

Superpowering Open-Vocabulary Object Detectors for X-ray Vision

Pablo Garcia-Fernandez^{1*} Lorenzo Vaquero^{1,2} Mingxuan Liu³ Feng Xue^{3*}
Daniel Cores¹ Nicu Sebe³ Manuel Mucientes¹ Elisa Ricci^{2,3}

¹University of Santiago de Compostela, Spain ²Fondazione Bruno Kessler, Italy

³University of Trento, Italy

{pablogarcia.fernandez, daniel.cores, manuel.mucientes}@usc.es lvaquero@fbk.eu
{mingxuan.liu, feng.xue, niculae.sebe, e.ricci}@unitn.it *Corresponding authors

Abstract

Open-vocabulary object detection (OvOD) is set to revolutionize security screening by enabling systems to recognize any item in X-ray scans. However, developing effective OvOD models for X-ray imaging presents unique challenges due to data scarcity and the modality gap that prevents direct adoption of RGB-based solutions. To overcome these limitations, we propose **RAXO**, a training-free framework that repurposes off-the-shelf RGB OvOD detectors for robust X-ray detection. RAXO builds high-quality X-ray class descriptors using a dual-source retrieval strategy. It gathers relevant RGB images from the web and enriches them via a novel X-ray material transfer mechanism, eliminating the need for labeled databases. These visual descriptors replace text-based classification in OvOD, leveraging intra-modal feature distances for robust detection. Extensive experiments demonstrate that RAXO consistently improves OvOD performance, providing an average mAP increase of up to 17.0 points over base detectors. To further support research in this emerging field, we also introduce **DET-COMPASS**, a new benchmark featuring bounding box annotations for over 300 object categories, enabling large-scale evaluation of OvOD in X-ray. Code and dataset available at: <https://pagf188.github.io/RAXO/>.

1. Introduction

Automated object detection technologies for X-ray imaging are essential to maintain public safety, enabling the identification of prohibited items at checkpoints in high-risk environments such as airports, train stations, museums and stadiums [29]. These systems improve security while simultaneously reducing the workload of human inspectors.

Conventional X-ray object detectors rely on supervised learning [2, 14, 20] and are inherently limited by the object categories present in their training datasets (see Fig. 1a). This limitation is exacerbated by the high cost of X-ray machinery and the requirement for expert annotation, often restricting these systems to fewer than 20 object classes [32],

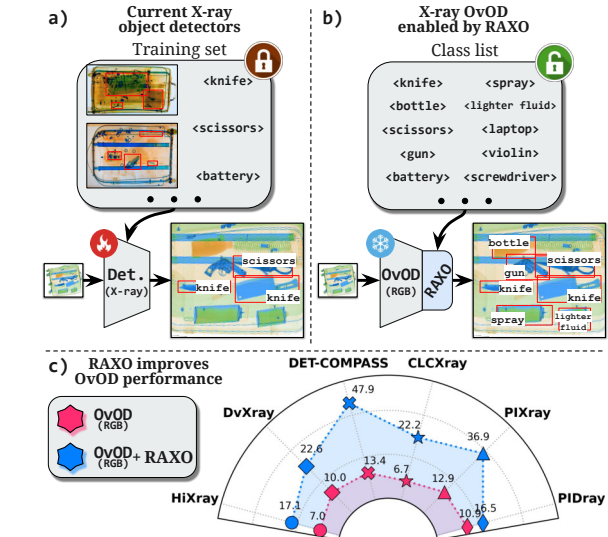


Figure 1. (a) Traditional X-ray object detectors are constrained by the limited categories in their training datasets. (b) We introduce the task of open-vocabulary object detection (OvOD) for X-ray imaging and propose RAXO, a training-free method that adapts off-the-shelf RGB OvOD models to X-ray data. (c) RAXO greatly improves detection performance across multiple benchmarks.

thereby hindering broader real-world applications.

In light of the expanding diversity of man-made objects and evolving security demands, an open-vocabulary object detection (OvOD) framework capable of recognizing arbitrary X-ray object categories defined by the user is imperative. Despite its importance, open-vocabulary detection in the X-ray domain remains largely unexplored. Concurrent works aiming to extend beyond base categories have managed to generalize to at most four unseen classes [16], thereby highlighting the inherent challenges of the task.

Recent advances in OvOD for conventional RGB images have been driven by large-scale annotated datasets, which facilitate effective alignment between visual and textual features [15, 18, 44]. However, these advancements do not directly transfer to X-ray imagery. In practice, applying state-of-the-art RGB-based OvOD models to X-ray scans leads to significant performance degradation, as illustrated

in Fig. 1c. Moreover, retraining these models on X-ray data is often impractical given the scarcity of large-scale annotated X-ray datasets.

Motivated by these challenges, this work opens up open-vocabulary object detection for X-ray imaging by repurposing robust, off-the-shelf RGB-based detectors. To this end, we propose **RAXO** (tRAining-free adaptation for X-ray Open-vocabulary detection), a training-free method that seamlessly adapts RGB OvOD models to X-ray. (Fig. 1b).

We find that the main reason RGB-based OvOD detectors fail in the X-ray domain is the disruption of text-visual feature alignment. The appearance disparity between the same object in RGB and X-ray modalities causes textual embeddings, aligned with RGB features, to mismatch with X-ray visual features. RAXO addresses this issue on three simple steps: (1) *visual sample acquisition*, which involves obtaining X-ray images that represent user-defined categories; (2) *class descriptor modeling*, which uses these images to construct descriptors that effectively encode the visual appearance of each category within the X-ray domain; and (3) *classifier construction*, where the computed descriptors are used to build a visual classifier that replaces the inter-modal (*i.e.*, text-to-visual) classification of conventional OvODs. This new classifier exploits the fact that, despite the modality shift, intra-modal (*e.g.*, visual-to-visual) feature distances remain reliable indicators for object identification. In this way RAXO enables any OvOD to successfully detect X-ray objects based on their inherent visual properties.

To facilitate comprehensive evaluation, we also introduce DET-COMPASS, a detection dataset comprising 370 distinct object classes with paired X-ray and RGB annotations. Extensive experiments demonstrate that our approach consistently improves detection performance by an average of $\uparrow 8.4$ AP across multiple benchmarks and OvOD models, with gains that scale as the underlying detectors become more powerful (Tab. 2). In summary, our contributions are:

- We formally introduce the problem of open-vocabulary object detection for X-ray imagery without training, addressing a critical need for security in real-world scenarios.
- We introduce DET-COMPASS, a novel benchmark with bounding box annotations across 370 object categories, enabling standardized evaluation of OvOD methods in the X-ray modality.
- We propose RAXO, a novel training-free approach that repurposes off-the-shelf RGB-based OvOD models for X-ray detection by constructing robust visual descriptors. RAXO achieves new state-of-the-art results across multiple benchmarks.

2. Related Work

Open-vocabulary object detection has advanced rapidly with the advent of Vision-Language Models (VLMs), enabling detectors to generalize beyond fixed categories.

OvOD detectors leverage weak supervision signals to improve detection accuracy. Based on the type of weak supervision utilized, OvOD methods can be categorized into (i) region-aware training, (ii) pseudo-labeling, (iii) knowledge distillation, and (iv) transfer learning.

Region-aware training methods aim to improve localization and feature representation by refining the alignment between image regions and their corresponding textual descriptions. Approaches such as DetCLIP [40], DetCLIPv2 [41], CORA [39], and VLDet [15] adopt this strategy. Pseudo-labeling methods, rely on large pretrained VLMs to generate pseudo-labels, effectively expanding the training set. Methods like RegionCLIP [43], PromptDet [6], CoDET [21], GLIP [13], Detic [44], and Grounding DINO [18] follow this approach. Knowledge distillation techniques, such as BARON [37], DK-DETR [12], CLIPSelf [38], and SIC-CADS [5], employ VLMs as teachers in a teacher-student framework, transferring knowledge from the VLM image encoder to enhance the detector backbone. In contrast, transfer learning approaches integrate the VLM encoder directly, either through fine-tuning, as seen in OWL-ViT [23], or by freezing the encoder, as in F-VLM [11].

Regardless of the training strategy, OvOD methods heavily depend on VLMs trained on large-scale datasets. However, such datasets are unavailable in the X-ray modality, making this approach impractical. To address this, RAXO seamlessly adapts existing RGB-based OvOD methods for X-ray object detection without requiring additional training.

X-ray object detection. Several datasets have been developed to tackle the challenge of object detection in X-ray imagery [1, 33]. SIXray [22], OPIXray [36], and CLCXray [42] focused on detecting occluded objects, introducing strategies to identify deliberately hidden prohibited object. To enhance dataset scale and diversity, PIDray [34] and HiXray [30] provided larger benchmarks, significantly increasing the number of annotated images. Meanwhile, PIXray [19] pioneered the introduction of an X-ray segmentation dataset and proposed a real-time framework for segmenting prohibited objects, further advancing automated threat detection in security screening applications.

Previous X-ray object detectors [28, 31, 35] considered mainly a closed-set paradigm, thus being limited to localize objects of a small set of predefined categories. More recently, Lin *et al.* [16] proposed fine-tuning a CLIP adapter to bridge the modality gap between CLIP’s training data and X-ray images, thereby achieving OvOD in X-ray data. They demonstrated generalization on four novel object classes, but both training and evaluation were limited by the lack of large-scale, well-annotated X-ray data. To address these issues, RAXO leverages data retrieval from the web to expand detection capabilities to a broader range of common objects. Furthermore, our proposed re-labeled DET-COMPASS dataset enables evaluation across a wider variety of classes.

	Venue	Images	Classes	Modality
DvXray [20]	TIFS'24	32,000	15	X-ray
PIXray [19]	TMM'22	5,046	15	X-ray
CLCXray [42]	TIFS'22	9,565	12	X-ray
FSOD [32]	ACMMM'22	12,333	20	X-ray
EDS [31]	CVPR'22	14,219	10	X-ray
PIDray [34]	ICCV'21	47,677	12	X-ray
HiXray [30]	ICCV'21	45,365	8	X-ray
DET-COMPASS (Ours)	–	1,928	370	X-ray+RGB

Table 1. **Existing X-ray detection datasets.** DET-COMPASS is the dataset with the highest number of categories and the only one providing pixel-level alignment between X-ray and RGB data.

3. DET-COMPASS

Object detection in security X-ray scans has advanced significantly in recent years. However, evaluating OvOD detectors in this modality remains challenging due to the limited number of annotated object categories in existing X-ray benchmarks. For instance, the largest X-ray detection dataset, FSOD [32], includes annotations for only 20 classes (see Tab. 1). This limitation severely constrains the comprehensive evaluation of OvOD methods, which require a broad and diverse category set (or say vocabulary) to assess generalization to unseen object semantics.

To address this gap, we introduce *DET-COMPASS*, a novel benchmark that repurposes the COMPASS-XP classification dataset [9] for object detection through meticulous *manual bounding box annotation*. DET-COMPASS comprises 370 distinct object classes, offering an *order-of-magnitude increase in vocabulary size* over previous X-ray detection benchmarks. Additionally, it provides pixel-aligned RGB images, ensuring precise spatial correspondence across modalities and facilitating the development of multimodal models. Each object is also labeled with a visibility attribute, indicating whether it produces a discernible signature in the X-ray spectrum. Further details are provided in Appendix C.

As summarized in Tab. 1, DET-COMPASS sets a new standard in class diversity and uniquely integrates multimodal, pixel-aligned X-ray and RGB data. The dataset will be released under an open license, serving as a valuable resource for advancing OvOD research in security screening and industrial inspection.

4. OvOD for X-ray Imaging

Preliminaries: OvOD in the RGB domain. Most RGB OvOD detectors [44, 45] follow a two-stage pipeline. During training, a region proposal network (RPN) is learned to yield a set of M proposals by $\{\mathbf{z}_m\}_{m=1}^M = \Phi_{\text{RPN}}(\mathbf{I}^{\text{RGB}})$, where $\mathbf{z}_m \in \mathbb{R}^D$ is a D -dimensional region-of-interest (RoI) feature embedding. Then, a bounding box regressor predicts coordinates for each proposed region via $\hat{\mathbf{b}}_m = \Phi_{\text{REG}}(\mathbf{z}_m)$. A set of text-based classifiers $\mathbf{W} = \{\mathbf{w}_c | \mathbf{w}_c \in \mathbb{R}^D\}_{c=1}^{|\mathcal{C}^{\text{train}}|}$ are used

to compute classification scores for each region as $\langle \mathbf{w}_c, \mathbf{z}_m \rangle$, where $\langle \cdot, \cdot \rangle$ is the cosine similarity function and $\mathcal{C}^{\text{train}}$ denotes the training vocabulary. In this way, each region’s class is determined by the class with the highest score. Here, the classifier \mathbf{W} is constructed by encoding class names in $\mathcal{C}^{\text{train}}$ using a pre-trained VLM text encoder, e.g., CLIP [26]. During training, OvOD models update all parameters while keeping \mathbf{W} frozen. This enables RGB-region-class alignment by leveraging large-scale RGB data and the pre-aligned vision-language semantic space of VLMs, facilitating open-vocabulary inference with any test vocabulary $\mathcal{C}^{\text{test}}$. The vocabularies $\mathcal{C}^{\text{train}}$ and $\mathcal{C}^{\text{test}}$ may be disjoint or overlapping.

Problem formulation. In this work, we study open-vocabulary object detection (OvOD) in X-ray modality. Specifically, given an input X-ray image \mathbf{I} and a vocabulary $\mathcal{C}^{\text{test}}$ defined by users at test time, an X-ray OvOD detector $\mathcal{F}_{\text{X-ray}}$ aims to detect objects specified in $\mathcal{C}^{\text{test}}$ from \mathbf{I} (e.g., a “Power bank” in a passenger’s backpack scan at an airport security checkpoint). Theoretically, this detection process can be formulated as $\mathcal{F}_{\text{X-ray}}: \mathbf{I} \rightarrow \{(\mathbf{b}_m, c_m)\}_{m=1}^M$, where $\mathbf{b}_m \in \mathbb{R}^4$ denotes the coordinates of each bounding box, and $c_m \in \mathcal{C}^{\text{test}}$ denotes the class label of each bounding box.

X-ray OvOD presents significant challenges due to the following reasons: *i)* Directly applying a pre-trained RGB OvOD detectors to X-ray images leads to suboptimal performance due to the visual modality gap, as shown in Fig. 1; *ii)* The scarcity of large-scale security X-ray datasets limits the application of RGB OvOD training techniques. These techniques typically rely on extensive image-text annotated data for strong supervision. To tackle these challenges, we introduce RAXO, a plug-and-play module that adapts any off-the-shelf RGB OvOD detectors to X-ray modalities. RAXO requires no training or in-domain detection annotations. Next, we present our approach.

5. RAXO: Training-free Modality Adaptation

As illustrated in Fig. 2, RAXO enables X-ray OvOD by constructing high-quality visual descriptors, \mathcal{X}_c , for each class c in the user-defined vocabulary, $\mathcal{C}^{\text{test}}$. To achieve this, RAXO first **acquires** X-ray samples in an open-ended manner, leveraging both *in-house* and *web-based* retrieval sources (Sec. 5.1). Then, it extracts the features of these samples and segments the relevant information, **modeling** the visual descriptor \mathcal{X}_c (Sec. 5.2). Once the visual descriptors are constructed offline, they can be directly applied to any *off-the-shelf* OvOD detector by replacing the conventional text-based classifier \mathbf{W} , with our **visual-based classifier** \mathcal{X} (Sec. 5.3). Thus, RAXO effectively overcomes the misalignment between X-ray features and text semantics, enabling OvOD detectors pre-trained with RGB data to accurately identify X-ray objects based on their intrinsic visual characteristics.

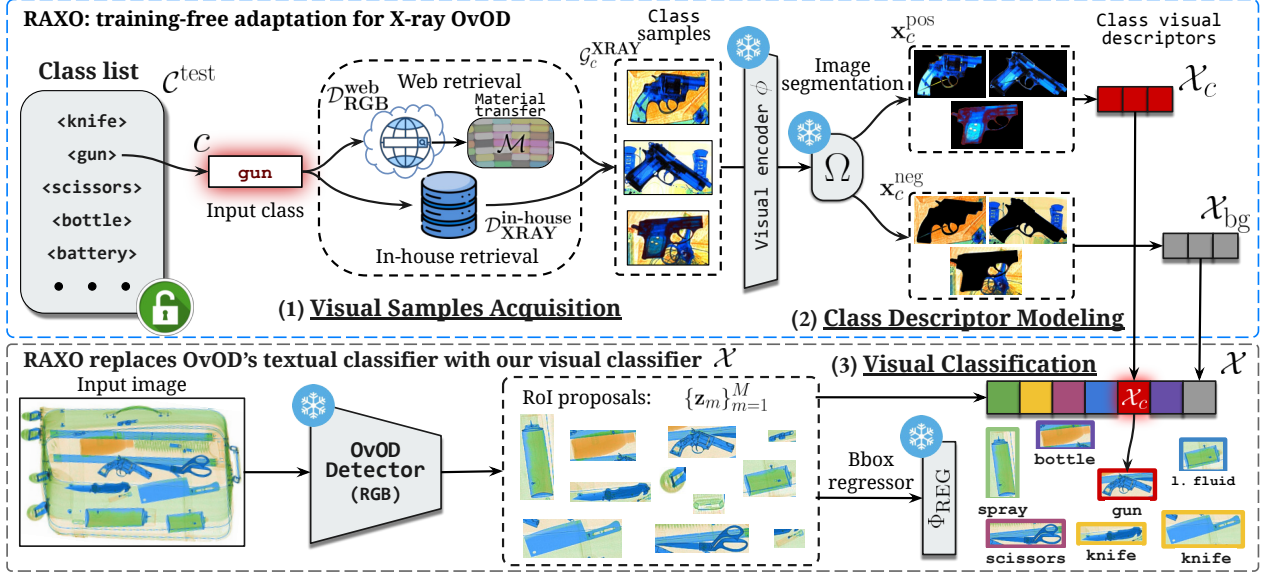


Figure 2. **Architecture of RAXO.** For a given user-defined class $c \in C^{\text{test}}$, RAXO first retrieves its corresponding X-ray images $\mathcal{G}_c^{\text{XRAY}}$ from in-house and web sources, using its (1) *Visual Samples Acquisition* pipeline (Sec. 5.1). Following this, RAXO extracts the features of the images and segments them with its (2) *Class Descriptor Modeling* module (Sec. 5.2), creating ensemble visual descriptors for the class \mathcal{X}_c and the background \mathcal{X}_{bg} . Finally, the text-based classifier from the baseline RGB OvOD detector is replaced with our (3) *Visual-based Classifier* (Sec. 5.3) \mathcal{X} , which yields accurate predictions on the X-ray modality.

5.1. Visual Sample Acquisition

Obtaining informative and representative images is crucial for generating robust visual descriptors. To this end, we propose a *dual-source* acquisition pipeline that retrieves a gallery of relevant X-ray samples $\mathcal{G}_c^{\text{XRAY}}$ for each user-specified class $c \in C^{\text{test}}$. This pipeline consists of two modules: an *in-house* retrieval module and a *web-powered* retrieval module.

In-house retrieval. Given an in-domain X-ray dataset, $\mathcal{D}_{\text{XRAY}}^{\text{in-house}}$, containing categories $C^{\text{in-house}}$, our sample acquisition pipeline first attempts to retrieve the K most relevant images for the user-specified class c based on class name matching: $\mathcal{G}_c^{\text{XRAY}} = \{\mathbf{I} | (\mathbf{I}, c) \in \mathcal{D}_{\text{XRAY}}^{\text{in-house}}\}^K$.

However, since the user-defined vocabulary C^{test} is open-ended, we cannot assume that every user-specified category will be present in $C^{\text{in-house}}$. Consequently, for some categories, the retrieved set $\mathcal{G}_c^{\text{XRAY}}$ may be empty. To overcome this limitation and fully support open-vocabulary user input, RAXO further incorporates a novel *web-powered* retrieval module. This module retrieves RGB images from the web and applies a material transfer mechanism to synthesize X-ray-style representations, which we introduce next.

Web-powered retrieval. To obtain high-quality visual samples for a category c that is not present in the in-house vocabulary $C^{\text{in-house}}$, we leverage the vast availability of web-based RGB image data $\mathcal{D}_{\text{RGB}}^{\text{web}}$ as an open-ended auxiliary source. Specifically, we perform text-based web retrieval using the class name c as a search

query, retrieving the top K results from Google Images as $\tilde{\mathcal{G}}_c^{\text{web}} = \{\mathbf{I}^{\text{RGB}} | (\mathbf{I}^{\text{RGB}}, c) \in \mathcal{D}_{\text{RGB}}^{\text{web}}\}^K$.

The raw web-retrieved results $\tilde{\mathcal{G}}_c^{\text{web}}$ are often noisy and may not always contain clear instances of the target class c . To refine these results, we apply a filtering step using an RGB OvOD detector \mathcal{F}_{RGB} . Specifically, we discard images in the raw web-retrieved results where class c is *not* confidently detected, retaining only those where the detection confidence exceeds a threshold τ as $\mathcal{G}_c^{\text{web}} = \text{Filter}(\tilde{\mathcal{G}}_c^{\text{web}}, \mathcal{F}_{\text{RGB}}, c, \tau)$.

Material-transfer mechanism for web-retrieved images. The substantial visual disparity between RGB and X-ray modalities prevents direct use of web-retrieved RGB samples $\mathcal{G}_c^{\text{web}}$ for constructing X-ray class descriptors. Style transfer methods [7] fall short in bridging this gap, as they fail to capture the underlying material properties of objects (as shown in Tab. 4). To address this, we introduce a novel material-transfer mechanism for generating synthetic X-ray samples from web-retrieved images. As shown in Fig. 3, our approach consists of two key steps: *i*) constructing an offline material database \mathcal{M} that encapsulates the expected X-ray appearance of various *materials*, and *ii*) adapting RGB samples $\mathcal{G}_c^{\text{web}}$ to X-ray style by applying the corresponding material properties from \mathcal{M} .

To construct \mathcal{M} , we employ a Large Language Model (LLM) to cluster the class names of $C^{\text{in-house}}$ into subsets $C_m \subset C^{\text{in-house}}$, where C_m contains classes that share the same m material – *i.e.*, “ $C_{\text{metal}} = \{\text{gun}, \text{knife}, \text{fork}\}$ ” or “ $C_{\text{leather}} = \{\text{boot}, \text{belt}\}$ ”. Each material m is then associated

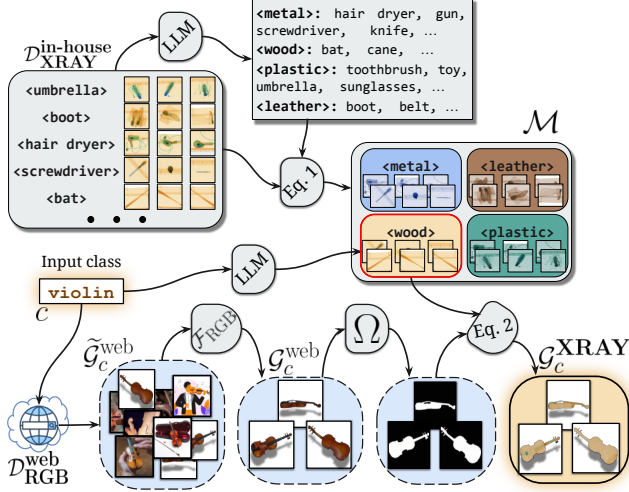


Figure 3. **Web-powered retrieval and material-transfer mechanism** for the class “violin”. We retrieve violin samples from the web, filter them using \mathcal{F}_{RGB} , and inpaint the retrieved appearance into the object masks to generate synthetic X-ray samples.

with the corresponding set of images $\mathcal{K}_m = \{\mathbf{I} \mid (\mathbf{I}, c_k) \in \mathcal{D}_{\text{XRAY}}^{\text{in-house}}, c_k \in \mathcal{C}_m\}$. These images are used to compute the appearance, $\mathcal{A}_m \in \mathbb{R}^3$, of the material as:

$$\mathcal{A}_m = \frac{1}{|\mathcal{K}_m|} \sum_{\mathbf{I} \in \mathcal{K}_m} \left(\frac{\sum_{i,j}^{H,W} (\mathbf{I}_{i,j} \odot \Omega(\mathbf{I}_{i,j}))}{\sum \Omega(\mathbf{I})} \right), \quad (1)$$

where \odot denotes element-wise multiplication and $\Omega : \mathbf{I} \rightarrow \{0, 1\}^{H \times W}$ is a segmentation function that produces a binary mask, to isolate the object of interest in \mathbf{I} . The resulting material-appearance pairs (m, \mathcal{A}_m) collectively form our material database \mathcal{M} .

Once \mathcal{M} is computed offline, it can be used to adapt the web-retrieved RGB samples, $\mathcal{G}_c^{\text{web}}$, to the X-ray modality. To achieve this, we first prompt the LLM to link the class c with its corresponding material $m_c \in \mathcal{M}$ (e.g., $c = \text{“violin”}$ is linked with the material $m_c = \text{“wood”}$). Subsequently, the X-ray appearance \mathcal{A}_m^c of the linked material is used to render the characteristic X-ray style on each web-retrieved RGB samples as:

$$\mathcal{G}_c^{\text{XRAY}} = \{\Omega(\mathbf{I}^{\text{RGB}}) \odot (\mathcal{A}_m^c \cdot \mathbf{1}) \mid \mathbf{I}^{\text{RGB}} \in \mathcal{G}_c^{\text{web}}\}, \quad (2)$$

Finally, the X-ray visual gallery for every class freely specified by the user is built as $\mathcal{G} = \bigcup_{c \in \mathcal{C}^{\text{test}}} \mathcal{G}_c^{\text{XRAY}}$ through our dual-source retrieval pipeline, seamlessly integrating high-quality in-house X-ray samples with synthetic images derived from web data.

5.2. Class Descriptor Modeling

The effectiveness of RAXO hinges on leveraging its visual gallery \mathcal{G} to construct a robust visual descriptor \mathcal{X}_c that accurately captures the visual X-ray properties of class c . A

naïve approach, such as averaging the feature representations of all $\mathcal{G}_c^{\text{XRAY}}$ instances, fails to account for intra-class variability (e.g., distinct shapes of utility knives versus chef knives) and does not differentiate between foreground and background regions. To address these limitations, we propose a novel class descriptor modeling strategy shown in Fig. 2(2).

First, we process each sample $\mathbf{I} \in \mathcal{G}_c^{\text{XRAY}}$ with Ω to separate the X-ray object from its background. Simultaneously, a feature extractor ϕ extracts per-patch embeddings $\phi(\mathbf{I}) \in \mathbb{R}^{H' \times W' \times D}$, where $H' \times W'$ represents the number of spatial tokens, and D is the embedding dimension. Following [10], we resize (denoted by ζ) the segmentation mask to match the spatial dimensions of the feature map, and subsequently we use it to compute both a positive and a negative prototype. The *positive prototype*, designed to capture the significant visual features of the object without background interference, is computed as the average embedding over the foreground region as:

$$\mathbf{x}_{\mathbf{I}}^{\text{pos}} = \frac{\sum_{i,j}^{H',W'} \zeta(\Omega(\mathbf{I}_{i,j})) \odot \phi(\mathbf{I}_{i,j})}{\sum \zeta(\Omega(\mathbf{I}))}, \quad (3)$$

On the other hand, the *negative prototype* is obtained as the average embedding over the complementary region (background) as:

$$\mathbf{x}_{\mathbf{I}}^{\text{neg}} = \frac{\sum_{i,j}^{H',W'} (1 - \zeta(\Omega(\mathbf{I}_{i,j}))) \odot \phi(\mathbf{I}_{i,j})}{\sum (1 - \zeta(\Omega(\mathbf{I})))}. \quad (4)$$

Subsequently, to construct the final visual descriptor \mathcal{X}_c of class c , we compute the average positive prototype $\bar{\mathbf{x}}_c^{\text{pos}}$ from $\mathcal{G}_c^{\text{XRAY}}$ and unite it with all the individual positive prototypes as:

$$\mathcal{X}_c = [\bar{\mathbf{x}}_c^{\text{pos}}, \{\mathbf{x}_{\mathbf{I}}^{\text{pos}} \mid \mathbf{I} \in \mathcal{G}_c\}]. \quad (5)$$

This formulation effectively captures both fine-grained object details and a holistic class-level representation, handling intra-class variability. Additionally, we also construct a global *background descriptor* by using the negative prototypes across the entire visual sample gallery as $\mathcal{X}_{bg} = [\bar{\mathbf{x}}^{\text{neg}}, \{\mathbf{x}_{\mathbf{I}}^{\text{neg}} \mid \mathbf{I} \in \mathcal{G}\}]$. This negative prototype is used to further improve detection reliability by filtering out low-quality proposals (e.g., a region proposed on the background) during inference. A key advantage of this approach is its modularity: visual descriptors \mathcal{X}_c are computed *offline* and can be incrementally expanded with new object categories, seamlessly integrating with the OvOD paradigm without training or requiring X-ray detection data.

5.3. Classification is All You Need

As shown in Fig. 2(3), once the visual descriptors are constructed for each class in the user-specified vocabulary, along with the background class, the RAXO classifier

$\mathcal{X} = \{[\mathcal{X}_c, \mathcal{X}_{bg}] \mid c \in \mathcal{C}^{\text{test}}\}$ can be directly applied to any OvOD detector to classify proposals \mathbf{z}_m as:

$$\hat{c} = \arg \max_{c \in \mathcal{C}^{\text{test}'}} \max_{\mathbf{x} \in \mathcal{X}} \langle \mathbf{z}_m, \mathbf{x} \rangle, \quad (6)$$

where $\mathcal{C}^{\text{test}'}$ denotes the test-time vocabulary $\mathcal{C}^{\text{test}}$ extended with the additional ‘‘background’’ class. If a proposal matches the background, it is removed. This simple-yet-effective strategy enables any pre-trained OvOD detectors to achieve strong performance on X-ray object detection with minimal modifications and no need for re-training.

Descriptor consistency criterion. To further reduce incorrect proposals from the OvOD detector, RAXO introduces a novel *Descriptor Consistency Criterion* (DCC) to enforce class consistency in predictions. DCC evaluates how well a proposal aligns with its predicted class relative to others, suppressing weakly aligned proposals. For a proposal \mathbf{z}_m , we measure its similarity to the closest prototype as $s_1 = \max_{\mathbf{x} \in \mathcal{X}_{\hat{c}}} \langle \mathbf{z}_m, \mathbf{x} \rangle$. We also compute its mean similarity to the average prototypes of all other classes as $s_2 = \text{avg}_{c \in \mathcal{C}^{\text{test}}, c \neq \hat{c}} \langle \mathbf{z}_m, \bar{\mathbf{x}}_c^{\text{pos}} \rangle$. The difference $\Delta = s_1 - s_2$ serves as a confidence measure, where a higher value indicates greater alignment with its predicted class than with any other. Proposals with Δ below a threshold σ are discarded.

6. Experiments

Evaluation protocol. Conventional OvOD models are typically assessed using a *Cross-Dataset Transfer Evaluation* (CDTE) protocol, where the model is trained on one dataset and evaluated on a different dataset in a zero-shot manner [45]. In contrast, our setting involves a more challenging scenario where both the datasets and their underlying modalities differ. To address this, we introduce a novel *Cross-Modality Transfer Evaluation* (CMTE) protocol. Under CMTE, an OvOD model is trained on a *source RGB* dataset and subsequently evaluated on *target X-ray* datasets without any additional training or fine-tuning. For performance assessment, we employ the standard MS COCO [17] metrics: AP, AP50, and AP75.

Implementation details. Unless stated otherwise, RAXO’s default configuration employs SAM 2 [27] as the object segmentation module Ω , ViT-B/14 [3] pretrained with DINOv2 [25] as the visual encoder ϕ , and GPT-4 [24] as the LLM. We set the consistency threshold to $\sigma = 0.15$, the similarity threshold to $\tau = 0.5$, and use $K = 30$ samples to construct the visual descriptors. $\mathcal{D}_{\text{XRAY}}^{\text{in-house}}$ is composed of hold-out samples sourced from the evaluation X-ray datasets and is completely disjoint from the test split. $\mathcal{D}_{\text{RGB}}^{\text{web}}$ is constructed by collecting images from the web using the public Google Custom Search API [8]. A detailed analysis of the impact of hyperparameter choices is provided in Sec. 6.2. Further implementation details can be found in Supp. D.

Datasets and baselines. We evaluate RAXO by integrating it with four different state-of-the-art OvOD detectors that were exclusively trained on RGB images: GroundingDINO [18], Detic [44], VLDet [15], and CoDet [21]. Our CMTE evaluation is performed across six diverse X-ray datasets: PIXray [19], PIDray [34], CLCXray [42], DvXray [20], HiXray [30], and our proposed DET-COMPASS, which together comprise 343 visible unique classes and over 140k images. A detailed description of these datasets is provided in Tab. 1.

6.1. Open-Vocabulary Detection Results

RAXO equips OvOD models with X-ray vision. We evaluate the X-ray adaptation capabilities of RAXO on the aforementioned datasets and baselines in Tab. 2. Since RAXO is training-free, we report the results directly on the test splits, accounting for all available categories. Furthermore, we analyze the influence of the gallery composition by varying \mathcal{G} from a *0/100* configuration (*i.e.*, all samples in \mathcal{G} are retrieved from the web, as $\mathcal{C}^{\text{test}} \cap \mathcal{C}^{\text{in-house}} = \emptyset$) to a *100/0* configuration (*i.e.*, all samples in \mathcal{G} are retrieved from the in-house database, as $\mathcal{C}^{\text{test}} \subset \mathcal{C}^{\text{in-house}}$). Each experiment is repeated three times with different random distributions of in-domain and web categories, and we report the averaged results.

As shown in Tab. 2, off-the-shelf OvOD detectors trained on RGB images perform poorly on X-ray data. In contrast, even without access to any in-domain data from $\mathcal{C}^{\text{test}}$ (*0/100* setting), RAXO enhances baseline OvOD methods, yielding an average improvement of $\uparrow 2.1$ points. This gain is further amplified when in-domain samples are incorporated into \mathcal{G} , which facilitates the construction of more robust materials in \mathcal{M} and yields higher-quality visual descriptors. Notably, even a limited number of in-domain samples (*20/80* setting) results in an average improvement of $\uparrow 4.8$ points. Moreover, increasing the in-domain samples (*50/50* setting) yields even more significant gains, particularly in datasets with a large number of classes, such as DET-COMPASS, where RAXO offers an average boost of $\uparrow 14.9$ points. Finally, when \mathcal{G} consists entirely of in-domain samples, RAXO delivers an average improvement of $\uparrow 14.7$ points across all OvOD methods and datasets, all in a training-free manner.

RAXO works with any OvOD detector. The consistent gains of RAXO across all evaluated detectors in Tab. 2, along with its simple integration (Sec. 5.3), confirm the generalization of our approach to any RGB OvOD model. Moreover, RAXO scales with the baseline method, yielding larger gains when integrated with more robust detectors (*e.g.*, RAXO enhances G-DINO $\uparrow 9.4$ points on average across all \mathcal{G} settings, compared to $\uparrow 7.5$ points for Detic). This highlights the advantage of our training-free method, which can leverage the rapid advancements in RGB OvOD detectors without needing large amounts of labeled X-ray data.

\mathcal{G}	Method	D-COMPASS	PIXray	PIDray	CLCXray	DvXray	HiXray	Avg.
G-DINO [18]		13.4	12.9	10.9	6.7	10.0	7.0	10.2
$\mathcal{D}_{\text{XRAY}}^{\text{in-h}}$ 100/0 80/20 50/50 20/80 0/100 $\mathcal{D}_{\text{RGB}}^{\text{web}}$	+ RAXO	47.9 \uparrow 34.5	36.9 \uparrow 24.0	16.5 \uparrow 5.6	22.2 \uparrow 15.5	22.6 \uparrow 12.6	17.1 \uparrow 10.1	27.2 \uparrow 17.0
		41.0 \uparrow 27.6	33.8 \uparrow 20.9	15.4 \uparrow 4.5	18.0 \uparrow 11.3	21.0 \uparrow 11.0	14.5 \uparrow 7.5	24.0 \uparrow 13.8
		31.4 \uparrow 18.0	25.4 \uparrow 12.5	15.5 \uparrow 4.6	17.0 \uparrow 10.3	16.1 \uparrow 6.1	13.4 \uparrow 6.4	19.8 \uparrow 9.6
		20.5 \uparrow 7.1	21.6 \uparrow 8.7	13.9 \uparrow 3.0	10.0 \uparrow 3.3	15.0 \uparrow 5.0	9.8 \uparrow 2.8	15.1 \uparrow 4.9
		14.0 \uparrow 0.6	16.1 \uparrow 3.2	13.4 \uparrow 2.5	7.1 \uparrow 0.4	12.4 \uparrow 2.4	7.9 \uparrow 0.9	11.8 \uparrow 1.6
VLDet [15]		10.6	9.8	6.9	4.4	7.4	5.1	7.4
$\mathcal{D}_{\text{XRAY}}^{\text{in-h}}$ 100/0 80/20 50/50 20/80 0/100 $\mathcal{D}_{\text{RGB}}^{\text{web}}$	+ RAXO	36.4 \uparrow 25.8	32.3 \uparrow 22.5	11.7 \uparrow 4.8	15.4 \uparrow 11.0	20.1 \uparrow 12.7	14.8 \uparrow 9.7	21.8 \uparrow 14.4
		31.8 \uparrow 21.2	29.2 \uparrow 19.4	11.0 \uparrow 4.1	12.7 \uparrow 8.3	16.8 \uparrow 9.4	13.1 \uparrow 8.0	19.1 \uparrow 11.7
		23.7 \uparrow 13.1	24.0 \uparrow 14.2	10.4 \uparrow 3.5	11.1 \uparrow 6.7	12.1 \uparrow 4.7	11.2 \uparrow 6.1	15.4 \uparrow 8.0
		16.2 \uparrow 5.6	21.6 \uparrow 11.8	9.4 \uparrow 2.5	5.2 \uparrow 0.8	10.6 \uparrow 3.2	9.3 \uparrow 4.2	12.1 \uparrow 4.7
		11.1 \uparrow 0.5	14.1 \uparrow 4.3	8.9 \uparrow 2.0	4.4 \uparrow 0.0	9.0 \uparrow 1.6	8.3 \uparrow 3.2	9.3 \uparrow 1.9
Detic [44]		11.5	9.3	7.1	4.7	7.0	4.8	7.4
$\mathcal{D}_{\text{XRAY}}^{\text{in-h}}$ 100/0 80/20 50/50 20/80 0/100 $\mathcal{D}_{\text{RGB}}^{\text{web}}$	+ RAXO	35.3 \uparrow 23.8	27.3 \uparrow 18.0	11.3 \uparrow 4.2	14.0 \uparrow 9.3	19.4 \uparrow 12.4	14.2 \uparrow 9.4	20.3 \uparrow 12.9
		30.7 \uparrow 19.2	23.9 \uparrow 14.6	10.8 \uparrow 3.7	12.3 \uparrow 7.6	18.0 \uparrow 11.0	12.1 \uparrow 7.3	18.0 \uparrow 10.6
		24.4 \uparrow 12.9	19.5 \uparrow 10.2	10.3 \uparrow 3.2	9.2 \uparrow 4.5	14.6 \uparrow 7.6	11.0 \uparrow 6.2	14.8 \uparrow 7.4
		16.4 \uparrow 4.9	15.2 \uparrow 5.9	9.6 \uparrow 2.5	8.0 \uparrow 3.3	12.7 \uparrow 5.7	9.9 \uparrow 5.1	12.0 \uparrow 4.6
		11.9 \uparrow 0.4	13.4 \uparrow 4.1	9.1 \uparrow 2.0	5.2 \uparrow 0.5	9.4 \uparrow 2.4	7.9 \uparrow 3.1	9.5 \uparrow 2.1
CoDet [21]		8.4	7.3	5.7	3.1	5.6	3.4	5.6
$\mathcal{D}_{\text{XRAY}}^{\text{in-h}}$ 100/0 80/20 50/50 20/80 0/100 $\mathcal{D}_{\text{RGB}}^{\text{web}}$	+ RAXO	35.8 \uparrow 27.4	27.9 \uparrow 20.6	10.3 \uparrow 4.6	14.8 \uparrow 11.7	17.6 \uparrow 12.0	13.2 \uparrow 9.8	19.9 \uparrow 14.3
		32.2 \uparrow 23.8	25.1 \uparrow 17.8	9.5 \uparrow 3.8	12.0 \uparrow 8.9	15.4 \uparrow 9.8	11.7 \uparrow 8.3	17.7 \uparrow 12.1
		24.0 \uparrow 15.6	20.0 \uparrow 12.7	9.5 \uparrow 3.8	9.2 \uparrow 6.1	11.5 \uparrow 5.9	9.9 \uparrow 6.5	14.0 \uparrow 8.4
		17.8 \uparrow 9.4	14.8 \uparrow 7.5	8.5 \uparrow 2.8	5.1 \uparrow 2.0	9.4 \uparrow 3.8	8.1 \uparrow 4.7	10.6 \uparrow 5.0
		12.2 \uparrow 3.8	11.5 \uparrow 4.2	8.1 \uparrow 2.4	4.0 \uparrow 0.9	6.9 \uparrow 1.3	6.5 \uparrow 3.1	8.2 \uparrow 2.6

Table 2. **X-ray OvOD performance under the Cross-Modality Transfer Evaluation (CMTE) setting** on DET-COMPASS (ours), PIXray [19], PIDray [34], CLCXray [42], DvXray [20], and HiXray [30] datasets. We integrate RAXO into different baselines using different gallery \mathcal{G} compositions, from using only $\mathcal{D}_{\text{XRAY}}^{\text{in-house}}$ data (100/0) to exclusively $\mathcal{D}_{\text{RGB}}^{\text{web}}$ samples (0/100). RAXO consistently improves the performance of all baseline OvOD detectors across every dataset. AP is used.

	PIXray				
	Scissors	Wrench	Battery	Pliers	AP50
OVXD [†]	16.9	46.2	14.6	6.4	21.0
BARON + RAXO	18.2	39.4	8.3	22.0	22.0
CoDet + RAXO	43.6	52.6	9.0	18.6	30.9
Detic + RAXO	46.3	60.5	7.9	36.6	37.8
VLDet + RAXO	48.3	62.7	10.6	38.5	40.0
G-DINO + RAXO	49.7	61.6	11.3	51.5	43.5

Table 3. **Comparison under OVXD [16] setting**, where the methods do not have access to in-domain data from the displayed categories. [†]OVXD is a supervised method explicitly trained on X-ray images for alignment. AP50 is used.

RAXO surpasses training-based approaches. We also compare RAXO with OVXD [16], the only concurrent work that extends X-ray detection beyond base categories. OVXD is a *fully-supervised* method trained directly on X-ray data, and its generalization capabilities are evaluated on a hold-out set comprising only four categories (*i.e.*, Scissors, Wrench, Battery, and Pliers). In Tab. 3, we follow this setting and evaluate RAXO on PIXray [19], excluding the aforementioned classes from the in-house dataset $\mathcal{D}_{\text{XRAY}}^{\text{in-house}}$. Our results demonstrate that RAXO outperforms OVXD even when both approaches utilize the same BARON detector [37]. Notably, unlike OVXD, our method does not retrain BARON or any other component. Yet, our training-free RAXO performs remarkably well, achieving an AP50

	Mod.	Filt.	Trans.	AP	AP50	AP75
G-DINO [18]				12.9	14.9	13.4
+ RAXO	X-ray			7.9 \downarrow 5.0	9.8 \downarrow 5.1	8.3 \downarrow 5.1
	RGB			13.9 \uparrow 1.0	17.4 \uparrow 2.5	14.3 \uparrow 0.9
	RGB	✓		14.3 \uparrow 1.4	17.6 \uparrow 2.7	14.7 \uparrow 1.3
	RGB	✓	\mathcal{S}	14.6 \uparrow 1.7	18.3 \uparrow 3.4	15.1 \uparrow 1.7
	RGB	✓	\mathcal{M}	16.1\uparrow3.2	19.8\uparrow4.9	16.8\uparrow3.4

Table 4. **Ablation of web-based retrieval** on PIXray [19] using only web-based samples (setting 0/100). **Mod.** indicates retrieval of a specific modality from the web, **Filt.** refers to filtering retrieved images with \mathcal{F}_{RGB} , and **Trans.** specifies style-transfer via: \mathcal{S} a diffusion-based method [7] or \mathcal{M} our material-transfer mechanism.

of 43.5 when paired with G-DINO [18].

6.2. Ablation Study

In this section we study the core components of RAXO. We conduct our experiments on PIXray [19], using GroundingDINO [13] as our baseline OvOD. Consistent findings are reported in Supp. E and H.

Web-based samples. Tab. 4 presents an ablation study on the components of our web-based visual sample acquisition pipeline (Sec. 5.1). Notably, directly retrieving X-ray samples from the web (Mod. = X-ray) leads to a \downarrow 5.0 point decrease in AP. In contrast, retrieving RGB images from the web (Mod. = RGB) provides a modest AP improvement of

	Ω	x^{neg}	\mathcal{X}	DCC	AP	AP50	AP75
G-DINO [18]					12.9	14.9	13.4
+ RAXO	✓				25.2 \uparrow 12.3	31.4 \uparrow 16.5	26.2 \uparrow 12.8
	✓	✓			27.1 \uparrow 14.2	33.6 \uparrow 18.7	28.4 \uparrow 15.0
	✓		✓		27.8 \uparrow 14.9	34.5 \uparrow 19.6	29.3 \uparrow 15.9
	✓	✓	✓		27.5 \uparrow 14.6	34.0 \uparrow 19.1	28.8 \uparrow 15.4
	✓			✓	28.5 \uparrow 15.6	35.3 \uparrow 20.4	29.9 \uparrow 16.5
	✓	✓	✓	✓	36.9\uparrow24.0	45.0\uparrow30.1	39.0\uparrow25.6

Table 5. **Ablation of class representation construction** on PIXray [19] using in-domain samples (setting 100/0). Ω indicates building positive prototypes using segmentation masks. x^{neg} indicates use of negative prototypes. \mathcal{X} indicates ensemble multiple prototypes in class descriptors. DCC indicates use of prototype consistency criterion.

$\uparrow 1.0$ point. We hypothesize that directly using X-ray examples obtained from the web is ineffective due to the limited availability of high-quality X-ray images online. Filtering the retrieved RGB images using \mathcal{F}_{RGB} increases AP by an additional $\uparrow 0.4$ points. Nonetheless, they still exhibit a significant visual gap compared to their X-ray counterparts. A style-transfer approach using StyleShot [7] (Trans. = \mathcal{S}) provides a marginal AP gain of $\uparrow 0.3$ points, likely due to its limited capacity to understand the intrinsic material properties of the objects. Conversely, our proposed material-transfer mechanism (Trans. = \mathcal{M}) yields a substantial AP improvement of $\uparrow 1.8$ points, underscoring its effectiveness in bridging the X-ray modality gap.

Class descriptor modeling. Tab. 5 presents an ablation study analyzing the impact of the various components used to construct class representations. As shown, using a simple per-class average of all the features from each sample already results in a significant AP improvement of $\uparrow 12.3$ points over the baseline OvOD. Refining the representation by applying an object segmentation method to isolate the foreground features (Ω) further boosts the AP by $\uparrow 1.9$ points. Moreover, leveraging background features to construct negative prototypes (x^{neg}) adds an additional $\uparrow 0.7$ points.

Replacing the averaged class prototypes with our proposed visual descriptor (\mathcal{X}) yields another $\uparrow 0.7$ point improvement, demonstrating its enhanced capability to capture intra-class visual variability. Finally, incorporating our descriptor consistency criterion to suppress low-quality proposals (DCC) provides a further AP increase of $\uparrow 8.4$ points, underscoring the effectiveness of our proposed framework.

Hyperparameter study. Fig. 4 analyzes the effect of varying the number of samples K used to construct the visual descriptors. Even with just a single sample per class, RAXO improves AP by $\uparrow 6.5$ points over the baseline. As K increases, performance continues to improve, reaching saturation at $K = 30$ with a total AP gain of $\uparrow 24.0$ points.

Furthermore, Fig. 5 explores the influence of the threshold σ in the *Descriptor Consistency Module*. Higher values of σ impose stricter consistency requirements, de-

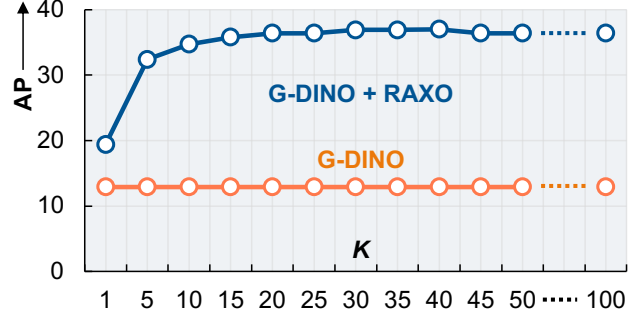


Figure 4. **Impact of K in class representations** evaluated on PIXray [19] using a G-DINO [18] baseline in the 100/0 setting.

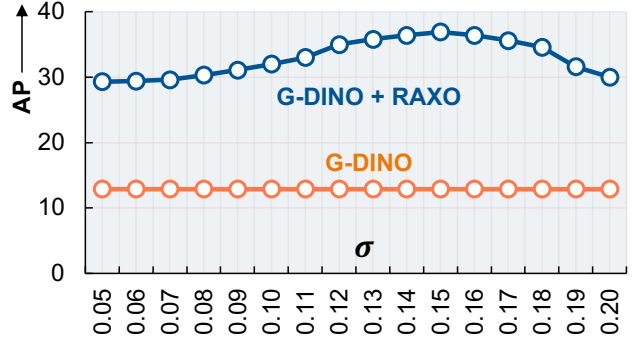


Figure 5. **Impact of σ in Descriptor Consistency Module** on PIXray [19] using a G-DINO [18] baseline in the 100/0 setting.

manding greater separation between the predicted class and other classes before accepting a proposal. We observe that $\sigma = 0.15$ provides an optimal balance between false positives and false negatives, yielding an AP of 36.9 points.

7. Conclusions

In this work, we pioneered the task of open-vocabulary object detection (OvOD) for X-ray imagery, a challenge shaped by the unique characteristics of X-ray data and the scarcity of annotated examples. To address this, we introduced RAXO, a training-free method that repurposes off-the-shelf RGB-based OvOD detectors for the X-ray domain. By leveraging intra-modal feature distances, a novel material-transfer mechanism, and robust class descriptor modeling, RAXO effectively bridges the modality gap between RGB and X-ray imagery. Our extensive experiments, conducted across multiple benchmarks and the newly proposed DET-COMPASS dataset, demonstrate that RAXO consistently enhances detection performance, achieving average mAP improvements of up to 17.0 points over baseline methods.

Looking ahead, the modularity of RAXO opens promising avenues for future research, including further refinement of visual descriptors and adaptation to additional modalities. Overall, our work not only sets a new state-of-the-art for X-ray open-vocabulary detection but also lays the groundwork for a thriving research direction in this emerging domain.

Acknowledgements

We thank CINECA and the ISCRA initiative for the availability of high-performance computing resources. This work was partially supported by the EU HORIZON IAMI (HORIZON-CL3-2023-FCT-01-04-101168272) project, the EU HORIZON ELIAS (HORIZON-CL4-2022-HUMAN-02-101120237) project, the EU HORIZON ELLIOT (HORIZON-CL4-2024-HUMAN-03-101214398) project, the MUR PNRR FAIR (PE00000013) project funded by the NextGenerationEU, the Spanish Ministerio de Ciencia e Innovación (grant numbers PID2020-112623GB-I00, PID2023-149549NB-I00), and the Galician Consellería de Cultura, Educación e Ordenación Universitaria (2024-2027 ED431G-2023/04). Some of these grants are co-funded by the European Regional Development Fund (ERDF). Pablo Garcia-Fernandez is supported by the Spanish Ministerio de Universidades under the FPU national plan (grant number FPU21/05581).

References

- [1] Samet Akcay and Toby Breckon. Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. *Pattern Recognition*, 122:108245, 2022. 2, 3
- [2] An Chang, Yu Zhang, Shunli Zhang, Leisheng Zhong, and Li Zhang. Detecting prohibited objects with physical size constraint from cluttered x-ray baggage images. *Knowledge-Based Systems*, 2022. 1
- [3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6
- [4] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. Lami-detr: Open-vocabulary detection with language model instruction. In *ECCV*, 2024. 4, 5
- [5] Ruohuan Fang, Guansong Pang, and Xiao Bai. Simple image-level classification improves open-vocabulary object detection. In *AAAI*, 2024. 2
- [6] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Prompt-det: Towards open-vocabulary detection using uncurated images. In *ECCV*, 2022. 2
- [7] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv:2407.01414*, 2024. 4, 7, 8, 5
- [8] Google LLC. Custom search JSON API reference. <https://developers.google.com/custom-search/v1>, 2024. Accessed: 2025-03-07. 6, 3
- [9] Lewis D. Griffin, Matthew Caldwell, and Jerone T. A. Andrews. COMPASS-XP. Zenodo, 2019. 3, 1
- [10] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *ECCV*, 2024. 5
- [11] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vm: Open-vocabulary object detection upon frozen vision and language models. *arXiv:2209.15639*, 2022. 2
- [12] Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *ICCV*, 2023. 2
- [13] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 2, 7
- [14] Mingyuan Li, Tong Jia, Hao Wang, Bowen Ma, Hui Lu, Shuyang Lin, Da Cai, and Dongyue Chen. Ao-detr: Anti-overlapping detr for x-ray prohibited items detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 1
- [15] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *ICLR*, 2023. 1, 2, 6, 7, 5
- [16] Shuyang Lin, Tong Jia, Hao Wang, Bowen Ma, Mingyuan Li, and Dongyue Chen. Detection of novel prohibited item categories for real-world security inspection. *Eng. Appl. Artif. Intell.*, 2025. 1, 2, 7
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6
- [18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024. 1, 2, 6, 7, 8, 4, 5
- [19] 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. *TMM*, 2022. 2, 3, 6, 7, 8, 4, 5
- [20] 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 TIFS*, 2024. 1, 3, 6, 7, 5
- [21] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. CoDet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In *NeurIPS*, 2023. 2, 6, 7, 5
- [22] 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 *CVPR*, 2019. 2
- [23] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022. 2
- [24] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 6

- [25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafranec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. 6
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3
- [27] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024. 6
- [28] Haifeng Sima, Bailiang Chen, Chaosheng Tang, Yudong Zhang, and Junding Sun. Multi-scale feature attention-detection transformer: Multi-scale feature attention for security check object detection. *IET Computer Vision*, 2024. 2
- [29] Archana Singh and Dhiraj. Advancements in machine learning techniques for threat item detection in x-ray images: a comprehensive survey. *Int. J. Multim. Inf. Retr.*, 2024. 1
- [30] 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 *ICCV*, 2021. 2, 3, 6, 7, 5
- [31] 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 *CVPR*, 2022. 2, 3
- [32] 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 *ACMMM*, 2022. 1, 3
- [33] Divya Velayudhan, Taimur Hassan, Ernesto Damiani, and Naoufel Werghi. Recent advances in baggage threat detection: A comprehensive and systematic survey. *ACM Computing Surveys*, 55(8):1–38, 2022. 2, 3
- [34] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *ICCV*, 2021. 2, 3, 6, 7, 5
- [35] Ruxue Wang, Yuliang Shi, and Mingyu Cai. Optimization and research of suspicious object detection algorithm in x-ray image. In *2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 2023. 2
- [36] 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 *ACMMM*, 2020. 2
- [37] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023. 2, 7
- [38] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *ICLR*, 2024. 2
- [39] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, 2023. 2
- [40] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *NeurIPS*, 2022. 2
- [41] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *CVPR*, 2023. 2
- [42] 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 TIFS*, 2022. 2, 3, 6, 7, 5
- [43] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 2
- [44] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 1, 2, 3, 6, 7, 5
- [45] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE TPAMI*, 2024. 3, 6