Dual-view X-ray Detection: Can AI Detect Prohibited Items from Dual-view X-ray Images like Humans?

Renshuai Tao[1], Haoyu Wang[1], Yuzhe Guo[1], Hairong Chen[1], Li Zhang[3,4]
Xianglong Liu[2], Yunchao Wei[1*], Yao Zhao[1]
[1]Visual Intellgence +X International Cooperation Joint Laboratory of MOE, Beijing Jiaotong University
[2]State Key Laboratory of Complex & Critical Software Environment, Beihang University
[3]Department of Engineering Physics, Tsinghua University
[4]National Engineering Laboratory for Dangerous Articles and Explosives Detection Technologies
{rstao, wanghy23, yuzhe.guo, hairong.chen, yunchao.wei, yzhao}@bjtu.edu.cn
zli@mail.tsinghua.edu.cn, xlliu@buaa.edu.cn

Abstract

*To detect prohibited items in challenging categories, human inspectors typically rely on images from two distinct views (vertical and side). Can AI detect prohibited items from dual-view X-ray images in the same way humans do? Existing X-ray datasets often suffer from limitations, such as single-view imaging or insufficient sample diversity. To address these gaps, we introduce the Large-scale Dual-view X-ray (LDXray), which consists of 353,646 instances across 12 categories, providing a diverse and comprehensive resource for training and evaluating models. To emulate human intelligence in dual-view detection, we propose the Auxiliary-view Enhanced Network (AENet), a novel detection framework that leverages both the main and auxiliary views of the same object. The main-view pipeline focuses on detecting common categories, while the auxiliary-view pipeline handles more challenging categories using "expert models" learned from the main view. Extensive experiments on the LDXray dataset demonstrate that the dual-view mechanism significantly enhances detection performance, e.g., achieving improvements of up to **+24.7%** for the challenging category of umbrellas. Furthermore, our results show that AENet exhibits strong generalization across seven different detection models for X-ray Inspection*[1].

1. Introduction

Security inspectors typically use X-ray scanners to detect prohibited items hidden in luggage, a process that is both time-consuming and labor-intensive. The development of deep learning technologies[9, 10, 27, 28], particularly in computer vision[3, 12, 18, 36, 40, 44], is expected to reduce the workload through automatic prohibited items detection.

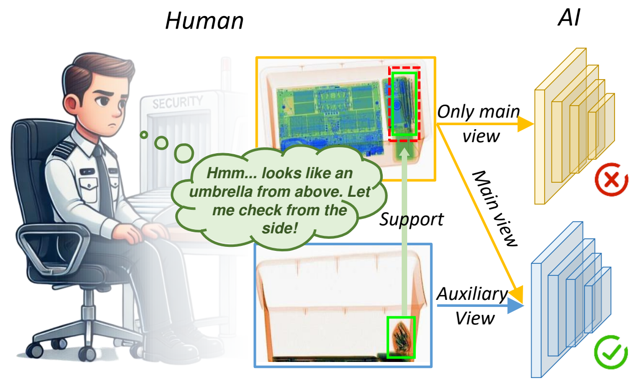


Figure 1. Illustration of the dual-view X-ray detection task. This paper aims to develop AI capable of reasoning like human experts when detecting prohibited items, leveraging insights from dual-view X-ray images to improve accuracy and interpretative ability.

In recent years, detecting prohibited items in X-ray security scenes has become a research hotspot, leading to the development of numerous datasets and detection methods[22, 24, 25, 33, 34, 38]. Datasets include a considerable number of categories and instances, offering a comprehensive resource for training and evaluating models[5, 8, 13, 15–17, 20, 29, 30]. These efforts have attracted the attention of researchers in this field, resulting in significant advancements and increased collaboration.

In real-world scenarios, human inspectors typically rely on images from two distinct views (vertical and side) to effectively detect prohibited items, particularly those in challenging categories. This dual-view approach is crucial because single-view images often fail to capture vital information that can be seen from the other perspective. The limitation of single-view images becomes especially significant

*Corresponding author.
[1]Webpage: https://github.com/rstao-bjtu/LDXray/.

when objects are randomly placed in luggage, leading to heavy occlusions and making it difficult to identify hidden items. In such cases, relying solely on a single view can result in missed detections. The question then arises: **Can AI detect prohibited items from dual-view X-ray images in the same way humans do, leveraging both views for enhanced detection?** Achieving this goal requires not only a large-scale evaluation dataset but also the development and deployment of sophisticated, professional models that can effectively process and interpret dual-view data.

Despite these challenges, as illustrated in Table 1, existing X-ray datasets often suffer from the limitations of single-view imaging or insufficient sample diversity. To address these gaps, in this paper, we introduce a Large-scale Dual-view X-ray (LDXray) dataset for prohibite items detection task, sourced directly from real-world security environments. LDXray contains 146,997 paired images, making it approximately 20 times larger than similar existing X-ray datasets, and represents 353,646 instances across 12 categories. We employed professional human security inspectors to ensure high annotation accuracy. We anonymized all personal information, such as time and location, to prevent linkage of the luggage to owners, guaranteeing data privacy.

To further improve detection performance, we propose a novel framework called the Auxiliary-view Enhanced Network (AENet). AENet is designed to leverage both main and auxiliary views of the same object, improving object detection in X-ray security inspection. The AENet framework consists of two distinct detection pipelines: (1) The main view pipeline is responsible for detecting simple categories, using a universal model applicable to all standard categories. (2) The auxiliary view pipeline utilizes saliency detection techniques to identify challenging objects and then hands over these candidates to specialized "expert models" for further analysis and confirmation. The outputs from both pipelines are combined to produce the final detection results. **The proposed dual-pipeline approach enhances the detection of challenging categories while maintaining stable performance for other simpler categories**.

We conduct extensive experiments on the proposed LDXray dataset and compare the performance of various widely-used detection models. Our results demonstrate that the dual-view mechanism significantly enhances detection accuracy, especially for challenging categories, *e.g.*, improving from 33.9% to 58.6% (**+24.7%**) for the umbrella category. Additionally, AENet shows strong generalization across different base models, with seven detection models implemented. We hope that future research will further explore the unique characteristics of dual-view data, leading to the development of more intelligent security inspection solutions. In summary, the key contributions are as follows:

- We introduce the Large-scale Dual-view X-ray (LDXray) dataset, a significant contribution to the field of X-ray security inspection. LDXray consists of a diverse and comprehensive set of dual-view X-ray images, specifically designed for the detection of prohibited items in luggage.
- We propose a novel detection framework, AENet, that leverages both the main and auxiliary views of the same object to significantly enhance detection, especially for challenging categories. The core idea is to combine the advantages of both main and auxiliary view.
- We conduct extensive experiments on the newly introduced LDXray dataset and compare our method with popular methods. Experimental results clearly show that this dual-view mechanism provides a significant performance improvement in detecting the challenging categories.

The remainder of the paper is organized as follows: Section 2 reviews related work on X-ray prohibited items detection task. Section 3 provides an overview of the proposed LDXray dataset. Section 4 details the proposed AENet method. In Section 5, we present the experimental setup and results. Finally, Section 6 offers concluding remarks.

2. Related Work

The development of X-ray image analysis for security and inspection has been greatly supported by a range of specialized datasets. Foundational datasets like GDXray [24] and SIXray [25] initially established resources for detection,

Dataset	2-view	Open	N_c	N_p	N_a	Task	Year
GDXray [24]	✗	✓	3	8,150	8,150	Detection	2015
SIXray [25]	✗	✓	6	8,929	8,929	Cls & Det	2019
OPIXray [38]	✗	✓	5	8,885	8,885	Detection	2020
PIDray [37]	✗	✓	12	47,677	47,677	Det & Seg	2021
HiXray [33]	✗	✓	8	45,364	102,928	Detection	2021
PIXray [21]	✗	✓	15	5,046	15,201	Det & Seg	2022
CLCXray [43]	✗	✓	12	9,565	9,565	Detection	2022
EDS [34]	✗	✓	10	14,219	31,654	Detection	2022
FSOD [35]	✗	✓	20	12,333	41,704	Detection	2022
LPIXray [11]	✗	✗	18	60,950	84,785	Detection	2023
MV-Xray [31]	✓	✗	2	3,231	3,231	Detection	2019
deei6 [1]	✓	✗	6	7,022	7,022	Det & Seg	2021
DBF [14]	✓	✗	4	2,528	3,034	Detection	2021
Dualray [39]	✓	✗	6	4,371	4,371	Detection	2022
DvXray [22]	✓	✓	15	5,000	5,496	Classification	2024
LDXray (**ours**)	✓	✓	12	**146,997**	**353,646**	Detection	2024

Table 1. A comprehensive overview of various X-ray datasets utilized in security inspection. The number of categories (N_c), images containing prohibited items (N_p), and annotations (N_a) associated are displayed. Additionally, we summary the task, year, and availability of each X-ray dataset for a more complete comparison.

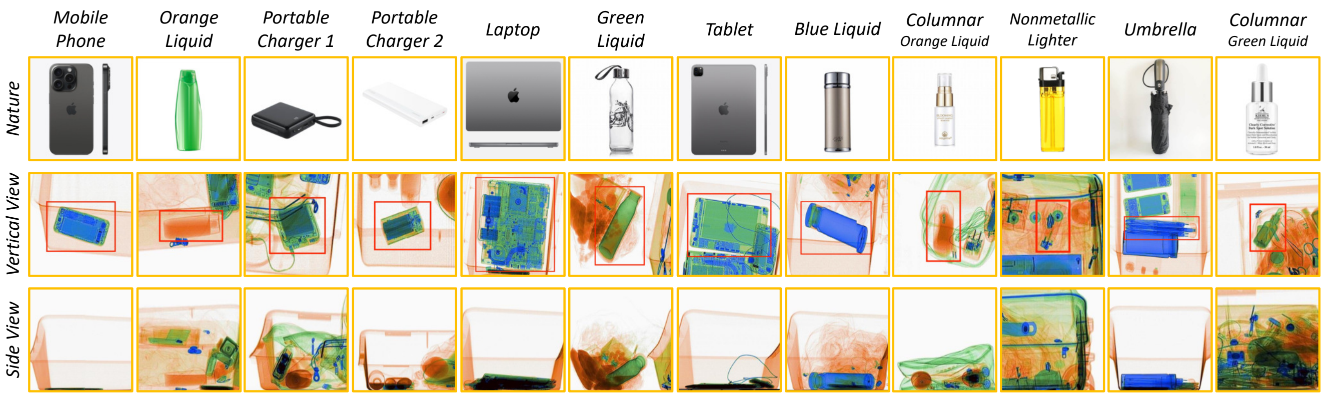


Figure 2. Various categories of objects and corresponding X-ray images from two different views in LDXray. The top row displays natural images of the objects. The middle row shows the X-ray images taken from the vertical view, while the bottom row shows the side view.

while more recent datasets such as PIDray [37], HiXray [33], and PIXray [21] have expanded task diversity with comprehensive annotations. Dual-view datasets, including MV-Xray [31] and Dualray [39], introduced multi-view data, enhancing depth in scene analysis. Our proposed dataset, LDXray, offers the largest annotated image set to date, supporting both dual-view and open-access needs for data diversity and task applicability.

3. LDXray Dataset

We introduce LDXray, the first large-scale dual-view X-ray dataset designed for prohibited items detection in X-ray images. In this section, we present a comprehensive overview of the construction details and statistics of LDXray dataset.

3.1. Construction Details

Data Collection. Due to significant differences between the distribution of artificial and real-world scenarios, such as the appearance frequency of certain categories and their co-occurrence with specific items, we collected the images from actual transportation hubs (to protect privacy, hubs are anonymous). When packages undergo inspections, the X-ray machine generates and stores corresponding X-ray images. We collect a large number of images, subsequently removing low-quality and invalid samples. The long-tail distribution accurately matched the real-world scenario.

Privacy Protect. We conduct a two-step process to safeguard privacy. First, during the period of collection, an obvious indicator was hung to inform the passengers that the dataset may be used for public academical research after personal information removed. **And security inspectors also asked passengers for their explicit consent.** If consent was not given, the X-ray image was deleted. Additionally, we have de-identified the collected data by removing details such as time and location that could potentially identify an individual. This ensures that the data does not contain Personally Identifiable Information (PII) [23].

Extensive Diversity. As illustrated in Figure 2, LDXray contains a wide range of prohibited items, categorized into 12 different classes, including "Mobile Phone", "Orange Liquid", "Portable Charger 1 (lithium-ion prismatic cell)", "Portable Charger 2 (lithium-ion cylindrical cell)", "Laptop", "Green Liquid", "Tablet", "Blue Liquid", "Columnar Orange Liquid", "Nonmetallic Lighter", "Umbrella" and "Columnar Green liquid" ("MP", "OL", "PC1", "PC2", "LA", "GL", "TA", "BL", "CO", "NL", "UM" and "CG", for short). This diversity mirrors the variety of items that security inspectors encounter during inspections. We offer a comprehensive dataset that facilitates the development of detection models, to manage diverse scenarios effectively.

Annotation by Professional Human X-ray Security Inspectors. Recognizing objects in X-ray images necessitates specialized training and expertise. The LDXray dataset is manually annotated at the bounding-box level by professional human X-ray security inspectors with extensive daily experience. To guarantee high-quality annotation, we followed the similar quality control procedure of annotation as the famous Pascal VOC [7]. Specifically, all X-ray security inspectors received detailed annotation guidelines that explained what to annotate, how to draw bounding boxes, and how to handle occlusions. Subsequently, another inspector verified the accuracy of each annotation, checking for omitted objects and ensuring comprehensive labeling.

3.2. Data Statistics

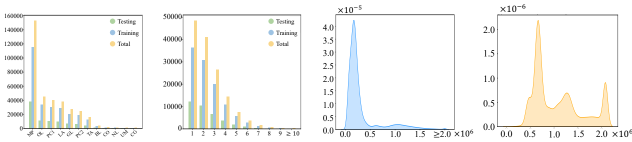
Image Quality. All images of LDXray are stored with an average resolution of 1200x900 pixels, while the maximum resolution of some samples reaches 2000x1040 pixels. As illustrated in Figure 4, the image quality of LDXray is superior, attributed to more advanced imaging techniques and higher resolution. These results demonstrate that LDXray dataset significantly outperforms those popular datasets, enabling a more realistic evaluation for X-ray detection.

Category Distribution. LDXray contains 146,997 X-ray image pairs, with 353,646 instances of 12 common categories. The category names and their abbreviations are depicted in "Construction Details" section. The dataset is partitioned into a training set and a testing set, where the ratio is about 3:1. The statistics of category distribution of the training set and testing set are also shown in Table 2 and Figure 3a, providing a understanding of scale and diversity.

Other Distributions. Several other distributions need to

Category	MP	OL	PC1	PC2	LA	GL	TA	BL	CO	NL	UM	CG	Total
Training	115,016	33,691	30,222	18,471	28,832	20,199	12,094	3,138	1,357	981	683	647	265,331
Testing	37,993	11,357	10,049	6,200	9,484	6,856	4,083	1,079	445	347	196	226	88,315
Total	153,009	45,048	40,271	24,671	38,316	27,055	16,177	4,217	1,802	1,328	879	873	353,646

Table 2. The number of instances in the training and testing sets.



(a) (b) (c) (d)

Figure 3. Various distributions. (a) Instance per Category Distribution. (b) Instance per Image Distribution. (c) Instance Area Density Distribution. (d) Image Area Density Distribution.

be introduced to further elucidate the characteristics of the LDXray dataset: (1) **Instance per Image Distribution.** As illustrated in Figure 3b, the average number of instances per image in the LDXray dataset is 2.27, with most images not exceeding seven instances. (2) **Instance Area Density Distribution.** Figure 3c illustrates the distribution of instance areas in terms of pixel count. The horizontal axis represents the area of bounding boxes (in square pixels), while the vertical axis indicates density. Most instances have an area below 500,000 square pixels, with only a small fraction exceeding 1,500,000 square pixels. (3) **Image Area Density Distribution.** Figure 3d shows the distribution of image areas, measured in pixel count. The horizontal axis corresponds to the total pixel area of images, and the vertical axis represents density. The majority of images have areas ranging from 500,000 to 2,000,000 square pixels.

4. Method

In this section, we elaborate on the details of the proposed **A**uxiliary-view **E**nhanced **N**etwork (AENet). We begin with an analysis of the dual-view detection task setting in Section 4.1. Next, we introduce the framework of AENet in Section 4.2. We then describe the Main View Detection Pipeline and Auxiliary View Detection Pipeline in Sections 4.3 and 4.4, respectively. Finally, in Section 4.5, we provide a detailed illustration of the network training process.

4.1. Task Overview

In the proposed dual-view detection task, the input images consist of a main view and an auxiliary view, denoted as I_1 and I_2. The goal is to predict a set of bounding boxes $B = \{b_1, b_2, \dots, b_n\}$ and corresponding class labels $C = \{c_1, c_2, \dots, c_n\}$, where each bounding box b_i defines the spatial extent of an object in the image, and each class label c_i identifies the object category, similar to the traditional object detection task described above. It is important to note that due to the difficulty for inspectors in annotating the auxiliary view, there is no ground truth available for I_2. Additionally, due to the positional invariance of the imaging devices, a corresponding horizontal coordinate relationship exists between the objects in these two perspective images.

Figure 4. Image quality comparison of current X-ray datasets. The images of LDXray are superior, attributed to more advanced imaging techniques, enhanced color vibrancy, and higher resolution.

$$b_i^*(x_1, x_2)(I_1) = \lambda b_i^*(x_1, x_2)(I_2) \tag{1}$$

$$c_i^*(I_1) = c_i^*(I_2) \tag{2}$$

where $b_i^*(x_1, x_2)$ denotes the abscissa of the ground truth bounding box, and λ is a constant transformation factor.

The objective of the proposed dual-view object detection task in this paper can be formally expressed as follows:

$$(B, C) = \mathbf{M}(I_1, B_{I_1}^*, C_{I_1}^*, I_2, B_{I_2}^*(x_1, x_2), C_{I_2}^*) \tag{3}$$

where $\mathbf{M}$ represents the detection model, which is designed to take as input a pair of views, namely I_1 and I_2, along with their associated ground-truth bounding boxes $B_{I_1}^*$, $B_{I_2}^*(x_1, x_2)$, and class labels $C_{I_1}^*$, $C_{I_2}^*$. The primary output of this model comprises two components: a set of predicted bounding boxes $B = \{b_1, b_2, \dots, b_n\}$ and their corresponding class labels $C = \{c_1, c_2, \dots, c_n\}$, both of which are associated with the main-view image I_1.

4.2. Framework

Figure 5 provides an overview of the proposed framework. The architecture of AENet consists of two distinct pipelines: the main view and the auxiliary view. In the main view pipeline, the training process is conducted independently of the auxiliary view, resembling the traditional object detection model paradigm. Moreover, we employ a specialized approach that involves cropping individual instances from the main view image using the ground-truth boxes. These expert models, which are trained to specialize in detecting "challenging categories", are stored for subsequent use in the auxiliary view detection pipeline. In the auxiliary view detection pipeline, we utilize saliency detection techniques to generate approximate location predictions for objects in the auxiliary view image. These location predictions, in combination with the category labels obtained from the main view pipeline, are used to train the auxiliary view detector. Then according to the location relation, we crop the essential instance of the main view image. The cropped image patch is then passed through the expert models trained earlier of main view to refine the prediction.

Finally, we integrate the detection results generated by both the main view and auxiliary view pipelines. By combining the outputs from both branches, AENet is able to deliver enhanced detection performance, with improved accuracy and robustness, particularly in handling occlusions, complex backgrounds, and difficult object categories.

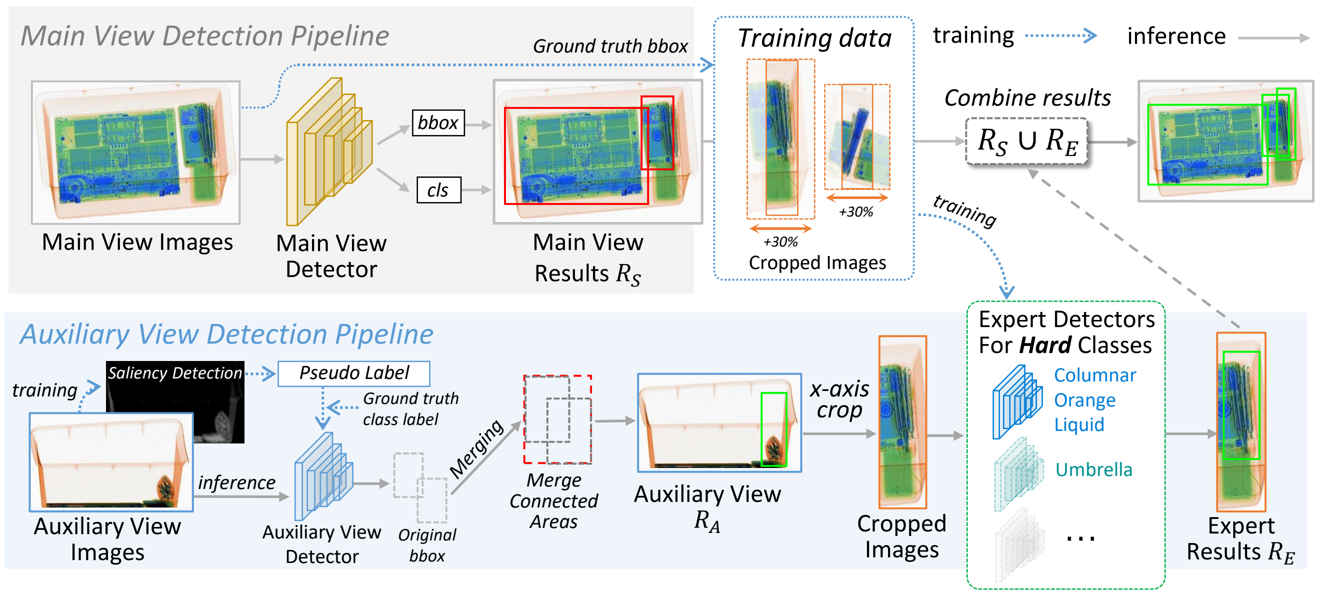


Figure 5. Overview of the AENet framework. This architecture consists of two distinct pipelines: the main view, which follows a traditional object detection paradigm, and the auxiliary view, which uses saliency detection and cross-view correspondence for localization.

4.3. Main View Detection Pipeline

This pipeline focuses on leveraging the traditional object detection paradigm, utilizing the main view image I_1 for training and detection. The primary objective of this pipeline is to predict bounding boxes $B = \{b_1, b_2, \ldots, b_n\}$ and class labels $C = \{c_1, c_2, \ldots, c_n\}$ for the objects present in I_1. The feature extraction backbone F_{main} processes the main view image I_1 to generate feature maps, $F_{main}(I_1) \rightarrow M_{main}$, where M_{main} represents the feature maps extracted from I_1. The detection heads, denoted by $\mathcal{D}_{main}$, then predict the bounding boxes and class labels:

$$(B, C) = \mathcal{D}_{main}(M_{main}) \tag{4}$$

The objective loss function $\mathcal{L}_{main}$ to minimize, consists of a classification loss L_{cls} and a regression loss L_{reg}:

$$\mathcal{L}_{main} = \sum_{i=1}^{n} L_{cls}(c_i, c_i^*) + \sum_{i=1}^{n} L_{reg}(b_i, b_i^*) \tag{5}$$

where c_i^* and b_i^* represent the ground-truth class label and bounding box for the annotated prohibited item object i.

To handle "hard categories", we employ a specialized cropping strategy. During training, individual object instances are cropped from I_1 using the ground-truth bounding boxes $B_{I_1}^* = \{b_1^*, b_2^*, \ldots, b_n^*\}$. The cropped image patches, denoted by $\{I_1^{crop}(b_1^*), I_1^{crop}(b_2^*), \ldots, I_1^{crop}(b_n^*)\}$, are passed through expert models $\{E_1, E_2, \ldots, E_m\}$, each trained specifically to detect "challenging categories":

$$\hat{c}_i, \hat{b}_i = E_j(I_1^{crop}(b_i^*)), \quad \text{for } i \in [1, n], j \in [1, m] \tag{6}$$

The expert models output refined predictions, denoted by $\hat{B} = \{\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_n\}$ and $\hat{C} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n\}$ and the expert models will later be utilized during inference of the auxiliary view pipeline for more accurate detection.

4.4. Auxiliary View Detection Pipeline

Since the auxiliary view lacks direct ground-truth labels, we adopt an indirect training approach based on saliency detection and the cross-view correspondence. First, a saliency detection function $\mathcal{S}$ is applied to the auxiliary view image to obtain approximate object locations:

$$B_{aux} = \mathcal{S}(I_2) = \{b_{aux_1}, b_{aux_2}, \ldots, b_{aux_m}\} \tag{7}$$

These bounding boxes are assigned class labels transferred from the main view using a correspondence relationship. Specifically, each auxiliary bounding box b_{aux_i} is matched to its corresponding main view bounding box b_{main_i}, and the class labels are assigned accordingly:

$$c_{aux_i} = c_{main_i}, \text{for corresponding } b_{aux_i} \text{ and } b_{main_i} \tag{8}$$

The auxiliary detector, denoted by $\mathcal{D}_{aux}$, then learns to predict bounding boxes and labels from the auxiliary view feature maps M_{aux} extracted by a feature extractor F_{aux}:

$$M_{aux} = F_{aux}(I_2), (\hat{B}_{aux}, \hat{C}_{aux}) = \mathcal{D}_{aux}(M_{aux}) \tag{9}$$

The training objective for the auxiliary view is to minimize the loss $\mathcal{L}_{aux}$, which consists of a saliency-based regression loss L_{sal} and a classification loss L_{cls}:

$$\mathcal{L}_{aux} = \sum_{i=1}^{m} L_{sal}(b_{aux_i}, \hat{b}_{aux_i}) + \sum_{i=1}^{m} L_{cls}(c_{aux_i}, \hat{c}_{aux_i}) \tag{10}$$

During inference, the auxiliary view bounding boxes may overlap. A merging strategy is employed to combine overlapping boxes, leading to more accurate predictions:

$$B_{aux}^{merged} = \mathbf{Merge}(\hat{B}_{aux}) \tag{11}$$

where the function $\mathbf{Merge}(\cdot)$ represents an algorithm (such as non-maximum suppression or a weighted average technique) for merging overlapping prediction bounding boxes.

The merged bounding boxes B_{aux}^{merged} and their class labels $\hat{C}_{aux}$ are then used to crop the corresponding regions from the main-view image I_1 (due to positional invariance of the imaging devices, a corresponding horizontal coordinate relationship exists between objects in these two perspective images). The cropped instances are passed through the corresponding expert models to refine the detection:

$$I_1^{crop}(b_{aux_i}^{merged}) \quad \text{for } b_{aux_i}^{merged} \in B_{aux}^{merged} \tag{12}$$

$$(\tilde{c}_i, \tilde{b}_i) = E_j(I_1^{crop}(b_{aux_i}^{merged})), \text{for } i, j \in [1, m] \tag{13}$$

The final set of bounding boxes B_{final} is obtained by combining the bounding boxes from the main view (B) and the auxiliary view $\{\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_m\}$. The final set of class labels C_{final} is obtained by combining the class labels from the main view (C) and the auxiliary view $\{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_m\}$:

$$B_u, C_u = B \cup \{\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_m\}, C \cup \{\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_m\} \tag{14}$$

Thus, the final detection results (B_{final}, C_{final}) represent the complete set of bounding boxes and labels by combining the contributions from both pipelines, enhancing detection quality by leveraging complementary information.

Algorithm 1: Overall Procedure of AENet

Input: Main view image I_1, auxiliary view image I_2, ground-truth boxes $B_{I_1}^*$, class labels $C_{I_1}^*$.
Output: Trained detection model $\mathbf{M}$, final bounding boxes B_u, class labels C_u.
Main View Training:
Extract feature maps M_{main} from I_1;
Predict (B, C) using $\mathcal{D}_{main}$ based on M_{main};
Compute loss $\mathcal{L}_{main}$ based on Eq. (5).
for *each* $b_i^* \in B_{I_1}^*$ **do**
  Crop $I_1^{crop}(b_i^*)$;
  Train expert models E_j for hard categories.
Auxiliary View Training:
Obtain $B_{aux} \leftarrow \mathcal{S}(I_2)$;
Assign class labels $c_{aux_i} = c_{main_i}$ for related boxes.
Extract $M_{aux} \leftarrow F_{aux}(I_2)$;
Predict $(\hat{B}_{aux}, \hat{C}_{aux})$ using $\mathcal{D}_{aux}$;
Compute loss $\mathcal{L}_{aux}$ based on Eq. (10).
Inference and Refinement:
Merge boxes: $B_{aux}^{merged} = \textbf{Merge}(\hat{B}_{aux})$.
for *each* $b_{aux_i}^{merged} \in B_{aux}^{merged}$ **do**
  Crop $I_1^{crop}(b_{aux_i}^{merged})$;
  Refine $(\tilde{c}_i, \tilde{b}_i) = E_j(I_1^{crop}(b_{aux_i}^{merged}))$.
Combine final results: $(B_u, C_u) = (B \cup \tilde{B}, C \cup \tilde{C})$.
Obtain trained model $\mathbf{M}$ and final outputs (B_u, C_u).

4.5. Overall Procedure

Algorithm 1 summarizes the whole process. The overall training and inference process of AENet involves training both main and auxiliary view pipelines independently and then combining their results. During main view training, I_1 is used to predict bounding boxes B and class labels C, with hard categories refined through expert models. Auxiliary view training utilizes saliency detection on I_2 to estimate object locations, which are labeled based on correspondence with the main view. During inference, auxiliary bounding boxes are refined, merged, and used to crop instances from I_1 for further refinement via expert models. Finally, outputs from both views are unified to generate the final bounding boxes and labels (B_u, C_u), ensuring improved accuracy and robustness by leveraging complementary information from both perspectives.

5. Experiments

In this section, we present a comprehensive evaluation of the proposed AENet, utilizing the LDXray dataset. We begin by detailing the experimental setup in Section 5.1, followed by an investigation of the integration of AENet with various baseline detection models in Section 5.1.1. Subsequently, Section 5.2 presents four groups of ablation studies designed to isolate and assess the impact of key components of the system. These studies include: (1) An analysis of the integration of different expert models for detecting cropped images in the main-view pipeline (Section 5.2.1). (2) The effect of different approximate location techniques used in the auxiliary-view pipeline (Section 5.2.2). (3) The impact of various detection models as auxiliary view detectors in the auxiliary-view pipeline (Section 5.2.3). (4) A study of how varying confidence score thresholds influence the results in the auxiliary-view module (Section 5.2.4).

5.1. Experimental Setup

To evaluate the performance of AENet, we use Average Precision (AP) for each category and mean Average Precision (mAP) for overall performance. The AP for each category is calculated using an Intersection over Union (IoU) threshold of 0.5. The mAP is then computed by taking the mean of these AP values across all categories.

All experiments were conducted using the open-source toolbox `mmdetection` [4], ensuring a consistent and reproducible environment. For all the baseline detection models, we used the default configuration settings provided by `mmdetection`, and all models were trained and evaluated on two NVIDIA GeForce RTX 4090 GPUs. The LDXray dataset, used in all our experiments, includes both simple and challenging categories, allowing us to evaluate the robustness and generalizability of the AENet across tasks.

Method	$\mathbf{AP_{50}}$	Simple Categories (**Same Main-view Detector**)								Challenging Categories			
		MP	OL	PC1	PC2	LA	GL	TA	BL	NL*	CO	UM	CG
F-RCNN [26]	67.8	95.8	67.0	96.8	97.0	76.0	96.4	93.9	90.4	30.1	12.6	45.6	12.2
+ours	**69.7**$_{\uparrow 1.9\%}$	95.8	67.0	96.8	97.0	76.0	96.4	93.9	90.4	30.1	**13.3**$_{\uparrow 0.7\%}$	**65.7**$_{\uparrow 20.1\%}$	**14.2**$_{\uparrow 2.0\%}$
C-RCNN [2]	66.6	95.6	63.7	96.6	97.3	72.7	96.0	93.0	89.4	21.9	14.1	42.1	16.5
+ours	**68.6**$_{\uparrow 2.0\%}$	95.6	63.7	96.6	97.3	72.7	96.0	93.0	89.4	21.9	**14.5**$_{\uparrow 0.4\%}$	**63.8**$_{\uparrow 21.7\%}$	**18.5**$_{\uparrow 2.0\%}$
S-RCNN [32]	69.6	97.3	65.1	96.4	98.2	75.5	95.7	94.8	90.2	1.0	30.4	61.0	29.2
+ours	**70.4**$_{\uparrow 0.8\%}$	97.3	65.1	96.4	98.2	75.5	95.7	94.8	90.2	1.0	30.4	**67.6**$_{\uparrow 5.6\%}$	**31.7**$_{\uparrow 2.5\%}$
RetinaNet [19]	62.7	96.5	63.9	96.1	97.8	73.4	95.2	94.3	88.4	2.7	7.0	33.9	3.1
+ours	**65.1**$_{\uparrow 2.4\%}$	96.5	63.9	96.1	97.8	73.4	95.2	94.3	88.4	2.7	**9.2**$_{\uparrow 2.2\%}$	**59.0**$_{\uparrow 25.1\%}$	**5.1**$_{\uparrow 2.0\%}$
CenterNet [6]	72.1	97.4	66.0	96.0	98.2	76.0	94.9	95.3	89.8	10.0	39.7	62.2	39.9
+ours	**72.4**$_{\uparrow 0.3\%}$	97.4	66.0	96.0	98.2	76.0	94.9	95.3	89.8	10.0	39.7	**65.4**$_{\uparrow 3.2\%}$	39.9
RepPoints [41]	71.9	97.3	69.8	97.4	98.1	78.3	96.9	96.1	92.0	20.8	31.6	53.6	31.2
+ours	**73.0**$_{\uparrow 1.1\%}$	97.3	69.8	97.4	98.1	78.3	96.9	96.1	92.0	20.8	31.6	**65.8**$_{\uparrow 12.2\%}$	**32.4**$_{\uparrow 1.2\%}$
ATSS [42]	72.8	97.0	67.7	96.9	97.9	76.6	95.7	95.0	89.4	37.7	21.9	60.0	37.4
+ours	**73.4**$_{\uparrow 0.6\%}$	97.0	67.7	96.9	97.9	76.6	95.7	95.0	89.4	37.7	21.9	**67.6**$_{\uparrow 7.6\%}$	37.4

Table 3. Performance of various state-of-the-art detection methods with and without the integration of AENet. The first line reports performance on the main-view data, and **the second line includes the auxiliary view with AENet integrated**. As illustrated, our model improves the detection performance in the challenging categories while maintaining stable performance on the other simple categories. *Non-metallic lighters (NL) represent a unique category, as they almost blend in with the background under X-ray imaging, making them exceptionally difficult to detect. Nevertheless, we have chosen to present these results and look forward to further exploration.

5.1.1 Integration with Various Detection Models

In this experiment, we evaluate the performance of AENet when integrated with various SOTA detection models. The goal is to assess the generalizability of AENet and its ability to enhance different detection architectures. We conduct experiments using several popular detection models, including Faster R-CNN (F-RCNN)[26], Cascade R-CNN (C-RCNN)[2], Sparse R-CNN (S-RCNN)[32], RetinaNet[19], CenterNet [6], RepPoints [41], and ATSS [42].

The results, presented in Table 3, show that the integration of AENet consistently improves the mAP across all detection models, with substantial gains in both simple and challenging categories. For example, when integrated with F-RCNN, AENet boosts the mAP by 1.9%, with significant improvements in challenging categories, such as UM (+20.1%) and CG (+2.0%). Similarly, integration with C-RCNN results in a 2.0% increase in mAP, with even greater improvements in UM (+21.7%) and CG (+2.0%). Even for models like RetinaNet, which show lower performance in challenging categories, AENet results in substantial improvements, particularly in UM (+25.1%) and CG (+2.0%).

In addition to the improvements in challenging categories, AENet also enhances the overall performance of region-based models such as Faster R-CNN and Cascade R-CNN, as well as anchor-free models like CenterNet and ATSS. For example, ATSS achieves a 0.6% improvement in mAP when integrated with AENet, with a 7.6% gain in the challenging UM category. These results demonstrate that AENet is a highly effective and flexible enhancement that can be applied to a wide variety of object detection models, regardless of their underlying architecture.

5.2. Ablation Studies

In this section, we present several ablation studies to analyze the impact of key components of the AENet framework. These studies evaluate the contribution of different techniques used in the auxiliary view module, the integration with various expert detection models, and the effect of varying confidence score thresholds. The results of these experiments are presented in the following sections.

5.2.1 Expert Models in Main-view Pipeline

The results of this experiment are shown in Table 4. The primary goal of this study is to identify the most suitable expert model for detecting objects in cropped images within the main-view pipeline. To accomplish this, we evaluated a range of detection models, including all the base models listed in Table 3, with the aim of determining which one performs best in the context of cropped images. Among the models considered, ATSS emerged as the top performer, achieving a mAP of 51.5%. This result not only outperforms C-RCNN, which achieved a mAP of 50.7%, and F-RCNN with a mAP of 48.9%, but also demonstrates a substantial improvement in accuracy across several challenging categories. For instance, ATSS shows a notable improvement in the CO category (+7.4%), a modest gain in UM (+0.2%), and a significant boost in CG (+5.4%). These improvements are particularly important for real-world applications, where detecting objects in difficult or tail categories is often a critical challenge. Taken together, these findings position ATSS as the effective model for analyzing cropped image patches within the main-view pipeline.

Method	Overall			Categories		
	mAP	AP_{75}	AP_{50}	CO	UM	CG
F-RCNN [26]	48.9	49.4	80.3	76.8	94.6	66.1
C-RCNN [26]	50.7	52.0	81.7	78.0	95.8	67.5
S-RCNN [26]	42.7	42.3	76.0	72.1	93.8	58.2
RetinaNet [26]	34.7	28.4	67.2	70.4	92.8	41.7
CenterNet [26]	45.6	44.1	78.9	69.6	94.1	63.7
RepPoints [26]	31.3	30.6	55.5	64.0	94.4	16.7
ATSS [26]	**51.5**	**55.1**	**83.9**	**84.2**	**94.8**	**72.9**

Table 4. Performance comparison of different expert models trained on the cropped images in the main-view pipeline.

5.2.2 Location Approximation in Auxiliary-view

The results of this experiment are presented in Table 5. In this study, we evaluate the impact of different location approximation techniques used in the auxiliary-view module of AENet. Specifically, we compare the performance of several methods, including the "Saliency" technique for approximating object locations. As shown in the table, the saliency technique, which utilizes a grey map approach, outperforms other methods. By incorporating this technique, the mAP increases from 38.2% to 39.1%, with notable improvements in challenging categories such as CO (+2.2%), UM (+25.1%), and CG (+2.0%). These results demonstrate that the Saliency technique provides more accurate spatial information, significantly enhancing the auxiliary-view module's ability to refine object localization. As a result, the overall detection performance is improved, confirming the effectiveness of the saliency detection for better location approximation.

Threshold	Overall			Categories		
	mAP	AP_{75}	AP_{50}	CO	UM	CG
Single-view	38.2	41.0	62.7	7.0	33.9	3.1
Conversion	38.2	41.1	62.7	7.5	33.9	3.1
Saliency (**ours**)	**39.1**	**41.4**	**65.1**	**9.2**	**59.0**	**5.1**

Table 5. Performance comparison of detection methods with different approximate location techniques in auxiliary-view module.

5.2.3 Auxiliary View Detectors

The results are shown in Table 6. In this experiment, we explore how the choice of detection model as the auxiliary view detector impacts performance. We evaluate several models for this task, assessing how their performance in the auxiliary-view pipeline influences the overall results.

The results show that using a specialized detector in the auxiliary view module can improve the mAP performance. The best results were observed when using the C-RCNN detector, which led to a 26.7% mAP performance, particularly in challenging categories like UM (78.6%) and CG (29.0%). These results underscore the importance of selecting an appropriate detector for the auxiliary view, emphasizing the proposed model's ability to detect challenging categories, which leads to a higher recall of items during detection.

Method	Overall			Categories		
	mAP	AP_{75}	AP_{50}	CO	UM	CG
F-RCNN [26]	24.3	20.2	49.0	9.0	75.8	22.3
C-RCNN [26]	**26.7**	**24.2**	**53.8**	**8.7**	**78.6**	**29.0**
S-RCNN [26]	8.2	7.2	16.1	4.8	25.3	6.8
RetinaNet [26]	20.1	16.7	43.4	4.5	76.3	10.4
CenterNet [26]	22.3	19.6	47.0	9.2	78.2	15.7
RepPoints [26]	4.2	3.1	10.1	4.5	16.7	3.5
ATSS [26]	25.6	23.2	49.7	11.5	79.3	20.1

Table 6. Comparison of performance across different auxiliary view detectorss **integrated with the proposed AENet**.

5.2.4 Confidence Score Thresholds

The results are shown in Figure 6. In this section, we analyze how varying confidence score thresholds influence the auxiliary-view module's performance. By adjusting the confidence threshold, we examine how the trade-off between precision and recall impacts the final results.

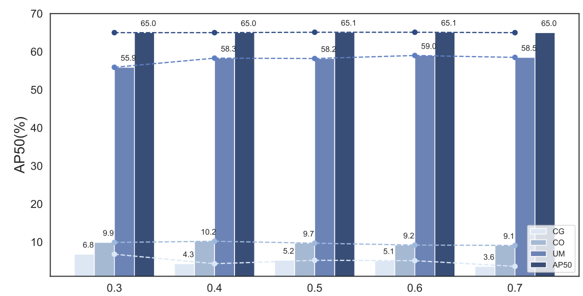


Figure 6. Impact of varying confidence score thresholds on the performance of the auxiliary view module, showing the optimal threshold for improving mAP and category-specific results.

As shown in Figure 6, setting the threshold to 0.6 yields the highest mAP of 39.1, with significant improvements in the CO and UM categories. Lower thresholds (*e.g.*, 0.3 or 0.4) result in a decrease in overall performance, likely due to an increase in false positives, while higher thresholds (e.g., 0.7) result in a drop due to missing true positives. These findings highlight the importance of setting an optimal confidence score threshold to maximize the contribution of the auxiliary-view module to the final detection results.

6. Conclusion

This paper introduces the Large-scale Dual-view X-ray (LDXray) dataset and the Auxiliary-view Enhanced Network (AENet). By leveraging both main and auxiliary views of X-ray images, AENet addresses the limitations of single-view detection, particularly for challenging categories. The extensive experiments conducted on the LDXray dataset demonstrate the effectiveness of the dual-view mechanism, showcasing its ability to enhance detection accuracy. The proposed dataset and framework not only offer a valuable resource for further research but also hold the potential to improve real-world security inspection.

7. Acknowledgment

Supported by National Key R&D Program of China under Grant No. 2022YFC3310200, Beijing Natural Science Foundation (No.L242021, No. L242022), National NSF of China (No. 62336001), and Talent Fund of Beijing Jiaotong University (No.2024XKRC047, No. 2024XKRC011).

References

[1] Neelanjan Bhowmik, Yona Falinie A Gaus, and Toby P Breckon. On the impact of using x-ray energy response imagery for object detection via convolutional neural networks. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1224–1228. IEEE, 2021. 2

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 7

[3] Yakun Chang, Yeliduosi Xiaokaiti, Yujia Liu, Bin Fan, Zhaojun Huang, Tiejun Huang, and Boxin Shi. Towards hdr and hfr video from rolling-mixed-bit spikings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25117–25127, 2024. 1

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[5] Adam Cobb, Anirban Roy, Daniel Elenius, Frederick Heim, Brian Swenson, Sydney Whittington, James Walker, Theodore Bapty, Joseph Hite, Karthik Ramani, et al. Aircraftverse: A large-scale multimodal dataset of aerial vehicle designs. *Advances in Neural Information Processing Systems*, 36:44524–44543, 2023. 1

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 7

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 3

[8] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[9] Jinyang Guo, Dong Xu, and Guo Lu. Cbanet: Toward complexity and bitrate adaptive deep image compression using a single network. *IEEE Transactions on Image Processing*, 32:2049–2062, 2023. 1

[10] Jinyang Guo, Dong Xu, and Wanli Ouyang. Multidimensional pruning and its extension: a unified framework for model compression. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 1

[11] Chengquan He, Tong Mu, Weiping Ren, and Bohua Zhao. Lpixray: A large-scale logistics prohibited item x-ray dataset for the application of deep learning in security inspection. In *2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, pages 481–485, 2023. 2

[12] Lingzhi He, Yakun Chang, Runmin Cong, Hongyu Liu, Shujuan Huang, Renshuai Tao, and Yao Zhao. Rethinking depth guided reflection removal. *IEEE Transactions on Multimedia*, 2025. 1

[13] Wei He, Kai Han, Ying Nie, Chengcheng Wang, and Yunhe Wang. Species196: A one-million semi-supervised dataset for fine-grained species recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[14] Brian K. S. Isaac-Medina, Chris G. Willcocks, and Toby P. Breckon. Multi-view object detection using epipolar constraints within cluttered x-ray security imagery. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9889–9896, 2021. 2

[15] Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. Climateset: A large-scale climate model dataset for machine learning. *Advances in Neural Information Processing Systems*, 36: 21757–21792, 2023. 1

[16] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Quanyi Li, Zhenghao Mark Peng, Lan Feng, Zhizheng Liu, Chenda Duan, Wenjie Mo, and Bolei Zhou. Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling. *Advances in neural information processing systems*, 36, 2024. 1

[18] Ziqiang Li, Muhammad Usman, Rentuo Tao, Pengfei Xia, Chaoyue Wang, Huanhuan Chen, and Bin Li. A systematic survey of regularization and normalization in gans. *ACM Computing Surveys*, 55(11):1–37, 2023. 1

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[20] Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[21] Bowen Ma, Tong Jia, Min Su, Xiaodong Jia, Dongyue Chen, and Yichun Zhang. Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake. *IEEE Transactions on Multimedia*, 2022. 2, 3

[22] Bowen Ma, Tong Jia, Mingyuan Li, Songsheng Wu, Hao Wang, and Dongyue Chen. Towards dual-view x-ray baggage inspection: A large-scale benchmark and adaptive hierarchical cross refinement for prohibited item discovery. *IEEE Transactions on Information Forensics and Security*, 2024. 1, 2

[23] Erika McCallister, Tim Grance, and Karen Scarfone. Identifiable information (pii). *NIST Special Publication*, 800:122, 2010. 3

[24] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of x-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34(4):42, 2015. 1, 2

[25] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2119–2128, 2019. 1, 2

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7, 8

[27] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 1

[28] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. *arXiv preprint arXiv:2501.09781*, 2025. 1

[29] Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, and Lianwen Jin. M5hisdoc: A large-scale multi-style chinese historical document analysis benchmark. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[30] Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[31] Jan-Martin O Steitz, Faraz Saeedan, and Stefan Roth. Multi-view x-ray r-cnn. In *German Conference on Pattern Recognition*, pages 153–168. Springer, 2018. 2, 3

[32] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021. 7

[33] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu. Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10923–10932, 2021. 1, 2, 3

[34] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Bowei Jin, Hongping Zhi, Xianglong Liu, and Aishan Liu. Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21157–21167. IEEE, 2022. 1, 2

[35] Renshuai Tao, Tianbo Wang, Ziyang Wu, Cong Liu, Aishan Liu, and Xianglong Liu. Few-shot x-ray prohibited item detection: A benchmark and weak-feature enhancement network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2012–2020, 2022. 2

[36] Renshuai Tao, Manyi Le, Chuangchuang Tan, Huan Liu, Haotong Qin, and Yao Zhao. Oddn: Addressing unpaired data challenges in open-world deepfake detection on online social networks. *arXiv preprint arXiv:2410.18687*, 2024. 1

[37] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5412–5421, 2021. 2, 3

[38] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM international conference on multimedia*, pages 138–146, 2020. 1, 2

[39] Modi Wu, Feifan Yi, Haigang Zhang, Xinyu Ouyang, and Jinfeng Yang. Dualray: Dual-view x-ray security inspection benchmark and fusion detection framework. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 721–734. Springer, 2022. 2, 3

[40] Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1366–1375, 2021. 1

[41] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9657–9666, 2019. 7

[42] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 7

[43] Cairong Zhao, Liang Zhu, Shuguang Dou, Weihong Deng, and Liang Wang. Detecting overlapped objects in x-ray security imagery by a label-aware mechanism. *IEEE transactions on information forensics and security*, 17:998–1009, 2022. 2

[44] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5697–5707, 2022. 1