen after pre-training. If not specified, 5 basic blocks are included in all DenseFPN modules. In order to stabilize the training process, synchronized batch normalization (SyncBN) [54] layers are used among intermediate layers. When testing, we also adopt Soft-NMS [55] to suppress the duplicate results with IoU larger than 0.5 since most objects are heavily overlapped in remote sensing images.

We use stochastic gradient descent (SGD) optimizer with initial learning rate 0.01, momentum 0.9, and weight decay 0.0001 to learn the parameters for all models. Each training batch contains 8 images. For iSAID dataset, we follow the standard 1\(\times\) training schedule that drop the learning rate by 1/10 at epoch 8 and 11, and stop training at epoch 12. For DIOR, NWPU VHR-10, and HRSID datasets, we adopt the 3\(\times\), 6\(\times\), and 3\(\times\) training schedules respectively.

**C. Instance Segmentation Results in Aerial Images**

We first evaluate our approach on the instance segmentation task. Table I shows the comparison of the proposed modules on iSAID dataset when being incorporated into different frameworks. The class-wise instance segmentation mAPs are also reported. For both object detection and instance segmentation tasks, our methods can steadily boost the 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.
TABLE V
Detailed comparisons on ISAI D dataset. The baseline model is Mask R-CNN with ResNet-50-FPN as backbone and neck.

(a) Types of Multi-Scale Feature Propagation Modules

| Method                  | Depth | APb | APm | #Params | FLOPs   |
|-------------------------|-------|-----|-----|---------|---------|
| Baseline                | –     | 43.8 | 36.5| 43.82M  | 114.96G |
| PAFPN [52]              | –     | 44.4 | 37.0| 47.36M  | 121.31G |
| HRFPN [50]              | –     | 44.1 | 36.8| 44.41M  | 129.09G |
| CARAFE [51]             | –     | 44.1 | 36.9| 49.43M  | 115.71G |
| NAS-FPN [50]            | 1     | 44.2 | 36.9| 45.66M  | 119.86G |
| NAS-FPN                 | 3     | 45.3 | 37.4| 53.93M  | 155.30G |
| NAS-FPN                 | 5     | 45.5 | 37.6| 62.20M  | 190.74G |
| NAS-FPN                 | 7     | 45.6 | 37.7| 70.47M  | 226.18G |
| FPG@128 [52]            | 5     | 43.7 | 36.3| 38.74M  | 92.75G  |
| FPG@128                 | 7     | 44.2 | 36.6| 41.43M  | 102.23G |
| FPG@128                 | 9     | 44.5 | 36.9| 44.13M  | 111.70G |
| FPG@128                 | 11    | 44.6 | 37.1| 46.83M  | 121.18G |
| FPG@256 [52]            | 5     | 44.2 | 36.8| 60.56M  | 143.19G |
| FPG@256                 | 7     | 44.8 | 37.1| 71.33M  | 181.04G |
| FPG@256                 | 9     | 45.1 | 37.3| 82.09M  | 218.99G |
| FPG@256                 | 11    | 45.2 | 37.4| 92.86M  | 256.75G |
| DenseFPN (ours)         | 1     | 44.7 | 37.1| 44.19M  | 111.45G |
| DenseFPN (ours)         | 3     | 45.2 | 37.4| 53.92M  | 130.05G |
| DenseFPN (ours)         | 5     | 45.5 | 37.7| 52.75M  | 148.64G |
| DenseFPN (ours)         | 7     | 45.7 | 37.8| 57.03M  | 167.23G |

@ - Feature Channels

(c) Types of Region of Interest Extractors

| Method          | Direction | APb | APm | #Params | FLOPs   |
|-----------------|-----------|-----|-----|---------|---------|
| Baseline        | –         | 43.8 | 36.5| 43.82M  | 114.96G |
| GRoIE [40]      | –         | 44.0 | 36.8| 47.62M  | 574.92G |
| SUM             | –         | 43.9 | 36.7| 43.82M  | 114.96G |
| CONCAT          | –         | 44.0 | 36.7| 84.13M  | 188.18G |
| ATTENTION       | –         | 43.9 | 36.8| 45.92M  | 186.97G |
| HRoIE (ours)    | ↓ + +     | 44.2 | 37.1| 44.87M  | 151.00G |
| HRoIE (ours)    | ↓ + +     | 44.1 | 37.0| 44.87M  | 151.00G |
| HRoIE (ours)    | ↑ + +     | 44.3 | 36.9| 44.87M  | 151.00G |
| HRoIE (ours)    | ↑ + +     | 44.4 | 37.2| 44.87M  | 151.00G |

D. Object Detection Results in Aerial Images

Aside from instance segmentation, we observed that our approach can also benefit the task of object detection. To verify the effects, we also evaluate our model on DIOR and NWPU VHR-10 datasets. Table II and Table III show the comparison of our approach and previous state-of-the-arts on these two datasets. Note that for RetinaNet based methods, we only incorporate DenseFPN and SCP since RoI extractors are not used in single-stage models. The experiment results on DIOR and NWPU VHR-10 datasets show that our approach significantly outperforms all the previous methods. Moreover, the performances of our model with ResNet-50 backbones are even better than the previous state-of-the-art with ResNet-101 backbone by a noticeable margin.

E. Instance Segmentation Results in SAR Images

Beyond optical remote sensing images, we also evaluate our model on the more challenging SAR images. Specifically, SAR images are considered as single-channel grayscale images.

Each image is constructed by stacking three single-channel images at channel dimension. Experimental results in Table IV also demonstrate that compared with strong baselines for natural images, our method significantly works better, even using a lighter backbone. Note that in SAR images, we do not observe obvious gains using multi-scale training, so only multi-scale testing is used when performing data augmentation.

F. Visualizations

To demonstrate the effectiveness of SCP, we visualize the context aggregation weights in GCNet [36] and SCP in Fig. 7. Each row represents the weights of all classes that are aggregated into each class. The sizes and color depths of circles denote the weights in GCNet and SCP respectively. From the visualization, we may observe that our method tends to aggregate global spatial context from objects with the same category. Some similar (e.g. plane and helicopter) or semantically related (e.g. ship and harbor) categories can also help each other during training. Compared with our method, GCNet does not focus much on similar or semantically related
objects, leading to information confounding when aggregating features globally.

Fig. 8 shows qualitative results on iSAID, DIOR, and HRSID datasets. Each image patch is visualized by its feature aggregation weights in SCP and the final object detection or instance segmentation results. The results indicate that in most cases, our method focuses more on areas that contain objects, and aggregate them into regions with poor features. The final results show that our method can effectively detect and segment objects accurately in multiple scenes.

In order to study more about the training process, we also visualize the overall training losses in Fig. 9 (a). Combined with the proposed modules, the model converges faster and reach a lower final minima after training.

G. Detailed Comparisons and Ablation Study

In order to study the significance and effectiveness of the proposed modules individually, we conduct experiments on comparing them with some previous methods and with different module combinations. All the experiments are performed on iSAID dataset under the standard training and testing protocol. FLOPs are computed using $512 \times 512$ inputs.

Fig. 9 (b) and (c) compare the object detection and instance segmentation performances of different multi-scale feature propagation modules. Numerical results are reported in Table V (a). Compared with existing methods, DenseFPN works distinctly better on both object detection and instance segmentation tasks with fewer computational costs. We also observe that simply stacking more basic blocks can further boost the performance of DenseFPN.

Table V (b) shows the comparison between multiple spatial context modules. Compared with the baseline, NLNet [16] can effectively bring higher performance with a large computational cost. GCNet [36] solves the problem of computational complexity, but leads to another information confounding problem. With the help of re-weighting context, our proposed CABlock steadily outperforms GCNet using different channel reduction rates. Further experiments show that moving CABlocks from the backbone to after the multi-scale fusion module can better boost the performance with similar computational costs.

Table V (c) shows the comparison between multiple RoI extractors. The baseline model only crops the RoI features from a single feature map, leading to a severe information loss and achieving ordinary results. Simply computing the sum or concatenation of RoI features cropped from multiple layers can slightly boost the performance. Considering that object detection and instance segmentation tasks require different features, incorporating HRoIE for adaptive feature fusion can better generate appropriate RoI features for these tasks.

Table V (d) presents the ablation study results. All the three proposed modules can marginally bring better results on object detection and instance segmentation in remote sensing images. When collaborating with each other, the performance improvements are still stable, indicating that these modules do not interfere with each other. The best experimental results can be achieved by combining all these modules together, enabling aggregating multi-scale context from multiple domains simultaneously. Note that the $1 \times$ models are highly under-trained. With proper data augmentation or longer training schedules, our model can achieve much better performance.

V. Conclusion

In this paper, we reported an in-depth study on global visual context in remote sensing images and presented the proposed CATNet, a novel framework capable of leveraging three lightweight plug-and-play modules, i.e. dense feature pyramid network, spatial context pyramid, and hierarchical region of interest extractor, to aggregate the global visual context in feature, spatial, and instance domains. It has been shown that the collaboration among these three modules can effectively enhance the discriminative object features for promoting both the object detection and instance segmentation accuracies. Experimental results on iSAID, DIOR, NWPU VHR-10, and HRSID datasets show that the proposed approach significantly outperforms state-of-the-arts with similar computational costs. We expect that the new understanding of global visual context and the design of proposed modules will benefit future researches in this area.
ACKNOWLEDGEMENTS
This project is supported in part by National Natural Science Foundation of China (41871246) and National Key Research and Development Program of China (2019YFB2102901).

REFERENCES
[1] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 91–99.
[2] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961–2969.
[4] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang et al., “Hybrid task cascade for instance segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4974–4983.
[5] A. Kirillov, Y. Wu, K. He, and R. Girshick, “Pointrend: Image segmentation as rendering,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[6] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “Yolact: Real-time instance segmentation,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 9157–9166.
[7] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, “Solo: Segmenting objects by locations,” Tech. Rep., 2020.
[8] E. Xie, P. Sun, X. Song, W. Wang, D. Liang, C. Shen, and P. Luo, “Polarmask: Single shot instance segmentation with polar representation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[9] G. Cheng, J. Han, P. Zhou, and L. Guo, “Multi-class geospatial object detection and geographic image classification based on collection of part detectors,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 98, pp. 119–132, 2014.
[10] G.-S. Xia, X. Bai, J. Ding, Z. Zhi, S. Belongie, J. Luo, M. Datcu, M. Pelleli, and L. Zhang, “Dota: A large-scale dataset for object detection in aerial images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 3974–3983.
[11] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, “Isaid: A large-scale dataset for instance segmentation in aerial images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 28–37.
[12] 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 Journal of Photogrammetry and Remote Sensing, vol. 159, pp. 296–307, 2020.
[13] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, “Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation,” IEEE Access, vol. 8, pp. 120234–120254, 2020.
[14] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick, “Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2874–2883.
[15] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, “Contextlocnet: Context-aware deep network models for weakly supervised localization,” in Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 350–365.
[16] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7794–7803.
[17] G. Zhang, S. Lu, and W. Zhang, “Cad-net: A context-aware detection network for objects in remote sensing imagery,” IEEE Transactions on Geoscience and Remote Sensing, vol. 57, no. 12, pp. 10015–10024, 2019.
[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708.
[20] H. Zhang, Y. Tian, K. Wang, W. Zhang, and F.-Y. Wang, “Mask ssd: An effective single-stage approach to object instance segmentation,” IEEE Transactions on Image Processing, vol. 29, pp. 2078–2093, 2019.
[21] Y.-H. Wu, Y. Liu, L. Zhang, W. Gao, and M.-M. Cheng, “Regularized densely-connected pyramid network for salient instance segmentation,” IEEE Transactions on Image Processing, vol. 30, pp. 3897–3907, 2021.
[22] S. Liu, L. Qi, H. Qiu, N. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8759–8768.
[23] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelleli et al., “Object detection in aerial images: A large-scale benchmark and challenges,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
[24] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, “Learning roi transformer for oriented object detection in aerial images,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2849–2858.
[25] J. Zhang, C. Xie, X. Xu, Z. Shi, and B. Pan, “A contextual bidirectional enhancement method for remote sensing image object detection,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 4518–4531, 2020.
[26] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, “Cross-scale feature fusion for object detection in optical remote sensing images,” IEEE Geoscience and Remote Sensing Letters, 2021.
[27] H. Lin, J. Zhou, Y. Gan, C.-M. Vong, and Q. Liu, “Novel up-scale feature aggregation for object detection in aerial images,” Neurocomputing, vol. 411, pp. 364–374, 2020.
[28] G. Wang, Y. Zhuang, H. Chen, X. Liu, T. Zhang, L. Li, S. Dong, and Q. Sang, “Fsd-net: Full-scale object detection from optical remote sensing imagery,” IEEE Transactions on Geoscience and Remote Sensing, 2021.
[29] X. Yang, J. Yan, X. Yang, J. Tang, W. Liao, and T. He, “Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing,” Tech. Rep. arXiv:2004.13316, 2020.
[30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.
[31] G. Ghiasi, T.-Y. Lin, and Q. V. Le, “Nas-fpn: Learning scalable feature pyramid architecture for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7036–7045.
[32] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[33] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7132–7141.
[34] X. Zhu, D. Cheng, Z. Zhang, S. Lin, and J. Dai, “An empirical study of spatial attention mechanisms in deep networks,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 6688–6697.
[35] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Crcn: Cross-modal attention for semantic segmentation,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 603–612.
[36] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, “Global context networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.
[37] Z. Cai and N. Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6154–6162.
[38] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, “Mask scoring r-cnn,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6409–6418.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[40] L. Rossi, A. Karimi, and A. Prati, “A novel region of interest extraction layer for instance segmentation,” in Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), 2021, pp. 2203–2209.
[42] G. Cheng, P. Zhou, and J. Han, “Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016.

[43] K. Li, G. Cheng, S. Bu, and X. You, “Rotation-insensitive and context-augmented object detection in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2337–2348, 2017.

[44] G. Cheng, P. Zhou, and J. Han, “Rifd-cnn: Rotation-invariant and fisher discriminative convolutional neural networks for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2884–2893.

[45] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 21–37.

[46] J. Redmon and A. Farhadi, “Yolov3: An incremental improvement,” Tech. Rep. arXiv:1804.02767, 2018.

[47] J. Lei, X. Luo, L. Fang, M. Wang, and Y. Gu, “Region-enhanced convolutional neural network for object detection in remote sensing images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5693–5702, 2020.

[48] P. Sun, G. Chen, and Y. Shang, “Adaptive saliency biased loss for object detection in aerial images,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 10, pp. 7154–7165, 2020.

[49] E. Liu, Y. Zheng, B. Pan, X. Xu, and Z. Shi, “Dcl-net: Augmenting the capability of classification and localization for remote sensing object detection,” *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[50] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang et al., “Deep high-resolution representation learning for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[51] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, “Carafe++: Unified content-aware reassembly of features,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[52] K. Chen, Y. Cao, C. C. Loy, D. Lin, and C. Feichtenhofer, “Feature pyramid grids,” Tech. Rep. arXiv:2004.03580, 2020.

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2012, pp. 1097–1105.

[54] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[55] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, “Soft-nms – improving object detection with one line of code,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5561–5569.