Revisiting the Rationality Few-shot Detection Benchmarks

Tianbo Wang,[1] **Renshuai Tao,** [1*] **Jiaying Wang,** [2] **Bowei Jin** [3]

[1] State Key Laboratory of Software Development Environment, Beihang University
[2] Shanghai Aerospace Electronic Technology Institute
[3] IFLYTEK (Suzhou) Technology Co., Ltd

Abstract

Traditional training strategies for CNN-based detection models require a large quantity of samples with bounding-box annotations. However, in many open-world scenarios (e.g., MRI medical diagnosis, X-ray security inspection), preparing such high-quality training data is unrealistic, causing the performance drop on the categories with few training samples. Recently, researchers propose the Few-Shot Object Detection(FSOD) Task, aiming to address this dilemma that accurately detect objects with few annotated samples. And Existing FSOD benchmarks the researchers exploit are mainly based on the categories selected from the classical visual datasets like MS COCO and Pascal VOC. However, due to that the images from these datasets illustrate common scenarios in daily life, the samples are easy to acquire and these object are are bright in color and easy to detect. In this paper, we first point out that FSOD task is rational and meaningful only when they are based on the extreme scenario, that the training samples are hard to acquire and the objects are not easy to detect. Therefore, we select a typical scenario, X-ray security inspection, and present the Rational Few-shot (RFS) benchmark. The RFS dataset consists of 12,333 images containing 41,704 instances with bounding-box annotations of 20 categories. We introduce the construction principle and rationality in detail, hoping our work can serve a new perspective to the few-shot detection research community.

Introduction

With the development of deep learning, many state-of-art methods using deep convolutional neural networks (CNNs) have been proposed to resolve diverse visual tasks, including classification, semantic segmentation and object detection in recent years. For object detection, CNN-based methods have achieved a great success on various datasets (Lin et al. 2014; Everingham et al. 2010). The application of deep learning methods to object detection leads to the urgent requirement for a huge amount of training samples to achieve satisfactory performance. However, in some real-world scenarios like MRI medical diagnosis (Chaudhary, Hazra, and Chaudhary 2019; Guo et al. 2019; Lu and Tong 2019) and X-ray security inspection (Xiao and Marlet 2020; Kang et al. 2019; Karlinsky et al. 2019), acquiring sufficient training samples is usually time-consuming and labor-intensive. Thus, researchers have been paying their attention to exploring the task that how to learn from few samples.

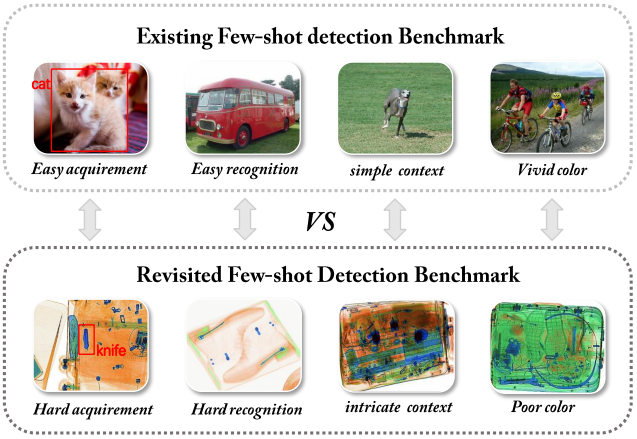

Figure 1: Comparison of samples in existing datasets and our datasets in the few-shot detection task.

Few-Shot Object Detection (FSOD), trying to accurately localize objects of novel classes in the case of few samples with annotated bounding-boxes, is a challenging task in computer vision, and attracts the attention of many researchers. Various CNN-based models achieving remarkable performance have been proposed, *i.e.*, optimizing the detection network architecture (Karlinsky et al. 2019; Fan et al. 2020) and proposing novel modules (Chen et al. 2021; Kang et al. 2019). The experiments of these traditional models are mainly conducted on the datasets whose categories and samples are selected from the classical visual datasets, like MS COCO (Lin et al. 2014) and Pascal VOC (Everingham et al. 2010). As showed in Figure 1, these categories in classical visual datasets are mainly acquired from common scenarios in daily life and the instances are usually bright in color and easy to recognize. It should be note that the few-shot task is aiming to solve the dilemma where the samples are difficult to obtain. Moreover, the instances of such scenarios are also accompanied by incomplete information, caused by occlusion, low-frequency appearance, *etc*. There-

*Corresponding author, rstao@buaa.edu.cn

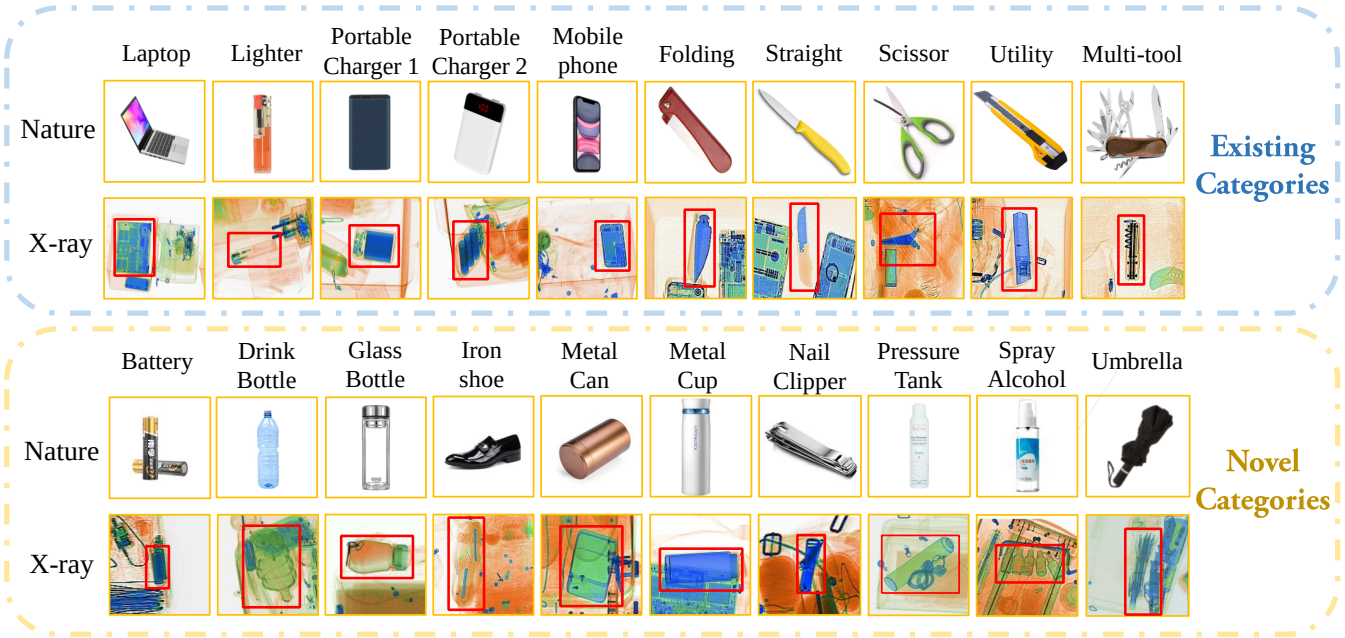

Figure 2: The X-ray images and nature images of whole 20 categories in RFS, including some common prohibited items like"Battery", 'Spray Alcohol',*etc*.

fore, the samples of traditional datasets cannot meet the requirement of real-world few-shot detection applications.

In this paper, aiming to exploring real-world few-shot detection task, we release the Rational Few-shot (RFS) dataset, providing a reasonable and fair benchmark for relevant researchers to evaluate their models. In RFS datset, we select the typical scenario in the real world, X-ray security inspection. There are mainly two advantages of selecting this typical scenario: 1) Some special categories of prohibited items appear at a very low frequency, making it extremely difficult to acquire the samples. 2) Due to the occlusion caused by objects overlapped with each other and the fuzziness of X-ray imaging, the X-ray security inspection is more similar to the real-world few-shot detection scenario.

RFS dataset consists of 12,333 X-ray images totally, with 41,704 annotated instances of 20 different categories. In these 20 categories, we select 5 original categories from the OPIXray dataset (Wei et al. 2020) and 5 from the HiXray dataset (Tao et al. 2021). Moreover, we add another 10 new categories to enrich the number of categories and meet the requirement of a standard few-shot object detection dataset. We hope the interesting RFS dataset we construct can serve a new perspective to Few-shot object detection.

Related Work

Datasets for Few-Shot Detection

Most of the FSOD works(Fan et al. 2020; Yan et al. 2019; Kang et al. 2019; Wang et al. 2020; Chen et al. 2021; Karlinsky et al. 2019; Xiao and Marlet 2020) evaluate their approach on traditional visual benchmarks(Lin et al. 2014; Everingham et al. 2010; Gupta, Dollar, and Girshick 2019). MS COCO(Lin et al. 2014) is one of the most well-known benchmark for object detection, segmentation and key-point detection. This benchmark is a large-scale dataset consisting of 328K images, 2.5 million labeled instances. The multiple classes of objects and complex background of images make MS COCO gain more favor from FSOD tasks. Pascal VOC(Everingham et al. 2010) contains 20 common object categories like car, bus and person. Each of the images in PASCAL VOC is well annotated in different ways, which give a huge support for various visual tasks. LVIS(Gupta, Dollar, and Girshick 2019) is another recently released benchmark commonly used for FSOD. This benchmark has annotations for over 1000 object categories in 164k images which are sufficient.

To ensure fair and direct comparison with previous works, most of the FSOD works stick to a consistent data construction and evaluation principle(Xiao and Marlet 2020; Kang et al. 2019; Karlinsky et al. 2019; Fan et al. 2020; Yan et al. 2019) to establish data for few-shot object detection on the datasets mentioned above. They separate all of classes contained in the dataset into two parts: one part is regarded as base classes with adequate annotations and the other part are novel classes with K-shot annotated instances. The K-shot means every novel class only remain k objects with bounding-box and label for adjusting models in training process. Just take Pascal VOC as example, the whole dataset(totally 20 classes) is divided into 15 base classes and 5 novel classes. The number of shots K is usually set to 1, 2, 3, 5 and 10. Only the annotations of base classes are available during the base training. The k annotated instances with bounding-box for novel classes are used for few-shot fine-tuning process.

Category	BA	DB	FK	GB	IS	LA	LI	MC1	MC2	MP	MK	NC	PC1	PC2	PT	SC	SA	SK	UM	UK	Total
Training	4,248	579	213	2,948	1,337	2,381	938	1,372	1,865	4,903	1,623	800	1,729	1,836	910	1,294	1,888	968	1,874	602	34,308
Testing	895	140	45	603	302	543	177	331	424	1,063	324	200	340	384	216	240	414	171	431	153	7,396
Total	5,143	719	258	3,551	1,639	2,924	1,115	1,703	2,289	5,966	1,947	1,000	2,069	2,220	1,126	1,534	2,302	1,139	2,305	755	41,704

Table 1: The statistics of category distribution of RFS.The name BA, DB, FK, GB, IS, LA, LI, MC1, MC2, MP, MK, NC, PC1, PC2, PT, SC, SA, SK, UM, UK denote “Battery”, “Drink Bottle”,“Folding Knife”,“Glass Bottle”,“Iron Shoe”,“Laptop”,“Lighter”,“Metal Can”,“Metal Cup”,“Mobile Phone”,“Multi-tool Knife”,“Nail Clippers”,“Portable Charger 1 ”, “Portable Charger 2 ”, “Pressure Tank”, “scissor”, “Spray Alcohol”, “Straight Knife”, “Umbrella” and “Utility Knife” respectively.

Rational Few-shot Benchmark

Without any objection that datasets of more rational and practical scenarios are of great significance to conduct experiment and evaluate methods. As mentioned above, existing datasets for the experiments of FSOD tasks are mainly classical visual datasets, which are considered lack evaluability for the real few-shot object detection task. However, little attention has been paid to constructing a satisfiable benchmark with rational scenario for FSOD. X-ray security inspection, whose context is more complicated and samples are difficult to obtain comparing to natural scene, could be a better choice. Thus, for FSOD tasks, we contribute the first X-ray benchmark named Rational Few-shot (RFS) benchmark mainly based on OPIXray(Wei et al. 2020) and HiXray(Tao et al. 2021), which are newly released X-ray benchmarks in recent years.

Construction Criterion

We will introduce the criteria when constructing RFS from the following three aspects.

Datasets	Released Year	Category	Task
OPIXray	2020	5	Detection
HiXray	2021	8	Detection
RFS	2021	**20**	**Few-Shot Detection**

Table 2: The comparison of detail information of different X-ray Datasets.

Image Source. RFS is a combination of OPIXray, HiXray and new source labeled X-ray images. As is shown in Table2, both of OPIXray and HiXray contains no more than 10 categories of instances. The number of categories contained in a single benchmark does not satisfy the required number(at least 20) of categories of an existed standard FSOD benchmark. Therefore, we combine the X-ray images meticulously selected from the two benchmarks and a batch of newly acquired images together to construct RFS. These newly acquired X-ray images are also produced by security inspection machine in our daily life, ensuring the authenticity of data source.

Category Selection. RFS totally contains 20 categories of instances. We first choose all 5 categories:“Folding Knife”, “Multi-tool Knif”, “Straight Knife”, “Utility Knife”, “Scissor” in OPIXray and 5 representative categories: “Portable Charger 1”, “Portable Charger 2”, “Mobile Phone”, “Lighter”, “Laptop” in HiXray as part of categories in RFS. The other 10 novel categories of RFS includes “Battery”, “Drink Bottle”, “Glass Bottle”, “Iron Shoe”, “Metal Can”, “Metal Cup”, “Nail Clippers”, “Pressure Tank”, “Spray Alcohol” and “Umbrella”, which are contained in the newly acquired images. Both nature and X-ray images of the whole 20 categories can be seen in Figure2. We choose the above 10 novel categories for two reasons: 1)some of these categories, like ’Spray Alcohol’, are prohibited items known by passengers. Thus the proba-

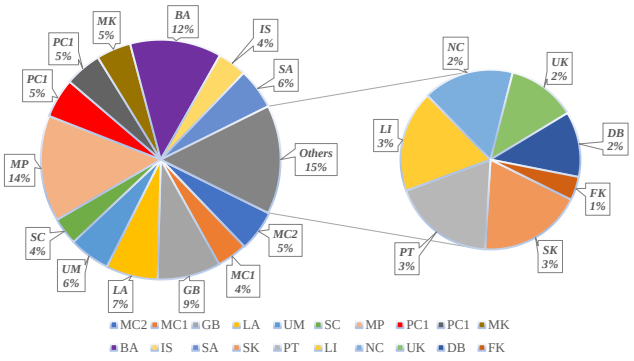

Figure 3: The proportion of category distribution in RFS. The left pie chart shows the overall proportion of all 20 categories. The right pie chart extracts data of 7 categories with the minimum proportion and combines into a new chart in order to make the smaller percentages more readable.

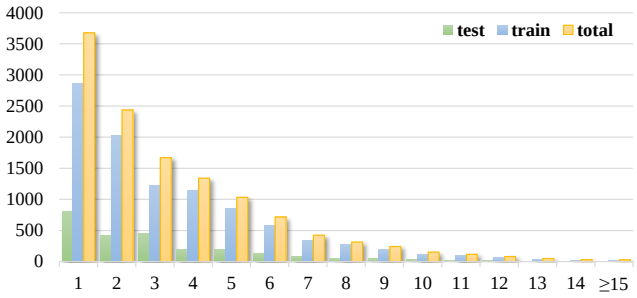

Figure 4: The distribution of numbers of instances per image.

bility of occurrence of these items is relatively low. And 2) instances of some classes, like 'Iron Shoe', may be heavy in weight or large in size, causing the inconvenience to carry. These two reasons make the sample acquisition harder, corresponding with the applicable scenario for FSOD.

Annotation Standard. To ensure the completeness of the annotation, we re-annotate the images selected from OPIXray and HiXray due to the reason that there could be object which belongs to the 10 novel categories but not be annotated in original images. Both the newly acquired images and the re-annotated images are annotated by professional security inspectors manually. Also, the annotating procedure follows the similar standards with Pascal VOC, including how to annotating bounding-box and how to resolve occlusion, to guarantee the quality of annotations.

Data Attribute

We further analyse the detail information of data attributes.

Instances per category. RFS contains 12,333 X-ray images, which are separated into training set with 9,867 images and testing set with 2,466 images, and 20 categories of totally 41,704 annotated objects. The number of instances for each category is shown in Table 1, and the proportion of different categories can be seen in Figure 3.

Instances per image. According to the statistics, each image in RFS has at least one instance. The image with the most objects has totally 23 labeled instances. On average there are approximately 3.38 instances per image, which is higher than the number of OPIXray and HiXray. Figure 4 shows the detail information of Instances per image.

Conclusion

In this paper, we point out that FSOD task is rational and meaningful when built on the real scene where the sample is more difficult to obtain and the context is more complicated, which indicates that the traditional datasets for various visual tasks might not be rational for the FSOD experiments. Thus we choose X-ray security inspection as a typical scenario for FSOD whose context is more complicated and samples are more difficult to obtain comparing to natural scene, and present our Rational Few-shot (RFS) benchmark consisting of 12,333 images with 41,704 annotated instances of 20 different categories to build up a more rational scenario for Few-Shot Object Detection. We hope our work can serve a new perspective to FSOD research community.

References

Chaudhary, A.; Hazra, A.; and Chaudhary, P. 2019. Diagnosis of Chest Diseases in X-Ray images using Deep Convolutional Neural Network. In *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. IEEE.

Chen, T.-I.; Liu, Y.-C.; Su, H.-T.; Chang, Y.-C.; Lin, Y.-H.; Yeh, J.-F.; Chen, W.-C.; and Hsu, W. H. 2021. Dual-Awareness Attention for Few-Shot Object Detection. *arXiv preprint arXiv:2102.12152*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-shot object detection with attention-RPN and multi-relation detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4013–4022.

Guo, S.; Tang, S.; Zhu, J.; Fan, J.; Ai, D.; Song, H.; Liang, P.; and Yang, J. 2019. Improved U-Net for Guidewire Tip Segmentation in X-ray Fluoroscopy Images. In *Proceedings of the 2019 3rd International Conference on Advances in Image Processing*, 55–59.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5356–5364.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8420–8429.

Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; and Bronstein, A. M. 2019. Repmet: Representative-based metric learning for classification and few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5197–5206.

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

Lu, J.; and Tong, K.-y. 2019. Towards to Reasonable Decision Basis in Automatic Bone X-Ray Image Classification: A Weakly-Supervised Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9985–9986.

Tao, R.; Wei, Y.; Jiang, X.; Li, H.; Qin, H.; Wang, J.; Ma, Y.; Zhang, L.; and Liu, X. 2021. Towards Real-world X-ray Security Inspection: A High-Quality Benchmark And Lateral Inhibition Module For Prohibited Items Detection. In *IEEE ICCV*.

Wang, X.; Huang, T. E.; Darrell, T.; Gonzalez, J. E.; and Yu, F. 2020. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*.

Wei, Y.; Tao, R.; Wu, Z.; Ma, Y.; Zhang, L.; and Liu, X. 2020. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM International Conference on Multimedia*, 138–146.

Xiao, Y.; and Marlet, R. 2020. Few-shot object detection and viewpoint estimation for objects in the wild. In *European Conference on Computer Vision*, 192–210. Springer.

Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; and Lin, L. 2019. Meta r-cnn: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9577–9586.