

Attentional Feature Refinement and Alignment Network for Aircraft Detection in SAR Imagery

Yan Zhao, Lingjun Zhao, Zhong Liu, Dewen Hu, Gangyao Kuang, Li Liu

Abstract—Aircraft detection in Synthetic Aperture Radar (SAR) imagery is a challenging task in SAR Automatic Target Recognition (SAR ATR) areas due to aircraft's extremely discrete appearance, obvious intraclass variation, small size and serious background's interference. In this paper, a single-shot detector namely Attentional Feature Refinement and Alignment Network (AFRAN) is proposed for detecting aircraft in SAR images with competitive accuracy and speed. Specifically, three significant components including Attention Feature Fusion Module (AFFM), Deformable Lateral Connection Module (DLCM) and Anchor-guided Detection Module (ADM), are carefully designed in our method for refining and aligning informative characteristics of aircraft. To represent characteristics of aircraft with less interference, low-level textural and high-level semantic features of aircraft are fused and refined in AFFM throughly. The alignment between aircraft's discrete back-scattering points and convolutional sampling spots is promoted in DLCM. Eventually, the locations of aircraft are predicted precisely in ADM based on aligned features revised by refined anchors. To evaluate the performance of our method, a self-built SAR aircraft sliced dataset and a large scene SAR image are collected. Extensive quantitative and qualitative experiments with detailed analysis illustrate the effectiveness of the three proposed components. Furthermore, the topmost detection accuracy and competitive speed are achieved by our method compared with other domain-specific, *e.g.*, DAPN, PADN, and general CNN-based methods, *e.g.*, FPN, Cascade R-CNN, SSD, RefineDet and RPDet.

Index Terms—Attentional feature refinement and alignment network (AFRAN), attention feature fusion module (AFFM), deformable lateral connection module (DLCM), alignment detection module (ADM), aircraft detection, SAR images, synthetic aperture radar (SAR).

I. INTRODUCTION

SYNTHETIC Aperture Radar (SAR) has attractive imaging capabilities in nearly all weather and illumination conditions and SAR image interpretation has numerous applications in civil and military fields including surface monitoring [1]–[3], transportation management [4], [5], military surveillance [6], [7], *etc.*. As a fundamental problem in SAR image interpretation fields, target detection aims at accurately

Yan Zhao, Lingjun Zhao and Gangyao Kuang are with the State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System, National University of Defense Technology (NUDT), Changsha, 410073 China {zy18@nudt.edu.cn, nudtzhj@163.com, kuangyeats@hotmail.com}.

Dewen Hu is with the College of Intelligent Science, NUDT, Changsha, China {dwhu@nudt.edu.cn}.

Li Liu and Zhong Liu are with the College of System Engineering, NUDT, Changsha, 410073, China. {dreamliu2010@gmail.com; liuzhong@nudt.edu.cn}.

This work was supported in part by the National Natural Science Foundation of China under Grant 61872379, 62001480, 62022091, 61825305 and 61806218.

Corresponding author: Lingjun Zhao

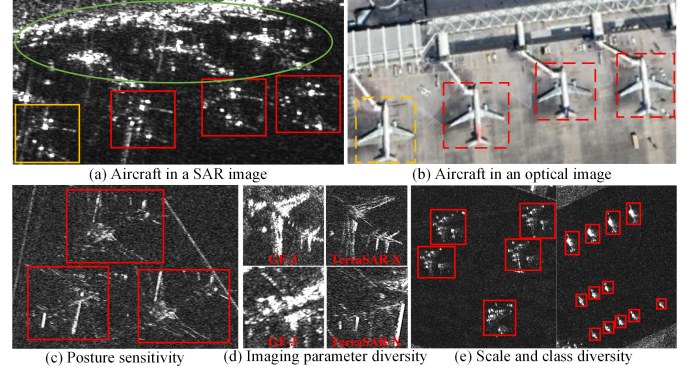


Fig. 1: Aircraft in SAR and optical images.

locating and identifying targets of interests from SAR imagery and has been researched for several decades [8]–[11]. As one of the key problems in SAR target detection, aircraft detection has potential applications in modern airport management, military reconnaissance, *etc.* and has become an independent research direction. Especially with the rapid development of SAR imaging technique, high resolution SAR images can be accessed more easily than before, enabling even more research opportunities for challenging fine-grained SAR target recognition tasks including fine-grained aircraft recognition.

The problem of SAR aircraft detection has the following main challenges.

- **Extremely discrete appearance.** Due to the smooth characteristic of aircraft's surface, the appearance of aircraft in SAR images is discrete compared to those in optical images, leading to the lack of appearance information like geometry and contour cues (see aircraft marked by the real and dotted yellow rectangles in Fig. 1 (a),(b)). Textural information and back-scattering points are important visual cues for SAR aircraft detection.
- **Large intraclass variation.** Variations in aircraft's posture, scale, category and SAR imaging parameter cause significant impact on the appearance of aircraft in SAR imagery (see Fig. 1 (c) to (e) for examples), leading to large intraclass variation. Thus, detecting aircraft with large intraclass variation calls for a precise representation of their salient and constant back-scattering information.
- **Small targets and large scene.** Different from object detection from natural images, *e.g.*, ImageNet [12] and Microsoft COCO [13], SAR images usually cover very large ranges especially those imaged by satellites and aircraft are with small size and parked sparsely (see

Fig. 11 for example), leading to very large searching spaces. Therefore, how to explore context information to locate regions of interest to increase detection speed plays a key role.

- **Uninformative background clutter.** As illustrated by the green ellipse of Fig. 1 (a), complicated background clutters like the scattering clusters of buildings are uninformative and could lead to a number of false alarm aircraft. Therefore, suppressing the disturbance of background clutters is important for accurate aircraft detection.

Earlier methods for SAR aircraft detection [14]–[16] rely on handcrafted gray-scale features and expert interpretation knowledge of aircraft, *e.g.*, gradients, structures, contours, and are sensitive to model parameters and lack of generalization to diversified situations, which result in suboptimal detection performance. Recently, Convolutional Neural Networks (CNNs) have achieved remarkable successes for object detection in optical imagery [17], [18] and also provide great opportunities for aircraft detection in SAR imagery [14], [19]–[22].

However, most of current CNN-based aircraft detection methods are directly borrowed from the field of object detection in optical imagery [23]–[26], without fully taking into consideration the aforementioned domain knowledge and challenges of SAR aircraft. That is to say, such detectors are not tailored well for aircraft detection in SAR images, leading to poor interpretability and suboptimal performance. Specifically, aircraft’s low-level textural details are not fully considered as the high-level semantic features. Also, the ability for suppressing interference of surroundings is unsatisfied, which leads to insufficient and suboptimal representation for aircraft in SAR images. Besides, it’s inappropriate for capturing aircraft’s discrete and irregular back-scattering points by a traditional convolution, of which the convolutional kernels are inflexible. Additionally, detecting sparsely parked small aircraft in large scene images calls for a carefully designed framework with sophisticated detection heads to maintain a trade-off between accuracy and speed, however, is neglected in current methods.

In response to the aforementioned challenges, a single-shot detector namely Attentional Feature Refinement and Alignment Network (AFRAN) is proposed for detecting aircraft in SAR images with balanced detection accuracy and speed. With RefineDet [27], an excellent CNN-based detector for handling diversified natural objects, as a basic prototype, the novelty of our method stands on a careful consideration of unique characteristics of aircraft, *e.g.*, structures, appearance, textures, in SAR images, and is realized by specially designing three crucial modules in AFRAN for feature refinement and alignment. For feature refinement, Attention Feature Fusion Module (AFFM), which consists of several feature aggregation pathways followed by a Split-Attention (SA) block, is designed for representing both textural and semantic features as well as highlighting significant information of aircraft adaptively. For feature alignment, Deformable Lateral Connection Module (DLCM) applied at lateral connections of a three-layer fine-grained feature pyramid, focus on aligning discrete back-scattering points of aircraft to spots of convolutional kernels with less interference involved. Furthermore, Anchor-

guided Detection Module (ADM) is attached at multi-scale feature maps to align sampling points of convolutional kernels with features at unique areas indicated by revised anchors. Our method is evaluated on a self-built SAR aircraft sliced dataset and a large scene image. Extensive quantitative and qualitative experiments illustrate the contributions of the three proposed sub-modules and the comprehensive performance of our method compared with other CNN-based detectors. Additionally, several key factors within our sub-modules are also discussed in detail.

The contributions of this paper are summarized as follows:

- (1) A single-shot detector AFRAN is proposed in this paper for aircraft detection in SAR images by carefully taking the characteristics of aircraft into consideration.
- (2) To identify and locate aircraft accurately, three significant components, *i.e.*, Attention Feature Fusion Module (AFFM), Deformable Lateral Connection Module (DLCM) and Anchor-guided Detection Module (ADM), are designed for refining low-level textural details and high-level semantic features as well as aligning discrete information of aircraft progressively.
- (3) The topmost detection accuracy with competitive speed is achieved by our method on a self-built SAR aircraft sliced dataset and a large scene SAR image compared with other CNN-based methods, in which the domain knowledge of aircraft is underutilized.

The remainders of this paper are organized as follows. In Section II, the related works for aircraft detection in SAR images are provided. In Section III, motivations and our method are introduced in detail. Experiments and analysis are given in Section IV. Finally, Section V concludes the work of this article.

II. RELATED WORKS

SAR Automatic Target Recognition (SAR ATR). SAR Automatic Target Recognition (SAR ATR) generally refers to detect and recognize target signatures from SAR data [28]–[30]. A standard architecture of SAR ATR proposed by MIT Lincoln Laboratory [31] is provided in Fig. 2, which contains three progressive stages including detection, discrimination and classification. Among the three steps, target detection acting as a footstone of the downstream tasks, aims at coarsely figuring out potential regions of interests (targets) and eliminate background areas. A discriminator is adopted to identify non-objects and objects from outputs of the detector. Finally, a classifier is constructed to identify the classes of targets by using diversified domain-specific knowledge [32]–[34].



Fig. 2: The processing flow of SAR ATR.

Aircraft detection in SAR images. Currently, methods for aircraft detection in SAR images could be coarsely divided into knowledge-based and CNN-based paradigms.

Knowledge-based methods. In these ways, apron areas are firstly located by statistic models based on gray-scale features of backgrounds. Then, potential aircraft within these regions are identified by employing various hand-crafted features according to prior knowledge of aircraft, *e.g.*, shapes, contours, attributes of back-scattering points. Among these methods, Hu *et al.* [16] introduced a Generalized Gamma Mixture Distribution (GGMD) based detector to detect aircraft in non-homogeneous backgrounds. Dou *et al.* [14] proposed a similarity measurement to identify candidates by using Kullback–Leibler Divergence (KLD) of Gaussian Mixture Model (GMM) based on saliency maps and scattering structure features of aircraft. Besides, Zhang *et al.* [15] introduced an active shape model (ASM) for contour evaluation by utilizing geometric information of aircraft. Needless to say, aircraft could be detected by designing complicated statistic models and combining various knowledge of aircraft together. However, the detection performance and generalization of these methods usually degenerates drastically where obvious mismatches exist between inflexible hand-crafted features and diversified appearance of aircraft.

CNN-based methods. The CNN-based object detectors could be finely divided into two mainstream paradigms including two-stage and one-stage detection frameworks. Inspired by R-CNN [35] and Fast R-CNN [36], Ren *et al.* [23] proposed Faster R-CNN, which laid down an original end-to-end two-stage processing framework. And the detection accuracy has been improved continually by numerous successors [25], [37], [38]. In contrast of improving detection accuracy regardless of time consumption, A considerable detection speed without much performance degradation is pursued by one-stage detectors [24], [39]–[41].

With respect to aircraft detection in SAR images by using CNN-based methods, Wang *et al.* [19] and Guo *et al.* [20] employed a CNN-based classifier to identify aircraft's candidates within suspicious areas. To regress locations of aircraft accurately, Diao *et al.* [21] employed a Fast R-CNN to discriminate aircraft. Considering partial characteristic of aircraft, He *et al.* [22] introduced a component-based detector based on YOLOv2 [26], which consists of paralleled root and part detectors, to identify and match fuselages and wings of aircraft. Besides, Guo *et al.* [42] proposed an attention pyramid network based on Feature Pyramid Network to explore scattering information enhancement (SIE) of aircraft. To handle sparse and discrete scattering points of aircraft, Zhao *et al.* [43] introduced a Pyramid Attention Dilated Network (PADN) based on RetinaNet [44] by designing a Multi-Branch Dilated Convolutional Module (MBDCM). Also, an Attention Feature Fusion Network (AFFN) [45] is designed for aircraft detection in SAR images. Undoubtedly, aircraft could be detected by these methods stably benefiting from powerful feature representation ability of CNN. However, the lacks of fully considerations of aircraft's domain-knowledge and the challenges limit a further promotion boosted by these methods.

III. METHODOLOGY

In this section, key inspirations of our method for tackling the aforementioned problems are described in subsection A.

And the holistic architecture of our method and its inner modules including AFFM, DLCM and ADM, are depicted in subsections C, D and E, respectively. Loss functions are provided at the end of this section.

A. Motivations

Feature refinement and alignment are main concentrations of our method for tackling domain-knowledge and challenges of aircraft detection in SAR images.

Feature refinement. A balanced representation of aircraft's low-level textural details and high-level semantic features with powerful suppression of interference is essential for detecting aircraft effectively. Although detailed information of small aircraft could be extracted by a low-level feature map, the semantic information it contains is rare and the correlations among different aircraft's features are weak. A wider receptive field could be obtained by a deeper neural network. However, textural details of aircraft may be indistinguishable due to a long propagation pathway, in which several intermediate convolutions involved. Instead of employing additional bottom-up pathways [46], [47], combining low-level textural details and high-level features directly could acquire multi-level aircraft's features distinctly without losing much original textural information. Additionally, serious interference caused by complex surroundings should also be resolved. Although significant features could be refined and highlighted [48]–[50], a layer-aware attention mechanism should be specially cared considering the differences of aircraft's characteristics across multi-level feature channels.

Feature alignment. A highly constant alignment between discrete back-scattering features and convolutional spots is essential for locating and identifying aircraft precisely. In comparison to traditional convolution, of which convolutional kernels' sampling locations are fixed, *e.g.*, 3×3 , deformable convolution [51], [52] could extract features within irregular areas by using convolutional kernels, of which sampling locations are tuned by trainable offsets. Hence, it's naturally suitable for representing discrete deformable features of aircraft accurately with the alleviation of deformable convolution in our method. Besides, a progressive detector striking the trade-off between accuracy and speed is practical and essential for detecting sparsely parked small aircraft in SAR images.

Shortly, the low-level textural and high-level features of aircraft are fused and emphasized throughly by feature refinement. Also, the significant discrete features of aircraft are captured precisely by feature alignment. With the aforementioned two criteria as guidelines, our method could perform competitively.

B. Overall Architecture of AFRAN

Our method belongs to a one-stage detector, of which the architecture is illustrated in Fig. 3.

Bottom-up pathway. A truncated VGG-16 network [53] acted as a backbone, is employed to extract fundamental features from the input SAR images in the bottom-up pathway. Three intermediate layers of VGG-16 including Conv4_3, Conv5_3 and Conv7, are selected as hierarchically semantic

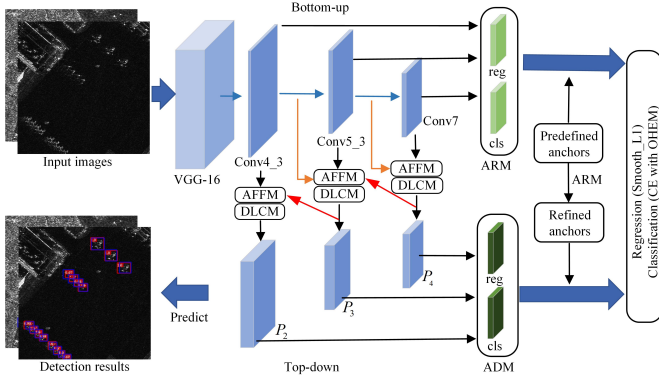


Fig. 3: The architecture of AFRAN.

features of aircraft. The spatial sizes of the three feature maps are 8, 16 and 32 times down-sampling of the original input images, which are 80×80 , 40×40 and 20×20 in pixels, respectively.

Top-down pathway. A fine-grained feature pyramid is established by combining and refining the three-layer features selectively in the top-down pathway. Three groups of AFFM are firstly adopted to fuse and filter semantic information from the three-layer feature maps. Here two feature forward propagation pathways (orange arrows in Fig. 3) are introduced to fully utilize low-level textural details of aircraft. Considering much interference around aircraft may be extracted by traditional convolution with axis-aligned kernels, three groups of DLCM are introduced following AFFM to capture aircraft's discrete characteristics adaptively. Most specifically, P_4 is set up at the first group of AFFM and DLCM by leveraging Conv_7 and Conv5_3. P_3 is constructed by the second group of AFFM and DLCM by using Conv5_3, Conv4_3 and P_4 together. And the low-level fine-grained feature map P_2 is established by fusing Conv4_3 and P_3 at the third group of AFFM and DLCM.

Coarse-to-fine detection head. Identifying aircraft progressively could alleviate mismatches between initial anchors and ground truths on location and quantity. Drawing from a coarse-to-fine detection strategy [27], an Anchor Refinement Module (ARM) and an improved object detection module namely Anchor-guided Detection Module (ADM) [54] are introduced to our method for predicting aircraft progressively. Specifically, refined anchors are firstly acquired by adjusting initial anchors based on parameterized locations and confidences, which are predicted by feeding Conv4_3, Conv5_3 and Conv7 to ARM. Aircraft are predicted eventually by feeding features, where locations at P_2 , P_3 and P_4 are calculated by refine anchors, to ADM subsequently.

To optimize our method, a multi-task loss including binary and multi-class cross entropy losses with Online Hard Example Mining (OHEM) [55] and Smooth L1 loss, is employed for supervising classification and regression branches of ARM and ADM, simultaneously.

C. Attention Feature Fusion Module

To leverage aircraft's low-level textural details and semantic features harmoniously, an Attention Feature Fusion Module (AFFM) as illustrated in Fig. 4, is introduced when building the fine-grained feature pyramid of AFRAN.

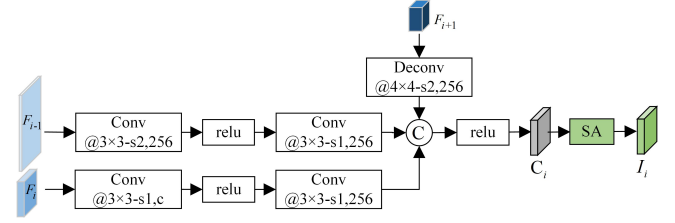


Fig. 4: Structure of AFFM.

Multi-level feature aggregation. To acquire multi-dimension features, a 3×3 convolution with stride of 2 followed by a relu activation function is firstly applied at the $i - 1$ th basic feature map F_{i-1} to reduce its spatial size for feature fusion. Meanwhile, the $i + 1$ th basic feature layer F_{i+1} is up-sampled by a 4×4 deconvolution to acquire abundant and discriminative semantic features, which are helpful for detecting small aircraft from low-level feature layers. To satisfy requirements of different modules, one and two layers of 3×3 convolution layers with a stride of 1 are attached at the $i - 1$ th feature map F_{i-1} and the i th feature map F_i , respectively, to maintain discrepancy as well as relaxing interference of different feature stages. After that, the modified F_{i+1} , F_{i-1} and F_i with the same spatial sizes are obtained. After concatenating (\odot) the three intermediate features along channel dimension, an intermediate feature map C_i containing abundant textural and semantic features of aircraft, is obtained. Specifically, Conv5_3 and Conv7 are adopted to build the topmost intermediate feature C_4 . The middle intermediate feature C_3 is constructed by leveraging Conv4_3, Conv5_3 and P_4 . The low-level intermediate feature map C_2 is obtained by combining Conv4_3 with P_3 finally.

Feature refinement by Split-Attention. Although abundant features of aircraft are accumulated after concatenation, gaps across different feature layers are still obvious. Also, interference of uninformative background around aircraft may also be collected. To highlight significant features of aircraft and suppress background interference at different feature layers carefully, a layer-aware attention mechanism namely Split-Attention (SA) [56] is employed following the concatenated feature C_i . The structure of SA is illustrated as Fig. 5.

Let $C_i \in \mathbf{R}^{h \times w \times c}$ denotes a concatenated feature block of r feature maps, where h , w , c and r refer to block's height, width, channel number and the number of concatenated features within AFFM, respectively in our method. To acquire refined feature block I_i , r groups of individual feature maps (K_1, K_2, \dots, K_r), of which numbers of channels all equal to c , are firstly generated by sending C_i to a convolution (Conv) with $c \times r$ filters and splitting along channel dimension. Then, these r groups of features are added element-wisely (\oplus) and sent into an adaptive two-dimension average pooling (Avgpool) and fully connection (FC) in sequence to produce

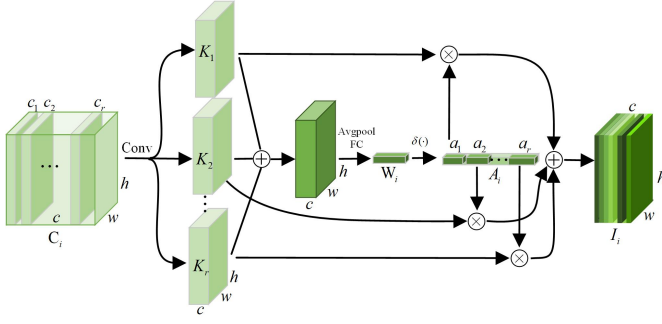


Fig. 5: Structure of SA.

a weighted vector \mathbf{W}_i . It represents contributions of different channels of the concatenated feature map. Then, r groups of channel attention weights a_1, a_2, \dots, a_r derived from \mathbf{A}_i , is acquired by applying a Softmax function (σ) on \mathbf{W}_i . Finally, the layer-aware refined feature map $\mathbf{I}_i \in \mathbb{R}^{h \times w \times c}$ could be acquired by an element-wise addition (\oplus) of r groups of productions (\otimes) between K_r and the corresponding vector a_r .

D. Deformable Lateral Connection Module

Representing discrete yet constant features of aircraft by a suitable convolution is essential for accurate aircraft detection in SAR images. As a toy example depicted in Fig. 6, red and blue points refer to sampling locations of convolutional kernels for aircraft and surroundings, respectively.

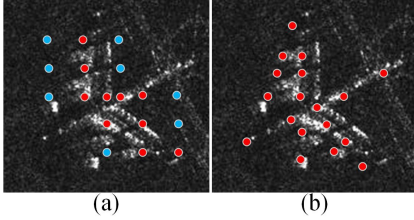


Fig. 6: Sampling locations of convolutional kernels for aircraft and surroundings by different types of convolution. (a) locations sampled by vanilla convolutional kernels. (b) locations sampled by deformable convolutional kernels.

As shown in Fig. 6 (a), back-scattering information of surroundings around aircraft is easily captured by a traditional convolutional operation due to its regular and rigid sampling strategies. However, as shown in Fig. 6 (b), a deformable convolution [51], [52] is naturally appropriate for capturing discrete and significant features of aircraft without introducing much background interference benefiting from its deformable convolutional strategy. Based on the superiority, three groups of Deformable Lateral Connection Module (DLCM) with several deformable convolution stacked, as depicted in Fig. 7, are designed in our method for capturing aircraft's discrete features following AFFM.

As illustrated by the blue and green rectangles of Fig. 7, two vanilla 3×3 convolution with stride 1 outputs 2 and 1 are firstly attached on the input feature map X to generate a two-dimension offset map Δp_k and a score mask

Δm_k , respectively. Therefore, revised sampling locations of the deformable convolution could be acquired by adjusting original axis-aligned convolutional spots through Δp_k . And discrete deformable features of aircraft are generated after modulating deformable values with the score mask Δm_k . In our method, flexible and irregular information of aircraft could be represented effectively by stacking several deformable convolution in DLCM sequentially.

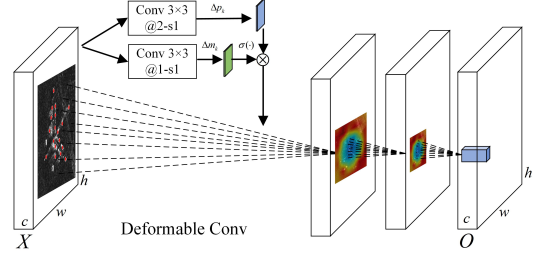


Fig. 7: Structure of DLCM.

Concisely, the deformable convolution is formulated in Eq. (1). Here $\mathbf{X}(p_n + p_k + \Delta p_k)$ refers to deformable input features. They are acquired by adjusting original eight-neighborhood features $\mathbf{X}(p_n + p_k)$ of $\mathbf{X}(p_n)$ using offsets Δp_k , where $k = (i, j)$ $i, j \in -1, 0, 1$ for a 3×3 convolutional kernel. The final output matrix $\mathbf{Y}(p_n)$ is obtained after mapping and modulating $\mathbf{X}(p_n + p_k + \Delta p_k)$ through trained weights \mathbf{W}_k and modulation factor Δm_k simultaneously.

$$\mathbf{Y}(\mathbf{p}_n) = \sum_k^K \mathbf{W}_k \cdot \mathbf{X}(\mathbf{p}_n + \mathbf{p}_k + \Delta \mathbf{p}_k) \cdot \Delta m_k \quad (1)$$

E. Anchor-guided Detection Module

Due to the lack of well-designed feature cropping strategies, *e.g.*, ROI Align [57], ROI Wrapping [23] leveraged in two-stage detectors, one-to-many mappings and obvious misalignments between anchors and features in one-stage detectors easily incur inaccurate identification and location for small aircraft. Instead of representing aircraft by a set of representative points [58] or features aligned by initial anchors [59], an Anchor-guided Detection Module (ADM) derived from [54], is adopted in our method for establishing a strict single-mapping between refined anchors and features. Most specifically, unique features calculated by refined anchors are utilized for identifying aircraft ultimately.

As illustrated in Fig. 8, blue rectangle refers to a refined anchor generated by adjusting locations and suppressing redundancy of initial anchors based on parameterized offsets (dx, dy, dw, dh) and confidence produced by ARM. It is distinct that features covered by a refined anchor are more representative than those covered by a initial anchor (the green rectangle in Fig. 8) resulting from an improved overlap between the revised anchor and the ground truth aircraft. Thus, a higher correspondence between features captured by the revised sampling points $S_{i,j}^*(X, Y)$ (9 blue points in Fig. 8) and the aircraft could be established than before. Based on advantages of deformable convolution, two-dimension offsets $O_{i,j}(X, Y)$ between revised sampling points ($S_{i,j}^*(X, Y)$) and

the related original sampling points ($S_{i,j}(X, Y)$) could be calculated and acted as deformable convolutional offsets for final regression and classification.

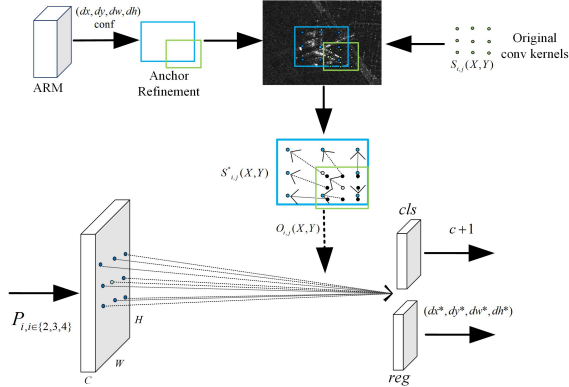


Fig. 8: Structure of ADM.

Concisely, the processing flows of ADM are formulated as Eq. (2), (3) and (4). The original sampling points $S_{i,j}(X, Y)$ are formulated as Eq. (2), where $i, j \in \{0, 1, k-1\}$ for a 3×3 convolution.

$$S_{i,j}(X, Y) = \left(X - \left\lfloor \frac{k}{2} \right\rfloor + i + 0.5, Y - \left\lfloor \frac{k}{2} \right\rfloor + j + 0.5\right) \quad (2)$$

The corresponding aligned sampling points $S_{i,j}^*(X, Y)$ at feature maps \mathbf{F} are formulated as Eq. (3), where x_1, y_1, x_2, y_2 refer to the locations of refined anchors.

$$S_{i,j}^*(X, Y) = \left(\frac{kx_1 + (x_2 - x_1)(i + 0.5)}{kS}, \frac{ky_1 + (y_2 - y_1)(j + 0.5)}{kS} \right), \quad i, j \in 0, 1, 2, 3 \dots k-1 \quad (3)$$

The two-dimension offsets for a deformable convolution formulated as Eq. (4), are obtained by subtracting Eq. (3) from Eq. (2).

$$\begin{aligned} O_i(X) &= S_{i,j}^*(X) - S_{i,j}(X) \\ &= \frac{x_1}{S} - X + \left\lfloor \frac{k}{2} \right\rfloor + \left(\frac{x_2 - x_1}{S} - 1 \right) (i + 0.5) \\ O_j(Y) &= S_{i,j}^*(Y) - S_{i,j}(Y) \\ &= \frac{y_1}{S} - Y + \left\lfloor \frac{k}{2} \right\rfloor + \left(\frac{y_2 - y_1}{wS} - 1 \right) (j + 0.5) \end{aligned} \quad (4)$$

F. Loss Functions

The overall loss for optimizing our network provided by Eq. (5), contains losses of ARM and ADM. As defined by Eq. (6), each part of Eq. (5) is a weighted sum of the classification loss (L_{conf}) between predicted classes (x) and truth labels (c) and the regression loss (L_{reg}) among predicted classes (x), bounding boxes (l) and ground truth locations (g).

$$L_{total} = L_{ARM} + L_{ADM} \quad (5)$$

$$L(x, c, l, g) = \frac{1}{N} (L_{conf}(x, c) + \alpha L_{reg}(x, l, g)) \quad (6)$$

Where N refers to the number of positive anchors matched to any ground truth objects. As defined by Eq. (7), L_{conf} is a binary or multi-class cross entropy loss between the predicted classes x and ground truth labels c .

$$L_{conf}(x, c) = - \sum_{i \in \text{Pos}} x_{i,j}^p \log(\hat{c}_i^p) \quad (7)$$

The Iverson bracket indicator function $x_{i,j}^p$ is 1 where the i th anchor matches to j th ground truth object for class p otherwise 0. \hat{c}_i^p refers to a Softmax loss defined as Eq. (8).

$$\hat{c}_i^p = \frac{\exp(c_i^p)}{\sum_p \exp(c_i^p)} \quad (8)$$

For regression, L_{reg} formulated as Eq. (9), refers to a Smooth L1 loss of offsets between predicted bounding boxes (l) and ground truths (g), where c_x, c_y, w and h are centers, widths and heights of initial or refined anchors and matched ground truths.

$$L_{reg}(x, l, g) = \sum_{i \in \text{Pos}} \sum_{m \in \{c_x, c_y, w, h\}} x_{i,j}^k \text{SmoothL1}(l_i^m - \hat{g}_j^m) \quad (9)$$

In training phase, the initial anchors are firstly revised by predictions of ARM. Only refined anchors with higher confidence than a preset threshold θ will contribute loss of ADM.

IV. EXPERIMENTS AND ANALYSIS

A. Datasets and Experimental Setup

Datasets descriptions. Since there is no publicly available dataset for aircraft detection in SAR images, a self-built aircraft sliced dataset and a large scene SAR image are collected for investigating the detection performance of our method in the experiments.

a) **Aircraft sliced dataset.** A self-built aircraft sliced dataset is constructed by using 174 large scene SAR images collected from Chinese GF-3 and German TerraSAR-X satellites. The large scene images captured by GF-3 satellite working in C-band HH and VV polarization and SpotLight (SL) observing mode are with nominal resolution of 1.0 meter. Another images captured by TerraSAR-X satellite working in X band and SpotLight (SL) mode are with nominal resolution of 0.5 meter. The ground truth aircraft were manually annotated by experts of SAR ATR by considering both prior knowledge and the corresponding optical images. After random cropping, 2317 non-overlapped 640×640 slices are collected with 6781 aircraft, of which structures, outlines and main components are clear and wings ranging from about 15 meters to 75 meters in total. The distributions of bounding boxes' sizes in pixels and aspect ratios are given by Fig. 9 (a), (b), respectively. Also, some slices of our dataset are provided as Fig. 10 (a) and (b), in which real aircraft encompassed by blue rectangles are imaged by GF-3 and TerraSAR-X satellites, respectively.

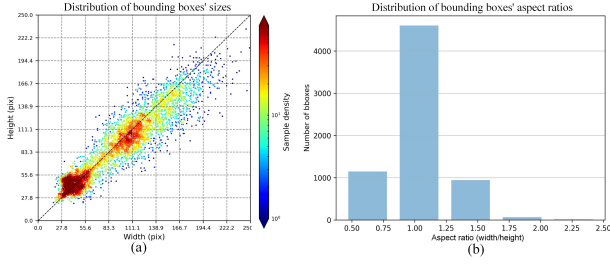


Fig. 9: Distributions of bounding boxes' sizes and aspect ratios of the self-built aircraft sliced dataset.

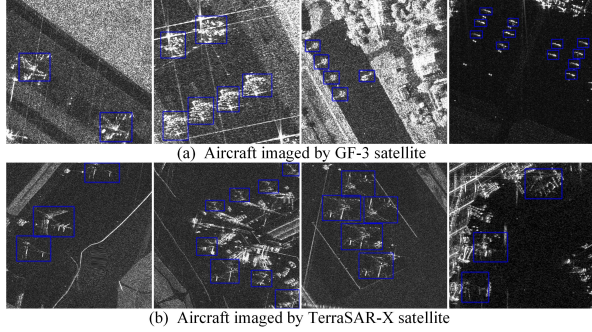


Fig. 10: Slices of SAR images in the self-built aircraft sliced dataset.

- b) **Large scene SAR image.** A large scene image is also adopted for evaluating the performance of our method. The left and the right subfigures of Fig. 11 illustrate the large scene image including an airport and corresponding optical image cropped from Google Earth software, respectively. Some imaging parameters are listed concisely in Table I



Fig. 11: A large scene for aircraft detection.

TABLE I: Imaging parameters of the large scene image

Parameter name	Value
Product level	Level 2
Product serial number	3253042
Product resolution	1.0 (meter)
Imaging central location	(99.82525 W, 32.412828 E)
Product height and width	25784 × 23161 (pixel)

Implementation details. The training, valid and test sets are constructed by dividing the original slices with a ratio of 5:2:3. Thus, numbers of images in the three sets are 1158, 463, and 696, respectively. Data augmentation strategies including contrast, illumination distortion, mirroring, random flipping, expanding and cropping, are adopted to increase diversity of aircraft in training set. To maintain considerable overlaps between initial anchors and aircraft, the basic anchor scales for P_2 , P_3 and P_4 of our method are set to 32, 64 and 128, respectively. Also, three aspect ratios $\{0.5, 1.0, 2.0\}$ are assigned to each anchor for all pyramid levels. Our method is trained for 200 epochs with a minibatch of 4. The initial learning rate is $1e-3$ and decayed at 75 and 150 epochs with rate (γ) 0.1. To achieve a stable convergence, warm-up is enabled at the first 5 epochs. Stochastic Gradient Descent (SGD) with weight decay rate $5e-4$ and momentum 0.9 is adopted to optimize network parameters. Other compared methods are implemented based on mmdetection framework [60].

Evaluation metrics. In all experiments, six average precision indicators from Microsoft COCO [13] are adopted including AP , $AP^{.5}$, $AP^{.75}$, AP^s , AP^m and AP^l , to judge performance of different methods. Moreover, $AP^{.5}$ and $AP^{.75}$ evaluate average precision scores at 0.5 and 0.75 Intersection of Union (IoU) thresholds between predictions and ground truths, respectively. AP^s , AP^m and AP^l refer to average precision scores of methods for detecting small, middle and large aircraft averaged by ten IoU thresholds (0.5, 0.55, ..., 0.95). Additionally, precision (P), recall (R) and F score (F_1) are also adopted. The six AP metrics are calculated at 0.05 confidence threshold. Precision, recall and F_1 are acquired at 0.5 confidence threshold. All these metrics are acquired at 0.5 IoU threshold. In terms of time and space complexities, frame-per-second (FPS), model parameter volumes (Params) and multiply-accumulate operations (MAC) are employed, of which the last two metrics are defined as Eq. (10) and (11), respectively.

$$\text{Params} = C_{out} \cdot (k_w \cdot k_h \cdot C_{in} + 1) \quad (10)$$

C_{out} , C_{in} , k_w and k_h in Eq. (10) are the output, input channels and kernel sizes of a convolution.

$$\text{MAC} = C_{out} \cdot C_{in} \cdot k_w \cdot k_h \cdot H_o \cdot W_o \quad (11)$$

Here k_w , k_h , C_{out} and C_{in} are width, height, output and input channels of a convolution. H_o , W_o are spatial sizes of an output feature map.

Additionally, precision-recall curves of different AP metrics and visual detection results are also provided for further judgements.

B. Ablation Studies

1) **Quantitative analysis:** The quantitative contributions of different inner modules are evaluated by enabling them progressively and the results are shown in Table II. A baseline detector is constructed by removing the three modules from AFRAN but maintaining the basic architecture, i.e., the three-layer fine-grained feature pyramid. Besides, we disentangle AFFM into feature fusion especially feature forward (orange

TABLE II: Effects of Inner Modules.

Models	FF	SA	DLCM	ADM	P	R	F_1	AP	$AP^{.5}$	$AP^{.75}$	AP^s	AP^m	AP^l	FPS
RefineDet	-	-	-	-	0.815	0.935	0.871	0.530	0.932	0.547	0.388	0.521	0.549	60
Baseline*	-	-	-	-	0.829	0.933	0.878	0.520	0.933	0.535	0.385	0.515	0.536	63
	✓	-	-	-	0.827	0.926	0.874	0.522	0.932	0.517	0.371	0.508	0.543	58
	✓	✓	-	-	0.871	0.931	0.900	0.536	0.941	0.563	0.427	0.524	0.555	50
	✓	✓	✓	-	0.899	0.927	0.913	0.538	0.929	0.572	0.394	0.526	0.564	46
AFRAN (Ours)	✓	✓	✓	✓	0.904	0.932	0.918	0.554	0.941	0.597	0.481	0.537	0.576	40

* The baseline detector is a modified AFRAN, which only contains three layers of multi-scale features for prediction. By enabling different sub-modules progressively, performance for detecting aircraft in SAR images is improved. Bold texts indicates the best value in each row.

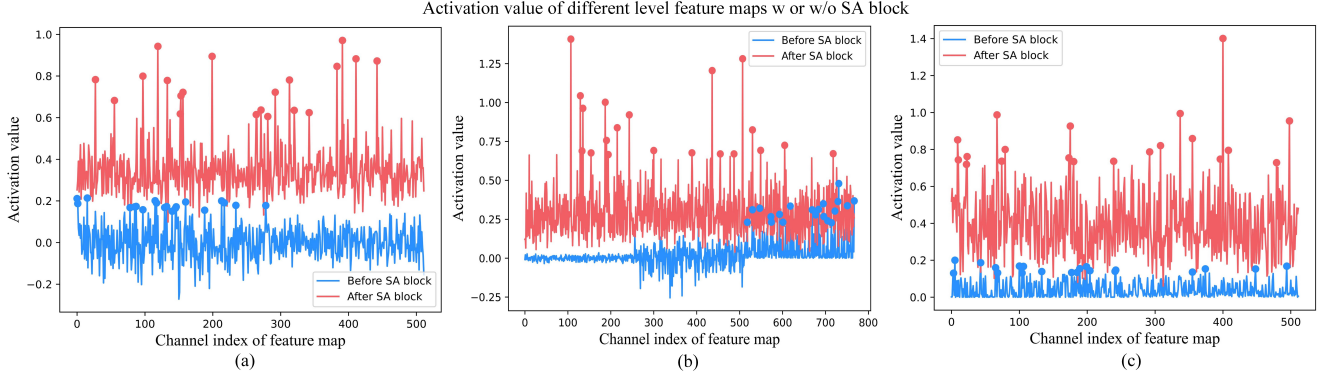


Fig. 12: Activation values of different levels of fine-grained feature maps w/ or w/o SA blocks. Activation values of low-level, middle-level and high-level fine-grained feature maps are plotted in sub-figures (a), (b) and (c), respectively.

arrows in Fig. 3) and Split-Attention block, donated as FF and SA, and explore their contributions to aircraft's textural feature extraction and significant feature refinement in detail. Considering the similar architecture and processing scheme between RefineDet and our method. The detection results of RefineDet are also provided at the first column of Table II. According to Table II, several insightful findings could be summarized as follows.

Baseline. The high-level semantic features play an important role for identifying and locating large aircraft. In comparison to the detection results of RefineDet, precision rate of the Baseline increases to 0.829 and is 1.4% higher than that of RefineDet at 0.5 IoU threshold. It may be because that less background information is accumulated at low-level feature layers through the top-down pathway, which reduces interference for aircraft detection to some extent. However, aircraft's high-level semantic features extracted by the Baseline is restricted resulting from the three-layer feature pyramid constructed by a truncated VGG-16 of the Baseline, which weaken its ability for identifying aircraft accurately. Therefore, distinct decreases emerge on AP , $AP^{.75}$ and AP^l when detecting aircraft by Baseline, which are 1.0%, 1.2% and 1.3% lower than those of RefineDet.

Effects of FF. By merely propagating aircraft's textural details forwardly, uninformative background interference may also be accumulated further, which leads to much confusion for aircraft discrimination. Specifically, score of $AP^{.75}$ achieved by Baseline (w/ FF) (the third column of Table II) decrease obviously and is 1.5% lower than that of Baseline.

Effects of SA. Aircraft's low-level textural details could be utilized effectively benefiting from powerful feature refinement

ability of SA blocks, which promote the detection performance of our method greatly. In comparison to the detection results of Baseline (w/ FF) and Baseline (w/ FF+SA) (the third and the fourth columns of Table II), scores of precision, F_1 , AP , $AP^{.75}$, AP^s , AP^m and AP^l achieved by Baseline (w/ FF+SA) are 4.4%, 1.6%, 1.4%, 4.6%, 5.6% 1.6% and 1.2% higher than those of Baseline (w/ FF). It fully verifies effectiveness of SA blocks for refining aircraft's significant features and suppressing uninformative background. To inspect effects of SA blocks further, values of different fine-grained feature channels w/o or w/ SA blocks are also provided in Fig. 12 in blue and red curves, respectively. Meanwhile, the top-k ($k=20$) corresponding feature channels sorted by activation values are also marked out with blue and red real dots. Obviously, peaks of feature channels after SA blocks (red points) distribute sparsely, however, densely and locally for peaks of initial feature channels, which illustrates that a wider range of aircraft's significant features across multi-level feature maps could be boosted. Especially for building the middle-level fine-grained feature map, the first 256 channel indexes at Fig. 12 (b) are enhanced and a more balanced feature map could be obtained after enabling SA block, which proves powerful feature refinement ability of SA block.

Effects of DLCM. Feature alignment by DLCM at lateral connections of the fine-grained feature pyramid promotes our method's ability for locating aircraft especially small aircraft further. Benefiting from powerful deformable modeling ability of DLCM, aircraft's high-level discrete features could be perceived accurately, leading to improvements on precision (P), $AP^{.75}$ and AP^l . However, scores of AP^s encounters a sharp decrease and is 3.3% lower than that of Baseline

(w/ FF+SA). It might be because of insufficient semantic information and complex background interference exists at low-level feature maps, which restricts DLCM's ability for capturing aircraft's discrete and significant information.

Effects of ADM. A tight correspondence between refined anchors and features is constructed by Baseline (w/ FF+SA+DLCM+ADM), which significantly boosts performance of our method for detecting aircraft. Compared with detection results of Baseline (w/ FF+SA+DLCM) (the fifth row of Table II), of which the detection head is the same as that of RefineDet, remarkable improvements on all indicators are acquired by Baseline (w/ FF+SA+DLCM+ADM) (the last column of Table II). Specifically, scores of AP , $AP^{.75}$, AP^s are 1.6%, 2.5% and 8.7% superior to those of Baseline (w/ FF+SA+DLCM), respectively.

Furthermore, number of positive anchors, values of average and maximal IoU between initial anchors and ground truths or refined anchors and ground truths are given by Fig. 13. Obviously, more positive anchors with higher IoU score are acquired by refined anchors than those of initial anchors. The average maximal IoU score between ground truths and refined anchors is 0.870 yet 0.653 for initial anchors, which further indicates positive guidance of refined anchors for ADM.

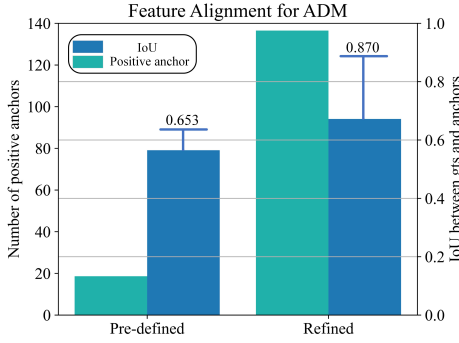


Fig. 13: Feature alignment of refined anchors for ADM. Blue and green bars show average IoU scores and positive anchor numbers achieved by initial and refined anchors at 0.5 IoU threshold, respectively. The horizontal lines indicate the maximal IoU scores acquired by initial and refined anchors.

2) **Qualitative analysis:** To further investigate evolution of features maps acquired by the baseline equipped with the three sub-modules progressively, aircraft detection results and heat maps of specific fine-grained feature maps including P_3 and P_4 , are provided in Fig. 14. Some observations could be summed up as follows.

Clear boundaries between aircraft and background are obtained by Baseline (w/ FF). Specifically, activation boundaries of small aircraft and surroundings at P_3 of Baseline (w/ FF) (the third column of Fig. 14) are more distinct than those of the Baseline (the second column of Fig. 14). Also, aircraft's activation areas are converged at P_3 of Baseline (w/ FF) yet divergent at identical feature areas generated by RefineDet (the first column of Fig. 14). It might be because that fine-grained features could be enhanced at P_3 further by fusing aircraft's local textural details from P_2 . Besides, less uninformative background information is propagated from

topper feature maps, *i.e.*, P_4 in our method, which improves our method's ability for identifying aircraft and surroundings clearly.

Significant components of aircraft are specifically highlighted by Baseline (w/ FF+SA). According to the fourth column of Fig. 14, activation areas of small aircraft (dark blue areas encompassed by orange dotted rectangle) are distinct and the background are homogeneous (areas in green color) by enabling SA blocks. It might be because of the effective layer-aware feature refinement ability provided by SA blocks that consistency of background and specificity of aircraft at different feature maps could be maintained when fusing and refining low-level textural details and high-level semantic characteristics of aircraft.

Clear separations among densely arranged aircraft are obtained by Baseline (w/ FF+SA+DLCM). According to the fifth column of Fig. 14, separations among densely parked small aircraft are clearer than those at the same feature map P_3 of Baseline (w/ FF+SA). Besides, activation areas of large aircraft at P_4 are more centered than those extracted by Baseline (w/ FF+SA), which promotes our model to locate aircraft precisely (also proved in Table II). It might be because that DLCM could capture irregular information of aircraft adaptively and avoid extracting much uninformative background interference by employing deformable convolutions instead of traditional convolutions equipped with axis-aligned kernels.

Hierarchy and consistency of aircraft's features at multi-scale feature maps are promoted by Baseline (w/ FF+SA+DLCM+ADM). As illustrated by the last column of Fig. 14, activation areas of small aircraft are more consistent than those at the same feature maps of Baseline (w/ FF+SA+DLCM). Also, aircraft with different scales could be represented hierarchically by different feature levels, which decreases aliasing across different feature maps. Specifically, activation areas of small and large aircraft are represented distinctly at P_3 and P_4 , respectively.

C. Comparing with CNN-based methods

1) **Quantitative analysis:** The performance of our method are also evaluated by comparing with other CNN-based detectors on test set of the self-built aircraft sliced dataset and a large scene image. Moreover, two methods specially designed for object detection in SAR images including PADN [10] and DAPN [61], are also re-implemented and compared together.

Detection accuracy. The detection accuracy of different methods on the test set of the self-built aircraft sliced dataset are provided in Table III. Specifically, AP and F_1 achieved by our method are 1.1% and 2.5% higher than those of the second-place methods (RPDet, Cascade R-CNN), respectively. Although a slight superiority on $AP^{.75}$ and AP^l is achieved by RPDet benefiting from its point-based aircraft representation strategy. However, its ability for identifying small aircraft is still unsatisfied. Besides, $AP^{.5}$ and AP^s scores achieved by RPDet are 5.5% and 6.6% lower than those achieved by our method, which are 0.941 and 0.481, respectively. It might be because of insufficient features of small aircraft extracted by the class-agnostic backbone. Obviously, competitive

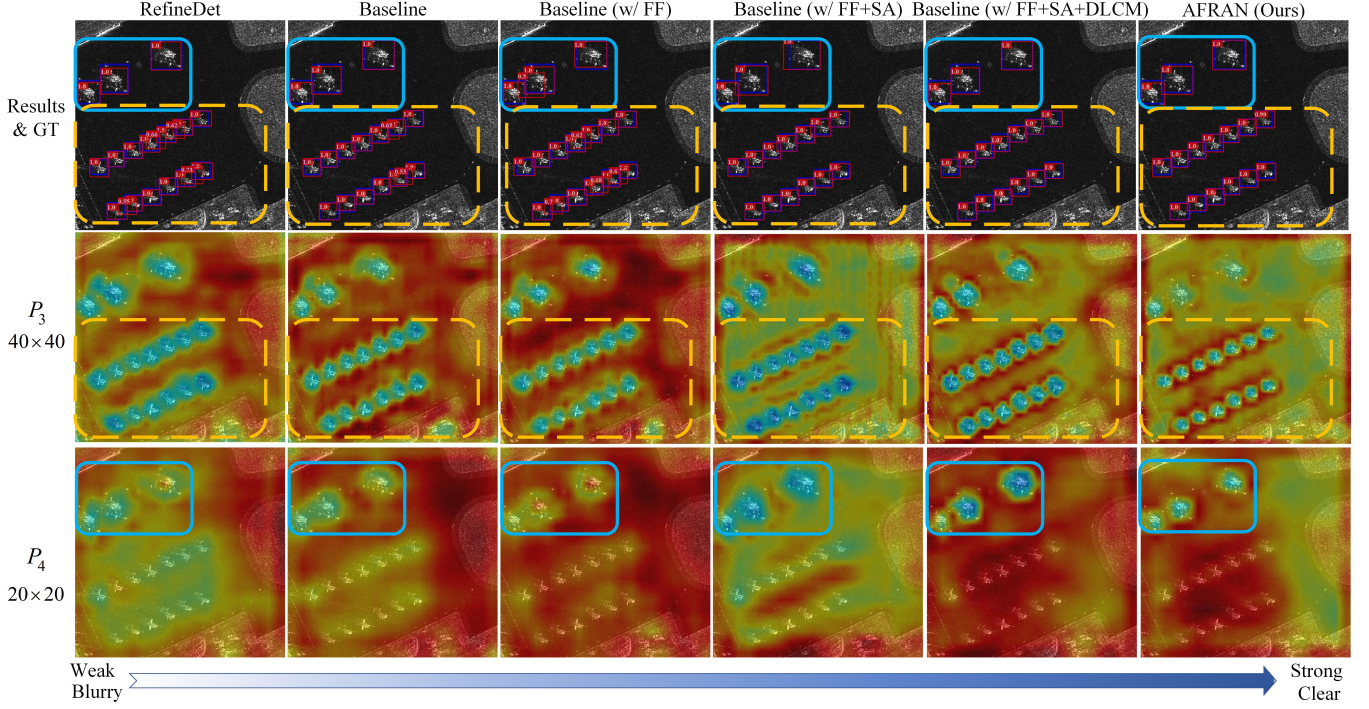


Fig. 14: Detection results and activation maps of specific fine-grained feature maps produced by our method equipped with the proposed sub-modules progressively.

TABLE III: Detection results of different CNN-based methods on test-set of aircraft sliced dataset.

Models	P	R	F_1	AP	$AP^{.5}$	$AP^{.75}$	AP^s	AP^m	AP^l
Two-stage methods									
Faster R-CNN [23]	0.893	0.890	0.892	0.520	0.877	0.567	0.331	0.495	0.561
FPN [25]	0.870	0.909	0.889	0.530	0.886	0.598	0.366	0.509	0.562
Cascade R-CNN [37]	0.898	0.889	0.893	0.539	0.876	0.594	0.398	0.517	0.577
DAPN [61]	0.885	0.830	0.857	0.448	0.851	0.413	0.250	0.420	0.489
One-stage methods									
SSD [24]	0.889	0.889	0.889	0.532	0.932	0.562	0.371	0.497	0.575
PADN [10]	0.842	0.916	0.878	0.512	0.906	0.518	0.308	0.480	0.554
RefineDet [27]	0.815	0.935	0.871	0.530	0.932	0.547	0.388	0.521	0.549
RPDet [58]	0.867	0.917	0.891	0.543	0.886	0.612	0.415	0.517	0.584
Ours (AFRAN)	0.904	0.932	0.918	0.554	0.941	0.597	0.481	0.537	0.576

TABLE IV: Running time and model complexity of different CNN-based methods

Models	Backbone	Input size	FPS	Params(M)	MAC(G)
Two-stage methods					
Faster R-CNN [23]	ResNet-101	$\sim 1333 \times 800$	6	51.75	921.07
FPN [25]			16	60.04	290.35
Cascade R-CNN [37]			15	60.4	296.42
DAPN [61]			12	128.23	335.46
One-stage methods					
SSD [24]	VGG-16	$\sim 512 \times 512$	90	25.22	88.84
PADN [10]	ResNet-101	$\sim 640 \times 640$	33	102.16	247.12
RefineDet [27]	VGG-16	$\sim 640 \times 640$	63	34.05	150.00
RPDet [58]	ResNet-101	$\sim 1333 \times 800$	17	55.59	278.44
Ours (AFRAN)	VGG-16	$\sim 640 \times 640$	45	35.82	150.59

performance on metrics for evaluating location accuracy are achieved by most of the two-stage detectors benefiting from their advanced feature alignment strategies in comparison with classical one-stage detectors, *e.g.*, SSD, RefineDet. However, our method could still locate aircraft competitively benefiting from its powerful feature alignment achieved by DLCM and ADM. Additionally, an obvious promotion is obtained by our method comparing with the two domain-specific methods including DAPN and PADN at all indicators. It might be because of their sub-optimal architectures and limited utilization for aircraft's low-level details, which incurs a sharp degeneration for aircraft detection in SAR images.

Time and spatial complexities. The detection speed and spatial complexities are evaluated and given by Table IV. Clearly, benefiting from truncated VGG-16 backbone as well as balanced three-layer feature pyramid, our method only takes up 35.82M and 150.59G on parameter volumes (Params) and multiply-accumulate operations (MAC), which is lower than those required by numerous two-stage detectors (Faster R-CNN, FPN, Cascade R-CNN, DAPN) and some one-stage detectors (RPDet, PADN). Meanwhile, our method achieves higher detection speed than RPDet and all two-stage detectors, which demonstrates its feasibility for real-time applications. However, our method runs slower than SSD and RefineDet causing by numerous parameters introduced by its complex sub-modules.

2) **Qualitative analysis:** To inspect trends of detection performance, precision-recall curves of all compared CNN-based methods under various conditions of IoU thresholds and aircraft's sizes are given by Fig. 15. Undoubtedly, curve of our method (red curves in Fig. 15 (a) to (f)) decreases much slightly than those of other methods with the increase of recall rate. Specifically, distinct advantages of our method exist at the conditions of 0.5 and 0.5@0.05:0.95 IoU thresholds, *i.e.*, the Fig. 15 (a) and (b). Besides, performance of our method is superior to other methods by a large margin for detecting small aircraft according to Fig. 15 (d). It might be because of the fully exploration and refinement of aircraft's multi-level features in AFFM, which further illustrates the powerful discrimination ability for aircraft in SAR images achieved by our method.

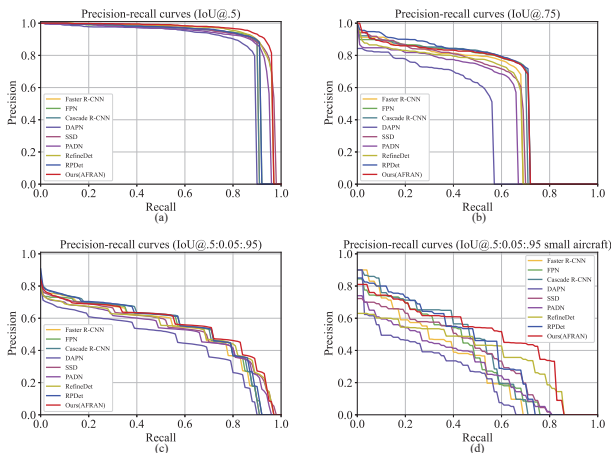


Fig. 15: Continued.

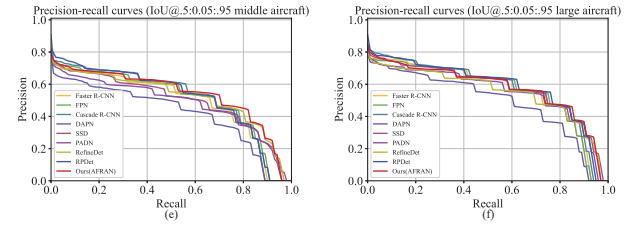


Fig. 15: Precision-recall curves of the CNN-based methods for aircraft detection in SAR images.

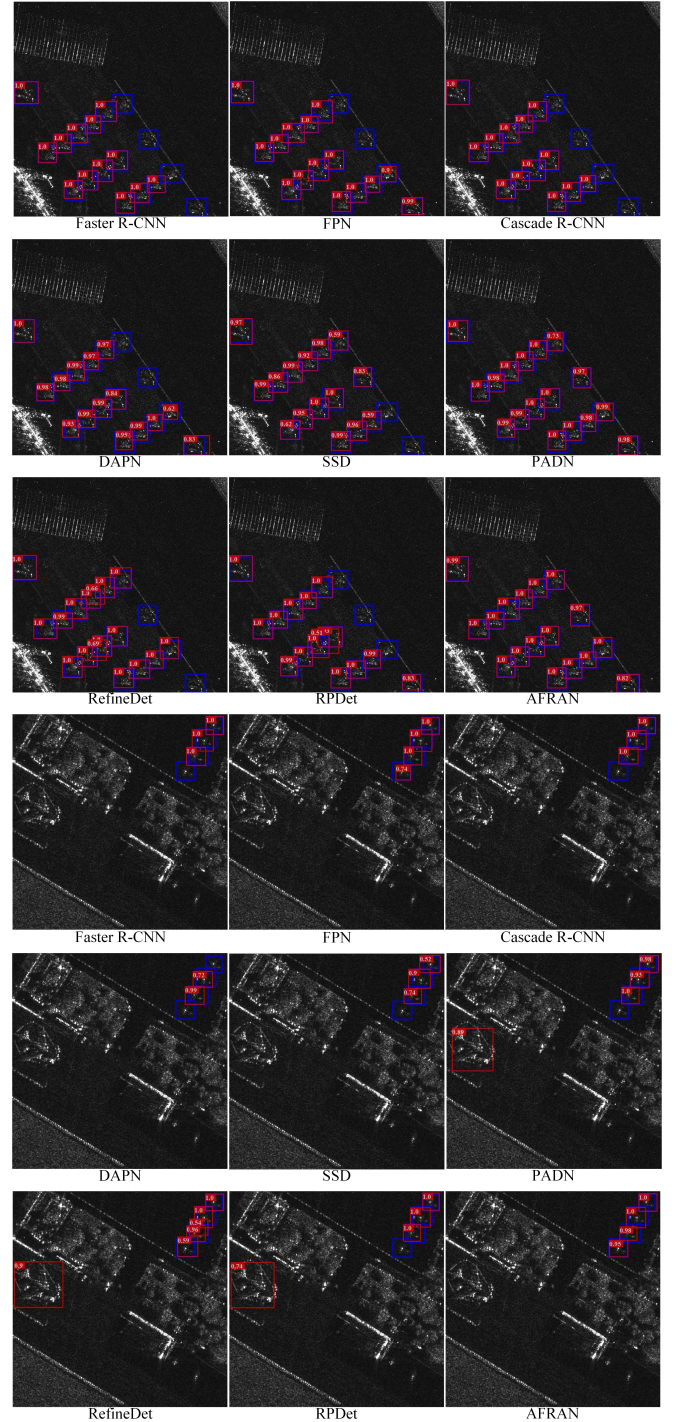


Fig. 16: Continued.

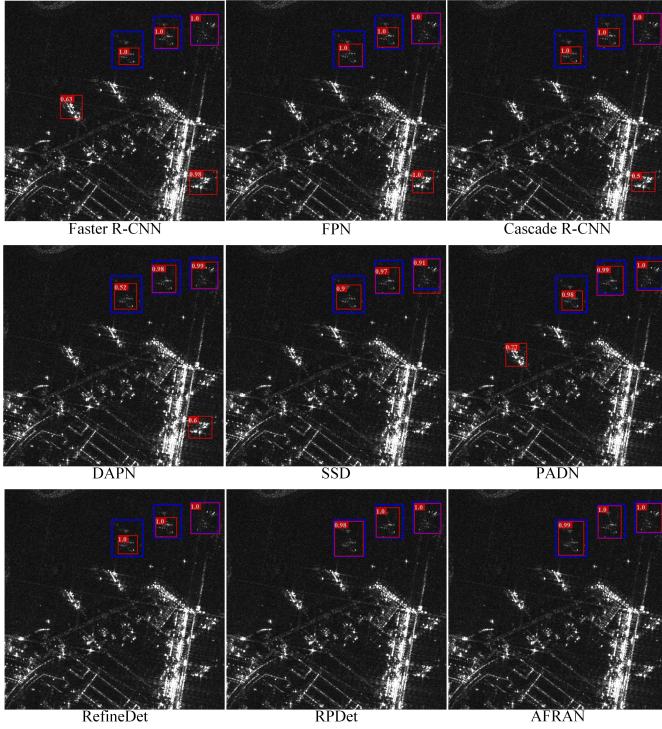


Fig. 16: Aircraft detection results of different CNN-based methods at three different situations.

Additionally, we further inspect the visual detection performance of different CNN-based methods at three representative situations and the results are given by Fig. 16. In the first scene, aircraft with indistinct back-scattering information are densely parked together. In comparison with obvious false alarms detected by RefineDet and RPDet, and false negatives (the aircraft on the right side of the scene) missed by all methods except for PADN, our method detects these aircraft accurately with distinct separations. In the second scene, buildings on the left side of scene are detected as an aircraft by PADN, RefineDet, RPDet. Also, aircraft with poor back-scattering intensities (the first aircraft near the buildings) are missed by most compared methods. However, no false alarms are detected and false negatives are missed by our method. In the third scene, strong interference exists due to complex and angular lounge bridges as well as other ground facilities. In comparison to partial aircraft encompassed by other methods, *e.g.*, Cascade R-CNN, PADN and RefineDet, our method and RPDet all detect and encompass aircraft integrally benefiting from their flexible representation strategies for aircraft's discrete characteristics.

Apart from the above comparisons, the detection performance of different CNN-based methods over a large scene are investigated as shown in Fig. 17. To evaluate detection performance of different methods purely, only sliced cropping is adopted to satisfy input sizes of methods in advance. And the final detection results are acquired by mapping aircraft's coordination from partial slices to the large scene. Obviously, our method could detect all aircraft with fewer false alarms than other compared methods. Most specifically, due to the lack of efficient feature refinement operations after feature

aggregation in FPN, Cascade R-CNN, DAPN and PADN, surroundings within area A (the blue rectangle in Fig. 17) are easily recognized as aircraft. However, our method could discriminate aircraft and interference accurately. In terms of detecting aircraft with highly discrete back-scattering information, *e.g.*, aircraft parked at area B marked by the green rectangle in Fig. 17, aircraft's partial bodies are recognized as new ones by SSD, RefineDet and RPDet. Also, ground facilities within airport are detected as aircraft by FPN, DAPN and PADN, however, discriminated accurately by our method without any false alarms and negatives.

D. More Discussion

Diversified parameter configurations also affect the performance of our method, which are discussed in this subsection.

Feature forward of AFFM. In our method, two feature forward propagation pathways plotted as orange arrows in Fig. 3, are specially designed for enhancing representation ability of the fine-grained middle P_3 and the top P_4 feature maps for aircraft's low-level details. Table V illustrates the detection results of our method configured with different feature forward pathways in AFFM. For a fair reference, the detection results of AFRAN without any feature forward branches in AFFM acting as a baseline, are given by the first row of Table V. Obviously, a 2.2% promotion on AP^s is achieved after combining low-level details of aircraft at Conv4_3 with semantic features existing at the middle layer Conv5_3 base on the baseline. However, a sharp decrease emerges on all indicators when merely introducing middle-level features to the topmost feature layer Conv7 due to the limited supervision of low-level details. When equipped with the two feature forward branches, a competitive performance, *e.g.*, $AP^{.5}$ is 4.1% higher than that of the baseline, is acquired by our method benefiting from abundantly detailed and semantic features represented by the multi-scale fine-grained feature pyramid.

Locations and groups of SA blocks. Due to various and complex semantic information at multi-scale feature maps, how the SA blocks placed and the number of SA blocks are also worth exploration. Table VI shows the contributions of SA blocks located at different feature layers. Most specifically, enabling SA blocks at a single feature layer, *i.e.*, the first four rows of Table VI, could boost the detection performance of our method, *e.g.*, AP and F_1 , resulting from powerful feature refinement donated by their well-designed layer-aware attentions. Especially, an obvious improvement is achieved after enabling SA blocks at the middle feature map when building P_3 . It may be because of fully consideration of aircraft's low-level and high-level features remaining at the three feature maps. Besides, as can be seen from the fifth to the seventh columns of Table VI, most AP metrics increase when enabling SA blocks at two feature layers benefiting from effective utilization of complementary semantic features at different feature layers. Eventually, by applying SA blocks at all feature levels, AP , $AP^{0.75}$, AP^s achieved by our method, are 1.4%, 1.8% and 7.6% higher than those of the baseline, which proves superior characteristics of SA blocks for feature

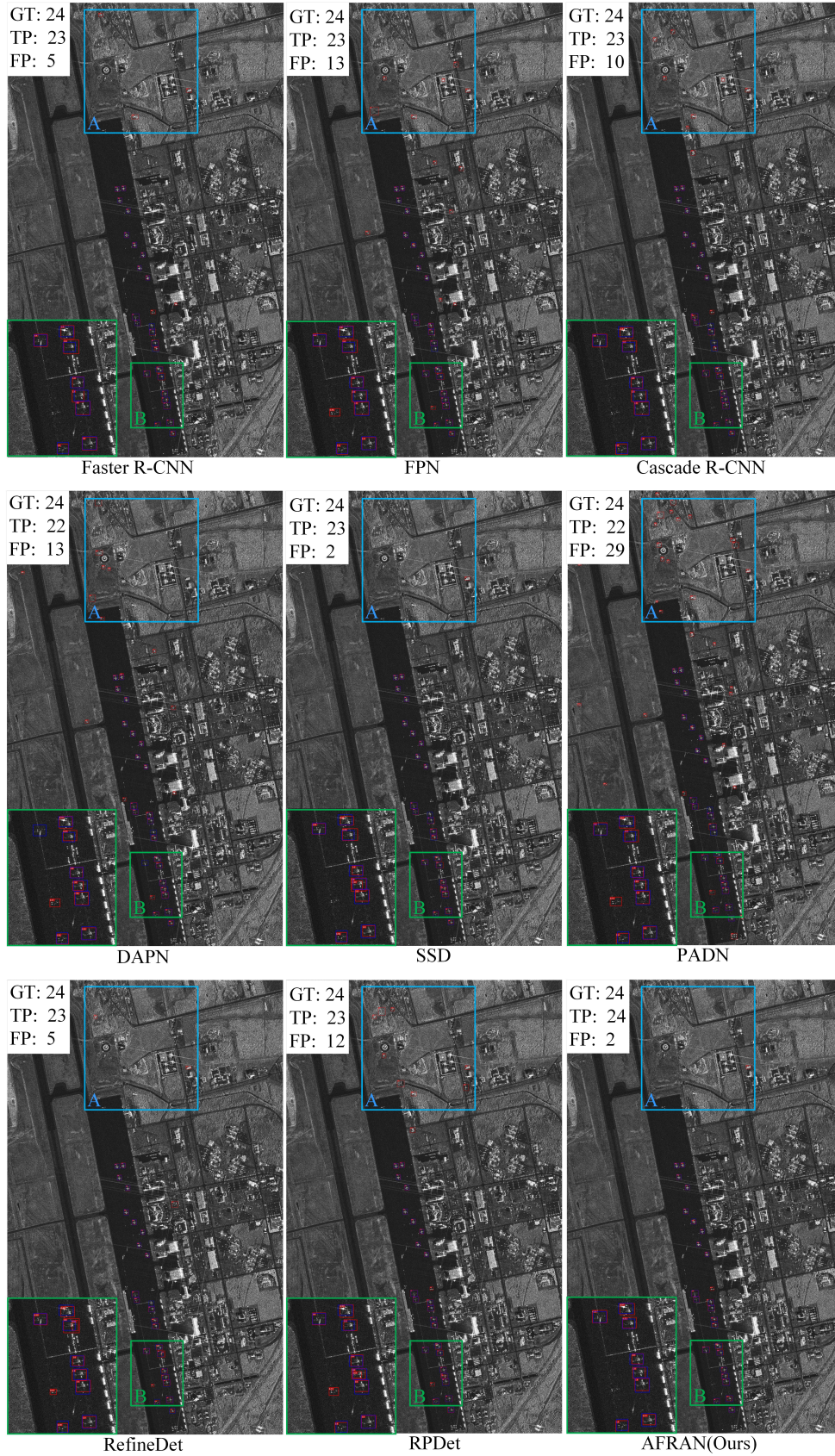


Fig. 17: Aircraft detection results over a large scene by different CNN-based methods.

TABLE V: Feature forward configuration of AFFM

B → M	M → T	P	R	F_1	AP	$AP^{.5}$	$AP^{.75}$	AP^s	AP^m	AP^l
\times	\times	0.900	0.923	0.911	0.548	0.930	0.607	0.440	0.533	0.573
\checkmark	\times	0.902	0.931	0.917	0.547	0.939	0.602	0.462	0.538	0.561
\times	\checkmark	0.963	0.762	0.851	0.499	0.856	0.532	0.380	0.488	0.519
\checkmark	\checkmark	0.904	0.932	0.918	0.554	0.941	0.597	0.481	0.537	0.576

TABLE VI: Locations and group configurations of SA blocks

B	M	T	P	R	F_1	AP	$AP^{.5}$	$AP^{.75}$	AP^s	AP^m	AP^l
-	-	-	0.883	0.932	0.907	0.540	0.939	0.579	0.405	0.520	0.568
\checkmark	-	-	0.887	0.924	0.905	0.544	0.928	0.589	0.441	0.534	0.561
-	\checkmark	-	0.897	0.930	0.913	0.550	0.940	0.590	0.453	0.538	0.571
-	-	\checkmark	0.899	0.929	0.914	0.548	0.939	0.576	0.427	0.532	0.576
\checkmark	\checkmark	-	0.893	0.922	0.907	0.550	0.932	0.602	0.422	0.541	0.568
\checkmark	-	\checkmark	0.896	0.930	0.913	0.542	0.931	0.569	0.431	0.531	0.566
-	\checkmark	\checkmark	0.893	0.922	0.907	0.553	0.934	0.601	0.418	0.539	0.577
$\checkmark(1)$	$\checkmark(1)$	$\checkmark(1)$	0.911	0.925	0.918	0.547	0.924	0.606	0.446	0.536	0.574
$\checkmark(2)$	$\checkmark(3)$	$\checkmark(2)$	0.904	0.932	0.918	0.554	0.941	0.597	0.481	0.537	0.576

TABLE VII: Different configurations of dilation rates of DLCM

B → M → T	P	R	F_1	AP	$AP^{.5}$	$AP^{.75}$	AP^s	AP^m	AP^l
1,1,1	0.904	0.932	0.918	0.554	0.941	0.597	0.481	0.537	0.576
3,2,1	0.894	0.931	0.912	0.551	0.933	0.593	0.454	0.533	0.576
6,4,2	0.892	0.930	0.911	0.550	0.941	0.596	0.427	0.528	0.578

refinement. Additionally, the effects of the number of attention groups are shown at the eighth and ninth rows of Table VI. Due to a unspecific attention mechanism provided by the unified weights when setting attention groups of SA blocks to 1, a slight decrease on AP appears leashing their ability for refining semantic information at specific feature levels.

Receptive fields of DLCM. Capturing discrete features of aircraft accurately calls for a sophisticated design of convolutional receptive fields. Table VII shows the detection results of our method configured with various dilation rates in DLCM. The larger the dilation rates are, the wider the receptive fields captured by DLCM. It could be observed that most of specific and comprehensive metrics, *e.g.*, precision (P), AP^s , AP^m , F_1 , AP , obtained by our method decreases continually with the increase of dilation rates from bottom to top feature layers. It might be because of much uninformative background captured by enlarged receptive fields, which causes to serious interference on feature extraction for small aircraft. Therefore, setting all dilation rates of deformable convolution within DLCM to 1 is a moderate trade-off for extracting discrete characteristics of aircraft.

V. CONCLUSION

In this paper, an Attentional Feature Refinement and Alignment Network (AFRAN) is proposed for aircraft detection in SAR images by carefully taking aircraft's domain knowledge and challenges into consideration. Three significant components including Attention Feature Fusion Module (AFFM), Deformable Lateral Connection Module (DLCM) and Anchor-guide Detection Module (ADM), were introduced for adaptively refining and aligning significant discrete features of aircraft. Ablation studies conducted on a self-built aircraft sliced dataset verified the powerful feature aggregation and refinement abilities of AFFM and effective feature alignment

ability of DLCM and ADM by adopting them at the lateral connections and the detection heads of a three-layer fine-grained feature pyramid, respectively. Extensive experiments conducted on test set of the aircraft sliced dataset and a large scene SAR image demonstrated the topmost detection accuracy as well as competitive speed and spatial complexities achieved by our method in comparison with other domain-specific, *e.g.*, DAPN, PADN, and general CNN-based methods, *e.g.*, Faster R-CNN, FPN, Cascade R-CNN, RefineDet and RPDet. In the future, a further combination between characteristics of aircraft and network design will be investigated to make great promotions for aircraft detection in SAR images.

REFERENCES

- [1] M. C. Dobson, L. E. Pierce, and F. T. Ulaby, "Knowledge-based land-cover classification using ERS-1/JERS-1 SAR composites," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 34, no. 1, pp. 83–99, 1996. 1
- [2] S. Saha, F. Bovolo, and L. Bruzzone, "Building change detection in VHR SAR images via unsupervised deep transcoding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 1917–1929, 2020. 1
- [3] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Collaborative Attention-based Heterogeneous Gated Fusion Network for Land Cover Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 3829–3845, 2020. 1
- [4] T. Zhang, X. Zhang, J. Shi, and S. Wei, "Hyperli-Net: A hyper-light deep learning network for high-accurate and high-speed ship detection from synthetic aperture radar imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 123–153, 2020. 1
- [5] F. Biondi, "Low-rank plus sparse decomposition and localized radon transform for ship-wake detection in SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 1, pp. 117–121, 2017. 1
- [6] J. Cheng, F. Zhang, D. Xiang, Q. Yin, and Y. Zhou, "Polar image classification with multiscale superpixel-based graph convolutional network," *IEEE Transactions on Geoscience and Remote Sensing*, 2021. 1
- [7] G. Dong, H. Liu, G. Kuang, and J. Chanussot, "Target recognition in SAR images via sparse representation in the frequency domain," *Pattern Recognition*, vol. 96, p. 106972, 2019. 1

- [8] G. Gao, L. Liu, L. Zhao, G. Shi, and G. Kuang, "An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 6, pp. 1685–1697, 2008. 1
- [9] Z. Cui, X. Wang, N. Liu, Z. Cao, and J. Yang, "Ship detection in large-scale SAR images via spatial shuffle-group enhance attention," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 379–391, 2020. 1
- [10] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 2738–2756, 2020. 1, 9, 10
- [11] X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler, "Deep learning meets SAR: Concepts, Models, Pitfalls, and Perspectives," *IEEE Geoscience and Remote Sensing Magazine*, pp. 0–0, 2021. 1
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255. 1
- [13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755. 1, 7
- [14] 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 *IEEE International Geoscience and Remote Sensing Symposium*, 2016, pp. 1575–1578. 2, 3
- [15] X. Zhang, B. Xiong, and G. Kuang, "Aircraft segmentation in sar images based on improved active shape model," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-3, pp. 2331–2335, 2018. 2, 3
- [16] H. Hu, L. Huang, and W. Yu, "Aircraft detection for HR SAR images in non-homogeneous background using GGMD-based modeling," *Chinese Journal of Electronics*, vol. 28, no. 6, pp. 1271–1280, 2019. 2, 3
- [17] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikainen, "From BoW to CNN: Two decades of texture representation for texture classification," *International Journal of Computer Vision*, vol. 127, p. 74–109, 2019. 2
- [18] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikainen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020. 2
- [19] S. Wang, X. Gao, H. Sun, X. Zheng, and X. Sun, "An aircraft detection method based on convolutional neural networks in high resolution SAR images," *Journal of Radars*, vol. 6, no. 2, pp. 195–203, 2017. 2, 3
- [20] q. Guo, H. Wang, and F. Xu, "Aircraft target detection from spaceborne synthetic aperture radar image," *Aerospace Shanghai*, vol. 35, no. 6, pp. 57–64, 2018. 2, 3
- [21] W. Diao, F. Dou, K. Fu, and X. Sun, "Aircraft detection in SAR images using saliency based location regression network," in *IEEE International Geoscience and Remote Sensing Symposium*, 2018, pp. 2334–2337. 2, 3
- [22] 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 Sensing*, vol. 10, no. 7, p. 1016, 2018. 2, 3
- [23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017. 2, 3, 5, 10
- [24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37. 2, 3, 10
- [25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944. 2, 3, 10
- [26] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271. 2, 3
- [27] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4203–4212. 2, 4, 10
- [28] T. D. Ross, J. J. Bradley, L. J. Hudson, and M. P. O'connor, "SAR ATR: so what's the problem? an MSTAR perspective," in *Algorithms for Synthetic Aperture Radar Imagery VI*, vol. 3721. International Society for Optics and Photonics, 1999, pp. 662–672. 2
- [29] L. Novak, "State-of-the-art of SAR automatic target recognition," in *IEEE International Radar Conference*. IEEE, 2000, pp. 836–843. 2
- [30] L. M. Novak, G. J. Owirka, W. S. Brower, and A. L. Weaver, "The automatic target-recognition system in SAIP," *Lincoln Laboratory Journal*, vol. 10, no. 2, 1997. 2
- [31] D. E. Dudgeon and R. T. Lacoss, "An overview of automatic target recognition," *Lincoln Laboratory Journal*, 1993. 2
- [32] Q. Zhao and J. C. Principe, "Support Vector Machines for SAR automatic target recognition," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 2, pp. 643–654, 2001. 2
- [33] S. Chen, H. Wang, F. Xu, and Y. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4806–4817, 2016. 2
- [34] C. F. Olson and D. P. Huttenlocher, "Automatic target recognition by matching oriented edge pixels," *IEEE Transactions on Image Processing*, vol. 6, no. 1, pp. 103–113, 1997. 2
- [35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587. 3
- [36] R. Girshick, "Fast R-CNN," in *IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448. 3
- [37] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162. 3, 10
- [38] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 821–830. 3
- [39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788. 3
- [40] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 89–95. 3
- [41] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 9627–9636. 3
- [42] Q. Guo, H. Wang, and F. Xu, "Scattering enhanced attention pyramid network for aircraft detection in SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 9, pp. 7570–7587, 2021. 3
- [43] Y. Zhao, L. Zhao, C. Li, and G. Kuang, "Pyramid attention dilated network for aircraft detection in SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 4, pp. 662–666, 2021. 3
- [44] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2020. 3
- [45] Y. [Zhao, L. Zhao, and G. Kuang, "Attention feature fusion network for rapid aircraft detection in sar images," *ACTA ELECTRONICA SINICA*, vol. 49, no. 9, pp. 1665–1674, 2021. 3
- [46] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768. 3
- [47] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 781–10 790. 3
- [48] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141. 3
- [49] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision*. Springer, 2018, pp. 3–19. 3
- [50] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *British Machine Vision Conference*. British Machine Vision Association, 2018. 3
- [51] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 764–773. 3, 5
- [52] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9308–9316. 3, 5
- [53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. 3

- [54] S. Zhang, L. Wen, Z. Lei, and S. Z. Li, "RefineDet++: Single-shot refinement neural network for object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 674–687, 2020. 4, 5
- [55] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 761–769. 4
- [56] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "ResNeSt: Split-attention networks," *arXiv:2004.08955*, 2020. 4
- [57] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988. 5
- [58] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *IEEE International Conference on Computer Vision*, 2019, pp. 9657–9666. 5, 10
- [59] Y. Chen, C. Han, N. Wang, and Z. Zhang, "Revisiting feature alignment for one-stage object detection," *arXiv:1908.01570*, 2019. 5
- [60] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019. 7
- [61] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8983–8997, 2019. 9, 10