Superpowering Open-Vocabulary Object Detectors for X-ray Vision*

Pablo Garcia-Fernandez[1], Lorenzo Vaquero[1,2], Mingxuan Liu[3], Feng Xue[3], Daniel Cores[1], Nicu Sebe[3], Manuel Mucientes[1], and Elisa Ricci[2,3]

[1] University of Santiago de Compostela, Spain
{pablogarcia.fernandez, daniel.cores, manuel.mucientes}@usc.es
[2] Fondazione Bruno Kessler, Italy
lvaquerootal@fbk.eu
[3] University of Trento, Italy
{mingxuan.liu, feng.xue, niculae.sebe, e.ricci}@unitn.it

Abstract. This keynote paper summarizes **RAXO**, a work published in the IEEE/CVF International Conference on Computer Vision 2025. 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.

Keywords: Open-vocabulary object detection · X-ray imaging · Security screening · Training-free adaptation

1 Introduction

X-ray object detection is a key technology for security screening in high-risk environments. However, most existing detectors are trained in a closed-set setting and can only recognize a limited number of annotated categories [4]. This restriction is especially problematic in X-ray imaging, where collecting and annotating large-scale datasets requires specialized equipment and expert supervision.

* Original paper: *Superpowering Open-Vocabulary Object Detectors for X-ray Vision*, IEEE International Conference on Computer Vision, 2025. ICCV is ICORE A* conf.

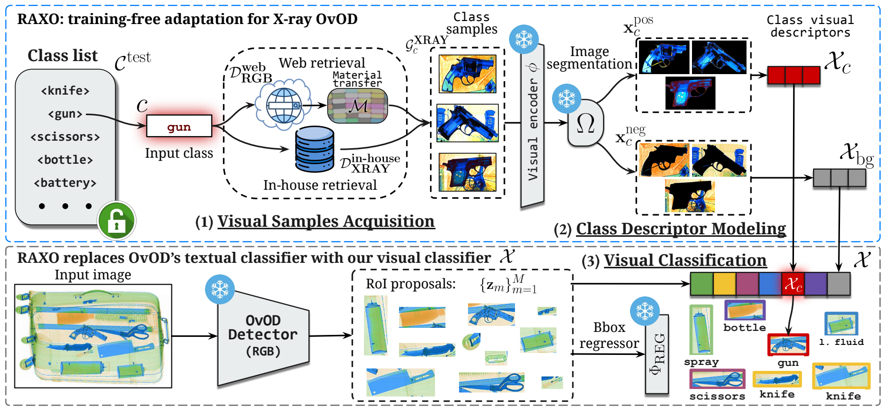

Fig. 1. Overview of RAXO.

Open-vocabulary object detection offers a promising alternative by allowing detectors to recognize categories specified at inference time. While this paradigm has achieved strong results in RGB images [10, 2, 3], transferring it to X-ray imagery remains challenging because of the strong modality gap between natural and X-ray images.

In this work, we propose **RAXO** (t**RA**ining-free adaptation for **X**-ray **O**pen-vocabulary detection), a training-free approach that adapts RGB OvOD detectors to the X-ray domain without additional training. We also present **DET-COMPASS**, a large-scale benchmark for evaluating open-vocabulary detection in X-ray imagery. Together, these contributions advance flexible and scalable object detection for security screening scenarios.

2 Method

As shown in Fig. 1, RAXO adapts RGB OvOD detectors to X-ray imagery without training. Given a test vocabulary $\mathcal{C}^{test}$, it constructs a visual descriptor $\mathcal{X}_c$ for each class c and uses these descriptors as classifiers for region proposals generated by an off-the-shelf detector.

Visual sample acquisition. For each class c, RAXO first retrieves K X-ray samples from an in-house database when available. Since the vocabulary is open-ended, some classes may not exist in this database. In such cases, RAXO retrieves RGB images from the web using the class name as a query and filters them with an RGB OvOD detector. To bridge the RGB–X-ray gap, we introduce a material-transfer mechanism: an LLM associates each class with a material, and the corresponding X-ray appearance is estimated from in-house X-ray samples. This appearance is then rendered inside the object mask of the web image, producing synthetic X-ray-style samples.

Class descriptor modeling. Given the X-ray-style gallery $\mathcal{G}_c^{XRAY}$, RAXO extracts patch-level features using a visual encoder ϕ and segments the object with

Ω. For each sample $\mathbf{I}$, foreground and background prototypes are computed as

$$\mathbf{x}_{\mathbf{I}}^{pos} = \frac{\sum \zeta(\Omega(\mathbf{I})) \odot \phi(\mathbf{I})}{\sum \zeta(\Omega(\mathbf{I}))}, \qquad \mathbf{x}_{\mathbf{I}}^{neg} = \frac{\sum (1 - \zeta(\Omega(\mathbf{I}))) \odot \phi(\mathbf{I})}{\sum (1 - \zeta(\Omega(\mathbf{I})))}. \tag{1}$$

The final class descriptor combines the average foreground prototype with all individual foreground prototypes:

$$\mathcal{X}_c = \left[\bar{\mathbf{x}}_c^{pos}, \{\mathbf{x}_{\mathbf{I}}^{pos} \mid \mathbf{I} \in \mathcal{G}_c^{XRAY}\}\right]. \tag{2}$$

A global background descriptor $\mathcal{X}_{bg}$ is built from negative prototypes.

Visual classification. For a proposal feature $\mathbf{z}_m$, RAXO replaces the original text classifier with nearest visual-descriptor matching:

$$\hat{c} = \underset{c \in \mathcal{C}^{test} \cup \{bg\}}{\arg\max} \; \max_{\mathbf{x} \in \mathcal{X}_c} \langle \mathbf{z}_m, \mathbf{x} \rangle. \tag{3}$$

Proposals classified as background are discarded. We further apply a descriptor consistency criterion that suppresses proposals whose similarity to the predicted class is not sufficiently larger than their similarity to other classes.

3 DET-COMPASS

Existing X-ray detection datasets have few annotated categories, limiting open-vocabulary evaluation. We introduce **DET-COMPASS**, a benchmark derived from COMPASS-XP [1], with manual bounding boxes for 370 object categories and pixel-aligned RGB/X-ray pairs for multimodal evaluation.

4 Experiments

Setup. We follow a **Cross-Modality Transfer Evaluation** protocol, training detectors on RGB data and testing them directly on X-ray datasets without fine-tuning. RAXO is evaluated with GroundingDINO [3], Detic [10], VLDet [2], and CoDet [6] on PIXray [5], PIDray [8], CLCXray [9], DvXray [4], HiXray [7], and DET-COMPASS.

Results. As shown in Tab. 1, RGB OvOD detectors perform poorly when directly transferred to X-ray imagery. RAXO consistently improves all evaluated detectors and datasets by replacing text-based classification with X-ray visual descriptors. Even in the fully web-based setting, where no in-domain examples of the target classes are available, RAXO improves AP by 2.1 points on average. In the fully in-domain setting, RAXO achieves an average improvement of 14.7 AP points. These results show that intra-modal visual matching is substantially more reliable than RGB text-to-X-ray feature alignment.

Acknowledgments. This work was partially supported by EU HORIZON IAMI, ELIAS, and ELLIOT; MUR PNRR FAIR/NextGenerationEU; Spanish grants PID2020-112623GB-I00, PID2023-149549NB-I00 and FPU21/05581; and Galician Consellería grant ED431G-2023/04, with some support co-funded by ERDF. **The authors have no competing interests to declare.**

$\mathcal{G}$		Method	D-COMPASS	PIXray	PIDray	CLCXray	DvXray	HiXray	Avg.
		G-DINO [3]	13.4	12.9	10.9	6.7	10.0	7.0	10.2
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}} \downarrow \mathcal{D}^{\text{web}}_{\text{RGB}}$	*100/0*	+ **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** $_{\uparrow 17.0}$
	80/20	+ **RAXO**	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 $_{\uparrow 13.8}$
	50/50	+ **RAXO**	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 $_{\uparrow 9.6}$
	20/80	+ **RAXO**	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 $_{\uparrow 4.9}$
	0/100	+ **RAXO**	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 [2]	10.6	9.8	6.9	4.4	7.4	5.1	7.4
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}} \downarrow \mathcal{D}^{\text{web}}_{\text{RGB}}$	*100/0*	+ **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** $_{\uparrow 14.4}$
	80/20	+ **RAXO**	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 $_{\uparrow 11.7}$
	50/50	+ **RAXO**	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 $_{\uparrow 8.0}$
	20/80	+ **RAXO**	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 $_{\uparrow 4.7}$
	0/100	+ **RAXO**	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 [10]	11.5	9.3	7.1	4.7	7.0	4.8	7.4
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}} \downarrow \mathcal{D}^{\text{web}}_{\text{RGB}}$	*100/0*	+ **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** $_{\uparrow 12.9}$
	80/20	+ **RAXO**	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 $_{\uparrow 10.6}$
	50/50	+ **RAXO**	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 $_{\uparrow 7.4}$
	20/80	+ **RAXO**	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 $_{\uparrow 4.6}$
	0/100	+ **RAXO**	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 [6]	8.4	7.3	5.7	3.1	5.6	3.4	5.6
$\mathcal{D}^{\text{in-h}}_{\text{XRAY}} \downarrow \mathcal{D}^{\text{web}}_{\text{RGB}}$	*100/0*	+ **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** $_{\uparrow 14.3}$
	80/20	+ **RAXO**	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 $_{\uparrow 12.1}$
	50/50	+ **RAXO**	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 $_{\uparrow 8.4}$
	20/80	+ **RAXO**	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 $_{\uparrow 5.0}$
	0/100	+ **RAXO**	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 1. X-ray OvOD performance under the CMTE setting

References

1. Griffin, L.D., Caldwell, M., Andrews, J.T.A.: COMPASS-XP. Zenodo (May 2019). https://doi.org/10.5281/zenodo.2654887, `https://doi.org/10.5281/zenodo.2654887`
2. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. In: ICLR (2023)
3. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In: ECCV (2024)
4. Ma, B., Jia, T., Li, M., Wu, S., Wang, H., Chen, D.: Towards dual-view x-ray baggage inspection: A large-scale benchmark and adaptive hierarchical cross refinement for prohibited item discovery. IEEE TIFS (2024)
5. Ma, B., Jia, T., Su, M., Jia, X., Chen, D., Zhang, Y.: Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake. TMM (2022)
6. Ma, C., Jiang, Y., Wen, X., Yuan, Z., Qi, X.: CoDet: Co-occurrence guided region-word alignment for open-vocabulary object detection. In: NeurIPS (2023)
7. Tao, R., Wei, Y., Jiang, X., Li, H., Qin, H., Wang, J., Ma, Y., Zhang, L., Liu, X.: Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In: ICCV (2021)
8. Wang, B., Zhang, L., Wen, L., Liu, X., Wu, Y.: Towards real-world prohibited item detection: A large-scale x-ray benchmark. In: ICCV (2021)
9. Zhao, C., Zhu, L., Dou, S., Deng, W., Wang, L.: Detecting overlapped objects in x-ray security imagery by a label-aware mechanism. IEEE TIFS (2022)
10. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV (2022)