Few-shot X-ray Prohibited Item Detection: A Benchmark and Weak-feature Enhancement Network

Renshuai Tao
State Key Lab of Software Development Environment, Beihang University
iFLYTEK Research
Beijing, China

Tianbo Wang
State Key Lab of Software Development Environment, Beihang University
Beijing, China

Ziyang Wu
iFLYTEK Research
Hefei, China

Cong Liu
iFLYTEK Research
Hefei, China

Aishan Liu
State Key Lab of Software Development Environment, Beihang University
Beijing, China

Xianglong Liu[†]
State Key Lab of Software Development Environment, Beihang University
Beijing, China

ABSTRACT

X-ray prohibited items detection of security inspection plays an important role in protecting public safety. It is a typical few-shot object detection (FSOD) task because some categories of prohibited items are highly scarce due to low-frequency appearance, *e.g.*, pistols, which has been ignored by recent X-ray detection works. In contrast to most FSOD studies that rely on rich feature correlations from natural scenarios, the more practical X-ray security inspection usually faces the dilemma of only weak features learnable due to heavy occlusion, color fading, *etc.*, which causes a severe performance drop when traditional FSOD methods are adopted. However, professional X-ray FSOD evaluation benchmarks and effective models of this scenario have been rarely studied in recent years. Therefore, in this paper, we propose the first X-ray FSOD dataset on the typical industrial X-ray security inspection scenario consisting of 12,333 images and 41,704 instances from 20 categories, which could benchmark and promote FSOD studies on such more challenging scenarios. Further, we propose the **W**eak-feature **E**nhancement **N**etwork (WEN) containing two core modules, *i.e.*, Prototype Perception (PR) and Feature Reconciliation (FR), where PR first generates a prototype library by aggregating and extracting the basis feature from critical regions around instances, to generate the basis information for each category; FR then adaptively adjusts the impact intensity of the corresponding prototype and forces the model to precisely enhance the weak features of specific objects through the basis information. This mechanism is also effective in traditional FSOD tasks. Extensive experiments on X-ray FSOD and Pascal VOC datasets demonstrate that WEN outperforms other baselines in both X-ray and common scenarios. The code and dataset have been released.

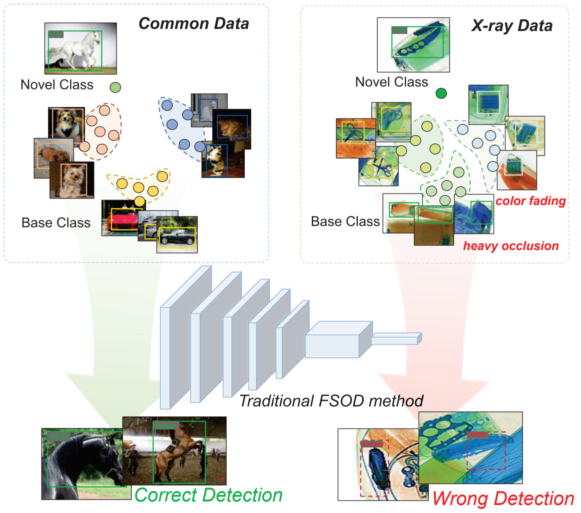

Figure 1: Comparison of few-shot object detection tasks for common data in natural scenario and X-ray data in industrial scenario. Due to heavy occlusion and color fading of X-ray data, the features of novel classes and base classes lack of distinctiveness, causing wrong detection results.

CCS CONCEPTS

- **Computing methodologies** → **Object detection**.

KEYWORDS

few-shot detection, dataset evaluation, X-ray object detection

ACM Reference Format:
Renshuai Tao, Tianbo Wang, Ziyang Wu, Cong Liu, Aishan Liu, and Xianglong Liu[†]. 2022. Few-shot X-ray Prohibited Item Detection: A Benchmark and Weak-feature Enhancement Network. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548075

[†]Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548075

1 INTRODUCTION

With the increasing population density in public transportation hubs, security inspection plays an important role in protecting public security. Inspectors usually adopt X-ray scanners to check the luggage of passengers for judging whether there exist prohibited items, which may cause severe danger to the public. Recent works [9, 17–19, 30, 31, 34, 40] mainly focused on improving the detection performance based on large-scale training samples. However, these works ignored that X-ray security inspection is a typical few-shot detection task since the samples of many prohibited items categories are highly rare due to low-frequency appearance, *e.g.*, pistols.

Recently, researchers have devoted great efforts to solving the few-shot detection problem: generating accurate representations of objects from limited samples available during training [16, 20–22, 24, 25, 28, 42–44]. Most of the existing few-shot learning methods [11, 15, 32, 33, 39, 41] focus on common scenarios and heavily rely on rich feature correlations contained in the natural dataset. However, X-ray few-shot detection of security inspection usually faces the dilemma of very weak features available in training samples due to heavy occlusion, color fading, *etc.*, which causes severe performance drop when traditional FSOD models are adopted. The aforementioned dilemma raises a challenging, meaningful, yet unexplored task: achieving satisfactory performance in extreme few-shot detection scenarios, where learnable features are extremely weak. Currently, relatively little progress has been devoted to this field due to the lack of professional evaluation benchmarks, and simply extending existing natural FSOD datasets is non-trivial owing to the significant scenarios gap. Thus, this task requires researchers to make breakthroughs in both constructing high-quality industrial dataset and designing effective baselines.

To support the study of this important issue, in this paper, we contribute the first industrial FSOD evaluation benchmark, named X-ray FSOD dataset, by selecting the typical scenario, X-ray security inspection. All the samples are generated by X-ray machines, where the texture information of prohibited items is almost eliminated by X-ray. X-ray FSOD dataset consists of 12,333 images, including 41,704 instances with bounding-box annotations of 20 common categories, which are prohibited in aviation security inspection. The number of categories and samples of the X-ray FSOD dataset are consistent with the classical FSOD setting (*e.g.*, Pascal VOC dataset), and all the bounding-boxes are annotated by professional security inspectors. We hope this dataset could serve a comprehensive and reliable evaluation benchmark for models to overcome the performance drop in X-ray FSOD scenario.

Besides, this paper also proposes the **W**eak-features **E**nhancement **N**etwork (WEN) as the baseline, which contains two core modules, *i.e.*,, Prototype Perception (PR) and Feature Reconciliation (FR). The key motivation to overcoming the performance drop caused by extremely weak features is to precisely enhance the weak features by exploiting the basis information. Specifically, in the first step, PR module generates a prototype library by aggregating and extracting the basis feature from critical regions around instances. In the second step, FR module then adaptively adjusts the impact intensity of the corresponding prototype and forces the model precisely to enhance the weak features of specific objects, through exploiting the knowledge stored in the library above.

We summarize the contributions of this study as follows:

- We point out that the significant X-ray security inspection is a typical FSOD task with weak features learnable. We conduct experiments to demonstrate that weak features could cause severe performance drop in FSOD. Further, we propose a challenging but interesting task that achieving satisfactory performance in X-ray FSOD scenario.
- To support the research of this task, we contribute the first practical benchmark, named X-ray FSOD dataset, by gathering and annotating the images generated by X-ray inspection machines. The category distribution follows the standard settings of classical FSOD evaluation benchmark.
- To overcome the performance drop caused by the weak-feature dilemma, we propose the WEN model, aggregating and extracting the basis features from critical regions around instances and precisely enhancing the weak features of specific objects by fusing the basis features extracted.
- We evaluate our method comprehensively on both the X-ray FSOD dataset and Pascal VOC dataset, and the extensive results demonstrate that the WEN model outperforms SOTA methods on both X-ray and common scenarios.

2 RELATED WORK

2.1 X-ray Prohibited Items Detection

X-ray imaging offers powerful ability in many tasks such as medical image analysis [3, 6, 14] and security inspection. As a matter of fact, obtaining X-ray images is difficult, so rare studies touch security inspection in computer vision due to the lack of specialized high-quality datasets. Several recent efforts [1, 2, 13, 17, 30, 35**?**] have been devoted to constructing such datasets. A released benchmark, GDXray [17] contains 19,407 gray-scale images, part of which contain three categories of prohibited items including gun, shuriken and razor blade. SIXray is a large-scale X-ray dataset which is about 100 times larger than the GDXray dataset but the positive samples are less than 1% to mimic a similar testing environment and the labels are annotated for classification. Other relevant works [1, 2, 13] have not made their data available to download. Recently, [35] proposed the first X-ray prohibited items detection dataset, OPIXray dataset, which contains 8,885 X-ray images of 5 categories of cutters and each instance is annotated by professional inspectors. [30] constructed a larger one, the HiXray dataset, which contains 102,928 common prohibited items with bounding-boxes of 8 categories. In addition, [29] proposed the EDS dataset, which consists of 14,219 images including 31,654 common instances from three domains (X-ray machines), with annotations from 10 categories.

2.2 Few-shot Object Detection

Since the few-shot research had been proposed, existing studies towards few-shot classification can be separated into two types of methods, metric-learning-based and meta-learning-based. However, in contrast to classification, few-shot object detection (FSOD) has been rarely studied. Recently, some researchers have made preliminary attempts, such as MetaDet [33], FSDetView [37] *etc.*, trying to attach a meta-learning strategy to existing object detection networks. There are also a few methods based on metric learning

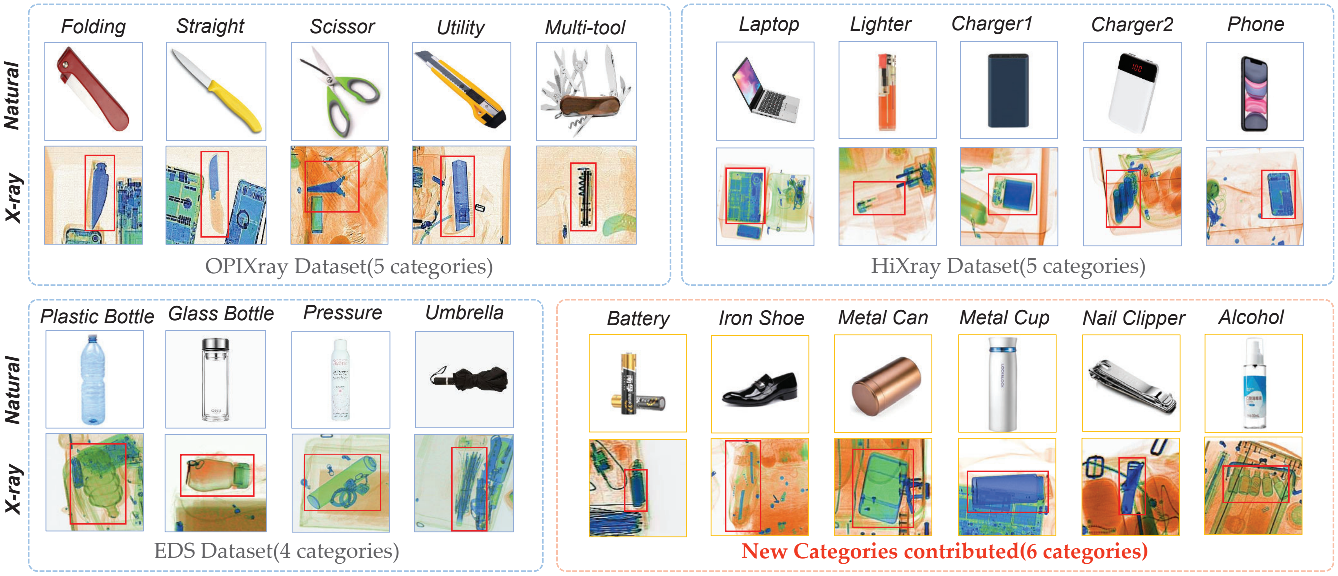

Figure 2: The natural and X-ray examples of all categories in X-ray FSOD dataset. Following the setting (20 classes totally, including 15 base classes and 5 novel classes) of the classical FSOD dataset, Pascal VOC, we select 16 categories from existing public dataset, including 5 from OPIXray, 5 from HiXray and 4 from EDS Dataset. Note that in this paper, we contribute additional 6 categories to construct a standard X-ray FSOD evaluation benchmark.

[11, 41]. At first, finetune-based approaches are considered as baselines with better performance than meta-learning-based approaches. LSTD [4] adopts transfer knowledge and background depression regularization to avoid overfitting. TFA [32] only finetunes the last few layers of the detector and outperforms all the prior meta-learning-based approaches. MPSR [36] develops an auxiliary branch to generate multi-scale positive samples and to refine the prediction at various scales. However, all of these methods are exhausted all the efforts to generate accurate feature representation as the basis information to localize and recognize the objects. Therefore, in real industrial scenario where the basis features are extremely weak, these traditional methods cannot generate accurate representation and fail to achieve satisfactory performance. As a result, exploring how to guide models enhance the basis features and generate accurate representation, to achieve robust and satisfactory performance in real industrial FSOD scenarios, is meaningful.

3 THE X-RAY FSOD DATASET

As illustrated above, a dataset is significant to boost a research and few researchers are involved in this field because there lacks of professional benchmark for evaluating the performance of models. Thus, in this section, we contribute the first weak-feature FSOD evaluation benchmark, X-ray FSOD dataset, by selecting the typical scenario, X-ray security inspection.

3.1 Construction Details

Category Selection. As the famous Pascal VOC dataset demonstrates, a standard FSOD evaluation benchmark usually consists of 20 categories, while 15 are the base classes and other 5 are novel classes during evaluation. Thus, after the X-ray security inspection scenario is selected, we investigate all the public X-ray prohibited items detection dataset, OPIXray [35], HiXray [30] and EDS [29] dataset. Considering the number of instances is no less than 1000, as Figure 2 illustrated, we select 16 categories from the three dataset, including 5 from OPIXray, 5 from HiXray and 4 from EDS Dataset. Additionally, we contribute other 6 categories to construct a standard industrial FSOD evaluation benchmark. The motivation of choosing these 6 categories of objects is due to the fact that they are the most commonly-witnessed prohibited items in airport.

Data Cleaning. During our exploration, we found that some categories in one public dataset also exist in other datasets, but not annotated. For example, the laptop is a category that annotated in HiXray, but it exists in the samples of OPIXray but not annotated. This phenomenon may cause potentially unfair evaluation results. Thus, we conduct the data cleaning progress by annotating all the instances of the total 20 categories, including 14 of the three public dataset and 6 we contributed in this paper.

Annotation Quality Control. All the bounding-boxes are manually annotated. Specifically, there is a preprocess pipeline (following Pascal VOC [5]) to control the image quality as follows, (1) we first hosted a 4-hour training course to teach 5 annotators skills to identify prohibited items from X-ray images accurately; (2) the 5 annotators then labeled our dataset using the “labelme" [1] tool (each image annotation took 2 minutes, and each annotator spent 8 hours per weekday); (3) each image is annotated twice and checked by a third inspector so that the errors are minimized.

3.2 Data Properties

Instances per category. Instances per category refer to “the number of instances of specific class in the dataset". X-ray FSOD dataset contains 12,333 X-ray images, 20 categories of totally 41,704 annotated instances. We separate the images into a training set with 9,867 and a testing set with 2,466 images. The number of instances for each category is shown in Table 1, and the proportion

[1] http://labelme.csail.mit.edu/Release3.0/

Category	FO	ST	SC	UT	MT	LA	LI	CH1	CH2	PH
Training	213	968	1,294	602	1,623	2,381	938	1,729	1,836	4,903
Testing	45	171	240	153	324	543	177	340	384	1,063
Total	258	1,139	1.534	755	1,947	2,924	1,115	2,069	2,220	5,966
Category	PB	GB	PR	UM	**BA**	**IS**	**MC**	**MB**	**NC**	**AL**
Training	579	2,948	910	1,874	**4,248**	**1,337**	**1,372**	**1,865**	**800**	**1,888**
Testing	140	603	216	431	**895**	**302**	**331**	**424**	**200**	**414**
Total	719	3,551	1,126	2,305	**5,143**	**1,639**	**1,703**	**2,289**	**1,000**	**2,302**

Table 1: The statistics of category distribution. The short names in this table correspond to the full names in Figure 2 in order. The names marked in red are the new categories.

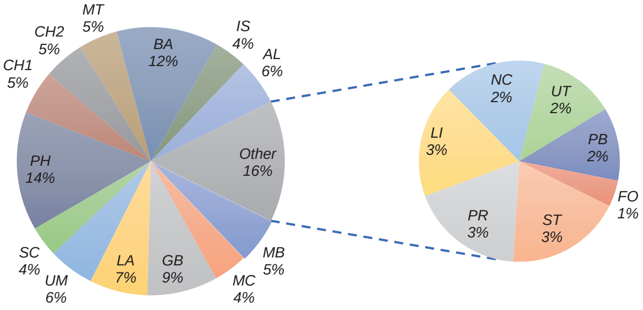

Figure 3: The proportion of category distribution. To make the small percentages more readable, the right pie illustrates the 7 categories with small proportion from the left.

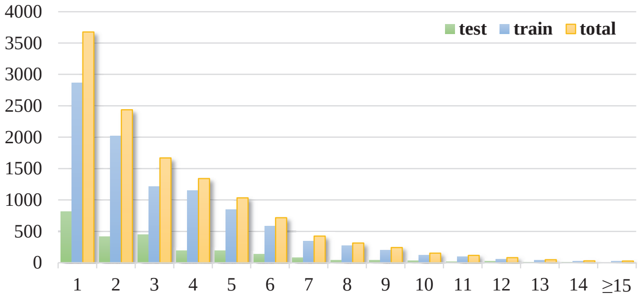

Figure 4: The distribution of numbers of instances per image.

of different categories can be seen in Figure 3. For diversity, we prepare ~30 different objects for each category we contributed.

Instances per image. Instances per image refer to "the number of instances for all classes contained in an image". To simulate the distribution of objects in luggage in real scenario to the greatest extent, the number of instances annotated in each image is not equal. Each image contains at least one instance, up to 23, on average of 3.38. Figure 4 illustrates the distribution of numbers of images containing different numbers of instances.

4 METHOD

In this section, we elaborate on the details of the proposed **W**eak-features **E**nhancement **N**etwork, *i.e.*, WEN model, for weak-feature FSOD scenarios, where usually face the dilemma of heavy occlusion, color fading, *etc.*Previous FSOD methods failed to achieve satisfactory performance due to the basis features are extremely weak. Inspired by the fact that prototype learning can aggregate the basis features, we propose the WEN, containing two core modules, *i.e.*, Prototype Perception (PR) and Feature Reconciliation (FR), which generates a prototype library by aggregating and extracting the basis feature from critical regions around instances, and adaptively adjust the impact intensity of the corresponding prototype and forces the model to precisely enhance the weak features of specific objects, receptively.

We will start with the problem analysis of the industrial few-shot object detection setting. Then we will introduce the architecture of our WEN in Section 4.2, including two core modules, Prototype Perception (PR) and Feature Reconciliation (FR) in Section 4.2.1 and Section 4.2.2, respectively. Finally, in 4.3, we illustrate the training process of the network in detail.

4.1 Problem Analysis

FSOD methods based on fine-tuning mainly include a backbone network $\mathbb{E}$ for extracting representative features for an input image, a Region Proposal Network (RPN) $\mathbb{R}$ for producing and selecting proposals, and a detector denoted as $\mathbb{D}$. These methods exploits the abundant characteristics of base classes to guide the network extract accurate representations, which are the basis features $\mathbf{f}^{(b)}$. The ability learned to represent basis features from base classes can be symbolized as $\theta^{(\mathbb{E},\mathbb{R},\mathbb{D})}_{\mathcal{L}_{(b)}}$. In the process of training novel classes, the ability learned of representing novel classes by leveraging knowledge from base classes can be formulated as follows:

$$\theta^{(\mathbb{R},\mathbb{D})}_{(n)} = \theta^{(\mathbb{R},\mathbb{D})}_{\mathcal{L}_{(b)}} + \nabla\mathcal{L}_{(n)},$$

where $\theta^{(\mathbb{R},\mathbb{D})}_{\mathcal{L}_{(b)}}$ refers to that the parameters of the backbone network $\mathbb{E}$ are frozen after the base samples trained. $\nabla\mathcal{L}_{(n)}$ refers to distinctive knowledge of the novel classes.

In FSOD task, the basis features generated for the novel and base classes, $\mathbf{f}^{(n)}$ and $\mathbf{f}^{(b)}$, are both extracted by the backbone $\theta^{(\mathbb{E})}_{\mathcal{L}_{(b)}}$. Note that in X-ray scenario with weak feature learnable, due to heavy occlusion, color fading, *etc.*$\mathbf{f}^{(n)}$ outputted by $\theta^{(\mathbb{E})}_{\mathcal{L}_{(b)}}$ and $\mathbf{f}^{(b)}$ cannot be discriminated directly. This dilemma confuses the detector to make a distinction between novel and base classes, resulting in the performance drop. Therefore, the key to handle this challenge is to enhance the basis features of novel classes, making them having enough distinctiveness to base classes.

4.2 The Architecture of WEN

In figure 5, the model we proposed is implemented by adding two core and efficient modules, *i.e.*, Prototype Perception (PR) and Feature Reconciliation (FR), in the base detection framework. We select the base framework including a backbone network $\mathbb{E}$ for extracting representative features for an input image, a Region Proposal Network (RPN) $\mathbb{R}$ for producing and selecting proposals, and a detector denoted as $\mathbb{D}$.

Specifically, PR module first generates a prototype library by aggregating and extracting the basis feature from critical regions around instances. FR module then adaptively adjusts the impact intensity of the corresponding prototype and forces the model to precisely enhance the weak features of specific objects.

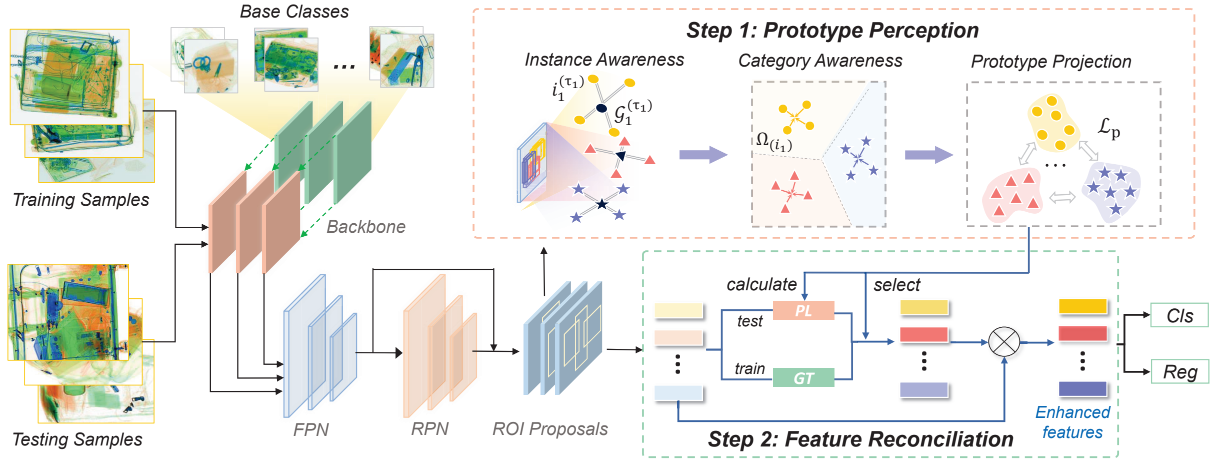

Figure 5: The network structure of the Weak-features Enhancement Network (WEN). WEN contains two core modules, *i.e.*, Prototype Perception (PR) and Feature Reconciliation (FR). In step 1, training samples are exploited to generate the prototype library by aggregating critical regions around instances. In step 2, the weak features of input instances are prewisely enhanced by adaptively adjusting the impact intensity of the corresponding prototype from the library generated in step 1.

4.2.1 Prototype Perception. To aggregate the basis information to enhance the weak features, we try to discover a proper dimension to distinct novel classes from base classes. We guide the network $\mathbb{R}$ project these features, $\mathbf{f}^{(\mathrm{b})}$ and $\mathbf{f}^{(\mathrm{n})}$, into a suitable feature space. Inspired by the fact that prototype learning can aggregate discriminative representations that are robust against variation, we exploits the graph-based prototype aggregation method to extract the new basis features. By optimizing the loss of these prototypes, we make the category prototype vector tend to be orthogonal, so as to highly distinguish all categories.

First, given a batch of images $\mathbf{X}$, for each object inside the images, we construct a relation graph set $\mathbf{G}^{(i)} = \{\mathrm{G}^{(i_1)}, \mathrm{G}^{(i_2)}, \ldots, \mathrm{G}^{(i_n)}\}$, by structuring the proposals generated around each object i generated by the network $\mathbb{R}$. For a specific graph $\mathrm{G}^{(i)} = \{\mathrm{V}^{(i)}, \mathrm{E}^{(i)}\}$, where $\mathrm{V}^{(i)} = \{i^{(\tau_1)}, i^{(\tau_2)}, \ldots, i^{(\tau_n)}\}$ is the proposal set of the object i and $\mathrm{E}^{(i)}$ is the edge set of each proposal and its corresponding ground-truth in $\mathrm{V}^{(i)}$. In the first step, we filter out unqualified proposals and aggregate the rest proposals of the object i to generate the most accurate feature map, *i.e.*,, the object prototype $\Omega^{(i)}$, that we consider to represent the object i. This process is formulated as:

$$\Omega^{(i)} = \left(\sum_{n=1}^{N_{(\tau)}} \mathbf{IoU}(i^{(\tau_n)}, \mathcal{G}_{(i)}) \cdot i^{(\tau_n)} \right) / \sum_{n=1}^{N_{(\tau)}} \mathbf{IoU}(i^{(\tau_n)}, \mathcal{G}_{(i)}) \tag{1}$$

where $i^{(\tau_n)}$ refers to the n-th proposal around of the object i, $\mathcal{G}_{(i)}$ refers to the ground-truth of the object i. $N_{(\tau)}$ refers to the number of proposals around the object i and $\Omega^{(i)}$ refers to the prototype of object i. Inspired by [38], we choose IoU of the proposal $i^{(\tau_n)}$ and the the ground-truth $\mathcal{G}_{(i)}$ as the measured metric. Note that we exploit the same operation to get category probability prototype $P^{(i)}$ for each object i.

In the second step, for each class k of the input image $\mathbf{x}$, similarly, we construct a relation graph set. This step generates the category prototype by the similar operation as the first step. After the two steps operation, the graph generates a prototype library $\Omega^{(k)} = \{\Omega^{(k_1)}, \Omega^{(k_2)}, \ldots, \Omega^{(k_n)}\}$. Each element in $\Omega^{(k)}$ represents the prototype of one category of both base classes and novel classes. In each epoch of training, the prototypes generated from the input image updates the prototype library $\Omega^{(k)}$. Thus, at the end of each round of training, the process of updating the library can be formulated as follows:

$$\Omega_{(l)}^{(k)} = \begin{cases} \Omega_{(l)}^{(k)}, & l = 1 \\ \alpha \cdot \Omega_{(l)}^{(k)} + (1-\alpha) \cdot \Omega_{(l-1)}^{(k)}, & l > 1 \end{cases} \tag{2}$$

where $\alpha \in [0, 1]$ refers to a super parameter. $\Omega_{(l)}^{(k)}$ refers to the category prototype of k in the l-th training.

Finally, to achieving the goal that discovering a proper dimension to distinct novel classes from base classes, we try to guide the network $\mathbb{R}$ to project these features into a suitable feature space by minimizing the cosine distance of each two categories. By minimizing the cosine distance, different categories of prototype vectors can gradually be orthogonal, so as to search the proper feature dimension that can fully separate all of these categories. The process of projection can be formulated as follow:

$$\mathcal{L}_p = \frac{\sum_{i=1}^{N_k} \sum_{j=1, j \neq i}^{N_k} \phi\left(\Omega^{(i)}, \Omega^{(j)}\right)}{N_k (N_k - 1)} \tag{3}$$

where $\mathcal{L}_p$ refers to the average distance of each two categories in the whole categories, which is the loss of the PR module. $\phi\,(\dot{})$ refers to the distance calculated by the consine function and N_k refers to the number of elements of the $\mathbf{\Omega}$. Thus, by minimizing $\mathcal{L}_p$, elements in $\mathbf{\Omega}^{(k)}$ tend to be orthogonal gradually. After optimized, the final prototype library is generated as $\mathbf{\Omega} = \{\Omega^{(1)}, \cdots, \Omega^{(K)}\}$.

4.2.2 Feature Reconciliation. Inspired by the fact that the amount of information lost for different categories of objects under X-ray is quite different, we consider that different categories of objects have different requirements to obtain distinguishable features from prototypes. Therefore, in order to avoid the over-fitting problem caused by excessive basis features, it is necessary to design a fusion method that absorbing different amounts of the basis features according to the category to which the object belongs. In this module, the first step is to determine which prototype feature of the prototype library $\mathbf{\Omega}$ should be absorbed for a proposal feature $\mathbf{f}_{(\mathbb{R})}$ generated by the network $\mathbb{R}$. Here, we adopt the method that comparing $\mathbf{f}_{(\mathbb{R})}$ with each element $\Omega^{(i)} \in \mathbf{\Omega}$ and selecting the one $\Omega^{(k)}$ whose distance is closest.

Regrading the fusion methods, there are mainly two types, linear and non-linear. We have tried these two methods respectively. The linear method can be formulated as follows:

$$\mathbf{f}_{(r)} = \alpha \cdot \Omega^{(k)} + \mathbf{f}_{(\mathbb{R})} \tag{4}$$

where $\mathbf{f}_{(r)}$ refers to the feature map outputted by the FR module, which is the final feature map extracted from the input image $\mathbf{x}$ and fed into the detector $\mathbb{D}$. $\mathbf{f}_{(\mathbb{R})}$ is the feature map outputted by the region proposal network $\mathbb{R}$, which represents initial object feature. α is a linear parameter, which is artificially set.

In view of the complexity of various feature under X-ray, linear methods are unable to achieve satisfactory performance. Thus, we exploit the convolutional network with activation function layer to adaptively assign the impact intensity of the corresponding prototype and forces the model to precisely enhance the weak features of specific objects. Specifically, we adopt a 1*1 convolutional network $\mathbb{F}$, to non-linearly absorb the information inside the prototypes generated by the PR module. For a category k, the absorb information $ReLU(Conv_{\mathbf{w},\mathbf{b}}(\Omega^{(k)}))$ represents the amount of basis feature that the model requires from the prototype when this category detection achieves the best performance. This nonlinear projection process can be formulated as follows:

$$\mathbf{f}_{(r)} = ReLU(Conv_{\mathbf{w},\mathbf{b}}(\Omega^{(k)})) + \mathbf{f}_{(\mathbb{R})} \tag{5}$$

where $\mathbf{f}_{(r)}$ refers to the feature map outputted by the FR module, which is fed into the detector $\mathbb{D}$. $\mathbf{w}, \mathbf{b}$ refers to the parameters of the 1*1 convolutional network. $ReLU$ refes to the activation function layer. $\mathbf{f}_{(\mathbb{R})}$ is the feature map outputted by the region proposal network $\mathbb{R}$, which represents input instance feature.

4.3 Network Training

Our WEN adopts the classical two-stage fine-tuning approach[32] for few-shot detection as basic training strategy, which is the most widely adopted in previous studies. The two-stage fine-tuning approach usually consists of the base-training stage and the fine-tuning stage. In this section, we elaborate the detailed loss function of the both two stages.

In the base-training stage, the WEN is trained only on the base classes with the standard loss function of Faster R-CNN[26]. This process can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{loc}, \tag{6}$$

where$\mathcal{L}_{rpn}$ refers to a binary cross-entropy loss for RPN to distinguish foreground from backgrounds and refine the anchors,$\mathcal{L}_{cls}$ refers to a cross-entropy for box classifier and $\mathcal{L}_{loc}$ is a smoothed L_1 loss for box regressor.

In the fine-tuning stage, the prototype library in PR module will be generated and updated after each training iteration, to extract the basis information for each category. FR module is applied to assign the impact intensity of the corresponding prototype and forces the model to precisely enhance the weak features of specific objects through the basis information. Different from TFA[32], we only freeze the parameters of backbone network to preserve the basic ability of extracting the preliminary representations. The total loss in fine-tuning stage can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{rpn} + \mathcal{L}_{cls} + \mathcal{L}_{loc} + \lambda \mathcal{L}_{p} \tag{7}$$

where $\mathcal{L}_{p}$ refers to the loss of the PR module and λ refers to a super parameter to balance the impact intensity of $\mathcal{L}_{p}$. Specifically, the fine-tuning stage of the whole network training procedure can be viewed as Algorithm 1.

Algorithm 1 Fine-tuning Stage of WEN's training procedure

```
Input: Images with K-shot samples, the number of categories $k$.
Output: The total loss value $\mathcal{L}_{total}$
  Generate the feature map set $\mathbf{\Phi}$.
  Generate the proposal set $\mathrm{V}^{(i)}$ with $m$ proposals.
  Calculate the loss value $\mathcal{L}_{rpn}$.
  for all $a = 1, 2, \ldots, \tau_n$ do
    Calculate each instance prototype $\Omega^{(a)}$.
    Calculate each probability prototype $P^{(a)}$.
  end for
  for all $b = 1, 2, \ldots, k_n$ do
    Calculate each category prototype $\Omega^{(b)}$.
    Update the prototype library $\mathrm{L}_{(b)}$.
  end for
  Calculate the loss value $\mathcal{L}_{p}$.
  Generate the enhanced feature through Fussing operation.
  Calculate the loss value $\mathcal{L}_{reg}$ and $\mathcal{L}_{cls}$.
  Calculate the total loss value $\mathcal{L}_{total}$
```

5 EXPERIMENT

In this section, we will introduce the extensive experiments to verify the effectiveness of our proposed methods.

5.1 Experimental Settings

5.1.1 Experimental Scenarios. **Firstly**, to demonstrate the fact that the weakening of features will lead to the decline of detection performance in FSOD task, we simulate this weakening of features by exploiting the edge detection method to generate the edge images from the classical FSOD dataset, Pascal VOC, whose samples are gathered from natural scenario. We conduct experiments on both natural and outline datasets to observe the performance drop. **Secondly**, to compare with SOTA methods comprehensively, we conduct the extensive experiments on both weak-feature scenario and common natural scenario. In weak-feature scenario, we conduct the experiments on both the X-ray FSOD dataset and the outline dataset. In common natural scenario, we conduct the experiments on the famous Pascal VOC dataset. **Finally**, in Experiment 4, ablation studies are conducted.

5.1.2 Benchmark Settings. In Section 5.2 and Section 5.3 (the third dataset), as for the Pascal VOC dataset, we follow the classical data partition[10, 23, 27, 32, 37] with three random split groups and each of the split randomly divide all 20 categories into 15 base classes and 5 novel classes. Sufficient samples are available for each base class during base-training, while only K = 1, 2, 3 ,5, 10 objects sampled from the combination of the trainval sets of VOC2007 and

Method	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
FRCN+ft+full[32]	1.9	8.3	8.9	10.6	18.3	1.6	8.1	9.5	15.3	22.3	12.1	16.5	18.9	20.5	27.2
TFA (w/fc)[32]	12.1	16.8	22.3	30.7	36	14.7	20.6	21.4	27.7	35.3	15.6	21.3	22.4	30.6	38.2
TFA (w/cos)[32]	18.4	20.0	22.0	27.5	34.4	13.7	17.5	18.4	26.4	33.5	17.6	21.4	22.3	29.7	37.2
DeFRCN[23]	20.2	23.2	32.8	36.3	41.6	13.7	23.5	25.7	30.8	**39.4**	18.7	29.4	32.3	36.5	48.8
FSCE[27]	23.7	27.8	32.7	37.4	42.1	12	22.3	23	29.4	37.5	16.9	27.6	29.9	36.0	49.7
DCNet[8]	22.4	24.4	29.1	33.6	39.5	14.9	22.1	24.8	29.3	39.2	19.3	29.1	30.9	36.8	44.8
WEN (Ours)	**30.5**	**32.2**	**38.9**	**43.9**	**42.3**	**17.2**	**27.7**	**28.2**	**31.2**	39.0	**22.2**	**32.4**	**35.2**	**41.6**	**50.8**
TFA (w/cos)*[32]	11.4	17.4	20.0	27.0	34.3	10.4	16.9	21.1	28.1	34.6	13.6	23.0	25.6	33.1	38.9
FSCE*[27]	17.6	25.8	29.9	37.6	41.7	10.6	21.1	27.3	31.5	39.9	17.9	27.8	30.9	38.7	47.5
WEN (Ours)*	**26.0**	**30.1**	**35.6**	**40.0**	**42.1**	**18.3**	**25.9**	**31.5**	**34.3**	**41.2**	**20.5**	**32.0**	**34.7**	**42.6**	**47.8**

Table 2: The mAP50 results for novel classes of various few-shot detection methods on the X-ray FSOD dataset. All of the results are averaged over multiple times of evaluations. * denotes the average results over multiple random seeds.

VOC2012 versions can be utilized for each novel class. In addition, the test set of VOC2007 is used for evaluation. In Section 5.3 (the second dataset), we stick to the same data partition with Pascal VOC. In Section 5.3 (the first dataset) and 5.4, consistent with the Pascal VOC, we separate all 20 categories of X-ray FSOD dataset into two parts randomly, where 5 categories are chosen as novel classes with K=1, 2, 3, 5, 10 shot training samples, and the left 15 are base classes. We evaluate methods on three different split groups, *i.e.*, {“charger 1”, “utility”, “phone”, “metal bottle”, “plastic bottle”}, {“charger 1”, “multi-tool”,“metal bottle”,“pressure”,“alcohol”}, and {“laptop”, “multi-tool”,“glass bottle”,“metal bottle”,“nail clippers”}.

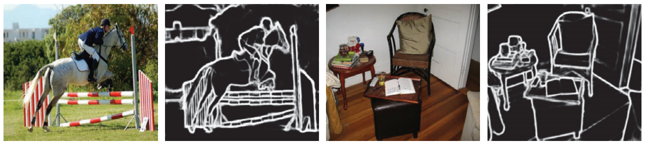

Figure 6: Two examples of images in VOC-outline dataset.

5.1.3 Implement Details. We employ the famous two-stage detection model Faster R-CNN[26] with FPN[12] and ResNet-101[7] as the basic network. The parameters of the network are optimized by a standard SGD with momentum 0.9 and weight decay $1e^{-4}$. The learning rate is set to 0.02 during the base-training stage and 0.01 during the fine-tuning stage. The IoU threshold when filtering low-quality proposals in PR module is set to 0.7 and the α for updating prototype library is set to 0.2. The initial weight for $\mathbf{w}$ in Formula 6 is set to 0.1 and the λ of $\mathcal{L}_p$ in Formula 8 is set to 0.5.

5.2 Effect of Weakening of Features

To verify that the weakening of features will lead to the decline of detection performance in FSOD task, we conduct experiments on both natural and edge samples dataset generated (named VOC-outline dataset) to observe the performance drop. We first make a performance comparison on original Pascal VOC and VOC-outline by adopting three advanced FSOD methods and the results are shown in Table 3. In Table 3, in all settings of instance shot, the performance drops of the three methods are obvious. For example, in 1 shot setting, the performance drops reach about 24.5%, 50.5% and 51.5%, respectively. Thus, it is obvious that there exists a huge performance drop caused by the weakening of features.

Method	Dataset	Novel Set 1				
		1	2	3	5	10
TFA (w/cos)[32]	VOC	**25.3**	**36.4**	**42.1**	**47.9**	**52.8**
	outline	19.1	20.7	28.7	35.2	43.5
FSCE[27]	VOC$^{\#}$	**37.8**	**42.6**	**49.7**	**60.1**	**60.8**
	outline	18.7	23.2	31.2	43.2	46.5
DeFRCN[23]	VOC	**40.2**	**53.6**	**58.2**	**63.6**	**66.5**
	outline	19.5	23.8	32.8	41.5	46.5

Table 3: Performance drop caused by feature weakening. $^{\#}$ denotes the result of the model in our implementation.

5.3 Comparison with SOTA methods

To evaluate the effectiveness of our method comprehensively, we conduct the extensive experiments on both weak-feature scenario and common natural scenario. In weak-feature scenario, we conduct the experiments on both the X-ray FSOD dataset and the VOC-outline dataset mentioned in Section 5.2. In common natural scenario, we conduct the experiments on the famous Pascal VOC.

5.3.1 X-ray FSOD dataset. We present the mAP50 results of the novel classes on OFSD with three different data splits, which is shown in Table 2. Table 2 demonstrates that WEN outperforms other SOTA FSOD methods obviously. On Novel Set 1, WEN outperforms the second best result by **4.8%** on average in all shot settings and especially **6.8%** in 1 shot setting. Our WEN especially achieves a remarkable performance increase of **16.9%** in maximum on 3 shot and **13.2%** in average over our baseline TFA (w/cos). Moreover, the fact that the performances are achieved for multiple random seed settings demonstrates the robustness of our method. Therefore, our WEN model is capable to alleviate the performance drop of weak features, caused by heavy occlusion, color fading in real industrial scenario.

Another interesting phenomenon is that the performance improvement of WEN on the setting of fewer shot is larger than the setting of more shots. For example, in Novel Set 1, the performance improvement on the setting of 1 shot reaches **6.8%** while **0.2%** on the setting of 10 shot. It is mainly because due to the scarcity of fine-tuning samples, the demand for basis information on the setting of fewer shots is more urgent than the setting of more shots.

5.3.2 VOC-outline dataset. As mentioned in Experiment 1 above, the VOC-outline dataset we generated is a simulation of the feature weakening, compared to the natural Pascal VOC. Therefore, to evaluate the effectiveness of our WEN model in weak-feature scenario, we conduct the experiments on the VOC-outline dataset, following the same setting of the experiments conducted on the

Method	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
TFA (w/cos)[32]	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
TFA (w/cos)+WEN	**35.6**	**38.7**	**45.8**	**52.9**	**59.7**	**20.1**	**29.1**	**32.1**	**37.3**	**45.2**	**26.2**	**34.8**	**38.0**	**49.4**	**53.7**
DeFRCN[23]	40.2	53.6	58.2	63.6	66.5	29.5	39.7	43.4	48.1	**52.8**	35.0	38.3	52.9	57.7	60.8
DeFRCN+WEN	**44.3**	**54.8**	**60.9**	**64.9**	**66.8**	**30.4**	**41.3**	**44.7**	**48.6**	52.3	**39.5**	**45.3**	**54.3**	**58.9**	**61.3**
FSCE$^{\#}$[27]	37.8	**42.6**	49.7	60.1	60.8	20.1	24.5	40.7	43	**48.2**	33	40.5	45.8	53.9	57.6
FSCE+WEN	**39.3**	42.5	**50.5**	**60.9**	**61.6**	**22.4**	**26.1**	**41.0**	**43.9**	48.1	**34.6**	**42.3**	**46.5**	**54.2**	**58.5**

Table 4: Comparisons of traditional few-shot detection approaches and module-inserted approaches on three different data splits of VOC benchmark. $^{\#}$ denotes the results of the approach are recorded under our experimental settings.

X-ray FSOD dataset. The results are illustrated in Table 5. From Table 5, our method achieves a stable performance improvement over various FSOD baselines. The maximum performance improvement is achieved in the setting of 3 shots.

Method	Novel Set 1				
	1	2	3	5	10
FRCN+ft[32]	11.5	12	20.2	30.2	39.6
TFA (w/fc)[32]	16.7	20.8	26.5	33.3	43.3
TFA (w/cos)[32]	19.1	20.7	28.7	35.2	43.5
DeFRCN[23]	19.5	23.8	32.8	41.5	46.5
FSCE[27]	18.7	23.2	31.2	**43.2**	46.5
WEN (Ours)	**19.9**	**24.3**	**33.1**	42.5	**47.4**

Table 5: The results of various models on the VOC-outline.

5.3.3 *Pascal VOC dataset.* On the two datasets above, we have evaluated the ability of our WEN model and other FSOD methods of handling the weak-feature dilemma and demonstrated the effectiveness of our WEN model. In this section, we conduct more interesting experiments to explore the performance of our WEN model in common natural scenarios, *i.e.*, the instance features are strong due to bright color and enough texture information, *etc.*Thus, we conduct the experiments on the classical FSOD datset, Pascal VOC, whose samples are gathered from natural scenario. Different from the Experiment 2, in the case of strong features in Experiment 3, our feature-enhanced WEN model plays an auxiliary role in detection. Therefore, we integrate our model on three different FSOD methods, *i.e.*, TFA (w/cos), DeFRCN and FSCE, and compare these integrated models to their corresponding base models, instead of directly comparing with these FSOD methods. The experimental results are illustrated in Table 4.

From Table 4, we can draw the observation that there exist various extents of improvement performance in most setting of shots and splits on the integrated models, compared with the corresponding baselines. In particular, WEN-integrated TFA (w/cos) outperforms the baseline TFA (w/cos) by **5.6%**, **2.7%** and **7.3%** averaged on three different split sets of the Pascal VOC dataset, respectively. Thus, the feature-enhanced mechanism of our WEN model not only outperforms other FSOD methods in weak-feature scenario, but also the WEN-integrated model plays an auxiliary role in strong-feature scenario, achieving impressive performance improvement in both industrial and common scenarios.

5.4 Ablation Studies

In this section, we conduct several ablation studies to further investigate our method on the X-ray FSOD dataset, to verify the effectiveness of each module and analysis more details.

Firstly, to evaluate the effectiveness of PR module, we separate the prototype mechanism and the $\mathcal{L}_p$ of Formula 4 (guide each category prototype tend to be orthogonal to each other). Secondly, to evaluate the effectiveness of FR module, we try two different fusion methods, linear and non-linear type (described in section 4.2.2). All combinations and results are shown in Table 6.

From Table 6, we can observe that the the prototype mechanism helps improve the performance by "**10.3%**" on average and the $\mathcal{L}_p$ promotes the detection precision by **7.4%** on average in three different settings of shots. Moreover, after adopting the $\mathcal{L}_p$ to the whole Network (including FW), there has another **0.7%** increase. Additionally, the non-linear fusion method we adopt in FW module further enhances the performance by **1.3%** on average, compared to the linear fusion method, and achieves a remarkable increase of **12.3%** comparing with the pure base model.

Due to length limitation, the ablation studies of the super parameters mentioned above are illustrated in supplemental materials.

PR module		FR module		Novel Set 1		
Prototypes	$\mathcal{L}_p$	Linear	Nonlinear	1	3	10
				18.4	22	34.4
✓		✓		27.9	36.9	40.9
✓	✓			23.3	35.6	38
✓	✓	✓		28.4	37.5	41.8
✓	✓		✓	**30.5**	**38.9**	**42.2**

Table 6: The results of several groups of ablation studies.

6 CONCLUSION

In this paper, we first point out the significant X-ray security inspection is a typical FSOD task, which usually faces the dilemma with only weak features due to heavy occlusion, color fading, *etc.*, which causes a severe performance drop for traditional FSOD methods. To support this vital study, we contribute the first X-ray FSOD dataset by gathering and annotating the images generated by X-ray inspection machines. Further, we propose the WEN model, aggregating and extracting the basis features from critical regions around instances and precisely enhancing the weak features of specific objects by fusing the basis features extracted. We evaluate our method comprehensively on both the X-ray FSOD dataset and Pascal VOC dataset, and the extensive results demonstrate that the WEN model outperforms SOTA methods on both X-ray and common scenarios. We hope our work could provide a new view to the FSOD community.

7 ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 62022009, the Beijing Nova Program of Science and Technology under Grant Z191100001119050 and the Research Foundation of iFLYTEK, P.R. China.

REFERENCES

[1] Samet Akcay and Toby P Breckon. 2017. An evaluation of region based object detection strategies within x-ray baggage security imagery. In *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1337–1341.

[2] Samet Akcay, Mikolaj E Kundegorski, Chris G Willcocks, and Toby P Breckon. 2018. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE transactions on information forensics and security* 13, 9 (2018), 2203–2215.

[3] Arjun Chaudhary, Abhishek Hazra, and Prakash Chaudhary. 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)*. IEEE, 1–6.

[4] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. 2018. Lstd: A low-shot transfer detector for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[6] Shuai Guo, Songyuan Tang, Jianjun Zhu, Jingfan Fan, Danni Ai, Hong Song, Ping Liang, and Jian Yang. 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.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. 2021. Dense relation distillation with context-aware aggregation for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10185–10194.

[9] Yao Jin, Guang Yang, Ying Fang, Ruipeng Li, Xiaomei Xu, Yongkai Liu, and Xiaobo Lai. 2021. 3D PBV-Net: an automated prostate MRI data segmentation method. *Computers in Biology and Medicine* 128 (2021), 104160.

[10] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. 2019. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8420–8429.

[11] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M Bronstein. 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.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[13] Jinyi Liu, Xiaxu Leng, and Ying Liu. 2019. Deep Convolutional Neural Network Based Object Detector for X-Ray Baggage Security Imagery. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 1757–1761.

[14] Jianjie Lu and Kai-yu Tong. 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*, Vol. 33. 9985–9986.

[15] Yuqing Ma, Wei Liu, Shihao Bai, Qingyu Zhang, Aishan Liu, Weimin Chen, and Xianglong Liu. 2020. Few-shot Visual Learning with Contextual Memory and Fine-grained Calibration. In *IJCAI*.

[16] Xiangxin Meng, Xu Wang, Hongyu Zhang, Hailong Sun, and Xudong Liu. 2022. Improving Fault Localization and Program Repair with Deep Semantic Features and Transferred Knowledge. In *ICSE*.

[17] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. 2015. GDXray: The database of X-ray images for nondestructive testing. *Journal of Nondestructive Evaluation* 34, 4 (2015), 42.

[18] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. 2019. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2119–2128.

[19] Ahmed Naglah, Fahmi Khalifa, Reem Khaled, Ayman El-Baz, et al. 2021. Thyroid cancer computer-aided diagnosis system using MRI-based multi-input CNN model. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 1691–1694.

[20] Yongri Piao, Zhengkun Rong, Miao Zhang, and Huchuan Lu. 2020. Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11865–11873.

[21] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. 2020. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *CVPR*.

[22] Binhang Qi, Hailong Sun, Wei Yuan, Hongyu Zhang, and Xiangxin Meng. 2021. DreamLoc: A Deep Relevance Matching-Based Framework for bug Localization. *IEEE Transactions on Reliability* (2021).

[23] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. 2021. DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8681–8690.

[24] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. 2020. Binary neural networks: A survey. *Pattern Recognition* 105 (2020), 107281.

[25] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. 2020. Forward and backward information retention for accurate binary neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2250–2259.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[27] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. 2021. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7352–7362.

[28] Mingxing Tan, Ruoming Pang, and Quoc V Le. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10781–10790.

[29] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Hongping Zhi Bowei Jin and, Xianglong Liu, and Aishan Liu. 2022. Exploring Endogenous Shift for Cross-domain Detection: A Large-scale Benchmark and Perturbation Suppression Network. In *IEEE CVPR*.

[30] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu*. 2021. Towards Real-world X-ray Security Inspection: A High-quality Benchmark and Lateral Inhibition Module for Prohibited Items Detection. In *IEEE ICCV*.

[31] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. 2021. Towards Real-World Prohibited Item Detection: A Large-Scale X-ray Benchmark. *arXiv preprint arXiv:2108.07020* (2021).

[32] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. 2020. Frustratingly simple few-shot object detection. *Proceedings of the 37th International Conference on Machine Learning* (2020).

[33] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2019. Meta-learning to detect rare objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9925–9934.

[34] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. 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.

[35] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. 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.

[36] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. 2020. Multi-scale positive sample refinement for few-shot object detection. In *European Conference on Computer Vision*. Springer, 456–472.

[37] Yang Xiao and Renaud Marlet. 2020. Few-shot object detection and viewpoint estimation for objects in the wild. In *European Conference on Computer Vision*. Springer, 192–210.

[38] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. 2020. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12355–12364.

[39] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. 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.

[40] Yan Yang, Na Wang, Heran Yang, Jian Sun, and Zongben Xu. 2020. Model-Driven Deep Attention Network for Ultra-fast Compressive Sensing MRI Guided by Cross-contrast MR Image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 188–198.

[41] Yukuan Yang, Fangyun Wei, Miaojing Shi, and Guoqi Li. 2020. Restoring negative information in few-shot object detection. *Annual Conference on Neural Information Processing Systems* (2020).

[42] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. 2021. Center-based 3d object detection and tracking. In *CVPR*.

[43] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. 2021. Diversifying sample generation for accurate data-free quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15658–15667.

[44] Zhijie Zhang, Yan Liu, Junjie Chen, Li Niu, and Liqing Zhang. 2021. Depth Privileged Object Detection in Indoor Scenes via Deformation Hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3456–3464.