Augmentation Matters: A Mix-Paste Method for X-Ray Prohibited Item Detection under Noisy Annotations

Ruikang Chen, Yan Yan, Jing-Hao Xue, Yang Lu, Hanzi Wang

***Abstract*—Automatic X-ray prohibited item detection is vital for public safety. Existing deep learning-based methods all assume that the annotations of training X-ray images are correct. However, obtaining correct annotations is extremely hard if not impossible for large-scale X-ray images, where item overlapping is ubiquitous. As a result, X-ray images are easily contaminated with noisy annotations, leading to performance deterioration of existing methods. In this paper, we address the challenging problem of training a robust prohibited item detector under noisy annotations (including both category noise and bounding box noise) from a novel perspective of data augmentation, and propose an effective label-aware mixed patch paste augmentation method (Mix-Paste). Specifically, for each item patch, we mix several item patches with the same category label from different images and replace the original patch in the image with the mixed patch. In this way, the probability of containing the correct prohibited item within the generated image is increased. Meanwhile, the mixing process mimics item overlapping, enabling the model to learn the characteristics of X-ray images. Moreover, we design an item-based large-loss suppression (LLS) strategy to suppress the large losses corresponding to potentially positive predictions of additional items due to the mixing operation. We show the superiority of our method on X-ray datasets under noisy annotations. In addition, we evaluate our method on the noisy MS-COCO dataset to showcase its generalization ability. These results clearly indicate the great potential of data augmentation to handle noise annotations. The source code is released at https://github.com/wscds/Mix-Paste.**

***Index Terms*—Object Detection, Noisy Annotation, Data Augmentation, X-Ray Prohibited Item Detection.**

I. Introduction

OVER the past few years, automatic X-ray prohibited item detection, which can assist security inspectors to quickly identify the locations and categories of prohibited items, has attracted much attention. A large number of prohibited item detection methods [1]–[7] have been developed.

Generally, existing X-ray prohibited item detection methods depend heavily on a large-scale dataset for model training. Unfortunately, obtaining correct annotations with clean category labels as well as accurate bounding boxes is labor-expensive and requires the expertise of professionals. Notably, the ubiquitous item overlapping in X-ray images renders the annotation of an X-ray dataset a challenging task. In many practical applications, machine-assisted annotations or crowd-sourcing are often employed to annotate large-scale data, reducing the expensive cost of high-quality human annotations. The machine-assisted process or the crowd-sourcing labeling process easily leads to noisy annotations. As a result, some existing X-ray datasets involve annotations with both *category noise* (i.e., noisy category labels) and *bounding box noise* (i.e., inaccurate ground-truth bounding boxes). Fig. 1 gives some examples with noisy annotations in a public X-ray dataset [1]. These noisy annotations greatly decrease the model performance.

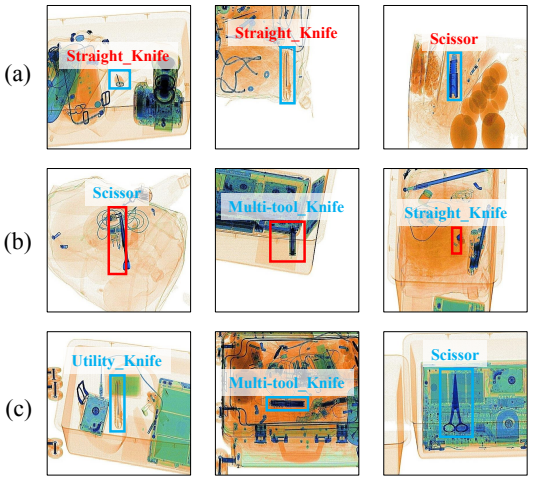

Fig. 1. Examples in an X-ray dataset [1]. (a) Examples with noisy category labels (the correct labels from left to right are folding knife, utility knife, and utility knife, respectively). (b) Examples with inaccurate bounding boxes. (c) Examples with correct annotations in X-ray images, where item overlapping is ubiquitous.

To address the problem of learning with label noise, existing methods [8]–[10] often adopt a label refinement or loss correction paradigm. However, most of these methods work on the image classification task without considering the existence or the location of the objects/items. Unlike the image classification task, the X-ray prohibited item detection task introduces additional challenges caused by inaccurate ground-truth bounding boxes. As a consequence, label noise learning methods do not work well on the prohibited item detection

This work was partly 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.)*

R. Chen, Y. Yan, Y. Lu, and 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: 23020221154074@stu.xmu.edu.cn; yanyan@xmu.edu.cn; luyang@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).

task (a specialized task of object detection). Recently, some works [11]–[15] have studied object detection under noisy annotations. But they are designed for common object detection and are not well-suited for prohibited item detection due to the presence of ubiquitous item overlapping in X-ray images.

To effectively train a robust prohibited item detector in noisy scenarios, we revisit the fundamental aspect of learning with noisy annotations (i.e., reducing the noise during training) and the inherent characteristics of X-ray images (i.e., the ubiquitous overlapping between items) from the perspective of data augmentation. In particular, for an X-ray dataset contaminated with noisy annotations, a collection of item patches that share the same category label is more likely to contain one correct prohibited item than the individual patch in the collection. Inspired by this observation, we mix such multiple item patches to generate a mixed patch and paste it back into the original image for data augmentation. Thus, the generated image can involve the correct prohibited item with a high probability, reducing the negative influence of noisy annotations. In fact, the mixing process of these multiple item patches also effectively mimics item overlapping in X-ray images, enabling the detector to gain a deeper understanding of X-ray images and improve the detection performance.

Although the mixed patch can effectively alleviate the noisy annotations, it may introduce additional noisy-labeled (caused by category noise) prohibited items during training. In such a case, the model tends to give predictions for all the possible prohibited items in the mixed patch before overfitting noisy labels. When the conventional classification loss is used for model optimization, some accurate predictions of additional prohibited items may be mistakenly considered as false predictions (since the category label of additional prohibited items is noisy) and generate large losses. Such a way can be harmful to model training. Hence, we should remove the large losses corresponding to these potentially positive predictions during loss calculation.

Based on the above analysis, we propose a simple yet effective data augmentation method, called label-aware mixed patch paste augmentation (**Mix-Paste**) to address the problem of training a robust X-ray prohibited item detector under noisy annotations. Specifically, for each item patch (corresponding to a ground-truth bounding box) in the training image, we first randomly choose several item patches (according to their ground-truth bounding boxes) with the same category label from different images. Then, we mix these patches and replace the original patch in the image with the mixed patch, obtaining a new image. By doing this, the probability of containing the correct prohibited item within the generated image is increased. It is worth pointing out that our method randomly selects item patches with the same category label, where the category label of some patches can be contaminated with noise. In other words, it does not require the label of the selected item patches to be clean. In fact, such a selection can increase the probability of containing the correct prohibited item in the mixed patch. To effectively optimize the model on augmented data, we design an item-based large-loss suppression (LLS) strategy, suppressing the large losses corresponding to potentially positive predictions of additional items caused by the mixing operation in Mix-Paste.

In summary, our contributions are given as follows:

- We propose a new data augmentation method by mixing item patches with the same category label for X-ray prohibited item detection. Our method can significantly reduce category noise and bounding box noise during training, obtaining a noise-robust prohibited item detector. To the best of our knowledge, we are the first to address the problem of noisy annotations in X-ray prohibited item detection from a novel perspective of data augmentation.
- We design an effective loss suppression strategy for loss calculation. Such a strategy overcomes the limitations of the small-loss criterion [16], [17] for label noise learning in our task. This greatly reduces the adverse influence of the large losses corresponding to potentially positive predictions of additional items caused by the mixing process of item patches and enhances the detection performance.

Our extensive experiments on public X-ray datasets validate the effectiveness of our method under noisy annotations. We also perform experiments on the noisy MS-COCO dataset [18], which exists a certain level of object overlapping. Interestingly, our method shows performance improvements on the common object detection task. These results clearly indicate the advantage of data augmentation to address the problem of learning with noisy annotations.

The remainder of this paper is organized as follows. We first review the related work in Section II. Then, we elaborately introduce our proposed method in Section III. Next, we conduct the experiments on the noisy X-ray datasets and the noisy MS-COCO dataset in Section IV. Finally, we draw the conclusion in Section V.

II. Related Works

In this section, we briefly review several related works. We first introduce X-ray prohibited item detection methods in Section II-A. Then, we review data augmentation methods in Section II-B. Finally, we review the methods of learning with label noise and learning with noisy annotations for object detection in Section II-C and Section II-D, respectively.

A. X-Ray Prohibited Item Detection

With the development of deep learning technology, automatic X-ray prohibited item detection has been widely applied in security inspection. A variety of methods [1], [3], [4], [6], [19]–[25] have been developed to address the severe occlusion and item overlapping problems in X-ray images by introducing attention mechanisms or specifically-designed modules. Wei *et al.* [1] propose an attention mechanism to enhance the edge and material information of prohibited items. Zhang *et al.* [3] apply spatial- and channel-wise attention mechanisms to extract discriminative features and incorporate a dependency refinement module to explore long-range dependencies within the feature map. Tao *et al.* [4] identify the object regions for prohibited item detection by removing the noisy information from neighboring regions and activating the boundary information. Ma *et al.* [6] leverage dual-view X-ray images

as the input and exploit non-overlapping information of two images to enhance feature representations of prohibited items, effectively mitigating background overlapping. Zhao *et al.* [19] introduce a label-aware mechanism to tackle the item overlapping problem by establishing the associations between feature channels and labels. Based on this, they refine and adjust the features to enhance prediction results according to the assigned pseudo labels. Shao *et al.* [20] propose a foreground and background separation (FBS) method, which can handle the severe overlapping problem in X-ray images by separating prohibited items from other irrelevant items. Velayudhan *et al.* [21] introduce a baggage threat detection framework based on broad learning. This framework leverages low-rank features to identify and localize concealed and cluttered baggage threats.

Due to the ubiquitous item overlapping in X-ray images, annotating an X-ray dataset becomes a challenging task. As a result, some X-ray datasets involve noisy annotations with both category noise and bounding box noise. Hence, it is crucial to develop a noise-robust prohibited item detector.

B. Data Augmentation

Data augmentation aims to improve the generalization capability of models by artificially increasing the diversity of the training data. Cutout [26] randomly applies masking to square patches in the image, effectively enforcing the model to learn from incomplete information. Mix-Up [27] generates new training examples by linearly interpolating between two images and their corresponding labels. This encourages the model to generalize beyond the training data and reduce sensitivity to adversarial examples. AlignMix [28] improves representation learning by geometrically aligning and interpolating features from multiple images. Such a way enhances the model's generalization and robustness. Mosaic [29] combines four images into one, enabling the model to observe multiple contexts in a single training step. This method increases the diversity of the dataset and exposes the model to more complex, multi-object scenes, enhancing robustness to variations in object scale and occlusions. Channel augmentation [30] explores the relationship between visible and infrared images to obtain modality-invariant features.

Recently, some methods leverage CLIP [31] or the diffusion model [32] to generate new data by using prompt words. Fang *et al.* [33] propose a data augmentation pipeline based on controllable diffusion models and CLIP for object detection. Gannamaneni *et al.* [34] generate safety critical scenes by inpainting with diffusion models conditioned on text and pose, offering fine-grained control over pedestrian attributes.

Some of the above methods perform well in clean X-ray datasets [35]. However, when applied to the X-ray datasets involving noisy annotations, the generated X-ray images contain significant disturbances, potentially degrading model performance. Different from the above data augmentation methods, we design a data augmentation method to effectively alleviate noisy annotations in the X-ray dataset and improve the training performance of the model in the noisy dataset.

C. Learning with Label Noise

A number of methods [8]–[10], [16], [36], [37] have been developed for learning with label noise. Some methods [38]–[41] address the label noise problem by employing a noise transition matrix to refine predictions. Goldberger *et al.* [38] adopt both an s-model and a c-model to effectively obtain a noise transition matrix. Patrini *et al.* [39] explicitly model the noise transition matrix to correct the loss. Several works [10], [36], [42], [43] reveal that a loss function involving symmetric properties exhibits enhanced robustness against label noise. However, these methods may only be capable of handling certain noisy rates. Recent methods focus on new learning paradigms [9], [44]–[50]. For example, MentorNet [44] leverages a teacher-student framework to learn a robust student model by exploiting the knowledge of a teacher model. Co-teaching [9] trains the two models simultaneously, where each model selects the samples with the small-loss criterion to update the other model. Co-teaching+ [45] improves the performance of Co-teaching by training on disagreement data. JoCoR [46] allows the two models to reach an agreement by minimizing the distance loss predicted by the two models. PurifyNet [48] introduces a hard-aware instance re-weighting strategy to focus on hard samples in the noisy dataset. Ye *et al.* [49] propose an online label co-refinement framework, which progressively refines noisy labels during model optimization.

The above methods mainly handle label noise and focus on the image classification task. In this paper, our method addresses noisy annotations involving both category noise and bounding box noise for training a robust prohibited item detector. Moreover, unlike the above methods that select clean data by designing different strategies, our method addresses noisy annotations from the perspective of data augmentation. In this way, our method does not require estimating/predefining the noise rates or selecting a clean subset as did in conventional label noise learning methods.

D. Learning with Noisy Annotations for Object Detection

Recently, some methods [11]–[14] have been developed to address robust training on noisy annotations for object detection. Chadwick *et al.* [11] extend the Co-teaching strategy to the field of object detection. Li *et al.* [12] decouple bounding box noise from category noise and then leverage the predicted output for category noise correction and bounding box refinement. Yang *et al.* [13] reduce the influence of noisy labels by employing diverse loss functions. Liu *et al.* [14] treat each object as a bag of instances and select accurate instances from the object bags for training. Wang *et al.* [15] develop a novel Bayesian filter-based prediction ensemble method to address noisy bounding box annotations within a teacher-student learning framework.

The aforementioned methods are designed for common object detection. In contrast, our method mixes multiple item patches to mimic the characteristics of X-ray images (i.e., the ubiquitous overlapping between items) for prohibited item detection. Surprisingly, experiments show that our method is also beneficial in improving the performance of common object detection (which exists a certain level of object overlapping).

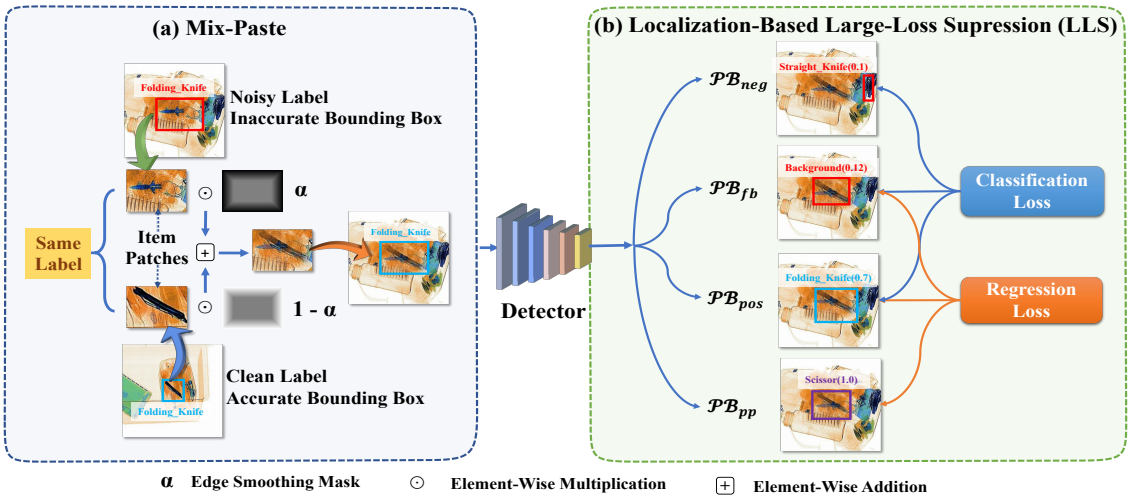

Fig. 2. Overview of our proposed method for training a robust prohibited item detector under noisy annotations. (a) illustrates Mix-Paste which mixes multiple item patches with the same category label (the correct label in the upper item patch is the scissor) for data augmentation. (b) illustrates the LLS strategy which suppresses the large losses corresponding to potentially positive predictions of additional items during loss calculation. $\mathcal{PB}_{neg}$: the predicted bounding boxes whose IoUs between them and the ground-truth bounding boxes are less than a threshold; $\mathcal{PB}_{fb}$: the predicted bounding boxes whose IoUs are greater than a threshold and the predicted label is the background; $\mathcal{PB}_{pos}$: the predicted bounding boxes whose IoUs are greater than a threshold and the predicted label is the same as the ground-truth category label; $\mathcal{PB}_{pp}$: the predicted bounding boxes whose IoUs are greater than a threshold and the predicted label (not the background) is different from the ground-truth category label.

III. Methodology

In this section, we first give the problem formulation in Section III-A. Then, we provide an overview of our method in Section III-B. Next, we describe our Mix-Paste method in detail in Section III-C. Finally, we introduce an LLS strategy, which can be effectively combined with Mix-Paste to alleviate the influence of noisy annotations during model training, in Section III-D.

A. Problem Formulation

Some X-ray datasets involve noisy annotations due to the difficulty of obtaining high-quality human annotations in X-ray images, where item overlapping is prevalent. In this paper, we address the problem of training a robust prohibited item detector on the noisy X-ray dataset, where the noise contains a mixture of category noise and bounding box noise. In addition, we do not assume that a subset with clean annotations is available.

Given a noisy dataset $\mathcal{D} = \{(\mathbf{x}_i, \widetilde{\mathbf{y}}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is the i-the training image and $\widetilde{\mathbf{y}}_i = \{\widetilde{c}_j, \widetilde{\mathbf{b}}_j\}_{j=1}^{J_i}$ denotes the annotation of $\mathbf{x}_i$. Here, $\widetilde{c}_j$ denotes the label of the j-th prohibited item, $\widetilde{\mathbf{b}}_j = (\widetilde{x}, \widetilde{y}, \widetilde{w}, \widetilde{h})$ represents the ground-truth bounding box coordinates ($\widetilde{x}$ and $\widetilde{y}$ represent the coordinates of the top-left corner, and $\widetilde{w}$ and $\widetilde{h}$ represent the width and height, respectively) of the j-th prohibited item, and J_i is the number of prohibited items for $\mathbf{x}_i$. Unlike the image classification task, the prohibited item detection task often involves two types of noise: category noise and bounding box noise. In this way, the dataset contains class-corrupted instances where the category labels are noisy, and position-corrupted instances where the ground-truth bounding boxes are inaccurate. Hence, we aim to train a noise-robust model based on $\mathcal{D}$ and evaluate its performance on the test set.

B. Overview

In this paper, we develop a simple yet effective data augmentation method called Mix-Paste for training on the noisy X-ray dataset. Mix-Paste is a plug-and-play data augmentation method that can be directly applied to the training of different prohibited item detectors. The overview of our method is shown in Fig. 2.

Specifically, for each item patch corresponding to a ground-truth bounding box in the training image, we first randomly select several item patches (specified by the ground-truth bounding boxes with the same category label) from different images. Then, we resize these item patches to the same size and mix them. Finally, we can paste the mixed patch back into the original item location in the image. In this way, the mixture of item patches can generate a new training image with reduced noise interference. Note that instead of applying Mix-Paste to all the images, we apply our proposed Mix-Paste to the randomly selected subset of the training set (with a probability), maintaining consistency between the training and test samples. In this way, some X-ray images are generated by Mix-Paste while the other images are unchanged.

To obtain a robust detector on the augmented data, we further design an item-based large-loss suppression (LLS) strategy, which suppresses the large losses corresponding to potentially positive predictions of additional items during loss calculation. Technically, we first select the predicted bounding boxes, for which the Intersection over Unions (IoUs) between them and the ground-truth bounding boxes are larger than a threshold. Then, we identify those predicted bounding

boxes whose predicted label is different from the ground-truth category label (except for predicted bounding boxes whose predicted label is the background). Finally, the classification losses corresponding to these identified bounding boxes are suppressed and not counted for loss calculation. By doing this, we effectively reduce the negative influence caused by the patch mixing operations (which may involve several different prohibited items due to category noise) for model optimization.

C. Mix-Paste

We generate a new patch by mixing multiple item patches that share the same category label. To mix these item patches, we crop the item patches from the selected images according to the ground-truth bounding boxes and resize them to the same size. Finally, the mixed patch is used to replace the original patch. Note that we only randomly select patches with the same category label from the whole dataset without assuming that the labels of the selected patches are clean.

Mathematically, the process of generating the mixed patch $\hat{\mathbf{B}}$ is formulated as

$$\hat{\mathbf{B}} = \boldsymbol{\alpha} \odot \mathbf{B}_a + \sum_{n=2}^{K} \frac{1-\boldsymbol{\alpha}}{K-1} \odot \text{resize}(\mathbf{B}_n), \tag{1}$$

where K is the total number of patches for mixing (including the original item patch); $\mathbf{B}_a$ is the original item patch in $\mathbf{x}$; $\mathbf{B}_n$ is the n-th item patch randomly selected from the whole dataset; '$\odot$' is the element-wise multiplication; $\text{resize}(\cdot)$ denotes the function that resizes the item patch to the same size as $\mathbf{B}_a$; $\boldsymbol{\alpha}$ is an edge smoothing mask to make the mixed patch more natural.

The edge smoothing mask is defined as

$$\boldsymbol{\alpha}(i,j) = \begin{cases} 1-(1-\lambda)(\boldsymbol{d}_{i,j}/(\beta \cdot w)), & \boldsymbol{d}_{i,j} \leq \beta \cdot w, \\ \lambda, & \boldsymbol{d}_{i,j} > \beta \cdot w, \end{cases} \tag{2}$$

where $\boldsymbol{d}_{i,j}$ is the distance between the pixel (with the spatial location of (i,j) in the patch) to the nearest boundary of the patch; β denotes a threshold to control the smoothing area (we empirically set β to 10%); w is the width of the bounding box; $\lambda \in [0,1]$ is a random number generated from a Beta distribution. Although the edge smoothing mask can allow for the natural appearance of the mixed patch, we also observe that simply merging the patches with linear combinations can also achieve similar performance.

Note that some methods apply threat image projection for image fusion. However, it is difficult to apply threat image projection in our method due to the following several reasons. First, some threat image projection methods [51], [52] require X-ray images with plain backgrounds to segment prohibited items and superimpose isolated prohibited items onto normal images. As most X-ray datasets do not have X-ray images with plain backgrounds, it is not trivial to obtain isolated prohibited items. Second, traditional threat image projection methods [53], [54] only work on the fusion of gray images. However, most current X-ray datasets are color images (notice that the color information of each prohibited item plays an important role in detection due to the penetration characteristics of X-rays). Therefore, these methods cannot be directly used in our method. Although recent methods [52] extend threat image projection to color X-ray images by superimposing pixel-level isolated prohibited items onto normal images, the pixel-level annotations are not available in our X-ray datasets. Third, most threat image projection methods compute the image intensity based on the X-ray energy, object material, and object thickness. In this way, some parameters related to the X-ray scanners (such as the X-ray energy) are required as a prerequisite. But these parameters are not provided in existing public X-ray datasets.

Subsequently, the mixed patch is pasted back into the original image, which can be formulated as

$$\mathbf{x}[\widetilde{x} : \widetilde{x}+\widetilde{w}, \widetilde{y} : \widetilde{y}+\widetilde{h}] = \hat{\mathbf{B}}, \tag{3}$$

where $\mathbf{x}$ denotes the original image corresponding to the item patch; $[\widetilde{x} : \widetilde{x}+\widetilde{w}, \widetilde{y} : \widetilde{y}+\widetilde{h}]$ denote the bounding box region of the original item patch $\mathbf{B}_a$.

Why does Mix-Paste work? We analyze the reasons why our Mix-Paste can work on the training of X-ray prohibited item detection under noisy annotations. First, Mix-Paste can reduce category noise and bounding box noise explicitly. For an annotated bounding box with the prohibited item label $\widetilde{c}_j$ in the dataset involving the category noise rate of P_c, the probability of the prohibited item within the bounding box region is estimated as $1-P_c$. When K item patches with the same category label $\widetilde{c}_j$ are mixed, the probability of the existence of the item with the label $\widetilde{c}_j$ (which is computed as $1-P_c^K$) is increased. Analogously, suppose that the bounding box noise rate is P_b, the probability of the K mixed patch that can accurately bound a correct prohibited item (which is computed as $1-P_b^K$) is also increased. Second, Mix-Paste can effectively mimic item overlapping in X-ray images, thereby enabling the detector to enhance its awareness of overlapping. Third, Mix-Paste can generate more diverse training samples, thereby enhancing the generalization ability of the model.

Can the mixing operation perfectly mimic item overlapping in X-ray images? Some existing data augmentation methods (such as Mix-Up [27]/CutMix [55]) fail to generate data perfectly as the original dataset. However, these data augmentation methods can significantly enhance the model performance in various tasks by substantially increasing the diversity of the dataset. In the same spirit, although the mixing operation in Mix-Paste cannot perfectly mimic item overlapping in X-ray images, it still can encourage the model to learn some characteristics of X-ray images under item overlapping conditions. More importantly, our Mix-Paste is shown to be effective in alleviating the influence of noisy annotations during model training.

Can Mix-Paste be applied to the segmentation task? Unfortunately, our method is difficult to be applied to the segmentation task. The core idea behind our method is to increase the probability of the target prohibited item appearing within a bounding box by mixing different item patches. It is straightforward to adjust the different sizes of bounding boxes for patch mixing since these boxes are rectangular. However, it is not easy to align the object with different shapes at the pixel level for the segmentation task. As a result, our method is more suitable for object detection than segmentation.

Comparison against traditional data augmentation methods. Both our method and traditional data augmentation methods combine the training samples to generate new samples. However, there are some intrinsic differences (in terms of the motivations and methodological details) between our method and traditional methods. First, Mix-Up encourages the model to behave linearly, reflecting a good inductive bias [27]. In contrast, Mix-Paste aims to synthesize more training samples with less annotation noise. Second, Mix-Up, which combines the two images at the image level, is developed for image classification. On the contrary, Mix-Paste, which mixes item patches with the same category label at the patch level, is designed for prohibited item detection. Our experiments further validate that Mix-Paste significantly outperforms Mix-Up for prohibited item detection under noisy annotations. CutMix [55] randomly replaces a patch in the image with another patch from another image. SaliencyMix [56] and Attentive CutMix [57] enhance CutMix by pasting the most salient region onto the corresponding location in the target image. Unlike the above methods, Mix-Paste mixes multiple item patches with the same category label.

D. Item-Based Large-Loss Suppression (LLS) Strategy

After the mixing operation, the probability of the mixed patch containing the correct target prohibited item is increased (the detailed analysis about why our Mix-Paste can work on the training of X-ray prohibited item detection under noisy annotations is given in Section III-C). However, the mixing operation is likely to introduce additional noisy-labeled prohibited items (caused by category noise in some selected patches) during training. In fact, the probability that the selected patches consist of all correct prohibited items is only $(1-P_c)^K$, where P_c denotes the category noise rate and K is the number of patches. Consequently, the mixing operation may introduce additional noisy-labeled prohibited items during training. In such a case, the model tends to predict these items for the newly generated image during training. However, these predicted bounding boxes will be mistakenly considered as false predictions since their corresponding correct labels are not available. Hence, these potentially positive predictions give large losses in the conventional classification loss calculation, resulting in a negative influence on model training.

To alleviate this problem, we propose an item-based large-loss suppression (LLS) strategy. As illustrated in Fig. 2, we categorize the prediction results into four parts, including (1) (1) $\mathcal{PB}_{neg}$: the predicted bounding boxes whose IoUs between them and the ground-truth bounding boxes are less than a threshold (e.g., there is no matching between the ground-truth bounding box and the predicted bounding box in Fig. 2(b)); (2) $\mathcal{PB}_{fb}$: the predicted bounding boxes whose IoUs are greater than a threshold and the predicted label is the background (e.g., the predicted bounding box position is correct but the category label is predicted to the background in Fig. 2(b)); (3) $\mathcal{PB}_{pos}$: the predicted bounding boxes whose IoUs are greater than a threshold and the predicted label is the same as the ground-truth category label (e.g., both the predicted bounding box position and category label are correct in Fig. 2(b)); (4) $\mathcal{PB}_{pp}$: the predicted bounding boxes whose IoUs are greater than a threshold and the predicted label (not the background) is different from the ground-truth category label (e.g., the predicted bounding box position is correct but the predicted category label is incorrect in Fig. 2(b)).

When calculating the classification loss, we suppress $\mathcal{PB}_{pp}$ from the prediction results and focus on the remaining three parts. Hence, the final loss is calculated as

$$\mathcal{L} = \mathcal{L}_{bbox} + \mathcal{L}_{cls_{neg}} + \mathcal{L}_{cls_{pos}} + \mathcal{L}_{cls_{fb}}, \tag{4}$$

where $\mathcal{L}_{bbox}$ denotes the bounding box regression loss; $\mathcal{L}_{cls_{neg}}$, $\mathcal{L}_{cls_{fb}}$, and $\mathcal{L}_{cls_{pos}}$ denote the classification losses for $\mathcal{PB}_{neg}$, $\mathcal{PB}_{fb}$, and $\mathcal{PB}_{pos}$, respectively.

For learning with label noise on image classification, the popular small-loss criterion [16], [17] treats samples with small losses as clean samples and considers samples with large losses as noisy samples. However, for prohibited item detection, the loss calculation contains both foreground and background predictions, where the background predictions account for the majority of the total loss. As a result, the small-loss criterion mainly focuses on background predictions and may ignore foreground predictions in this task. In contrast, our LLS strategy is highly effective in handling category noise by removing only the potentially positive predictions.

Why does the LLS strategy work? In the LLS strategy, we ignore the predicted bounding boxes whose predicted labels are different from the ground-truth category labels (except for those whose predicted label is the background) when calculating the classification loss. In noisy scenarios, the mixed patches may contain multiple prohibited items because of category noise. Consequently, when these newly generated images are used for training, the model tends to give correct predictions for additional items. However, since these items are not associated with correct labels, they are considered as false predictions and consequently give large losses during model training. This will lead to incorrect model optimization, reducing the overall performance of the model. To mitigate this and enhance the model performance, we suppress the large losses corresponding to potentially positive predictions of additional items for loss calculation. Note that we still compute the loss for those predicted bounding boxes whose predicted labels are the background since the predicted bounding box region contains a prohibited item.

IV. Experiments

In this section, we first introduce the datasets and evaluation metrics in Section IV-A. Then, we present the noise rate estimation and implementation details 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 noisy X-ray datasets in Section IV-D and the noisy MS-COCO dataset in Section IV-E. After that, we conduct ablation studies in Section IV-F. Finally, we give some visualization results in Section IV-G.

A. Datasets

In this paper, we conduct experiments on two popular X-ray datasets, i.e., OPIXray [1] and PIDray [3]. OPIXray

TABLE I
COMPARISON RESULTS (%) ON THE OPIXRAY DATASET. P_c AND P_b DENOTE THE CATEGORY NOISE RATE AND BOUNDING BOX NOISE RATE, RESPECTIVELY.

Method	$P_c = 20\%$ $P_b = 20\%$		$P_c = 40\%$ $P_b = 40\%$		$P_c = 60\%$ $P_b = 60\%$	
	mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	81.1 (+0.0)	31.5 (+0.0)	70.0 (+0.0)	25.7 (+0.0)	56.7 (+0.0)	18.4 (+0.0)
LIM (ICCV, '21) [4]	83.7 (+2.6)	34.6 (+3.1)	80.2 (+10.2)	31.3 (+5.6)	72.7 (+16.0)	26.4 (+8.0)
SDANet (IJCV, '23) [3]	83.9 (+2.8)	32.6 (+1.1)	71.2 (+1.2)	25.1 (-0.6)	52.4 (-4.3)	17.1 (-1.3)
GADet (SENS J., '24) [25]	81.2 (+0.1)	34.5 (+3.0)	77.5 (+7.5)	32.5 (+6.8)	69.7 (+13.0)	27.9 (+9.5)
SCE (ICCV, '19) [10]	81.8 (+0.7)	32.5 (+1.0)	72.1 (+2.1)	25.2 (-0.5)	48.6 (-8.1)	16.1 (-2.3)
LNCIS (ECCV, '20) [13]	84.3 (+3.2)	34.2 (+2.7)	80.4 (+10.4)	29.6 (+3.9)	65.9 (+9.2)	23.4 (+5.0)
OA-MIL (ECCV, '22) [14]	82.2 (+1.1)	31.3 (-0.2)	70.2 (+0.2)	27.5 (+1.8)	56.4 (-0.3)	21.6 (+3.2)
Ours	**87.0 (+5.9)**	**38.3 (+6.8)**	**86.7 (+16.7)**	**37.0 (+11.3)**	**81.8 (+25.1)**	**33.7 (+15.3)**

TABLE II
COMPARISON RESULTS (%) ON THE PIDRAY DATASET. P_c AND P_b DENOTE THE CATEGORY NOISE RATE AND BOUNDING BOX NOISE RATE, RESPECTIVELY. WE REPORT MAP@[.5, .95] AS THE EVALUATION METRIC.

Method	$P_c = 40\%$ $P_b = 40\%$				$P_c = 60\%$ $P_b = 60\%$			
	easy	hard	hidden	Avg	easy	hard	hidden	Avg
FRCNN (PAMI, '17) [58]	42.5 (+0.0)	40.1 (+0.0)	19.9 (+0.0)	34.2 (+0.0)	29.4 (+0.0)	27.4 (+0.0)	13.3 (+0.0)	23.4 (+0.0)
LIM (ICCV, '21) [4]	53.6 (+11.1)	49.2 (+9.1)	27.6 (+7.7)	43.5 (+9.3)	43.8 (+14.4)	40.5 (+13.1)	19.9 (+6.6)	34.7 (+11.3)
SDANet (IJCV, '23) [3]	40.8 (-1.7)	37.8 (-2.3)	19.3 (-0.6)	32.6 (-1.6)	26.0 (-3.4)	25.0 (-2.4)	11.1 (-2.2)	20.7 (-2.7)
GADet (SENS J., '24) [25]	47.9 (+5.4)	42.9 (+2.8)	18.9 (-1.0)	36.6 (+2.4)	41.3 (+11.9)	34.5 (+7.1)	16.7 (+3.4)	30.8 (+7.4)
SCE (ICCV, '19) [10]	40.3 (-2.2)	37.9 (-2.2)	21.5 (1.6)	33.2 (-1.0)	24.7 (-4.7)	23.3 (-4.1)	11.1(-2.2)	19.7(-3.7)
LNCIS (ECCV, '20) [13]	48.2 (+5.7)	45.2 (+5.1)	25.9 (+6.0)	39.8 (+5.6)	33.4 (+4.0)	30.9 (+3.5)	14.5 (+1.2)	26.3 (+2.9)
OA-MIL (ECCV, '22) [14]	44.6 (+2.1)	39.6 (-0.5)	23.8 (+3.9)	36.0 (+1.8)	30.5 (+1.1)	25.0 (-2.4)	10.6 (-2.7)	22.0 (-1.4)
Ours	**57.4 (+14.9)**	**53.2 (+13.1)**	**31.9 (+12.0)**	**47.5 (+13.3)**	**51.3 (+21.9)**	**47.9 (+20.5)**	**21.9 (+8.6)**	**40.4 (+17.0)**

contains 8,885 images with 5 categories of prohibited items (i.e., different types of cutters). Following [1], we use 7,109 images for training and 1,776 images for testing. We report mAP@.5 and mAP@[.5, .95] as the evaluation metrics. PIDray contains 29,457 images for training and 18,220 images for testing, covering 12 different categories. The images in the test set are further divided into 3 subsets (i.e., easy, hard, and hidden) according to their detection difficulty. Following [3], we use 29,457 images for training and 18,220 images for testing. We report mAP@[.5, .95] as the evaluation metric.

To validate the generalization ability of our method to common object detection, we also conduct experiments on MS-COCO [18]. MS-COCO is a public common object detection dataset, which contains more than 135k images for training and 5k images for testing, covering 80 different categories. Following [12], we use $train2017$ as training data, and report mAP@.5 and mAP@[.5, .95] on $val2017$.

B. Noise Rate Estimation

Our research has shown variability in noise rates across different X-ray datasets. Specifically, while some X-ray datasets (such as the PIDray dataset) exhibit minimal noisy annotations, we identify that some X-ray datasets (such as the OPIXray dataset) contain a number of noisy annotations (including both category noise and bounding box noise). In these datasets, many bounding box annotations are larger than the actual position of prohibited items while the category labels of some prohibited items are mislabeled due to the great similarity between some prohibited items.

TABLE III
BOUNDING BOX NOISE RATE ESTIMATION OF THE OPIXRAY DATASET.

Category	Total Samples	Noise Samples	Noise Rate
Scissor	1494	73	4.89%
Utility Knife	1635	70	4.28%
Multi-Tool Knife	1612	81	5.02%
Straight Knife	809	34	4.20%
Folding Knife	1589	33	2.08%
Total	7139	291	4.08%

We conduct a quantitative analysis of the noise rate on the OPIXray dataset. Specifically, for bounding box noise, we first train a model on the original dataset by treating all the categories as one category. In this way, the influence of category noise is removed. Then, we compare the detection results with the ground-truth labels and filter out samples whose IoUs are less than 0.70. These samples are manually checked to identify noisy-labeled samples. For category noise, we randomly select a certain number of samples (500 samples

in total) in the dataset and manually check whether they are labeled incorrectly. Based on the above steps, the category noise rate in the OPIXray dataset is estimated as about 5%. The bounding box noise rate in the OPIXray dataset is estimated as about 4%. We also estimate the bounding box noise rate for each class in the OPIXray dataset in Table III. The above analysis validates the existence of noisy annotations in some X-ray datasets.

C. Implementation Details

To effectively evaluate the performance of our method under noisy annotations, we introduce different types of noise to the original dataset. For category noise, we randomly replace the original category label with another category label, with a replacement probability of P_c. For bounding box noise, we randomly perturb the original bounding box with a probability of P_b. Specifically, for a bounding box with coordinates (x, y, w, h), we randomly perturb the coordinates with a probability of P_b by shifting and scaling the box as follow:

$$\begin{aligned} \widetilde{x} &= x + \Delta_x \times w, \\ \widetilde{y} &= y + \Delta_y \times h, \\ \widetilde{w} &= w \times (1 + \Delta_w), \\ \widetilde{h} &= h \times (1 + \Delta_h), \end{aligned} \tag{5}$$

where Δ_x, Δ_y, Δ_w, and Δ_h are randomly sampled from a uniform distribution $U(-\delta, \delta)$ (δ is the perturbation level). We set δ to 0.3 in all experiments.

For all the datasets, we adopt Faster R-CNN (FRCNN) [58] with ResNet-50 as the backbone network. The backbone is initialized with the weights pretrained on ImageNet [59]. The whole network is optimized by the stochastic gradient descent (SGD) algorithm with a momentum of 0.9 and a weight decay of 0.0001. The batch size is set to 2. The initial learning rate is set to 0.005 and decreased by a factor of 10 at the 17th and 21st epochs. The total number of training epochs is 24. The number of item patches K for mixing is set to 2. We apply Mix-Paste to the training set with a probability of 0.6. We only employ the random flip to all the comparison methods. All the competing methods are trained on a machine with an NVIDIA RTX 3090 GPU.

D. Experiments on the X-Ray Datasets

To verify the effectiveness of our method, we perform experiments on two X-ray datasets with different levels of noise rates for both category noise and bounding box noise. We compare our method with several state-of-the-art methods, including the baseline method (FRCNN [58]), prohibited item detection methods (LIM [4], SDANet [3], and GADet [25]), and learning with noisy annotations methods (SCE [10], LNCIS [13], and OA-MIL [14]). The results are shown in Table I and Table II.

For the **OPIXray** dataset, we can observe that the noise-robust loss function-based method SCE does not perform well. Compared with the baseline, SCE only gives marginal performance improvements at low noise rates. Moreover, when the noise rates are large, the performance obtained by SCE is even lower than that obtained by the baseline method. The performance degradation of SCE can be attributed to its limited ability to handle the bounding box noise. In other words, when both the bounding box noise rate and the category noise rate are high, the performance of SCE is severely affected. The X-ray prohibited item detector LIM shows relatively good anti-noise ability. At some noise rates, it even performs better than LNCIS, a method specifically designed to deal with object detection noise. This is because LIM can filter out irrelevant noisy information in features, making it less prone to overfit the noise. The X-ray prohibited item detector GADet also demonstrates good performance in terms of mAP@[.5,.95]. This can be attributed to its IoU-aware label assignment strategy, which selects high-quality and precise positive samples while ignoring potentially noisy low-quality predictions. Among all the competing methods, our Mix-Paste method achieves the best results at all noise rates. Specifically, our method achieves 81.8% mAP@.5 and 33.7% mAP@[.5, .95] (at the noise rates of $P_c = 60\%$ and $P_b = 60\%$), which is 25.1% and 15.3% higher than the baseline, respectively.

TABLE IV
COMPARISON RESULTS (%) ON THE MS-COCO DATASET. P_c AND P_b DENOTE THE CATEGORY NOISE RATE AND BOUNDING BOX NOISE RATE, RESPECTIVELY.

Method	$P_c = 40\%$	$P_b = 40\%$	$P_c = 60\%$	$P_b = 60\%$
	mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	50.4 (+0.0)	30.2 (+0.0)	43.2 (+0.0)	23.3 (+0.0)
SCE (ICCV, '19) [10]	44.6 (-5.8)	27.1 (-3.1)	36.8 (-6.4)	20.5 (-2.8)
LNCIS (ECCV, '20) [13]	50.9 (+0.5)	30.9 (+0.7)	44.2 (+1.0)	24.8 (+1.5)
OA-MIL (ECCV, '22) [14]	47.4 (-3.0)	28.2 (-2.0)	43.8 (+0.6)	24.7 (+1.4)
Ours	**51.2** (+0.8)	**31.5** (+1.3)	**45.3** (+2.1)	**26.2** (+2.9)

For the **PIDray** dataset, the X-ray prohibited item detector LIM and GADet shows good performance. LNCIS can also alleviate the negative influence of noisy annotations to a certain extent. However, it exhibits only marginal performance improvements at high noise rates. At some high noise rates, the performance obtained by OA-MIL is inferior to the baseline, indicating its instability. Compared with the other competing methods, our method consistently gives the best results across all noise rates. Specifically, at the noise rates of $P_c = 60\%$ and $P_b = 60\%$, our method achieves a mAP@[.5, .95] of 51.3%, 47.9%, and 21.9% in the easy, hard, and hidden test sets, which is 21.9%, 20.5%, and 8.6% higher than the baseline, respectively.

The above results show that our method can greatly improve the performance of the model at both low and high noise rates and enhance the robustness of the model.

E. Experiments on the MS-COCO Dataset

Our method is mainly designed for prohibited item detection under noisy annotations by considering the characteristics of X-ray images, where item overlapping is ubiquitous. Interestingly, object overlapping also exists in some natural images

TABLE V
ABLATION STUDY RESULTS (%) ON THE KEY COMPONENTS OF OUR METHOD ON THE OPIXRAY DATASET.

Method	Mix-Paste	LLS	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]			56.7	18.4
Ours	✓		80.3	31.5
	✓	✓	81.8	33.7

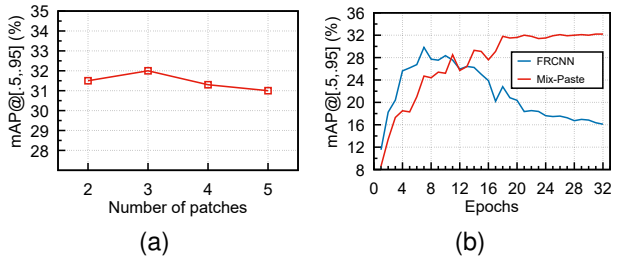

Fig. 3. (a) Ablation study results on the influence of the number of patches for Mix-Paste on the OPIXray dataset. (b) Training curve comparison between our Mix-Paste and the baseline method on the OPIXray dataset.

for the common object detection task. Hence, the idea of increasing the probability of target objects in the generated images by fusing the same category label can be also applied to common object detection.

To evaluate the generalizability of our proposed method, we conduct experiments on the widely used common object detection dataset, the MS-COCO dataset. The evaluation results are shown in Table IV. From Table IV, we observe that SCE performs poorly. LNCIS shows moderate performance improvements over the baseline method (the vanilla FRCNN). The performance of OA-MIL is unstable, and its performance is even worse than the baseline at some noise rates (e.g., $P_c = 40\%$ and $P_b = 40\%$). Compared with the other competing methods, our method gives better performance at different noise rates. Specifically, at the noise rates of $P_c = 60\%$ and $P_b = 60\%$, our method achieves 45.3% mAP@.5 and 26.2% mAP@[.5, .95], which is 2.1% and 2.9% higher than the baseline, respectively. These results demonstrate the effectiveness of our method on the noisy MS-COCO dataset, indicating the robustness and broad applicability of our method to data beyond X-ray images.

F. Ablation Studies

We conduct ablation studies to study the effectiveness of each component in our method. Unless otherwise specified, the noise rates are set to P_c = 60% and P_b = 60% and the LLS strategy is not used (we focus on the evaluation of Mix-Paste). The OPIXray dataset is used.

Effectiveness of Mix-Paste and LLS. The ablation study results on the key components of our method are shown in Table V. We can see that the performance of our method with only Mix-Paste is better than the baseline by 13.1% mAP@[.5, .95]. This demonstrates the effectiveness of Mix-Paste, which mixes multiple item patches with the same category label to alleviate the influence of noisy annotations. After further applying the LLS strategy, the performance of our method is further improved by 2.2%, verifying the importance of the LLS strategy, which ignores the potentially positive predictions during the mixing process.

TABLE VI
ABLATION STUDY RESULTS (%) ON THE INFLUENCE OF THE EDGE SMOOTHING MASK ON THE OPIXRAY DATASET WITHOUT INTRODUCING ANY ADDITIONAL NOISE. ONLY MIX-PASTE IS USED IN THIS EXPERIMENT.

Method	Linear Combination	Edge Smoothing Mask
mAP@[.5,.95]	31.3	31.5

TABLE VII
COMPARISON OF TRAINING TIME, INFERENCE TIME AND PERFORMANCE (%) ON THE ORIGINAL OPIXRAY DATASET WITHOUT INTRODUCING ADDITIONAL NOISE.

Method	mAP@.5	mAP@[.5,.95]	Training Time	Inference Time(fps)
FRCNN (PAMI, '17) [58]	86.1	36.9	4h 27min	18.7
LIM (ICCV, '21) [4]	88.6	38.9	19h 18min	7.3
SDANet (IJCV, '23) [3]	88.1	38.3	6h 20min	16.2
SCE (ICCV, '19) [10]	85.9	36.7	4h 28min	19.3
LNCIS (ECCV, '20) [13]	88.1	37.6	4h 30min	19.3
OA-MIL (ECCV, '22) [14]	87.3	37.4	4h 33min	19.4
Mix-Paste + LLS	**90.1**	**40.3**	4h 30min	19.0

Influence of the Number of Patches K. We investigate the influence of the number of patches K used in Mix-Paste, as shown in Fig. 3(a). Our method gives a good performance when the values of K are set to 2 and 3. However, when the value of K becomes large (e.g., 4 or 5), the performance obtained by our method decreases. When more patches are mixed, the likelihood of capturing the correct prohibited item is increased. However, such a way also raises the probability of introducing additional prohibited items. As a result, excessive patch mixing can negatively influence the extraction of relevant information in the original patch, hindering the learning of correct prohibited items.

Training Curve. To verify whether our Mix-Paste can help alleviate the overfitting of noise during model training, we plot the training curve in Fig. 3(b). We can see that the performance obtained by the baseline model increases at the early training stage but gradually decreases at the later training stage. In contrast, the performance obtained by our method is relatively stable at the later training stage. This demonstrates that our method can effectively alleviate the overfitting of the model to noise during training.

Effectiveness of the Patch Mixing Strategy. In Mix-Paste, we mix multiple item patches with the same category label to generate a mixed patch and paste it back into the original image for data augmentation. To show the effectiveness of our patch mixing strategy, we compare the performance between our strategy and a variant (which mixes multiple randomly selected item patches with different category labels). The

TABLE VIII
ABLATION STUDY RESULTS (%) ON THE EFFECTIVENESS OF THE PATCH MIXING STRATEGY. RANDOM-MIX DENOTES THE METHOD THAT SELECTS TWO RANDOM ITEM PATCHES FOR THE MIXING OPERATION. THE NOISE RATES ARE SET TO P_c = 60% AND P_b = 60%

Method	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	56.7	18.4
Random-Mix	55.7	21.0
Mix-Paste	**80.3**	**31.5**

results are shown in Table VIII.

We can see that our patch mixing strategy achieves better performance than the variant. Mixing item patches with different category labels not only brings additional disturbances caused by different prohibited items, but also does not increase the probability of containing the correct prohibited item in the generated image, thereby decreasing the final performance. The above results further validate the superiority of our patch mixing strategy.

Effectiveness of the Edge Smoothing Mask. For our Mix-Paste, we use an edge smoothing mask to mix multiple item patches, generating images with more natural appearances (some generated images are illustrated in Fig. 4). We conduct experiments to investigate the influence of the edge smoothing mask. We compare the edge smoothing mask with the simple linear combination of multiple item patches (i.e., all the pixels in the edge smoothing mask are set to a fixed value). The results are given in Table VI. We can see that the performance obtained by our method with the edge smoothing mask is only slightly better than that with the linear combination. This can be ascribed to the fact that the model focuses on the appearance of the prohibited items, and thus it does not pay too much attention to the edges of the mixed patches during the learning process. This result also aligns with the principle of Occam's razor, where the linear combination indicates a simple fusion method.

Effectiveness on the Original X-ray Dataset. We conduct experiments to investigate the effectiveness of our method on the original X-ray dataset. We test all the competing methods on the OPIXray dataset without introducing additional noise. We also report the training time and inference time of different methods. The results are shown in Table VII. From Table VII, we can observe that our method achieves better performance than the other competing methods. This is because the original dataset also contains a certain amount of noisy annotations, and our method can effectively alleviate the negative influence of noisy annotations during the model training. Note that although the LIM and SDANet methods also outperform the baseline method (FRCNN), the training time and inference time of these methods is significantly longer.

Effectiveness on Different Category Noise Rates and Bounding Box Noise Rates. We conduct experiments to investigate the effectiveness of our method on different category noise rates and bounding box noise rates on the OPIXray dataset. The results are shown in Table IX. From the results, we can see that OA-MIL is good at addressing category noise,

TABLE IX
COMPARISON RESULTS (%) OF DIFFERENT CATEGORY NOISE RATES AND BOUNDING BOX NOISE RATES ON THE OPIXRAY DATASET. P_c AND P_b DENOTE THE CATEGORY NOISE RATE AND THE BOUNDING BOX NOISE RATE, RESPECTIVELY.

Method	$P_c = 20\%$ $P_b = 40\%$		$P_c = 40\%$ $P_b = 20\%$	
	mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	78.4 (+0.0)	28.9 (+0.0)	72.7 (+0.0)	28.5 (+0.0)
LIM (ICCV, '21) [4]	82.0 (+3.6)	32.2 (+3.3)	81.5 (+8.8)	32.7 (+4.2)
SDANet (IJCV, '23) [3]	80.1 (+1.7)	28.3 (-0.6)	74.4 (+1.7)	29.6 (+1.1)
GADet (SENS J., '24) [25]	77.5 (-0.9)	32.1 (+3.1)	77.9 (+5.2)	34.0 (+5.5)
SCE (ICCV, '19) [10]	76.7 (-1.7)	27.3 (-1.6)	75.4 (+2.7)	29.8 (+1.3)
LNCIS (ECCV, '20) [13]	78.9 (+0.5)	28.5 (-0.4)	81.9 (+9.2)	33.3 (+4.8)
OA-MIL (ECCV, '22) [14]	54.8 (+3.8)	20.8 (+2.4)	73.3 (+0.6)	28.2 (-0.3)
Mix-Paste + LLS	**86.2** (+7.8)	**36.9** (+8.0)	**86.3** (+13.6)	**37.7** (+9.2)

TABLE X
ABLATION STUDY RESULTS (%) ON THE INFLUENCE OF THE PROBABILITY OF APPLYING MIX-PASTE ON THE OPIXRAY DATASET. ONLY MIX-PASTE IS USED IN THIS EXPERIMENT.

Probability	0	0.2	0.4	0.6	0.8	1
mAP@[.5,.95]	18.4	31.2	31.4	31.5	30.8	0.4

and LNCIS works well on handling bounding box noise. Among all the competing methods, our method can effectively deal with both category noise and bounding box noise, and achieve the best results in all the cases. This proves that our method has a strong anti-noise ability.

Influence of the Probability of Applying Mix-Paste. We investigate the influence of the probability p of applying Mix-Paste for augmentation during training. The results are shown in Table X. We can see that when the value of p is 0.6, our method can achieve the best performance. When the value of p approaches 1, the performance obtained by our method decreases or even the training fails. This is because when p is close to 1, all samples are artificially generated. Such a manner can lead to significant inconsistency between the training samples and the test samples, making the distribution of the training set greatly inconsistent with that of the test set. As a consequence, our method is unable to learn the true data distribution from the training set. In this paper, we fix $p = 0.6$ in all experiments.

Influence of the Perturbation Level. We conduct experiments to investigate the influence of perturbation level (which is used to generate bounding box noise). We test our method under two perturbation levels (i.e., $\delta = 0.5$ and $\delta = 0.3$) on the OPIXray dataset. The results are shown in Table XI. We can see that even when the perturbation level is large ($\delta = 0.5$), our method can still achieve the best results, while other methods suffer from a serious performance decline. This shows that our method has a strong anti-noise ability in the case of high perturbation levels.

Influence of Different Detectors. Our Mix-Paste, which is a data augmentation method, can be applied to different object detectors. We evaluate the performance obtained by our

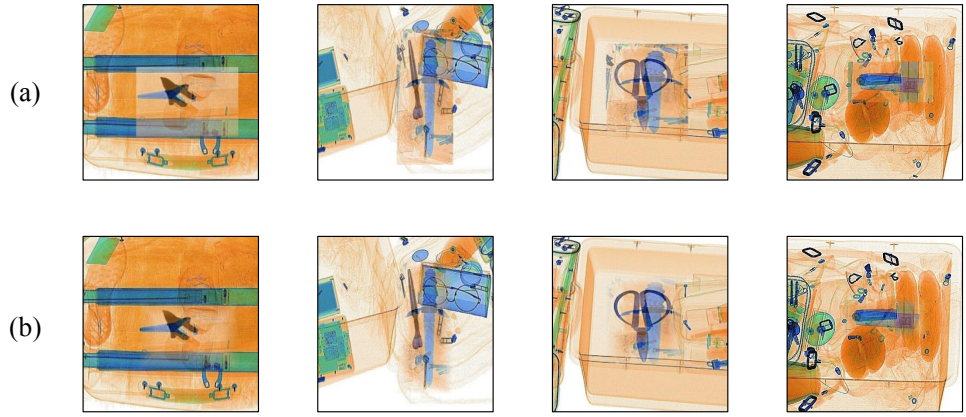

Fig. 4. Examples of generated images on the OPIXray dataset. (a) The generated images with the linear combination. (b) The generated images with the edge smoothing mask.

TABLE XI
COMPARISON RESULTS (%) UNDER DIFFERENT PERTURBATION LEVELS ON THE OPIXRAY DATASET.

Method	$\delta = 0.3$		$\delta = 0.5$	
	$P_c = 60\%$	$P_b = 60\%$	$P_c = 60\%$	$P_b = 60\%$
	mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	56.7 (+0.0)	18.4 (+0.0)	42.7 (+0.0)	12.7 (+0.0)
LIM (ICCV, '21) [4]	72.7 (+15.0)	26.4 (+8.0)	65.7 (+23.0)	23.6 (+10.9)
SDANet (IJCV, '23) [3]	52.4 (-4.3)	17.1 (-1.3)	38.2 (-4.5)	10.7 (-2.0)
GADet (SENS J., '24) [25]	69.7 (+13.0)	27.9 (+9.5)	59.6 (+16.9)	24.1 (+7.4)
SCE (ICCV, '19) [10]	48.6 (-8.1)	16.1 (-2.3)	35.0 (-7.7)	10.5 (-2.2)
LNCIS (ECCV, '20) [13]	65.9 (+9.2)	23.4 (+5.0)	46.6 (+3.9)	14.0 (+1.3)
OA-MIL (ECCV, 22) [14]	56.4 (-0.3)	21.6 (+3.2)	44.9 (+2.2)	15.0 (+2.3)
Mix-Paste+LLS	**81.8** (+25.1)	**33.7** (+15.3)	**76.1** (+33.4)	**29.9** (+17.2)

TABLE XII
THE MAP@[.5, .95] (%) AND MAP@.5 OBTAINED BY DIFFERENT DETECTORS ON THE OPIXRAY DATASET WITH NOISE RATE OF $P_c = 60\%$ AND $P_b = 60\%$. ONLY MIX-PASTE IS USED IN THIS EXPERIMENT.

Method	Original		+Mix-Paste	
	mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	56.7	18.4	**80.3**	**31.5**
RetinaNet (ICCV, '17) [60]	56.1	20.8	61.7	26.1
Cascade RCNN (CVPR, '18) [61]	47.6	16.0	78.6	31.6
ATSS (CVPR, '20) [62]	55.5	18.7	68.2	28.2
LIM (ICCV, '21) [4]	**72.7**	**26.4**	78.1	31.1
SDANet (IJCV, '23) [3]	52.4	17.1	77.8	31.6

method on different detectors, including two-stage detectors (FRCNN [58] and Cascade RCNN [61]), one-stage detectors (RetinaNet [60] and ATSS [62]), and X-ray prohibited item detectors (SDANet [3] and LIM [4]). For all the detectors, we use the default hyper parameters in the MMDetection framework [63] for training. The results are shown in Table XII. We can observe that all the detectors with our method

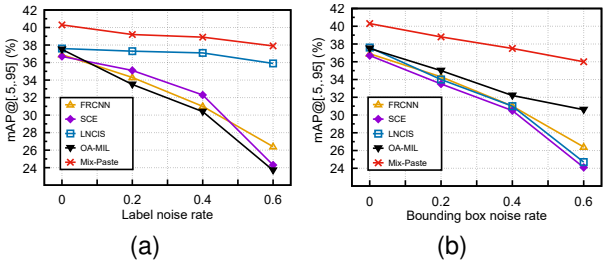

Fig. 5. Performance comparison between Mix-Paste and the competing methods at different noise rates under (a) category noise and (b) bounding box noise on the OPIXray dataset.

outperform those without our method. These results show the superiority of our proposed Mix-Paste. Note that the experimental settings in Table XII and Table VII are different. In Table XII, the results are obtained on the noisy OPIXray dataset under the settings that both the bounding box noise rate and the label noise rate are 60%. Hence, the mAP obtained by FRCNN is low (18.4%). In Table VII, the results are obtained on the original OPIXray dataset without introducing any synthetic noise. Hence, the mAP obtained by FRCNN in Table VII is higher than that in Table XII.

Robustness to Different Types of Noise. We investigate the robustness of our method at different noise rates under two types of noise, including category noise and bounding box noise. The results are shown in Fig. 5. For category noise, it is evident that both SCE and OA-MIL struggle to mitigate the adverse influence of category noise on the model. In contrast, both LNCIS and our method show great effectiveness in dealing with category noise. Notably, our method exhibits the best performance at different noise rates, demonstrating its effectiveness in handling category noise. For bounding box noise, our method outperforms the baseline method by a large margin at the high bounding box noise rates. Moreover, the performance obtained by some competing methods (such as SCE and LNCIS) shows a significant performance decline

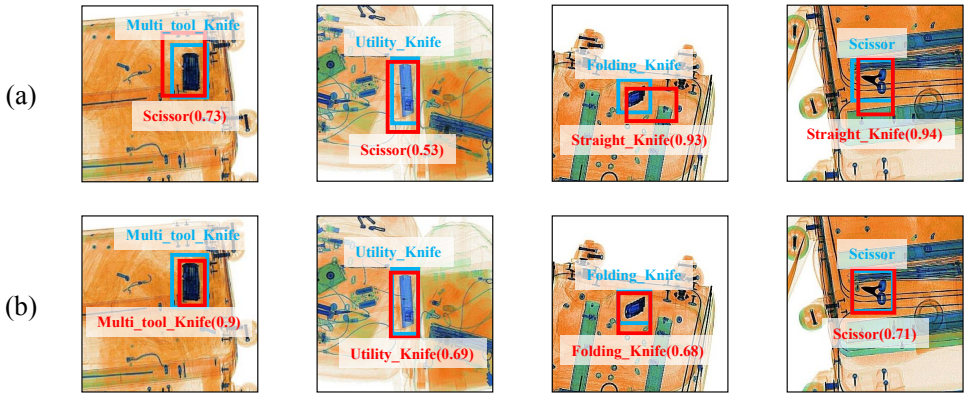

Fig. 6. Some detection results on the OPIXray dataset. (a) The detection results obtained by the baseline method (FRCNN). (b) The detection results obtained by our method. The model is trained under the noise rates of $P_l = 60\%$ and $P_b = 60\%$. The blue annotation and bounding box denote the ground-truth annotation and bounding box, and the red annotation and bounding box denote the detection result.

TABLE XIII
COMPARISON RESULTS (%) AGAINST SOME POPULAR DATA AUGMENTATION METHODS ON THE OPIXRAY DATASET.

Method	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	56.7	18.4
Mix-Up (arXiv, '17) [27]	46.2	15.5
Cutout (arXiv, '17) [26]	58.7	21.0
Mosaic (arXiv, '20) [29]	58.7	19.9
Color jitter (Commun. ACM, '17) [59]	57.6	19.4
Blur	61.7	21.8
Mix-Paste	**80.3**	**31.5**

TABLE XIV
COMPARISON RESULTS (%) AGAINST THE SMALL-LOSS CRITERION ON THE OPIXRAY DATASET.

Method	mAP@.5	mAP@[.5, .95]
Mix-Paste	80.3	31.5
Mix-Paste+Small-Loss V1	64.2	23.6
Mix-Paste+Small-Loss V2	62.8	23.1
Mix-Paste+Small-Loss V3	62.4	21.0
Mix-Paste+LLS	**81.8**	**33.7**

at high bounding box noise rates. This further validates the superiority of our method in handling bounding box noise.

Comparison against Popular Data Augmentation Methods. Table XIII shows the comparison results between our method and some popular data augmentation methods. We observe that the performance obtained by Mix-Up is even lower than that obtained by the baseline method. Our method is specifically designed to address the challenge of noisy annotations while Mix-Up does not consider such a challenge. In fact, the mixing process in Mix-Up cannot reduce the noise, and can significantly introduce additional noisy samples, leading to model overfitting to noisy annotations. Moreover, our method greatly outperforms other competing methods (such as Mosaic, Cutout and Color jitter), showing its effectiveness.

Comparison against the Small-Loss Criterion. We compare our LLS with the small-loss criterion. The small-loss criterion is a widely used method in label noise learning for the image classification task. Due to the difference between the object detection task and the image classification task, we implement three different versions of the small-loss criterion. The first version (Small-Loss V1) is that we select samples with small losses from all classification losses as clean samples and add them to the loss calculation. The second version (Small-Loss V2) is that we divide all classification results into two categories: positive predictions (whose IoUs between predicted boxes and the ground truth are greater than a threshold) and negative predictions (whose IoUs between predicted boxes and the ground truth are less than a threshold), and select a certain proportion of small-loss predictions from the two parts and add them to the loss calculation. The third version (Small-Loss V3) is that we select a certain proportion of samples with small losses from the total loss (both the classification loss and the regression loss) as clean samples and add them to the loss calculation. For a fair comparison, both the small-loss criterion and our LLS are based on our Mix-Paste. The results are shown in Table XIV. The clean sample proportion in the small-loss criterion is set to $1 - \tau \cdot min(T/5, 1)$ as done in [9], where τ is the noise rate (we set it to 0.6) and T is the training epoch.

We can see that the three versions of the small-loss criterion fail to improve the performance of Mix-Paste. On the contrary, our LLS can be effectively combined with Mix-Paste to improve the performance, which validates the importance of LLS in learning with noisy annotations.

Influence of Bounding Box Noise Distribution. Existing noisy-robust object detection methods (such as [12], [14]) often apply the uniform distribution to generate bounding box noise. In this paper, we follow the same settings as these methods [12], [14]. In this subsection, we also evaluate the effectiveness of our method by applying the Gaussian distribution to generate bounding box noise. Specifically, we

TABLE XV
ABLATION STUDY RESULTS (%) ON THE INFLUENCE OF THE BOUNDING BOX NOISE DISTRIBUTION ON THE OPIXRAY DATASET. WE ADD NOISE TO THE BOUNDING BOX USING THE GAUSSIAN DISTRIBUTION.

Method	$P_c = 60\%$ $\mu = 0$ $\sigma = 0.1$		$P_c = 60\%$ $\mu = 0$ $\sigma = 0.2$	
	mAP@.5	mAP@[.5, .95]	mAP@.5	mAP@[.5, .95]
FRCNN (PAMI, '17) [58]	57.6 (+0.0)	20.0 (+0.0)	39.1 (+0.0)	10.4 (+0.0)
LIM (ICCV, '21) [4]	72.6 (+15.0)	28.1 (+8.1)	60.4 (+21.3)	19.3 (+8.9)
SDANet (IJCV, '23) [3]	57.2 (-0.4)	19.9 (-0.1)	34.8 (-4.3)	8.9 (-1.5)
GADet (SENS J., '24) [25]	69.4 (+11.8)	28.3 (+8.3)	57.3 (+18.2)	20.5 (+10.1)
SCE (ICCV, '19) [10]	53.1 (-4.5)	18.4 (-1.6)	58.7 (+19.6)	21.0 (+10.6)
LNCIS (ECCV, '20) [13]	71.3 (+13.7)	26.5 (+6.5)	42.2 (+3.1)	11.7 (+1.3)
OA-MIL (ECCV, '22) [14]	54.8 (-2.8)	20.8 (+0.8)	49.0 (+9.9)	16.5 (+6.1)
Mix-Paste+LLS	**84.2** (+26.6)	**35.8** (+15.8)	**72.1** (+33.0)	**25.7** (+15.3)

apply Gaussian noise to the training dataset and evaluate the detection performance. The shifting and scaling are changed as $\widetilde{x} = x + N(\mu, \sigma^2) \times w$, $\widetilde{y} = y + N(\mu, \sigma^2) \times h$, $\widetilde{w} = w \times (1 + N(\mu, \sigma^2))$, $\widetilde{h} = h \times (1 + N(\mu, \sigma^2))$, where $N(\cdot)$ denotes the Gaussian distribution which is characterized by the mean (μ) and the standard deviation (σ). Note that in our experiments $\widetilde{x}, \widetilde{y}, \widetilde{w}$ and $\widetilde{h}$ are constrained to be positive numbers. The results are given in Table XV.

From Table XV, we can observe that our method can also effectively reduce the influence of bounding box noise generated by the Gaussian distribution, showing the robustness of our method.

G. Visualization Results

We visualize some detection results obtained by the baseline method (i.e., FRCNN) and our method on the OPIXray dataset, as shown in Fig. 6. We can see that the baseline method gives false predictions with high confidence. In some cases, both the predicted category labels and predicted bounding box coordinates are inaccurate (see the images in the first row). This is because the training of the baseline model is easily affected by noisy annotations. On the contrary, our method can give more accurate and correct predictions. This indicates that our method is a simple yet effective data augmentation method, which can significantly improve the detection performance of X-ray prohibited items for learning with noisy annotations.

V. CONCLUSION

In this paper, we address the problem of training a robust X-ray prohibited item detector under noisy annotations from the novel perspective of data augmentation. We propose Mix-Paste by mixing multiple item patches with the same category label and generating a new image involving a mixed patch. Such a manner not only effectively increases the probability of containing the correct prohibited item but also mimics item overlapping in X-ray images. Moreover, we design an LSS strategy for loss calculation. Our strategy alleviates the negative influence of mistakenly treating potentially positive predictions as false predictions caused by the mixing process of item patches. We perform extensive experiments on two X-ray datasets to demonstrate the effectiveness of our method in training a noise-robust detector. We also conduct experiments on the MS COCO dataset to verify the generalization ability of our method on the common object detection task. These results demonstrate the benefits of data augmentation in tackling the challenges posed by learning with noisy annotations.

REFERENCES

[1] 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.

[2] 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 Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2114–2123.

[3] L. Zhang, L. Jiang, R. Ji, and H. Fan, "PIDray: A large-scale X-ray benchmark for real-world prohibited item detection," in *Proc. Int. J. Comput. Vis.*, 2023, pp. 3170–3192.

[4] 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. Int. Conf. Comput. Vis.*, 2021, pp. 10 923–10 932.

[5] L. D. Griffin, M. Caldwell, J. T. A. Andrews, and H. Bohler, "'unexpected item in the bagging area': Anomaly detection in X-ray security images," *IEEE Trans. Inf. Forensics Security*, pp. 1539–1553, 2019.

[6] 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 Trans. Inf. Forensics Security*, pp. 3866–3878, 2024.

[7] 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 Trans. Inf. Forensics Security*, pp. 3510–3524, 2024.

[8] J. Li, R. Socher, and S. C. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[9] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Adv. Neural Inform. Process. Syst.*, 2018, pp. 8536–8546.

[10] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 322–330.

[11] S. Chadwick and P. Newman, "Training object detectors with noisy data," in *Proc. IEEE Intelligent Vehicles Symp.*, 2019, pp. 1319–1325.

[12] J. Li, C. Xiong, R. Socher, and S. Hoi, "Towards noise-resistant object detection with noisy annotations," *arXiv preprint arXiv:2003.01285*, 2020.

[13] L. Yang, F. Meng, H. Li, Q. Wu, and Q. Cheng, "Learning with noisy class labels for instance segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 38–53.

[14] C. Liu, K. Wang, H. Lu, Z. Cao, and Z. Zhang, "Robust object detection with inaccurate bounding boxes," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 53–69.

[15] S. Wang, J. Gao, B. Li, and W. Hu, "Narrowing the gap: Improved detector training with noisy location annotations," *IEEE Trans. Image Process.*, pp. 6369–6380, 2022.

[16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *Commun. ACM*, pp. 107–115, Feb 2017.

[17] D. Arpit, S. Jastrzebski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 233–242.

[18] 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 *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[19] 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*, pp. 998–1009, 2022.

[20] 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 Recognition*, p. 108261, 2022.

[21] D. Velayudhan, T. Hassan, A. H. Ahmed, E. Damiani, and N. Werghi, "Baggage threat recognition using deep low-rank broad learning detector," in *IEEE Mediterranean Electrotechnical Conf.*, 2022, pp. 966–971.
[22] S. Akcay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging," *Pattern Recognition*, p. 108245, 2022.
[23] M. Rafiei, J. Raitoharju, and A. Iosifidis, "Computer vision on x-ray data in industrial production and security applications: A comprehensive survey," *IEEE Access*, pp. 2445–2477, 2023.
[24] D. Velayudhan, T. Hassan, E. Damiani, and N. Werghi, "Recent advances in baggage threat detection: A comprehensive and systematic survey," *ACM Computing Surveys*, pp. 1–38, 2022.
[25] M. Li, B. Ma, H. Wang, D. Chen, and T. Jia, "Gadet: A geometry-aware x-ray prohibited items detector," *IEEE Sensors Journal*, pp. 1665–1678, 2024.
[26] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.
[27] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
[28] S. Venkataramanan, E. Kijak, L. Amsaleg, and Y. Avrithis, "Alignmixup: Improving representations by interpolating aligned features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2022, pp. 19 174–19 183.
[29] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
[30] M. Ye, Z. Wu, C. Chen, and B. Du, "Channel augmentation for visible-infrared re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 2299–2315, 2024.
[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[32] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 3836–3847.
[33] H. Fang, B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye, "Data augmentation for object detection via controllable diffusion models," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2024, pp. 1257–1266.
[34] S. S. Gannamaneni, F. Klein, M. Mock, and M. Akila, "Exploiting clip self-consistency to automate image augmentation for safety critical scenarios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2024, pp. 3594–3604.
[35] T. W. Webb, N. Bhowmik, Y. F. A. Gaus, and T. P. Breckon, "Operationalizing convolutional neural network architectures for prohibited object detection in x-ray imagery," in *Proc. IEEE Int. Conf. Mach. Learn. Appl.*, 2021, pp. 610–615.
[36] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Adv. Neural Inform. Process. Syst.*, 2018, pp. 8792–8802.
[37] C. Tan, J. Xia, L. Wu, and S. Z. Li, "Co-learning: Learning from noisy labels with self-supervision," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 1405–1413.
[38] J. Goldberger and E. Ben-Reuven, "Training deep neural-networks using a noise adaptation layer," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–9.
[39] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1944–1952.
[40] X. Xia, T. Liu, B. Han, N. Wang, M. Gong, H. Liu, G. Niu, D. Tao, and M. Sugiyama, "Part-dependent label noise: Towards instance-dependent label noise," in *Adv. Neural Inform. Process. Syst.*, 2020, pp. 7597–7610.
[41] X. Li, T. Liu, B. Han, G. Niu, and M. Sugiyama, "Provably end-to-end label-noise learning without anchor points," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 6403–6413.
[42] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 1919–1925.
[43] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, "Normalized loss functions for deep learning with noisy labels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6543–6553.
[44] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2304–2313.
[45] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7164–7173.
[46] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 13 726–13 735.
[47] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," *arXiv preprint arXiv:1910.09217*, 2020.
[48] M. Ye and P. C. Yuen, "PurifyNet: A robust person re-identification model with noisy labels," *IEEE Trans. Inf. Forensics Security*, pp. 2655–2666, 2020.
[49] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. H. Hoi, "Collaborative refining for person re-identification with label noise," *IEEE Trans. Image Process.*, pp. 379–391, 2022.
[50] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, pp. 2352–2364, 2022.
[51] N. Bhowmik, Q. Wang, Y. F. A. Gaus, M. Szarek, and T. P. Breckon, "The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composited x-ray imagery," *arXiv preprint arXiv:1909.11508*, 2019.
[52] L. Duan, M. Wu, L. Mao, J. Yin, J. Xiong, and X. Li, "Rwsc-fusion: Region-wise style-controlled fusion network for the prohibited x-ray security image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 22 398–22 407.
[53] D. Mery and A. K. Katsaggelos, "A logarithmic x-ray imaging model for baggage inspection: Simulation and object detection," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017, pp. 57–65.
[54] T. W. Rogers, N. Jaccard, E. D. Protonotarios, J. Ollier, E. J. Morton, and L. D. Griffin, "Threat image projection (tip) into x-ray images of cargo containers for training humans and machines," in *IEEE Int. Carnahan Conf. Security Technol.*, 2016, pp. 1–7.
[55] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.
[56] A. Uddin, M. Monira, W. Shin, T. Chung, S.-H. Bae *et al.*, "Saliencymix: A saliency guided data augmentation strategy for better regularization," in *Proc. Int. Conf. Learn. Represent.*, 2021.
[57] D. Walawalkar, Z. Shen, Z. Liu, and M. Savvides, "Attentive cutmix: An enhanced data augmentation approach for deep learning based image classification," in *ICASSP*, 2020.
[58] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1137–1149, 2017.
[59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, pp. 84–90, 2017.
[60] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[61] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 6154–6162.
[62] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 9759–9768.
[63] 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.