Detecting Overlapped Objects in X-Ray Security Imagery by a Label-Aware Mechanism

Cairong Zhao, Liang Zhu, Shuguang Dou, Weihong Deng, Member, IEEE, and Liang Wang, Fellow, IEEE

Abstract—One of the key challenges to the X-ray security check is to detect the overlapped items in backpacks or suitcases in the X-ray images. Most existing methods improve the robustness of models to the object overlapping problem by enhancing the underlying visual information such as colors and edges. However, this strategy ignores the situations that the objects have similar visual clues as to the background, and objects overlapping each other. Since the two cases rarely appear in existing datasets, we contribute a novel dataset – Cutters and Liquid Containers X-ray Dataset (CLCXray) to complete the related research. Furthermore, we propose a novel Label-Aware Mechanism (LA) to tackle the object overlapping problem. Particularly, LA establishes the associations between feature channels and different labels and adjusts the features according to the assigned labels (or pseudo labels) to help improve the prediction results. Extensive experiments demonstrate that the LA is accurate and robust to detect overlapped objects, and also validate the effectiveness and the good generalization of the LA for arbitrary state-of-the-art (SOTA) methods. Furthermore, experimental results show that the network constructed by the LA is superior to the SOTA models on OPIXray and CLCXray, especially solving the challenges of the subset of the highly overlapped objects.

Index Terms—Object detection, X-ray dataset, overlap.

I. INTRODUCTION

In the past few decades, security check is generally recognized as an effective preventive measure for terrorist attacks and crimes worldwide. The X-ray-based package security check system has been widely used in subways, airports, customs, and other public places to check possible threat objects in packages for years. Although this system has achieved great success, it still suffers from low stability and low accuracy due to the reliance on manual human operator review. To tackle this problem, many researchers have studied the application of object detection algorithms to X-ray security images to assist the staff in identifying threat objects. Schmidt-Hackenberg et al. [1] proposed the use of two visual cortex-inspired features, SLF-HMAX and V1-like, combined with the bag of visual words method. Flitton et al. [2] explored 3D feature descriptors with application to object detection in 3D CT security imagery. Baştan [3] proposed two dense sampling methods as keypoint detectors for textureless objects and extended the SPIN color descriptor to utilize the material information for multi-view imagery. Kundegorski et al. [4] benchmarked various feature point descriptors in combination with the bag of visual words method. Jaccard et al. [5] first used convolutional neural networks (CNN) in X-ray images of cargo containers. Subsequently, Jaccard et al. [6] proposed a machine learning framework for X-ray cargo inspection. Petroziello and Jordanov [7] used image augment to remove noisy and fuzzy images, and evaluated the performance of CNN and Autoencoder. Akçay et al. [8]–[13] evaluated the performance of YoloV2 [14], R-CNN [15] and other deep learning methods on X-ray security images.

Existing works [6], [8] show the advantage of deep learning methods against traditional methods such as the bag of visual words method. However, deep learning methods require a large number of samples to achieve good generalization. From 2015 to 2019, there were only one public dataset [16], of which only 1552 X-ray baggage images are labeled with bounding boxes. In order to improve the generalization of the model with the limited available data, the researchers adopted techniques such as data augment and transfer learning. Jain et al. [9] employed an imaging model for the generation of new X-ray images. Cui and Oztan [12] used threat image projection (TIP) to generate training data. Bhowmik et al. [17] investigated the difference in detection performance achieved using real and synthetic X-ray training imagery. Gaus et al. [18] evaluated the transferability of deep learning networks. Wei and Liu [19] designed a transfer learning network based on SSD. However, the improvement of these technologies in terms of generalization is limited. Caldwell and Griffin [20] pointed out that data transfer from optical image data to X-ray security images is only beneficial when the data is scarce. Bhowmik et al. [17] showed the limitations of synthetic training data for prohibited object detection in X-ray security imagery. Cubuk et al. [21] pointed

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on March 21,2022 at 03:59:36 UTC from IEEE Xplore. Restrictions apply.
out that the magnitude of data augmentation is limited by the size of the model and the training set. More importantly, in order to study a specific problem, there must be a customized dataset. Miao et al. [22] published a dataset Security Inspection X-ray (SIXray), which contains a large number of pictures without threat objects, to study the imbalance of positive and negative samples. Wei et al. [23] released a dataset Occluded Prohibited Items X-ray (OPIXray) to study the overlap problem, in which images generally have complex backgrounds.

Due to the X-ray imaging principle, the images of the objects stacked in the baggage often overlap with each other. Unlike the occlusion problem in optical images, overlapped objects are still visible in the X-ray security images. However, due to the overlap of the images, the detection of the overlapped object is disturbed. According to the difference of overlapped objects, the overlap problem can be divided into three types, the overlap between the object and the irrelevant background, the overlap between the object and the similar background, and the overlap between multiple objects. Previous works mainly studied the overlap between threat objects and irrelevant backgrounds. Liu et al. [24] proposed a two-stage method, which firstly used color information to segment the target image from the input image, and then performed detection on the target image. Hassan et al. [25] also proposed a two-stage method, which firstly used contour information to segment the regions of interest (ROI) from the image, and then performed detection on the ROI. Instead of segmenting objects from backgrounds, Wei et al. [23] proposed to use the attention mechanism to make the network focus on the colors and contours of the objects in the image. Besides, Cao et al. [26] proposed to use partial appearance to identify threat objects, which required additional partial appearance labels.

However, the real scene is complicated. In some scenes, the color of the background and the object are similar, and the object does not have a clear and separable outline. Besides, there are overlaps between different objects. In this paper, we contribute a new dataset Cutters and Liquid Containers X-ray Dataset (CLCXray) to further study the overlap problem. Unlike OPIXray [23], CLCXray focuses more on the overlap between objects and similar backgrounds, as well as the overlap between multiple objects. In terms of categories, there are two types of threat objects in the CLCXray dataset, cutters and liquid containers, which are widespread but have been ignored in previous studies. Samples of CLCXray are shown in Fig. 1.

To solve the overlap problem, we propose a novel Label-aware Mechanism (LA), which uses the gradient to establish the relationship between the feature channels and the assigned label, and weights the feature channels according to the assigned label. Unlike previous strategies based on underlying visual information, which do not distinguish between different foregrounds, LA is based on high-level features. Extensive experiments demonstrate that LA is accurate and robust to detect overlapped objects, and also validate the effectiveness and the good generalization of LA for arbitrary networks on both OPIXray and CLCXray.

We summarize the contributions of this work as follows:

- We contribute a new dataset CLCXray for the overlap problem. Different from all existing datasets, CLCXray provides a large number of overlapped objects based on real scenes, which provides a good foundation for the research of the overlap problem. Besides, CLCXray takes hazardous liquids into consideration, expanding the scope of research on threat objects. Moreover, CLCXray provides high-precision annotations, which makes up for the current lack of high-quality bounding box (bbox) annotations.
- We propose a new Label-aware Mechanism (LA) for the overlap problem. Different from all existing methods, LA separates overlapped objects in high-level feature maps. By adaptively adjusting the corresponding features through the labels assigned to different anchors (sampling points), the LA can handle the overlap between objects and similar backgrounds, as well as the overlap between multiple objects.
- We evaluate several SOTA object detection methods on CLCXray and OPIXray, and evaluate the performance of LA on different methods. Extensive experiments demonstrate that LA is accurate and robust to detect overlapped objects, and also validate the effectiveness and the good generalization of LA for arbitrary networks on both OPIXray and CLCXray.

II. RELATED WORK

A. X-Ray Security Image Datasets

Mery et al. [29], [30] summarized the datasets appearing in the papers for object detection within X-ray security imagery. As shown in Table I, Durham Baggage Patch/Full Image Dataset [8], MV-Xray Dataset [27], and SASC Dataset [28] have not yet been publicly released. There are three published
datasets, GDXray, SIXray, and OPIXray. Among them, Grima X-ray Dataset (GDXray) [16] contains multi-view images and is usually used for classification tasks. GDXray contains 5 groups: castings, welds, baggage, nature, settings, where the group baggage is the dataset required for X-ray security image object detection. The group baggage contains 8,150 X-ray images arranged in 77 series. The X-ray images are taken from different containers such as backpacks, pen cases, wallets, etc. Series B0046, B0047 and B0048 contains 600 X-ray images that can be used for object detection of handguns, shuriken, razor blades. To study the multi-view problem, the experiments can be conducted on series B0049, B0050, and B0051 which includes X-ray images of individual handguns, shuriken, razor blades respectively taken from different points of view.

Security Inspection X-ray (SIXray) [22] is used to study the problem of class imbalance. SIXray contains a total of 1,059,231 X-ray images, of which 8,929 images are labeled. These images were collected from several subway stations with the original meta-data indicating the presence or absence of prohibited items. There are six common categories of prohibited items, namely, gun, knife, wrench, pliers, scissors, hammer. The distribution of these objects aligns with the real-world scenario, in which there are much fewer positive samples compared to negative samples. To study the impact brought by training data imbalance, Miao et al. constructed three subsets of this dataset, and named them SIXray10, SIXray100, and SIXray1000, respectively, with the number indicating the ratio of negative samples over positive samples.

Occluded Prohibited Items X-ray (OPIXray) is the first high-quality object detection dataset for security inspection. OPIXray contains a total of 8885 Xray images of 5 categories of cutters, namely, folding knife, straight knife, scissor, utility knife, multi-tool knife. The backgrounds of all samples are scanned by the security inspection machine and the prohibited items are synthesized into these backgrounds by the professional software. In order to study the impact brought by occlusion levels, Wei et al. divided the testing set into three subsets and named them Occlusion Level 1 (OL1), Occlusion Level 2 (OL2), and Occlusion Level 3 (OL3), where the number indicates occlusion level of prohibited items in images.

### B. Label Assignment
Label assignment is a step in the object detection pipeline to match labels and spatially distributed predictions. Currently, most label assignment strategies are based on prior knowledge. For example, Faster-RCNN [31], SSD [32], YOLOv3 [33], RetinaNet [34] are based on the anchor-based IoU prior, which assigns the label to each spatial location according to the Intersection over Union (IoU) of the preset anchor box and ground truth bbox. FCOS [35] is based on the center prior, which assigns the labels to each sampling point according to the distance from the sampling point to the center of the ground truth bbox. However, prior-based label assignment strategy ignores the actual content of the intersecting region, which may contain noisy background, nearby objects or a few meaningful parts of the target object to be detected. Since these actual contents are reflected in the prediction results, there have been many studies on the dynamic strategies of label assignment based on the prediction in recent years. FSAF [36] explored the dynamic strategy of assigning labels to different FPN layers. In order to determine the optimal FPN layer, FSAF designed a new module, which assigns labels by comparing the loss between the predictions and labels in different FPN layers. FreeAnchor [37] further explored the dynamic strategy of assigning labels to all anchors. FreeAnchor formulated detector training as a maximum likelihood estimation (MLE) procedure, which selects the most representative anchor from a “bag” of anchors for each object. PAA [38] proposed a novel anchor assignment strategy that adaptively separates anchors into positive and negative samples for a ground truth bbox according to the model’s learning status such that it is able to reason about the separation in a probabilistic manner.

### C. Solutions to Overlap Problem
The previous works mainly studied the overlap between objects and irrelevant backgrounds. Miao et al. [22] tried to use the information of different FPN layers to solve the overlap problem. From this perspective, they proposed to use foreground information between different FPN layers to eliminate background information. Liu et al. [24] tried to solve the

### TABLE I

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baggage Images</th>
<th>Annotated with bbox</th>
<th>Classes</th>
<th>Source</th>
<th>Task</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDXray [16]</td>
<td>8,150</td>
<td>1,552</td>
<td>3</td>
<td>Unknown</td>
<td>Classification and Object Detection</td>
<td>✓</td>
</tr>
<tr>
<td>Dnp2 [8]</td>
<td>11,627</td>
<td>11,627</td>
<td>6</td>
<td>Synthetic</td>
<td>Object Detection</td>
<td>✗</td>
</tr>
<tr>
<td>Liu et al. [24]</td>
<td>32,253</td>
<td>12,683</td>
<td>6</td>
<td>Real</td>
<td>Object Detection</td>
<td>✗</td>
</tr>
<tr>
<td>MV-Xray [27]</td>
<td>16,724</td>
<td>12,924</td>
<td>2</td>
<td>Unknown</td>
<td>Object Detection</td>
<td>✗</td>
</tr>
<tr>
<td>SASC [28]</td>
<td>3,250</td>
<td>3,250</td>
<td>2</td>
<td>Unknown</td>
<td>Object Detection</td>
<td>✗</td>
</tr>
<tr>
<td>SIXray10 [22]</td>
<td>88,372</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIXray100 [22]</td>
<td>882,802</td>
<td>8,929</td>
<td>6</td>
<td>Mostly real</td>
<td>Classification and Localization</td>
<td>✓</td>
</tr>
<tr>
<td>SIXray1000 [22]</td>
<td>1,054,911</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPIXray [23]</td>
<td>8,885</td>
<td>8,885</td>
<td>5</td>
<td>Synthetic</td>
<td>Object Detection</td>
<td>✓</td>
</tr>
<tr>
<td>Ours</td>
<td>9,565</td>
<td>9,565</td>
<td>12</td>
<td>Simulated and Real</td>
<td>Object Detection</td>
<td>✓</td>
</tr>
</tbody>
</table>
overlap problem from the perspective of image processing. Specifically, they segmented the foreground and background in the original image based on the color statistics of threat objects. Instead of using color information, Hassan et al. [25] chose to use contour information to separate the front and back backgrounds. They converted the input image into a contour image and used a novel structure tensor to separate the contours of the foreground and background. Wei et al. [23] considered both color and contour information, and introduced the attention mechanism to solve the overlap problem. In order to make the network pay attention to the color and contour of the image, they designed a DOAM module, which generates an attention map based on the color and contour. The generated attention map is used to enhance the input image.

III. THE CLCXRAY DATASET

The overlap problem is a challenging problem for X-ray security images. In order to study this problem, a suitable dataset is needed. Although one dataset OPIXray has been proposed for the overlap problem, it does not cover the overlap between multiple objects. In addition, the images of OPIXray are synthetic by TIP, which is different from the data of the real scene. For the above reasons, we propose a new dataset, CLCXray. Compared with all existing datasets, CLCXray has the most labeled images, labeled threat objects, threat categories, and accurate annotations of bbox. The following subsections introduce the CLCXray in details.

A. Motivation

At present, the research on the overlap problem is limited to the overlap between the object and the background, and there are few images with multiple objects overlapping each other in the existing datasets. In order to expand the research on the overlap problem, we propose the CLCXray dataset. Fig. 2 shows the different types of overlap in CLCXray. In addition, in the early datasets, highly lethal weapons are the main research objects, while toxic, corrosive, flammable, explosive liquids and various knives are neglected. Therefore, in CLCXray, we labeled cutters and liquid containers as threat objects, to promote research on cutters and liquid containers. Moreover, as shown in Fig. 3, the bbox annotations in SIXray and OPIXray are relatively rough, which is not conducive to the study of more precise positioning of object detection.

B. Pre-Processing

The CLCXray we provide has been pre-processed, which is approved by professionals. The raw data for each sample comprises two 16-bit grey-scale images, with values ranging from 0 to 65535. To transform the raw data to the three-channel image for training, testing, and visualization, we first divide the high-energy image and the low-energy image by 256. The resulting images constitutes the first channel and the second channel in the three-channel image. Then we use the ratio \( R \) of the high-energy image to the low-energy image to fill the third channel. Since \( \text{Sigmoid}(0) = 0.5 \), and the value of \( R \) is greater than 0, we use the following formula to project \( R \) to the interval between 0 and 255:

\[
\text{Channel}_3 = 510 \cdot \text{Sigmoid}(R) - 255. \tag{1}
\]

After transformation, we obtain the samples shown in Fig. 2.

C. Data Properties

The CLCXray dataset contains 9,565 X-ray images, in which 4,543 X-ray images (real data) are obtained from the real subway scene and 5,022 X-ray images (simulated data) are scanned from manually designed baggages. All images were acquired using the same type of X-ray scanner (TECHIK, model TH-XS6550). All labels were separately marked by 8 junior staff (less than 5 years working experience and students) and reviewed by 2 Senior staff (more than 5 years working experience). There are 12 categories in the CLCXray dataset, including 5 types of cutters and 7 types of liquid containers. Five kinds of cutters include \textit{blade}, \textit{dagger}, \textit{knife}, \textit{scissors}, \textit{swiss army knife}. Seven kinds of liquid containers include \textit{cans}, \textit{carton drinks}, \textit{glass bottle}, \textit{plastic bottle}, \textit{vacuum cup}, \textit{spray cans}, \textit{tin}. The distribution of each category is shown in Table II. The CLCXray dataset contains more than 20,000 potentially dangerous items and each X-ray
image contains more than two potentially dangerous items on average. The resolutions of images are between 373 × 200 and 732 × 1280. The labels were made into COCO format. With reference to the general division, CLCXray is divided into the training set, validation set, and testing set at a ratio of 8:1:1. We first construct the test set with a ratio of 1:9 between simulated data and real data through random sampling. Then we use the remaining samples to form the training set and the test set at a ratio of 8:1. The test set contains a much higher proportion (90%) of real samples than the proportion (43%) of real samples in the training set and the validation set.

Compared with GDXray, SIXray and OPIXray, CLCXray has the following unique properties: First, there are more overlaps between multiple objects in CLCXray, as the result of more labeled objects per image on average. As shown in Fig. 5, nearly 60% of X-ray images in the CLCXray dataset contain at least two or more foregrounds. In SIXray and OPIXray, only a small number of X-ray images contain more than one object. Fig. 2 shows the different overlaps in CLCXray. Second, the category in CLCXray contains liquid containers, which has not been seen in previous studies. Liquid containers may contain toxic, corrosive, flammable, and explosive liquids, which are dangerous but easily overlooked. Third, CLCXray has more accurate bbox annotations. Fig. 4 shows the line graphs obtained by training and testing the baseline model, ATSS [39], on different datasets. The steep decline that occurs on OPIXray and SIXray shows the difficulty for the model to learn accurate positioning from the bbox annotations. Furthermore, we visualize the annotations of different datasets, as shown in Fig. 3. Compared with SIXray and OPIXray, CLCXray has annotations that visually fit the object edge more closely.

D. Availability
The images and the corresponding annotation results can only be used for ACADEMIC PURPOSES. NO COMMERCIAL USE is allowed. Copyright ©Visual and Intelligent Learning lab, Tongji University. All rights reserved. Download the dataset from here:
https://github.com/GreysonPhoenix/CLCXray

IV. OUR APPROACH
A. Overall Framework
In this paper, we combine LA and ATSS to build our network. ATSS has the following improvements based on
Fig. 6. Overall framework. It has the same structure as FCOS [35] and ATSS [39] but with an additional branch, the Label-aware branch. The input X-ray image is processed with ResNet-50 as the Backbone and the five-layer FPN as the Neck to generate features. The features generate prediction results in the Head part.

RetinaNet. Structurally, ATSS uses five-layer FPN and predicts the regression quality score, center-ness, on the regression branch. In the choice of regression loss, ATSS uses GIoU loss. In the label assignment strategy, ATSS changes the fixed threshold for assigning positive or negative labels to the dynamic threshold based on the statistics of IoUs of anchors and ground truth bbox. As shown in Fig. 6, our overall network structure is similar to ATSS, except that a new branch Label-aware is added to the Head part. The Label-aware branch forms a reverse network from the prediction results and the label assignment to the feature map. Thus, there is a loop in the Head part. In our setting, this loop happens only once. Specifically, the network makes two predictions, and the regression branch repeats twice in the second prediction. When calculating the loss function, the first prediction is not considered.

B. Label-Aware Mechanism

Fig. 7 shows the overlap among the vacuum cup (A), the plastic bottle (B), and the unlabeled keyboard (background). P is a sampling point located in the green grid. Since P is in the overlapping area, P extracts the low-level visual feature from both A, B, and the background. When P is responsible for predicting A, the information from B and the background is redundant. And when P is responsible for predicting B, the information from A and the background is redundant. Redundant information causes the high-level features of P to be near the decision boundary on the feature manifold. LA distinguishes redundant information according to the label assigned to P, and adjusts the high-level feature to keep P away from the decision boundary on the feature manifold. This mechanism can be expressed as searching for an adjustment weight with the lowest task loss $L$:

$$\hat{\omega} = \arg \min_{\omega} \sum_{i}^{N} \mathcal{L}(y_{i}, \hat{y}_{\sigma(i)}),$$

$$y = \mathcal{F}(wx; \theta), \quad (2)$$

where $\hat{y}$ is the ground truth set of objects, and $y = \{y_{i}\}_{i=1}^{N}$ is the prediction set of N sampling points. $\sigma$ is the mapping from the sampling point subscript to the ground truth subscript, which is determined by the label assignment. $\theta$ is a set of parameters of the head part of the network. We notice that similar objects usually cause the wrong predictions of networks, which indicates that similar objects may have common features. To decrease the wrong prediction, the above formula is modified as follows:

$$\hat{\omega} = \arg \min_{\omega} \sum_{i}^{N} (\mathcal{L}(y_{i}, \hat{y}_{\sigma(i)}) - \mathcal{L}(y_{i}, \hat{y}_{\hat{j}})), \quad \hat{j} = \arg \min_{j} \mathcal{L}(y_{i}, \hat{y}_{j}, j \neq \sigma(i)), \quad y = \mathcal{F}(wx; \theta). \quad (3)$$

It can be seen from the formula that in order to obtain $\hat{\omega}$, the label information of each sampling point is needed to obtain through label assignment. Different from the methods of dynamic label assignment, which usually redistributes labels on the basis of static label assignment, LA adjusts features on the basis of static label assignment. Label information can
be divided into classes and regressions. In the next sections, we introduce the implementation of LA with these two types of label information.

C. LA Using Class Labels

In this section, we introduce using the class labels to implement LA. Our original idea is to allow the network to learn the weight \( \hat{w} \) directly through the category of the assigned label and the multi-class confidences of predictions. Thus, the early version of LA first calculates the cross-entropy of the confidence of the predicted category and the class label and then uses a set of \( 1 \times 1 \) convolutions to learn weights based on the calculated results. The network structure of this method is shown in the (A) of Fig. 8.

Although the early version of LA can improve the performance of the model, it is unstable and lacks interpretability. To allow the generated weight to reasonably reflect the correspondence between features and labels, we take advantage of the gradient. For the predicted confidence of a specific category, the corresponding gradient of the feature map reflects the importance of different positions on the feature map to improve the confidence. Thus this gradient is consistent with our goal. Considering that the form of residuals is conducive to identity mapping, the formula for generating the new feature map has the following form:

\[
x_{\text{new}} = x + w_1 \cdot x.
\]

However, there may be intersections between different feature channels required to predict different categories. Enhancing common channels will not only increase the confidence in the correct category but also the confidence in the wrong category. To decrease wrong predictions, we generate a second weight based on the category with the highest confidence other than the correct category. By subtracting the second weight from the first weight, we obtain the current version, which is LAccls. The new feature map generated by LAccls has the following form:

\[
x_{\text{new}} = x + w_1 \cdot x - w_2 \cdot x,
\]

\[
w_1 = \text{Sigmoid}(\nabla_x (y \cdot \hat{y})),
\]

\[
w_2 = \text{Sigmoid}(\nabla_x (y \cdot y^*)),
\]

where \( \hat{y} \) represents the correct category label using one-hot encoding, and \( y^* \) represents the misleading category label using one-hot encoding. The misleading category refers to the category with the highest predicted probability other than the correct category. \( y \) represents the predicted multi-category confidence, which has the same shape as \( \hat{y} \) and \( y^* \).

D. LA Without Labels

LA mainly works in the training phase, using the assigned labels to adjust features. During the testing phase, labels are not available for LA. However, recent studies \[42\], \[49\] show the importance of the consistency between the training phase and the testing phase. To tackle this problem, we use pseudo-labels generated by the predicted category to replace the role of labels in the testing phase. When the network predicts correctly, the pseudo labels are equivalent to the ground truth labels. When the network predicts wrongly, LA does not change the decision-making. And compared to the original network, the network trained with LA learns feature extraction and feature-to-label mapping more effectively. So overall, the network with LA produces better predictions than the original network in the testing phase.

E. LA Using Regression Labels

Compared with using class labels to construct the LA mechanism, it is more difficult to use regression labels. Because the pseudo-labels used in the testing phase cannot be generated in the current general regression form. To tackle this problem, we refer to the strategies of GFL \[49\] and Scope head \[50\] to discretize continuous regression representation. In our method, we first change the original regression representation into the regression representation of FCOS, which regresses the four distances from the center point to the four borders of the bbox. Then we turn predicting distances in the four directions into predicting probabilities that the distance values fall in different numerical ranges. By discretizing the regressor in this way, we can take the same strategy as LAccls to construct LA using regression labels. Besides, this method can still obtain the continuous predicted distance in the four directions by calculating the expected value. In our experiment, we set the maximum distance to 16 times the stride, and divide the maximum distance into 16 intervals evenly. The LA using regression labels is named LARreg. The new feature map is generated by the following form:

\[
x_{\text{new}} = 0.5 \cdot x + \text{Sigmoid}(\nabla_x (y \cdot \hat{y})) \cdot x,
\]

where \( y \) represents the confidence of the predicted regression target in different intervals. \( \hat{y} \) represents the interval where the ground truth regression target is located.

V. EXPERIMENTS

In this section, we first set a baseline much stronger than the SOTA method on OPIXray and verify the effectiveness of LA by applying LA to the baseline. Then we compare the performance of multiple SOTA object detection methods on CLCXray and compare LA with other methods to further verify the effectiveness of our method. We also apply LA on other networks to verify the generality of LA. Since
TABLE III
DETECTION PERFORMANCE IN TERMS OF mAP (%) ON CLCXray. THE RESULTS ARE SHOWN AS MEANS ± STDS OF THREE TRAINING RUNS

<table>
<thead>
<tr>
<th>Models</th>
<th>Backbone</th>
<th>Memo(GB)</th>
<th>mAP</th>
<th>mAP50</th>
<th>mAP75</th>
<th>mAPs</th>
<th>mAPm</th>
<th>mAPl</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>two – stages:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Empirical Att[40]</td>
<td>R-50-FPN</td>
<td>8.0</td>
<td>55.1±0.2</td>
<td>70.2±0.4</td>
<td>65.7±0.3</td>
<td>2.8±1.2</td>
<td>27.3±0.6</td>
<td>60.9±0.1</td>
<td>ICCV2019</td>
</tr>
<tr>
<td>LIBRA RCNN[41]</td>
<td>R-50-FPN</td>
<td>4.6</td>
<td>55.7±0.6</td>
<td>70.4±0.5</td>
<td>66.5±1.3</td>
<td>13.4±4.2</td>
<td>28.8±0.7</td>
<td>61.2±0.5</td>
<td>CVPR2019</td>
</tr>
<tr>
<td>Cascade RCNN[42]</td>
<td>R-50-FPN</td>
<td>4.4</td>
<td>58.4±0.1</td>
<td>71.4±0.1</td>
<td>68.4±0.2</td>
<td>6.8±7.1</td>
<td>31.5±0.3</td>
<td>63.8±0.3</td>
<td>TPAMI2019</td>
</tr>
<tr>
<td>Dynamic RCNN[43]</td>
<td>R-50-FPN</td>
<td>3.8</td>
<td>56.7±0.5</td>
<td>70.9±5</td>
<td>66.9±0.8</td>
<td>2.7±1.7</td>
<td>28.5±1</td>
<td>62.8±0.3</td>
<td>ECCV2020</td>
</tr>
<tr>
<td>Double head[44]</td>
<td>R-50-FPN</td>
<td>6.8</td>
<td>57.6±0.5</td>
<td>71.6±0.1</td>
<td>67.9±1.1</td>
<td>12.5±1</td>
<td>28.9±0.2</td>
<td>63.8±0.6</td>
<td>CVPR2020</td>
</tr>
<tr>
<td>one – stages:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSD[32]</td>
<td>VGG16</td>
<td>10.2</td>
<td>51.1±0.3</td>
<td>66.4±0.3</td>
<td>59.8±0.4</td>
<td>0.7±1</td>
<td>22.0±5</td>
<td>57.5±0.4</td>
<td>ECCV2016</td>
</tr>
<tr>
<td>YOLOv3[33]</td>
<td>Darknet53</td>
<td>3.8</td>
<td>53±1</td>
<td>67.2±0.3</td>
<td>63.0±0.2</td>
<td>0±0</td>
<td>25.9±2.1</td>
<td>58.6±0.6</td>
<td>-</td>
</tr>
<tr>
<td>GHM[45]</td>
<td>R-50-FPN</td>
<td>4.0</td>
<td>53.2±0.1</td>
<td>71.6±0.1</td>
<td>68.0±0.2</td>
<td>13.4±8.5</td>
<td>28.6±0.6</td>
<td>63.9±0.2</td>
<td>AAAI2019</td>
</tr>
<tr>
<td>FCOS[35]</td>
<td>R-50-FPN</td>
<td>6.5</td>
<td>56.2±0</td>
<td>70.7±2.0</td>
<td>66.6±0.4</td>
<td><strong>36.3±6.0</strong></td>
<td>27.3±2.3</td>
<td>62.1±0.3</td>
<td>ICCV2019</td>
</tr>
<tr>
<td>GA-RCNN[46]</td>
<td>R-50-FPN</td>
<td>4.0</td>
<td>58.0±0.1</td>
<td>71.6±0.1</td>
<td>68.0±0.2</td>
<td>13.4±8.5</td>
<td>28.6±0.6</td>
<td>63.9±0.2</td>
<td>CVPR2019</td>
</tr>
<tr>
<td>RepPoints[47]</td>
<td>R-50-FPN</td>
<td>3.9</td>
<td>55.8±0.4</td>
<td>69.8±0.2</td>
<td>66.8±0.6</td>
<td>18.9±1.8</td>
<td>26.3±2.6</td>
<td>60.9±0.2</td>
<td>ICCV2019</td>
</tr>
<tr>
<td>Free anchor[37]</td>
<td>R-50-FPN</td>
<td>4.9</td>
<td>57.2±0.1</td>
<td>70.8±0.6</td>
<td>67.1±0.4</td>
<td>21±7.4</td>
<td>27.3±2.3</td>
<td>63.0±0.4</td>
<td>-</td>
</tr>
<tr>
<td>FSAF[36]</td>
<td>R-50-FPN</td>
<td>3.15</td>
<td>55.8±0.6</td>
<td>70.0±0.8</td>
<td>66.5±0.8</td>
<td>15.1±7.1</td>
<td>24.4±2.4</td>
<td>61.7±0.7</td>
<td>CVPR2019</td>
</tr>
<tr>
<td>NAS-FCOS[48]</td>
<td>R-50-FPN</td>
<td>-</td>
<td>57.3±0.1</td>
<td><strong>72.3±0.4</strong></td>
<td>67.7±0.5</td>
<td>30.3±2.8</td>
<td>28.8±0.8</td>
<td>63.3±0.2</td>
<td>CVPR2020</td>
</tr>
<tr>
<td>FCOS+DOAM[23]</td>
<td>R-50-FPN</td>
<td>-</td>
<td>54.3±1</td>
<td>68.5±1.3</td>
<td>63.5±1.5</td>
<td>31.2±1.5</td>
<td>27.3±0.3</td>
<td>59.9±1.1</td>
<td>ACM2020</td>
</tr>
<tr>
<td>PAA[38]</td>
<td>R-50-FPN</td>
<td>3.7</td>
<td>58.3±0.1</td>
<td>71.6±0.4</td>
<td><strong>68.5±0.3</strong></td>
<td>19.8±3.4</td>
<td>29.4±1.2</td>
<td>63.9±0.1</td>
<td>ECCV2020</td>
</tr>
<tr>
<td>ATSS(baseline)[39]</td>
<td>R-50-FPN</td>
<td>3.7</td>
<td>58.0±0.2</td>
<td>70.8±0.2</td>
<td>67.2±0.2</td>
<td>17.9±4.1</td>
<td>31.0±0.4</td>
<td>63.5±0</td>
<td>CVPR2020</td>
</tr>
<tr>
<td>ATSS+LAreg(Ours)</td>
<td>R-50-FPN</td>
<td>3.7</td>
<td>58.5±0.1</td>
<td>70.9±0.1</td>
<td>67.7±0.5</td>
<td>12.6±4.7</td>
<td>30.5±0.8</td>
<td>63.8±0.2</td>
<td>-</td>
</tr>
<tr>
<td>ATSS+LAcls(Ours)</td>
<td>R-50-FPN</td>
<td>3.7</td>
<td><strong>59.3±0.2</strong></td>
<td>71.8±0.2</td>
<td>68.2±0.1</td>
<td>23.0±1.5</td>
<td><strong>32.4±0.5</strong></td>
<td><strong>64.5±0.1</strong></td>
<td>-</td>
</tr>
</tbody>
</table>

CLCXray contains a number of images without overlapped objects, we build a much more challenging subset for further evaluation.

A. Experiment Details

We conduct experiments on two datasets, OPIXray and CLCXray. For OPIXray, we adopt the evaluation metric in the article of OPIXray [23], which is the mean average precision computed at the Intersection over Union (IoU) threshold of 0.5. For CLCXray, we adopt COCO evaluation metrics [51]. As shown in Table III, \( mAP \) represents the mean average precision computed across 10 IoU thresholds of 0.5:0.05:0.95, which is the primary challenge metric. \( mAP_{50} \) represents the mean average precision computed at a single IoU threshold of 0.5. \( mAP_{75} \) represents the mean average precision computed at a single IoU threshold of 0.75. \( mAP_{s} \) represents the \( mAP \) for small objects (\( area < 32^2 \)). Due to the small number of small targets in the CLCXray test set, the standard deviation of the results in the mAPs column is large. \( mAP_{m} \) represents the \( mAP \) for medium objects (\( 32^2 < area < 96^2 \)). \( mAP_{l} \) represents the \( mAP \) for large objects (\( 96^2 < area \)). We use two Nvidia RTX 3090 GPUs to conduct experiments and use pre-trained weights in all models. The epoch of all models with backbone R-50-FPN is uniformly set to 12. The batch size, learning rate, momentum, weight decay and other parameters refer to the configuration of each method in the paper. The configuration of our network is consistent with the baseline, in which the batch size per GPU is 4, the type of optimizer is SGD, the epoch is 12, the learning rate is 0.01, the momentum is 0.9, and the weight decay is 0.0001. Besides, we run LAcls and LAreg with the same regression and classification branches.

TABLE IV
DETECTION PERFORMANCE IN TERMS OF mAP (%) ON OPIXray

<table>
<thead>
<tr>
<th>Models</th>
<th>FO</th>
<th>ST</th>
<th>SC</th>
<th>UT</th>
<th>MU</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD</td>
<td>70.89</td>
<td>76.91</td>
<td>35.02</td>
<td>93.41</td>
<td>65.87</td>
</tr>
<tr>
<td>SSD+DOAM</td>
<td>74.01</td>
<td>81.37</td>
<td>41.50</td>
<td>95.12</td>
<td>68.21</td>
</tr>
<tr>
<td>YOLO</td>
<td>78.21</td>
<td>92.53</td>
<td>36.02</td>
<td>97.34</td>
<td>70.81</td>
</tr>
<tr>
<td>YOLO+DOAM</td>
<td>79.25</td>
<td>90.23</td>
<td>41.73</td>
<td>96.96</td>
<td>72.12</td>
</tr>
<tr>
<td>FCOS</td>
<td>82.02</td>
<td>86.41</td>
<td>68.47</td>
<td>90.22</td>
<td>78.39</td>
</tr>
<tr>
<td>FCOS+DOAM</td>
<td>82.41</td>
<td>86.71</td>
<td>68.58</td>
<td>90.23</td>
<td>78.84</td>
</tr>
<tr>
<td>ATSS</td>
<td>86.59</td>
<td>92.31</td>
<td>72.04</td>
<td>96.58</td>
<td>80.38</td>
</tr>
<tr>
<td>ATSS+DOAM</td>
<td>85.58</td>
<td>90.66</td>
<td>66.78</td>
<td>96.17</td>
<td>81.83</td>
</tr>
<tr>
<td>ATSS+LAreg(Ours)</td>
<td>87.39</td>
<td><strong>92.78</strong></td>
<td>71.17</td>
<td>96.61</td>
<td>83.45</td>
</tr>
<tr>
<td>ATSS+LAcls(Ours)</td>
<td><strong>88.26</strong></td>
<td>90.04</td>
<td><strong>74.99</strong></td>
<td><strong>97.60</strong></td>
<td><strong>85.70</strong></td>
</tr>
</tbody>
</table>
TABLE V  
COMPARING WITH SOTA METHODS. MAPS (%) ARE REPORTED ON CLCXRAY

<table>
<thead>
<tr>
<th>Models</th>
<th>mAP</th>
<th>mAP_{50}</th>
<th>mAP_{75}</th>
<th>mAP_{s}</th>
<th>mAP_{m}</th>
<th>mAP_{l}</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATSS (baseline)</td>
<td>58.1</td>
<td>70.9</td>
<td>66.9</td>
<td>13.5</td>
<td>31.5</td>
<td>63.3</td>
</tr>
<tr>
<td>ATSS + Sel[52]</td>
<td>58.0</td>
<td>71.3</td>
<td>68.0</td>
<td>36.5</td>
<td>30.3</td>
<td>63.4</td>
</tr>
<tr>
<td>ATSS + CBAM [53]</td>
<td>58.1</td>
<td>70.9</td>
<td>67.8</td>
<td>36.5</td>
<td>32.0</td>
<td>62.7</td>
</tr>
<tr>
<td>ATSS + RCCA [54]</td>
<td>58.3</td>
<td>70.2</td>
<td>67.4</td>
<td>2.4</td>
<td>30.5</td>
<td>63.1</td>
</tr>
<tr>
<td>ATSS + FCANet [55]</td>
<td>58.3</td>
<td>71.5</td>
<td>67.7</td>
<td>14.2</td>
<td>31.1</td>
<td>63.6</td>
</tr>
<tr>
<td>ATSS + SPA [36]</td>
<td>58.4</td>
<td>71.2</td>
<td>68.1</td>
<td>20.2</td>
<td>31.4</td>
<td>63.8</td>
</tr>
<tr>
<td>ATSS + LArg (ours)</td>
<td>58.6</td>
<td>71.1</td>
<td>68.2</td>
<td>13.5</td>
<td>31.2</td>
<td>64.0</td>
</tr>
<tr>
<td>ATSS + LAcLs (ours)</td>
<td>59.5</td>
<td>71.5</td>
<td>68.0</td>
<td>9</td>
<td>33</td>
<td>64.6</td>
</tr>
</tbody>
</table>

B. Comparing With SOTA Methods

In the OPIXray dataset, we use ATSS as the baseline and test its performance. In addition, we test the performance of using DOAM [19], LArg, and LAcLs on ATSS, where DOAM is the method proposed in the OPIXray article. The configuration of all tested methods is consistent with that on the CLCXray dataset. We use the metric in the article [19], which is the mean average precision computed at a single IoU threshold of 0.5. As shown in the mAP_{50} column of Table IV, ATSS is 4.17% higher than the SOTA model FCOS + DOAM. On such a strong baseline, LAcLs improves mAP_{50} by 1.67%. In the CLCXray dataset, we test the SOTA model of object detection, FCOS + DOAM, and other SOTA models of general object detection methods in recent years. Similarly, we use ATSS as the baseline and test its performance. We also test the improvement of ATSS by LA. As shown in the mAP column of Table III, the baseline, ATSS, is 3.7% higher than the SOTA model of the overlap problem, FCOS + DOAM. On such a strong baseline, the proposed method LAcLs increases mAP by 1.3%, and LArg increases mAP by 0.6%. Compared with the earlier methods, SOTA methods have small improvements on CLCXray, indicating that CLCXray is challenging. At the same time, LA’s improvement to the baseline is significant. Moreover, among all models, ATSS + LAcLs achieves the highest scores on mAP.

C. Comparing With Self-Attention Methods

The self-attention methods [52], [53] and LA both adjust the feature map based on the generated weight. The difference is that self-attention methods generate weights from features themselves while LA generates weights from assigned labels. To explore the difference between these two strategies, we substitute LAcLs with self-attention methods in the network and evaluate these methods on CLCXray. As shown in Table V, the performance of LAcLs is 1.4% higher than CBAM [53], 1.5% higher than SE [52], and 1.2% higher than RCCA [54]. In general, self-attention methods achieve similar results. Compared with learning the weight by the network itself, our strategy of generating the weight based on labels and gradients is more effective on the CLCXray.

D. Generalization Ability of LA

In order to test the generality of our method, we choose a static label assignment model FCOS and a dynamic label assignment model PAA to apply our method. Both of these are state-of-the-art models in the past two years. Experiments are conducted on CLCXray. As shown in Table VI, our method LAcLs improves FCOS from 56.3 to 57.4 and improves PAA from 58.5 to 59.3. Compared with the static label assignment model, LA has relatively little improvement to the dynamic label assignment model PAA. As mentioned before, the essence of dynamic label assignment is to select the optimal label assignment according to the state of the extracted features, while LA adjusts the extracted features according to the label assignment. They are the two sides of the coin. Therefore, the performance improvements they bring are mutually diluted. In addition, PAA is based on ATSS. Compared with the improvement of ATSS by PAA, LA has a greater improvement to ATSS, which shows that LA is better for the overlap problem.

E. Ablation Studies

We add coefficients to the three terms of the Eq. 5 to study the different role of components in LAcLs. The generalized Eq. 5 is as follows:

\[ x_{new} = a \cdot x + b \cdot w_1 \cdot x - c \cdot w_2 \cdot x, \]

\[ w_1 = \text{Sigmoid}(\nabla_x (y \cdot \hat{y})), \]

\[ w_2 = \text{Sigmoid}(\nabla_y (x \cdot y^*)), \]

In order to make \( x_{new} = x \) when LA lose effectiveness (values of \( w_1 \) and \( w_2 \) are close to 0.5), we make the following constraints:

\[ a + 0.5 \cdot b - 0.5 \cdot c = 1. \]

By changing the values of the coefficients, we obtain Table VII, where the first row of the data corresponds to the baseline. The experiment shows that when the three coefficients are not all 0, the performance is the highest.

F. Analysis

Since CLCXray still contains a number of images without overlap. We select 300 images with highly-overlapped objects from CLCXray to build a challenging subset. Comparisons are made among our method LA, the baseline ATSS, and the

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on March 21,2022 at 03:59:36 UTC from IEEE Xplore. Restrictions apply.
SOTA object detection model FCOS + DOAM. To build the challenging subset, we select and reserve the images with the overlap between multiple objects, or the overlap between objects and similar backgrounds in the test set. Comparisons are shown in Table VIII. LAcls increases ATSS by 3.9%, LAregr increases ATSS by 2.4%, and ATSS + LAcls is 3.8% higher than SOTA model FCOS + DOAM. Fig. 9 shows the visual test results of the baseline model ATSS and the model ATSS + LAcls. As shown in the set of images on the left, there are many liquid containers that have not been successfully detected. These liquid containers usually either overlap with other objects, or they are very similar to the background. When there is an overlap between the object and a similar background, the sampling point extracts too much background information, which leads to the prediction of the background. When there is an overlap between multiple objects, the sampling points extract too many features of other objects, leading to the prediction of low-quality bboxes of other objects, and then these bboxes are removed by NMS. As shown in the set of images on the right, several objects with overlap problems are correctly detected. At the same time, since LA adjusts the features, the detected objects generally have higher confidence in predictions. Experimental data and visualization results show that by optimizing the feature extraction of sampling points in the overlapping area, LA improves the robustness and accuracy of the model to overlap problems.

VI. CONCLUSION

The overlapping problem is significant and challenging for threat detection in X-ray images. In this paper, we first publicly release a high-quality dataset CLCXray as the research foundation for the overlapping problem. Then we propose a new method LA to address the overlapping problem. Different from previous methods, LA adjusts high-level features rather than low-level visual features, which manages to separate overlapped objects in the high-dimensional space. The visualizations show that LA generally improves the detection confidence of overlapped objects and avoids a large number of missed detections due to overlapping problems.
The experiments show that LA generally improves the detection performance of the models, and the combination of ATSS and LA achieves the highest mAP. For further study, we sample some highly overlapped samples to form a more challenging subset. Experiments on the subset show that LA provides a larger performance boost for the models, further demonstrating the effectiveness of LA for detecting overlapped objects.

REFERENCES


Cairong Zhao received the B.Sc. degree from Jilin University in 2003, the M.Sc. degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2006, and the Ph.D. degree from the Nanjing University of Science and Technology in 2011. He is currently a Professor with Tongji University. He is the author of more than 30 scientific papers in pattern recognition, computer vision, and related areas. His research interests include computer vision, pattern recognition, and visual surveillance.

Liang Wang (Fellow, IEEE) received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CAS), China, in 2004, From 2004 to 2010, he was working as a Research Assistant with Imperial College London, U.K., and Monash University, Australia; a Research Fellow with The University of Melbourne, Australia; and a Lecturer with the University of Bath, U.K. He is currently a Full Professor with the Hundred Talents Program, National Laboratory of Pattern Recognition (NLPR), Institute of Automation, CAS, and the Deputy Director of NLPR. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published at highly ranked international journals, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON IMAGE PROCESSING and leading international conferences, such as CVPR, ICCV, and ICDM. He is currently an IAPR Fellow and a member of BMVA. He has obtained several honors and awards, such as the Special Prize of the Presidential Scholarship of the Chinese Academy of Sciences.

Shuguang Dou is currently pursuing the Ph.D. degree with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision, deep learning, X-ray, and person re-identification.

Liang Zhu is currently a Graduate Student in computer science with Tongji University. His research fields focus on deep learning and computer vision, specifically on object detection in X-ray images.

WeiHong Deng (Member, IEEE) received the B.E. degree in information engineering and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004 and 2009, respectively. From 2007 to 2008, he was a Post-Graduate Exchange Student with the School of Information Technologies, The University of Sydney, Sydney, NSW, Australia, under the support of the China Scholarship Council. He is currently a Professor with the School of Information and Telecommunications Engineering, BUPT. His research interests include statistical pattern recognition and computer vision, with a particular emphasis in face recognition. He has authored over 30 technical papers in international journals and conferences, including a technical comment on face recognition in Science magazine. His Ph.D. dissertation was awarded the Outstanding Doctoral Dissertation by the Beijing Municipal Commission of Education in 2011. He was elected the New Century Excellent Talents Program by the Ministry of Education of China in 2013.