I^{2}OL-Net: Intra-Inter Objectness Learning Network for Point-Supervised X-Ray Prohibited Item Detection

Chenyang Wang, Yan Yan, *Senior Member, IEEE*, Jing-Hao Xue, *Senior Member, IEEE*,
Hanzi Wang, *Senior Member, IEEE*

***Abstract*—Automatic detection of prohibited items in X-ray images plays a crucial role in public security. However, existing methods rely heavily on labor-intensive box annotations. To address this, we investigate X-ray prohibited item detection under labor-efficient point supervision and develop an intra-inter objectness learning network (I^{2}OL-Net). I^{2}OL-Net consists of two key modules: an intra-modality objectness learning (intra-OL) module and an inter-modality objectness learning (inter-OL) module. The intra-OL module designs a local focus Gaussian masking block and a global random Gaussian masking block to collaboratively learn the objectness in X-ray images. Meanwhile, the inter-OL module introduces the wavelet decomposition-based adversarial learning block and the objectness block, effectively reducing the modality discrepancy between natural images and X-ray images and transferring the objectness knowledge learned from natural images with box annotations to X-ray images. Based on the above, I^{2}OL-Net greatly alleviates the severe problem of part domination caused by large intra-class variations in X-ray images. Experimental results on four X-ray datasets show that I^{2}OL-Net can achieve superior performance with a significant reduction of annotation cost, thus enhancing its accessibility and practicality. The source code is released at https://github.com/houjoeng/I2OL-Net.**

***Index Terms*—X-ray prohibited item detection, point-supervised learning, objectness knowledge transfer.**

I. Introduction

OVER the past few decades, automatic X-ray prohibited item detection, which can greatly facilitate security inspectors to identify prohibited items (such as knives and guns), has attracted considerable attention. This is mainly due to its crucial role in public security. Accordingly, a variety of X-ray prohibited item detection methods [3]–[7] have been proposed and made significant progress.

Existing prohibited item detection methods usually train models with box annotations involving both bounding boxes and categories of prohibited items. However, fully annotating X-ray images is very time-consuming and labor-intensive since the annotations usually require the expertise of professionals. Hence, semi-supervised learning [8]–[10] and weakly-supervised learning methods [11]–[14] have been developed to reduce the annotation cost. In this paper, we study X-ray prohibited item detection under the point-supervised setting, where only point annotations are given.

Point annotations can only provide limited position information of prohibited items and lack the scale information. To address this, existing methods [15]–[17] often leverage class activation maps or instance classifier refinement to estimate the scale information. However, due to the distinct penetration capabilities of different materials in X-ray, the items in X-ray images often exhibit large intra-class variations and differences. As a result, when applied to X-ray images, existing weakly-/point-supervised methods cannot capture the objectness of prohibited items and easily suffer from the severe problem of part domination (i.e., the predicted bounding boxes tend to be dominated only by the most discriminative parts (e.g., central regions) of a prohibited item) [18]–[20], as shown in Fig. 1. For example, for the class knife, the model is often trained to identify only the handguard as the detection results, greatly deteriorating the final detection performance.

To alleviate the problem of part domination in X-ray images, we propose to learn the objectness information of the whole object from natural images and transfer this knowledge to X-ray prohibited item detection. Our proposal is inspired by the observation that, compared with X-ray images, natural images with box annotations are more easily accessible. Nevertheless, the natural images and X-ray images have large modality discrepancy. Hence, how to perform knowledge transfer in the modality-irrelevant space merits further investigation.

To this end, we propose an intra-inter objectness learning network (I^{2}OL-Net), which can effectively estimate the scales of prohibited items, to train an X-ray prohibited item detector under point supervision. Our method successfully transfers the knowledge learned from large-scale natural images to X-ray images and takes advantage of wavelet decomposition-based adversarial learning to reduce the modality discrepancy between natural images and X-ray images. Therefore, our method can significantly address the problem of part domination for point-supervised X-ray prohibited item detection.

Specifically, our method introduces two key modules (in-

This work was in part supported by the National Natural Science Foundation of China under Grants 62372388, 62071404, and U21A20514 and by the Fundamental Research Funds for the Central Universities under Grant 20720240076. *(Corresponding author: Yan Yan.)*

C. Wang, Y. Yan, and H. Wang are with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China and the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: chenyangwang@stu.xmu.edu.cn; yanyan@xmu.edu.cn; hanzi.wang@xmu.edu.cn).

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, UK (e-mail: jinghao.xue@ucl.ac.uk).

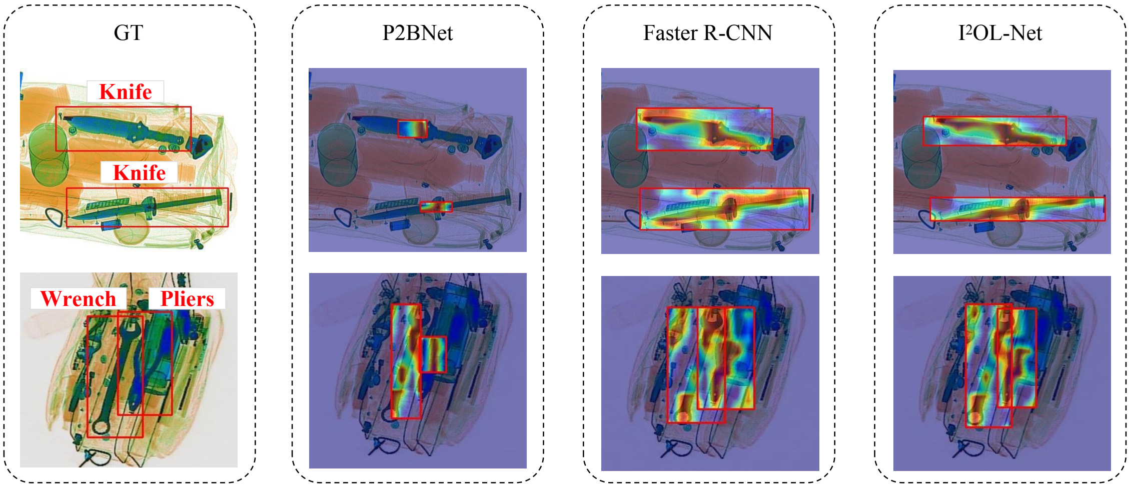


Fig. 1. Illustration of part domination. From left to right are the ground truth (GT), detection results obtained by P2BNet (point supervision) [1], Faster R-CNN [2] (box supervision), and our I^2OL-Net (point supervision), respectively. The attention maps with high-confidence proposals are overlaid on input images. P2BNet is severely affected by part domination, where the detection results only concentrate around the central regions. Our I^2OL-Net can achieve closed results to Faster R-CNN but with a much lower annotation cost. The images are taken from the SIXray dataset [3].

volving an intra-modality objectness learning (intra-OL) module and an inter-modality objectness learning (inter-OL) module). In the intra-OL module, we propose a local focus Gaussian masking (LFGM) block and a global random Gaussian masking (GRGM) block to collaboratively prevent the model from focusing solely on the most discriminative regions of the item, thereby enforcing the model to explore more potential regions that can represent the whole prohibited item. In the inter-OL module, we design a wavelet decomposition-based adversarial learning (WDAL) block and an objectness block. The WDAL block first applies wavelet decomposition to both natural images and X-ray images, decoupling the style and content information. Then, it performs adversarial learning between the styles from the two modalities, greatly minimizing the modality differences between natural images and X-ray images. In this way, the objectness block can incorporate the objectness information learned from natural images into a multiple instance learning (MIL) branch, substantially increasing the weights of proposal boxes containing the whole items.

The main contributions of this paper are as follows :

- We perform point-supervised X-ray prohibited item detection by leveraging more accessible natural images with box annotations. We propose a novel I^2OL-Net that effectively estimates the scale information of prohibited items in X-ray images with low annotation cost.
- We design an intra-OL module to jointly perform local focus Gaussian masking and global random Gaussian masking. Meanwhile, we design an inter-OL module to perform wavelet decomposition-based adversarial learning. In this way, our method can learn the objectness consistently from both the X-ray images and natural images, greatly reducing the problem of part domination.
- We conduct extensive experiments on four X-ray datasets. Compared with the state-of-the-art point-supervised object detection model P2BNet [1], I^2OL-Net (with the ResNet-50 backbone) achieves 35.12%, 50.50%, 56.96%, and 35.66% performance improvements in terms of AP_{50} on OPIXray [21], HIXray [22], SIXray [3], and PIDray [23], respectively. This clearly shows the benefits of exploiting natural images for X-ray prohibited item detection.

The remainder of this paper is organized as follows. First, we review the related work in Section II. Then, we present our proposed method in Section III. Next, we perform extensive experiments on four X-ray datasets in Section IV. Finally, we draw the conclusion in Section V.

II. Related Work

In this section, we review some related work. First, we introduce X-ray prohibited item detection methods in Section II-A. Then, we review weakly-supervised and point-supervised object detection methods in Section II-B. Finally, we review domain adaption object detection methods in Section II-C.

A. X-Ray Prohibited Item Detection

As a special type of object detection task, X-ray prohibited item detection aims to detect the locations and categories of prohibited items in X-ray images. Existing prohibited item detection methods [24]–[27] are often based on popular object detection paradigms (including one-stage detectors [28]–[31] and two-stage detectors [2], [32], [33]). Unlike object detection in natural images, prohibited item detection in X-ray images usually suffers from heavy occlusion and item overlapping. Wei *et al.* [21] develop a de-occlusion attention module (DOAM) containing an edge attention module and a region attention module to detect occluded items. Miao *et al.* [3] assume that each input image is sampled from a mixed distribution and introduce a class-balanced hierarchical refinement (CHR) method. Tao *et al.* [22] employ a locally inhibitory module (LIM) for recognizing distinguishable features while ignoring irrelevant information in occluded items. Wang *et al.* [23] propose a selective dense attention network (SDANet) with dense attention and attention dependency refinement modules.

The above methods depend heavily on box annotations. Although such a way can give promising performance, it usually requires extensive human efforts to collect accurate annotations. To balance the annotation cost and the detection performance, we study prohibited item detection under point supervision, which only provides an anchor point for locating the area near the center of the item and thus greatly reduces the total annotation time.

B. *Weakly-Supervised and Point-Supervised Object Detection*

Instead of using box supervision, weakly-supervised and point-supervised object detection methods leverage weaker forms of supervision (such as image supervision and point supervision) to reduce the annotation cost. According to [1], the average time of annotating a single point is about 1.87s per image, which approximates that of image-level annotation (1.50s per image) but is much lower than that of annotating a bounding box (34.5s per image) on VOC [34].

For weakly-supervised object detection, Bilen *et al.* [35] first introduce weakly-supervised deep detection networks (WSDDN) which incorporate MIL into weakly-supervised object detection. Tang *et al.* [17] develop online instance classifier refinement (OICR) to alleviate the problem of part domination by iterative refinement. Tang *et al.* [36] design proposal cluster learning (PCL) by expanding detection regions through proposal clustering. Seo *et al.* [37] introduce object discovery by leveraging contrastive learning to improve pseudo-labeling accuracy. The above methods detect objects in an image using a model trained on image-level annotations. However, due to the absence of location information and the challenges in distinguishing densely packed objects, these methods struggle to perform effectively in complex scenarios. In contrast, point-supervised object detection offers location cues for objects and is still significantly more label-efficient than box-supervised object detection.

For point-supervised object detection, Papadopoulos *et al.* [38] first introduce center-click annotation and use the error between two clicks to estimate the scale. Ren *et al.* [39] propose a unified object detection framework that is capable of handling various forms of supervision. Chen *et al.* [1] propose P2BNet to bridge the performance gap between point-supervised and box-supervised detectors by generating high-quality proposal bags for MIL. Ge *et al.* [40] develop a point-teaching method by combining Hungarian-based point matching, multiple instance learning, and point-guided copy-paste data augmentation. Chen *et al.* [41] propose a point DETR method by introducing a point encoder to DETR, encouraging full exploitation of point annotations. Zhang *et al.* [42] design a group R-CNN method, which generates proposals for point annotations via instance-level grouping and enhances precision with instance-aware learning. Wang *et al.* [43] study weakly semi-supervised X-ray prohibited item detection with points and propose a BCR-Net that requires both box and point annotations. Wu *et al.* [44] propose a rotation-modulated relational graph matching method for weakly semi-supervised oriented object detection. Luo *et al.* [45] introduce a PointOBB method to learn oriented object detection via single-point supervision. The above methods either leverage additional unlabeled data or work on oriented object detection.

Existing point-supervised object detection methods usually focus on natural image detection. Although many of these methods can successfully address the problem of part domination in natural images, their performance in X-ray images is not satisfactory due to the great differences between natural images and X-ray images. In this paper, we develop a method tailored for point-supervised X-ray prohibited item detection and adequately learn the objectness from different modalities, largely addressing the severe problem of part domination in X-ray images.

C. *Domain Adaptation Object Detection*

Domain adaptation object detection aims to transfer a model trained on the source domain to the target domain for testing. Existing methods can be roughly divided into three categories. The first category of methods [46]–[49] focuses on adjusting feature distributions from different domains by either applying adversarial learning or minimizing maximum mean discrepancy between source and target domains. Saito *et al.* [46] propose an adaptive detector method based on strong local alignment and weak global alignment. The second category of methods [50]–[53] adopts self-training strategies, aiming to generate high-quality pseudo-labels in the target domain. RoyChowdhury *et al.* [53] automatically labels target data according to high-confident detection results. The third category of methods [54]–[56] leverages teacher-student frameworks to achieve domain adaptation detection through consistency constraints predicted by the detector. Deng *et al.* [57] introduce an unbiased mean teacher (UMT) model for cross-domain distillation.

In this paper, we investigate the way of transferring the objectness information learned from natural images to X-ray images based on the simple yet effective wavelet decomposition. This clearly shows the great potential of making use of natural images for addressing prohibited item detection in X-ray images.

III. Method

In this section, we first describe the problem formulation in Section III-A. Then, we give the overview of our proposed method in Section III-B. Finally, we introduce the two key modules of our method in Section III-C and Section III-D.

A. *Problem Formulation*

In this paper, we aim to train a prohibited item detector by exploiting the objectness knowledge from both X-ray images (with point annotations) and easily accessible natural images (with box annotations). Such a manner greatly alleviates the

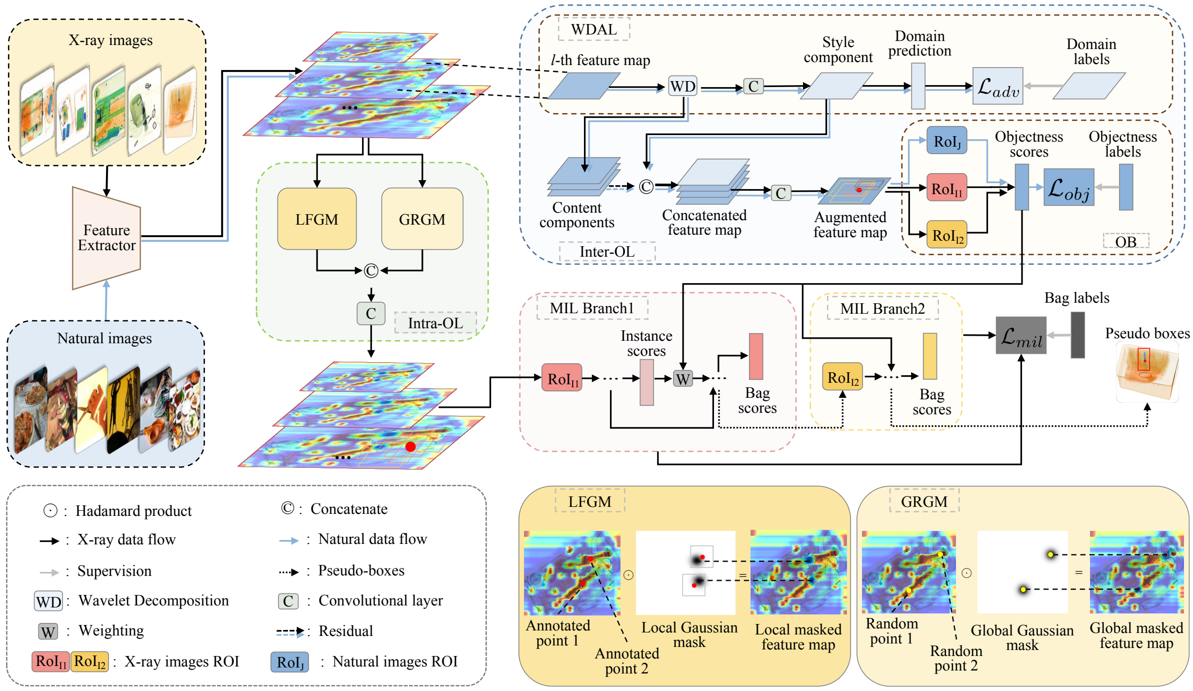


Fig. 2. Overview of our proposed I^{2}OL-Net, which consists of an intra-OL module and an inter-OL module. The intra-OL module contains a local focus Gaussian masking (LFGM) block and a global random Gaussian masking (GRGM) block. The inter-OL module contains a wavelet decomposition-based adversarial learning (WDAL) block and an objectness (OB) block.

severe problem of part domination and thus generates high-quality pseudo bounding boxes in X-ray images for training. In this way, our method significantly reduces the annotation cost of labeling an X-ray dataset.

Mathematically, we have an X-ray training set $\mathcal{I}_{train} = \{\mathbf{I}_n, \mathcal{P}_n\}_{n=1}^{N_X}$ annotated with point annotations $\mathcal{P}_n$ (i.e., the quasi-center (QC) points of prohibited items and their corresponding category labels as P2BNet [1]), where $\mathbf{I}_n$ denotes the n-th X-ray image, N_X is the total number of X-ray images in the training set and $\mathcal{P}_n = \{\mathbf{p}_n^k\}_{k=1}^K$. Here, $\mathbf{p}_n^k = (p_n^{c_x,k}, p_n^{c_y,k})$ denotes the coordinates of the k-th point annotation and K is the number of point annotations in the n-th image. We also have an additional natural image dataset $\mathcal{I}_{extra} = \{\mathbf{J}_n, \mathcal{B}_n\}_{n=1}^{N_A}$ (such as COCO [58]) with box annotations $\mathcal{B}_n$, where $\mathbf{J}_n$ and N_A denote the n-th natural image and the total number of natural images, respectively. Based on the above, we learn a prohibited item detection model and evaluate the learned model on the X-ray test set $\mathcal{I}_{test}$.

B. Overview

The overview of our intra-inter objectness learning network (I^{2}OL-Net) is shown in Fig. 2. I^{2}OL-Net is designed based on P2BNet [1]. P2BNet is a state-of-the-art method for point-supervised object detection, where the model is only trained with point annotations. Meanwhile, P2BNet has demonstrated excellent performance for the natural image detection task, making it a good base model for our proposed method. Note that some recently-developed object detection methods [40]–[42], [44], [45] also exploit point supervision. But these methods either leverage unlabeled data or work on oriented object detection. P2BNet consists of a backbone, a feature pyramid network (FPN), and several multiple instance learning (MIL) branches. Each MIL branch takes the region of interest (RoI) features generated by the feature extractor (including the backbone and FPN) as the input and feeds them into the classification subbranch and the instance subbranch. These two subbranches give classification scores and instance scores, respectively. Based on P2BNet, I^{2}OL-Net introduces two key components: an intra-OL module and an inter-OL module. Generally, our I^{2}OL-Net is first trained to generate pseudo bounding boxes. Then, a prohibited item detector is trained based on X-ray images and their corresponding generated pseudo bounding boxes.

Specifically, we randomly sample an image batch from $\mathcal{I}_{train}$ and $\mathcal{I}_{extra}$, and input them into the feature extractor. Then, we design an intra-OL module consisting of a local focus Gaussian masking (LFGM) block and a global random Gaussian masking (GRGM) block to explicitly encourage the model to focus on more potential discriminative regions of the item other than only the most discriminative region in X-ray images. Meanwhile, we design an inter-OL module consisting of a wavelet decomposition-based adversarial learning (WDAL) block and an objectness (OB) block. The inter-OL module first decouples the style and content features of both the X-ray image and the natural image using wavelet decomposition and then applies adversarial learning between the style features from X-ray and natural images, reducing the modality discrepancy. Subsequently, the transformed style features are concatenated with the content features, obtaining modality-agnostic features. After obtaining modality-agnostic features, we further leverage an objectness block to predict the objectness score for each proposal box. To optimize the objectness block, we minimize the objectness loss for natural images (with box annotations). In this way, the objectness score is further fused with the instance scores in the MIL branch to give more reliable pseudo bounding boxes for X-ray images. By designing intra-OL and inter-OL modules, I^{2}OL-Net can effectively learn both intra-modality and inter-modality objectness information to accurately estimate the scale of items, which can be used to train the detector. During inference, the detector only takes the X-ray image as the input for prohibited item detection.

C. Intra-OL Module

1) Local Focus Gaussian Masking (LFGM) Block: Considering that proposal boxes are generated according to annotations around the centers of prohibited items, almost all proposal boxes include the annotated center region of the items. For example, when the model is trained to detect the hammer, the proposal boxes are generated around the annotated point. Such a way may lead the model to concentrate on handles around the annotated points while neglecting other non-central discriminative regions (such as heads). To address this, we search the points with local maximum activation values and generate local Gaussian masks, explicitly enforcing the model to learn non-central regions of prohibited items.

Technically, given an X-ray image $\mathbf{I}_n$ and its corresponding point annotations $\mathcal{P}_n = \{\mathbf{p}_n^k\}_{k=1}^K$. The image $\mathbf{I}_n$ is first fed into the backbone and FPN to obtain different scales of feature maps $\{\mathbf{F}_{n,l}\}_{l=1}^L$, where $\mathbf{F}_{n,l} \in \mathbb{R}^{H_l \times W_l \times C_l}$, L denotes the number of feature maps, and H_l, W_l, and C_l represent the height, width, and the number of channels of feature map $\mathbf{F}_{n,l}$, respectively. For each channel of the feature map $\mathbf{F}_{n,l}$, we first find the point with the highest activation within a window of size $(\beta W_l) \times (\beta H_l)$ (β denotes the scaling factor), centered at the annotated point $\mathbf{p}_n^k$. This point serves as the mean position for generating the local Gaussian mask. The above process can be expressed as

$$\begin{aligned} \mu_{n,l}^{c,k} = \text{argmax}_{x',y'} \mathbf{F}_{n,l}^c \left(p_n^{c_x,k} + x', p_n^{c_y,k} + y'\right), \\ |x'| \leq \frac{\beta W_l}{2},\ |y'| \leq \frac{\beta H_l}{2} \end{aligned} \tag{1}$$

where $\mu_{n,l}^{c,k}$ denotes the mean position on the c-th channel feature map $\mathbf{F}_{n,l}^c$.

Meanwhile, for each $\mathbf{F}_{n,l}$, we also define a covariance matrix $\Sigma_l = \alpha \mathbf{R}_l$, where $\mathbf{R}_l = \begin{bmatrix} W_l & 0 \\ 0 & H_l \end{bmatrix}$ is the 2D diagonal matrix and α denotes the scaling factor.

Finally, we apply local Gaussian masks to each channel of the feature map,

$$\mathbf{F}_{n,l}^{'c} = \mathbf{F}_{n,l}^c (1 - \sum_{k=1}^{K} \mathbf{G}_{n,l}^{c,k}), \tag{2}$$

where $\mathbf{F}_{n,l}^{'c}$ denotes the c-th channel of the masked feature map $\mathbf{F}_{n,l}^{'}$ and $\mathbf{G}_{n,l}^{c,k} \sim N(\mu_{n,l}^{c,k}, \Sigma_l)$ denotes the local Gaussian mask.

2) Global Random Gaussian Masking (GRGM) Block: By applying the LFGM block, the model is enforced to learn non-central regions. However, the model may still focus on local regions. Hence, we further design the GRGM block, which randomly applies several Gaussian masks on feature maps, to prevent the model from overly focusing on local discriminative regions and learn the whole item (including both central and non-central regions) globally. Such a way is beneficial to exploit intra-modality objectness knowledge. Unlike the LFGM block, we randomly generate points for each feature map in the GRGM block.

Specifically, for each feature map $\mathbf{F}_{n,l}$, a set of points $\{\mathbf{O}_{n,l}^m\}_{m=1}^M$ are randomly generated, where $\mathbf{O}_{n,l}^m = (x_{n,l}^m, y_{n,l}^m)$ represents the m-th randomly generated point (x coordinate and y coordinate) on the feature map $\mathbf{F}_{n,l}$ and M denotes the total number of generated points. Hence, a Gaussian mask $\mathbf{G}_{n,l}^{'m} \sim N(\mathbf{O}_{n,l}^m, \Sigma_l)$ is generated for each point $\mathbf{O}_{n,l}^m$.Finally, we apply these Gaussian masks to the feature map $\mathbf{F}_{n,l}$ and subtract them from the original feature map,

$$\mathbf{F}_{n,l}^{''} = \mathbf{F}_{n,l} (1 - \sum_{m=1}^{M} \mathbf{G}_{n,l}^{'m}), \tag{3}$$

where $\mathbf{F}_{n,l}^{''}$ denotes the masked feature map.

After obtaining $\mathbf{F}_{n,l}^{'}$ and $\mathbf{F}_{n,l}^{''}$, we concatenate the two feature maps and reduce their channels using a 1×1 convolutional layer, resulting in the feature $\mathbf{E}_{n,l}$. This feature is then used as the input feature of the MIL branches.

D. Inter-OL Module

1) Wavelet Decomposition Based Adversarial Learning (WDAL) Block: Due to significant modality discrepancy between natural images and X-ray images, directly transferring the objectness knowledge extracted from natural images to X-ray images is not desirable. To address this issue, we propose a WDAL block, which first decouples the feature maps of both natural and X-ray images into style and content using wavelet decomposition and then reduces the modality differences caused by style differences through adversarial learning. Our WDAL block is motivated by the fact that wavelet decomposition is effective in preserving content information for domain adaption with fixed parameters [59].

Technically, we use Haar wavelets for wavelet decomposition. Haar wavelets consist of four kernels, including $\{\mathbf{L}\mathbf{L}^{\mathrm{T}}, \mathbf{L}\mathbf{H}^{\mathrm{T}}, \mathbf{H}\mathbf{L}^{\mathrm{T}}, \mathbf{H}\mathbf{H}^{\mathrm{T}}\}$, where $\mathbf{L}$ and $\mathbf{H}$ are low-pass and high-pass filters, respectively, defined as

$$\mathbf{L}^{\mathrm{T}} = \frac{1}{\sqrt{2}}[1, 1], \mathbf{H}^{\mathrm{T}} = \frac{1}{\sqrt{2}}[-1, 1]. \tag{4}$$

Wavelet decomposition decomposes the feature map $\mathbf{M}_{n,l}$ (extracted from the X-ray image $\mathbf{I}_n$ or the natural image $\mathbf{J}_n$)

into four components, that is,

$$\begin{aligned}\mathbf{A}_{n,l} &= \mathbf{M}_{n,l} * (\mathbf{LL}^{\mathrm{T}}),\\ \mathbf{H}_{n,l} &= \mathbf{M}_{n,l} * (\mathbf{LH}^{\mathrm{T}}),\\ \mathbf{V}_{n,l} &= \mathbf{M}_{n,l} * (\mathbf{HL}^{\mathrm{T}}),\\ \mathbf{D}_{n,l} &= \mathbf{M}_{n,l} * (\mathbf{HH}^{\mathrm{T}}),\end{aligned} \tag{5}$$

where '$*$' denotes the convolutional operation; $\mathbf{A}_{n,l}$ represents the low-frequency component; $\mathbf{H}_{n,l}$, $\mathbf{V}_{n,l}$, and $\mathbf{D}_{n,l}$ denote the high-frequency components. Typically, the low-frequency component mainly corresponds to the style feature, capturing detailed style information of the X-ray image. The high-frequency components correspond to the content features, focusing on local details and edges.

$\mathbf{A}_{n,l}$ is then fed into a convolutional layer (with learnable parameters) to obtain the transformed style feature $\mathbf{A}'_{n,l}$. To mitigate the modality discrepancy, we perform adversarial learning on the transformed style feature. We consider the natural image dataset and the X-ray dataset as the source domain and the target domain, respectively. In this paper, we train a domain classifier on each activation value of the transformed style feature to predict the modality of each style feature. Such a way is beneficial to enlarging the number of training samples for adversarial learning and reducing global image differences [60].

Mathematically, let $\hat{D}_n$ be the domain label for the n-th image, where $\hat{D}_n = 0$ if the n-th image belongs to the source domain and $\hat{D}_n = 1$ otherwise. We represent the activation value at the coordinate (x, y) for the transformed style feature as $\phi_{x,y}(\mathbf{A}'_{n,l})$. The predicted value $d^{x,y}_{n,l}$ of each position on the corresponding style feature $\phi_{x,y}(\mathbf{A}'_{n,l})$ is utilized to compute the adversarial loss $\mathcal{L}_{adv}$ based on the cross-entropy, that is,

$$\mathcal{L}_{adv} = -\sum_{n,l,x,y} [\hat{D}_n \log d^{x,y}_{n,l} + (1 - \hat{D}_n) \log(1 - d^{x,y}_{n,l})]. \tag{6}$$

To align modality distributions, we simultaneously optimize the parameters of the modality classifier by minimizing the adversarial loss and the parameters of the base network by maximizing this loss. As done in [60], we also employ a gradient reversal layer (GRL) to train the modality classifier.

Then, the content features $\mathbf{H}_{n,l}$, $\mathbf{V}_{n,l}$, $\mathbf{D}_{n,l}$, and the transformed style feature $\mathbf{A}'_{n,l}$ are concatenated to obtain the modality-agnostic feature $\mathbf{K}_{n,l}$, that is,

$$\mathbf{K}_{n,l} = \mathbf{A}'_{n,l} \oplus \mathbf{H}_{n,l} \oplus \mathbf{V}_{n,l} \oplus \mathbf{D}_{n,l}, \tag{7}$$

where '$\oplus$' denotes the concatenation operation.

2) Objectness (OB) Block: After transferring the objectness information from natural images to X-ray images with the WDAL block, we design an objectness block to incorporate this information into the MIL branch, enhancing the probability of proposal boxes containing the whole prohibited item.

We compute an objectness score, which reflects the probability of each RoI feature containing a whole item. Let $\{\mathbf{RoI}^{k,r}_n\}^R_{r=1}$ denote the RoI features corresponding to the proposals generated by the k-th annotated point of the n-th image (the X-ray or natural image). Here, R denotes the total number of proposals. Then, we perform RoI pooling on each feature map and compute the final objectness score through fully-connected layers. The objectness block can be formulated as

$$obj^{k,r}_n = \mathrm{FC}(\mathrm{RoI\_Pooling}(\mathbf{RoI}^{k,r}_n)), \tag{8}$$

where RoI_Pooling$(\cdot)$ and FC$(\cdot)$ denote the RoI pooling and the fully-connected layers, respectively; $obj^{k,r}_n$ represents the objectness score corresponding to the r-th proposal box generated by the k-th point.

For natural images, we can leverage rich box annotations to train the objectness block. Specifically, for proposal boxes generated from a point annotation, we compute the intersection over union $\mathrm{IoU}^{k,r}_n$ between them and the ground-truth box annotations, where we set the proposal boxes with $\mathrm{IoU}^{k,r}_n$ greater than or equal to 0.5 as positive samples, labeled as 1, and those below 0.5 as negative samples, labeled as 0. Based on this, we can employ the cross-entropy loss function to adjust the model parameters, that is,

$$\mathcal{L}_{obj} = -\sum_{n,k,r} (\mathrm{BIoU}^{k,r}_n \log(obj^{k,r}_n) + (1 - \mathrm{BIoU}^{k,r}_n) \log(1 - obj^{k,r}_n)), \tag{9}$$

$$\mathrm{BIoU}^{k,r}_n = \begin{cases} 1, & \text{if } \left(\frac{\mathrm{Area}(obj^{k,r}_n) \cap \mathrm{Area}(GT^k_n)}{\mathrm{Area}(obj^{k,r}_n) \cup \mathrm{Area}(GT^k_n)}\right) \geq 0.5 \\ 0, & \text{if } \left(\frac{\mathrm{Area}(obj^{k,r}_n) \cap \mathrm{Area}(GT^k_n)}{\mathrm{Area}(obj^{k,r}_n) \cup \mathrm{Area}(GT^k_n)}\right) < 0.5. \end{cases} \tag{10}$$

Here, $\mathrm{Area}(obj^{k,r}_n)$ and $\mathrm{Area}(GT^k_n)$ denote the region of the r-th proposal box and its corresponding ground truth box annotation, respectively.

For X-ray images, we multiply $obj^{k,r}_n$ by the instance score of the k-th point annotation in the MIL branch (i.e., the k-th bag of MIL). Hence, the proposal boxes will be assigned larger weights for higher objectness scores. The classification scores and subsequent calculation of the MIL loss remain unchanged. This process can be represented as

$$S^{k,r}_n = \mathrm{Softmax}(\mathrm{FC}_{\mathrm{ins}}(\mathbf{RoI}^{k,r}_n)) \times obj^{k,r}_n, \tag{11}$$

where $S^{k,r}_n$ denotes the weighted instance score; Softmax$(\cdot)$ and $\mathrm{FC}_{\mathrm{ins}}(\cdot)$ denote the softmax function and the instance subbranch, respectively.

Finally, the joint loss of I^{2}OL-Net is given as

$$\mathcal{L} = \mathcal{L}_{mil} + \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{obj}, \tag{12}$$

where $\mathcal{L}_{mil}$ represents the total MIL loss defined in P2BNet; λ_1 and λ_2 are the balancing weights.

IV. Experiments

In this section, we first introduce the datasets in Section IV-A. Then, we give implementation details and evaluation metrics of our method in Section IV-B and Section IV-C, respectively. Next, we compare our method with state-of-the-art methods on the four X-ray datasets in Section IV-D. Finally, we conduct ablation studies in Section IV-E.

A. Datasets

We evaluate our method on 4 commonly used X-ray datasets: OPIXray [21], HIXray [22], SIXray [3], and PIDray [23]. In this paper, we adopt the division of the training and test sets according to the default evaluation protocols provided by the datasets. These datasets provide predefined

TABLE I
SUMMARY OF THE FOUR X-RAY DATASETS USED IN THE EXPERIMENTS.

Dataset	OPIXray	HIXray	SIXray	PIDray
# of Categories	5	8	6	12
# of Training	7,109	36,295	7,127	29,457
# of Test	1,776	9,069	1,802	18,220
Total	8,885	45,364	8,929	47,677

training and test sets, which can be directly used for model training and evaluation, ensuring data validity and consistency in our experiments. Specifically, the OPIXray dataset consists of 5 categories with a total of 8,885 images, including 7,109 images in the training set and 1,776 images in the test set. The HIXray dataset consists of 8 categories with a total of 45,364 images, including 36,295 images in the training set and 9,069 images in the test set. The SIXray dataset consists of 6 categories with a total of 1,059,231 images. We use 8,929 images containing prohibited items for training and testing, including 7,127 images in the training set and 1,802 images in the test set. The PIDray dataset consists of 12 categories with a total of 124,486 images. Among these images, we use 47,677 images containing prohibited items, including 29,457 images in the training set and 18,220 images in the test set. Table I summarizes the detailed information of all the four X-ray datasets.

During training, we also use the COCO14 training set [58] as the extra dataset. To alleviate the unbalanced number of images between the natural image dataset and the X-ray dataset, we randomly select the same number of images as the X-ray dataset from the COCO14 training set. In addition, we select the objects whose object sizes are close to/smaller than those of prohibited items. Hence, most of the classes with small object sizes are chosen from the COCO14 dataset while some large object sizes are not selected. Finally, we combine the selected sampled subset with the X-ray training set to construct the final training set.

B. Implementation Details

Our I^{2}OL-Net is implemented based on MMDetection [64]. As done in [1], [11], we adopt ResNet-50 [65] pretrained on ImageNet as our backbone for model training. We train our model on 2 NVIDIA RTX 3090 GPUs using the SGD algorithm with a learning rate of 0.005. The total training epochs are set to 12 and the batch size is set to 4. We randomly select each batch from both $\mathcal{I}_{train}$ and $\mathcal{I}_{extra}$. The balancing weights λ_1 and λ_2 in (12) are empirically set to 0.1 and 1, respectively. The scaling factors α and β are set to 0.1 and 0.15, respectively. The number of random points M in GRGM is set to 2. Except for the parameters mentioned above, all the other settings remain the same as P2BNet.

We follow the standard procedure of saving the model weights after completing the final epoch of training. This ensures that we employ the trained model after the full training process, avoiding potential inconsistencies that may arise from saving the model at intermediate stages.

C. Evaluation Metrics

In all our experiments, we evaluate the performance of our method using two widely-used object detection metrics: Average Precision (AP) and Average Precision at an intersection over union (IoU) threshold of 0.50 (AP$_{50}$). AP measures the overall detection performance by averaging precision across multiple IoU thresholds (from 0.50 to 0.95 with increments of 0.05). AP provides a balanced evaluation by considering both strict and lenient IoU thresholds, reflecting the model's overall ability to localize objects. AP$_{50}$ calculates the average precision at an IoU threshold of 0.50. AP$_{50}$ measures performance with a relatively lenient IoU threshold, highlighting the model's ability to detect objects without requiring precise localization. In addition, we also use F1 scores and ROC curves for performance evaluation. The F1 score is the balance between precision and recall at a fixed IoU threshold. The ROC curve visualizes the trade-off between the true positive rate and the false positive rate across different decision thresholds. The F1 score provides precision-recall balance. The ROC curve analyzes model performance across varying confidence thresholds, offering a broader view of detection capability.

The above four metrics provide a comprehensive evaluation of the model's detection performance in X-ray images.

D. Comparison with State-of-the-Art Methods

We compare our I^{2}OL-Net with several box-supervised object detection methods (including two one-stage detectors (RetinaNet [30]and YOLOv8-s [61]), a two-stage detector (Faster R-CNN [2]), an anchor-free detector (FCOS [31]), and a Transformer-based detector (Deformable DETR [62])), image-supervised object detection methods (including PCL [36], WSOD2 [63] and OD-WSCL [37]), weakly semi-supervised object detection methods (including Group R-CNN [42], BCR-Net [43]), and a point-supervised object detection method (including P2BNet [1]). Due to the limited availability of point-supervised methods, we choose P2BNet for comparison. To demonstrate the superiority of our method, we evaluate our method on different backbones. In addition, we also train a baseline model (called P2BNet*) on natural images and fine-tune it on the X-ray dataset, where only the intra-OL module and the objectness block are used. Unless otherwise specified, all the object detection methods except for I^{2}OL-Net and P2BNet* are trained only on the X-ray dataset. The comparison results are shown in Table II. Fig. 3 visualize some detection results obtained by P2BNet and our I^{2}OL-Net on four X-ray datasets.

From Table II, we can see that the box-supervised detectors can achieve better performance than both image-supervised detectors, weakly semi-supervised detectors, and point-supervised detectors. This is because of the scale and position information provided by the box annotations. Image-supervised detectors fail to accurately detect prohibited items on X-ray datasets. This can be ascribed to the fact that the selective search strategy used by these detectors cannot effectively generate proposals covering the whole prohibited items under complex background conditions of X-ray datasets. Moreover, these detectors rely solely on MIL and may not

TABLE II

PERFORMANCE COMPARISON RESULTS (%) BETWEEN OUR METHOD AND STATE-OF-THE-ART METHODS ON FOUR X-RAY DATASETS.

Method	Backbone	OPIXray			HIXray			SIXray			PIDray		
		AP	AP_{50}	F1	AP	AP_{50}	F1	AP	AP_{50}	F1	AP	AP_{50}	F1
Box-supervised detectors													
Faster R-CNN [2]	ResNet-50	40.06	89.87	**88.70**	45.06	84.42	**83.29**	**56.31**	**89.65**	**86.87**	64.73	82.37	77.31
RetinaNet [30]	ResNet-50	**40.53**	**90.02**	88.20	46.88	83.85	81.47	51.17	84.88	79.61	66.78	80.95	75.55
YOLOv8-s [61]	CSPDarkNet	37.90	83.40	80.84	43.60	77.00	74.77	55.70	84.20	79.14	66.00	77.50	71.43
FCOS [31]	ResNet-50	38.36	89.60	85.37	45.80	**86.80**	82.06	53.80	86.80	81.75	67.00	**83.10**	77.39
Deformable DETR [62]	ResNet-50	37.70	88.90	87.51	**49.80**	82.05	80.35	56.10	89.20	85.70	**70.12**	82.97	**78.99**
Image-supervised detectors													
PCL [36]	ResNet-50	-	2.87	-	-	4.21	-	-	1.48	-	-	7.76	-
WSOD2 [63]	ResNet-50	-	3.11	-	-	5.18	-	-	1.65	-	-	**9.22**	-
OD-WSCL [37]	ResNet-50	-	**4.21**	**2.81**	-	**6.87**	**4.09**	-	**1.87**	**1.01**	-	9.17	**6.43**
Weakly semi-supervised detectors													
Group R-CNN [42]	ResNet-50	28.30	76.65	74.98	44.80	75.31	70.82	25.32	61.39	59.41	54.62	74.83	69.61
BCR-Net [43]	ResNet-50	**29.04**	**79.21**	**77.18**	**47.01**	**78.42**	**73.77**	**31.82**	**69.73**	**67.57**	**57.92**	**79.04**	**73.27**
Point-supervised detectors													
P2BNet [1]	ResNet-34	1.04	5.22	4.78	6.61	18.41	16.55	2.11	8.21	4.07	6.51	15.83	9.20
P2BNet [1]	ResNet-50	3.32	15.50	13.93	10.87	25.23	23.19	3.43	10.57	7.47	8.98	19.77	13.85
P2BNet* [1]	ResNet-50	5.08	22.55	21.11	13.92	35.61	34.67	6.09	16.48	12.05	11.74	27.97	21.22
I^2OL-Net	ResNet-34	10.72	38.12	36.35	29.70	74.50	72.81	22.45	61.84	57.17	29.91	51.88	48.06
I^2OL-Net (Ours)	ResNet-50	**15.09**	**50.62**	**49.10**	**30.62**	**75.73**	**76.04**	**26.24**	**67.53**	**63.30**	**31.82**	**55.43**	**54.01**

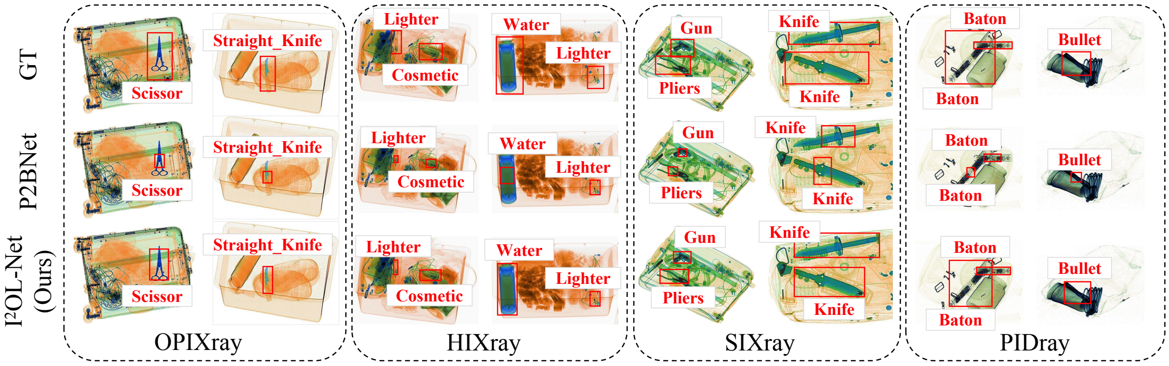
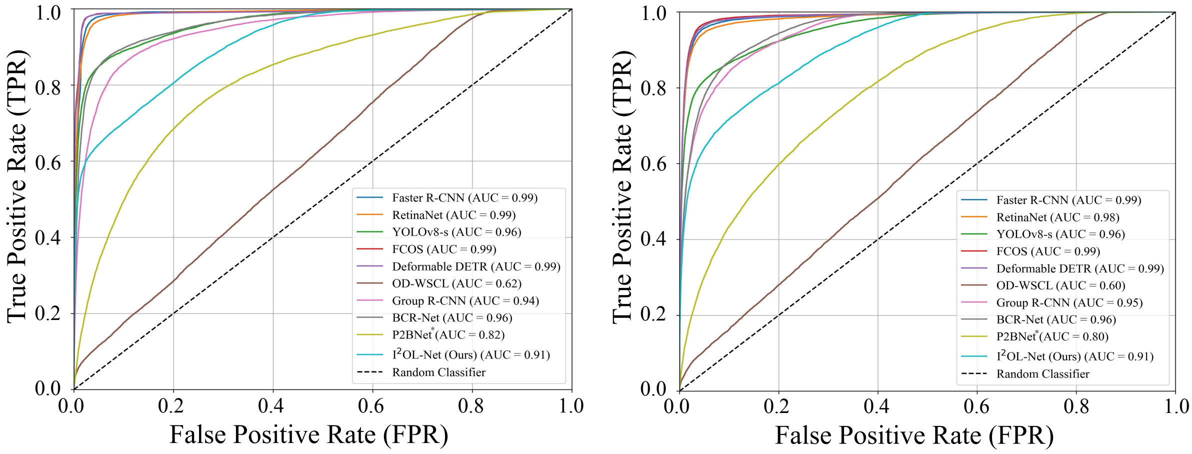


Fig. 3. Visualization of detection results of I^2OL-Net and P2BNet on four datasets. I^2OL-Net significantly alleviates the problem of part domination caused by intra-class variations in X-ray images. By introducing intra-OL and inter-OL modules, our method can detect prohibited items more accurately than P2BNet.



(a) HIXray

(b) PIDray

Fig. 4. ROC Curves on the HIXray and PIDray datasets. Box-supervised detectors achieve the best performance on the HIXray and PIDray datasets. The performance of weakly semi-supervised detectors follows closely behind. Our point-supervised detector, I^2OL-Net, shows competitive results, closely trailing the box-supervised detectors. In contrast, the baseline P2BNet demonstrates a significant performance drop, while the image-supervised detector OD-WSCL performs poorly.

TABLE III
CLASS-WISE AP_{50} FOR EACH PROHIBITED ITEM CATEGORY ON THE HIXRAY DATASET.

Method	Mobile_Phone	Cosmetic	Laptop	Water	Portable_Charger_1	Portable_Charger_2	Tablet	Nonmetallic_Lighter	Mean
# of Samples	43204	7969	8046	2471	9919	6216	3921	706	10306.5
Faster R-CNN [2]	97.04	66.93	97.94	92.26	96.37	94.64	95.22	34.94	84.42
RetinaNet [30]	98.19	65.09	98.89	93.09	96.59	95.19	96.39	27.39	83.85
OD-WSCL [37]	9.58	4.07	7.15	8.70	5.68	8.35	11.43	0.00	6.87
Group R-CNN [42]	96.51	52.45	97.84	85.30	90.68	86.82	92.88	0.00	75.31
BCR-Net [43]	96.81	64.66	98.10	90.18	92.38	90.09	95.01	0.13	78.42
P2BNet* [1]	43.45	27.34	64.22	30.06	21.71	26.78	66.20	5.12	35.61
I^2OL-Net(Ours)	92.01	71.56	97.11	76.18	71.64	86.22	94.77	16.38	75.73

TABLE IV
CLASS-WISE AP_{50} FOR EACH PROHIBITED ITEM CATEGORY ON THE PIDRAY DATASET.

Method	Baton	Pliers	Hammer	Powerbank	Scissors	Wrench	Gun	Bullet	Sprayer	HandCuffs	Knife	Lighter	Mean
# of Samples	1513	4236	3546	5171	4352	4350	2178	1837	2970	2096	3290	4169	3309
Faster R-CNN [2]	98.45	99.81	97.23	94.71	96.01	98.09	15.83	94.13	80.11	98.78	34.89	80.40	82.37
RetinaNet [30]	98.73	99.56	97.34	96.10	98.19	93.09	10.27	91.57	82.87	94.88	28.13	80.67	80.95
OD-WSCL [37]	0.01	23.93	9.81	14.01	9.78	8.77	0.45	19.26	4.51	9.44	1.15	8.89	9.17
Group R-CNN [42]	92.66	98.54	94.70	93.40	78.69	95.71	23.74	92.91	36.01	97.68	14.51	79.41	74.83
BCR-Net [43]	97.44	99.01	97.13	95.69	84.37	96.44	19.81	95.41	63.99	97.90	21.43	79.86	79.04
P2BNet* [1]	0.67	68.21	11.34	48.01	21.51	39.33	7.82	51.03	6.70	49.70	3.31	28.01	27.97
I^2OL-Net (Ours)	1.33	95.61	19.71	89.64	77.72	85.23	13.44	93.01	22.49	94.35	11.57	61.11	55.43

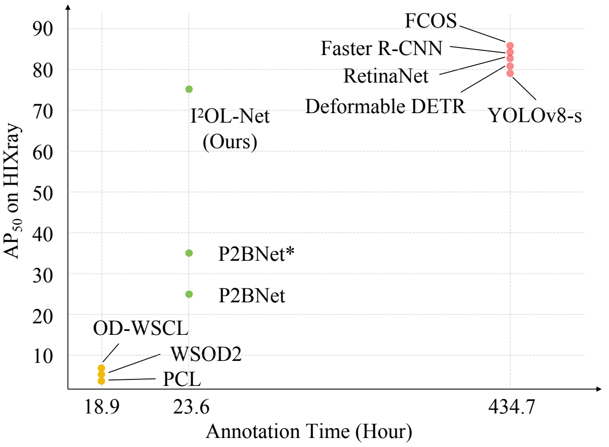


Fig. 5. The detection accuracy (AP_{50}) and the total annotation time obtained by several competing methods in the HIXray dataset.

adequately capture the scale information of prohibited items due to the heavy overlapping in X-ray images. Since the AP metric of these image-supervised detectors is nearly 0, we use '-' to represent the results. Weakly semi-supervised detectors can achieve good performance, but they leverage a small number of box annotations. Such a way enables them to obtain more object location information, thereby improving detection accuracy. Compared with P2BNet, I^2OL-Net achieves significant improvements in terms of AP, AP_{50}, and F1 scores. For instance, on the OPIXray dataset, I^2OL-Net achieves 15.09% in terms of AP, a remarkable improvement over P2BNet* (5.08% in terms of AP) when ResNet-50 is used as the backbone. Similarly, on the HIXray dataset, I^2OL-Net shows 30.62% in terms of AP, significantly surpassing P2BNet* (13.92% in terms of AP) when ResNet-50 is used as the backbone. In addition, I^2OL-Net substantially narrows the performance gap between point-supervised detectors and box-supervised detectors. This indicates that I^2OL-Net can effectively learn the objectness by the intra-OL and inter-OL modules. P2BNet* leverages a pre-trained objectness predictor to predict the scores of RoIs on X-ray images. It improves performance by utilizing the additional natural images. However, the performance improvement obtained by P2BNet* is trivial because of the significant modality discrepancy (in terms of image appearance and content) between natural images and X-ray images. By introducing our WDAL block, the performance of I^2OL-Net can be greatly improved, indicating the importance of reducing the modality discrepancy.

Generally, AP, AP_{50} and F1 scores of all categories reflect the overall detection performance. Thus, they cannot fairly reflect the model performance for each class when there is a class imbalance in the dataset. This is because the model might favor predicting the majority classes and neglect minority classes. Hence, we also give the class-wise AP_{50} obtained by several competing methods on the HIXray and PIDray datasets, as shown in Table III and Table IV. We can see that our method can achieve better performance than P2BNet* on all the prohibited items. Even when the number of training samples for some prohibited item categories (such as 'Nonmetallic Lighter' in HIXray or 'Bullet' in PIDray) is limited, our method still shows promising performance. These results further validate the advantage of our method against P2BNet* when dealing with imbalanced class distribution. Note that the class-wise AP_{50} of some prohibited item categories (such as 'Water' in HIXray or 'Baton' in PIDray) is much lower than the box-supervised methods. This indicates the great challenge of detecting some prohibited item categories when the training samples of these categories are limited.

In addition, we give the ROC curves obtained by some competing methods in HIXray and PIDray datasets, as given in Fig. 4. Our method gives much better results than P2BNet in terms of ROC curves in both HIXray and PIDray datasets.

TABLE V
THE TRAINING TIME, INFERENCE TIME, AND GPU MEMORY OBTAINED BY P2BNET AND OUR I^{2}OL-NET. HERE, THE TRAINING TIME REFERS TO THE TIME REQUIRED FOR THE UPSTREAM DETECTOR TO CONVERGE, THE INFERENCE TIME REFERS TO THE TIME TAKEN BY THE UPSTREAM DETECTOR TO PERFORM INFERENCE ON A SINGLE IMAGE, AND THE GPU MEMORY REFERS TO THE GPU MEMORY CONSUMPTION ON 1 GPU DURING TRAINING.

Method	Dataset	Training Time (h)	Inference Time (ms)	GPU Memory (GB)
P2BNet	OPIXray	6.0	16.1	7
	HIXray	15.0	16.1	11
	SIXray	6.0	16.1	8
	PIDray	11.5	16.1	10
I^{2}OL-Net	OPIXray	8.5	44.2	12
	HIXray	19.0	44.2	18
	SIXray	8.5	44.2	15
	PIDray	19.0	44.2	18

TABLE VI
ABLATION STUDY RESULTS (%) ON THE INFLUENCE OF INTRA-OL AND INTER-OL.

intra-OL	inter-OL	OPIXray		HIXray		SIXray		PIDray	
		AP	AP$_{50}$	AP	AP$_{50}$	AP	AP$_{50}$	AP	AP$_{50}$
×	×	3.32	15.50	10.87	25.23	3.43	10.57	8.98	19.77
✓	×	5.08	22.55	13.92	35.61	6.09	16.48	11.74	27.97
×	✓	13.34	42.65	27.10	65.16	23.88	60.23	28.32	46.61
✓	✓	**15.09**	**50.62**	**30.62**	**75.73**	**26.24**	**67.53**	**31.82**	**55.43**

These results further validate the effectiveness of our method.

We also visualize the detection accuracy and the total annotation time obtained by the image-supervised detectors, box-supervised detectors, a point-supervised detector, and our method on the HIXray dataset. The results are given in Fig. 5. Our method achieves a good tradeoff between the detection accuracy and annotation cost.

Finally, we compare the training time, inference time per image, and the GPU memory obtained by P2BNet and our method on the four datasets. The results are shown in Table V. Our method achieves higher training and inference time than P2BNet*. The GPU memory consumed by our method is also higher than that obtained by P2BNet*. The increased training/inference time and GPU memory consumption primarily arises from the two proposed modules. However, these modules significantly enhance detection performance. Note that the inference time of our method still satisfies real-time requirements in practical applications.

E. Ablation Studies

1) Influence of the Intra-OL and Inter-OL Modules: The ablation study results of the intra-OL and inter-OL modules are shown in Table VI. We can observe that when the intra-OL or inter-OL module is used, the model performance is improved. Specifically, when the intra-OL module is used alone, the model achieves performance improvements of 1.76%, 3.05%, 2.66%, and 2.76% in terms of AP across the four datasets. In contrast, using the inter-OL module alone results in more significant performance gains, with improvements of 10.02%, 16.23%, 20.45%, and 19.34% in terms of AP across the four datasets. When both modules are jointly used, the model achieves the best performance, with AP improvements of 11.77%, 19.75%, 22.81%, and 22.84% across the four datasets. These results indicate that the baseline method (where both the intra-OL and inter-OL modules are not used) cannot effectively learn the objectness in the X-ray dataset. Our intra-OL and inter-OL modules enforce the model to explore holistic information in X-ray images by generating Gaussian masks and transferring helpful information from box annotations in the COCO dataset, respectively. This greatly enhances the model performance.

2) Influence of the LFGM and GRGM Blocks: The ablation study results of the LFGM block and the GRGM block are given in Table VII. Both LFGM and GRGM blocks have a positive influence on the model performance on all four X-ray datasets. When the LFGM block or the GRGM block is incorporated into the baseline, the models can achieve better performance in terms of AP and AP$_{50}$. When both LFGM and GRGM blocks are combined with the baseline, the performance can be further improved, with AP improvements of 1.75%, 3.52%, 2.36%, and 3.00% across the four datasets. These results indicate that the combination of the LFGM and GRGM blocks in the intra-OL module can give performance improvements. This is because the LFGM block encourages the model to focus on non-central discriminative regions, while the GRGM block enables the model to learn the global structure of the whole prohibited item.

3) Influence of Key Components in the Inter-OL Module: We validate the effectiveness of wavelet decomposition (WD), adversarial learning (AL), and the objectness block (OB) in the inter-OL module. The results are given in Table VIII. When only the objectness block is used, the model performance is similar to the baseline, which achieves 7.28%, 15.21%, 9.65% and 13.40% in AP on four datasets. When either WD or AL

TABLE VII

Ablation study results (%) on the influence of the LFGM block and the GRGM block in the intra-OL module.

LFGM	GRGM	OPIXray		HIXray		SIXray		PIDray	
		AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
×	×	13.34	42.65	27.10	65.16	23.88	60.23	28.82	46.61
✓	×	14.89	48.87	29.64	72.03	25.46	65.98	30.71	53.34
×	✓	14.11	45.78	28.45	69.77	24.81	63.24	30.19	51.67
✓	✓	**15.09**	**50.62**	**30.62**	**75.73**	**26.24**	**67.53**	**31.82**	**55.43**

TABLE VIII

Ablation study results (%) on the influence of wavelet decomposition (WD), adversarial learning (AL), and the objectness (OB) block.

WD	AL	OB	OPIXray		HIXray		SIXray		PIDray	
			AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
×	×	×	5.08	22.55	13.92	35.61	6.09	16.48	11.74	27.97
×	×	✓	7.28	27.70	15.21	40.01	9.65	25.41	13.40	31.12
✓	×	✓	8.22	30.01	16.73	43.56	10.45	29.31	15.68	34.06
×	✓	✓	11.22	37.01	24.32	56.30	21.85	51.45	22.47	41.66
✓	✓	✓	**15.09**	**50.62**	**30.62**	**75.73**	**26.24**	**67.53**	**31.82**	**55.43**

TABLE IX

Ablation study results (%) on the influence of applying adversarial learning to different components obtained by wavelet decomposition. Low-C means only low-frequency components are used for adversarial learning, High-C means all 3 high-frequency components are used for adversarial learning.

Low-C	High-C	OPIXray		HIXray		SIXray		PIDray	
		AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
×	✓	13.89	43.03	28.77	70.00	25.05	65.22	30.84	53.84
✓	×	**15.09**	**50.62**	**30.62**	**75.73**	**26.24**	**67.53**	31.82	55.43
✓	✓	14.13	46.02	29.97	74.52	26.03	67.25	**32.01**	**56.33**

TABLE X

Ablation study results (%) on the influence of the scaling factor α in the GRGM block.

α	OPIXray		HIXray		SIXray		PIDray	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
0.05	15.01	49.40	29.72	72.59	25.34	65.70	31.22	54.86
0.10	**15.09**	**50.62**	**30.62**	**75.73**	**26.24**	**67.53**	31.82	55.43
0.15	14.75	48.40	29.69	72.34	25.41	65.82	**31.91**	**55.62**
0.20	13.94	43.27	29.03	70.28	24.85	63.28	30.47	52.13

is added, the model performance can be improved. The model achieves the best performance (13.94%, 29.03%, 24.85%, and 30.47% in terms of AP across the four datasets) when WD, AL, and OB are all used. These results indicate that applying AL based on wavelet decomposition can reduce the modality difference between X-ray images and natural images. Such a way is beneficial to learning holistic information from natural images.

4) Influence of Low or High Frequency Components in WDAL: The WDAL block decomposes the feature map into low-frequency style components and high-frequency content components using wavelet decomposition and then employs adversarial learning to bring the style features of X-ray images closer to those of natural images. We investigate the influence of applying adversarial learning to different components obtained by wavelet decomposition on the model performance. The results are given in Table IX. Our method achieves the best performance on most datasets when only low-frequency components are used for adversarial learning. This can be ascribed to the fact that the low-frequency components contain the essential information for extracting the object structure. When high-frequency components or both high-frequency and low-frequency components are used for adversarial learning, the performance obtained by our method is decreased. This is because the high-frequency components contain more object details, which can affect objectness learning.

5) Influence of α in the GRGM block: We evaluate the influence of the scaling factor α in the GRGM block. α is used

TABLE XI
ABLATION STUDY RESULTS (%) ON THE INFLUENCE OF M IN THE GRGM BLOCK. $M = 0$ INDICATES THAT THE GRGM BLOCK IS NOT USED.

M	OPIXray		HIXray		SIXray		PIDray	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
0	14.89	48.87	29.64	72.03	25.46	65.98	30.71	53.34
1	14.98	49.19	30.31	74.99	25.91	66.80	30.99	54.04
2	**15.09**	**50.62**	**30.62**	**75.73**	**26.24**	**67.53**	**31.82**	**55.43**
3	14.32	47.08	30.52	75.40	26.12	67.01	31.34	55.03

TABLE XII
ABLATION STUDY RESULTS (%) ON THE INFLUENCE OF THE SCALING FACTOR β IN THE LFGM BLOCK.

β	OPIXray		HIXray		SIXray		PIDray	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
0.10	**15.21**	**51.09**	30.34	75.11	25.93	66.89	31.53	55.17
0.15	15.09	50.62	**30.62**	**75.73**	**26.24**	**67.53**	31.82	55.43
0.20	14.70	48.21	29.85	73.46	25.57	66.30	**31.99**	**56.24**
0.25	14.44	47.75	29.10	71.21	25.01	65.10	31.10	54.33

TABLE XIII
PERFORMANCE COMPARISON RESULTS (%) BETWEEN P2BNET AND OUR I^2OL-NET WITH DIFFERENT DOWNSTREAM DETECTORS ON FOUR X-RAY DATASETS.

Upstream	Downstream	OPIXray		HIXray		SIXray		PIDray	
		AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
P2BNet	Faster R-CNN	3.32	**15.50**	**10.87**	25.23	3.43	10.57	8.98	19.77
	RetinaNet	2.45	12.01	10.64	25.01	1.98	5.67	8.21	18.90
	YOLOv8-s	1.99	8.78	8.82	14.56	3.00	9.21	7.45	16.42
	FCOS	3.87	14.64	10.66	24.67	3.33	**10.76**	9.05	18.89
	Deformable DETR	**4.44**	13.43	9.34	**26.33**	**5.02**	9.88	**11.76**	**20.10**
I^2OL-Net	Faster R-CNN	**15.09**	**50.62**	30.62	**75.73**	**26.24**	**67.53**	31.82	55.43
	RetinaNet	14.27	44.88	29.41	74.57	21.88	59.72	30.22	53.13
	YOLOv8-s	11.00	36.20	25.80	63.10	23.50	57.00	29.70	48.90
	FCOS	13.55	48.57	30.65	72.77	22.62	61.25	30.02	54.75
	Deformable DETR	13.95	50.36	**32.77**	74.88	25.93	65.46	**35.15**	**56.31**

to generate the covariance matrix of the Gaussian distribution. The results are given in Table X.

It can be observed that different values of α can significantly affect the model performance. On the OPIXray, HIXray, and SIXray datasets, the model achieves the optimal performance when the value of α is set to 0.10. Specifically, the model achieves 50.62%, 75.73%, 67.53%, and 55.43% in terms of AP_{50} across the four datasets. When the value of α is increased or decreased, the model performance gradually declines on all datasets. Our method achieves the best performance (31.91% in terms of AP) on the PIDray dataset when the value of α is set to 0.15. In comparison, when α is set to 0.10, the AP is 31.82%.. This can be ascribed to the large sizes of prohibited items in the PIDray dataset, requiring a larger value of α.

TABLE XIV
THE RECALL RATES (%) OF PROHIBITED ITEMS WHEN DIFFERENT NUMBERS OF PROPOSALS GENERATED BY SELECTIVE SEARCH ARE USED AS FOREGROUND ON THE FOUR X-RAY DATASETS.

N	OPIXray	HIXray	SIXray	PIDray
500	4.99	6.90	6.12	16.18
1,000	8.96	11.53	6.97	19.79
1,500	11.86	14.66	7.12	20.72
2,000	14.17	16.18	7.34	20.96

6) Influence of M in the GRGM block: We also evaluate the influence of M in the GRGM block. M is used to determine the number of generated Gaussian masks. The results are given in Table XI.

Our method achieves the best performance when the value of M is set to 2 on all the datasets. A smaller value of results in insufficient Gaussian masks, which leads to 0.2% decrease in AP on the OPIXray dataset when M is set to 0 compared with the case when M is set to 2. While a larger value of

TABLE XV
ABLATION STUDY RESULTS (%) ON THE INFLUENCE OF DIFFERENT SAMPLED SUBSETS OF THE COCO DATASET.

No.	OPIXray		HIXray		SIXray		PIDray	
	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}	AP	AP_{50}
1	13.50	42.53	30.28	75.10	25.95	65.64	29.91	51.58
2	12.45	41.16	30.47	75.35	23.90	60.82	31.33	54.56
3	14.81	48.76	**30.76**	74.62	**26.28**	65.96	31.65	55.21
4	**15.09**	**50.62**	30.62	**75.73**	26.24	**67.53**	**31.82**	**55.43**

M leads to excessive masking of critical features, resulting in 0.7% decrease in AP on the same dataset when M is increased. Both cases cause performance degradation.

7) Influence of β in the LFGM Block: We evaluate the influence of the scaling factor β in the LFGM block. The results are given in Table XII.

We can see that the model achieves the optimal performance (with the AP=15.21% and AP_{50}=51.09%) when the value of β is set to 0.10 on the OPIXray dataset. For the HIXray and SIXray datasets, the model performs the best when the value of β is set to 0.15, achieving AP_{50} of 75.73% and 67.53%. For the PIDray dataset, the best performance is achieved when the value of β is set to 0.20, achieving AP_{50} of 56.24%. The optimal value of β is different on different datasets. This is because of the different characteristics of the prohibited items in X-ray datasets. For the OPIXray dataset, the prohibited items are relatively small. Hence, using a smaller value of β helps search for local maximum activation values. Conversely, for the HIXray, SIXray, and PIDRary datasets, the prohibited items are larger. Thus, a larger value of β is required to cover a wider area. However, the performance difference is small when the value of β is in the range of [0.10, 0.20]. In this paper, we set the value of β to 0.15 in the other experiments.

8) Performance Results on Different Downstream Detectors: We respectively train different downstream prohibited item detectors using pseudo-labels generated by P2BNet or I^2OL-Net. The results are shown in Table XIII.

It can be observed that the downstream detectors of I^2OL-Net outperform those of P2BNet on all four datasets. When comparing the best performance of P2BNet with the worst performance of I^2OL-Net, the latter still achieves performance gains of 6.56%, 14.93%, 16.86%, and 17.94% in terms of AP across the four datasets. Additionally, YOLOv8-s (with fewer parameters) trained with pseudo-labels generated by I^2OL-Net achieves better performance than the detectors (such as deformable DETR with more parameters) trained with pseudo-labels generated by P2BNet. These results show the high-quality pseudo-labels given by I^2OL-Net, which effectively learns objectness information from the intra-OL and inter-OL modules.

9) Recall Rate of Selective Search on X-ray Datasets: We compute the recall rates of prohibited items on the four X-ray datasets when different numbers of proposals generated by the selective search method are used as foreground. The results are shown in Table XIV.

When 2,000 proposals are used as foreground detection results, the recall rates of prohibited items on the four X-ray datasets are higher than those using a smaller number of proposals (e.g., 500). However, even when 2,000 proposals are used as foreground, the recall rates on various datasets are still low, which is undesirable for X-ray prohibited item detection. This also explains the reasons why image-supervised prohibited item detectors fail to achieve satisfactory performance on X-ray datasets.

10) Influence of Different Sampled Subsets of the COCO Dataset: In this subsection, we evaluate the influence of different sampled subsets of the COCO dataset on the final performance. We give the results on 4 randomly selected subsets of the COCO dataset in Table XV.

Incorporating randomly sampled subsets of the COCO dataset as an extra dataset introduces some performance variations, but the overall influence remains minor. Specifically, different subsets of the COCO dataset exhibit fluctuations across metrics, such as AP ranging from 12.45% to 15.09% for OPIXray and from 23.90% to 26.28% for SIXray. However, these variations are relatively small, demonstrating the model's robustness to diverse sampled subsets. Notably, HIXray and PIDray, which have a larger number of point annotations, show even smaller variations in AP and AP_{50}. This highlights the model's increased stability when handling X-ray datasets with sufficient point annotations. Overall, our method does not rely on specific class information from the COCO dataset. As a result, the training process is not greatly influenced by the class distribution in the COCO dataset. The influence of class-wise sampling on model performance is expected to be minimal.

V. Conclusion and Future Work

In this paper, we develop a novel I^2OL-Net for point-supervised X-ray prohibited item detection, which greatly reduces the annotation cost. In I^2OL-Net, we design two key modules, an intra-OL module and an inter-OL module, to learn the objectness in X-ray images and from natural images. By transferring knowledge from natural images and leveraging wavelet decomposition, we successfully alleviate the problem of part domination. Extensive experiments on four X-ray datasets show significant performance improvements of our I^2OL-Net over existing weakly-supervised methods.

Our current experiments show that the models trained with point annotations perform worse than those trained with box annotations when the same training set is used. However, the annotation cost of point annotations is significantly reduced. As techniques evolve, the performance gap between point and box annotations is expected to narrow through algorithmic optimization and data augmentation. In particular, vision-language models (VLMs) [66] have demonstrated great effectiveness in a variety of vision tasks. In this paper, we have validated the feasibility of transferring the knowledge learned from natural images to X-ray images. Meanwhile, VLMs often learn rich vision-language relationships from web-scale image-text pairs that are from the Internet, enabling zero-shot predictions on various vision tasks. Hence, how to apply VLMs to prohibited item detection under point annotations for performance improvements merits further investigation in future work.

References

[1] P. Chen, X. Yu, X. Han, N. Hassan, K. Wang, J. Li, J. Zhao, H. Shi, Z. Han, and Q. Ye, "Point-to-box network for accurate object detection via single point supervision," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 51–67.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[3] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, "SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2119–2128.

[4] S. Akcay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within X-ray security imaging," *Pattern Recognition*, vol. 122, p. 108245, 2022.

[5] H. D. Nguyen, R. Cai, H. Zhao, A. C. Kot, and B. Wen, "Towards more efficient security inspection via deep learning: A task-driven X-ray image cropping scheme," *Micromachines*, vol. 13, no. 4, p. 565, 2022.

[6] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, "EAOD-Net: Effective anomaly object detection networks for X-ray images," *IET Image Processing*, vol. 16, no. 10, pp. 2638–2651, 2022.

[7] D. Liu, J. Liu, P. Yuan, F. Yu *et al.*, "A lightweight dangerous liquid detection method based on depthwise separable convolution for X-ray security inspection," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[8] X. Wang, X. Yang, S. Zhang, Y. Li, L. Feng, S. Fang, C. Lyu, K. Chen, and W. Zhang, "Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3240–3249.

[9] Q. Guo, Y. Mu, J. Chen, T. Wang, Y. Yu, and P. Luo, "Scale-equivalent distillation for semi-supervised object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 522–14 531.

[10] J. Zhang, X. Lin, W. Zhang, K. Wang, X. Tan, J. Han, E. Ding, J. Wang, and G. Li, "Semi-DETR: Semi-supervised object detection with detection transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 809–23 818.

[11] Z. Huang, Y. Zou, B. Kumar, and D. Huang, "Comprehensive attention self-distillation for weakly-supervised object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 797–16 807, 2020.

[12] Y. Gao, B. Liu, N. Guo, X. Ye, F. Wan, H. You, and D. Fan, "C-MIDN: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9834–9843.

[13] Y. Xu, C. Zhou, X. Yu, B. Xiao, and Y. Yang, "Pyramidal multiple instance detection network with mask guided self-correction for weakly supervised object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3029–3040, 2021.

[14] Z. Wu, C. Liu, J. Wen, Y. Xu, J. Yang, and X. Li, "Selecting high-quality proposals for weakly supervised object detection with bottom-up aggregated attention and phase-aware loss," *IEEE Transactions on Image Processing*, vol. 32, pp. 682–693, 2022.

[15] A. Diba, V. Sharma, A. Pazandeh, H. Pirsiavash, and L. Van Gool, "Weakly supervised cascaded convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 914–922.

[16] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1325–1334.

[17] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2843–2851.

[18] D. Kim, D. Cho, D. Yoo, and I. So Kweon, "Two-phase learning for weakly supervised object localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 3534–3543.

[19] J. Mai, M. Yang, and W. Luo, "Erasing integrated learning: A simple yet effective approach for weakly supervised object localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8766–8775.

[20] S. Yang, Y. Kim, Y. Kim, and C. Kim, "Combinational class activation maps for weakly supervised object localization," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2941–2949.

[21] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An X-ray security inspection benchmark and de-occlusion attention module," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 138–146.

[22] R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, and X. Liu, "Towards real-world X-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 923–10 932.

[23] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards real-world prohibited item detection: A large-scale X-ray benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5412–5421.

[24] C. Zhao, L. Zhu, S. Dou, W. Deng, and L. Wang, "Detecting overlapped objects in X-ray security imagery by a label-aware mechanism," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 998–1009, 2022.

[25] B. Ma, T. Jia, M. Li, S. Wu, H. Wang, and D. Chen, "Toward dual-view X-ray baggage inspection: A large-scale benchmark and adaptive hierarchical cross refinement for prohibited item discovery," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3866–3878, 2024.

[26] F. Yang, R. Jiang, Y. Yan, J.-H. Xue, B. Wang, and H. Wang, "Dual-mode learning for multi-dataset X-ray security image detection," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3510–3524, 2024.

[27] S. Akçay, M. E. Kundegorski, C. G. Willcocks, and T. Breckon, "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 2203–2215, 2018.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 21–37.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017, pp. 2980–2988.

[31] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9627–9636.

[32] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[33] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[34] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[35] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2846–2854.

[36] P. Tang, X. Wang, S. Bai, W. Shen, X. Bai, W. Liu, and A. Yuille, "PCL: Proposal cluster learning for weakly supervised object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 176–191, 2018.

[37] J. Seo, W. Bae, D. J. Sutherland, J. Noh, and D. Kim, "Object discovery via contrastive learning for weakly supervised object detection," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 312–329.

[38] D. P. Papadopoulos, J. R. Uijlings, F. Keller, and V. Ferrari, "Training object class detectors with click supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6374–6383.

[39] Z. Ren, Z. Yu, X. Yang, M.-Y. Liu, A. G. Schwing, and J. Kautz, "UFO^2: A unified framework towards omni-supervised object detection," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 288–313.

[40] Y. Ge, Q. Zhou, X. Wang, C. Shen, Z. Wang, and H. Li, "Point-teaching: Weakly semi-supervised object detection with point annotations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 667–675.

[41] L. Chen, T. Yang, X. Zhang, W. Zhang, and J. Sun, "Points as queries: Weakly semi-supervised object detection by points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8823–8832.

[42] S. Zhang, Z. Yu, L. Liu, X. Wang, A. Zhou, and K. Chen, "Group R-CNN for weakly semi-supervised object detection with points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9417–9426.

[43] X. Li, C. Wang, Y. Yan, J. Wu, Y. Huang, and H. Wang, "BCR-Net: Boundary-category refinement network for weakly semi-supervised X-ray prohibited item detection with points," *arXiv:2412.18918*, 2024.

[44] W. Wu, H.-S. Wong, S. Wu, and T. Zhang, "Relational matching for weakly semi-supervised oriented object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 800–27 810.

[45] J. Luo, X. Yang, Y. Yu, Q. Li, J. Yan, and Y. Li, "PointOBB: Learning oriented object detection via single point supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 730–16 740.

[46] K. Saito, Y. Ushiku, T. Harada, and K. Saenko, "Strong-weak distribution alignment for adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6956–6965.

[47] M. A. Munir, M. H. Khan, M. Sarfraz, and M. Ali, "SSAL: Synergizing between self-training and adversarial learning for domain adaptive object detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 770–22 782, 2021.

[48] Y. Pan, A. J. Ma, Y. Gao, J. Wang, and Y. Lin, "Multi-scale adversarial cross-domain detection with robust discriminative learning," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1324–1332.

[49] Z. He and L. Zhang, "Multi-adversarial Faster R-CNN for unrestricted object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6668–6677.

[50] J. Maurya, K. R. Ranipa, O. Yamaguchi, T. Shibata, and D. Kobayashi, "Domain adaptation using self-training with mixup for one-stage object detection," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 4178–4187.

[51] Z. Zhang and M. Hoai, "Object detection with self-supervised scene adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 589–21 599.

[52] F. Yu, D. Wang, Y. Chen, N. Karianakis, T. Shen, P. Yu, D. Lymberopoulos, S. Lu, W. Shi, and X. Chen, "SC-UDA: Style and content gaps aware unsupervised domain adaptation for object detection," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 382–391.

[53] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller, "Automatic adaptation of object detectors to new domains using self-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 780–790.

[54] V. VS, P. Oza, and V. M. Patel, "Towards online domain adaptive object detection," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2023, pp. 478–488.

[55] Y.-J. Li, X. Dai, C.-Y. Ma, Y.-C. Liu, K. Chen, B. Wu, Z. He, K. Kitani, and P. Vajda, "Cross-domain adaptive teacher for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7581–7590.

[56] L. Zhao and L. Wang, "Task-specific inconsistency alignment for domain adaptive object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 217–14 226.

[57] J. Deng, W. Li, Y. Chen, and L. Duan, "Unbiased mean teacher for cross-domain object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4091–4101.

[58] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014, pp. 740–755.

[59] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9036–9045.

[60] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. V. Gool, "Domain adaptive Faster R-CNN for object detection in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2018, pp. 3339–3348.

[61] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[62] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[63] Z. Zeng, B. Liu, J. Fu, H. Chao, and L. Zhang, "WSOD2: Learning bottom-up and top-down objectness distillation for weakly-supervised object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8292–8300.

[64] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 770–778.

[66] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.