Dual-Mode Learning for Multi-Dataset X-Ray Security Image Detection

Fenghong Yang, Runqing Jiang, Yan Yan, *Senior Member, IEEE,* Jing-Hao Xue, *Senior Member, IEEE,* Biao Wang, and Hanzi Wang, *Senior Member, IEEE*

***Abstract*—With the recent advance of deep learning, a large number of methods have been developed for prohibited item detection in X-ray security images. Generally, these methods train models on a single X-ray image dataset that may contain only limited categories of prohibited items. To detect more prohibited items, it is desirable to train a model on the multi-dataset that is constructed by combining multiple datasets. However, directly applying existing methods to the multi-dataset cannot guarantee good performance because of the large domain discrepancy between datasets and the occlusion in images. To address the above problems, we propose a novel Dual-Mode Learning Network (DML-Net) to effectively detect all the prohibited items in the multi-dataset. In particular, we develop an enhanced RetinaNet as the architecture of DML-Net, where we introduce a lattice appearance enhanced sub-net to enhance appearance representations. Such a way benefits the detection of occluded prohibited items. Based on the enhanced RetinaNet, the learning process of DML-Net involves both common mode learning (detecting the common prohibited items across datasets) and unique mode learning (detecting the unique prohibited items in each dataset). For common mode learning, we introduce an adversarial prototype alignment module to align the feature prototypes from different datasets in the domain-invariant feature space. For unique mode learning, we take advantage of feature distillation to enforce the student model to mimic the features extracted by multiple pre-trained teacher models. By tightly combining and jointly training the dual modes, our DML-Net method successfully eliminates the domain discrepancy and exhibits superior model capacity on the multi-dataset. Extensive experimental results on several combined X-ray image datasets demonstrate the effectiveness of our method against several state-of-the-art methods. Our code is available at https://github.com/vampirename/dmlnet.**

***Index Terms*—X-ray security image detection, domain discrepancy, occlusion, feature distillation, multi-dataset learning.**

I. Introduction

With the popularity of public transportation, security inspection plays a critical role in protecting public safety. Security inspection usually adopts X-ray scanners, based on which the security inspectors can quickly identify the prohibited items in the passenger luggage. However, security inspectors may struggle to accurately detect all the prohibited items after long-term observation of extensive X-ray security images without distraction. Although changing shifts can alleviate this problem, the high cost of human resources is still not desirable.

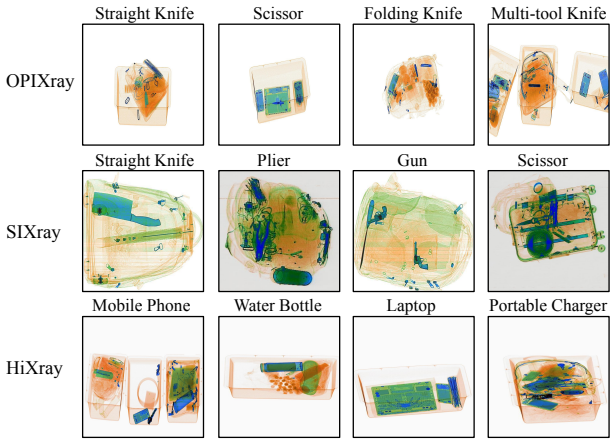

Fig. 1: Samples of the representative categories of prohibited items in OPIXray [4], SIXray [7], and HiXray [8]. There are some common prohibited items (such as straight knife and scissor) across the OPIXray and SIXray datasets and some unique prohibited items in each dataset. Some datasets (e.g., SIXray and HiXray) show large differences in color imaging.

To reduce expensive labor costs, a prevailing strategy in X-ray security checks is to leverage automatic detection techniques to assist security inspectors. Over the past few years, automatic prohibited item detection in X-ray security images (also called X-ray security image detection) has attracted considerable attention. Recently, much progress has been made in X-ray security image detection due to the rapid development of deep learning. Many efforts [1]–[6] have been devoted to designing dedicated models to boost the performance. Wei *et al.* [4] develop a De-Occlusion Attention Module (DOAM), which applies attention maps generated by the appearance information of prohibited items to address heavy occlusion in X-ray images. Zhao *et al.* [5] propose to identify overlapped objects in high-level feature maps.

In practical applications, different transportation hubs (such as airports and subways) often use different types of X-ray scanners. Moreover, each transportation hub is concerned with

This work was partly supported by the National Natural Science Foundation of China under Grants 62372388, 62071404, U21A20514, and 61872307, by the Open Research Projects of Zhejiang Lab under Grant 2021KG0AB02. (*Corresponding author: Yan Yan.*)

F. Yang, R. Jiang, Y. Yan, H. Wang are with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: yangfh@stu.xmu.edu.cn; jiangrunqing@stu.xmu.edu.cn; yanyan@xmu.edu.cn; hanzi.wang@xmu.edu.cn).

F. Yang and R. Jiang have contributed equally to this work.

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

B. Wang is with the Zhejiang Lab, Hanzhou 311101, China (e-mail: wangbiao@zhejianglab.com).

specific threat profiles. For airports, there is a broad range of prohibited items, especially considering international flights for the security threat. For subways, the primary prohibited items are improvised explosive devices and weapons. Therefore, the data from different hubs can significantly differ. Accordingly, many public X-ray image datasets [4], [7]–[11] are often collected under different conditions and target at detecting different categories of prohibited items. For example, the OPIXray dataset [4] consists of different types of cutters (e.g., straight knife, scissor, folding knife, and multi-tool knife). The SIXray dataset [7] contains five classes of prohibited items (e.g., straight knife, plier, gun, and scissor). The HiXray dataset [8] includes seven classes of prohibited items (e.g., mobile phone, water bottle, laptop, and portable charger). Some samples in these datasets are shown in Fig. 1.

The above X-ray image datasets involve only limited categories of prohibited items. In real-world X-ray security checks, it is preferable to detect as many prohibited items as possible. Notably, considering real-world scenarios such as airport security checks, where a list of prohibited items is essential, an ideal X-ray security inspection system should recognize all the items on this list. Therefore, it is necessary to construct a multi-dataset by combining multiple X-ray image datasets, thereby enlarging the categories of prohibited items.

Based on the multi-dataset, we can directly apply existing X-ray security image detection methods. Unfortunately, the domain discrepancy between datasets can be large since different datasets are captured by different X-ray scanners (which have significant differences in color imaging, as illustrated in Fig. 1). In addition, the occlusion in X-ray security images can be severe, hindering the extraction of target-specific features. Note that the above two problems are closely related (i.e., alleviating the occlusion problem in X-ray security images is greatly helpful for addressing the problem of domain discrepancy between datasets). As a result, existing methods cannot learn effective models and achieve satisfactory detection accuracy on the multi-dataset due to the problems of domain discrepancy and occlusion. Hence, it is important to develop a universal model that can adapt to multiple datasets and is robust to occlusion.

Till now, some works investigate the problem of multi-dataset object detection. For instance, Chen *et al*. [12] use variational attention to propagate domain-specific knowledge for multi-dataset learning in crowd counting. Zhou *et al*. [13] propose to train a universal object detector on the multi-dataset via dataset-specific training protocols and losses based on a shared backbone.

Generally, conventional multi-dataset object detection methods mainly work on natural images. X-ray security image detection is intrinsically different from natural image detection on the multi-dataset. On the one hand, natural image detection tries to detect all the potential objects, while X-ray security image detection aims to identify only the prohibited items. As a result, the background clutters or heavy occlusion may greatly influence the X-ray security image detection performance. On the other hand, most multi-dataset object detection methods focus on addressing the label inconsistency problem [13]. In contrast, multi-dataset X-ray security image detection intends to detect the common prohibited items across datasets and the unique prohibited items in each dataset. These differences necessitate the development of different learning methods for multi-dataset X-ray security image detection.

In this paper, we develop a novel Dual-Mode Learning Network (DML-Net) for multi-dataset X-ray security image detection. The architecture of DML-Net is based on a novel enhanced RetinaNet, where we design a Lattice Appearance Enhanced sub-net (LAE) to enhance appearance representations. This benefits the detection of occluded prohibited items that are ubiquitous in X-ray security images.

The learning process of DML-Net involves dual-mode learning leveraging both common mode learning and unique mode learning, which are developed to detect the common prohibited items across datasets and the unique prohibited items in each dataset, respectively. For common mode learning, we design an adversarial prototype alignment module to align the feature prototypes from different datasets in the domain-invariant feature space. Meanwhile, for unique mode learning, we adopt feature distillation to enforce the student model to learn from multiple teachers pre-trained on individual datasets. Based on the above designs, our method effectively alleviates the domain discrepancy between datasets while largely reducing the negative impact of occlusion in X-ray security images.

In summary, our main contributions are given as follows:

- We propose DML-Net to accurately learn a universal prohibited item detector on the multi-dataset. In DML-Net, we develop dual-mode learning, consisting of common mode learning and unique mode learning, to address the domain discrepancy and category differences between datasets. To the best of our knowledge, we are the first to investigate multi-dataset X-ray security image detection.
- We design LAE to enhance the feature representations of X-ray security images. In particular, LAE explores the potential of rich combinations of edge and texture feature maps, enlarging the representation space of the model. In this way, the occlusion problem can be greatly relieved, facilitating dual-mode learning.
- We extensively evaluate DML-Net on different combinations of popular X-ray image datasets. Without any whistles and bells, DML-Net consistently outperforms several state-of-the-art methods. This clearly demonstrates the superiority of our method on the multi-dataset.

The remainder of this paper is organized as follows. First, we briefly review the related work in Sec. II. Then, we present the details of our proposed DML-Net method in Sec. III. Next, we evaluate the performance of DML-Net and compare DML-Net with several state-of-the-art methods in Sec. IV. Finally, we conclude our work in Sec. V.

II. Related Work

In this section, we briefly review deep learning-based methods for X-ray security image detection and multi-dataset object detection, which are closely related to our method.

A. X-Ray Security Image Detection

With the rapid development of deep learning, a large number of X-ray security image detection methods [4], [5], [14]–[18]

have been proposed to learn effective feature representations of prohibited items and improve the detection accuracy. Jaccard *et al.* [14] first develop a deep learning scheme for the detection of small metallic threats in X-ray cargo images and validate the superior performance of Convolutional Neural Networks (CNNs) over traditional methods. Later, Akcay *et al.* [15] introduce CNN to detect prohibited items in X-ray security images. They employ a transfer learning paradigm, where the detector is first pre-trained on the natural image classification dataset and then fine-tuned on the domain-specific X-ray image dataset.

Due to the characteristics of security inspection, occlusion, which is prevalent in X-ray security images, poses a great challenge for detection. Wei *et al.* [4] propose DOAM, which combines different appearance information of prohibited items to generate an attention map, to address the heavy occlusion problem. Zhao *et al.* [5] introduce a label-aware mechanism to tackle the object overlapping problem. This mechanism models the relationship between feature channels and labels, and then refines the features according to the assigned labels. Wang *et al.* [16] present a selective dense attention network to detect the prohibited items hidden in messy objects. A dense attention module and a dependency refinement module are introduced to extract discriminative features. Wang *et al.* [18] propose a material-aware cross-channel interaction attention module, which takes advantage of the material information to handle the inter-class occlusion.

Existing X-ray security image detectors are usually designed based on the general object detectors (such as YOLO series [19], RetinaNet [20], and FCOS [21]) on a single X-ray image dataset. However, many existing X-ray image datasets [4], [7], [8] involve only limited categories of prohibited items because of different capturing conditions and objectives. This heavily prevents these X-ray security image detection methods from real-world applications, which require to detect various prohibited items. Moreover, simply applying these methods to a multi-dataset cannot achieve satisfactory results due to the significant domain discrepancy between datasets [22], [23] and the severe occlusion in images. In this paper, we are concerned with the little-studied but important task of training a universal detector to effectively identify all the prohibited items in the multi-dataset. Hence, our method can easily adapt to the multi-dataset captured from various scenarios in practical applications.

B. Multi-Dataset Object Detection

Multi-dataset learning, which aims to learn a universal model from multiple datasets, has received increasing attention in various computer vision tasks, including depth estimation [24]–[26], stereo matching [27], [28], pedestrian detection [29], [30], semantic segmentation [31], [32], and object detection [33]–[37]. In this subsection, we mainly review multi-dataset object detection.

To perform multi-dataset object detection, it is common practice to merge different semantic classes across datasets. Perrett *et al.* [33] concatenate labels from different datasets. However, they do not explicitly consider the domain discrepancy and category differences between datasets. Later, Wang *et al.* [38] design a universal object detector, which is capable of operating over multiple domains. They propose a new family of adaptation layers to compensate for domain shift. However, such a detector works only on small datasets and does not fully model the semantic relationship between datasets. Yao *et al.* [34] develop the dataset-aware losses for multi-dataset training. Zhou *et al.* [13] propose to train a detector on multiple large-scale datasets. They integrate dataset-specific outputs into a common semantic taxonomy. But this method requires dataset-specific training protocols, making it difficult to apply to the X-ray image datasets. Wang *et al.* [39] propose UniDectector to recognize a large number of categories in the open world without any finetuning. Chen *et al.* [40] introduce a scalable multi-dataset detector (ScaleNet) to learn across multiple datasets in a unified semantic label space.

Generally, the above methods focus on multi-dataset object detection on natural images. X-ray security images are significantly different from natural images due to different sensing techniques. In addition, X-ray image detection targets identifying only the prohibited items while natural image detection aims to detect all the objects of interest. Therefore, directly employing existing multi-dataset object detection methods on X-ray security images cannot guarantee desirable performance. In particular, the relationship between the common mode and the unique mode, which is critical for multi-dataset X-ray security image detection, is not well modeled in existing methods. In this paper, we develop an effective multi-dataset object detection method tailored for X-ray security images.

III. Methodology

In this section, we first give an overview of our proposed DML-Net in Section III-A. Then, we describe the key components of DML-Net from Section III-B to Section III-E. Finally, we summarize the overall training of our method in Section III-F.

A. Overview

Suppose that the multi-dataset is represented as $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K\}$, where $\mathcal{D}_k$ denotes the k-th X-ray image dataset and K is the total number of datasets. In this paper, we suppose the requisite number of datasets is at least two (i.e., $K \geq 2$). For the k-th dataset $\mathcal{D}_k$, its label space is represented as $\mathcal{L}_k$, which consists of the labels for the common prohibited items across datasets and the labels for the unique prohibited items in $\mathcal{D}_k$. Given a mini-batch $\mathcal{B}_k$ that is randomly selected from $\mathcal{D}_k$, an input image in $\mathcal{B}_k$ is denoted as $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and channel number of the input image, respectively.

Due to the large domain discrepancy between datasets and the occlusion in images, naively training an existing X-ray security image detection method on the multi-dataset may lead to poor performance. To address the problems of domain discrepancy and occlusion, we develop DML-Net to detect all the prohibited items in the multi-dataset. The overview of DML-Net is given in Fig. 2. Specifically, based on RetinaNet [20], we develop an enhanced RetinaNet as the backbone,

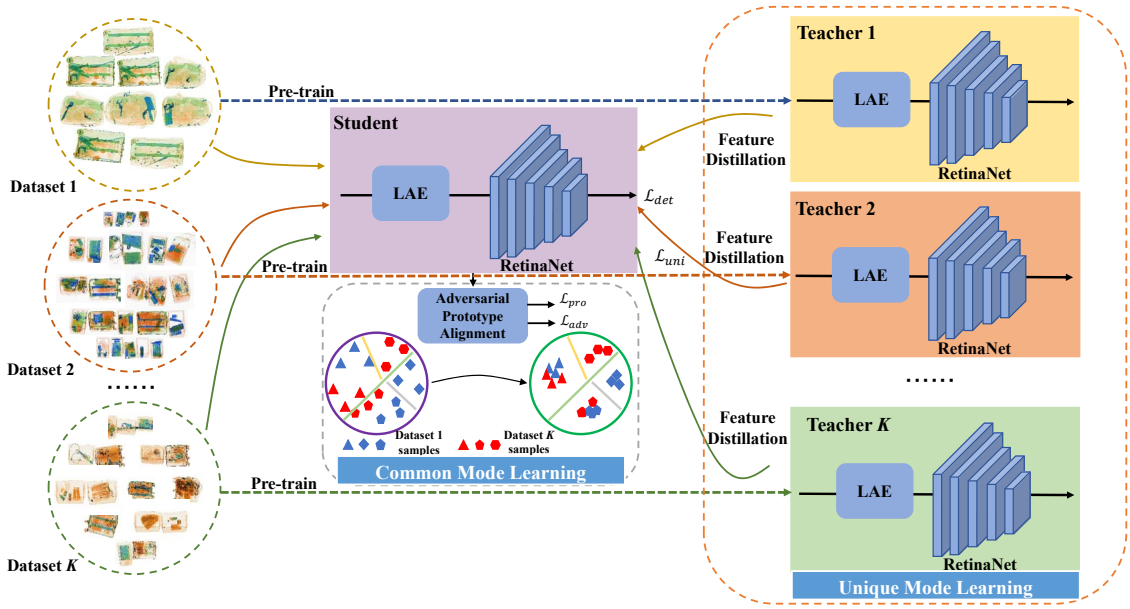

Fig. 2: Overview of DML-Net for multi-dataset X-ray security image detection. The architecture of DML-Net is based on a novel enhanced RetinaNet. The learning process of DML-Net leverages dual-mode learning: common mode learning and unique mode learning. For common mode learning, APA is developed for feature alignment in the domain-invariant feature space. For unique mode learning, feature distillation is used to enforce the student model to learn from multiple teacher models.

where we design and incorporate a Lattice Appearance Enhanced sub-net (LAE) into RetinaNet to tackle the occlusion problem in X-ray security images. Based on the enhanced RetinaNet, the overall learning process of DML-Net involves common mode learning and unique mode learning, which are proposed to detect the common and unique prohibited items, respectively. For common mode learning, we introduce an Adversarial Prototype Alignment module (APA) to align the feature prototypes from different datasets in the domain-invariant feature space. For unique mode learning, we leverage feature distillation to enforce the student model to learn from multiple teacher models. Here, each teacher model is pre-trained on a single dataset.

Overall, DML-Net is a simple yet effective method for multi-dataset X-ray security image detection. On the one hand, DML-Net introduces LAE to mitigate the occlusion problem in X-ray security images. Such a way largely facilitates dual-mode learning and thus benefits the detection task. On the other hand, DML-Net develops dual-mode learning to explicitly consider the domain discrepancy and category differences between datasets. Thus, our method effectively exploits the knowledge from different X-ray image datasets to train a universal detector.

In the following, we introduce the enhanced RetinaNet, common mode learning, unique mode learning, the total loss, and the overall training of our method.

B. Enhanced RetinaNet

1) RetinaNet: RetinaNet [20] is a mainstream one-stage object detection method, which not only shows impressive performance for general object detection but also maintains high inference efficiency. Notably, RetinaNet introduces a focal loss to handle the class imbalance problem by emphasizing the importance of hard examples.

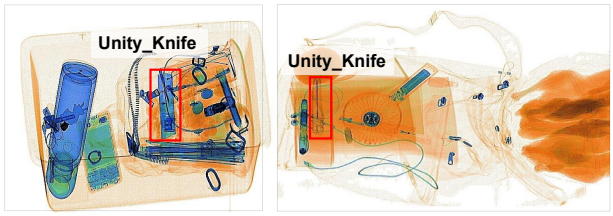

Fig. 3: Two occluded training images in the OPIXray dataset.

Generally speaking, RetinaNet involves three components: a base network and two task-specific subnetworks. The base network involves a ResNet architecture to extract features at different stages, and a Feature Pyramid Network (FPN) to extract a multi-scale convolutional feature pyramid with a top-down pathway and lateral connections. Specifically, given an image $\mathbf{X} \in \mathcal{B}_k$, we denote the features extracted by the last four residual blocks of ResNet as $C_1(\mathbf{X})$, $C_2(\mathbf{X})$, $C_3(\mathbf{X})$, and $C_4(\mathbf{X})$ (corresponding to the outputs of conv2, conv3, conv4, and conv5 in ResNet, respectively), while we define the corresponding multi-scale features extracted by FPN as $\mathcal{P} = \{P_l(\mathbf{X})\}_{l=1}^{M}$. Here, M (M=4) is the number of multi-scale features. Meanwhile, the two task-specific subnetworks are designed to perform object classification and bounding box regression in a convolutional fashion.

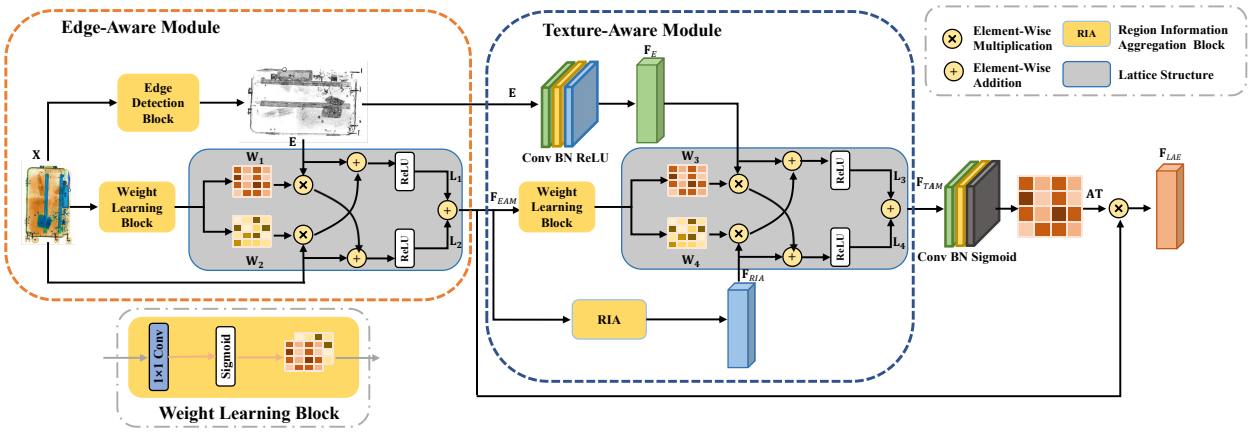

Fig. 4: The network architecture of LAE. In the figure, 'Conv' and 'BN' denote a convolutional layer and a batch normalization layer, respectively. 'ReLU' and 'Sigmoid' denote a ReLU activation function and a Sigmoid function, respectively.

2) Lattice Appearance Enhanced Sub-Net (LAE): In real-world applications, prohibited items are frequently overlapped by other objects. As a result, the appearance of prohibited items is not prominent due to occlusion. This results in the poor detection performance of traditional object detection methods since the feature representations of prohibited items are inferior. Some training examples are illustrated in Fig. 3. We can see that the appearance of prohibited items is greatly affected due to occlusion. In this paper, inspired by the lattice filter [41], we develop LAE to effectively enhance the appearance information of prohibited items, greatly alleviating the occlusion problem in X-ray security images and thus promoting dual-mode learning.

The network architecture of LAE is shown in Fig. 4. LAE mainly consists of an edge-aware module, a texture-aware module, and a convolutional block to enhance appearance representations from the perspectives of both edge and texture.

Technically, the input image $\mathbf{X}$ is first fed into the edge-aware module, which contains an edge detection block, a weight learning block, and a lattice structure. The edge detection block adopts the Sobel operator to generate two edge images ($\mathbf{E}_h$ and $\mathbf{E}_v$) along the horizontal and vertical directions, respectively. Then, the edge images $\mathbf{E}_h$ and $\mathbf{E}_v$ are combined to obtain the final edge image $\mathbf{E}$. Next, we fuse the original image $\mathbf{X}$ and its corresponding edge image $\mathbf{E}$ via a lattice structure. At the same time, the weight learning block, which contains a 1×1 convolutional layer followed by a Sigmoid function, takes the X-ray image as the input and adaptively generates two weighting tensors (denoted as $\mathbf{W}_1 \in \mathbb{R}^{H\times W\times 1}$ and $\mathbf{W}_2 \in \mathbb{R}^{H\times W\times 1}$) for the lattice structure. Note that the lattice structure allows the potential of various combinations of inputs [41], greatly enlarging the representation space of the model in an efficient manner. Mathematically, the above process can be formulated as

$$\begin{aligned} \mathbf{L}_1 &= \sigma(\mathbf{E} \oplus \mathbf{X} \otimes \phi(\mathbf{W}_2)), \\ \mathbf{L}_2 &= \sigma(\mathbf{X} \oplus \mathbf{E} \otimes \phi(\mathbf{W}_1)), \end{aligned} \tag{1}$$

where $\mathbf{L}_1 \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{L}_2 \in \mathbb{R}^{H\times W\times C}$ denote the output feature maps of the lattice structure in the edge-aware module; $\sigma(\cdot)$ represents the nonlinear ReLU activation function; '$\otimes$' means the element-wise multiplication operation; '$\oplus$' indicates the element-wise addition operation; $\phi(\cdot)$ represents the broadcast operation, where the weights are broadcast along the channel dimension.

The output of the edge-aware module is given as

$$\mathbf{F}_{EAM} = \mathbf{L}_1 \oplus \mathbf{L}_2, \tag{2}$$

where $\mathbf{F}_{EAM} \in \mathbb{R}^{H\times W\times C}$ represents the enhanced edge feature map.

From the edge-aware module, both $\mathbf{F}_{EAM}$ and $\mathbf{E}$ pass through the texture-aware module, which mainly involves a Region Information Aggregation (RIA) block [42], a weight learning block, and a lattice structure, to effectively improve the texture representations. On the one hand, the feature map $\mathbf{F}_{EAM}$ generated from the edge-aware module is first fed into a RIA block (consisting of an average pooling operation and an extension operation) to aggregate the local information and generate a texture feature map $\mathbf{F}_{RIA}$. Meanwhile, $\mathbf{F}_{EAM}$ is also used to generate two weighting tensors (denoted as $\mathbf{W}_3 \in \mathbb{R}^{H\times W\times 1}$ and $\mathbf{W}_4 \in \mathbb{R}^{H\times W\times 1}$) by the weight learning block. On the other hand, the edge image $\mathbf{E}$ is fed into a convolutional block (including a 3×3 convolutional layer, a batch normalization layer, and a ReLU activation function) to extract the edge feature map $\mathbf{F}_E$. Subsequently, the two feature maps are combined in the lattice structure. The above process can be given as

$$\begin{aligned} \mathbf{L}_3 &= \sigma(\mathbf{F}_E \oplus \mathbf{F}_{RIA} \otimes \phi(\mathbf{W}_4)), \\ \mathbf{L}_4 &= \sigma(\mathbf{F}_{RIA} \oplus \mathbf{F}_E \otimes \phi(\mathbf{W}_3)), \end{aligned} \tag{3}$$

where $\mathbf{L}_3 \in \mathbb{R}^{H\times W\times C}$ and $\mathbf{L}_4 \in \mathbb{R}^{H\times W\times C}$ denote the output feature maps of the lattice structure in the texture-aware module.

The output of the texture-aware module is given as

$$\mathbf{F}_{TAM} = \mathbf{L}_3 \oplus \mathbf{L}_4, \tag{4}$$

where $\mathbf{F}_{TAM} \in \mathbb{R}^{H\times W\times C}$ represents the enhanced texture feature map.

Unlike natural image detection which relies heavily on both edge and texture cues, X-ray security image detection is more dependent on the edge than the texture of the X-ray images. Based on this observation, instead of directly combining the enhanced edge and texture feature maps, we leverage the enhanced texture feature map to generate an attention map and impose the attention map on the enhanced edge feature map. Therefore, we employ a convolutional block (consisting of a 1×1 convolutional layer, a batch normalization layer, and a Sigmoid function) on $\mathbf{F}_{TAM}$ and obtain an attention map $\mathbf{AT} \in \mathbb{R}^{H\times W\times C}$. That is,

$$\mathbf{AT} = \tau(\mathrm{BN}(\mathrm{Conv}(\mathbf{F}_{TAM}))), \tag{5}$$

where $\tau(\cdot)$, $\mathrm{BN}(\cdot)$, and $\mathrm{Conv}(\cdot)$ represent the Sigmoid function, the batch normalization layer, and the convolutional layer, respectively.

Finally, the output of LAE is formulated as

$$\mathbf{F}_{LAE} = \mathbf{AT} \otimes \mathbf{F}_{EAM}, \tag{6}$$

where $\mathbf{F}_{LAE} \in \mathbb{R}^{H\times W\times C}$ is the final feature map, which is used as the input of RetinaNet.

Note that DOAM developed in [42] also leverages an RIA block and the Sobel edge detector to enhance feature representations. However, the differences between LAE and DOAM are significant. We specifically introduce two lattice structures in LAE to largely enlarge the representation space of the model, where the feature maps from different branches can boost the representations of each other in a mutual way. Moreover, we adaptively fuse the information from the original image and the edge image (or the texture feature map and the edge feature map) based on a weight learning block. Such a way can emphasize relevant information and suppress irrelevant information in the feature maps. Note that DML-Net is optimized by minimizing the detection loss, enabling LAE to enhance the appearance representations of prohibited items. Therefore, compared with DOAM, LAE has better feature learning capability. In a word, the design of LAE is beneficial for addressing the occlusion problem, thus facilitating effective dual-mode learning on the multi-dataset.

C. Common Mode Learning

As we mentioned previously, there exists large domain discrepancy between X-ray image datasets. For example, the color imaging in the SIXray dataset and the OPIXray/HiXray datasets is substantially different (see Fig. 1). Hence, prohibited items belonging to the same category may have essentially different feature distributions across different datasets. As a result, naively training models on the multi-dataset cannot achieve promising results for these common prohibited items.

In this subsection, we develop APA to align the feature prototypes from different datasets in the domain-invariant feature space. To achieve this, we take advantage of adversarial learning to obtain a domain-invariant feature space, based on which we can explicitly align the feature prototypes of common prohibited items across datasets. In this way, the common prohibited items from different datasets are enforced to have similar feature distributions. Therefore, we can largely reduce the domain discrepancy, thus benefiting the detection of the common prohibited items in multiple datasets.

More specifically, we first play an adversarial game between a feature extractor f_{Θ_G} (i.e., the enhanced RetinaNet) and a domain discriminator f_{Θ_D}. Here, the domain discriminator f_{Θ_D} (consisting of three fully-connected layers and a Sigmoid function) is designed to estimate the probability of a sample $\mathbf{X}$ coming from the mini-batch $\mathcal{B}_k$. f_{Θ_D} takes the feature map generated by the feature extractor f_{Θ_G} as the input and outputs a probability distribution vector (i.e., f_{Θ_D} classifies the input feature map into one of K classes). By adversarial learning, we can extract domain-invariant features by confusing the domain discriminator. The adversarial loss in $\mathcal{B}_k$ is defined as

$$(f_{\Theta_D}, f_{\Theta_G}) = \min_{f_{\Theta_D}} \max_{f_{\Theta_G}} \mathcal{L}^k_{adv}(f_{\Theta_D}, f_{\Theta_G}), \tag{7}$$

where

$$\mathcal{L}^k_{adv}(f_{\Theta_D}, f_{\Theta_G}) = \frac{1}{B} \sum_{\mathbf{X}\in\mathcal{B}_k} \boldsymbol{y}^{\mathrm{T}} \log f_{\Theta_D}(f_{\Theta_G}(\mathbf{X})), \tag{8}$$

and $\boldsymbol{y}$ is a one-hot vector; $\boldsymbol{y}[i] = 1$ if $i = k$, and 0 otherwise; B denotes the size of the mini-batch $\mathcal{B}_k$.

To facilitate model training, a Gradient Reversal Layer (GRL) [43] is used between the feature extractor and the domain discriminator. Note that GRL has no learning parameters, and can serve as a simple identity function during the forward propagation while reversing the sign of the passing gradient in the back-propagation. Therefore, based on GRL, a domain-invariant feature space can be effectively learned.

Adversarial learning is leveraged to reduce the domain discrepancy. However, the distances between the same-category samples from different datasets can still be large, increasing the difficulty of identifying the common prohibited items in the multi-dataset. Therefore, we further adopt prototype alignment to explicitly minimize the feature prototype distances between common prohibited items across different datasets. Mathematically, for a category of common prohibited item $c \in \mathcal{L}_k$ in the mini-batch $\mathcal{B}_k$, we iteratively calculate the prototype in an average-moving manner [44],

$$\mathbf{p}_{k,c} \leftarrow \alpha \mathbf{p}_{k,c} + (1-\alpha)\mathbf{q}_{\mathbf{X},c}, \tag{9}$$

where $\mathbf{p}_{k,c}$ denotes the prototype of the prohibited item from the c-th category in $\mathcal{B}_k$; $\mathbf{q}_{\mathbf{X},c}$ denotes the feature map at the spatial position indicated by the anchor (which corresponds to the ground-truth bounding box with the label c in the image $\mathbf{X}$; α is a momentum coefficient (which is set as 0.999 in all our experiments).

Based on the above, we define the prototype alignment loss as

$$\mathcal{L}_{pro} = \frac{1}{K(K-1)C} \sum_{k_1=1}^{K} \sum_{k_2=1,k_2\neq k_1}^{K} \sum_{c\in\mathcal{L}_{k_1}\cap\mathcal{L}_{k_2}} ||\mathbf{p}_{k_1,c} - \mathbf{p}_{k_2,c}||_{\mathrm{F}}, \tag{10}$$

where C is the category number of common prohibited items and $||\cdot||_{\mathrm{F}}$ represents the Frobenius norm.

By minimizing the prototype alignment loss $\mathcal{L}_{pro}$, the feature distributions of prohibited items from the same category are aligned across datasets. Thus, the domain discrepancy

between different datasets is further reduced and the discrimination between common prohibited items is enhanced.

The detection loss $\mathcal{L}^k_{det\text{-}com}$ for the common prohibited items in $\mathcal{B}_k$ is defined as

$$\mathcal{L}^k_{det\text{-}com} = \frac{1}{B|\mathcal{C}_{\mathbf{X}}|}\sum_{\mathbf{X}\in\mathcal{B}_k}\sum_{i=1}^{|\mathcal{C}_{\mathbf{X}}|}\mathcal{L}_{cls}(f_{\Theta_C}(\mathbf{C}^i_{\mathbf{X}}), y^i_{com}) + \mathcal{L}_{reg}(f_{\Theta_R}(\mathbf{C}^i_{\mathbf{X}}), b^i_{com}), \tag{11}$$

where $\mathcal{C}_{\mathbf{X}}$ denotes the anchor bag (which includes the top few anchors according to the IoUs between the anchors and the ground-truth boxes) of common prohibited items in the image $\mathbf{X}$; $\mathbf{C}^i_{\mathbf{X}}$ represents the i-th anchor in $\mathcal{C}_{\mathbf{X}}$; $|\mathcal{C}_{\mathbf{X}}|$ represents the number of anchors; $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ denote the focal loss and the smooth L_1 loss, respectively; $f_{\Theta_C}(\cdot)$ and $f_{\Theta_R}(\cdot)$ denote the classification subnetwork and the regression subnetwork, respectively; y^i_{com} and b^i_{com} respectively represent the anchor label and the ground-truth bounding box w.r.t. $\mathbf{C}^i_{\mathbf{X}}$.

D. Unique Mode Learning

Generally, each X-ray image dataset often involves its unique prohibited items. To train a universal detector on the multi-dataset, we also develop unique mode learning to detect the unique prohibited items in each dataset.

Suppose that we have a set of teacher models $\{\mathcal{M}^{(1)}_t, \mathcal{M}^{(2)}_t, ..., \mathcal{M}^{(K)}_t\}$ and a student model $\mathcal{M}_s$. Here, $\mathcal{M}^{(k)}_t$ represents the k-th teacher model, which is pre-trained to detect the unique prohibited items in the dataset $\mathcal{D}_k$. Both the teacher and student models adopt the enhanced RetinaNet as the network architecture. As indicated in [45], intermediate feature representations from the teacher models provide rich information to improve the training of the student model. Inspired by the above observations, we take advantage of multi-scale feature distillation to effectively transfer the knowledge from multiple teachers to the student.

Technically, given an image $\mathbf{X} \in \mathcal{B}_k$, we first feed $\mathbf{X}$ into $\mathcal{M}_s$ and $\mathcal{M}^{(k)}_t$, and obtain a set of multi-scale features $\{P^{\mathcal{M}_s}_l(\mathbf{X})\}^M_{l=1}$ and $\{P^{\mathcal{M}^{(k)}_t}_l(\mathbf{X})\}^M_{l=1}$ from FPN. Then, we enforce the student model $\mathcal{M}_s$ to mimic the feature maps obtained by its corresponding teacher model $\mathcal{M}^{(k)}_t$ pre-trained on $\mathcal{D}_k$. Therefore, the feature distillation loss $\mathcal{L}^k_{uni}$ in $\mathcal{B}_k$ is defined as

$$\mathcal{L}^k_{uni} = \frac{1}{BM}\sum_{\mathbf{X}\in\mathcal{B}_k}\sum_{l=1}^{M}||\mathrm{T}_l(P^{\mathcal{M}_s}_l(\mathbf{X})) - P^{\mathcal{M}^{(k)}_t}_l(\mathbf{X})||_{\mathrm{F}}, \tag{12}$$

where $\mathrm{T}_l(\cdot)$ denotes the feature transformation operation (we use the up-sampling operation) that resizes the feature map of the student model to that of the teacher model at the l-th layer and $||\cdot||_{\mathrm{F}}$ denotes the Frobenius norm.

By optimizing $\mathcal{L}^k_{uni}$, the rich information in the multi-scale feature maps can be transferred from the k-th teacher model to the student model. This is helpful in detecting the unique prohibited items in the k-th dataset.

The detection loss $\mathcal{L}^k_{det\text{-}uni}$ for the unique prohibited items in $\mathcal{B}_k$ is defined as

$$\mathcal{L}^k_{det\text{-}uni} = \frac{1}{B|\mathcal{U}_{\mathbf{X}}|}\sum_{\mathbf{X}\in\mathcal{B}_k}\sum_{i=1}^{|\mathcal{U}_{\mathbf{X}}|}\mathcal{L}_{cls}(f_{\Theta_C}(\mathbf{U}^i_{\mathbf{X}}), y^i_{uni}) + \mathcal{L}_{reg}(f_{\Theta_R}(\mathbf{U}^i_{\mathbf{X}}), b^i_{uni}), \tag{13}$$

where $\mathcal{U}_{\mathbf{X}}$ denotes the anchor bag of unique prohibited items in the image $\mathbf{X}$; $\mathbf{U}^i_{\mathbf{X}}$ represents the i-th anchor in $\mathcal{U}_{\mathbf{X}}$; $|\mathcal{U}_{\mathbf{X}}|$ represents the number of anchors; y^i_{uni} and b^i_{uni} respectively represent the anchor label and the ground-truth bounding box w.r.t. $\mathbf{U}^i_{\mathbf{X}}$.

E. Total Loss

Based on the above formulations, the total loss of DML-Net is given as

$$\mathcal{L}_{joint} = \mathcal{L}_{det} + \lambda_1\mathcal{L}_{pro} + \lambda_2\mathcal{L}_{adv} + \lambda_3\mathcal{L}_{uni}, \tag{14}$$

where $\mathcal{L}_{det} = \mathcal{L}_{det\text{-}com} + \mathcal{L}_{det\text{-}uni}$ denotes the joint detection loss, in which $\mathcal{L}_{det\text{-}com} = 1/K\sum_{k=1}^{K}\mathcal{L}^k_{det\text{-}com}$ and $\mathcal{L}_{det\text{-}uni} = 1/K\sum_{k=1}^{K}\mathcal{L}^k_{det\text{-}uni}$ denote the detection losses for the common prohibited items and the unique prohibited items, respectively; $\mathcal{L}_{adv} = 1/K\sum_{k=1}^{K}\mathcal{L}^k_{adv}$ and $\mathcal{L}_{uni} = 1/K\sum_{k=1}^{K}\mathcal{L}^k_{uni}$ are the joint adversarial loss and the joint feature distillation loss, respectively; λ_1, λ_2, and λ_3 are the weighting parameters.

In the total loss, we leverage multiple loss items (a joint detection loss (containing the detection losses for common prohibited items and unique prohibited items), a prototype alignment loss, a joint adversarial loss, and a joint feature distillation loss) to optimize the model. Every loss item is important to ensure the excellent performance for the challenging multi-dataset detection task. In particular, the joint detection loss is optimized to localize both the common and unique prohibited items. The joint adversarial loss is leveraged to reduce the domain discrepancy. The prototype alignment loss is designed to optimize the distances between the same category samples from different datasets. Both the joint adversarial loss and the prototype alignment loss are designed for common mode learning. Finally, the joint feature distillation loss is employed to transfer the knowledge from multiple teacher models to the student model for unique mode learning. By minimizing the total loss, DML-Net can effectively learn discriminative feature representations for prohibited item detection on the multi-dataset.

F. Overall Training

We give the overall training procedure of DML-Net in Algorithm 1. Generally, the learning process of DML-Net contains both common mode learning and unique mode learning based on the enhanced RetinaNet. Common mode learning focuses on the detection of common prohibited items across datasets while unique mode learning focuses on the detection of unique prohibited items on each dataset. Finally, a universal detector can be effectively trained.

Algorithm 1: Overall Training of DML-Net

Input: An enhanced RetinaNet $\mathcal{M}_s$; a domain discriminator f_{Θ_D}; multiple training datasets $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K\}$; a set of pre-trained teacher models $\{\mathcal{M}_t^{(1)}, \mathcal{M}_t^{(2)}, ..., \mathcal{M}_t^{(K)}\}$; total training epochs E.

for $e \leftarrow 1$ *to* E **do**
  Shuffle $\{\mathcal{D}_k\}_{k=1}^{K}$ into $\{N_k\}_{k=1}^{K}$ mini-batches;
  Set $N \leftarrow \sum_{k=1}^{K} N_k$;
  Set $b \leftarrow 1$;
  while $b \leq N$ **do**
    for $k \leftarrow 1$ *to* K **do**
      Select a mini-batch of images $\mathcal{B}_k$ from $\mathcal{D}_k$;
      Feed $\mathcal{B}_k$ into $\mathcal{M}_s$ and its corresponding teacher $\mathcal{M}_t^{(k)}$ and obtain a set of multi-scale features;
      Compute the detection loss for the common prohibited items according to Eq. (11);
      Pass the multi-scale features through f_{Θ_D}, and **compute** the adversarial loss according to Eq. (8);
      Update the prototypes of the common prohibited items according to Eq. (9);
      Compute the detection loss for the unique prohibited items according to Eq. (13);
      Compute the feature distillation loss for the unique prohibited items according to Eq. (12);
    end
    Compute the prototype alignment loss according to Eq. (10);
    Compute the total loss according to Eq. (14);
    Update $\mathcal{M}_s$,f_{Θ_D} by stochastic gradient descent;
    $b \leftarrow b + K$;
  end
end

Output: A well-trained detector $\mathcal{M}_s$.

IV. Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed DML-Net method. First, we introduce public X-ray image datasets and implementation details in Section IV-A. Then, we perform ablation studies in Section IV-B. Finally, we compare our DML-Net method with several state-of-the-art methods in Section IV-C.

A. Datasets and Implementation Details

1) Datasets: To demonstrate the effectiveness of our DML-Net, we use three widely-used X-ray image datasets (including OPIXray [4], SIXray [7], and HiXray [8]) and perform experiments on the combinations of these datasets. Specifically, the OPIXray dataset involves a total of 8,885 X-ray images with five prohibited items (including folding knife, straight knife, scissor, utility knife, and multi-tool knife). The SIXray dataset consists of 8,929 X-ray images with five prohibited items (including gun, straight knife, wrench, plier, and scissor). Compared with the OPIXray and SIXray datasets, the HiXray dataset is a larger X-ray image dataset. HiXray contains 45,364 high-quality X-ray images of seven prohibited items (including portable charger, mobile water bottle, laptop, mobile phone, tablet, cosmetic, and metallic-lighter). In this paper, we aim to train a universal detector on the X-ray image multi-dataset. Therefore, we adopt different combinations of these datasets to construct the multi-datasets.

TABLE I: Ablation studies for LAE on the OPIXray+SIXray dataset.

Methods	mAP	mAP_{50}	mAP_{75}
SSD	37.7	75.7	33.8
SSD+DOAM	40.9	79.9	36.0
SSD+LAE	43.1	82.4	37.3
YOLOv3	42.5	84.2	38.2
YOLOv3+DOAM	43.6	85.2	39.0
YOLOv3+LAE	44.8	86.4	39.9
FCOS	43.6	81.4	40.0
FCOS+DOAM	44.7	82.7	41.0
FCOS+LAE	46.3	84.1	42.3
RetinaNet	46.1	83.5	45.0
RetinaNet+DOAM	47.3	84.6	45.9
RetinaNet+LAE	**48.7**	**86.3**	**47.0**

2) Implementation Details: In our DML-Net, the ResNet-50 followed by FPN is used in our backbone network. We implement our method based on PyTorch and conduct all the experiments on an NVIDIA RTX 3090 GPU. We use the stochastic gradient descent optimizer with a momentum of 0.9. For unique mode learning, a set of teacher models are pre-trained on each individual dataset, where the training epochs are set as 120. The training epochs of the student model are also set as 120 for all the datasets. The batch size is uniformly set as 4. The weighting parameters λ_1, λ_2, and λ_3 in Eq. (14) are set as 0.5, 0.5, and 1.0, respectively.

Following [42], we adopt Average Precision (AP) and mean Average Precision (mAP) to measure the performance in each category and all the categories, respectively. We also use mAP_{50} and mAP_{75}, which represent the mAP computed at the IoU thresholds of 0.50 and 0.75, respectively.

B. Ablation Studies

In this section, we perform ablation studies to evaluate the influence of the key components of DML-Net on the final performance. The OPIXray+SIXray dataset is used for evaluation, where the common prohibited items are the scissor and the straight knife.

1) Influence of LAE: We incorporate LAE into four popular object detection methods (including SSD [46], YOLOv3 [47], FCOS [21], and RetinaNet [20]), where we add LAE into the backbones of these methods. For a fair comparison, we also evaluate the performance obtained by combining DOAM [4] with these object detection methods. The comparison results are given in Table I.

RetinaNet achieves higher accuracy than SSD, YOLOv3, and FCOS. This shows the superiority of RetinaNet for X-ray security image detection. Moreover, compared with DOAM-based methods, LAE-based methods give higher mAP, mAP_{50}, and mAP_{75}. LAE can greatly enhance appearance representations and improve the detection performance on X-ray security images. Among all the variants, RetinaNet+LAE achieves the best mAP, mAP_{50}, and mAP_{75}. This is mainly due to the fact

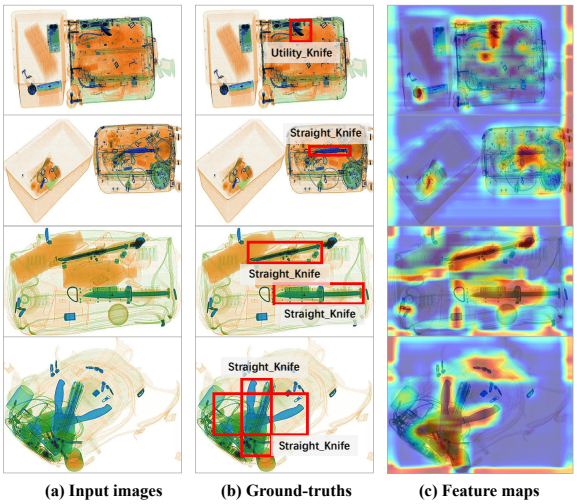

Fig. 5: Visualization of (a) input images, (b) ground-truths, and (c) the feature maps generated by LAE. The first and second rows show the results on OPIXray, while the third and fourth rows show the results on SIXray.

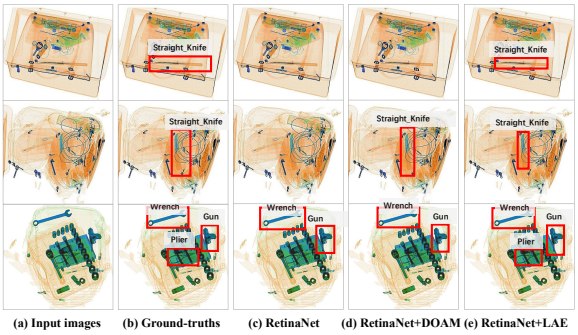

Fig. 6: Visualization of (a) input images, (b) ground-truths, and the detection results obtained by (c) RetinaNet, (d) RetinaNet+DOAM, and (e) RetinaNet+LAE. The first and second rows show the results on OPIXray, while the third row shows the results on SIXray.

that RetinaNet+LAE effectively relieves both the occlusion problem (by using LAE) and the class imbalance problem (by using the focal loss). Note that the above problems are ubiquitous in the X-ray image datasets.

To further illustrate the advantage of LAE, we show some visualization results in Fig. 5. Specifically, we visualize the feature maps obtained by LAE. LAE effectively enhances the appearance of prohibited items, especially the occluded ones. Therefore, the heavy occlusion problem can be greatly alleviated, thus improving the detection performance.

Fig. 6 shows the detection results obtained by RetinaNet, RetinaNet+DOAM, and RetinaNet+LAE. RetinaNet fails to detect some occluded pliers and straight knives, while RetinaNet+DOAM cannot detect some prohibited items in the complex background. In contrast, most prohibited items can be correctly identified by RetinaNet+LAE. The above results

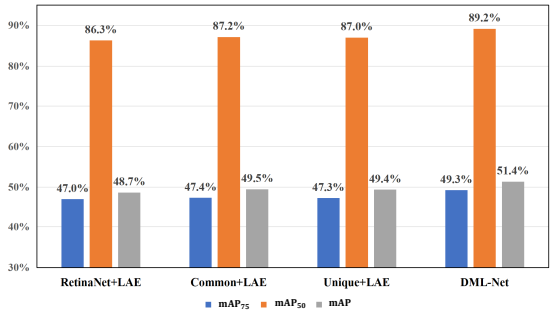

Fig. 7: Performance comparisons between different variants of DML-Net on the OPIXray+SIXray dataset.

TABLE II: Ablation studies for the common mode learning on the OPIXray+SIXray dataset.

Methods	mAP	mAP_{50}	mAP_{75}
RetinaNet+LAE	48.7	86.3	47.0
Unique+LAE	49.4	87.0	47.3
DML-Net_Adv	50.3	88.2	48.8
DML-Net_Pro	50.2	88.0	48.7
DML-Net	**51.4**	**89.2**	**49.3**

validate the effectiveness of LAE, which enhances the appearance information by two lattice structures.

2) Influence of Unique Mode Learning: We evaluate the variant (denoted as Common+LAE) of DML-Net, which is trained by removing the feature distillation loss and keeping only the detection loss for unique mode learning. The results are given in Fig. 7.

We can observe that the Common+LAE method results in 1.9%, 2.0%, and 1.9% drops (in terms of mAP, mAP_{50}, and mAP_{75}, respectively) from DML-Net. This is mainly because the performance of Common+LAE greatly drops for detecting the unique prohibited items.

Moreover, we evaluate another variant (denoted as Unique+LAE) of DML-Net, which is trained by mainly using unique mode learning and removing common mode learning (except for the detection loss for the common prohibited items). The Unique+LAE method obtains 0.7% higher mAP than RetinaNet+LAE. This further demonstrates the effectiveness of unique mode learning.

3) Influence of Common Mode Learning: We evaluate some variants of our method: 1) enhanced RetinaNet (denoted as RetinaNet+LAE); 2) Unique+LAE; 3) DML-Net with only adversarial learning in common mode learning (denoted as DML-Net_Adv); 4) DML-Net with only prototype alignment in common mode learning (denoted as DML-Net_Pro); 5) DML-Net which is based on LAE and dual-mode learning. The evaluated results are given in Table II.

Among all the variants, RetinaNet+LAE gets the worst performance. This is because it ignores the domain discrepancy between datasets. DML-Net_Adv performs slightly better than DML-Net_Pro. This can be ascribed to the fact that adversarial learning is helpful in obtaining a domain-invariant feature

TABLE III: Ablation studies for the different weighting parameters on the OPIXray+SIXray dataset.

(a) Influence of λ_1.

λ_1	mAP	mAP_{50}	mAP_{75}
0.0	50.3	88.2	48.8
0.5	**51.4**	**89.2**	**49.3**
1.0	51.0	88.7	49.0
1.5	50.8	88.5	48.9
2.0	50.6	88.4	48.8

(b) Influence of λ_2.

λ_2	mAP	mAP_{50}	mAP_{75}
0.0	50.2	88.0	48.7
0.5	**51.4**	**89.2**	**49.3**
1.0	50.9	88.7	49.1
1.5	50.6	88.4	48.9
2.0	50.4	88.1	48.7

(c) Influence of λ_3.

λ_3	mAP	mAP_{50}	mAP_{75}
0.0	49.5	87.2	47.4
0.5	50.6	88.5	48.3
1.0	**51.4**	**89.2**	**49.3**
1.5	51.2	88.9	49.1
2.0	50.9	88.5	48.7

TABLE IV: Detection performance comparisons in terms of mAP (%) and AP (%) on the OPIXray+SIXray dataset.

Methods	mAP	Prohibited Items							
		Common Classes		Unique Classes on OPIXray			Unique Classes on SIXray		
		SKnife	*scissor*	*FKnife*	*UKnife*	*MTKnife*	*gun*	*wrench*	*plier*
RetinaNet	46.1	37.6	46.7	34.1	33.7	32.7	66.4	57.3	60.8
RetinaNet+DOAM	47.3	38.7	47.9	34.7	34.9	33.7	67.5	58.9	62.2
SSD	37.7	27.5	38.5	30.4	29.9	32.1	61.3	36.7	45.3
SSD+DOAM	40.9	32.3	44.6	34.0	31.6	36.0	63.2	38.6	47.1
YOLOv3	42.5	31.4	45.7	31.5	32.8	32.1	63.4	48.4	54.8
YOLOv3+DOAM	43.6	31.2	47.9	33.9	33.5	34.7	62.7	50.3	54.6
FCOS	43.6	30.1	49.6	27.9	29.1	29.2	65.8	55.5	62.3
FCOS+DOAM	44.7	31.1	50.7	28.7	29.6	30.4	67.1	56.8	63.6
YOLOv7	45.1	33.2	47.9	32.6	34.3	33.7	64.1	50.3	57.2
YOLOv7+DOAM	47.3	34.1	49.0	33.8	35.6	35.1	65.4	52.8	57.4
LA	48.0	35.8	49.2	36.5	36.6	33.7	69.9	59.2	63.4
Unified	35.3	32.9	41.3	34.3	37.6	33.5	33.7	36.5	32.4
UniDet	35.9	33.2	42.9	35.2	35.0	34.0	35.8	37.4	34.0
Unidetector	40.7	29.9	45.2	21.3	**39.9**	34.6	56.3	47.3	51.0
DML-Net (ours)	**51.4**	**46.5**	**55.7**	**37.3**	37.5	**35.5**	**70.9**	**61.7**	**65.9**

1 'SKnife', 'FKnife', 'UKnife', and 'MTKnife' represent 'straight knife', 'folding knife', 'utility knife', and 'multi-tool knife', respectively.

space. DML-Net outperforms both DML-Net_Adv and DML-Net_Pro, validating the importance of APA. Moreover, DML-Net obtains better performance than Unique+LAE. These results show the effectiveness of common mode learning for handling the domain discrepancy problem on the multi-dataset.

4) Influence of Different Weighting Parameters: In this subsection, we illustrate the influence of different weighting parameters in Eq. (14) on the final performance.

We first fix $\lambda_2 = 0.5$ and $\lambda_3 = 1.0$ and vary the value of λ_1 from 0.0 to 2.0. The results are shown in Table III. From Table III(a), we can observe that when λ_1 is set to 0.0 (which indicates that the prototype alignment loss is not used during model training), DML-Net gets the lowest mAP. Moreover, DML-Net gives the best performance when $\lambda_1 = 0.5$. Next, we fix $\lambda_1 = 0.5$ and $\lambda_3 = 1.0$ and change the value of λ_2 from 0.0 to 2.0. As shown in Table III(b), DML-Net achieves the best results when $\lambda_2 = 0.5$ while obtaining the worst performance when $\lambda_2 = 0.0$. This verifies the effectiveness of adversarial learning used in DML-Net. Finally, we fix $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ and vary the value of λ_3 from 0.0 to 2.0. The results are given in Table III(c). From Table III(c), DML-Net gives the best results when $\lambda_3 = 1.0$. When $\lambda_3 = 0.0$, the joint feature distillation loss is not used and DML-Net gets the worst results in terms of mAP, mAP_{50}, and mAP_{75} among all the variants.

C. Comparisons with State-of-the-Art Methods

In this section, we evaluate our proposed DML-Net on different multi-datasets. We compare our proposed DML-Net with several state-of-the-art object detection methods (SSD [46], YOLOv3 [47], FCOS [21], and YOLOv7 [48]), X-ray security image detection methods (LA [5] and DOAM-based methods [4], including SSD+DOAM, YOLOv3+DOAM, FCOS+DOAM, YOLOv7+DOAM) and representative multi-dataset object detection methods (Unified [49], UniDet [13], and Unidector [39]) on several combinations of X-ray image datasets. For all the competing methods, we report their results by running the source codes provided by their original papers.

TABLE V: Detection performance comparisons in terms of mAP (%) and AP (%) on the SIXray+HiXray dataset.

Methods	mAP	Prohibited Items											
		Unique Classes on SIXray					Unique Classes on HiXray						
		SKnife	*scissors*	*gun*	*wrench*	*plier*	*charger*	*bottle*	*laptop*	*phone*	*tablet*	*cosmetic*	*lighter*
RetinaNet	53.2	51.7	57.7	68.3	60.1	62.9	53.4	52.9	70.5	62.5	66.2	26.2	5.9
RetinaNet+DOAM	54.1	52.9	59.3	69.9	61.4	63.9	54.3	53.6	71.4	62.9	66.8	26.6	6.4
SSD	44.1	42.0	43.2	62.9	38.5	46.8	48.4	45.6	64.3	56.2	61.0	17.6	2.6
SSD+DOAM	45.4	44.2	44.5	64.1	39.5	48.3	49.5	46.6	65.8	57.4	62.6	18.9	3.9
YOLOv3	45.0	41.4	48.4	61.8	50.0	54.3	44.1	45.5	63.9	54.2	59.1	15.0	2.7
YOLOv3+DOAM	46.2	42.9	49.5	63.3	51.2	55.4	45.3	46.2	64.8	55.4	59.9	16.5	3.8
FCOS	52.1	52.8	60.0	67.2	56.9	64.1	54.3	53.1	63.2	60.9	65.1	24.5	3.0
FCOS+DOAM	53.7	54.1	61.4	68.5	58.3	65.4	55.9	54.7	64.9	62.8	67.5	26.2	4.2
YOLOv7	47.2	45.6	50.1	64.8	52.4	57.1	46.5	48.0	65.1	55.4	61.2	18.1	3.2
YOLOv7+DOAM	48.6	46.9	52.3	65.7	53.5	57.9	47.3	48.7	66.2	56.2	62.0	19.9	5.1
LA	54.4	54.3	60.7	70.8	59.7	63.9	55.9	56.3	71.9	62.3	66.4	25.2	5.9
Unified	37.1	32.2	43.0	35.1	35.8	32.4	41.9	40.8	54.5	54.2	50.1	19.6	5.7
UniDet	38.6	35.2	43.2	36.0	37.1	33.9	43.4	41.0	58.9	54.3	51.7	20.7	7.2
Unidetector	45.2	42.3	55.3	56.0	48.2	48.9	55.2	53.3	58.7	48.5	48.9	20.5	7.0
DML-Net (ours)	**60.6**	**61.7**	**67.9**	**74.2**	**64.5**	**68.0**	**62.9**	**64.6**	**79.7**	**68.9**	**72.4**	**28.4**	**13.4**

1 'SKnife', 'charger', 'bottle', 'phone', and 'lighter' represent 'straight knife', 'portable charger', 'mobile water bottle', 'mobile phone', and 'metallic-lighter', respectively.

TABLE VI: Detection performance comparisons in terms of mAP (%) and AP (%) on the OPIXray+SIXray+HiXray dataset.

Methods	mAP	Prohibited Items													
		Common Classes on OPIXray and SIXray		Unique Classes on OPIXray			Unique Classes on SIXray			Unique Classes on HiXray						
		SKnife	*scissor*	*FKnife*	*UKnife*	*MTKnife*	*gun*	*wrench*	*plier*	*charger*	*bottle*	*laptop*	*phone*	*tablet*	*cosmetic*	*lighter*
RetinaNet	46.2	36.4	45.2	33.2	32.6	31.9	65.2	56.3	60.1	52.7	51.9	69.8	61.6	65.4	25.6	4.8
RetinaNet+DOAM	47.0	37.5	46.2	34.6	34.2	32.8	66.2	57.8	61.3	53.4	52.2	70.3	61.6	65.2	26.1	5.5
SSD	40.3	41.8	42.4	29.7	29.2	31.2	60.1	35.8	44.7	47.5	44.5	63.8	55.3	60.1	16.9	2.0
SSD+DOAM	42.0	43.3	43.2	33.2	30.4	35.1	62.1	38.1	46.3	48.6	45.3	64.6	56.6	61.7	18.1	3.2
YOLOv3	41.0	37.3	44.8	30.7	31.6	31.4	61.2	47.6	53.1	43.0	44.4	63.1	53.1	57.8	14.2	2.3
YOLOv3+DOAM	42.7	42.0	48.8	32.8	32.8	33.9	61.7	49.4	53.8	44.7	45.4	64.0	54.1	58.8	15.7	3.2
FCOS	46.2	51.9	59.1	27.1	27.9	28.3	64.5	54.7	61.4	53.2	52.2	62.4	59.5	64.3	23.8	2.4
FCOS+DOAM	47.7	53.3	60.3	27.9	28.5	29.5	66.6	56.0	62.8	54.8	53.8	64.1	62.2	66.8	25.6	3.7
YOLOv7	43.1	38.9	46.2	33.8	34.6	33.8	63.0	38.5	46.2	45.1	46.3	65.7	56.9	58.1	17.4	4.1
YOLOv7+DOAM	45.2	41.4	50.7	36.1	35.2	35.4	63.8	40.3	46.9	46.9	47.4	66.6	57.8	59.0	19.2	5.9
LA	47.4	34.6	48.1	35.2	35.1	32.4	70.3	58.1	62.3	54.7	55.3	70.7	61.4	63.2	23.9	5.1
Unified	35.9	32.6	40.7	34.0	36.1	32.6	33.5	35.4	31.7	41.3	39.9	54.2	53.1	49.8	18.8	5.5
UniDet	37.8	33.7	43.9	34.6	38.5	33.3	37.0	35.2	34.3	42.9	40.3	59.6	53.3	53.0	19.9	7.1
Unidetector	40.2	28.8	44.3	21.9	37.2	32.8	55.2	47.2	49.4	54.7	52.0	57.8	48.1	48.2	19.9	6.1
DML-Net (ours)	**52.7**	**42.7**	**54.7**	**38.1**	**38.9**	**35.9**	**72.4**	**62.7**	**67.2**	**61.2**	**62.9**	**77.6**	**67.2**	**71.1**	**26.6**	**11.4**

1 'SKnife', 'FKnife', 'UKnife', 'MTKnife', 'charger', 'bottle', 'phone', and 'lighter' represent 'straight knife', 'folding knife', 'utility knife', 'multi-tool knife', 'portable charger', 'mobile water bottle', 'mobile phone', and 'metallic-lighter', respectively.

1) Results on OPIXray+SIXray: Table IV shows the comparison results on the OPIXray+SIXray dataset. We can observe that conventional object detection methods with DOAM achieve higher mAP than those without DOAM. This validates the effectiveness of DOAM for detecting X-ray security images. Our proposed DML-Net outperforms the RetinaNet-based, SSD-based, YOLOv3-based, FCOS-based, and YOLOv7-based methods. This is because DML-Net is designed for multi-dataset X-ray security image detection by explicitly modeling the intrinsic relationship between different datasets. In contrast, the RetinaNet-based, SSD-based, YOLOv3-based, FCOS-based, and YOLOv7-based methods do not fully consider the large domain discrepancy between datasets, leading to a performance drop. Moreover, our DML-Net outperforms the Unified, UniDet, Unidetector methods (which are designed for multi-dataset object detection in natural images). In particular, DML-Net achieves 10.7% higher mAP than the Unidector method. For common prohibited items (such as straight knife and scissor), our DML-Net increases the AP by 16.6% and 10.5%, respectively. For unique prohibited items (such as gun), DML-Net improves the AP by 14.6%. These results show the superiority of dual-mode learning and LAE.

2) Results on SIXray+HiXray: The comparison results on the SIXray+HiXray dataset are given in Table V. Note that there are no common prohibited items between SIXray and HiXray. Thus, only unique mode learning is employed in DML-Net. Our DML-Net outperforms the second-best detector LA by a moderate margin (about 6.2% higher mAP). This can be ascribed to the effectiveness of unique mode learning and LAE. YOLOv3 has three detection layers, making it good at detecting natural objects of various sizes. However, it may struggle with very small or overlapping X-ray items. YOLOv7 offers better accuracy and speed than YOLOv3. Note that the RetinaNet method achieves 8.2% and 6.0% higher mAP than YOLOv3 and YOLOv7, respectively. This is because YOLOv3 and YOLOv7 do not address the heavy class imbalance problem in the X-ray image datasets. In a word, our DML-Net excels in X-ray prohibited item detection, while YOLO models are more suited for general object detection.

3) Results on OPIXray+SIXray+HiXray: We further evaluate the performance of our method on the OPIXray+SIXray+HiXray dataset, which is a more challenging multi-dataset than the previous two multi-datasets. The comparison results are given in Table VI. Our DML-Net achieves better performance on the common prohibited items (such as straight knife and scissor) than the other competing methods. Moreover, DML-Net also outperforms the other competing methods on all the unique prohibited items. Hence, our method can improve the performance of detecting both the common and unique prohibited items on the multi-dataset.

Some detection results obtained by RetinaNet, RetinaNet+DOAM, Unified, and our developed DML-Net are illustrated in Fig. 8. We can observe that RetinaNet and the Unified method may fail to detect some occluded prohibited items. RetinaNet+DOAM gives more accurate detection results than RetinaNet since DOAM is specifically designed for X-ray security images. Our DML-Net is capable of detecting more prohibited items than other competing methods. By introducing LAE into RetinaNet, the appearance information of prohibited items can be enhanced while the negative effect of background clutters is suppressed, facilitating dual-mode learning. Meanwhile, dual-mode learning effectively alleviates the large domain discrepancy and category differences between datasets, improving the final detection performance on X-ray security images.

TABLE VII: The number of parameters (Params), Giga floating-point operations per second (GFLOPs), training time (hour), and inference latency (frame per second) obtained by different methods on the OPIXray+SIXray dataset. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

Method	Params (M)	GFLOPs	Training Time	Latency
RetinaNet	36.19	18.79	1.16	0.0169
RetinaNet+DOAM	36.21	20.94	1.33	0.0182
SSD	24.28	30.69	1.10	0.0120
SSD+DOAM	24.30	32.84	1.16	0.0132
YOLOv3	61.55	19.39	1.35	0.0137
YOLOv3+DOAM	61.57	21.84	1.53	0.0156
FCOS	32.02	18.29	1.22	0.0186
FCOS+DOAM	32.04	20.44	1.42	0.0196
YOLOv7	42.56	18.96	1.30	0.0130
YOLOv7+DOAM	42.58	21.11	1.50	0.0140
LA	37.03	23.54	1.69	0.0212
Unified	40.20	35.67	1.45	0.0235
UniDet	45.49	40.32	2.34	0.0439
Unidetector	46.87	42.58	2.55	0.0526
DML-Net (ours)	36.20	20.65	1.28	0.0177

4) Computational Complexity Analysis: We report the number of parameters, Giga floating-point operations per second, training time, and inference latency obtained by different competing methods on the OPIXray+SIXray dataset, as shown in Table VII. Compared with RetinaNet, the number of parameters of DML-Net slightly increases, but the performance gains are more significant (see Table IV). Besides, our DML-Net outperforms RetinaNet+DOAM with fewer parameters (36.20M vs. 36.21M). This demonstrates the efficiency of LAE. Additionally, the training time of DML-Net is remarkably lower than Unified, UniDet, and Unidetector. For the inference latency, compared with RetinaNet+DOAM, SSD+DOAM, YOLOv3+DOAM, YOLOv7+DOAM, and FCOS+DOAM, our DML-Net achieves similar results. This can be ascribed to the fact that DML-Net adopts a lightweight lattice-structure based appearance enhancement module.

5) Failure Case Analysis: Fig. 9 presents some failure cases of our DML-Net on the OPIXray+SIXray dataset. The failure cases can be roughly summarized into two aspects: 1) Small item size: For prohibited items with small image sizes, our method sometimes struggles to detect them. Some false negative examples are given in the first three columns of Fig. 9. This is because of the limited information on small prohibited items, increasing the difficulty of extracting discriminative features for detection. Hence, DML-Net cannot accurately detect the small prohibited items. To address this issue, we can incorporate contextual information (such as local pixel context or semantic context) in the detection network. 2) Data limitations: Our method may fail to detect the prohibited

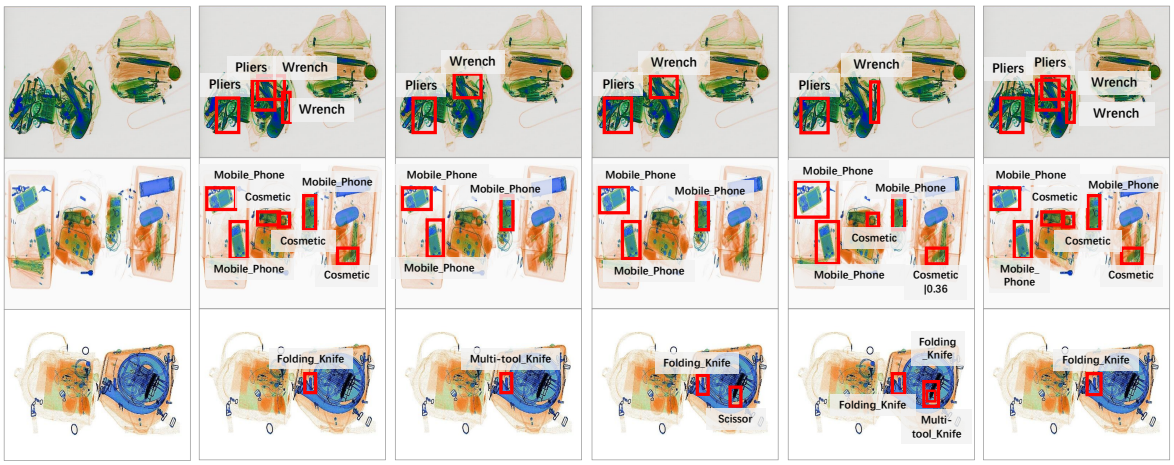

Fig. 8: Visualization of (a) input images, (b) ground-truths, and the detection results obtained by (c) RetinaNet, (d) RetinaNet+DOAM, (e) Unified, and (f) our DML-Net. The first, second, and third rows show the results on SIXray, HiXray, and OPIXray, respectively.

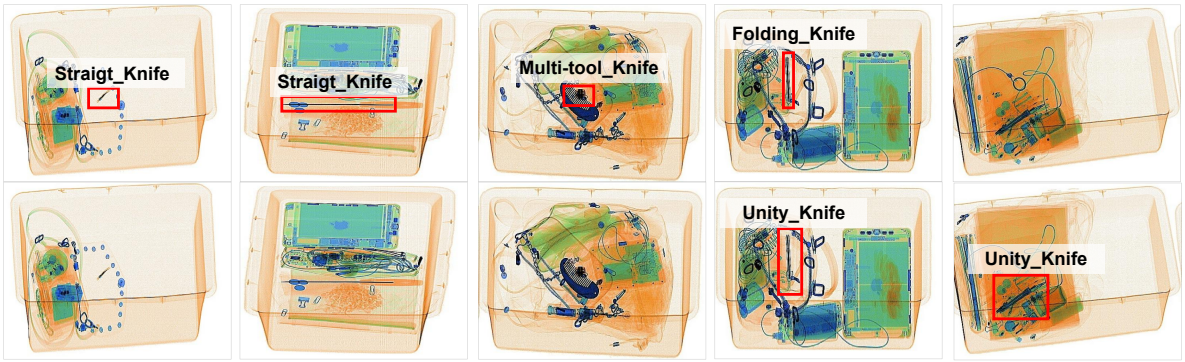

Fig. 9: Examples of failure cases on the OPIXray+SIXray dataset. The first row shows the original image and the corresponding ground-truths. The second row gives the detection results obtained by our method.

items when the training data of one prohibited item category are limited. Some false positive examples are given in the last two columns of Fig. 9. In such cases, there exists the severe class imbalance between classes. As a result, the trained detector tends to focus on the detection of majority classes while ignoring minority classes. To tackle the issue posed by data limitations, we can either balance the data distribution during training or employ more advanced loss functions.

V. Conclusion

In this paper, we investigate an under-explored but important task, which targets training a universal X-ray image detector on the multi-dataset. To address this task, we propose a novel DML-Net based on an enhanced RetinaNet. Specifically, we introduce an enhanced RetinaNet by designing LAE to enhance appearance representations of prohibited items, mitigating the occlusion problem in X-ray security images. Based on the enhanced RetinaNet, the learning process of DML-Net involves both common mode learning and unique mode learning to detect the common and unique prohibited items, respectively. By tightly combining the dual modes, we significantly eliminate the domain discrepancy between datasets and are able to effectively detect all the prohibited items across datasets. Extensive experimental results on combined X-ray image datasets show the superiority of our DML-Net method in comparison with several state-of-the-art methods.

References

[1] S. Akcay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within X-ray security imaging," *Pattern Recognit.*, vol. 122, no. 108245, pp. 1–12, 2022.
[2] F. Shao, J. Liu, P. Wu, Z. Yang, and Z. Wu, "Exploiting foreground and background separation for prohibited item detection in overlapping X-ray images," *Pattern Recognit.*, vol. 122, no. 108261, pp. 1–11, 2022.
[3] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, "Prohibited object detection in X-ray images with dynamic deformable convolution and adaptive IoU," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 3001–3005.

[4] 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 *Proc. ACM Int. Conf. Multimedia.*, 2020, pp. 138–146.
[5] 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 Trans. Inf. Forensics Security*, vol. 17, pp. 998–1009, 2022.
[6] L. D. Griffin, M. Caldwell, J. T. Andrews, and H. Bohler, "Unexpected item in the bagging area: Anomaly detection in X-ray security images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1539–1553, 2018.
[7] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2119–2128.
[8] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10 923–10 932.
[9] D. Mery, V. Riffo, U. Zscherpel, G. Mondragón, I. Lillo, I. Zuccar, H. Lobel, and M. Carrasco, "GDXray: The database of X-ray images for nondestructive testing," *J. Nondestruct. Eval.*, vol. 34, no. 4, pp. 1–12, 2015.
[10] J.-M. O. Steitz, F. Saeedan, and S. Roth, "Multi-view X-ray R-CNN," in *Proc. German Conf. Pattern Recognit.*, 2018, pp. 153–168.
[11] Z. Liu, J. Li, Y. Shu, and D. Zhang, "Detection and recognition of security detection object based on YOLO9000," in *Proc. 5th Int. Conf. Syst. Informat.*, 2018, pp. 278–282.
[12] B. Chen, Z. Yan, K. Li, P. Li, B. Wang, W. Zuo, and L. Zhang, "Variational attention: Propagating domain-specific knowledge for multi-domain learning in crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16 065–16 075.
[13] X. Zhou, V. Koltun, and P. Krähenbühl, "Simple multi-dataset detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7571–7580.
[14] N. Jaccard, T. W. Rogers, E. J. Morton, and L. D. Griffin, "Automated detection of smuggled high-risk security threats using deep learning," in *Proc. Int. Conf. Imaging Crime Detect. Prev.*, 2016, pp. 1–6.
[15] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2203–2215, 2018.
[16] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards real-world prohibited item detection: A large-scale X-ray benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 5412–5421.
[17] A. Chang, Y. Zhang, S. Zhang, L. Zhong, and L. Zhang, "Detecting prohibited objects with physical size constraint from cluttered X-ray baggage images," *Knowledge-Based Syst.*, vol. 237, no. 107916, pp. 1–14, 2022.
[18] M. Wang, H. Du, W. Mei, S. Wang, and D. Yuan, "Material-aware cross-channel interaction attention (MCIA) for occluded prohibited item detection," *Visual Comput.*, pp. 1–13, 2022.
[19] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of YOLO algorithm developments," *Procedia Comput. Sci.* vol. 199, pp. 1066–1073, 2022.
[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[21] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
[22] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3339–3348.
[23] Y. Wang, R. Zhang, S. Zhang, M. Li, Y. Xia, X. Zhang, and S. Liu, "Domain-specific suppression for adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9603–9612.
[24] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, 2022.
[25] J. L. G. Bello and M. Kim, "Self-supervised deep monocular depth estimation with ambiguity boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9131–9149, 2022.
[26] Y. Wang, X. Song, T. Xu, Z. Feng, and X.-J. Wu, "From RGB to depth: Domain transfer network for face anti-spoofing," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4280–4290, 2021.
[27] G. Yang, J. Manela, M. Happold, and D. Ramanan, "Hierarchical deep stereo matching on high-resolution images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5515–5524.
[28] T.-J. Song, J. Jeong, and J.-H. Kim, "End-to-end real-time obstacle detection network for safe self-driving via multi-task learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16 318–16 329, 2022.
[29] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable pedestrian detection: The elephant in the room," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11 328–11 337.
[30] K. Pfeiffer, A. Hermans, I. Sárándi, M. Weber, and B. Leibe, "Visual person understanding through multi-task and multi-dataset learning," in *Proc. German Conf. Pattern Recognit.*, 2019, pp. 551–566.
[31] D. Kim, Y.-H. Tsai, Y. Suh, M. Faraki, S. Garg, M. Chandraker, and B. Han, "Learning semantic segmentation from multiple datasets with label shifts," *arXiv:2202.14030*, 2022.
[32] L. Wang, D. Li, H. Liu, J. Peng, L. Tian, and Y. Shan, "Cross-dataset collaborative learning for semantic segmentation in autonomous driving," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2487–2494.
[33] T. Perrett and D. Damen, "Recurrent assistance: Cross-dataset training of lstms on kitchen tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2017, pp. 1354–1362.
[34] Y. Yao, Y. Wang, Y. Guo, J. Lin, H. Qin, and J. Yan, "Cross-dataset training for class increasing object detection," *arXiv:2001.04621*, 2020.
[35] T. Zhao, P. Liu, X. Lu, and K. Lee, "OMDET: Language-aware object detection with large-scale vision-language multi-dataset pre-training," *arXiv:2209.05946*, 2022.
[36] N. Jiang, K. Wang, X. Peng, X. Yu, Q. Wang, J. Xing, G. Li, J. Zhao, G. Guo, and Z. Han, "Anti-UAV: A large multi-modal benchmark for UAV tracking," *arXiv:2101.08466*, 2021.
[37] X. F. Zhu, T. Xu, J. Zhao, J. W. Liu, K. Wang, G. Wang, J. Li, Z. Zhang, Q. Wang, L. Jin, and Z. Zhu, "Evidential detection and tracking collaboration: New problem, benchmark and algorithm for robust anti-UAV system," *arXiv:2306.15767*, 2023.
[38] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7289–7298.
[39] Z. Wang, Y. Li, X. Chen, S. N. Lim, A. Torralba, H. Zhao, and S. Wang, "Detecting everything in the open world: Towards universal object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11433–11443.
[40] Y. Chen, M. Wang, A. Mittal, Z. Xu, P. Favaro, J. Tighe, and D. Modolo, "ScaleDet: A scalable multi-dataset object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7288–7297.
[41] A. Gray and J. Markel, "Digital lattice and ladder filter synthesis," *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 6, pp. 491–500, 1973.
[42] 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 *Proc. ACM Int. Conf. Multimedia.*, 2020, pp. 138–146.
[43] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[44] J. Li, C. Xiong, and S. Hoi, "MOPRO: Webly supervised learning with momentum prototypes," in *Proc. Int. Conf. Learn. Represent.*, 2020.
[45] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.
[46] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[47] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv:1804.02767*, 2018.
[48] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–7475.
[49] X. Zhao, S. Schulter, G. Sharma, Y. H. Tsai, M. Chandraker, and Y. Wu, "Object detection with a unified label space from multiple datasets," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 178–193.