

Detecting Fine-Grained Airplanes in SAR Images With Sparse Attention-Guided Pyramid and Class-Balanced Data Augmentation

Wei Bao ^{1b}, Jingjing Hu ^{1b}, Meiyu Huang ^{1b}, Yao Xu ^{1b}, Nan Ji, and Xueshuang Xiang ^{1b}

Abstract—Airplane detection in synthetic aperture radar (SAR) images has drawn much attention owing to the success of deep learning methods. However, the development of fine-grained airplane detection in SAR images is still in a dilemma due to the small interclass variance and the large intraclass variance in complex scenes with strong interference from the background. In addition, the class imbalance problem in multiclass fine-grained airplane recognition also significantly limits the direct application of general deep-learning-based airplane detectors. This article proposes two effective methods to tackle the above two problems, respectively. First, we propose a sparse attention-guided fine-grained pyramid module to simultaneously sample discriminative local features scattered in multiscale layers and adaptively aggregate them with fine-grained attention to better classify subordinate-level airplanes with multiple scales. Second, a simple class-balanced copy-paste data augmentation strategy, which randomly copies an airplane of one category and pastes it onto an image according to the classwise probability, is proposed for class balance. Finally, extensive experiments on one public dataset and three representative deep-learning-based detection benchmarks are conducted to show the effectiveness and generalization of the two proposed methods. The combination of these two methods based on the cascade R-CNN benchmark also won the fifth place in fine-grained airplane detection in SAR images in the 2021 GaoFen Challenge.

Index Terms—Class-balanced copy-paste data augmentation (CC-DA), fine-grained airplane detection, large intraclass variance, small interclass variance, sparse attention-guided fine-grained pyramid.

I. INTRODUCTION

SYNTHETIC aperture radar (SAR) is an active microwave remote sensing imaging radar with the capability of working

Manuscript received 9 June 2022; revised 24 July 2022 and 27 August 2022; accepted 9 September 2022. Date of publication 27 September 2022; date of current version 12 October 2022. This work was supported by the National Key Research and Development Program of China under Grant 2020YFB1709503 and the Beijing Nova Program of Science and Technology under Grant Z191100001119129. (Corresponding authors: Jingjing Hu; Meiyu Huang; Xueshuang Xiang.)

Wei Bao is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100811, China, and also with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing 100094, China (e-mail: baowei@bit.edu.cn).

Jingjing Hu is with the School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100811, China (e-mail: hujingjing@bit.edu.cn).

Meiyu Huang, Yao Xu, Nan Ji, and Xueshuang Xiang are with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing 100094, China (e-mail: huangmeiyu@qxslab.cn; xuyao@qxslab.cn; jinan@qxslab.cn; xiangxueshuang@qxslab.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3208928

in all-day and all-weather conditions and has triggered many SAR image processing tasks, including target recognition [1], detection [2], and segmentation [3], [4], [5]. As an important task, airplane detection in SAR images is highly significant on modern battlefields and military intelligence acquisition. The traditional SAR airplane detection method [6] leverages the constant false-alarm rate to distinguish objects from the background and suffers from tremendous difficulties in accurate detection due to weak feature extraction capabilities.

Benefiting from the rapid development of deep learning and the surge of satellite data, remarkable breakthroughs have been made in deep convolutional neural network (CNN) [7] based detection methods [8], [9], [10], [11], [12], [13], [14], [15], [16]. Based on these excellent CNN-based detectors, some effective network structures [17], [18], [19], [20], [21], [22], [23] are designed to detect airplanes with special structures and complex imaging mechanism in SAR images. Diao et al. [17] present a saliency-based target prelocating algorithm to reduce the false alarms in the region proposal stage. He et al. [18] propose a component-based multilayer parallel network to detect the overall aircraft by introducing corresponding component information. Zhang et al. [19] propose a cascaded three-look network that contains three stages: airport detection, aircraft detection in several airport chips, and airfield runway elimination. Zhao et al. [20] design a pyramid attention dilated network to enhance the relationship among discrete back-scattering features of aircraft. Other methods [21], [22], [23] design various attention mechanisms to improve detection performance. In addition, some SAR ship detection methods [24], [25], [26] and target recognition methods [27], [28] are also proposed to tackle with the specific imaging mechanism in SAR images.

Despite the success in general object detection above, the development of fine-grained airplane detection, which aims to locate and distinguish objects from different subordinate-level categories within a general category, is still in a dilemma. One significant challenge of fine-grained airplane detection lies in the small interclass variance and the large intraclass variance in complex scenes. As depicted in Fig. 1(a), the small interclass variance comes from two airplanes of different types with blue and white rectangles marked. What causes this phenomenon is the similar scattering mechanism with certain parts of the surrounding area and only subtle differences between some local areas of these two airplanes. The large intraclass variance comes from two airplanes of the same category with the orange

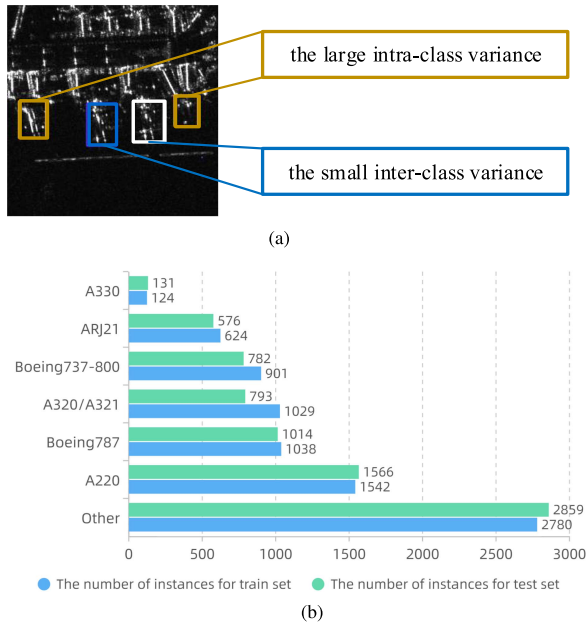


Fig. 1. Two serious problems for fine-grained airplane detection in SAR images. (a) Small interclass variance and the large intraclass variance. (b) the class imbalance problem in MFAD dataset.

rectangles marked in Fig. 1(a). These two airplanes of the same class present different scattering power distributions from the background caused by different orientations, scales, and appearances. Inspired by successful works [29], [30], [31], [32], [33], [34], [35], [36], [37] in fine-grained object recognition [38], several researchers [39], [40] leverage or find more discriminative information to overcome the small interclass and the large intraclass variance in fine-grained object detection. Han et al. [39] propose a concept of upper-level class to mine the potential interclass relationships among multiple ship categories hierarchically in optical remote sensing images. This hierarchically class annotation information is more discriminative to distinguish subordinate-level ships but domain-specific. Instead of performing fine-grained classification tasks separately, Song et al. [40] design a fine-grained detection head to select pixel-level features from different scales for each instance and aggregates them for the finer representation via a dynamic routing mechanism. However, the fine-grained information interaction only exists in adjacent scales, and simply elementwise aggregation in a dynamic routing mechanism is less efficient.

In addition to the small interclass and the large intraclass variance in complex scenes, another challenge for fine-grained object detection comes from class imbalance, especially the imbalance between different subcategories. As depicted in Fig. 1(b), the multiclass fine-grained SAR airplane detection (MFAD) dataset (the 2021 GaoFen Challenge dataset, and we will describe it in Section IV-A for more details) presents the class imbalance problem where 2780 airplanes for category “other,” while only 124 instances for category “A330” in the train set. A big performance drop would be observed when directly adopting detectors designed for a fairly balanced dataset to a long-tail distribution dataset. Reweighting-based methods [11],

[41], [42] and resampling-based methods [43], [44], [45] are two main branches of methods to deal with the class imbalance problem in object detection. Compared to reweighting-based methods, which are vulnerable to the interference of special background class with many instances, resampling-based methods are more suitable for detection tasks. However, the existing image-level resampling-based methods aim to augment images rather than instances, while instance-level resampling-based methods directly resample more proposals and do not increase the instance diversity, which is less efficient. As for instance augmentation, copy-paste data augmentation [46], [47], [48] performs well in the fairly balanced dataset but not designed to address the class imbalance problem in the fine-grained object detection task.

According to the above analysis, we propose a sparse attention-guided fine-grained pyramid (SA-FP) module and a class-balanced copy-paste data augmentation (CC-DA) strategy to deal with the small interclass variance and the large intraclass variance and class imbalance in SAR fine-grained airplane detection, respectively. The overall process is depicted in Fig. 2. Given the input image, the proposed CC-DA strategy randomly copies an airplane of one category (“ARJ21” with a blue rectangle marked) and pastes it onto this image according to the class-balanced probability derived from the number of airplanes in this category. If the frequency of instances of one class in the whole dataset is small, the CC-DA strategy will perform the copy-paste strategy with higher probability and vice versa. After the CC-DA strategy alleviates the class imbalance, the proposed SA-FP module simultaneously samples discriminative pixel-level local features scattered in multiscale layers and adaptively aggregates them to better classify different subordinate-level categories with the small interclass variance and the large intraclass variance, enhancing the multiscale feature representation as well. More specifically, in Fig. 2, the SA-FP module contains two submodules: the multiscale sparse sampling (MSSS) module and the attention-guided fine-grained fusion (AFF) module. The MSSS block leverages deformable convolution [49] to simultaneously sample several discriminative pixel-level features (six red squares) through learning the offsets according to the current reference point (the orange square) in spatial dimensions for all the feature maps. Then, the AFF block adaptively aggregates these discriminative local features and the current reference point with fine-grained attention to form the new pixel (the yellow square), enhancing the fine-grained representation. Finally, the SA-FP module and the CC-DA strategy can be easily combined to improve the performance of fine-grained airplane detection in SAR images. The main contributions of our work can be summarized as follows.

- 1) We analyze and conclude two main challenges in fine-grained airplane detection in SAR images: the small interclass variance and the large intraclass variance in complex scenes and the class imbalance problem.
- 2) Considering the small interclass variance and the large intraclass variance, we propose the SA-FP module to simultaneously sample discriminative pixel-level local features scattered in multiscale layers and adaptively aggregate them with fine-grained attention to better classify different subordinate-level airplanes with multiple scales.

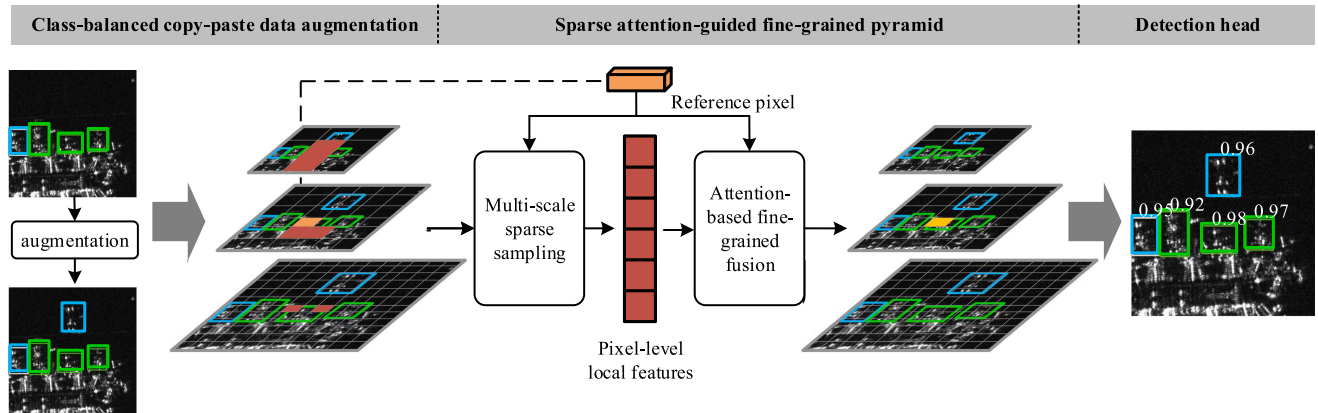


Fig. 2. Overall process of the fine-grained airplane detector in SAR images. It consists of CC-DA strategy, SA-FP module, and detection head. The CC-DA strategy randomly copies an airplane of one category and pastes it onto the current image for class balance. Given a reference point (the orange square), the SA-FP module simultaneously samples discriminative pixel-level local features (six red squares) scattered in multiscale layers and adaptively aggregates them to form new pixel (the yellow square) to better classify different subordinate-level airplanes with multiple scales.

- 3) Different from the traditional resampling-based methods and simple instance-level data augmentation, the CC-DA strategy is proposed to randomly copy an airplane of one category and paste it onto an image according to the classwise probability for class balance.
- 4) Various experiments are conducted on one fine-grained airplane detection dataset and three representative CNN-based detection benchmarks [8], [11], [14] to demonstrate the effectiveness and generalization of the proposed methods.

The rest of this article is organized as follows. Section II introduces some related works, and Section III introduces our methods in detail. Section IV provides the experimental settings and results analysis. Finally, Section V concludes this article.

II. RELATED WORK

A. CNN-Based Detectors

Generally, CNN-based detection methods can be divided into two categories: anchor-based detection methods [8], [9], [10], [11], [12] and anchor-free detection methods [13], [14], [15], [16].

- 1) Anchor-based detection methods predefine a handful of anchor boxes with different scales and respect ratios as common references to search possible regions containing the object of interest. Faster R-CNN [8], Cascade R-CNN [9], and Libra R-CNN [10] preliminarily extract class-agnostic region proposals of the potential objects with negative locations filtered out and then further refine these proposals and classify them into different categories in a two-stage manner. RetinaNet [11] and YOLOv3 [12] omit the region proposal generation process and directly classify and regresses different objects in a one-stage manner.
- 2) Considering that the predefined anchors are domain-specific and highly sophisticated for detection heads, anchor-free detection methods, such as RepPoints [13] and FCOS [14], design considerably simpler detectors to

eliminate the predefined anchor boxes and reduce the number of design parameters. More specifically, DETR [15] and Deformable DETR [16] leverage the transformer mechanism to avoid the design of predefined anchor boxes by formulating object detection as a direct set prediction task.

B. Fine-Grained Object Recognition

Recent works for fine-grained object recognition [38] attempt to leverage or find more discriminative information to distinguish different subordinate-level categories and can be classified into three classes: feature-encoding methods, localization-based methods, and attention-based methods.

- 1) Feature-encoding methods [29], [30], [31] aim to learn more discriminative representation for modeling subtle differences by performing high-order feature interactions.
- 2) Localization-based methods [32], [33], [34] rely on additional annotation information to locate the subtle parts and then perform feature extraction and classification.
- 3) Attention-based methods [35], [36], [37] attempt to find the discriminative region in images by exploiting the powerful properties of attention, releasing the reliance on manual annotation.

C. Class-Balanced Object Detection

The class imbalance problem in object detection can be solved from two perspectives: 1) reweighting training examples to balance optimization direction in loss level [11], [41], [42] and 2) resampling training examples to balance the distribution in data level [43], [44], [45]. Reweighting-based methods elaborately design balanced loss to dynamically rebalance gradients of imbalanced classes. Resampling-based methods can be performed at the image and instance levels. RFS [43] proposes an image-level repeat factor sampling strategy to oversample the images that contain the category with small instances. Sim-Cal [44] and Forest R-CNN [45] focus on the instance-level sampler to balance the classes by selecting more proposals for tailed classes.

D. Multiscale Representation

A feature pyramid network (FPN) [50] has been proven to alleviate the scale variance in object representation by combining low-resolution semantic strongly features at the high level with high-resolution semantic weakly features at the low level. PANet [51] further enhances the entire feature hierarchy with bottom-up path augmentation. Aug-FPN [52] narrows the semantic gaps between features of different scales before feature fusion through consistent supervision. Considering the semantic information diluted in nonadjacent levels, BFP [10] integrates and refines the multilevel features to enhance the semantic hierarchy at the same time. Dyhead [53] presents a novel dynamic head framework to enhance scale awareness between feature levels. Deformable DETR [16] upgrades DETR [15] with a multiscale representation module to mitigate the limited feature spatial resolution.

E. Attention Mechanism

SEnet [54] adaptively recalibrates channelwise feature responses by explicitly modeling interdependencies between channels. CBAM [55] sequentially extracts cross-channel and spatial information by inferring attention maps along the channel and spatial dimensions for adaptive feature refinement, respectively. GCnet [56] forms a global context feature by aggregating the features of all the positions together, followed by channelwise interdependencies modeling. Transformer [57] captures long-range global context by directly computing each query position response as the weighted aggregation of the features at all the positions regardless of their distance.

III. METHODOLOGY

In this section, we first introduce three CNN-based object detection benchmarks: Faster R-CNN [8], RetinaNet [11], and FCOS [14]. Next, the SA-FP structure, including the MSSS and the AFF module, is described in detail. Finally, we will introduce how the CC-DA strategy solves the class imbalance problem.

A. CNN-Based Object Detection Methods

We select three representative methods: Faster R-CNN [8], RetinaNet [11], and FCOS [14], as the baselines in our work. Here, we only introduce the critical architecture of these three detection benchmarks, and we refer to their original paper [8], [11], [14] to see a more detailed introduction. It is noted that our proposed SA-FP module and CC-DA strategy can be easily applied to other state-of-the-art SAR airplane detectors to boost the performance of fine-grained recognition.

Faster R-CNN [8], on behalf of the two-stage anchor-based object detection methods, consists of three modules: feature embedding network extracting high-level features from the original images, region proposal network (RPN) preliminarily generating the object proposals and prediction network performing the final classification and regression task. We initially set three anchor boxes with one scale of size 8 and three aspect ratios of size $\{0.5, 1, 2.0\}$ at each spatial location of each feature map. After RPN generating proposals, we adopt the RoIAlign [58] operation to fix the misalignment of feature maps caused by coarse

spatial quantization. Furthermore, we select the cross-entropy and *smooth* L1 loss function to optimize the classification and regression task, respectively.

RetinaNet [11], the representative one-stage anchor-based object detector, eliminates the dense-to-sparse stage in the RPN and leverages a novel focal loss to resolve the extreme foreground-background class imbalance in a directly dense prediction manner. It is noted that the focal loss is not suitable to solve the foreground-foreground class imbalance in fine-grained object detection. Similar to the parameters settings in Faster R-CNN, we also set three anchor boxes with one scale and three aspect ratios at each spatial location. The parameters of focal loss are set as $\alpha = 0.5$ and $\gamma = 2$.

FCOS [14], the representative anchor-free object detector, explores fully convolutional networks to directly regress the distances from the location falling into the bounding box to the four sides in a per-pixel prediction manner and eliminates the design of domain-specific anchor boxes. Moreover, FCOS adds an additional “centerness” classification branch to suppress low-quality detected bounding boxes produced by locations far away from the center of an object.

B. Sparse Attention-Guided Fine-Grained Pyramid

The FPN is one of the most classic architectures in object detection and has been proven to alleviate the scale variance in object representation. However, the FPN ignores the fine-grained local features scattered in multilevel features and only performs elementwise addition to combine low-resolution semantic strongly features with high-resolution semantic weakly features. The proposed SA-FP method can well find and leverage these discriminative pixel-level local features for fine-grained representation. The same idea that conditionally selects a pixel-level combination of local features from different scales also occurs in [40]. However, the fine-grained feature was only selected from adjacent scales via spatial gate. Instead, our proposed MSSS module can simultaneously sample fine-grained local features across all the feature levels. Moreover, the proposed AFF module can aggregate discriminative features with dynamic fine-grained attention scores for different feature vectors other than simple elementwise accumulation in [40]. The proposed SA-FP method can also generate appropriate pyramidal representation, which is crucial to multiscale airplane detection. The idea that refines FPN features is also similar to those in [10], [16], [26], and [59]. The most related work Deformable DETR [16] proposes a multiscale deformable attention module to mitigate issues of slow convergence and limited feature spatial resolution in DETR [15]. However, the SA-FP method leverages MSSS to capture pixel-level local features and elaborately designs fine-grained dynamic attention instead of the inner products or a separated convolution layer in Deformable DETR [16]. Next, we will introduce the MSSS module and the AFF module in detail.

1) *MSSS Module*: It is difficult to directly extract effective fine-grained local features from the whole images due to complex spatial feature relationships and fixed receptive fields in standard convolution. Inspired by DCN [49] and Deformable DETR [16], the MSSS module is proposed to enable the network

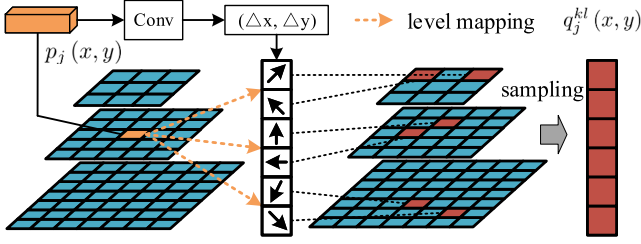


Fig. 3. Structure of the MSSS module.

to collect a small set of highly correlated pixel-level features by a separated convolution layer according to the current reference pixel, as depicted in Fig. 3. Specifically, the convolution layer takes in the reference pixel and outputs the offsets of pixels to be sampled in the x and y directions. Then, the fine-grained local features can be sampled in the multilevel features based on these offsets. It is noted that the MSSS module needs to traverse all the reference pixels on the feature map, and only one point is shown in Fig. 3 for convenience.

Take RetinaNet as an example; the backbone is divided into five stages according to the size of the feature maps in different layers. Based on the last three stages of the backbone, the FPN produces five-layer feature maps, which can be defined as $\{P_3, P_4, P_5, P_6, P_7\}$. More specifically, the feature map at pyramid level l is denoted as $P_l \in \mathbb{R}^{(W/s_l) \times (H/s_l) \times C}$ with $l = 3, 4, 5, 6, 7$, where $W \times H$ is the size of the input image and C is the channel dimension; $s_l = 2^l$ is the corresponding downsampling ratio to the input image.

The SA-FP module refines each pixel in pyramidal level j for finer representation. Suppose that the reference pixel p_j with location (x, y) has content feature $p_j(x, y) \in \mathbb{R}^{1 \times 1 \times C}$ in input feature map P_j . As depicted in Fig. 3, the MSSS module first samples KL sparse pixels for all the pyramid levels according to $p_j(x, y)$, where K enumerates all the sampled pixels in each pyramid level and L is the number of pyramid levels. Assume that $q_j^{kl}(x, y)$ represents pixels to be sampled. The input-specific offsets $(\Delta x^{kl}, \Delta y^{kl})$ between $p_j(x, y)$ and $q_j^{kl}(x, y)$ can be learned via a separate $1 \times 1 \times C$ convolution layer with $2KL$ output channels applied over $p_j(x, y)$. It is worth noting that the learned convolution kernels can be updated simultaneously with the whole detection network. $K \ll H \times W$ enables the network to focus on several discriminative pixel-level local features. The formula for obtaining sampled pixels $q_j^{kl}(x, y)$ from the reference point in location (x, y) can be written as follows:

$$\begin{aligned} q_j^{kl}(x, y) &= p_l(x^\dagger, y^\dagger) \\ &= p_l(\phi_j^l(x) + \Delta x^{kl}, \phi_j^l(y) + \Delta y^{kl}) \end{aligned} \quad (1)$$

where $\phi_j^l(\cdot)$ is the level mapping function that rescales the coordinates of $p_j(x, y)$ to the feature map in the l th level. For the clarity of scale formulation, we use normalized coordinates, in which the normalized coordinates $p_l(0, 0)$ and $p_l(1, 1)$ indicate the top-left and the bottom-right feature map corners, respectively. Moreover, the sampling offset $(\Delta x^{kl}, \Delta y^{kl})$ is typically fractional; the value $p_j^{kl}(x, y)$ remains to be determined. Bilinear

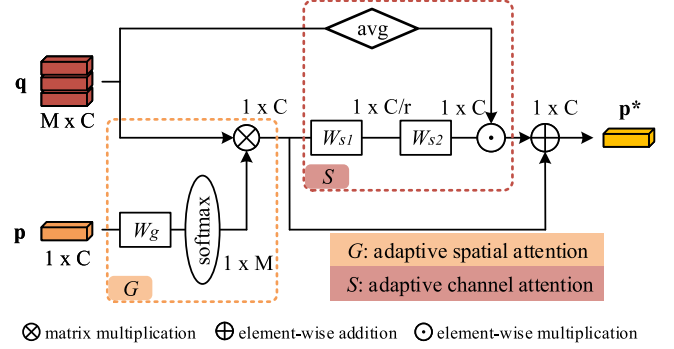


Fig. 4. Structure of the AFF module.

interpolation is implemented to determine the missed fractional value

$$\begin{aligned} q_j^{kl}(x, y) &= \sum_{(x^*, y^*) \in \tau} B((x^*, y^*), (x^\dagger, y^\dagger)) p_l(x^*, y^*) \\ &= \sum_{(x^*, y^*) \in \tau} \max(0, 1 - |x^* - x^\dagger|) \\ &\quad \max(0, 1 - |y^* - y^\dagger|) p_l(x^*, y^*) \end{aligned} \quad (2)$$

where (x^*, y^*) represents an arbitrary position and τ enumerates all integer spatial locations near (x^\dagger, y^\dagger) . B is the bilinear interpolation kernel and is separated into two 1-D kernel in the x and y directions. Then, KL sparsely sampled pixels will be concatenated to generate the corresponding sampled feature vector. Next, we will introduce the AFF module to fuse these pixel-level local features to improve the fine-grained representation.

2) *Attention-Based Fine-Grained Fusion Module*: The AFF module aims to aggregate the sampled local features to achieve fine-grained representation. The general approach to directly sum or average the sample pixels is simple but less efficient. To adaptively aggregate contexts and retain more semantic information, we propose to leverage the attention mechanism to enhance the fusion performance. As shown in Fig. 4, the procedure contains sequentially the adaptive spatial attention G and adaptive channel attention S to involve both spatial structures and channel semantics under the guidance of the reference point. Here, we use $M = KL$, \mathbf{q} , and \mathbf{p} to represent the number of sampled pixels, sampled pixels $q_j^{kl}(x, y)$, and the reference point $p_j(x, y)$ for convenience, respectively. Supposing that \mathbf{p}^* is the final fused feature value, the formula of the AFF module can be defined as follows:

$$\mathbf{p}^* = \text{AFF}(\mathbf{p}, \mathbf{q}) = G(\mathbf{p}, \mathbf{q}) + S(\mathbf{p}, \mathbf{q}). \quad (3)$$

As for the adaptive spatial attention G , we leverage one separated linear layer $W_g \in C \times M$ followed by a softmax operation to directly generate the spatial attention weights for \mathbf{q} based on \mathbf{p} . Different from general spatial attention in [56], the sampled pixels here are selected according to the reference pixel via a separated convolution layer in the MSSS module (see Fig. 3 for more details), which illustrates that \mathbf{q} and \mathbf{p} are already related to each other. Thus, the generated attention weights from \mathbf{p} can well involve spatial structures on \mathbf{q} . The corresponding

formula can be defined as follows:

$$G(\mathbf{p}, \mathbf{q}) = \frac{e^{(\mathbf{p}W_g)_m}}{\sum_{m=1}^M e^{(\mathbf{p}W_g)_m}} \mathbf{q}. \quad (4)$$

Consequently, the sampled local features can be reweighted in spatial dimension.

As for the adaptive channel attention S , we adopt two consecutive linear layers (W_{s1} and W_{s2}) after the adaptive spatial attention to adaptively recalibrate channelwise feature responses by explicitly modeling interdependencies between channels. Different from the channel attention directly recalibrating the whole feature maps with a fixed response in [54], the proposed AFF module recalibrates different sparsely sampled feature vectors with dynamic responses for different reference pixels. This dynamic channel attention enables the AFF module to extract more fine-grained semantic information for local features scattered in multiscale feature maps, which is constructive for improving the ability of fine-grained airplane recognition. The dynamic mechanism shares similar insights with the fine-grained attention mechanism for neural machine translation in [60]. Specifically, the corresponding formula can be defined as follows:

$$\begin{aligned} S(\mathbf{p}, \mathbf{q}) &= \delta(G(\mathbf{p}, \mathbf{q})) \cdot \mathbf{q} \\ &= W_{s2} \text{ReLU}(W_{s1}(G(\mathbf{p}, \mathbf{q}))) \cdot \mathbf{q} \\ &= W_{s2} \text{ReLU}\left(W_{s1}\left(\frac{e^{(\mathbf{p}W_g)_m}}{\sum_{m=1}^M e^{(\mathbf{p}W_g)_m}} \mathbf{q}\right)\right) \cdot \mathbf{q} \end{aligned} \quad (5)$$

where $\delta(\cdot) = W_{s2} \text{ReLU}(W_{s1}(\cdot))$ denotes the bottleneck transform used in [54]. $a \cdot b$ means the elementwise multiplication operation between a and b . Both $W_{s1} \in C \times C/r$ and $W_{s2} \in C/r \times C$ are linear layers to capture the channel dependence. The factor r is set as 4 in all the experiments by default. Consequently, the sampled local features can be reweighted in channel dimension.

After the AFF module, the SAR airplane detector can well extract fine-grained information to distinguish different subordinate-level categories. Moreover, we also adopt the multihead attention module in [16] to more adaptively aggregate fine-grained features from different representation subspaces. The formula of multihead attention operation can be defined as follows:

$$\begin{aligned} \text{MultiHead}(\text{AFF}(\mathbf{p}, \mathbf{q})) &= \text{Concat}(\text{head}_1, \dots, \text{head}_H) W_{h2} \\ \text{head}_i &= \text{AFF}(\mathbf{p}, \mathbf{q} W_{h1}^i) \end{aligned} \quad (6)$$

where Concat represents the concatenate operation, i indexes the attention head, and H is the number of heads (we use $H = 8$ by default). $W_{h1}^i \in C \times C/H$ and $W_{h2} \in C \times C$ are linear layers.

C. Class-Balanced Data Augmentation

The foreground–foreground class imbalance problem usually exists in fine-grained object detection, such as MFAD dataset, as depicted in Fig. 1(b). Resampling-based methods are more suitable for detection tasks than reweighting-based methods.

Algorithm 1: The Training Process of the CC-DA Strategy. I is the Number of Categories.

Input: the training dataset and its annotations.

Output: the class-balanced training process for fine-grained detection.

- 1: Initialize the set F_i to record the number of instances in each category i and the set E_i to record the airplane slice in each category i ;
 - 2: **for** each image in the training dataset **do**
 - 3: **for** each airplane in the current image **do**
 - 4: inquire the class index cls and bounding box $bbox$;
 - 5: copy the current airplane slice e from the current image according to the $bbox$;
 - 6: $E_{cls} = E_{cls} \cup e$;
 - 7: $F_{cls} = F_{cls} + 1$;
 - 8: **end for**
 - 9: **end for**
 - 10: compute the probability $prob_i$ for each category according to (7);
 - 11: start training;
 - 12: **for** each image in the training dataset **do**
 - 13: **for** i in $1, 2, \dots, I$ **do**
 - 14: random select a copied airplane slice e from E_i according to $prob_i$;
 - 15: random paste e onto the current image and obtain the new location $bbox_{new}$;
 - 16: add the class index i and the new location $bbox_{new}$ into the current annotation;
 - 17: **end for**
 - 18: perform forward and backward operation;
 - 19: **end for**
-

LVIS [43] proposes a repeat factor sampling strategy that increases the rate of tailed categories being observed by oversampling the images containing them. However, when an image containing tailed categories appears multiple times, the head class on this image can also appear multiple times. Moreover, the image-level sampling strategy will also significantly increase training time. The basic idea behind the proposed CC-DA strategy is to oversample the limited instances for small-sample classes at the instance level instead of the image level. More specifically, for the category i , the proposed CC-DA strategy randomly copies an airplane from the whole airplanes of the category i and pastes it onto an image with probability $prob_i$:

$$prob_i = e^{-\frac{f_i}{t}} \quad (7)$$

where f_i denotes the frequency of airplane of category i in the whole training dataset and t is the hyperparameter to adjust $prob_i$. If the number of airplanes of category i is small, the probability of random copy-paste operation is larger. As a result, the number of airplanes in different categories can be balanced. The pseudocode of the CC-DA strategy is provided in Algorithm 1. The proposed copy-paste data augmentation strategy is similar to that in [46], [47], and [48], which has been demonstrated to provide solid gains on a fairly balanced detection dataset. However, these methods are not dedicated to

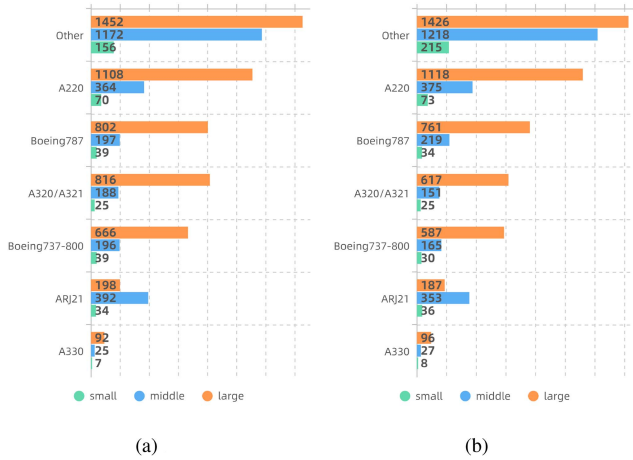


Fig. 5. Distribution of instances number and size for different classes in the MFAD dataset. (a) Number of instances for the train set. (b) Number of instances for the test set.

the long-tail distribution dataset and are unsuitable for solving class imbalance problems in the multiclass fine-grained object detection task.

IV. EXPERIMENTS

All the experiments are implemented in the PyTorch 1.7.0 framework and carried out over an NVIDIA 3070 GPU. The PC operating system is 64-bit Ubuntu 20.04. We conduct experiments on the MFAD dataset and three detection benchmarks [8], [11], [14] to demonstrate the effectiveness and generalization of the proposed methods. In addition, we use RetinaNet to perform the ablation study.

A. Datasets

The “Fine-Grained Airplane Recognition in High-Resolution SAR Images” track of “2021 GaoFen Challenge on automated high-resolution earth observation image interpretation” announces the world’s first MFAD for convenience. The MFAD contains 2000 high-resolution multiscale images collected from the GF-3 satellite. There are 1102 images with 600×600 pixels, 555 images with 1024×1024 pixels, and 343 images with 2048×2048 pixels. The MFAD consists of six categories: A220, A330, A320/A321, Boeing737-800, Boeing787, and ARJ21. We randomly select 1000 images as training data, and the remaining 1000 images are used for testing data. Each image of the training and testing dataset is cropped into 512×512 pixels with an overlap of 256 pixels, except that images with 600×600 pixels are directly resized into 512×512 pixels. Finally, the new version of MFAD (we use MFAD to indicate the new version later) contains 3352 training images and 3352 testing images with 512×512 pixels. Fig. 5(a) and (b) shows the distribution of instances number and size for different classes in the train and test dataset, respectively. We divide these instances into three different sizes: small size (area $< 32 \times 32$), middle size ($32 \times 32 < \text{area} < 64 \times 64$), and large size ($64 \times 64 < \text{area}$). It can be seen that the class imbalance and the multiscale instances are two major problems in the MFAD dataset.

B. Parameter Settings

The overall experiments are performed based on MMDetection (<https://github.com/open-mmlab/mmdetection>). For all three detectors, we use the pretrained ResNet-50 [61] on the ImageNet [62] to initialize the backbone network. All the baselines are trained with stochastic gradient descent for 24 epochs (usually called $2 \times$ schedule) with eight images per minibatch. The initial learning rate is set as 0.005 and then divided by 10 at the 16th and 22nd epochs. We use the weight decay of 0.0001 and the momentum of 0.9. Other parameters are set as the same as that in MMDetection. The intersection over union (IoU) threshold is set as 0.5 when training and testing for rigorous filtering of the bounding boxes with low precision. Warm-up [61] is introduced during the initial training stage to avoid gradient explosion, and the corresponding number of epochs is set as 2. As for the SA-FP module, the number of FPN levels for sparse sampling is set as $L = 5$, and the number of sampling points in every level is set as $K = 6$. We use the same settings for all the experiments for a fair comparison.

C. Evaluation Metrics

The mean average precision metrics mAP_{iou} under different IoU thresholds are employed to evaluate the performance of fine-grained SAR airplane detectors. mAP_{iou} is the mean of AP_{iou}^c for different classes

$$\text{mAP}_{\text{iou}} = \frac{1}{C} \sum_{c=1}^C \text{AP}_{\text{iou}}^c \quad (8)$$

where C is the number of categories and iou indicating different IoU thresholds. The average precision AP_{iou} is the area under the curve of precision–recall and usually be calculated as follows for convenience:

$$\text{AP}_{\text{iou}} = \frac{1}{101} \sum_{r \in S} \text{Precision}_{\text{iou}} \Big|_{\text{Recall}_{\text{iou}}=r} \quad (9)$$

where $S = \{0, 0.01, \dots, 1\}$ representing a set of equally spaced recall rates, and we use AP_{iou} to denote AP_{iou}^c for convenience. The $\text{Precision}_{\text{iou}}$ and $\text{Recall}_{\text{iou}}$ represent the precision rate and recall rate under different iou , respectively. For a given iou , they can be defined as

$$\text{Precision} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \quad (10)$$

$$\text{Recall} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \quad (11)$$

where TP is true positive, FP is false positive, TN is true negative, and FN is false negative. N_{TP} , N_{FP} , and N_{FN} is the number of TP, FP, and FN, respectively. More specifically, TP indicates the correctly detected airplanes, FP represents the false alarms, and FN denotes the missing airplanes. A predicted bounding box is considered a true positive if its IoU with the ground truth is higher than iou . Otherwise, it is regarded as a false positive. Moreover, the predicted bounding box with the highest confidence score is seen as the true positive, if the IoUs of several ones with the ground truth are all higher than the threshold. mAP_{iou} is a comprehensive evaluation metric for the

TABLE I
OVERALL PERFORMANCE OF DETECTORS WITH DIFFERENT COMPONENTS ON THE MFAD DATASET BASED ON FASTER R-CNN [8], RETINANET [11], AND FCOS [14] BENCHMARKS

Faster R-CNN [8]								
Methods	SA-FP	CC-DA	mAP _{0.5}	mAP _{0.75}	mAP	params(M)	FLOPs(G)	FPS
baseline	✗	✗	0.8817	0.7332	0.6275	41.15	63.28	57.1
SA-det	✓	✗	0.8932	0.7555	0.6442	41.42	68.92	47.6
CC-det	✗	✓	0.8880	0.7540	0.6410	41.15	63.28	57.1
SA-CC-det	✓	✓	0.8970	0.7725	0.6563	41.42	68.92	47.6
RetinaNet [11]								
baseline	✗	✗	0.8674	0.7260	0.6260	36.23	52.96	59.1
SA-det	✓	✗	0.8992	0.7562	0.6595	36.55	54.71	50.6
CC-det	✗	✓	0.8832	0.7493	0.6485	36.23	52.96	59.1
SA-CC-det	✓	✓	0.9015	0.7714	0.6696	36.55	54.71	50.6
FCOS [14]								
baseline	✗	✗	0.8870	0.7430	0.6450	31.90	51.61	67.2
SA-det	✓	✗	0.9032	0.7554	0.6578	32.22	53.36	57.5
CC-det	✗	✓	0.8890	0.7540	0.6580	31.90	51.61	67.2
SA-CC-det	✓	✓	0.9094	0.7521	0.6592	32.22	53.36	57.5

Bold entity represents the best.

quantitative performance of different models by simultaneously considering the precision rate and the recall rate. $AP_{0.5}$ denotes AP_i with the IoU threshold being 0.5. To evaluate the localization performance more accurately, we adopt $AP_{0.5}$, $AP_{0.75}$, and AP metrics. AP indicates the averaged AP_{iou} where iou is set from 0.50 to 0.95 with the step size set as 0.05, which can be defined as

$$AP = \frac{1}{10} \sum_{iou \in I} AP_{iou} \quad (12)$$

where $I = \{0.5, 0.55, \dots, 0.95\}$ representing a set of equally spaced IoU threshold.

D. Results Analysis

1) *Overall Performance*: Table I reports the detection performance of SAR fine-grained airplane detectors adopting the proposed SA-FP module and the CC-DA strategy based on different benchmarks. We use SA-det and CC-det to represent the baseline with only the SA-FP module and the CC-DA strategy applied, respectively. Similarly, the SA-CC-det denotes the baseline adopting both the SA-FP module and the CC-DA strategy. It can be seen that the SA-det can improve the detection performance under different mAP metrics with slightly increased model parameters, FLOPs, and running time per image. Specifically, the SA-det achieves 3.18% higher $mAP_{0.5}$ than the baseline based on RetinaNet [11] benchmark. Furthermore, when the IoU threshold becomes larger, which indicates that the requirement of localization accuracy gets higher, $mAP_{0.75}$ and mAP gain similar improvement of 3.02% and 3.35%, respectively. The quantitative detection performance increase demonstrates that the SA-det can conditionally select discriminative local features in similar scattering intensities and effectively fuse the fine-grained feature for fine representation. As for the CC-det

based on RetinaNet, 1.58%, 2.33%, and 2.25% performance improvement can be achieved in terms of $mAP_{0.5}$, $mAP_{0.75}$, and mAP metrics, respectively. More importantly, because the CC-DA is a data augmentation method, it is purely cost-free and does not increase model parameters. All these improved performances prove that the CC-DA strategy is very constructive for class balance in fine-grained airplane detection. Finally, the SA-CC-det based on RetinaNet can further enhance the detection results and bring 3.41%, 4.54%, and 4.136% gains over the baseline under the metric of $mAP_{0.5}$, $mAP_{0.75}$, and mAP , respectively. Similar improvements in the other two benchmarks also demonstrate the effectiveness and generalization capability of the proposed SA-FP module and CC-DA strategy. In addition to quantitative comparisons, we also visualize some detection results in Fig. 6 to show an intuitive understanding of our proposed methods. We can see that the baseline misses several airplanes and generates a few false alarms due to the influence of the background with similar scattering intensity. Moreover, the misclassification of fine-grained airplanes occurs in the scene of the second column for baseline because of the small interclass variance. In contrast, the proposed SA-CC-det performs better on fine-grained airplane detection in complex scenes. The SA-CC-det can accurately locate and classify different fine-grained airplanes and significantly increase the confidence score of the correctly detected target. Next, we will analyze the multiclass fine-grained airplane detection performance in more detail.

2) *Performance for Each Category*: To more clearly illustrate the performance improvement that the two proposed methods bring for multiclass fine-grained aircraft detection, we show the detection results of each category in terms of AP metric based on the RetinaNet [11] benchmark in Table II. We count the number and percentage of airplanes in each category and then arrange the classes in descending order for easy comparison. It can be observed that the SA-det can improve the detection results

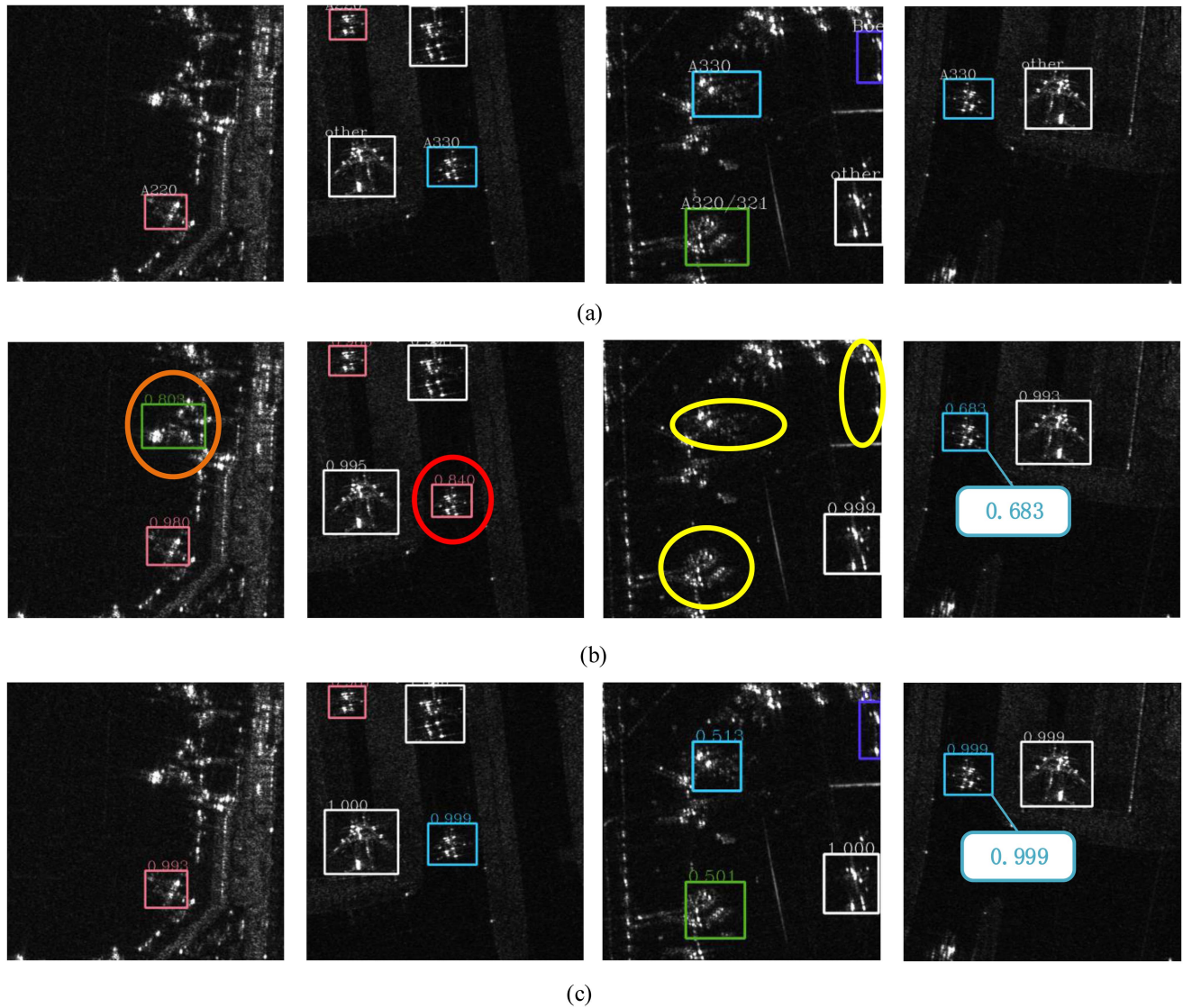


Fig. 6. Comparison results of baseline and the SA-CC-det on the MFAD dataset based on RetinaNet [11]. The yellow and orange circles represent the missing airplanes and false alarms, respectively. The red circles denote the misclassified airplanes, which means not only missing airplanes and also false alarms. The confidence score is shown in the blue rectangles. (a) Ground truth. (b) Results of baseline. (c) Results of SA-CC-det.

TABLE II
PERFORMANCE UNDER AP METRIC OF DETECTORS WITH DIFFERENT COMPONENTS ON THE MFAD DATASET BASED ON THE RETINANET [11] BENCHMARK FOR EACH CATEGORY

category	number (percentage)	baseline	SA-det	CC-det	SA-CC-det
other	2859 (37.03%)	0.6282	0.6509	0.6318	0.6507
A220	1566 (20.28%)	0.6470	0.6913	0.6571	0.6835
Boeing787	1014 (13.13%)	0.7076	0.7411	0.7126	0.7439
A320/321	793 (10.27%)	0.5670	0.6074	0.5931	0.6134
Boeing737-800	782 (10.13%)	0.6623	0.6856	0.6684	0.6882
ARJ21	576 (7.46%)	0.5748	0.5980	0.5826	0.6008
A330	131 (1.70%)	0.5950	0.6424	0.6883	0.7067
mean	—	0.6260	0.6595	0.6485	0.6696

Bold entity represents the best.

by a large margin ranging from 2.27% to 4.74% for all the categories, which demonstrates that the SA-FP module helps detect airplanes of a different class. As for the CC-det, the improvement becomes larger for small-sample classes, such as “A330” with 9.33% gains, while it becomes smaller for large-sample classes, such as “other” with only 0.36% improvements. We conjecture that this phenomenon is due to the different copy-paste probabilities for each category in the CC-DA strategy. It also illustrates that the CC-DA strategy can relieve the class-imbalanced problem in fine-grained airplane detection. When adopting the two proposed methods, the SA-CC-det achieves the highest performance for all the categories, except for “other” and “A220,” where performance drops slightly. This interesting phenomenon further illustrates that the SA-CC-det gives dynamic attention to different categories and little sacrifices the performance for the category with more instances to achieve

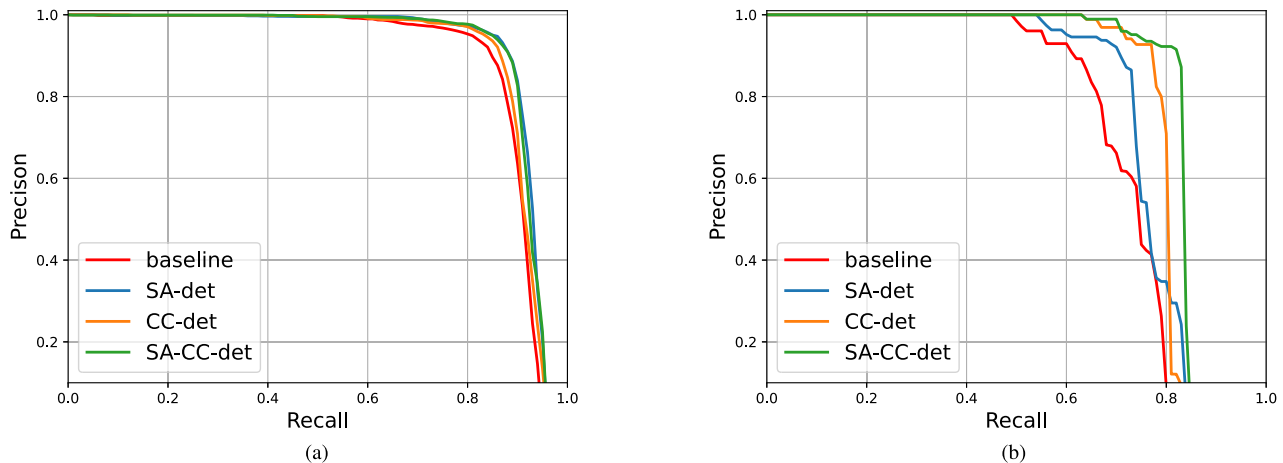


Fig. 7. P-R curves under $AP_{0.5}$ metric for different categories based on RetinaNet [11]. (a) P-R curve for category “other.” (b) P-R curve for category “A330.”

TABLE III
PRECISION AND RECALL RATE OF DETECTORS BASED ON THE RETINANET [11]
BENCHMARK FOR EACH CATEGORY

category	Precision rate / Recall rate			
	baseline	SA-det	CC-det	SA-CC-det
other	0.921/0.849	0.930/0.873	0.944 /0.849	0.936/ 0.876
A220	0.914/0.833	0.926/ 0.863	0.915/0.847	0.945 /0.856
Boeing787	0.946/0.867	0.963 / 0.898	0.936/0.881	0.956/0.896
A320/321	0.881/0.816	0.903/ 0.852	0.896/0.836	0.916 /0.842
Boeing737-800	0.927/0.862	0.945 / 0.880	0.921/0.866	0.927/0.873
ARJ21	0.904/0.852	0.915/0.873	0.924 /0.847	0.910/ 0.875
A330	0.779/0.672	0.906/0.733	0.962 /0.780	0.959/ 0.785

The confidence and IoU threshold of all the detectors are both set as 0.5. Bold entity represents the best.

the whole performance improvement, leading to the significant increase in mAP metric. We also report each category’s precision and recall rate under specific confidence and IoU threshold in Table III. Our methods can still achieve excellent performance improvement when the detector is used for a fixed scenario where the confidence and IoU threshold are both set as 0.5. We can also observe that the SA-CC-det does not consistently achieve the best precision and recall rate under specific confidence and IoU threshold. We conjecture that it is due to the hyperparameters in this experiment being obtained by monitoring the overall performance instead of the special performance. In addition to quantitative comparisons, we also plot P-R curves under $AP_{0.5}$ metric in Fig. 7 to intuitively show the improvement of the proposed methods for different categories. We select the category “other” and “A330” with the highest and lowest numbers of instances, respectively. In Fig. 7(a), the orange curve corresponding to the CC-det is always above the red curve corresponding to the baseline. However, the green curve corresponding to the SA-CC-det is almost identical to the blue curve corresponding to the SA-det. In contrast, the orange and green curves are consistently above the red and blue curves with a large margin in Fig. 7(b), respectively. This phenomenon also

TABLE IV
ABLATION STUDY FOR THE SA-FP MODULE

MSSS		AFF		mAP _{0.5}	mAP _{0.75}	mAP	param (M)	FLOPs (G)
SS	MS	ASA	ACA					
×	×	×	×	0.8674	0.7260	0.6260	36.23	52.96
✓	×	×	×	0.8781	0.7350	0.6401	36.39	53.81
✓	✓	×	×	0.8814	0.7434	0.6495	36.47	54.35
✓	✓	✓	×	0.8915	0.7494	0.6526	36.54	54.68
✓	✓	✓	✓	0.8992	0.7562	0.6592	36.55	54.71

Bold entity represents the best.

illustrates that the proposed SA-FP module is always constructive for fine-grained detection. The proposed CC-DA method is a strategy of maintaining or even sacrificing the performance for the category with more instances to improve the detection accuracy for the category with few instances.

3) *Ablation Study for the SA-FP Module:* We conducted a series of ablation studies for the SA-FP module to analyze the influence of each designed component in Table IV. Generally, the proposed SA-FP module consists of the MSSS submodule and the AFF submodule. Furthermore, the MSSS module includes sparse sampling in each level (denoted as “SS”) and regular sampling in multiple levels (denoted as “MS”). In contrast, the AFF module includes adaptive spatial attention (denoted as “ASA”) and adaptive channel attention (denoted as “ACA”). As for the baseline without the AFF module, we directly sum up all sampled pixels. It can be seen that the detection performance increases gradually with the parameters and FLOPs increasing slightly as each component is added to the baseline. This phenomenon illustrates the effectiveness of each proposed component from the SA-FP module. More specifically, the MSSS module leads to 1.4% higher $mAP_{0.5}$ than the baseline, which illustrates that the MSSS module can find fine-grained local regions from multilevel features. Finally, when adopting the MSSS module and the AFF module simultaneously, the SA-CC-det achieves 2.18%, 3.02%, and 3.32% improvement

TABLE V
IMPACT OF HYPERPARAMETERS IN THE CC-DA STRATEGY

K	mAP _{0.5}	mAP _{0.75}	mAP	params (M)	FLOPs (G)	FPS
baseline	0.8674	0.7260	0.6260	36.23	52.96	59.1
2	0.8929	0.7485	0.6534	36.42	54.04	51.5
4	0.8984	0.7552	0.6587	36.48	54.37	51.0
6	0.8992	0.7562	0.6592	36.55	54.71	50.6
8	0.9002	0.7549	0.6595	36.61	55.04	50.2

Bold entity represents the best.

TABLE VI
ABLATION STUDY FOR THE CC-DA STRATEGY

augmentation	class balance	flip	jitter	mAP _{0.5}	mAP _{0.75}	mAP
×	×	×	×	0.8674	0.7260	0.6260
✓	×	×	×	0.8775	0.7424	0.6272
✓	✓	×	×	0.8832	0.7493	0.6485
✓	✓	✓	×	0.8826	0.7546	0.6483
✓	✓	✓	✓	0.8820	0.7536	0.6477

Bold entity represents the best.

under mAP_{0.5}, mAP_{0.75}, and mAP metrics, respectively. This phenomenon demonstrates the dynamic spatial and channel attention in the AFF module, which greatly improves fine-grained representations.

4) *Hyperparameters in the SA-FP Module*: We conduct a series of experiments to verify the influence of the hyperparameters in the SA-FP module: the number of sampling points K in each feature level. As depicted in Table V, as K becomes larger, the detection performance in terms of APs metrics gradually increases. However, the parameters and FLOPs increase slightly, and the inference speed becomes slower. More importantly, when K grows up to 8, the detection performance of mAP_{0.75} drops, and that of mAP hardly changes anymore. We conjecture that this is due to the fine-grained representation no longer benefiting from the sampled local features, and the arbitrary collection of useless features does not improve the model much or even damage the feature representation. Hence, we adopt $K = 6$ in all the experiments to achieve the tradeoff between the speed and accuracy of the proposed method.

5) *Ablation Study for the CC-DA Strategy*: We conduct a series of ablation studies for the CC-DA strategy to verify whether the proposed CC-DA strategy is more efficient in dealing with the class imbalance problem than the general copy-paste data augmentation in Table VI. As a commonly used data augmentation method, the copy-paste strategy without class balance can improve detection performance. However, the improvement is relatively small, with only 0.12% gains in terms of mAP especially. After adopting the classwise probabilistic augmentation, the detector achieves 1.58%, 2.33%, and 2.25% gains under the metric of mAP_{0.5}, mAP_{0.75}, and mAP than the baseline, respectively. This improvement illustrates the effectiveness and superiority of class-balanced ideas behind the CC-DA strategy. We also adopt random flip and scale jitters used in [48] to verify the diversity of instance augmentation. It can be seen that the

TABLE VII
IMPACT OF HYPERPARAMETERS IN THE CC-DA STRATEGY

method	adjusting factor t	mAP _{0.5}	mAP _{0.75}	mAP
baseline	–	0.8674	0.7260	0.6260
CC-det	0.200	0.8750	0.7310	0.6260
	0.100	0.8790	0.7380	0.6410
	0.075	0.8810	0.7490	0.6470
	0.050	0.8832	0.7493	0.6485
	0.025	0.8790	0.7460	0.6410
	0.010	0.8766	0.7324	0.6358

Bold entity represents the best.

TABLE VIII
COMPARISON WITH SIMILAR METHODS FOR THE SA-FP MODULE BASED ON THE RETINANET [11] BENCHMARK

method	mAP _{0.5}	mAP _{0.75}	mAP	param (M)	FLOPs (G)
PANet [51]	0.8831	0.7384	0.6345	38.59	54.47
AugFPN [52]	0.8713	0.7327	0.6332	38.01	53.03
BFP [10]	0.8785	0.7374	0.6420	36.49	53.03
Dyhead [53]	0.8910	0.7406	0.6449	47.23	55.02
SA-det (ours)	0.8992	0.7562	0.6595	36.55	54.71

Bold entity represents the best.

random flip can only improve performance under mAP_{0.75} metric, while it is slightly harmful to the performance under mAP_{0.5} and mAP. After adopting the scale jitters, the performance under all the metrics further drops. We conjecture that it is mainly because these two enhancements are not suitable for airplane detection in SAR images with different imaging characteristics from natural images. Based on the above observation, we adopt the CC-DA strategy without any instance augmentation in all our experiments.

6) *Hyperparameters in the CC-DA Strategy*: Adjusting factor t is a significant hyperparameter in (7) to determine the final detection performance of CC-det. The higher the value of t , the higher the probability of copy-paste data augmentation for each category. We experimentally evaluate its impact based on the RetinaNet [11] benchmark with all the settings remaining identical except for the value of t and compare it with the baseline. We first adopt the grid search method where t is varied according to exponential decay to roughly search the range, then vary t with a step size of 0.025, and finally show detection results in Table VII. It is observed that the CC-det surpasses baseline with no class-balanced data augmentation within a large range of t , presenting high robustness. When t turns to 0.05, the CC-det achieves the highest results in different degrees.

7) *Comparison With Other Similar Methods*: We conduct comparison with other feature pyramid methods based on the RetinaNet [11] benchmark in Table VIII and other class-balanced methods based on the Faster R-CNN [8] benchmark in Table IX. For the SA-FP method, we compare it to some state-of-the-art FPN-style methods including more advanced architectures: PANet [51], AugFPN [52], and FPN with attention mechanisms for refining features: BFP [10] and Dyhead [53]. It

TABLE IX
COMPARISON WITH SIMILAR METHODS FOR THE CC-DA STRATEGY BASED ON
THE FASTER R-CNN [8] BENCHMARK

method	type	mAP _{0.5}	mAP _{0.75}	mAP
Seesaw loss [41]	Reweighting	0.8861	0.7256	0.6285
RFS [43]	Resampling	0.8851	0.7395	0.6350
Forest RCNN [45]	Resampling	0.8865	0.7389	0.6321
Instaboost [47]	Augmentation	0.8870	0.7453	0.6379
CC-det (ours)	Augmented Resampling	0.8880	0.7540	0.6410

Bold entity represents the best.









Ranking	Team	Submission Number	Submission Time	Running Time	Technical Report	Score
TOP1	 九天鹰隼 NWPU-ASGO	29	2021-10-22 12:18:29	1255s	✓	68.7004
TOP2	 星辰大海 西安电子科技...	27	2021-10-26 16:58:07	1360s	✓	68.6344
TOP3	 NWPU-RSAI 介绍: 团队主...	44	2021-10-27 16:29:36	5805s	✓	67.9002
4	 DayDream_det 来自北京航空...	33	2021-10-27 13:37:32	1946s	✓	67.3473
5	 TED&DET 真的热爱科研...	35	2021-10-25 19:46:57	1909s	✓	67.2888
6	 杨门虎将 西安电子科技...	44	2021-10-27 16:16:21	2462s	✓	67.0252
7	 MDIPL-Lab-SAR	28	2021-10-26 21:25:50	193s	✓	66.8363
8	 众里寻SAR千百度 本队伍依托于...	12	2021-10-26 11:34:56	304s	✓	66.2845

Fig. 8. Ranking of competition results for the “Fine-Grained Airplane Recognition in High-Resolution SAR Images” track in 2021 GaoFen challenge.

can be seen from Table VIII that our SA-det outperforms other methods by a large margin with only a slight increase in terms of parameters and FLOPs. Our method is elaborately designed to overcome the small interclass and the large intraclass variance which other methods do not focus on. As for the CC-DA strategy, we compare it to other class-balanced methods including Seesaw loss [41], RFS [43], Forest R-CNN [45], and Instaboost [47]. It can also be observed that our CC-det achieves the best performance under all the metrics. All these better performances can well demonstrate the superiority of the proposed two methods.

8) *Comparison With Other Methods in the GaoFen Challenge*: The 2021 GaoFen challenge provides the first related dataset in fine-grained airplane detection in SAR images. We compare the proposed SA-CC-det with other state-of-the-art methods in the “Fine-Grained Airplane Recognition in High-Resolution SAR Images” track. Fig. 8 shows the final detection performance and the ranking of our team, termed “TED&DET,” in the preliminary stage of this competition (<http://gaofen-challenge.com/indexpage>). Specifically, we adopt the proposed SA-FP and CC-DA method based on the Cascade RCNN [9] benchmark. In addition, we also adopt some commonly used tricks including mosaic [63], mix-up data augmentation [64], stochastic weight averaging [65], weighted box fusion [66], test-time augmentation, and so on. Finally, we achieved 67.28 scores and 1909 s in terms of mAP_{0.5} and inference speed in the validation dataset, respectively, and won the fifth place in the 2021 GaoFen Challenge. The ranking score in Fig. 8 can

further verify the effectiveness and superiority of the proposed two methods.

V. CONCLUSION

This article analyzes two main challenges in fine-grained airplane detection in SAR images: the small interclass variance and the large intraclass variance in complex scenes and the class imbalance problem. Correspondingly, we propose the SA-FP module and the CC-DA strategy to deal with the above two issues. Specifically, the proposed SA-FP module can simultaneously sample discriminative pixel-level local features scattered in multiscale layers and adaptively aggregate them with fine-grained attention to better classify subordinate-level airplanes with multiple scales, while the proposed CC-DA strategy randomly copies an aircraft of one category and pastes it onto an image according to the class-balanced probability for class balance in the instance level. Various experiments are conducted to demonstrate the effectiveness and superiority of these two proposed methods. We hope our method can serve as a strong baseline for future research in SAR fine-grained airplane detection. However, our proposed methods also have some limitations. Although SA-FP methods can find and leverage the discriminative pixel-level local features, the local features are only generated from the individual reference point without any explicit location information, which may be suboptimal. The CC-DA strategy directly pastes sampled instances onto an image randomly without considering the background interference and characteristics of SAR imaging. Therefore, we will consider how to effectively select more discriminative local features and generate more reasonable and higher quality SAR images in the future.

REFERENCES

- [1] F. Sharifzadeh, G. Akbarizadeh, and Y. S. Kaviani, “Ship classification in SAR images using a new hybrid CNN-MLP classifier,” *J. Indian Soc. Remote Sens.*, vol. 47, no. 4, pp. 551–562, 2019.
- [2] W. Bao, M. Huang, Y. Zhang, Y. Xu, X. Liu, and X. Xiang, “Boosting ship detection in SAR images with complementary pretraining techniques,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 8941–8954, 2021.
- [3] G. Akbarizadeh, “A new statistical-based Kurtosis wavelet energy feature for texture recognition of SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 11, pp. 4358–4368, Nov. 2012.
- [4] Z. Tirandaz and G. Akbarizadeh, “A two-phase algorithm based on Kurtosis curvelet energy and unsupervised spectral regression for segmentation of SAR images,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 3, pp. 1244–1264, Mar. 2016.
- [5] M. Rahmani and G. Akbarizadeh, “Unsupervised feature learning based on sparse coding and spectral clustering for segmentation of synthetic aperture radar images,” *IET Comput. Vis.*, vol. 9, no. 5, pp. 629–638, 2015.
- [6] Y. Tan, Q. Li, Y. Li, and J. Tian, “Aircraft detection in high-resolution SAR images based on a gradient textural saliency map,” *Sensors*, vol. 15, no. 9, pp. 23071–23094, 2015.
- [7] Tara N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for LVCSR,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8614–8618.
- [8] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
- [9] Z. Cai and N. Vasconcelos, “Cascade R-CNN: Delving into high quality object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
- [10] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, “Libra R-CNN: Towards balanced learning for object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 821–830.

- [11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
- [12] J. Redmon and A. Farhadi, "YOLOV3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [13] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "RepPoints: Point set representation for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9657–9666.
- [14] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
- [15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.
- [16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [17] W. Diao, F. Dou, K. Fu, and X. Sun, "Aircraft detection in SAR images using saliency based location regression network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2334–2337.
- [18] C. He, M. Tu, D. Xiong, F. Tu, and M. Liao, "A component-based multi-layer parallel network for airplane detection in SAR imagery," *Remote Sens.*, vol. 10, no. 7, 2018, Art. no. 1016.
- [19] L. Zhang, C. Li, L. Zhao, B. Xiong, S. Quan, and G. Kuang, "A cascaded three-look network for aircraft detection in SAR images," *Remote Sens. Lett.*, vol. 11, no. 1, pp. 57–65, 2020.
- [20] Y. Zhao, L. Zhao, C. Li, and G. Kuang, "Pyramid attention dilated network for aircraft detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 662–666, Apr. 2021.
- [21] Q. Guo, H. Wang, and F. Xu, "Scattering enhanced attention pyramid network for aircraft detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7570–7587, Sep. 2021.
- [22] Y. Zhao, L.-J. Zhao, and G.-Y. Kuang, "Attention feature fusion network for rapid aircraft detection in SAR images," *ACTA Electron. Sinica*, vol. 49, no. 9, 2021, Art. no. 1665.
- [23] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in sar images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.
- [24] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.
- [25] Z. Sun et al., "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7799–7816, 2021.
- [26] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.
- [27] F. Dou, W. Diao, X. Sun, S. Wang, K. Fu, and G. Xu, "Aircraft recognition in high resolution SAR images using saliency map and scattering structure features," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1575–1578.
- [28] S. Feng, K. Ji, L. Zhang, X. Ma, and G. Kuang, "SAR target classification based on integration of ASC parts model and deep learning algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10213–10225, 2021.
- [29] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1449–1457.
- [30] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 317–326.
- [31] S. Kong and C. Fowlkes, "Low-rank bilinear pooling for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 365–374.
- [32] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.
- [33] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1173–1182.
- [34] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-CNN: Part-based convolutional neural networks for fine-grained visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 579–590, Feb. 2022.
- [35] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 842–850.
- [36] M. Sun, Y. Yuan, F. Zhou, and E. Ding, "Multi-attention multi-class constraint for fine-grained image recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 805–821.
- [37] F. Zhang, M. Li, G. Zhai, and Y. Liu, "Multi-branch and multi-scale attention learning for fine-grained visual categorization," in *Proc. Int. Conf. Multimedia Model.*, 2021, pp. 136–147.
- [38] X.-S. Wei et al., "Fine-grained image analysis with deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, doi: 10.1109/TPAMI.2021.3126648.
- [39] Y. Han, X. Yang, T. Pu, and Z. Peng, "Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 5612318.
- [40] L. Song et al., "Fine-grained dynamic head for object detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 11131–11141.
- [41] J. Wang et al., "Seesaw loss for long-tailed instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9695–9704.
- [42] C. Feng, Y. Zhong, and W. Huang, "Exploring classification equilibrium in long-tailed object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3417–3426.
- [43] A. Gupta, P. Dollar, and R. Girshick, "LVIS: A dataset for large vocabulary instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5356–5364.
- [44] T. Wang et al., "The devil is in classification: A simple framework for long-tail instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 728–744.
- [45] J. Wu, L. Song, T. Wang, Q. Zhang, and J. Yuan, "Forest R-CNN: Large-vocabulary long-tailed object detection and instance segmentation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1570–1578.
- [46] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1301–1310.
- [47] H.-S. Fang, J. Sun, R. Wang, M. Gou, Y.-L. Li, and C. Lu, "InstaBoost: Boosting instance segmentation via probability map guided copy-pasting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 682–691.
- [48] G. Ghiasi et al., "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2918–2928.
- [49] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
- [50] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
- [51] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8759–8768.
- [52] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "AugFPN: Improving multi-scale feature learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12595–12604.
- [53] X. Dai et al., "Dynamic head: Unifying object detection heads with attentions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7373–7382.
- [54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [55] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [56] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 1971–1980.
- [57] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.
- [58] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
- [59] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*.
- [60] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, 2018.
- [61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

- [62] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [63] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOV4: Optimal speed and accuracy of object detection,” 2020, *arXiv:2004.10934*.
- [64] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” in *Proc. Int. Conf. Learn. Representations*, 2018.
- [65] H. Zhang, Y. Wang, F. Dayoub, and N. Stünderhuf, “SWA object detection,” 2020, *arXiv:2012.12645*.
- [66] R. Solovyev, W. Wang, and T. Gabruseva, “Weighted boxes fusion: Ensembling boxes from different object detection models,” *Image Vis. Comput.*, vol. 107, 2021, Art. no. 104117.



Wei Bao received the B.S. degree in communication engineering from Nanjing Tech University, Nanjing, China, in 2018, and the M.S. degree in information and communication engineering from the Beijing Institute of Technology, Beijing, China, in 2021, where he is currently working toward the Ph.D. degree.

He is also performing cooperation research with researchers with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing. His research interests include multimodal learning and remote sensing object detection.



Jingjing Hu received the Ph.D. degree in computer science from the Beijing Institute of Technology, Beijing, China.

She is currently an Associate Professor with the School of Computer, Beijing Institute of Technology. Her research interests include service computing, web intelligence, and information security.



Meiyu Huang received the B.S. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2010, and the Ph.D. degree in computer application technology from the University of Chinese Academy of Sciences, Beijing, China, in 2016.

She is currently an Assistant Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. Her research interests include machine learning, ubiquitous computing, human-computer interaction, computer vision, and image processing.



Yao Xu received the B.S. degree in electrical and computer engineering from Shanghai Jiao Tong University, Shanghai, China, in 2013, and the M.S. degree in electrical and computer engineering from the University of California Irvine, Irvine, CA, USA, in 2016.

He is currently an Assistant Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. His research interests include deep learning, data fusion, distributed systems, and computer architecture.



Nan Ji received the Ph.D. degree from the School of Mathematics and Systems Science, University of Chinese Academy of Sciences, Beijing, China, in 2019.

She is currently an Assistant Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China. Her research interests include the security of deep learning algorithm and image segmentation.



Xueshuang Xiang received the B.S. degree in computational mathematics from Wuhan University, Wuhan, China, in 2009, and the Ph.D. degree in computational mathematics from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, in 2014.

He was a Postdoctoral Researcher with the Department of Mathematics, National University of Singapore, Singapore, in 2016. He is currently an Associate Researcher with the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology. His research interests include numerical methods for partial differential equations, image processing, and deep learning.