This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

AO-DETR: Anti-Overlapping DETR for X-Ray Prohibited Items Detection

Mingyuan Li, Tong Jia, Hao Wang, Bowen Ma, Hui Lu, Shuyang Lin, Da Cai, and Dongyue Chen

***Abstract*—Prohibited item detection in X-ray images is one of the most essential and highly effective methods widely employed in various security inspection scenarios. Considering the significant overlapping phenomenon in X-ray prohibited item images, we propose an anti-overlapping detection transformer (AO-DETR) based on one of the state-of-the-art (SOTA) general object detectors, DETR with improved denoising anchor boxes (DINO). Specifically, to address the feature coupling issue caused by overlapping phenomena, we introduce the category-specific one-to-one assignment (CSA) strategy to constrain category-specific object queries in predicting prohibited items of fixed categories, which can enhance their ability to extract features specific to prohibited items of a particular category from the overlapping foreground–background features. To address the edge blurring problem caused by overlapping phenomena, we propose the look forward densely (LFD) scheme, which improves the localization accuracy of reference boxes in mid-to-high-level decoder layers and enhances the ability to locate blurry edges of the final layer. Similar to DINO, our AO-DETR provides two different versions with distinct backbones, tailored to meet diverse application requirements. Extensive experiments on the PIXray, OPIXray, and HIXray datasets demonstrate that the proposed method surpasses the SOTA object detectors, indicating its potential applications in the field of prohibited item detection. The source code will be available at: https://github.com/Limingyuan001/AO-DETR.**

***Index Terms*—Iterative refinement boxes, label assignment, object detection, X-ray inspection.**

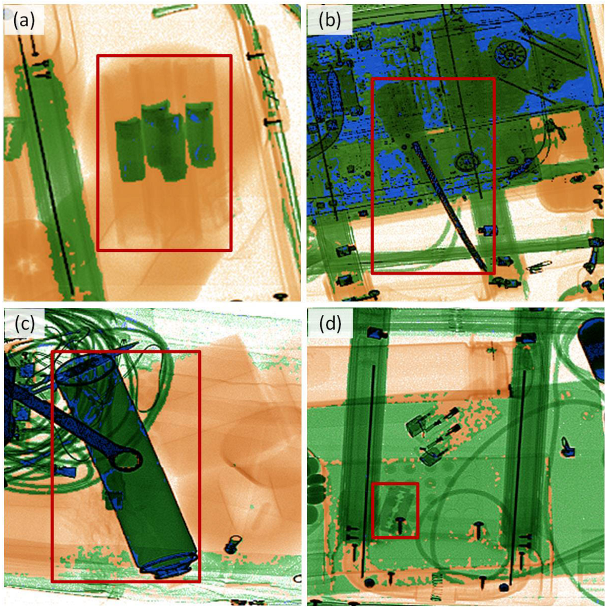


Fig. 1. Localized X-ray images with prohibited items. The phenomenon of overlap in images, to varying extents, leads to the overlapping of foreground and background as well as the blurring of object boundaries. (a)–(d) Index of X-ray image example.

I. Introduction

SECURITY inspection is one of the most vital and crucial measures to uncovering the potential risks in public spaces, such as airports, train stations, subway stations, and sensitive departments. Currently, a predominant approach to contraband detection involves the acquisition of X-ray images of luggage through a security scanning machine, followed by a meticulous manual inspection conducted by security staff who have undergone specialized training. With the advancement of computer vision technology, authors [1], [2], [3], [4], [5], [6], [7], [8], and [9] have attempted to apply models from the general field of image classification, object detection, and semantic segmentation to the realm of X-ray prohibited item detection, aiming to assist security staff in auxiliary inspections. However, as shown in Fig. 1, X-ray images exhibit the overlapping phenomenon. The overlapping phenomenon leads to two issues, including the feature coupling issue and the edge blurring problem. These, in turn, cause the model to inaccurately perceive the semantic information of the prohibited item categories and to inaccurately locate the edges.

Recently, several works have been studied for alleviating the above issue. Specifically, regarding the feature coupling issue, GADet [10] proposes an IAA label assignment strategy, taking a comprehensive consideration of the intersection over union (IoU) between positive samples and ground-truth

Received 31 January 2024; revised 12 June 2024 and 30 August 2024; accepted 26 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U22A2063, Grant 62173083, and Grant 62206043; in part by the National Key Research and Development Project of China under Grant 2022YFF0902401; in part by China Postdoctoral Science Foundation under Grant 2023M730517; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021B1515120064; in part by the Major Program of National Natural Science Foundation of China under Grant 71790614; and in part by the 111 Project under Grant B16009. *(Corresponding author: Tong Jia.)*

Mingyuan Li, Hao Wang, Bowen Ma, Hui Lu, Shuyang Lin, and Da Cai are with the College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110819, China (e-mail: 542027743@qq.com; ddsywh@yeah.net; 2010285@stu.neu.edu.cn; 2603813543@qq.com; 2210329@stu.neu.edu.cn; 1553317260@qq.com).

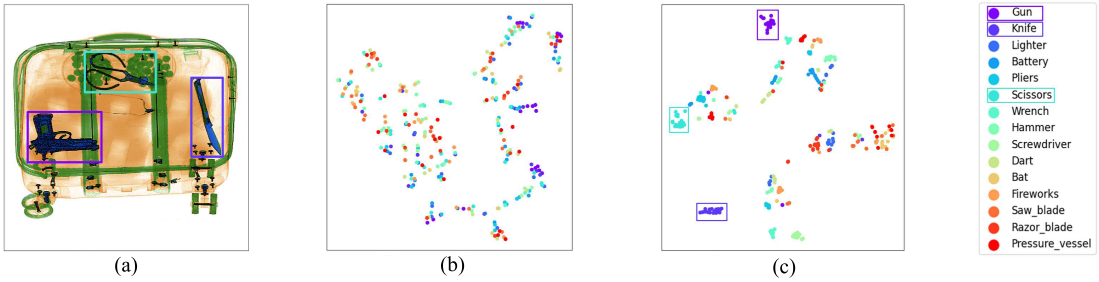
Tong Jia is with the College of Information Science and Engineering and the Key Laboratory of Data Analytics and Optimization for Smart Industry, Ministry of Education, Northeastern University, Shenyang, Liaoning 110819, China (e-mail: jiatong@ise.neu.edu.cn).

Dongyue Chen is with the College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110819, China, and also with the Foshan Graduate School of Innovation, Northeastern University, Foshan, Guangdong 528311, China (e-mail: chendongyue@ise.neu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2024.3487833

2162-237X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.



(a) (b) (c)

Fig. 2. t-SNE dimensionality reduction comparison. (a) Original X-ray image contains a gun, SCs, and a knife. (b) and (c) Distributions visualization of t-SNE dimensionality reduction of the object queries from the last decoder layer in DINO and AO-DETR, respectively.

labels. Xdet [4] proposes a hard-negative-sample selection scheme (HSS) label assignment strategy that effectively selects hard negative samples from complex backgrounds to train the proposed prohibited object detection network. The two label assignment methods mentioned above essentially aim at providing high-quality supervision information during the training phase, which directly augments the capacity of the model to discern features of all foreground categories amidst overlapping features. However, the two methods are relatively crude, as they treat each prohibited item category uniformly, whereas the nature and pattern of coupling between each category and the background differ. Regarding the edge blurring problem, some multistage regression methods are conducive for the model to achieve accurate positioning results. For example, deformable detection transformer (DETR) [11] proposes an iterative refinements boxes paradigm guided by the look forward once (LFO) scheme to obtain excellent localization results. DETR with improved denoising anchor boxes (DINO) [12] improves upon the LFO strategy in deformable DETR by proposing a look forward twice (LFT) scheme. These methods essentially involve repeatedly using the difference between predicted boxes and location labels to supervise and correct the model parameters. However, they have not clearly analyzed the reasons why LFT is superior to LFO, and the utilization of the location supervision signal is not sufficiently thorough.

In this article, we try to improve the anti-overlapping capability of the detector without changing the architecture of model DINO [12] and without adding any computational complexity. In addition, to pursue the optimal performance in X-ray image prohibited item detection field, we introduce a novel object detection model named Anti-overlapping DETR (AO-DETR), based on DINO, which is the state-of-the-art (SOTA) DETR-like model in natural image object detection field.

Specifically, to alleviate the adverse effects caused by the ambiguity of category semantics in object queries due to the overlapping of foreground and background in X-ray images, as illustrated in Fig 2(b), we propose a category-specific one-to-one assignment (CSA) strategy. Through CSA, the category-specific object queries will be stably assigned to ground-truth labels and reference boxes of the specific category of prohibited items, enabling it to specialize in extracting features of specific category of prohibited items from overlapping foreground and background. In addition, to mitigate the blurred boundary problem caused by overlapping phenomena, after conducting a detailed analysis of why LFT is more suitable for precise boundary localization than LFO, we propose the look forward densely (LFD) scheme, which can localize the edges of foreground objects more accurately. LFD is capable of transmitting gradients densely to multiple decoder layers, allowing low-level decoder layers to provide more reliable reference boxes to deformable attention in high-level decoder layers. This helps high-level decoder layers focus on learning how to predict accurate location information from blurry edges.

To prove the efficiency of our proposed methods, we conduct comprehensive experiments on the PIXray [3] dataset, OPIXray [2] dataset, and HIXray [13] dataset. The evaluation results demonstrate that our proposed model is superior to the SOTA object detector.

Our main contributions are summarized as follows.

1) We propose a powerful end-to-end object detector for overlapping phenomena in X-ray images, AO-DETR. To the best of the authors' knowledge, this is the first DETR-like model in the field of prohibited item detection. Experimental results show that our model achieved the best performance on multiple datasets.
2) We propose a CSA strategy, which enhances the anti-overlapping feature extraction capability for specific category foregrounds by constraining the object classes assigned to category-specific queries during the training phase.
3) The proposed LFD scheme improves the accuracy of reference boxes predicted by mid-level and high-level decoder layers through dense gradient transmission, ultimately enhancing the perception ability of blurry edges of models.

The remainder of this article is as follows. Initially, Section II provides a review of the relevant methods. This is followed by Section III, where we describe our proposed method in detail. Section IV then presents and discusses the results of our experiments. This article concludes with Section V, summarizing our key findings and observations.

II. Related Works

We briefly summarize some SOTA object detection methods that have achieved remarkable results. In addition, we introduce existing methods for enhancing anti-overlapping feature extraction capabilities from two pathways: accurate label assignment and multistage regression.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

A. General Object Detector

Recently, the object detectors based on convolutional neural networks, such as Faster regions with CNN features (R-CNN) [14], Cascade R-CNN [15], FCOS [16], ATSS [17], and YOLOX [18], are gradually overtaken by detectors with detection transformer framework. DETR [19] first proposes an end-to-end detection transformer framework while requiring no hand-crafted nonmaximum suppression (NMS) [20]. Deformable DETR [11] presents the deformable attention to accelerate convergence of DETR, which combines the advantage of the sparse spatial sampling of deformable convolution, and the global modeling capability of Transformers. DAB-DETR [21] presents a new query formulation and using dynamic anchor boxes for DETR. DN-DETR [22] and DINO [12] enhance DETR-like models from the perspectives of denoising queries and anchor boxes, respectively. Co-DETR [23] combines some traditional label assignment strategies and Hungarian matching to mitigate the sparse supervision problem caused by the one-to-one set matching strategy.

B. Label Assignment

Labeling anchors as positive or negative samples is crucial in detector training, which can mitigate the foreground–background class imbalance problem in X-ray images. Traditional anchor-based detectors like YOLOv3 [24], SSD [25], Faster R-CNN [14], and RetinaNet [26] use IoU for label assignment. Anchor-free detectors, such as FCOS [16] and Foveabox [27], employ center sampling strategies. However, these methods are suboptimal due to their fixed rules. Some advanced label assignment strategies, e.g., OTA [28], ATSS [17], PAA [29], and SimOTA [18] offer dynamic positive sample selection. OTA [28] views label assignment globally, treating it as an optimal transportation problem. ATSS [17] uses top-k anchors for threshold determination, while PAA [29] applies a probabilistic method. SimOTA [18] focuses on overall cost for better positive sample selection.

Recently, DETR [19] pioneered the use of global matching cost and the Hungarian algorithm to achieve a unique prediction result for each object in images. This represents the first successful application of a one-to-one label assignment scheme. However, the bipartite graph matching is unstable in the early stage of the training phase. DN-DETR [22] introduces noised bounding boxes and labels that bypass the need for Hungarian matching, thereby mitigating the issue of unstable assignments. DINO [12] further put forward contrastive denoising (CDN) training to accelerate convergence. Group-DETR uses a groupwise one-to-many label assignment, akin to the hybrid matching of H-DETR [30], for multiple positive object queries. Conversely, Co-DETR introduces a collaborative optimization approach for one-to-one set matching, differing from these follow-ups. Recently, some research, such as SP [31] and BCNet [32], has proposed an unsupervised pretraining strategy based on saliency prompt to provide prior knowledge for queries and kernels, thereby enhancing model performance.

However, in the object detection domain, the role of the learned queries in DETR is still not fully understood or utilized. We are experimenting with enabling queries to specifically target information from certain categories, thereby enhancing their ability to extract features in overlapping scenarios.

C. Multistage Regression

One-stage architectures such as YOLO [24], [33], [34] and SSD [25], [35], [36] series, designed for real-time performance. These models forgo separate region proposals of the R-CNN framework, processing the image in one pass. Although faster than region-based models, they initially faced challenges with lower detection accuracy. Two-stage detectors, such as Fast R-CNN [37], Faster R-CNN [14], and R-FCN [38] utilize a dual-step approach in object detection. Initially, they generate region proposals or candidate bounding boxes likely to contain objects. Subsequently, these proposals are classified and refined for the final detection outcome. Cascade R-CNN [15], a multistage detection system, tackles challenges by using detectors trained with increasing IoU thresholds. This approach allocates different IoU threshold proposals as positive samples at each stage, enabling the refinement of regression progressively. Deformable DETR [11] designed a simple iterative mechanism for bounding box refinement to improve detection performance, where each decoder layer refines the boxes based on the output of the prior layer. To overcome the shortsightedness of refining boxes in each decoder layer, while keeping the advantages of fast convergence, DINO introduces a novel LFT scheme, where the updated parameters are corrected using predictions from the current layer and the next lower layer. Recent works in other fields, such as EfficientNet [39] and DNTDF [40], have proposed utilizing dense connections to explore the efficient use of mid-to-high-level semantic features.

In this work, we seek to extend this approach, allowing the gradient of the prediction results from the current layer to propagate to the current layer and each lower decoder layer.

D. X-Ray Object Detection

Some CNN-based object detectors [1], [2], [3], [4], [5], [6], [7], [8], [9] have contributed significantly to security inspection. SIXray [1] introduces a class-balanced hierarchical refinement to address X-ray image complexities and class imbalance. OPIXray [2] develops a deocclusion attention module for X-ray image occlusions, enhancing feature maps with distinct item appearances. PIXray [3] offers a dataset with diverse prohibited items and proposes a dense deoverlap attention snake for segmentation. Xdet [4] first statistically analyzes the physical size distribution of different prohibited object categories and found that their physical sizes exhibit clear distinctions. However, GADet [10] points out that the areas of objects obtained by the Otsu [41] algorithm are not accurate enough in Xdet [4], and the area is not as robust as the diagonal length in X-ray images. Thereby, they introduce the physical diagonal length constraint (PDLC), which can utilize this underlying relationship to align classification and localization tasks in object detection. Although these methods utilize CNN frameworks, we aim to enhance DETR-like models specifically for X-ray prohibited item detection.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

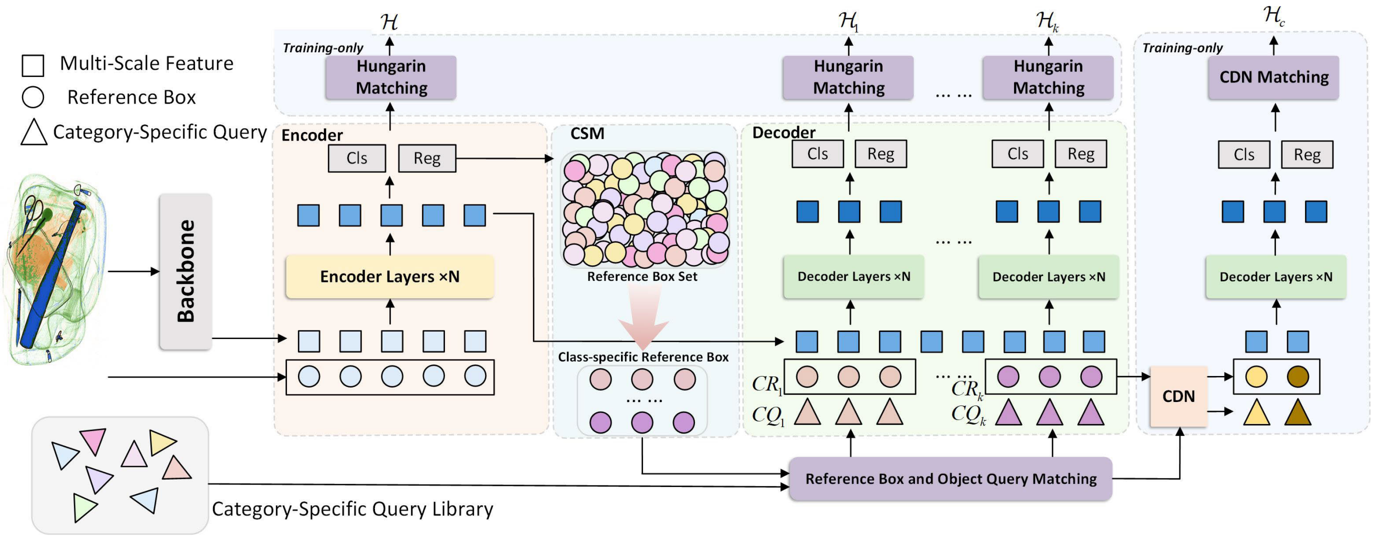


Fig. 3. Architecture of AO-DETR. The backbone, encoder, decoder, and CDN modules are the same as DINO [12]. For the CSA strategy, we match the category-specific high-quality reference boxes obtained from category-specific select mechanism (CSM) with their corresponding category-specific object queries before inputting them into the decoder module for prediction. We further employ an additional k-category-specific Hungarian matching mechanism to conduct one-to-one matching on the predicted results. This process serves to enhance the semantic clarity of the object query categories.

III. Proposed Method

The proposed AO-DETR incorporates CSA strategy on the basis of DINO. The overview of the model system is illustrated in Fig. 3. For the CSA strategy, we match high-quality reference boxes of specific categories obtained from CSM and their corresponding specific category object queries and then input them into the decoder module used for prediction. We further adopted an additional k-class-specific Hungarian matching mechanism to predict the results. This process is used to enhance the semantic clarity of object query categories.

A. Revisit the Match Part Label Assignment of DINO

In the DINO framework, the encoder outputs a large number of prediction results P, encompassing classification results C, and localization results or reference boxes R. Similar to the decoding process of most deformable DETR series models, they randomly initialize N_{pred} object queries Q^0, to guide the corresponding number of reference boxes R^0 selected from R to predict objects, where sets Q^0 and R^0 will be used by the decoder layer 0 for subsequent predictions. Then, the selection mechanism in DINO picks the top N_{pred} prediction results, denoted by P^0, which have the highest classification confidence scores among the set P. In addition, they are then assigned one-to-one to each q^0, where $q^0 \in Q^0$, in the descending order of classification confidence. This establishes the pairing relationship between the reference box and the object query for the decoder. For ease of description, we utilize the index i to represent the pairing relationship among reference box, classification confidence, object query, and ground truth. Elements with the same subscript i possess a pairing relationship. To obtain the ith pair of object query and reference box in layer l, where $q_i^l \in Q^l$ and $r_i^l \in R^l$, the decoder procedure of lth layer decoder layer can be expressed in the following form:

$$q_i^l, r_i^l, c_i^l = \mathcal{L}^l\left(q_i^{l-1}, r_i^{l-1}, X; \theta^l\right) \tag{1}$$

where $l \in \{x \mid x \in \mathbb{Z}, 0 < x \leq L\}$, representing the layer index, and $L = 6$ denotes the total number of layers. θ^l is the parameters in lth layer and X denotes the multiscale features extracted by the encoder.

To establish the one-to-one correspondence between the ground-truth labels G and the prediction results P^l from the lth layer of the decoder, it is necessary to compute a matching cost matrix $M \in \mathbb{R}^{N_{\text{pred}}, N_{\text{gt}}}$. N_{gt} and N_{pred} represent the numbers of ground truths and predictions, respectively. The specific method for computing this cost matrix can be referenced from DETR. Then, they apply the Hungarian algorithm for bipartite graph matching, and the process can be represented as follows:

$$\left\{p_i^l, q_i^l, g_i \mid q \in \mathbb{Z}, 0 \leq q < N_{\text{gt}}\right\} = \mathcal{H}^l\left(P^l, Q^l, G\right). \tag{2}$$

With this step, the algorithm establishes the pairing among the reference boxes, classification confidences, object queries, and ground truths, which can be represented by the subscript "i."

B. Category-Specific One-to-One Assignment

Although the aforementioned method has achieved significant success, the issue of the unclear categorical significance of queries remains unresolved. As shown in Fig. 2, upon processing X-ray images with significant overlap, the queries in the final layer of the decoder, after being reduced in dimensionality, exhibit a high degree of coupling and are scattered all over the plane. This indicates that the queries have extracted an extensive array of diverse backgrounds information and that a single query is unable to extract the features of a category-specific foreground object from the overlapping foreground and background. We propose a CSA strategy to alleviate the issue of unclear categorical significance in queries, while simultaneously enhancing the anti-overlapping capabilities of feature extraction of queries. The four main components of the strategy are as follows.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

1) Category-Specific Object Queries: Specifically, we initialize a category-specific object query prototype library at random, denoted by set CQ, each category-specific query prototype $q_k^0 \in CQ$. For the kth category, we define the category-specific query group Q_k^0, which contains N_k identical q_k^0. Due to the presence of positional encoding and the allocation of different ground truths, each query $q_{k,i}^0$ within the kth query group gradually diverges during the training phase. Since each of them is trained from the same prototype, this naturally leads to an effect where queries within the same category become similar, while those from different categories become increasingly dissimilar.

2) Category-Specific Select Mechanism: In decoders with deformable attention, queries guide networks in feature extraction from areas near reference box centers to refine these boxes. Aligning the categories of queries and reference boxes is crucial. Queries adept in detecting category A, denoted by $Q_{k=A}$, when paired with reference boxes of category B, denoted by $R_{k=B}$ might lead to network confusion. Specifically, the region within $R_{k=B}$ containing only class B foregrounds and overlapping backgrounds cannot provide the feature of class A for $Q_{k=A}$. Similarly, $Q_{k=A}$ cannot effectively direct the network to extract features of class B in the region with $R_{k=B}$ for object detection. Hence, keeping the category of reference boxes and queries consistent aids the network in extracting relevant foreground features from overlapping backgrounds, enhances the anti-overlapping capability of feature extraction of the model.

In mainstream DETR detectors such as deformable DETR and DINO, the select strategy directly chooses the top N reference boxes with the highest confidence scores from all encoder predictions and then arranges them in sequence with the object queries in the decoder. This approach, in fact, overlooks the consistency between the categories of the queries and the categories of the reference boxes. To address this issue, we propose a CSM that provides reference boxes corresponding to each category-specific query group, thereby resolving the aforementioned problem of category inconsistency. The specific process is as illustrated in Algorithm 1. Given all the results P predicted by the last layer of the encoder, the number of categories K, and the number of reference boxes N required for all categories, this algorithm can filter out the top N_k predictions with the highest confidence for each category, which are category-specific reference predictions of class k, denoted by P_k^0. This process can be denoted as follows:

$$P^0 = \left\{P_k^0 \mid k \in \mathbb{Z}, 0 \le k < K\right\} = \mathrm{CSM}(P, N, K) \quad (3)$$

where P^0 represents all category-specific prediction results, including category-specific reference boxes R^0 and category-specific classification score C^0.

3) Reference Box and Object Query Matching: Before decoding, to align the categories of reference boxes and object queries, we perform one-to-one matching between the category-specific reference boxes selected by CSM and the category-specific query group with the same category k. Then, we obtain the ith pair object query $q_{k,i}^0 \in Q_k^0$ and reference box $r_{k,i}^0 \in R_k^0$, and i represents the matching pair index. This matching relation ensures that a category-specific query for a specific class consistently guides the prediction of the reference box containing the corresponding class throughout the training process. Furthermore, we obtain the i-th pair of object query and reference box in category k for layer l, where $q_{k,i}^l \in Q_k^l$ and $r_{k,i}^l \in R_k^l$, and the procedure of lth each decoder layer can be expressed in the following form:

$$q_{k,i}^l, r_{k,i}^l, c_{k,i}^l = \mathcal{L}^l\left(q_{k,i}^{l-1}, r_{k,i}^{l-1}, X; \theta^l\right) \quad (4)$$

if $l = 1$, $q_{k,i}^{l-1} \in Q_k^0$ and $r_{k,i}^{l-1} \in R_k^0$, which can be obtained by category-specific query library and CSM.

4) Category-Specific Hungarian Matching: To effectively establish a one-to-one correspondence in category k between the ground truth G and the prediction results of the l-th decoder layer P_k^l, we compute the category-specific matching cost matrix $M_k \in \mathbb{R}^{N_{k,\mathrm{pred}}, N_{k,\mathrm{gt}}}$. Here, $N_{k,\mathrm{pred}}$ and $N_{k,\mathrm{gt}}$ represent the number of ground truths and predictions for category k, respectively. Then, we utilize the Hungarian algorithm for bipartite graph matching for each category, and the process can be represented as follows:

$$\left\{p_{k,i}^l, q_{k,i}^l, g_i \mid q \in \mathbb{Z}, 0 \le q < N_{\mathrm{gt}}\right\} = \mathcal{H}_k^l\left(P_k^l, Q_k^l, G\right). \quad (5)$$

At this point, our CSA strategy establishes the pairing among reference boxes, classification scores, object queries, and ground truths in each category k. As shown in Fig. 2(c), in technique is a variation of stochastic neighbor embedding (t-SNE) dimensionality reduction visualization of all category-specific object queries in the last decoder layer of AO-DETR, which utilizes CSA strategy, category-specific queries responsible for gun, scissor (SC), and knife in the last decoder layer converge to distinct corners. They exhibit both intra-class clustering and interclass repulsion characteristics. This indicates that our CSA strategy has trained queries to extract specific foreground features for particular categories, demonstrating strong antioverlapping features.

Algorithm 1 CSM

```
Require:
  $N$ is the number of all queries;
  $P$ is the prediction results of the last encoder layer, which
  including the classification scores $C$ and reference boxes $R$;
  $K$ is the number of categories;
Ensure:
  build empty set for classification scores: $C^0 \leftarrow \emptyset$;
  build empty set for reference boxes: $R^0 \leftarrow \emptyset$;
  $N_k = N/K$;
  for $\forall$ category-specific predication results $P_k \in P$ do
    $C_k^0 \leftarrow$ select $N_k$ reference box $p_{k,p}$ from $P_k$ with highest $c_{k,p}$;
    $I_k \leftarrow$ obtain the index of $c_k^0$ in Set $C_k$;
    $R_k^0 \leftarrow$ get the corresponding reference boxes by $I_k$;
    $C^0 = C^0 \cup C_k^0$;
    $R^0 = R^0 \cup R_k^0$;
  end for
  $P^0 = R^0 \cup C^0$;
  return $P^0$;
```

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

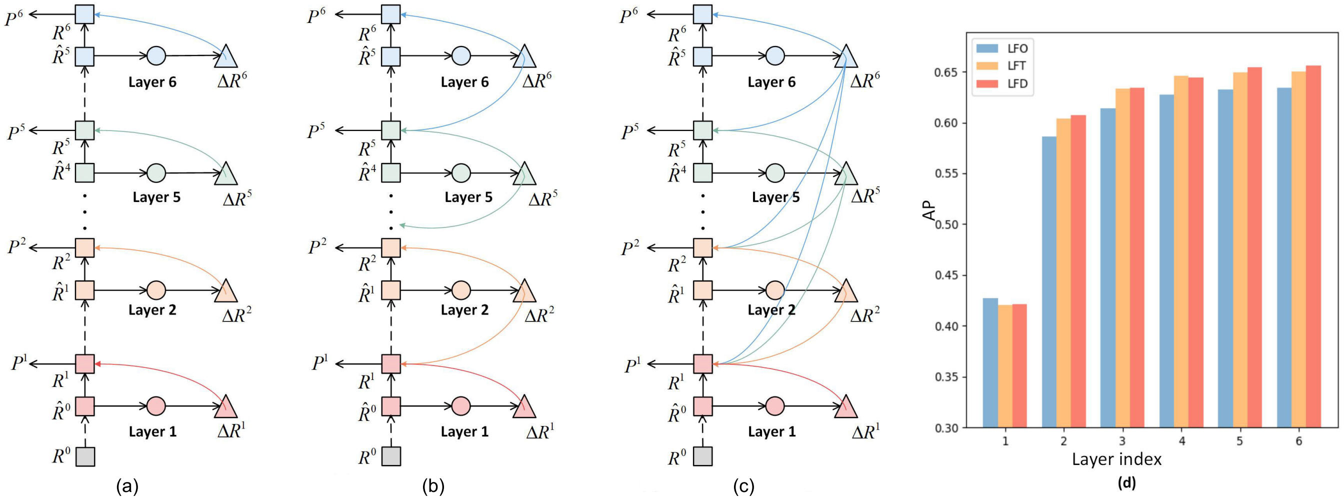


Fig. 4. (a)–(c) Comparing the structures of LFO, LFT, and LFD. (d) APs of LFO, LFT, and LFDD in each decoder layer. "LFO," "LFT," and "LFD" are the corresponding abbreviations.

C. Look Forward Densely

As shown in Fig. 1, due to the presence of overlapping phenomenon, the edges of the firework in (a), as well as the handle of the screwdriver in (b), exhibit significant edge blurring phenomenon. Therefore, the overlapping phenomenon in X-ray images leads to issues of inaccurate localization, and precise regression and localization of edges are of critical importance.

Deformable DETR [11] was the pioneer in introducing an iterative bounding box refinement mechanism in the decoder, assisting the decoder layers in iteratively locating the edges of objects. The changes in the predicted boxes during the process have been shown in Fig. 4(a), and can be represented by the following equations. Given normalized localization boxes R^{l-1} predicted by the $(l-1)$th decoder, they first stop the gradient as follows:

$$\hat{R}^{l-1} = \mathcal{B}\left(R^{l-1}\right) \tag{6}$$

where $\mathcal{B}(\cdot)$ represents blocking gradient propagation, a mainstream approach when paired with layer-specific auxiliary loss [42]. Then, the obtained $\hat{R}^{l-1}$ are used as the input reference box for the lth decoder layer as follows:

$$\Delta R^l = \mathcal{L}_{\text{reg}}^l\left(Q^{l-1}, \hat{R}^{l-1}, X; \theta^l\right) \tag{7}$$

where $\mathcal{L}_{\text{reg}}^l(\cdot)$ represents the regression prediction process at the lth layer. ΔR^l denotes the predicted offset, including x, y, w, and h [11]. Finally, $\hat{R}^{l-1}$ and ΔR^l are utilized for obtaining the normalized localization results of the lth decoder layer, denoted by R^l, as follows:

$$R^l = \sigma\left(\sigma^{-1}\left(\hat{R}^{l-1}\right) + \Delta R^l\right) \tag{8}$$

where $\sigma(\cdot)$ and $\sigma^{-1}(\cdot)$ denote the sigmoid and inverse sigmoid functions, respectively. It is important to note that this box update approach is designed to ensure that the updated boxes have normalized x, y, w, and h values, which range between 0 and 1. In the aforementioned method, to obtain lth layer reference boxes R^l, the deviation predicted by the lth layer ΔR^l serves only to correct the reference box of $(l-1)$th layer $\hat{R}^{l-1}$, and the parameters for lth layer θ^l are updated based exclusively on the corresponding layer-specific auxiliary loss. Consequently, this approach has been termed the LFO strategy by some researchers [12]. DINO introduces a better LFT scheme, which utilizes ΔR^l to guide the prediction processes of both R^l and R^{l-1}. LFT scheme tends to degrade the prediction results of the previous layer while improving those of the current layer. This refinement process of reference boxes of layer l can be described as follows:

$$R^l = \sigma\left(\sigma^{-1}\left(\hat{R}^{l-1}\right) + \Delta R^l + \Delta R^{l+l}\right). \tag{9}$$

Analyzing the influence of the predicted offset of a single layer that uses the LFT strategy, the offset of the current layer will interfere with the localization results of one lower layer, while directly improving the accuracy of the localization results of the current layer and indirectly improving the accuracy of the localization results of the all higher layers. Therefore, the current layer with the LFT strategy is negatively influenced by one higher layer while positively influenced by one lower layer. In a holistic view, the lowest layer is only negatively influenced, whereas the highest layer receives only positive guidance. The middle layers are subjected to both positive and negative influences. Since the offsets from the higher layers are generally smaller, for one layer, the negative guidance from the higher layer tends to be weaker than the positive guidance from the current layer. As a result, the lower layers suffer, the middle layers benefit marginally, and the upper layers gain significantly. During the network inference, the network outputs only the predictions of the highest layer, making the LFT strategy advantageous for precise regression at the edges of objects. However, LFT is relatively conservative. We attempt to further extend this strategy into a more dense form of guidance, termed LFD. The LFD allows the offset predicted by the current layer to participate in the prediction of the reference boxes of all lower layers. The prediction process of reference boxes in layer l can

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

be represented in the following form:

$$R^l_{S,E} = \sigma\left(\sigma^{-1}\left(\hat{R}^{l-1}\right) + \sum_{n=l}^{L} \Delta R^l\right) \tag{10}$$

where subscripts "S" and "E" represent the sum with equaling weighting factor. The current layer with the LFD strategy is negatively influenced by all higher layers while positively influenced by all lower layers. Compared with LFT, the lowest layer is negatively influenced by other all layers, and the highest layer receives positive guidance from other all layers. The middle layers obtain benefits more times than LFT. As a result, compared with LFT, the lower layers suffer more, the middle layers benefit more, and the upper layers gain more. This results in localization by the final layer being more precise, effectively countering the inaccuracies caused by edge blurriness. In comparison with the LFT strategy, as shown in Fig. 4(d), our LFD further enhances the predictive outcomes of both the middle and the higher layers. Furthermore, we further explore the effects of geometrically scaling the offsets from different layers, either by amplifying or diminishing them in a proportional manner, as demonstrated in the following equation:

$$R^l_{S,A} = \sigma\left(\sigma^{-1}\left(\hat{R}^{l-1}\right) + \sum_{n=l}^{L} \Delta R^l / 2^{L-l}\right) \tag{11}$$

$$R^l_{S,D} = \sigma\left(\sigma^{-1}\left(\hat{R}^{l-1}\right) + \sum_{n=l}^{L} \Delta R^l / 2^{l}\right) \tag{12}$$

where subscripts "A" and "D" represent the amplifying and diminishing weighting, respectively. Finally, we also attempted to apply the averaging operation separately to each of the aforementioned three methods, denoted by $R^l_{V,E}$, $R^l_{V,A}$, and $R^l_{V,D}$, where subscript "V" means the average operation. The most effective form among them is $R^l_{V,D}$. For detailed experimental results, please refer to Section IV, where a comprehensive analysis and presentation of the outcomes are provided.

D. Foreground Instability Score (FIS)

The Hungarian matching algorithm utilizes the globally optimal solution of the cost matrix for label assignment. However, during the training of the network, distinct ground truth objects may be assigned to a specific query at different epochs. This variability leads to instability in label assignments for queries, manifesting in two aspects: instability in foreground categories assignment and instability in foreground objects assignment. In order to quantitatively evaluate the instability of foreground label assignment, we designed a metric named FIS as follows. For one training image, decoders of model predict objects $P^j = \{P^j_0, P^j_1, \ldots, P^j_{N_{\text{pred}}-1}\}$, where N_{pred} represents the number of predicted objects, and the ground-truth objects are denoted as $G = \{G^j_0, G^j_1, \ldots, G^j_{N_{\text{gt}}-1}\}$, where N_{gt} represents the number of ground-truth objects. After label assignment, we compute an index vector $V^j = \{V^j_0, V^j_1, \ldots, V^j_{N_{\text{pred}}-1}\}$ to store the ground-truth object assignment results for epoch j as follows:

$$V^j_n = \begin{cases} m, & \text{if } P^j_n \text{ matches } G_m \\ -1, & \text{if } P^j_n \text{ matches nothing.} \end{cases} \tag{13}$$

Similar, we compute an index vector $T^j = \{T^j_0, T^j_1, \ldots, T^j_{N_{\text{pred}}-1}\}$ to store the ground-truth object assignment results for epoch j as follows:

$$T^j_n = \begin{cases} c, & \text{if } P^j_n \text{ matches } c\text{-th category } G_m \\ -1, & \text{if } P^j_n \text{ matches nothing.} \end{cases} \tag{14}$$

We define the foreground category instability of epoch j for one training image as the difference between its T^j_0 and T^{j-1}_1 as follows:

$$\text{FCS}^j = \sum_{n=0}^{N_{\text{pred}}} \mathbb{1}\left(T^j_n \neq T^{j-1}_n\right) \cdot \mathbb{1}\left(T^j_n \neq -1 \wedge T^{j-1}_n \neq -1\right) \tag{15}$$

where $\mathbb{1}(x)$ is 1 if x is true and 0 otherwise, and the symbol "$\wedge$" represents logical AND. $\mathbb{1}(T^j_n \neq -1 \wedge T^{j-1}_n \neq -1) = 1$ means that T^j_n and T^{j-1}_n are both responsible for foreground objects. Similar, we define the foreground objects instability of epoch j for one training image as the difference between its V^j_0 and V^{j-1}_1 as follows:

$$\text{FOS}^j = \sum_{n=0}^{N_{\text{pred}}} \mathbb{1}\left(V^j_n \neq V^{j-1}_n\right) \cdot \mathbb{1}\left(V^j_n \neq -1 \wedge V^{j-1}_n \neq -1\right). \tag{16}$$

Finally, we take the average of both and normalize it by the number of predicted objects N_{pred}

$$\text{FIS}^j = \frac{\text{FCS}^j + \text{FOS}^j}{2 \cdot N_{\text{pred}}}. \tag{17}$$

The instability of epoch j for the entire dataset is averaged over the instability numbers for all images. We omit the image index for notation simplicity in (13)–(17).

FIS comprehensively considers the instability of both foreground category assignments and object assignments, the lower the FIS value, the more stable the label assignment between object queries and foreground objects. In Section IV, we will analyze the impact of CSA on the model training process using our FIS metric and the instability score (IS) metric [22].

IV. Experiments

In this section, we first conduct comprehensive ablation experiments on ResNet-50 [43] and Swin-L [44] backbone networks to analyze the compatibility of CSA and LFD. Then, we analyze the impact of CSA on the model training process using our FIS metric and the IS metric [22]. Subsequently, we explore six schemes of dense guidance in LFD. Finally, we design extensive experiments on the PIXray, OPXray, and HIXray datasets to assess the performance of our model. We compare our model with SOTA models from both general detectors and prohibited items detectors in a unified environment. Furthermore, we commence with a visual analysis

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

of the sampling points of the decoder layer in AO-DETR equipped with CSA. Finally, we visualize the results of baseline on images with severe overlapping phenomena and draw conclusion.

A. Experimental Setup

1) Datasets and Evaluation Metrics: The PIXray [3] dataset consists of 5046 X-ray images of prohibited items, with 4046 images used for training and 1000 images for testing. The dataset covers 15 categories of prohibited items, including bat, knife, gun, wrench, pliers, hammer, scissors, saw blade, dart, razor blade, battery, screwdriver, lighter, fireworks, and pressure vessel.

The OPIXray [2] dataset contains 8885 X-ray images of prohibited items, with 7019 images used for training and 1776 images for testing. The dataset covers five categories of prohibited items, all of which are knives, including folding, straight, SC, utility, and multitool (MU).

The HiXray [13] dataset contains a total of 45 364 X-ray images, with 36 295 images used for training and 9069 images for testing. The dataset includes eight categories of items: portable charger 1, portable charger 2, mobile phone, laptop, tablet, cosmetic, water, and nonmetallic lighter.

For the PIXray dataset [3], we utilize the COCO [45] evaluation metrics. The primary challenge metric is the mean average precision (mAP), computed across ten IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. AP_{50} represents the mAP calculated at a single IoU threshold of 0.5, while AP_{75} represents the mAP at a single IoU threshold of 0.75. In addition, AP_S, AP_M, and AP_L represent the AP for small objects (area $< 32^2$), medium objects ($32^2 <$ area $< 96^2$), and large objects ($96^2 <$ area), respectively.

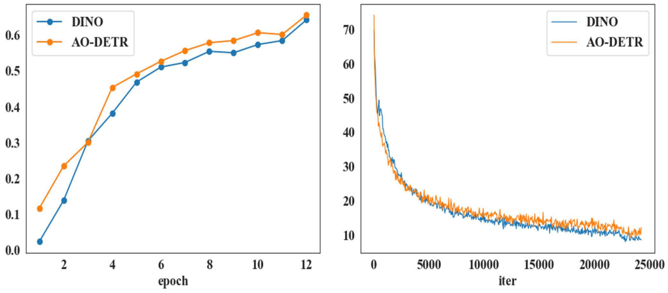
For OPIXray and HiXray, we adopt the VOC [46] evaluation metric. AP is calculated from the area under the precision–recall curve of one category at the IoU threshold of 0.5. The mAP (mAP) is then computed as the average of AP of all categories. mAP serves as a comprehensive evaluation criterion, effectively representing the accuracy and recall of the detector. It provides a holistic assessment that captures the performance strengths and weaknesses of the detector.

2) Implementation Details: For the sake of fair comparison, we ensure that all models are trained under identical conditions. Each model utilizes the ImageNet [47] pretrained model, including ResNet-50 [43], ResNeXt-101 [48], and Swin-L [44]. The convolutional models employ the SGD optimizer with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.1, while transformer models, such as DINO utilize the AdamW optimizer with a learning rate of 0.0001, and weight decay of 0.0001. All models undergo 12 training epochs and are implemented based on the MMDetection framework. For warm-up scheme of convolutional models, during the initial 500 iterations, the learning rate gradually increases in a linear fashion with a warming-up ratio of 0.001. Following this warm-up phase, the learning rate undergoes a stepwise decrement, with specific adjustments occurring at the 8-th and 11-th epoch. For DETR-like models, following Deformable DETR [11], the learning rate is decayed at the 11th epoch by a factor of 0.1. Our AO-DETR utilizes l_1 and GIoU [49] losses for box regression and quality focal loss [50] for classification. All training is conducted on a uniform computer platform equipped with an NVIDIA GeForce RTX 4090 GPU, an Intel Core i9-13900K CPU, 64-GB memory, Windows 10 system, and PyTorch 1.13.1.

TABLE I

ABLATION RESULTS OF THE PROPOSED CSA AND LFD ON THE PIXRAY [3] DATASET. "CSA" AND "LFD," RESPECTIVELY, REPRESENT THE PROPOSED CSA STRATEGY AND LFD SCHEME

Backbone	CSA	LFD	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50	✗	✗	64.3	86.5	71.0	19.3	48.9	73.9
	✓	✗	65.0	86.1	**72.7**	22.8	50.1	74.0
	✗	✓	65.2	**86.7**	71.5	23.7	49.2	74.5
	✓	✓	**65.6**	86.1	72.0	**23.9**	**50.7**	**74.8**
Swin-L	✗	✗	72.8	90.0	80.1	38.3	60.4	80.4
	✓	✗	73.2	90.3	79.9	38.6	62.4	80.6
	✗	✓	73.4	**90.6**	79.4	39.8	61.6	81.5
	✓	✓	**73.9**	89.9	**80.6**	**40.5**	**62.4**	**81.6**



(a) (b)

Fig. 5. (a) AP curve of DINO [12] and AO-DETR on PIXray [3] dataset. (b) Loss convergence curve of DINO and AO-DETR on PIXray [3] dataset.

B. Ablation Study

1) Ablation Study of AO-DETR: As shown in Table I, we conducted ablation experiments on the PIXray dataset using ResNet-50 and Swin-L backbones, respectively. These experiments aimed to analyze the impact of our CSA and LFD on the baseline detection performance and assess the compatibility between the two methods. CSA and LFD, respectively, improve the AP of DINO with ResNet-50 from 64.3% to 65.0% and 65.2%, which demonstrates their effectiveness. Simultaneously integrating both CSA and LFD, AO-DETR achieves the highest AP of 65.6%, AP_S of 23.9%, AP_M of 50.7%, and AP_L of 74.8%. This underscores the exceptional performance of AO-DETR and highlights the complementary nature of CSA and LFD.

In addition, we also compare the AP and loss convergence curves of AO-DETR and DINO, as shown in Fig. 5. In terms of the AP curve, AO-DETR consistently outperforms the baseline throughout the entire process. Regarding the loss curve, AO-DETR demonstrates a faster convergence in the early stages and stable convergence in the later stages.

2) Ablation Study of CSA: To comprehensively analyze the impact of CSA on the model training process, we analyze the

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

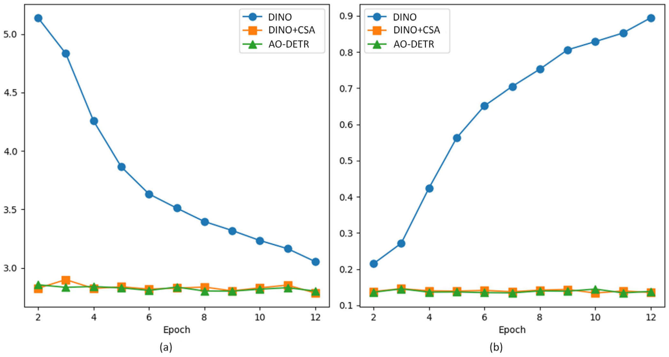


Fig. 6. (a) IS [22] of DINO, DINO+CSA, and DINO+CSA+LFD (AO-DETR) on PIXray [3] dataset. (b) FIS of them.

instability of models during the training process using the IS metric [22] in conjunction with our FIS metric. IS [22] is an indicator that evaluates the instability of object assignments without considering whether the predicted results of object queries are allocated to background objects. The lower the IS value, the more stable the label assignment between object queries and all objects including foreground and background. As illustrated in Fig. 6, with the training of DINO, the IS value decreases. This is attributed to the network's improved ability to discern whether a query is suitable for foreground object responsibility, reducing the probability of erroneously assigning low-capacity queries to predict foreground objects. However, due to the inherent randomness in the original Hungarian matching method, queries are uncertain about their responsibility for a specific category or object in the foreground. This confusion deepens as the network undergoes training, ultimately leading to an increase in the FIS metric.

In contrast, CSA alleviates this confusion significantly by constraining the object classes assigned to category-specific queries during the training phase. Therefore, with the incorporation of CSA, this confusion faced by object queries is greatly mitigated, and both IS and FIS metrics consistently maintain at lower levels. As the object queries become responsible for fixed categories, they can specialize in the features of objects belonging to those categories, thereby enhancing the anti-overlapping feature extraction capability of the network.

3) Ablation Study of LFD: To fully explore the potential of the LFD scheme, we design and compare six different dense guidance terms, employing equaling (E), amplifying (A), and diminishing (D) strategies in both averaging (V) and summing (S) modes. The specific formulas can be referred to in (10)–(12). As shown in Table II, we initially assess the impact of employing equaling, amplifying, and diminishing strategies under the summing strategy on the AP. We observe that $R^l_{S,D}$ achieves a superior AP of 64.9% compared with $R^l_{S,A}$ and $R^l_{S,E}$ with 63.9% and 64.8%, respectively. Subsequently, under the averaging strategy, we compare the AP for employing equaling, amplifying, and diminishing strategies. Here, the AP of $R^l_{V,D}$ at 65.2% outperforms the AP of $R^l_{V,A}$ and $R^l_{V,E}$, which are 64.8% and 64.5%, respectively. In summary, "V" is superior to "S," "D" is superior to both "A" and "E," and the overall performance of mode $R^l_{V,D}$ is the best, demonstrating the highest detection accuracy.

TABLE II

Ablations for Different Types of LFD Strategy on the PIXray [3] Dataset

Type	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
R^l	64.3	86.5	71.0	19.3	48.9	73.9
$R^l_{S,E}$	64.8	86.2	71.7	21.6	50.1	74.0
$R^l_{S,A}$	63.9	85.5	70.8	20.2	48.2	73.6
$R^l_{S,D}$	64.9	86.4	71.8	22.9	49.9	73.8
$R^l_{V,E}$	65.0	86.3	72.5	22.3	**50.6**	74.3
$R^l_{V,A}$	64.5	85.4	72.1	22.1	49.5	74.2
$R^l_{V,D}$	**65.2**	**86.7**	**72.8**	**23.7**	50.2	**74.5**

C. Comparison With SOTA Methods

1) Comparison With General Detectors: To validate the superior performance of the proposed model for X-ray image prohibited item detection compared with general object detectors, we compare AO-DETR with ResNet-50 and AO-DETR with Swin-L against SOTA detectors in mainstream prohibited item datasets PIXray [3], OPIXray [2], and HIXray [13], as well as general-domain detectors. These general detectors include models based on convolutional neural networks, such as Faster R-CNN [14], Mask R-CNN [51], Cascade R-CNN [15], GFLv1 [50], ATSS [17], and models based on Transformers, such as deformable DETR and DINO. For fairness, all models are trained for 12 epochs, which are all reimplemented by MMDetection [58], and images are resized to 320 × 320. The number of object queries in DINO and category-specific object queries of our AO-DETR remains consistent, both being twice the number of categories. As shown in Table III, our AO-DETR with Swin-L achieves the highest detection accuracy on the PIXray dataset, with an AP of 73.9%, surpassing other general detectors significantly. To balance real-time requirements, we introduce a lightweight version, AO-DETR with ResNet-50, reducing PARAMS from 229 to 58.38 M, GFLOPs from 156 to 26.89 G, and improving frame rate from 40 to 54 frames/s. The AP remains high at 65.6%, surpassing the best general detector GFLv1 at 57.5%, with the exception of DINO. Furthermore, AO-DETR, compared to the baseline model DINO, shows no increase in required GFLOPs and PARAMs during the inference process, and there is no decrease in FPS. This indicates that our approach does not require additional computational resources during inference. While maintaining inference speed, the detection accuracy of the smaller model improves from 64.3% and 72.8% to 65.6% and 73.9%, respectively.

To demonstrate the robustness of our method, we perform a comparative analysis on two other mainstream prohibited item detection datasets, OPIXray and HIXray. The results, as shown in Tables IV and V, align with the conclusions mentioned above, indicating the consistent performance of AO-DETR across different datasets. Specifically, the AO-DETR model, which uses the Swin-L model as its backbone, achieved the highest core indicator mAP values on the OPIXray and HIXray datasets, with 80.8% and 75.8%, respectively, demonstrating its stability and superiority.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III

Comparison With SOTAGeneral Detectors on PIXray [3]. BS, PARAMs, GFLOPs, and FPS Represent Batch Size, the Total Number of Parameters, the Giga Floating Point Operations, and the Number of Inferences the Model Can Perform Per Second, Respectively

Method	Backbone	BS	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	PARAMs	GFLOPs	Reference
Faster R-CNN [14]	ResNet-50	16	87	51.2	79.9	57.6	3.1	35.4	60.1	41.19 M	20.36	TPAMI17
Faster R-CNN [14]	ResNeXt-101-32x4d	16	70	53.6	82.3	60.8	3.9	37.7	62.7	59.83 M	28.35	TPAMI17
Mask R-CNN [51]	ResNet-50	16	86	50.0	79.9	56.2	3.8	34.8	59.1	41.42 M	20.36	ICCV17
Mask R-CNN [51]	ResNeXt-101-32x4d	16	73	52.4	81.9	59.4	4.2	36.2	61.3	60.04 M	28.35	ICCV17
Cascade R-CNN [15]	ResNet-50	16	61	56.5	81.3	63.2	8.0	41.0	65.9	68.97 M	22.37	CVPR18
ATSS [17]	ResNet-101	16	66	52.8	80.8	60.2	7.0	37.4	63.6	51.14 M	27.82	CVPR20
GFLv1 [50]	ResNeXt-101-32x4d	16	66	57.5	82.8	66.0	9.1	42.0	67.4	50.70 M	28.51	NeurIPS20
Deformable DETR [11]	ResNet-50	2	60	44.6	74.2	48.5	9.6	30.0	53.0	52.14 M	13.47	ICLR21
DINO [12]	ResNet-50	2	54	64.3	86.5	71.0	19.3	48.9	73.9	58.38 M	26.89	ICLR23
DINO [12]	Swin-L	2	40	72.8	**90.0**	80.1	38.3	60.4	80.4	229.0 M	156.0	ICLR23
RT-DETR+MMCL [52]	ResNet-50	2	64	63.6	85.9	71.4	24.0	49.9	72.6	42.81 M	17.07	arXiv24
AO-DETR **(ours)**	ResNet-50	2	54	65.6	86.1	72.0	23.9	50.7	74.8	58.38 M	26.89	—
AO-DETR **(ours)**	Swin-L	2	40	**73.9**	89.9	**80.6**	**40.5**	**62.4**	**81.6**	229.0 M	156.0	—

TABLE IV

Comparison With SOTA General Detectors on OPIXray [2]. FO, ST, SC, UT, and MU Represent Folding Knife, Straight Knife, Scissor, Utility Knife, and Multitool Knife, Respectively

Method	Backbone	BS	FPS	mAP	FO	ST	SC	UT	MU	PARAMs	GFLOPs	Reference
Faster R-CNN [14]	ResNet-50	16	87	70.9	75.1	45.8	88.4	67.7	77.3	41.19 M	20.36	TPAMI17
Faster R-CNN [14]	ResNeXt-101-32x4d	16	70	73.4	80.6	45.4	89.1	69.1	83.1	59.83 M	28.35	TPAMI17
Mask R-CNN [51]	ResNet-50	16	86	74.7	77.9	51.4	89.5	69.4	85.5	41.42 M	20.36	ICCV17
Mask R-CNN [51]	ResNeXt-101-32x4d	16	73	77.2	83.6	55.9	89.8	71.5	85.2	60.04 M	28.35	ICCV17
Cascade R-CNN [15]	ResNet-50	16	61	72.8	75.7	50.0	89.4	70.0	79.0	68.97 M	22.37	CVPR18
ATSS [17]	ResNet-101	16	66	67.5	72.8	38.0	88.6	58.0	80.2	51.14 M	27.82	CVPR20
GFLv1 [50]	ResNeXt-101-32x4d	16	66	75.6	80.0	53.6	89.3	71.7	83.4	50.70 M	28.51	NeurIPS20
Deformable DETR [11]	ResNet-50	2	60	63.4	70.1	29.0	86.0	55.7	76.4	52.14 M	13.47	ICLR21
DINO [12]	ResNet-50	2	54	78.2	83.2	58.8	89.4	72.7	86.7	58.38 M	26.89	ICLR23
DINO [12]	Swin-L	2	40	80.0	84.2	61.1	89.0	**78.9**	86.6	229.0 M	156.0	ICLR23
RT-DETR+MMCL [52]	ResNet-50	2	64	62.5	65.9	22.3	86.4	57.1	80.7	42.81 M	17.07	arXiv24
AO-DETR **(ours)**	ResNet-50	2	54	79.2	83.8	60.5	90.1	74.7	87.1	58.38 M	26.89	—
AO-DETR **(ours)**	Swin-L	2	40	**80.8**	**84.8**	**63.0**	**90.1**	77.7	**88.4**	229.0 M	156.0	—

TABLE V

Comparison With SOTA General Detectors on HIXray [13]. PO1, PO2, WA, LA, MP, TA, CO, and NL Denote "Portable Charger 1 (Lithium-Ion Prismatic Cell)," "Portable Charger 2 (Lithium-Ion Cylindrical Cell)," "Water," "Laptop," "Mobile Phone," "Tablet," "Cosmetic," and "Nonmetallic Lighter"

Method	Backbone	BS	FPS	mAP	PO1	PO2	WA	LA	MP	TA	CO	NL	PARAMs	GFLOPs	Reference
Faster R-CNN	ResNet-50	16	87	72.6	87.8	83.9	87.4	90.0	89.7	88.9	49.8	3.4	41.19 M	20.36	TPAMI17
Faster R-CNN	ResNeXt-101-32x4d	16	70	73.1	87.4	83.8	86.9	89.9	89.6	88.9	49.2	9.1	59.83 M	28.35	TPAMI17
Mask R-CNN	ResNet-50	16	86	72.4	88.0	84.3	86.9	89.7	89.6	88.9	49.5	2.4	41.42 M	20.36	ICCV17
Mask R-CNN	ResNeXt-101-32x4d	16	73	72.9	87.8	84.0	86.5	89.8	89.8	89.0	52.2	3.9	60.04 M	28.35	ICCV17
Cascade R-CNN	ResNet-50	16	61	72.4	87.8	84.7	87.1	90.1	89.7	89.1	50.8	0.0	68.97 M	22.37	CVPR18
ATSS	ResNet-101	16	66	65.5	76.6	68.1	83.6	89.8	85.7	86.8	33.7	0.0	51.14 M	27.82	CVPR20
GFLv1	ResNeXt-101-32x4d	16	66	70.3	83.6	74.3	86.0	89.7	88.1	87.0	44.4	9.1	50.70 M	28.51	NeurIPS20
Deformable DETR	ResNet-50	2	60	70.4	85.8	82.5	83.2	89.4	89.1	88.1	44.6	0.6	52.14 M	13.47	ICLR21
DINO	ResNet-50	2	54	74.1	87.8	85.2	87.1	90.1	89.9	88.1	60.0	4.6	58.38 M	26.89	ICLR23
DINO	Swin-L	2	40	75.4	88.1	85.9	86.8	90.3	**90.0**	**89.3**	63.8	**9.4**	229.0 M	156.0	ICLR23
AO-DETR **(ours)**	ResNet-50	2	54	74.5	87.9	85.8	87.8	90.0	89.4	89.1	56.9	9.1	58.38 M	26.89	—
AO-DETR **(ours)**	Swin-L	2	40	**75.8**	**88.3**	**86.9**	**88.4**	**90.4**	89.1	89.0	**64.7**	9.2	229.0 M	156.0	—

2) Comparison With Prohibited Items Detectors: To validate the superiority of AO-DETR over other prohibited item detectors, we conducted comparisons on the OPIXray dataset. In order to challenge the performance limits of AO-DETR, we extend the number of training epochs from 12 to 15, increase the object query quantity from the default twice the number of categories to 20 times the number of categories, and enlarge the image input size from 320 × 320 to 640 × 640. In addition, we change the pretrained model from ResNet-50 and Swin-L on the ImageNet dataset to the entire DINO (ResNet-50) and DINO (Swin-L) on the COCO dataset. The relevant parameters for other prohibited item detectors were referenced from their respective papers. The results, as shown in Table VI, demonstrate that our AO-DETR with Swin-L detector, trained on lower resolution images for only 15 epochs, achieves an mAP surpassing other SOTA models trained at higher resolutions, such as POD-F and MCIA-FPN.

D. Visualization Analysis

1) Visualization Analysis of Sampling Points: As shown in Fig. 7, we take three X-ray images containing different categories of prohibited items as examples. From the last

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VI

Comparison With SOTA Prohibited Item Detectors on OPIXray [2]. IJON Stands for the Journal Neurocomputing. MAX Represents the Model Trained Until It No Longer Converges. "—" Indicates That the Data Are Not Published or Cannot Be Obtained due to the Model Not Being Open Source

Method	Backbone	Epoch	Input Size	mAP	FO	ST	SC	UT	MU	PARAMs	FPS	Reference
DOAM [2]	ResNet-50	MAX	–	82.41	86.71	68.58	90.23	78.84	87.67	90.79 M	—	ACM MM20
XDet [4]	ResNet-50	MAX	1280	86.69	90.42	75.95	91.46	84.31	91.29	41.19 M	25	KBS22
ATSS+LAreg [53]	ResNet-50	12	1280	87.39	**92.78**	71.17	96.61	83.45	92.92	—	—	TIFS22
ATSS+LAcls [53]	ResNet-50	12	1280	88.26	90.04	74.99	97.60	85.70	**92.96**	—	—	TIFS22
MCIA-FPN [54]	ResNet-101	MAX	–	82.59	89.08	74.48	89.99	86.13	89.75	—	—	TVC23
POD-F-R [55]	ResNet-50	24	1333	84.9	88.7	76.0	88.9	82.8	88.1	118.32 M	7	IJON23
POD-F-X [55]	ResNeXt-50	24	1333	86.1	89.4	78.7	90.6	83.3	88.7	119.67 M	6	IJON23
AC-YOLOv4 [56]	CSPDarknet53	300	416	85.69	—	—	—	—	—	—	—	MTA23
GADet-S [10]	Modified CSP v5	60	320	69.6	72.6	43.6	86.6	67.5	77.5	8.94 M	116	JSEN24
GADet-L [10]	Modified CSP v5	60	320	77.7	81.8	54.0	89.8	77.5	85.2	54.16 M	75	JSEN24
GADet-X [10]	Modified CSP v5	60	320	78.1	83.1	56.3	89.8	75.7	85.5	99.01 M	56	JSEN24
FDTNet [57]	ResNeXt-101	12	512	82.04	87.90	60.20	96.10	78.90	87.10	66.17 M	—	arXiv24
FDTNet [57]	ResNeXt-101	12	1333	88.02	91.50	74.60	97.60	85.20	91.20	66.17 M	—	arXiv24
AO-DETR **(ours)**	ResNet-50	15	640	87.2	90.0	80.1	90.8	85.6	89.5	58.38 M	29	—
AO-DETR **(ours)**	Swin-L	15	640	**89.0**	89.4	**80.4**	**97.8**	87.4	90.0	229.0 M	15	—

layer of the decoder, we select one category-specific object query from each of the four category-specific object query groups responsible for batteries, pressure vessels, fireworks, and razor blades. Then, we visualize their corresponding sampling points, reference points, and localization results. Overall, the category-specific object query prioritizes perceiving and focusing on regions in the images that exhibit the highest similarity to the features of prohibited items of its responsible category. Taking the category-specific object query for batteries as an example, it recognizes batteries in row (a) and row (c), while in row (b), it attends to the top of a pressure vessel, which bears the highest similarity to a battery. Thanks to category-specific one-to-one matching, even though the category-specific object query for batteries observes the top of a pressure vessel in row (b), it can still discern that the features it attends to do not belong to a battery, leading to the decision to withhold output predictions. Furthermore, the stability of category-specific object queries for category matching is remarkably high. In Fig. 7, there is no occurrence of a category-specific object query for category A focusing on and predicting prohibited items of category B. In row (a) and row (c), fireworks overlap significantly with background features, yet they are still accurately covered by the sampling points of the corresponding category-specific query. Moreover, when looking at the overall distribution of sampling points, those for batteries and razor blades are consistently densely concentrated in small areas, while those for fireworks and pressure vessels are consistently sparsely distributed in larger regions. This suggests that category-specific object queries extract features based on the size and shape characteristics of their responsible categories during the prediction process. In conclusion, we have successfully clarified the category semantics of object queries, using this as an opportunity to assist the network in identifying target objects in overlapping foreground–background scenarios.

2) Visualization Analysis of SOTA Comparison: As shown in Fig. 8, we conduct detection tests on three images using the SOTA general model YOLOX [18], the SOTA prohibited item detector GADet [10], and our AO-DETR. The detection results have been visualized for comparative analysis. The SOTA general object detector YOLOX, after being fine-tuned, still produces numerous false positives and misses in detecting items against overlapping backgrounds. For instance, in row (a), miscellaneous items are incorrectly identified as razor blades; in row (b), a razor blade is overlooked; and in row (c), fireworks are missed, and a hammer is mistakenly identified as pliers. Moreover, there are cases of redundant detection results, such as a hammer being detected multiple times in row (c). The SOTA prohibited item detector GADet alleviates some of the false positives and misses in detecting contraband against overlapping backgrounds, but severe overlaps still lead to detection issues. For example, a battery in row (a) is still missed, and the location of the razor blade in row (b) is not accurate enough. In addition, GADet fails to detect fireworks in row (c), and the localization of the wrench is imprecise. In comparison, AO-DETR delivers highly accurate detection results, managing to accurately perform both classification and localization tasks even in the presence of overlapping backgrounds.

3) Visualization Analysis of Ablation Study: We conduct an ablation analysis of CSA and LFD using visualization of the detection results of four X-ray images containing different prohibited items, as shown in Fig. 9. We enumerate two adverse effects caused by overlapping phenomena. One is the feature coupling resulting from the overlap of prohibited items and the background, leading to missed detections. For instance, in column (a), a pair of pliers is missed by DINO due to the overlap. The other effect is the edge blurring caused by overlap, subsequently leading to inaccurate edge localization. In column (b), DINO inaccurately locates the overlapping part of fireworks and a screwdriver with a baseball bat. Column (c) and column (d) depicts scenarios where both missed detections and inaccurate localization occur simultaneously. The CSA strategy enhances the perception of category-specific object queries for particular types of contraband, thereby

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

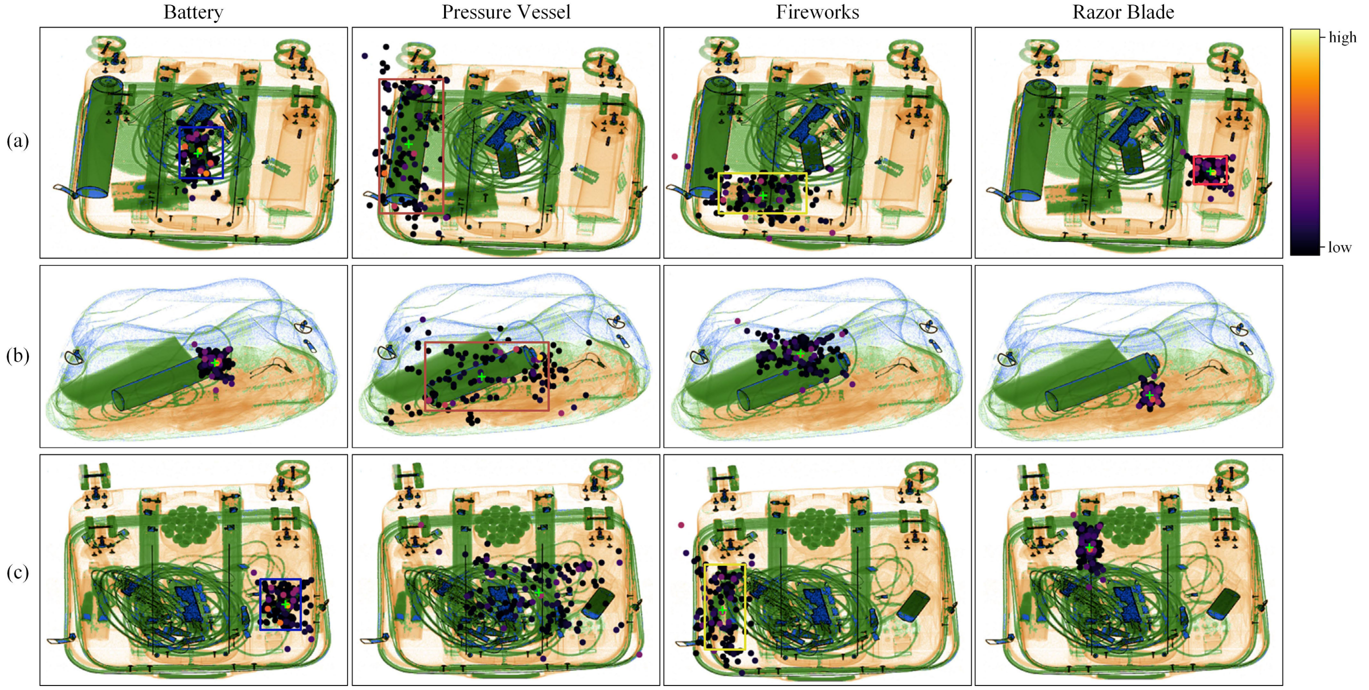


Fig. 7. Visualization of deformable attention sampling points and reference points of corresponding category-specific object queries in the last decoder layer. Each sampling point is depicted as a filled circle, with its color reflecting its corresponding attention weight. The reference point is represented by a green cross marker. The predicted bounding boxes whose confidence scores are over threshold value have been shown with category-specific color. (a)–(c) Index of X-ray image examples too.

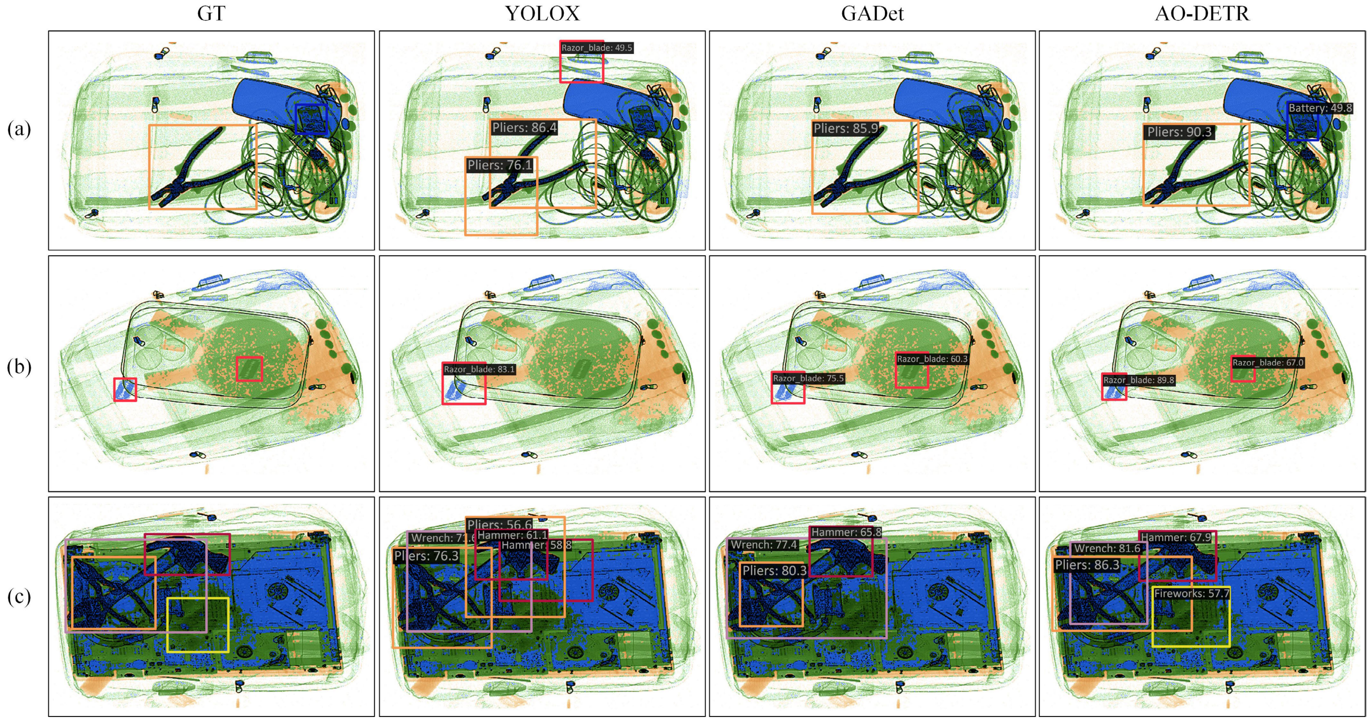


Fig. 8. Visualization comparison of detection results from SOTA models. In comparison with the SOTA general detector YOLOX and the SOTA prohibited item detector GADet, our AO-DETR achieves the best detection performance. (a)–(c) Index of X-ray image examples too.

reducing the model's false negative and false positive rates. "DINO+CSA" can detect Pliers and Razor blades in the column (a), pressure vessels in the column (c), and fireworks in the column (d), which cannot detected by the baseline model DINO. The LFD strategy improves the perception of the edges of contraband items, and "DINO+LFD" achieves more accurate localization results than DINO alone, as seen with the Fireworks in the column (b) and the lighter in the column (c). Moreover, the CSA and LFD strategies do not interfere with each other. AO-DETR demonstrates the best detection performance, avoiding false negatives and positives while providing more accurate localization.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

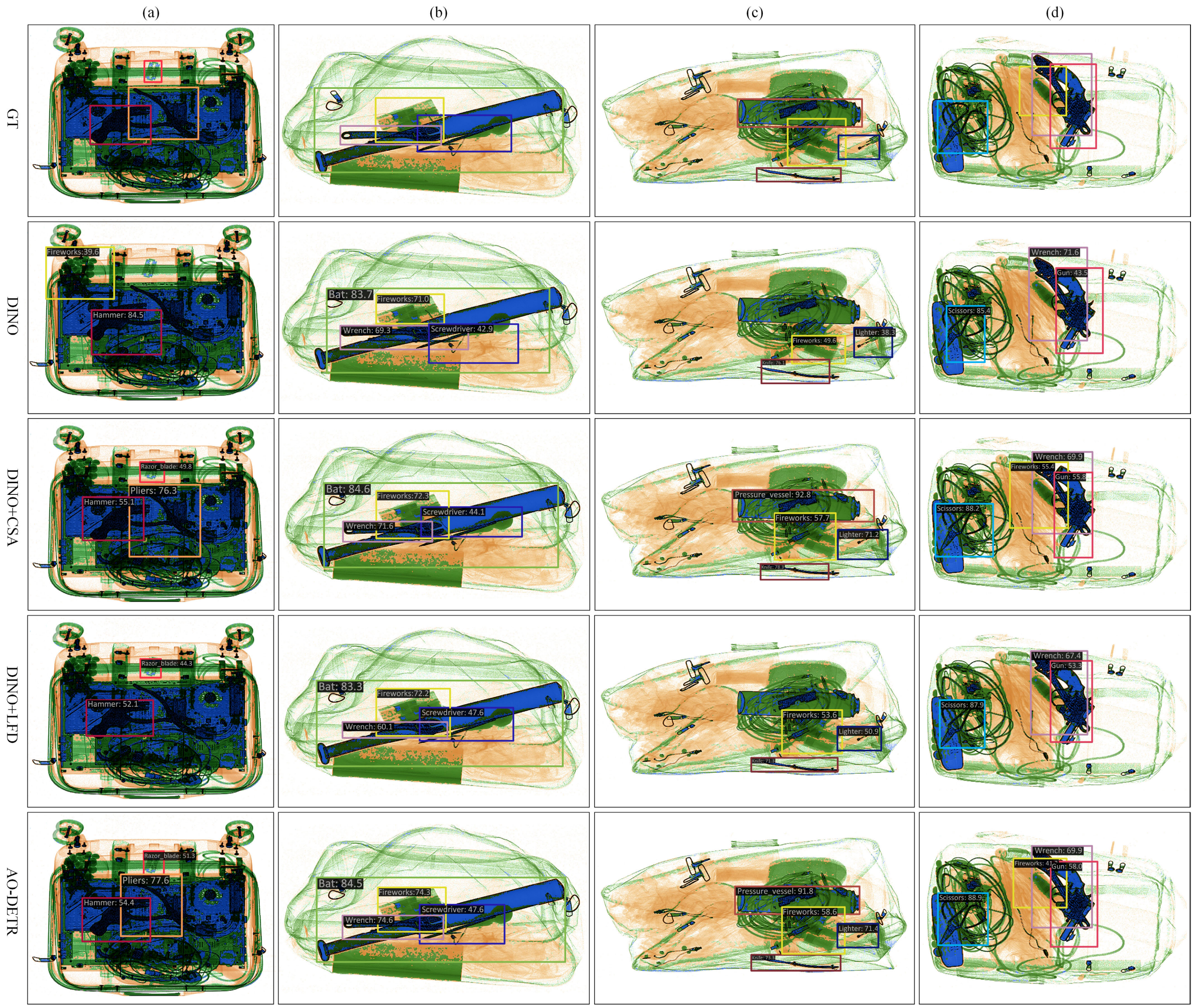


Fig. 9. Detection results visualization analysis for "CSA" and "LFD" on the PIXray [3] dataset. The row "GT" represents four typical X-ray prohibited item images with overlapping phenomenon, each with annotated ground truth boxes. The rows "DINO," "DINO+CSA," "DINO+LFD," and "AO-DETR" correspond to their respective detection results. To achieve optimal display effectiveness, we have standardized the color and category relationships between ground-truth boxes and predicted boxes. For instance, yellow boxes denote fireworks, while bright red boxes signify razor blades. (a)–(d) Index of X-ray image example.

V. Conclusion

In this article, we first conduct an in-depth analysis of the two major challenges in the field of X-ray image prohibited item detection. Subsequently, we explore how to enhance general object detectors based on the characteristics of X-ray images. Overall, we improve the SOTA DETR-like model in the general object detection domain, DINO, and introduce the AO-DETR series models. Specifically, we propose the CSA strategy to enhance the anti-overlapping feature extraction capability for specific category foregrounds by constraining the object classes assigned to category-specific queries during the training phase. Furthermore, by employing the proposed LFD scheme, we enhance the accuracy of reference boxes predicted by mid-level and high-level decoder layers through dense gradient transmission, ultimately improving the ability to perceive blurry edges of models. Extensive experiments on the PIXray, OPIXray, and HIXray datasets demonstrate that our two novel methods can significantly enhance detection performance. In addition, our AO-DETR series models outperform SOTA detectors for various requirements.

References

[1] C. Miao et al., "SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2119–2128.

[2] 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. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 138–146.

[3] B. Ma, T. Jia, M. Su, X. Jia, D. Chen, and Y. Zhang, "Automated segmentation of prohibited items in X-ray baggage images using dense de-overlap attention snake," *IEEE Trans. Multimedia*, vol. 25, pp. 4374–4386, 2022.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[4] A. Chang, Y. Zhang, S. Zhang, L. Zhong, and L. Zhang, "Detecting prohibited objects with physical size constraint from cluttered X-ray baggage images," *Knowl.-Based Syst.*, vol. 237, Feb. 2022, Art. no. 107916.

[5] L. Zhang, L. Jiang, R. Ji, and H. Fan, "PIDray: A large-scale X-ray benchmark for real-world prohibited item detection," 2022, *arXiv:2211.10763*.

[6] R. Tao, T. Wang, Z. Wu, C. Liu, A. Liu, and X. Liu, "Few-shot X-ray prohibited item detection: A benchmark and weak-feature enhancement network," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2012–2020.

[7] F. Shao, J. Liu, P. Wu, Z. Yang, and Z. Wu, "Exploiting foreground and background separation for prohibited item detection in overlapping X-ray images," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108261.

[8] M. Li, B. Ma, T. Jia, and Y. Zhang, "PIXDet: Prohibited items X-ray image detection in complex background," in *Proc. CECNet*, 2022, pp. 81–90.

[9] M. Li, B. Ma, H. Wang, Y. Li, D. Chen, and T. Jia, "PID-YOLOX: An X-ray prohibited items detector based on YOLOX," in *Proc. IEEE 13th Int. Conf. CYBER Technol. Autom., Control, Intell. Syst. (CYBER)*, Jul. 2023, pp. 413–418.

[10] M. Li, B. Ma, H. Wang, D. Chen, and T. Jia, "GADet: A geometry-aware X-ray prohibited items detector," *IEEE Sensors J.*, vol. 24, no. 2, pp. 1665–1678, Jan. 2024.

[11] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–16. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke

[12] H. Zhang et al., "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[13] R. Tao et al., "Towards real-world X-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10923–10932.

[14] 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.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[15] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[16] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.

[17] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[18] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 213–229.

[20] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, 2006, pp. 850–855.

[21] S. Liu et al., "DAB-DETR: Dynamic anchor boxes are better queries for DETR," 2022, *arXiv:2201.12329*.

[22] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 13619–13627.

[23] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Jun. 2023, pp. 6748–6758.

[24] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[25] W. Liu et al., "Ssd: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*. Amsterdam, The Netherlands: Springer, 2016, pp. 21–37.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[27] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, and J. Shi, "FoveaBox: Beyound anchor-based object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 7389–7398, 2020.

[28] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 303–312.

[29] K. Kim and H. S. Lee, "Probabilistic anchor assignment with iou prediction for object detection," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 355–371.

[30] D. Jia et al., "DETRs with hybrid matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19702–19712.

[31] H. Li et al., "Boosting low-data instance segmentation by unsupervised pre-training with saliency prompt," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15485–15494.

[32] Y. Liu, X. Dong, D. Zhang, and S. Xu, "Deep unsupervised part-whole relational visual saliency," *Neurocomputing*, vol. 563, Jan. 2024, Art. no. 126916.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[34] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Mark Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.

[35] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.

[36] Z. Li, L. Yang, and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.

[37] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[38] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[39] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.

[40] R. Cong et al., "Densely nested top-down flows for salient object detection," *Sci. China Inf. Sci.*, vol. 65, no. 8, 2022, Art. no. 182103.

[41] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[42] M. Hofinger, S. R. Bulo, L. Porzi, A. Knapitsch, T. Pock, and P. Kontschieder, "Improving optical flow on a pyramid level," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 770–786.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[44] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[45] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf. Comput. Vis. (ECCV)*. Zurich, Switzerland: Springer, 2014, pp. 740–755.

[46] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[48] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1492–1500.

[49] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[50] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, Dec. 2020, pp. 21002–21012.

[51] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[52] M. Li, T. Jia, H. Lu, B. Ma, H. Wang, and D. Chen, "MMCL: Boosting deformable DETR-based detectors with multi-class min-margin contrastive learning for superior prohibited item detection," 2024, *arXiv:2406.03176*.

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

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[54] M. Wang, H. Du, W. Mei, S. Wang, and D. Yuan, “Material-aware cross-channel interaction attention (MCIA) for occluded prohibited item detection,” *Vis. Comput.*, vol. 39, no. 7, pp. 2865–2877, Jul. 2023.

[55] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, “Occluded prohibited object detection in X-ray images with global context-aware multi-scale feature aggregation,” *Neurocomputing*, vol. 519, pp. 1–16, Jan. 2023.

[56] B. Wang, H. Ding, and C. Chen, “AC-YOLOv4: An object detection model incorporating attention mechanism and atrous convolution for contraband detection in X-ray images,” *Multimedia Tools Appl.*, vol. 83, no. 9, pp. 26485–26504, Aug. 2023.

[57] Z. Zhu, Y. Zhu, H. Wang, N. Wang, J. Ye, and X. Ling, “FDTNet: Enhancing frequency-aware representation for prohibited object detection from X-ray images via dual-stream transformers,” *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108076.

[58] K. Chen et al., “MMDetection: Open MMLab detection toolbox and benchmark,” 2019, *arXiv:1906.07155*.

Mingyuan Li received the B.E. degree from Northeastern University, Shenyang, China, in 2021, where he is currently pursuing the Ph.D. degree with the College of Information Science and Engineering.

His research interests include deep learning, object detection, and prohibited item detection in X-ray images.

Tong Jia received the B.E. degree in computer science and the Ph.D. degree in pattern identification and intelligent system from Northeastern University, Shenyang, China, in 1998 and 2008, respectively.

From 2012 to 2013, he was an International Visiting Scholar with the Department of Electronic Engineering, Michigan State University, East Lansing, MI, USA. He is currently a Professor with the College of Information Science and Engineering, Northeastern University. His research interests include computer/machine vision, image processing, and pattern identification.

Hao Wang received the Ph.D. degree from the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, in 2022.

He is currently a Lecturer with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests include object detection, object segmentation, and related problems.

Bowen Ma received the B.E. and M.S. degrees from Northeastern University, Shenyang, China, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the College of Information Science and Engineering.

His research interests include computer vision, deep learning, and image processing for X-ray testing.

Hui Lu received the B.E. degree from Northeastern University, Shenyang, China, in 2023, where she is currently pursuing the M.S. degree with the College of Information Science and Engineering.

Her research interests include deep learning, object detection, and prohibited item detection in X-ray images.

Shuyang Lin received the B.E. degree from Northeastern University, Shenyang, China, in 2022, where she is currently pursuing the Ph.D. degree with the College of Information Science and Engineering.

Her research interests include deep learning, X-ray prohibited item detection, open-vocabulary object detection, and related problems.

Da Cai received the M.S. degrees from Northeastern University, Shenyang, China, in 2021, where he is currently pursuing the Ph.D. degree with the College of Information Science and Engineering.

His research interests include computer vision, deep learning, and image processing for X-ray testing.

Dongyue Chen received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Fudan University, Shanghai, China, in 2002 and 2007, respectively.

From 2014 to 2015, he was an International Visiting Scholar with the Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. He is currently a Professor with the College of Information Science and Engineering, Northeastern University, Shenyang, China. His research interests include biologically motivated visual modeling, computer vision, and image processing.

Authorized licensed use limited to: Northeastern University. Downloaded on December 09,2024 at 03:02:51 UTC from IEEE Xplore. Restrictions apply.