FAIR1M: A Benchmark Dataset for Fine-grained Object Recognition in High-Resolution Remote Sensing Imagery

Xian Sun[a,b,c,*], Peijin Wang[a,b,c], Zhiyuan Yan[a,b,c], Cheng Wang[d,e], Wenhui Diao[a,b,c], Jin Chen[f], Jihao Li[a,b,c], Yingchao Feng[a,b,c], Tao Xu[a,b,c], Martin Weinmann[g], Stefan Hinz[g] and Kun Fu[a,b,c,*]

[a]*Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China*

[b]*School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China*

[c]*Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China*

[d]*Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China*

[e]*Fujian Collaborative Innovation Center for Big Data Applications in Governments, Fuzhou 350003, China*

[f]*Beijing Remote Sensing Information Institute, Beijing 100011, China*

[g]*Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany*

ARTICLE INFO

Keywords:
Remote sensing images
Fine-grained object detection and recognition
Deep learning
Benchmark dataset
Convolutional Neural Network (CNN)

ABSTRACT

With the rapid development of deep learning, many deep learning based approaches have made great achievements in object detection task. It is generally known that deep learning is a data-driven method. Data directly impact the performance of object detectors to some extent. Although existing datasets have included common objects in remote sensing images, they still have some limitations in terms of scale, categories, and images. Therefore, there is a strong requirement for establishing a large-scale benchmark on object detection in high-resolution remote sensing images. In this paper, we propose a novel benchmark dataset with more than 1 million instances and more than 15,000 images for Fine-grAined object recognItion in high-Resolution remote sensing imagery which is named as **FAIR1M**. We collected remote sensing images with a resolution of 0.3m to 0.8m from different platforms, which are spread across many countries and regions. All objects in the FAIR1M dataset are annotated with respect to 5 categories and 37 sub-categories by oriented bounding boxes. Compared with existing detection datasets dedicated to object detection, the FAIR1M dataset has 4 particular characteristics: (1) it is much larger than other existing object detection datasets both in terms of the quantity of instances and the quantity of images, (2) it provides more rich fine-grained category information for objects in remote sensing images, (3) it contains geographic information such as latitude, longitude and resolution, (4) it provides better image quality owing to a careful data cleaning procedure. To establish a baseline for fine-grained object recognition, we propose a novel evaluation method and benchmark fine-grained object detection tasks and a visual classification task using several State-Of-The-Art (SOTA) deep learning based models on our FAIR1M dataset. Experimental results strongly indicate that the FAIR1M dataset is closer to practical application and it is considerably more challenging than existing datasets. Researchers can better investigate fine-grained object detection algorithms with the help of the FAIR1M dataset.

1. Introduction

Object detection and recognition aims to obtain the localization and categories of objects of pre-defined categories in an image. It is one of the most fundamental and important tasks in the field of earth observation, which serves various civil applications, such as geographic information system mapping, agriculture, traffic planning, and navigation [49, 48, 17, 6, 5, 10]. Due to the wide spatial coverage of remote sensing images, there are typically a large number of objects in a remote sensing image. It is a challenging task for machines to recognize and detect objects accurately in such images, but the development of deep learning-based approaches provides effective solutions characterized by a strong ability of feature extraction and feature expression.

However, deep learning is a data-driven concept in the field of computer vision, and the performance of respective deep learning-based approaches strongly depends on the quality and quantity of given data. A challenging and excellent dataset can accelerate the development of the field. For example, the ImageNet [9] and MSCOCO [30] datasets hasten the evolution of Convolutional Neural Networks (CNNs) on natural scene image classification and object detection tasks, UC Merced [60] and MSTAR[11] datasets separately promote the progress of optical remote sensing scene classification and Synthetic Aperture Radar (SAR) target recognition, the Cityscape [7] and Vaihingen [43] datasets facilitate the development of deep neural networks for semantic segmentation in natural scenes and remote sensing scenes respectively, the DOTA [53] and DIOR [27] datasets are proposed for generic object detection in remote sensing images, and the FGSD [3] and VEDAI [37] datasets inspire the research of fine-grained object detection in re-

*Corresponding author

sunxian@aircas.ac.cn (X. Sun); wangpj@aircas.ac.cn (P. Wang); yanzy@aircas.ac.cn (Z. Yan); cwang@xmu.edu.cn (C. Wang); diaowh@aircas.ac.cn (W. Diao); chenjin_wonder@hotmail.com (J. Chen); lijihao17@mails.ucas.edu.cn (J. Li); fengyingchao17@mails.ucas.edu.cn (Y. Feng); xutao17@mails.ucas.edu.cn (T. Xu); martin.weinmann@kit.edu (M. Weinmann); stefan.hinz@kit.edu (S. Hinz); fukun@mail.ie.ac.cn (K. Fu)

ORCID(s): 0000-0002-0038-9816 (X. Sun)

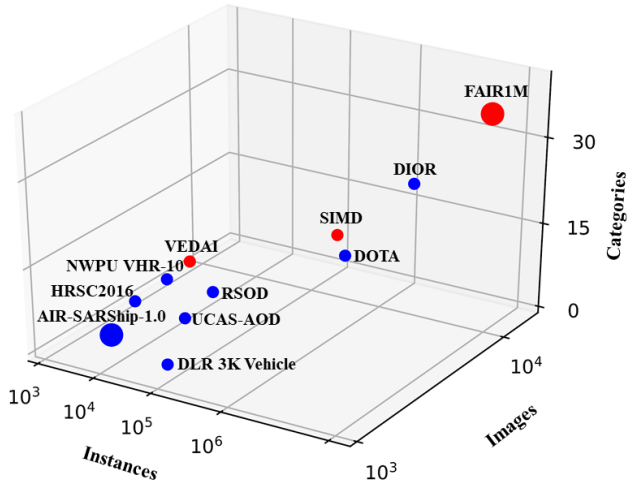


Figure 1: Multi-dimensional representation for typical object detection datasets in the field of remote sensing. Red and blue points denote the fine-grained datasets and generic datasets, respectively. The larger points denote the datasets which additionally contain geographic information.

mote sensing images. Among these tasks, object detection and recognition has attracted wide attention in the past few years [18, 16, 58, 51, 62, 4, 59]. However, compared with datasets focusing on natural scenes, some deficiencies of existing datasets limit the development of fine-grained object recognition in the field of remote sensing.

1. The scale of datasets can be further expanded. With the enhancement of the demand for remote sensing applications, object detectors need to own stronger generalization ability. Due to the over-fitting phenomenon, some seemingly excellent algorithms which perform well on small datasets are likely to obtain bad results in a larger dataset. Hence, in order to evaluate an algorithm more comprehensively, the scale of corresponding datasets is required to be relatively large in terms of object instances and image quantity. Currently, noteworthy achievements have been made in natural scene object detection datasets, such as MSCOCO [30]. Therefore, the scale of remote sensing scene object detection datasets still needs to be expanded eagerly.

2. Fine-grained information needs to enrich. Objects in remote sensing images usually have multiple fine-grained types. At present, generic object detection and recognition are difficult to meet the needs of applications and the demand for fine-grained recognition is rapidly growing. For example, an excellent algorithm must be able to not only correctly detect instances belonging to the category *Airplane*, but also recognize that the object belongs to a certain subcategory such as *Airbus 350*, *Boeing 747* or other type. As shown in Figure 1, most well-known existing object detection datasets contain coarse-grained annotation information, or a small amount of fine-grained information. For instance, DOTA [53] divides vehicles into *Large-vehicles* and *Small-vehicles*. Deep learning models trained on these datasets may not perform very well when they are faced with large-scale type recognition tasks.

3. Image quality needs to be improved. Taking into account the process of high-resolution satellite image acquisition, there may be some interference factors in the images, such as clouds and fog. No cleaning or improper cleaning will directly reduce the quality of images, and then influence the performance of object detection algorithms.

4. The dataset should contain more geographic characteristics. Temporal and spatial information are the two major geographic characteristics in the field of remote sensing. Remote sensing images collected in most of the existing object detection datasets are single-temporal, which means that there is no time dimension difference in the same remote sensing scene. It is relatively difficult for these datasets to represent the change of seasons and surroundings. This also has an effect on the generalization ability of deep learning models to some extent. As shown in Figure 1, the storage format of images in existing datasets is the same as for natural scene images and tends to lack geographic information. Geographic information in turn is vital for remote sensing image processing, referring to properties such as spatial resolution, longitude, and latitude.

Consequently, in order to better address the problems mentioned above, we propose a novel benchmark dataset for Fine-grAined object recognItion in high-Resolution remote sensing imagery which is named as **FAIR1M**. Some representative examples of images and their annotations are shown in Figure 2. In the FAIR1M dataset, we collect remote sensing images containing more than 15000 images and 1 million instances from the GaoFen satellites and Google Earth platform. Scenes in these remote sensing images are spread across many continents, such as Asia, America, and Europe. All FAIR1M images are annotated with oriented bounding boxes (OBB) and with respect to 5 categories and 37 sub-categories under the guidance of many experts in remote sensing. To the best of our knowledge, FAIR1M is the largest fine-grained oriented object detection dataset suitable for remote sensing scenes.

Due to the high quantity and quality of images and fine-grained categories, FAIR1M promotes the challenging tasks for fine-grained object detection and visual classification, which aims to obtain fine-grained categories and locations of objects. Compared with generic object detection, fine-grained object detection can recognize not only generic categories but also fine-grained types for objects in remote sensing images. In order to standardize the development of fine-grained detection on the FAIR1M dataset, we also propose a new evaluation metric and build a benchmark using representative algorithms.

In summary, the proposed FAIR1M benchmark dataset intends to provide a large-scale fine-grained object detection dataset to the remote sensing community. We hope that with the help of FAIR1M, a growing number of novel algorithms can be investigated in the field of remote sensing image interpretation. Our main contributions of this work are briefly summarized as follows:

- A large-scale public dataset has been proposed for object recognition in remote sensing images. To the best

Figure 2: Visualization of annotations in the FAIR1M dataset. In addition to fine-grained object categories, the FAIR1M dataset contains crowded scenes, complex background, and various sizes and angles of objects. Yellow boxes are the bounding boxes we annotated.

of our knowledge, the proposed FAIR1M benchmark dataset is the largest fine-grained object recognition dataset in remote sensing with 1 million instances. Moreover, multi-temporal images, geographic information and orientated annotations are provided in the FAIR1M dataset.

- To evaluate the performance of detection methods on the FAIR1M dataset, we design a novel evaluation metric for object detection in remote sensing images. Compared with generic object detection, fine-grained object detection pays more attention to the categories of objects. As a result, a score-aware and challenging evaluation metric is proposed and validated on the FAIR1M dataset.

- We propose a novel cascaded hierarchical object detection network and benchmark fine-grained object detection tasks and a visual classification task using several state-of-the-art object detection models on the FAIR1M dataset, which can be utilized as the baseline for future work. We believe that this benchmark will certainly promote the development of object detection in the field of remote sensing.

The rest of this paper is organized as follows. In Section 2, we review several existing popular object detection datasets in remote sensing. Section 3 describes the proposed new FAIR1M dataset in detail. In Section 4, we introduce the evaluation metrics and evaluate some excellent object detectors on the proposed FAIR1M dataset. Ultimately, Section 5 summarizes the paper and provides an outlook on the future of object detection in remote sensing imagery.

2. Related Work

2.1. Datasets for Object Recognition in Ground-level Natural Images

Object detection is one of the most important research directions in the fields of computer vision and remote sensing. Therefore, a series of datasets were proposed to accelerate the development of object detection. The PASCAL VOC [13] is one of the fundamental datasets, which is widely used in the ground-level natural object detection. It was proposed in 2005 and has been gradually extended until 2012. Finally, the Pascal VOC 2012 dataset became a large-scale dataset, which contains 20 object categories in 11.5k images resulting in 27k bounding boxes. However, the Pascal VOC 2012 dataset can not adequately represent the real world as most of the scenes are represented in iconic view. The MSCOCO [30] dataset proposed in 2014 is much larger than the PASCAL VOC dataset, and contains more object categories and object instances, especially regarding objects of smaller size. Specifically, the dataset contains 80 object categories and 896k object instances in more than 200k images.

Table 1
Comparison between FAIR1M dataset and other object detection datasets containing remote sensing images

Datasets	Source	Instances	Images	Image width	Categories	Annotation	Image format	Fine-grained
NWPU VHR-10 [6]	Google Earth	3,775	800	~1000	10	HBB	JPG	N
VEDAI [37]	Google Earth	3,640	1,210	512, 1024	9	OBB	PNG	Y
UCAS-AOD [63]	Google Earth	6,029	910	~1000	2	OBB	PNG	N
DLR 3K Vehicle [31]	Aerial images	14,235	20	5616	2	OBB	JPG	N
HRSC2016 [32]	Google Earth	2,976	1,070	~1100	1	OBB	BMP	N
AIR-SARShip-1.0 [54]	Gaofen-3	3,000	31	3000	1	HBB	TIFF	N
RSOD [55]	Google Earth, Tianditu	6,950	976	~1000	4	HBB	JPG	N
DOTA [53]	Google Earth, Satellite JL-1, GF-2	188,282	2,806	800-4000	15	OBB	PNG	N
xView [26]	WorldView-3	~1.0M	1,127	2000-4000	60	HBB	PNG	Y
DIOR [27]	Google Earth	192,472	23,463	800	20	HBB	JPG	N
SIMD [19]	Google Earth	45,096	5,000	1024	15	HBB	JPG	Y
FGSD [3]	Google Earth	5,634	2,612	930	43	OBB	JPG	Y
FAIR1M	Gaofen, Google Earth	**~1.0M**	15000	1000-10000	37	OBB	TIFF	Y

More recently, Megvii released Objects365 [44], a large-scale dataset containing 638k images, which is 5 times larger than MSCOCO, and which consists of 365 object classes and more than 10 million object instances. Compared with the prior datasets, the number of instances per image reaches 15.8 (2.4 for PASCAL VOC and 7.3 for MSCOCO). Furthermore, Google proposed the Open Images [25] dataset, which is now updated to version V6. The dataset contains a total of 16 million bounding boxes for 600 object classes on 1.9 million images, making it the largest existing dataset dedicated to object detection. The above large-scale datasets promote the research on general object detection in natural images [23, 21, 45, 28, 58, 40, 38]. However, considering the differences between the ground-level natural images and geo-spatial remote sensing images, the research on geo-spatial object detection requires specially designed datasets.

2.2. Datasets for Object Recognition in Remote Sensing Images

Due to the top view of the remote sensing images and the various spatial resolutions of sensors, the scale variations of object instances are huge and objects often appear in arbitrary orientations. In order to promote the development of the aerial object detection research, many remote sensing datasets have been proposed, such as NWPU VHR-10, HRRSD, DOTA and DIOR. The NWPU VHR-10 dataset [6] is composed of 10 geospatial object categories in 715 images, and the images were collected from Google Earth. However, the total number of object instances in NWPU VHR-10 is only 3775, which cannot adequately reflect the complexity of the problem in the real world. The HRRSD dataset [62] contains more than 55k object instances for 13 object categories in around 21k images, the images in the HRRSD dataset have been acquired from Google Earth and Baidu Map and their spatial resolution varies from 0.15m to 1.2m. However, the size of the images in the HRRSD dataset is too small, with only 227×227 pixels, which limits the range of the research. More recently, a large-scale dataset was proposed, named DIOR [27], which contains more than 23k images and 192k instances, covering 20 object categories. However, the objects in the DIOR dataset are annotated with horizontal bounding box (HBB), which cannot better enclose the objects and differentiate closeby objects from each other. The DOTA dataset [53] contains 15 different geospatial object categories and more than 188k oriented bounding boxes. The dataset consists of 2806 images collected from different platforms with multiple spatial resolutions and the sizes of images range from 800×800 to 4000×4000 pixels, which is widely used in the aerial object detection research.

There are some other remote sensing datasets dedicated to the research on important object categories. The RSOD dataset [55] was proposed in 2015, which contains four categories, including overpasses, oil tanks, airplanes and playgrounds. It consists of 976 images obtained from Google Earth and Tianditu with spatial resolutions ranging from 0.3m to 3m. However, the number of object instances in RSOD is only close to 7k. The UCAS-AOD dataset [63]

Figure 3: Data samples of each category in the FAIR1M dataset.

is designed for airplane and vehicle detection, which consists of an airplane dataset and a vehicle dataset. The former contains 600 images and 3210 airplanes and the latter contains 310 images and 2819 vehicles. The COWC dataset [35] only has one object category, which is designed for car detection. It contains around 32.7k instances and 58.2k unique negative examples. The LEVIR dataset [64] contains three object classes, including airplane, ship and oilpot. It consists of 21.9k images and 11k object bounding boxes. The size of the images is 600×800 pixels and the images have been collected from Google Earth and their spatial resolutions vary from 0.2m to 1m. There are also datasets dedicated to building detection, such as the Semcity toulouse dataset [41], ISPRS benchmark on urban object detection and 3D building reconstruction dataset [42], Inria Aerial Image Labeling dataset [34] and DeepGlobe 2018 dataset [8]. However, these datasets ignored the fine-grained category information for such important geo-spatial objects.

2.3. Datasets for Fine-grained Objects Detection and Recognition

Objects in remote sensing images require more fine-grained analysis. Therefore, there are some datasets for fine-grained object detection and recognition. The VEDAI dataset [37] is a fine-grained vehicle detection dataset, which allows the development and benchmarking of vehicle detection algorithms in aerial images. There are in total 1210 images and 3700 instances in VEDAI and the size of the images is 1024×1024 pixels. However, the VEDAI dataset does not include scenes with a large number of vehicles, such as large parking lots. The MTARSI dataset [52] is designed for the aircraft type recognition, which consists of 20 aircraft types. A total of 9385 remote sensing images were obtained from Google Earth satellite images, with spatial resolution ranging from 0.3m to 1.0m. However, the size of the images in

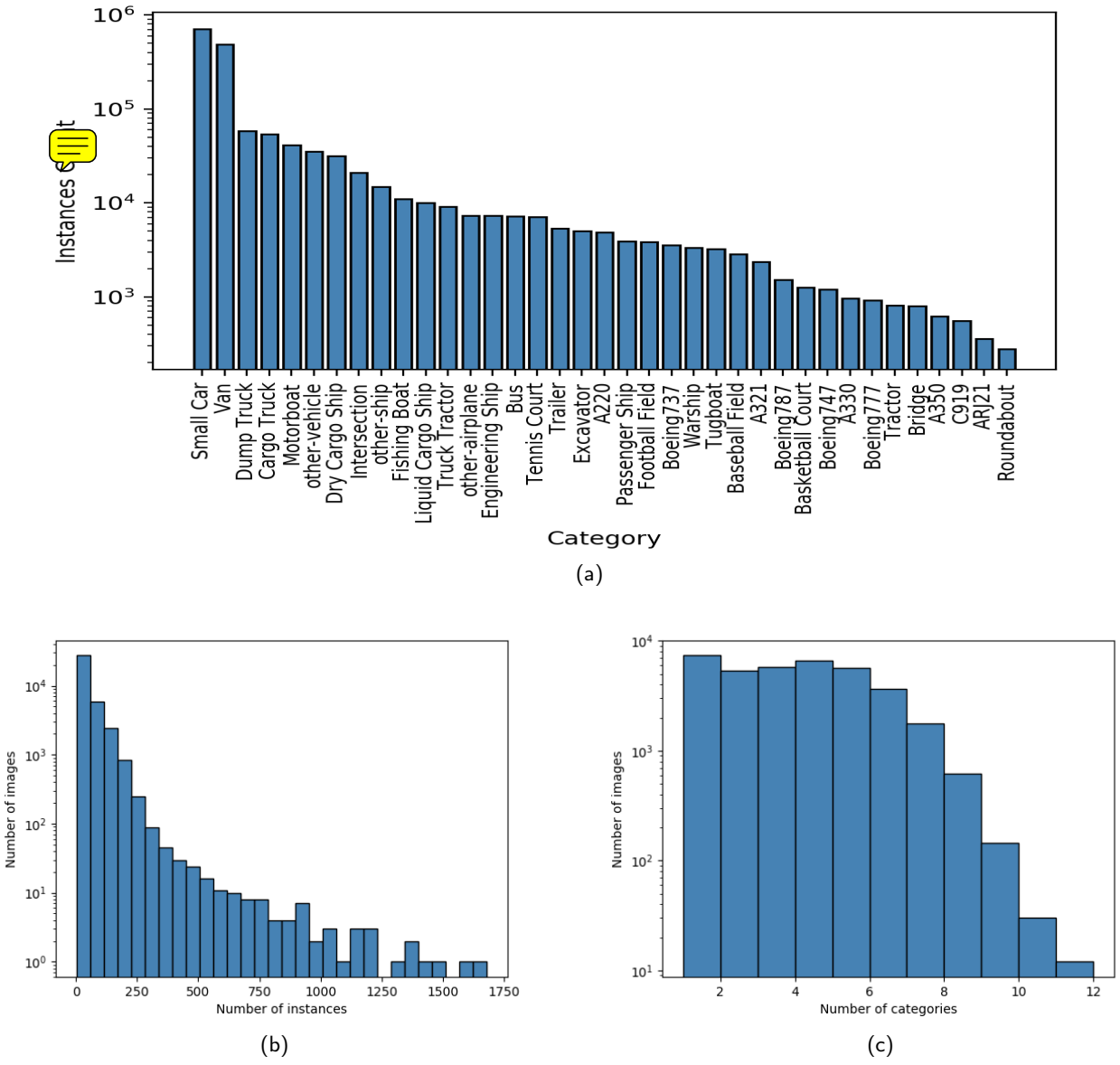


Figure 4: (a) The distribution of the number of instances per category. (b) The distribution of the number of instances per image. (c) The distribution of the number of categories per image.

MTARSI is only 256×256. The MTARSI dataset cannot be used for the research on object detection. The HRSC2016 dataset [33] is the most used dataset for ship detection in aerial remote sensing images, which contains 22 classes of ships and 2976 samples in 1061 images. The image resolutions are between 0.4m and 2m and the image sizes range from 300×300 to 1500×900 pixels. However, the images in HRSC2016 only cover a very limited number of ports and the samples of ships are limited. The FGSD dataset [3] is the latest ship detection dataset, which consists of high-resolution satellite images from 17 large ports in four countries and 43 classes of ships were labeled. It contains 5634 samples in 2612 images. However, the number of images and samples still limits its application.

Above datasets are designed for fine-grained single-class object detection and recognition. There are two datasets for fine-grained multi-class object detection and recognition, which are the xView dataset [26] and SIMD dataset [19]. The xView dataset contains over 1 million horizontal bounding boxes in 1127 images, labeled with respect to 60 object categories, which were organized in a class hierarchy. There are mainly seven categories, including fixed-wing aircraft, passenger cars, trucks, railway vehicles, engineering vehicles, maritime ships, and buildings. However, the image quality of the xView dataset is not very high and some significant categories are coarse, for example, airplanes only contain two types (i.e., Small Aircraft and Cargo Plane). Besides, the distribution of object instances in the xView

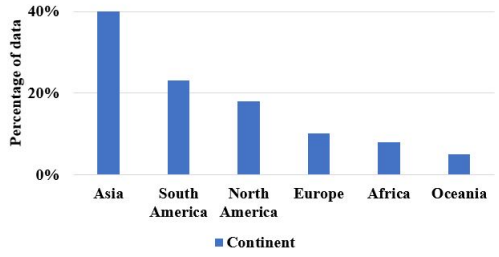


Figure 5: The distribution of the FAIR1M dataset across continents. It can be seen from the figure that the FAIR1M dataset is widely distributed all over the world.

dataset is relatively unbalanced, most of which focus on the building class and the small car class. The SIMD dataset contains 15 different object categories, including seven types of vehicles, six types of aircrafts, Boat and Others class. The images were acquired from Google Earth. Specifically, the dataset comprises 5000 images of resolution 1024×768 pixels and collectively contains around 45k objects. However, the SIMD dataset also applies the horizontal bounding box definition for object annotation and the distribution of object instances in SIMD is also unbalanced. Half of the instances belong to the Car class, there are only few instances in the other categories. The FGSD dataset consists of 43 fine-grained categories, but it only contains ships. Our dataset applies the oriented bounding box definition for annotation and focuses on a careful selection of fine-grained categories for significant objects in remote sensing images, including airplanes, ships, vehicles, courts and roads. It is a larger and more comprehensive dataset for object detection.

3. Details of the FAIR1M Dataset

3.1. Image collection and Pre-processing

Xia et al. [53] prove that a variety of sensors and resolutions can be used to eliminate biases. To meet the needs of practical applications, images in our dataset are collected from the Gaofen satellites and Google Earth, with a spatial resolution ranging from 0.3m to 0.8m. The diversity of data is of great significance to the study of transportation and humanities in different countries and regions. Therefore, we collect 15000 images from more than 60 civil airports and harbors all over the world. The distribution of the FAIR1M dataset with respect to different continents can be seen in Figure 5.

To obtain high-resolution remote sensing images, we adopt a series of pre-processing methods for Gaofen satellite images. Considering the poor quality of satellite images in some datasets, we first check the quality of the raw data and remove the images with many clouds, noise, and bright spots. In order to ensure that the images in the same region have the same positioning accuracy, we perform block adjustment for multi-temporal and multi-source images. Based on the results of block adjustment, we perform a rational function to generate the orthographic results of panchromatic and multi-spectral images. Then, we use the Pan Sharpening algorithms [36] to improve the spatial resolution of multispectral images by fusing panchromatic images. Finally, we use histogram equalization to adjust the hue component of the images. In this way, we can obtain high-resolution and high-quality remote sensing images.

3.2. Category Design

Most existing datasets pay more attention to static objects, such as bridges, baseball fields, storage tanks, basketball courts and so on. These datasets lack fine-grained information about objects, which plays an important role in real applications in the field of remote sensing. In the FAIR1M dataset, the objects we selected include 5 categories: airplanes, ships, vehicles, courts and roads. The selection of fine-grained types of each category in the FAIR1M dataset depends on practical application scenarios and the shape it presents. For airplanes, we set 10 fine-grained categories covering 34 airports around the world. The types of airplane contain Boeing 737, Boeing 777, Boeing 747, Boeing 787, Airbus A320, Airbus A220, Airbus A330, Airbus A350, COMAC C919, and COMAC ARJ21, which are the most common categories in the civil aviation. The categories of ships are defined according to their functions. There are 8 specific categories for ships, including passenger ship, motorboat, fishing boat, tugboat, engineering ship, liquid cargo ship, dry cargo ship, and warship. As well as for ships, the categories of vehicles are defined according to their functions. There are 9 specific categories for vehicles, including small car, bus, cargo truck, dump truck, van, trailer, tractor, truck tractor, and excavator. Furthermore, we selected 4 categories and 3 categories for courts and roads, respectively. As a result, there are 37 fine-grained categories in the FAIR1M dataset for object detection. Besides these specific categories, we also assigned the categories 'other-airplane', 'other-ship', and 'other-vehicle' for objects that do not belong to the previously defined specific object types. All categories and the number of instances per category can be seen in Figure 3 and Figure 4. It is well-known that the number of each category depends on its actual distribution in remote sensing scenarios. The distribution of instances can reflect the authenticity and challenge of the proposed dataset.

3.3. Image Annotation

3.3.1. Annotation Format

With the development of deep learning, most of the current detection methods in the field of remote sensing are transferred from natural scenes. As a result, most of existing object detection datasets in remote sensing are annotated with horizontal bounding boxes, such as UCAS-AOD [63], NWPU VHR-10 [6], DIOR [27], and SIMD [19]. Unlike the objects in natural scenes, which are usually in vertical directions, objects in remote sensing images have a variety of directions. Therefore, using horizontal bounding boxes cannot provide accurate spatial information for oriented objects. To wrap objects more accurately and develop more suitable algorithms for oriented objects in remote sensing

Figure 6: Visualization of annotations of FAIR1M dataset. (a), (b) and (c) show the oriented bounding box annotations of airplanes, ships, and vehicles, respectively. The red points represent the top-left points of the instances. (d) shows the traditional horizontal annotation in other datasets.

images, all instances in the FAIR1M dataset are annotated with oriented bounding boxes (OBB). Samples of annotated instances in the FAIR1M dataset can be seen in Figure 6.

Arbitrary rectangular bounding boxes $\{(x_i,y_i)\}$, i = 1, 2, 3, 4} are adopted for annotation, where (x_i,y_i) denotes the coordinates of the i-th vertices of the rectangular bounding boxes. As well as given for other oriented detection datasets, we arrange the vertices of a object in a clockwise direction. Further, the top-left point is highlighted as the first point (x_1,y_1), which also denotes the positive direction of the object. For categories that are difficult to recognize, we label them as 'other' category. For instances that are occluded or difficult to identify the positive direction, we add the corresponding label information.

Considering the down-sampling characteristics of deep neural networks, we set 16 pixels to be the thresholds for annotations. For example, ships which are longer than 13 meters need not be annotated in the images with spatial resolution of 0.8 meters.

3.3.2. Annotation Quality

To ensure the quality of annotations, we develop a complete and strict quality control process. After getting the initial annotations, we design three stages to check and correct the annotations. At the first stage, we divide annotators in pairs and they check the annotations of each other by annotating them again. Then, each annotator needs to merge the two annotations to obtain a more accurate result. Next, we invite a supervisor to further check the quality of annotations, including location, category, and orientation of bounding boxes. Finally, experts in the field of remote sensing imagery are invited to check the quality of the annotations and the dataset.

3.4. Characteristics of the Dataset

In the past few years, many object detection datasets have been proposed in the field of remote sensing. However, most of them have some common disadvantages, for example, the number of instances and categories are relatively small, and the variations and diversity within the dataset are relatively limited, and the selection of categories does not satisfy the practical application. These shortcomings affect the development of object detection for remote sensing imagery to some extent. As far as we know, we provide the most diverse and challenging object dataset in comparison to other datasets in the field of remote sensing. There are 37 categories and more than a million of instances of objects in the FAIR1M dataset. We perform a comprehensive analysis of the FAIR1M dataset with other available object detection datasets, which is shown in Table 1. In addition to these analysis, the diversity and challenges of the FAIR1M dataset can be reflected in the following characteristics.

1. ***Comprehensive fine-grained types.*** To improve the development of fine-grained object detection and recognition, experts in remote sensing imagery interpretation are invited to design logical fine-grained types for the proposed dataset. The types we finally select are the most common categories in the practical applications. Taking the airplane as an example, researchers and practical applications mainly focus on the types of airplanes, whereas existing airplane datasets only consider coarse categories of airplane.
2. ***Large range of sizes and orientations.*** Due to the imaging principles and spatial resolutions of used sensors, it is common that multi-scale objects widely exist in remote sensing images. To extend the range of size variations of instances, we collect images with multiple spatial resolutions. In comparison with between-class size variation in other datasets, there are size variations and orientation variations not only between multiple classes but also within fine-grained types. As shown in Figure 4, the size of instances in the FAIR1M dataset varies widely.
3. ***High within-class variation and between-class similarity.*** In addition to size variations and angle variations, high within-class variation and between-class similarity is one of the important characteristics of the FAIR1M dataset. For each group of objects, we select common fine-grained types to be different categories. However, the shapes and appearance of different fine-grained types are similar, which results in

(a) (b)

Figure 7: Two possible cases of prediction. (a) and (b) have the same IoU, but the predicted bounding box in (B) contains less confusing information. Green: prediction. Red: ground truth.

high between-class similarity. To obtain high within-class variation, we collect the same scenes with different seasons and weather conditions. Therefore, objects of the same category have different poses, background, colors and light. High within-class variation and between-class similarity make the FAIR1M dataset become more diverse and challenging.

4. ***Complex scenes with densely distributed objects.*** It is common that objects might be densely distributed in remote sensing images. To develop algorithms for densely distributed object detection in remote sensing images, we collect many images from complex scenes with densely distributed objects, such as large ports, parking lots, and crossroads, and we annotate the instances in the dense scenes one by one. Object detection in these scenes is a relatively enormous challenge for many of the current object detection algorithms.
5. ***Rich geographic information.*** The images in existing datasets in the field of remote sensing only contain the size information. Compared with these datasets, the FAIR1M dataset contains rich geographic information, such as spatial resolution, longitude, and latitude. Geographic information can provide researchers with more remote sensing information. In addition to the spatial information of the image, image scenes in the FAIR1M dataset are multi-temporal, which can be used to analyze the changes in the temporal dimension.

3.5. Dataset Splits

We divide the proposed dataset into training set, validation set and testing set. The proportions of the three subsets are 1/2, 1/6 and 1/3, respectively. We will provide all the images and ground truth for the training set and the validation set. For the testing set, we only provide images, but people can obtain detection results via the evaluation server.

4. Algorithm Analysis

We consider two detection tasks and a recognition task on the FAIR1M dataset. The detection tasks are fine-grained oriented object detection and horizontal object detection in remote sensing images. In these object detection tasks, we need to detect objects belonging to the 37 categories in each image. Based on this task, we propose a cascaded hierarchical object detection method to improve the accuracy of fine-grained object detection. The other task is fine-grained visual categorization in remote sensing images. In this task, we set three fine-grained classification subtasks based on the FAIR1M dataset.

Data: a dataset of C categories with ground truth boxes G^c and predicted bounding boxes D^c
Result: mAP_F
for *pairs of* $(G_i^c,\ D_j^c)$ **do**
  calculate $FIoU$ to determine TP, FP, FN;
end
for C *categories* **do**
  for *scores* **do**
    $Precision_F = \sum_i^{Num_G} \frac{FIoU \cdot TP_i \cdot score_{TP}}{TP_i \cdot score_{TP} + FP_i \cdot score_{FP}}$
    $Recall_F = Recall = \frac{FIoU \cdot TP}{TP + FN}$
  end
end
calculate mAP_F by obtained $Precision_F$ and $Recall_F$;

Algorithm 1: The calculation process of mAP_F

4.1. Detection

Object detection is a fundamental task in the field of remote sensing imagery. However, fine-grained categories, various sizes and orientations, multiple scenes severely impact the detection performance on the FAIR1M dataset. We have benchmarked several state-of-the-art object detection methods and our novel hierarchical object detection method.

4.1.1. Evaluation Metrics

The evaluation will be based on the quantitative comparison between predictions and ground-truth for the dataset. In this paper, we design Fine-grained Intersection-Over-Union (FIoU) and Fine-grained mean Average Precision (mAP_F) for the FAIR1M dataset. These metrics are detailed in following sections.

Fine-grained Intersection-Over-Union (FIoU). For the generic object detection task, each detected bounding box is labeled by the Intersection-Over-Union (IoU) metric with ground-truth for a specific category. The ratio of the area of the intersection to the union is the IoU metric of each pair of detected bounding box (D) and ground-truth (G). If the IoU is higher than the threshold of 0.5, the matched detection bounding box will be labeled as TP. Otherwise, it will be labeled as FP. If a ground-truth bounding box has no TP detection to match with, it will be labeled as FN. The IoU metric can represent the overlap of two bounding boxes, but it ignores the impact of confusing information inside or outside the detection boxes. As shown in Figure 3.4, the ground truth box and predicted box have the same IoU in

Table 2
The benchmark results of FAIR1M dataset on the task of oriented bounding boxes detection.

Coarse Category	Category	RetinaNet [29]		Faster RCNN [39]		ROI Transformer [12]		Cascade RCNN [2]		Gliding Vertex [56]	
		$AP(\%)$	$AP_F(\%)$	$AP(\%)$	$AP_F(\%)$	$AP(\%)$	$AP_F(\%)$	$AP(\%)$	$AP_F(\%)$	$AP(\%)$	$AP_F(\%)$
Airplane	Boeing737	38.46	13.75	36.43	10.10	39.58	17.60	40.42	12.53	35.43	11.54
	Boeing747	55.36	20.37	50.68	11.81	73.56	35.53	52.86	23.37	47.88	13.42
	Boeing777	24.75	5.62	22.50	3.20	18.32	7.51	29.07	11.06	15.67	3.32
	Boeing787	51.81	13.57	51.86	13.42	56.43	31.22	52.47	21.77	48.32	11.23
	C919	0.81	0.31	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01
	A220	40.50	14.33	47.81	21.88	47.67	30.21	44.37	20.69	40.11	12.33
	A321	41.06	10.29	43.83	11.81	49.91	25.95	38.35	9.09	39.31	9.08
	A330	18.02	4.73	17.66	2.25	27.64	13.98	26.55	10.89	16.54	2.21
	A350	19.94	5.54	19.95	4.17	31.79	14.15	17.54	2.10	16.56	3.23
	ARJ21	1.70	0.27	0.13	0.02	0.00	0.00	0.00	0.00	0.01	0.01
	other-airplane	62.75	28.11	66.15	19.29	68.28	35.51	65.64	30.83	61.04	20.11
Ship	Passenger Ship	9.57	4.38	9.81	5.67	14.31	10.00	12.10	9.09	9.12	4.56
	Motorboat	22.55	9.09	28.78	9.09	28.07	16.10	28.84	12.76	23.34	9.06
	Fishing Boat	1.33	0.39	1.77	1.01	1.03	1.00	0.71	0.36	1.23	0.34
	Tugboat	16.37	9.09	17.65	9.09	14.32	9.09	15.35	9.09	15.67	9.09
	Engineering Ship	19.11	9.09	16.47	3.46	15.97	11.09	18.53	10.26	15.43	7.76
	Liquid Cargo Ship	14.26	9.09	16.19	9.09	18.04	10.99	14.63	9.09	15.32	8.98
	Dry Cargo Ship	24.70	9.70	27.06	9.09	26.02	13.84	25.15	13.99	25.43	9.34
	Warship	15.37	9.09	13.16	6.67	12.97	7.77	14.53	9.09	13.56	8.56
	other-ship	2.63	0.77	3.04	2.02	2.25	1.66	1.54	0.55	2.45	1.23
Vehicle	Small Car	65.20	15.85	68.42	21.60	68.80	38.95	68.19	37.54	66.23	16.67
	Bus	22.42	9.09	28.37	9.09	37.41	14.29	28.25	8.66	23.43	8.45
	Cargo Truck	44.17	9.09	51.24	9.09	53.96	21.79	48.62	18.38	46.78	10.32
	Dump Truck	35.37	9.37	43.60	9.09	45.68	18.62	40.40	13.79	36.56	9.67
	Van	52.44	13.19	57.51	13.82	58.39	26.55	58.00	20.96	53.78	13.45
	Trailer	19.17	9.09	15.03	2.88	16.22	5.59	13.66	5.33	14.32	4.56
	Tractor	1.28	0.85	3.04	2.34	5.13	6.90	0.91	0.79	16.39	1.45
	Excavator	17.03	9.09	17.99	7.55	22.17	10.39	16.45	9.09	16.92	8.23
	Truck Tractor	28.98	9.09	29.36	9.09	46.71	23.62	30.27	9.09	28.91	10.21
	other-vehicle	8.91	4.22	5.23	1.45	11.62	9.09	11.65	9.09	8.98	4.58
Court	Basketball Court	50.58	9.09	58.26	9.09	54.84	32.02	38.81	16.35	48.41	8.79
	Tennis Court	81.09	18.09	82.67	9.09	80.35	71.34	80.29	53.55	80.31	32.46
	Football Field	52.50	9.09	54.50	9.09	56.68	30.37	48.21	21.49	53.46	11.34
	Baseball Field	66.76	9.09	71.71	8.30	69.07	51.68	67.90	42.70	66.93	11.59
Road	Intersection	60.13	9.09	59.86	9.09	58.44	31.18	55.67	20.87	59.41	8.41
	Roundabout	17.41	5.77	16.92	4.54	18.58	6.50	20.35	9.09	16.25	4.29
	Bridge	12.58	6.06	11.87	4.77	31.81	15.42	12.62	7.88	10.39	5.15
mAP/mAP_F (%)		30.19	8.99	31.53	7.92	34.65	19.12	30.78	14.09	29.46	8.51

(a) and (b). However, (a) contains more information about another object and background, which is bad for fine-grained classification. Therefore, we design a novel FIoU for the task of fine-grained object detection to penalize the exceptional results.

$$FIoU = \sqrt[3]{\frac{G \cap D}{G \cup D} \cdot \frac{G \cap D}{G} \cdot \frac{G \cap D}{D}} \tag{1}$$

Fine-grained mean Average Precision (mAP_F). Different from generic object detection in remote sensing im-

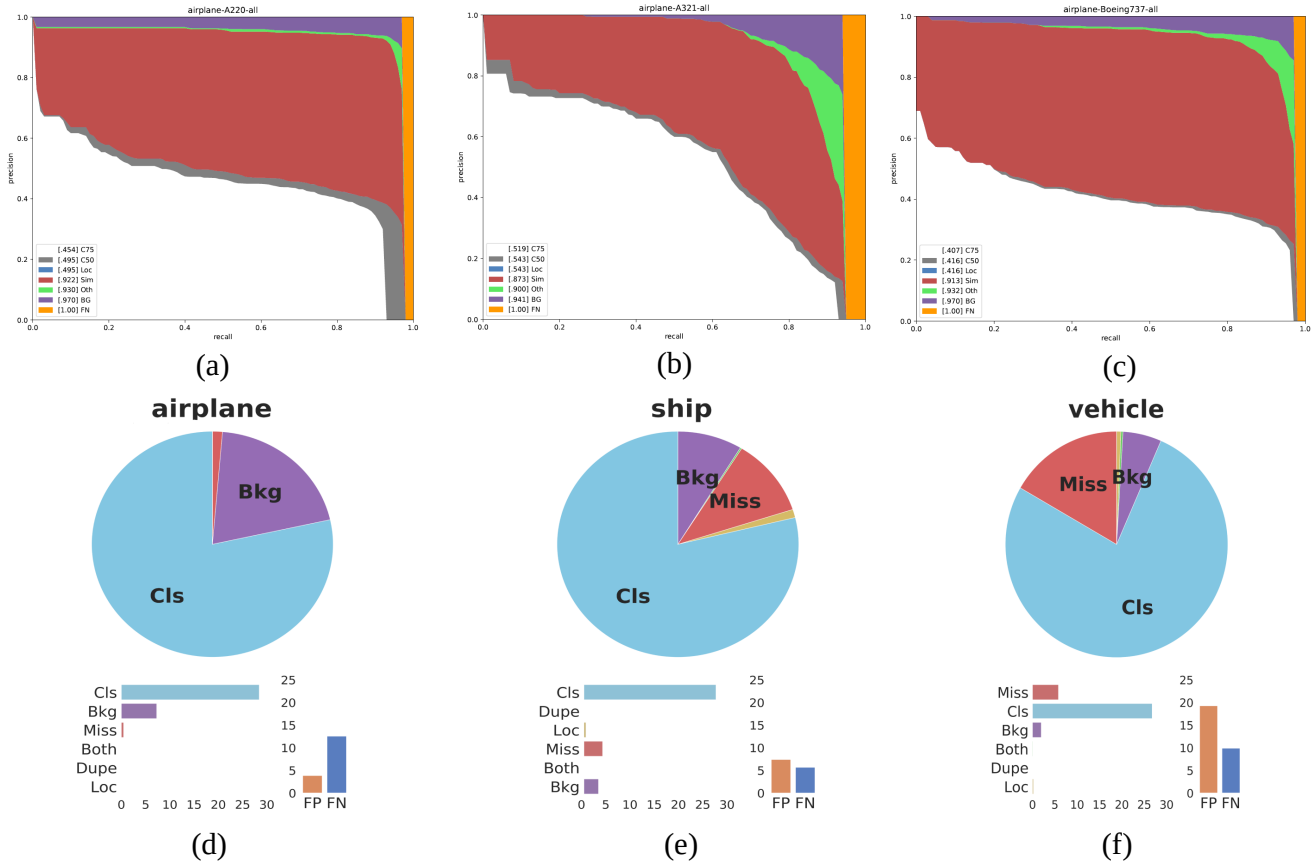


Figure 8: The false analysis of some categories in the FAIR dataset. (a), (b) and (c) show the false analyses between fine-grained airplanes using the COCO evalution tookit. (d), (e) and (f) show the false analyses of airplanes, ships and vehicles using TIDE toolbox.

Table 3
The benchmark results of FAIR1M dataset on the task of horizontal bounding boxes detection.

Method	mAP	mAP_F
RetinaNet [29]	30.68	8.45
Faster RCNN [39]	34.06	18.38
Cascade RCNN [2]	35.09	18.80

ages, the task of fine-grained object detection pays more attention to type recognition. As a result, we use $FIoU$ to obtain TP and FP. We define a detection box as TP if $FIoU$ is more than 0.5, otherwise it is FP. According to TP and FP, we can calculate recall and precision. We add the classification score as a constraint to the original formula of the precision. High scores indicate that we obtain a better detector.

$$Precision_F = \frac{FIoU \cdot TP \cdot score_{TP}}{TP \cdot score_{TP} + FP \cdot score_{FP}} \quad (2)$$

$$Recall_F = Recall = \frac{FIoU \cdot TP}{TP + FN} \quad (3)$$

where $score$ means the classification score of detected boxes. For each category, the precision is calculated with TP and FP as shown in Equation2, under the $FIoU$ threshold of 0.5 and a series of score thresholds from the VOC2012 [14]. Then the Average Precision (AP_F) is the mean of the collection of precisions. Finally, the mAP_F is calculated with the mean of AP_F over all categories. The mAP_F is a float value between 0.0 and 1.0. Compared with mAP in the task of generic object detection, mAP_F is more sensitive to fine-grained classification scores.

False analysis. In addition to quantitative metrics, we conduct two detailed false analyses: COCO evaluation toolkit [30] and the TIDE toolbox [1]. The COCO false analysis contains the curves of $C75$, $C50$, Loc, Sim, Oth, BG, and FN. We set the super-categories of all sub-categories to the corresponding generic categories. For example, the super-categories of Boeing737 and Boeing747 both are airplane. In this way, we can obtain the influence of similar objects and dissimilar objects. Compared with the COCO evaluation toolkit, the TIDE toolbox can calculate the contribution of each error type.

4.1.2. Benchmarks

Baseline models. We have investigated the state-of-the-art oriented object detection algorithms in the field of object detection from remote sensing imagery. We select RetinaNet

Figure 9: Visualization of detection results of testing on FAIR1M dataset using ROI Tranformer method.

[29], Faster R-CNN [39], Cascade R-CNN [2], Gliding Vertex [56], and ROI Transformer [12] as our baseline models with a ResNet-101 backbone [20]. To be specific, RetinaNet, Faster R-CNN and Cascade R-CNN represent region-based methods and proposal-based methods respectively, which are transferred from horizontal object detection in natural scenes. ROI Transformer and Gliding Vertex represent different oriented object detection methods designed for object detection from remote sensing images. To better analyze the results of models, we crop images into patches with 1024 × 1024 pixels. Since the dataset is still under construction, we use a part of the dataset for training and testing. The division of the dataset is the same as described in the section 3. To ensure that there is no duplicate data in different subsets, we divide datasets according to the cities.

Baseline results. We train and evaluate 5 detectors on the task of oriented bounding boxes detection and 3 detectors on the task of horizontal bounding boxes detection. The results in Table 2 and Table 3 show the difficulty of generic object detection methods in detecting fine-grained objects. We calculate AP and AP_F for each category. For the categories with obvious features, such as Boeing747, detectors

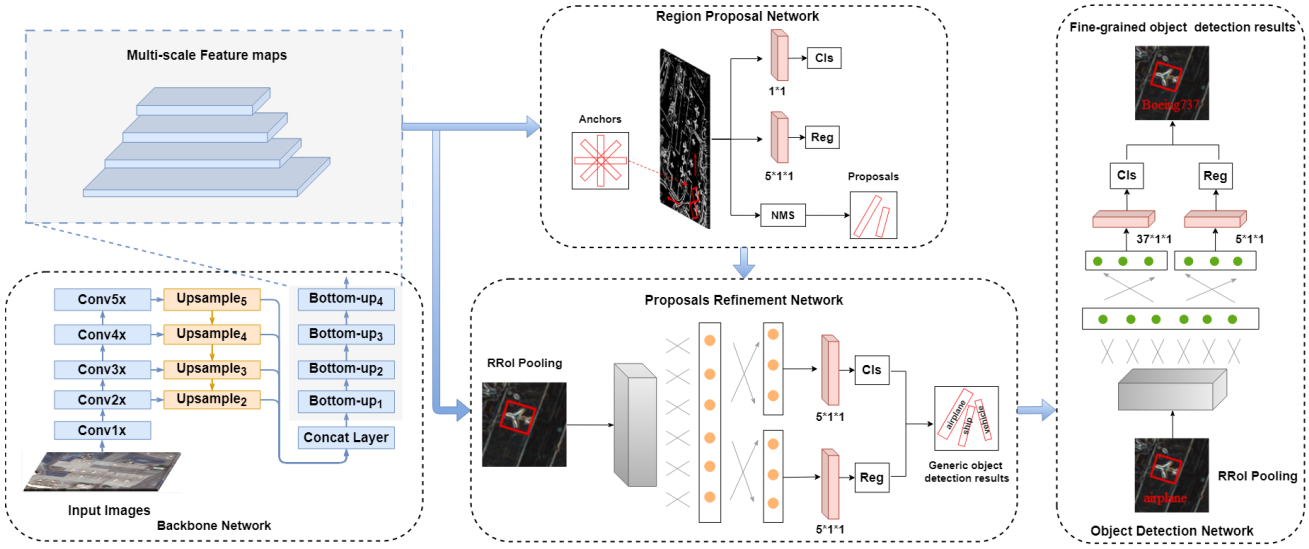


Figure 10: The structure of cascaded hierarchical object detection network.

obtain better results. The performance of detectors on most categories verifies the difficulty of our dataset. The number, distribution, and characteristics of different objects have caused the imbalance of the detection accuracy. Generally, objects with a large number of instances and obvious features are easier to obtain higher detection accuracy. While, some categories of object detection accuracy is quite low, for example, C919 and ARJ21. The main reason for this result is that these two types of airplanes are very rare, so it is necessary to study few-shot learning methods to improve the detection accuracy of these objects. Although the mAP values of the five algorithms are close, there are great differences in the mAP_F value. The combination of AP and AP_F is helpful for analyzing the performance of the detectors.

We conduct false analysis on each category to further verify the challenge of our dataset. Figure 8 shows two kinds of false analyzes of several categories on the result of the ROI Transformer method. We plot $C75$, $C50$, Loc, Sim, Oth, BG, and FN to analyze the within-class performance. Figure 8 (a), (b) and (c) in 8 show that:

1. Removing localization error (Loc) brings some improvement on AP, which shows the influence of the large range of sizes and orientations in our dataset.
2. Removing the Sim and FN results in more improvements than removing the Oth for each category, which means the influence of similar objects is larger than the influence of dissimilar objects for most categories. It is challenging that researchers need to design more suitable algorithms to detect fine-grained objects accurately.
3. In addition to Sim and FN, complex background also has a large impact on detection results. The complex background in our dataset is also challenging.

Figure 8 also shows the contribution of each error type for each category. It is obvious that the classification error accounts for most of the error. Moreover, missdetection has a great impact on ships and vehicles due to their small sizes. The results show more difficulty in detecting objects in the FAIR1M dataset.

Failure cases. Figure 9 shows the visualization of detection results using the ROI Transformer model. In addition to classification, there are undetected objects in the results. The complex background and small size of objects result in the misdetection of small objects.

4.1.3. Hierarchical Object Detection Method

According to the above experiments, although fine-grained detection is a detection task, the challenge lies in how to classify the objects. It is relatively difficult for detectors to learn the feature information of the 37 fine-grained categories. The categories of objects in the FAIR1M dataset are organized in a hierarchy, which is designed coarse to fine. Considering this characteristic of the FAIR1M dataset, we propose a cascaded hierarchical object detection network (CHODNet). It can learn external and internal representations independently from the dataset using a cascaded hierarchical structure. Figure 10 shows the structure of the CHODNet.

Framework. We build the CHODNet based on oriented Faster RCNN. Compared with Faster RCNN, CHODNet adds a training stage to learn the information of coarse categories. As shown in Figure 10, CHODNet consists of four stages: feature refinement network, region proposal network, proposals refinement network, and fine-grained detection network.

- **Feature Refinement Network.** We use the ResNet-

(a) (b) (c) (d)

Figure 11: The detection results of CHODNet on multi-temporal images. (a), (b) Gaoqi airport. (c), (d) Dalian port.

101 as our backbone. As we know, the high-level layers and low-level layers in a deep learning network contain semantic information and localization information [15], respectively. To generate more discriminative feature maps, we upsample multi-scale feature maps to the same scale, and build a bottom-up feature hierarchy.

- **Region Proposal Network.** Due to the oriented annotations in the FAIR1M dataset, we generate oriented anchors in the first stage. Each anchor can be represented as a five-tuple (x, y, w, h, θ), where (x, y) denotes the center coordinates of the anchor, (w, h) specifies the width and height of it, respectively, and θ means the angle between the long side and the horizontal direction. We generate K anchors with K_r aspect ratios, K_s scales, and K_a angles, where $K = K_r \times K_s \times K_a$. At each position of feature map, a regression output layer generates $(K \times 5)$ vectors to encode the offset of anchors, and a classification output layer generates $(K \times 2)$ scores to predict whether the anchor is positive. The loss function of this stage is a multi-

task loss similar to the one of Faster RCNN, which is defined as:

$$L_{orpn} = L_{cls}(p, u) + L_{loc}(t, t^*). \quad (4)$$

where L_{cls} and L_{loc} represent the classification loss and localization loss in the oriented region proposal network, respectively. For L_{cls}, the parameter u is 1 if the anchor is positive, otherwise it is 0. The parameter p is the predicted score for anchors over background and foreground. t and t^* denote the predicted regression offset and ground-truth box, respectively.

- **Proposals Refinement Network.** Compared with the R-CNN subnetwork in the Faster RCNN, the proposal refinement network is mainly used to output the coarse category classification result and localization offset for each proposal. There are 5 coarse categories in the FAIR1M dataset. Therefore, the output layers of classification and regression both are of size $K \times 5$. The loss function is still a multi-task loss about the classification and localization:

$$L_{prn} = L_{cls}(p, v) + L_{loc}(t, t^*) \quad (5)$$

where v denotes the score of coarse categories.

- **Fine-grained Detection Network.** Different from the previous stage, the fine-grained detection network will output scores of all fine-grained categories. The parameters in fine-grained detection network are learned specifically for fine categories. As shown in Figure 10, the output layers of classification and regression are $K \times 37$ and $K \times 5$, respectively. The loss of this stage is defined as:

$$L_{fdn} = L_{cls}(p, w) + L_{loc}(t, t^*) \quad (6)$$

Where, w denotes the score of fine categories.

Loss Function. There are 3 stages in our CHODNet. The loss function of CHODNet is a weighted summation of oriented region proposal network, proposals refinement network, and fine-grained object detection network. In other words, the total loss consists of foreground/background loss, coarse loss, and fine-grained loss.

$$L_{CHODNet} = L_{orpn} + \lambda_1 L_{prn} + \lambda_2 L_{fdn} \quad (7)$$

Where λ_1 and λ_2 denote the weights of different losses.

Staged Training Strategy. We have defined a weighted summation of different losses. When training the detector, we hope it can give priority to learning the features of coarse categories. After the parameters of the current two stages are trained, we focus on the learning of the third stage. Therefore, we develop a stage training strategy to focus on different losses in an end-to-end manner. The staged training strategy can modify the value of the weights λ_1 and λ_2 while training the detector, which denote the contributions of different losses. For the three-stage detecting, the initial weights of the loss are [0.7, 0.3], then they are adjusted to [0.3, 0.7] after two hundred thousand steps. The loss function with the highest weight can be regarded as the focus of the current step.

Results and Temporal analysis. After adopting the staged training strategy, CHODNet obtains 31.12% mAP. Adding an extra branch of learning coarse categories can improve the accuracy of fine-grained object detection. As shown in Figure 11, we test the CHODNet on multi-temporal images. The two rows of Figure 11 show the Gaoqi airport and Dalian port at different times. Figure 11 (a) and (b) are the Gaoqi Airport in 2015 and 2016, respectively. Figure (c) and Figure (d) are the Dalian port in 2016 and 2017, respectively. Although the lighting and timing of the images are different, the detectors can still detect the objects.

Table 4
The information of three FGVC datasets.

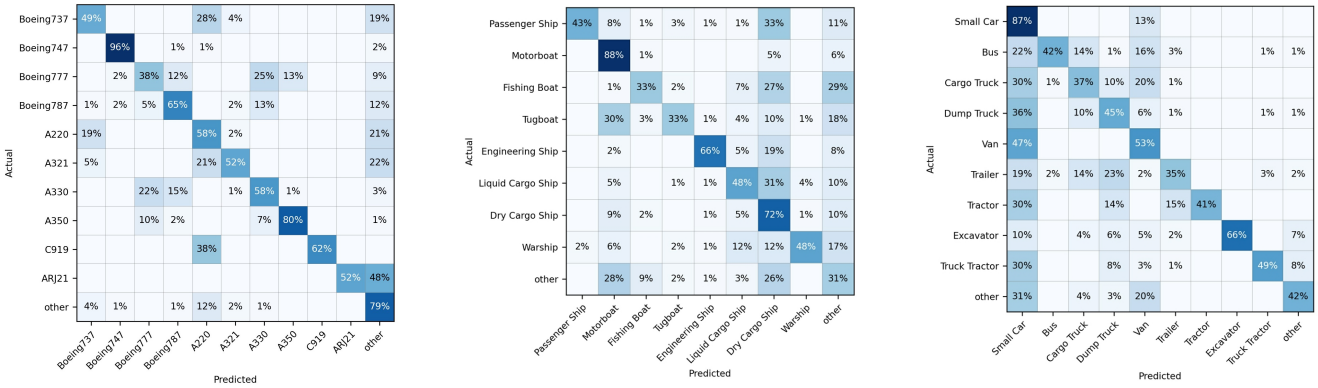
Category	Instance
Airplane_train	8000
Airplane_test	4000
Ship_train	8000
Ship_test	4000
Vehicle_train	10000
Vehicle_test	6000

4.2. Fine-grained Image Classification

Fine-grained visual categorization (FGVC) aims to obtain the category of an input image. Many FGVC datasets have been proposed in the field of natural images, such as the CUB-200-2011 dataset [47], the Stanford Cars dataset [24], the FGVC-Aircraft dataset [50], and Stanford Dogs dataset [22]. There are relatively few datasets about FGVC in the field of remote sensing imagery. HR-SAR and MR-SAR [57] are two SAR ship datasets. Sumbul et al. [46] propose a street tree dataset in aerial images. It is necessary to build FGVC datasets and develop the FGVC task for movable objects in remote sensing images. Based on the FAIR1M dataset, we generate three FGVC datasets: FAIR-Airplane dataset, FAIR-Ship dataset, and FAIR-Vehicle dataset.

Dataset Information. According to the bounding box annotations of the FAIR1M dataset, we select relatively large airplanes, ships, and vehicles and crop them from the original images. The information about the number of instances of the three datasets can be seen in Table 4.

Evaluation Metrics. To evaluate the performance of FGVC algorithms, we use the classification accuracy as the metric for this task. $N_{correct}$ and N_{all} denote the number of correctly predicted instances and total instances, respec-



(a) (b) (b)

Figure 12: The confusion matrices of MMAL-Net on three fine-grained datasets.

Table 5
The results of three FGVC methods on our datasets. The evaluation metric is mAP and its provided values refer to %.

Method	Airplane	Ship	Vehicle
ResNet-50	41.23	37.86	38.93
ResNet-101	43.46	38.64	39.76
MMAL-Net	45.18	40.27	41.43

tively. The definition of the classification accuracy is as follows:

$$Acc = \frac{N_{correct}}{N_{all}} \tag{8}$$

Baseline Classifiers and Results. We choose two fundamental image classifiers (*i.e.* ResNets) and a state-of-the-art fine-grained image classifier (*i.e.* MMAL-Net [61]) as our baseline classifiers. ResNet-50 and ResNet-101 are usually used for generic image classification. MMAL-Net has achieved state-of-the-art results on the CUB200-2011, FGVC-Aircraft and Stanford Cars datasets.

We train the three algorithms on three fine-grained datasets. As shown in Table 5, ResNet-50 and ResNet-101 show more difficulty in fine-grained classification in remote sensing images. To show the results of various categories more clearly, we provide the confusion matrices on the three fine-grained datasets using the MMAL-Net method. Except for the small number of objects (*i.e.* regarding the sub-categories *C919* and *ARJ21*), the MMAL-Net performs relatively well on the FAIR-Airplane dataset. However, the MMAL-Net shows more difficulty in classifying fine-grained ships and vehicles. Only *Motorboat* and *Small Car* have relatively high confidence. The remaining categories are confused with other categories to some extent. Therefore, it is challenging to design fine-grained classifiers on three datasets for researchers.

4.3. Cross-dataset Validations

In order to verify the generalization performance of our dataset, we use the DOTA dataset to do cross-dataset validation experiments, which is one of the largest and OBB-style object detection datasets in the field of remote sensing. For there are large-scale images in the DOTA dataset, we crop images into patches with 1024 × 1024 pixels. We only choose generic categories in the FAIR1M dataset to do cross-dataset validations because the DOTA dataset does not have fine-grained categories. ROI Transformer is used to be the testing detector for the experimental results presented in Table 6.

The difference between the two datasets is 27.92% mAP and 6.77% mAP, respectively. The results in Table 6 show that the FAIR1M dataset covers the characteristics of the DOTA dataset and has more types and patterns not contained in the DOTA dataset. ROI Transformer-D and ROI Transformer-F get lower results on FAIR1M, which verifies the challenge of the FAIR1M dataset.

5. Conclusion

In this paper, we propose a more challenging dataset for fine-grained object detection and recognition in high-resolution remote sensing images. We believe the diversity and challenge of the FAIR1M dataset will benefit from fine-grained types, large range of sizes and orientations, high within-class variation and between-class similarity, complex scenes, and geographic information. We introduce the collection, category, annotation and characteristics of our dataset. Finally, we implement a series of state-of-the-art algorithms to build an object detection benchmark to foster future research. With the development of remote sensing image interpretation technology, coarse object recognition cannot meet the requirements well. We hope that this dataset can enhance the development of fine-grained object recognition in the field of remote sensing images.

Table 6
Results of cross-dataset generalization. ROI Transformer-D and ROI Transformer-F are trained based on the DOTA dataset and FAIR1M dataset, respectively. The evaluation metric is mAP and its provided values refer to %.

Testing set	Detector	Airplane	Ship	Vehicle	Aug.
DOTA	ROI Transformer-D	90.28	81.31	71.85	81.15
	ROI Transformer-F	81.54	61.45	59.10	67.36
FAIR1M	ROI Transformer-D	80.95	33.66	45.09	53.23
	ROI Transformer-F	90.74	39.06	51.98	60.59

Acknowledgment

We are very grateful for the support of the ISPRS Scientific Initiatives 2021. To promote the academic research on object detection and recognition in high-resolution satellite images, the FAIR1M dataset will be a standard and large-scale dataset foundation for the ISPRS benchmark. It is also supported by the National Natural Science Foundation of China 61725105.

References

[1] Bolya, D., Foley, S., Hays, J., Hoffman, J., 2020. Tide: A general toolbox for identifying object detection errors. arXiv preprint arXiv:2008.08115 .

[2] Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162.

[3] Chen, K., Wu, M., Liu, J., Zhang, C., 2020. Fgsd: A dataset for fine-grained ship detection in high resolution satellite images. arXiv preprint arXiv:2003.06832 .

[4] Chen, X., Xiang, S., Liu, C., Pan, C., 2014. Vehicle detection in satellite images by hybrid deep convolutional neural networks. IEEE Geoscience and Remote Sensing Letters 11, 1797–1801.

[5] Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 117, 11–28.

[6] Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 54, 7405–7415.

[7] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The cityscapes dataset for semantic urban scene understanding, in: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[8] Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R., 2018. Deepglobe 2018: A challenge to parse the earth through satellite images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 172–181.

[9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 248–255.

[10] Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., Zou, H., 2018. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS Journal of Photogrammetry and Remote Sensing 145, 3–22.

[11] Diemunsch, J.R., Wissinger, J., 1998. Moving and stationary target acquisition and recognition (mstar) model-based automatic target recognition: Search technology for a robust atr, in: Algorithms for synthetic aperture radar Imagery V, International Society for Optics and Photonics. pp. 481–492.

[12] Ding, J., Xue, N., Long, Y., Xia, G.S., Lu, Q., 2019. Learning roi transformer for oriented object detection in aerial images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2849–2858.

[13] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88, 303–338.

[14] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[15] Fu, K., Chang, Z., Zhang, Y., Xu, G., Zhang, K., Sun, X., 2020. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 161, 294–308.

[16] Guo, W., Yang, W., Zhang, H., Hua, G., 2018. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. Remote Sensing 10, 131.

[17] Han, J., Zhang, D., Cheng, G., Guo, L., Ren, J., 2015. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Transactions on Geoscience and Remote Sensing 53, 3325–3337.

[18] Han, X., Zhong, Y., Zhang, L., 2017. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sensing 9, 666.

[19] Haroon, M., Shahzad, M., Fraz, M.M., 2020. Multi-sized object detection using spaceborne optical imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing .

[20] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

[21] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al., 2017. Speed/accuracy trade-offs for modern convolutional object detectors. computer vision and pattern recognition , 3296–3297.

[22] Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L., 2011. Novel dataset for fine-grained image categorization, in: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO.

[23] Kong, T., Yao, A., Chen, Y., Sun, F., 2016. Hypernet: Towards accurate region proposal generation and joint object detection. computer vision and pattern recognition , 845–853.

[24] Krause, J., Stark, M., Deng, J., Fei-Fei, L., 2013. 3d object representations for fine-grained categorization, in: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia.

[25] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al., 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 .

[26] Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., Bulatov, Y., McCord, B., 2018. xview: Objects in context in overhead imagery. arXiv preprint arXiv:1802.07856 .

[27] Li, K., Wan, G., Cheng, G., Meng, L., Han, J., 2020. Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS Journal of Photogrammetry and Remote Sensing 159, 296–307.

[28] Li, Z., Zhou, F., 2017. Fssd: Feature fusion single shot multibox detector. arXiv: Computer Vision and Pattern Recognition .

[29] Lin, T., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal loss for dense object detection , 2999–3007.

[30] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: European conference on computer vision, Springer. pp. 740–755.

[31] Liu, K., Mattyus, G., 2015. Fast multiclass vehicle detection on aerial images. IEEE Geoscience and Remote Sensing Letters 12, 1938–1942.

[32] Liu, Z., Wang, H., Weng, L., Yang, Y., 2016. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geoscience and Remote Sensing Letters 13, 1074–1078.

[33] Liu, Z., Yuan, L., Weng, L., Yang, Y., 2017. A high resolution optical satellite image dataset for ship recognition and some new baselines , 324–331.

[34] Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE. pp. 3226–3229.

[35] Mundhenk, T.N., Konjevod, G., Sakla, W.A., Boakye, K., 2016. A large contextual dataset for classification, detection and counting of cars with deep learning, in: European Conference on Computer Vision, Springer. pp. 785–800.

[36] Padwick, C., Deskevich, M., Pacifici, F., Smallwood, S., 2010. Worldview-2 pan-sharpening, in: Proceedings of the ASPRS 2010 Annual Conference, San Diego, CA, USA, pp. 1–14.

[37] Razakarivony, S., Jurie, F., 2016. Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation 34, 187–203.

[38] Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 .

[39] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: towards real-time object detection with region proposal networks 2015, 91–99.

[40] Ren, Y., Zhu, C., Xiao, S., 2018. Deformable faster r-cnn with aggregating multi-layer features for partially occluded object detection in optical remote sensing images. Remote Sensing 10, 1470.

[41] Roscher, R., Volpi, M., Mallet, C., Drees, L., Wegner, J.D., 2020. Semcity toulouse: A benchmark for building instance segmentation in satellite images. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences 5, 109–116.

[42] Rottensteiner, F., Sohn, G., Gerke, M., Wegner, J.D., Breitkopf, U., Jung, J., 2014. Results of the isprs benchmark on urban object detection and 3d building reconstruction. ISPRS Journal of Photogrammetry and Remote Sensing 93, 256–271.

[43] Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, pp. 293–298.

[44] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J., 2019. Objects365: A large-scale, high-quality dataset for object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 8430–8439.

[45] Shen, Z., Liu, Z., Li, J., Jiang, Y., Chen, Y., Xue, X., 2017. Dsod: Learning deeply supervised object detectors from scratch. international conference on computer vision , 1937–1945.

[46] Sumbul, G., Cinbis, R.G., Aksoy, S., 2017. Fine-grained object recognition and zero-shot learning in remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 56, 770–779.

[47] Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report.

[48] Wang, C., Bai, X., Wang, S., Zhou, J., Ren, P., 2018. Multiscale visual attention networks for object detection in vhr remote sensing images. IEEE Geoscience and Remote Sensing Letters 16, 310–314.

[49] Wang, P., Sun, X., Diao, W., Fu, K., 2019. Fmssd: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 58, 3377–3390.

[50] Wu, Q., Sun, H., Sun, X., Zhang, D., Fu, K., Wang, H., 2014. Aircraft recognition in high-resolution optical satellite remote sensing images. IEEE Geoscience and Remote Sensing Letters 12, 112–116.

[51] Wu, X., Hong, D., Tian, J., Chanussot, J., Li, W., Tao, R., 2019. Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features. IEEE Transactions on Geoscience and Remote Sensing , 1–13.

[52] Wu, Z.Z., Wan, S.H., Wang, X.F., Tan, M., Zou, L., Li, X.L., Chen, Y., 2020. A benchmark data set for aircraft type recognition from remote sensing images. Applied Soft Computing 89, 106132.

[53] Xia, G.S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., Zhang, L., 2018. Dota: A large-scale dataset for object detection in aerial images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3974–3983.

[54] Xian, S., Zhirui, W., Yuanrui, S., Wenhui, D., Yue, Z., Kun, F., 2019. Air-sarship–1.0: High resolution sar ship detection dataset. J. Radars 8, 852–862.

[55] Xiao, Z., Liu, Q., Tang, G., Zhai, X., 2015. Elliptic fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images. International Journal of Remote Sensing 36, 618–644.

[56] Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.S., Bai, X., 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[57] Xu, Y., Lang, H., 2020. Distribution shift metric learning for fine-grained ship classification in sar images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 2276–2285.

[58] Xu, Z., Xu, X., Wang, L., Yang, R., Pu, F., 2017. Deformable convnet with aspect ratio constrained nms for object detection in remote sensing imagery. Remote Sensing 9, 1312.

[59] Yan, J., Wang, H., Yan, M., Diao, W., Sun, X., Li, H., 2019. Iou-adaptive deformable r-cnn: Make full use of iou for multi-class object detection in remote orientation robustsensing imagery. Remote Sensing 11, 286.

[60] Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification, in: Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, pp. 270–279.

[61] Zhang, F., Li, M., Zhai, G., Liu, Y., 2020. Multi-branch and multi-scale attention learning for fine-grained visual categorization. arXiv:2003.09150.

[62] Zhang, Y., Yuan, Y., Feng, Y., Lu, X., 2019. Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection. IEEE Transactions on Geoscience and Remote Sensing 57, 5535–5548.

[63] Zhu, H., Chen, X., Dai, W., Fu, K., Ye, Q., Jiao, J., 2015. Orientation robust object detection in aerial images using deep convolutional neural network, in: 2015 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 3735–3739.

[64] Zou, Z., Shi, Z., 2017. Random access memories: A new paradigm for target detection in high resolution aerial remote sensing images. IEEE Transactions on Image Processing 27, 1100–1111.