X-ray illicit object detection using hybrid CNN-transformer neural network architectures

Jorgen Cani*, Christos Diou*, Spyridon Evangelatos†, Panagiotis Radoglou-Grammatikis‡§, Vasileios Argyriou¶,
Panagiotis Sarigiannidis‡, Iraklis Varlamis* and Georgios Th. Papadopoulos*

* Department of Informatics and Telematics, Harokopio University of Athens, Athens, Greece
† Research & Innovation Development Department, Netcompany-Intrasoft S.A., Luxembourg, Luxembourg
‡ Department of Electrical and Computer Engineering, University of Western Macedonia, Kozani, Greece
§ K3Y Ltd, Sofia, Bulgaria
¶ Department of Networks and Digital Media, Kingston University, London, United Kingdom
Email: *{cani,cdiou,varlamis,g.th.papadopoulos}@hua.gr, †sevangelatos@netcompany.com,
‡Vasileios.Argyriou@kingston.ac.uk, §psarigiannidis@uowm.gr, ¶pradoglou@k3y.bg

Abstract—In the field of X-ray security applications, even the smallest details can significantly impact outcomes. Objects that are heavily occluded or intentionally concealed pose a great challenge for detection, whether by human observation or through advanced technological applications. While certain Deep Learning (DL) architectures demonstrate strong performance in processing local information, such as Convolutional Neural Networks (CNNs), others excel in handling distant information, e.g., transformers. In X-ray security imaging the literature has been dominated by the use of CNN-based methods, while the integration of the two aforementioned leading architectures has not been sufficiently explored. In this paper, various hybrid CNN-transformer architectures are evaluated against a common CNN object detection baseline, namely YOLOv8. In particular, a CNN (HGNetV2) and a hybrid CNN-transformer (Next-ViT-S) backbone are combined with different CNN/transformer detection heads (YOLOv8 and RT-DETR). The resulting architectures are comparatively evaluated on three challenging public X-ray inspection datasets, namely EDS, HiXray, and PIDray. Interestingly, while the YOLOv8 detector with its default backbone (CSP-DarkNet53) is generally shown to be advantageous on the HiXray and PIDray datasets, when a domain distribution shift is incorporated in the X-ray images (as happens in the EDS datasets), hybrid CNN-transformer architectures exhibit increased robustness. Detailed comparative evaluation results, including object-level detection performance and object-size error analysis, demonstrate the strengths and weaknesses of each architectural combination and suggest guidelines for future research. The source code and network weights of the models employed in this study are available at https://github.com/jgenc/xray-comparative-evaluation.

Index Terms—Object detection, X-ray imaging, convolutional neural networks, vision transformers, hybrid architectures

I. INTRODUCTION

Automated X-ray screening for prohibited item detection constitutes a crucial task for public safety and yet remains open to several challenges, such as object occlusion, cluttered

scenes, high intra-class variation, and the inherent visual ambiguity of grayscale X-ray imagery [1]. Under these conditions, relying on manual inspection can lead to fatigue and errors in high-throughput environments. In recent years, deep neural networks have driven breakthroughs in generic object detection. Consequently, deep learning methods, such as Convolutional Neural Networks (CNNs), have introduced notable advances in automatic object detection in X-ray images [2] [3] [4].

In parallel, Vision Transformers (ViTs) have shown increased capabilities in modeling global visual context, by attending to whole image patches via multi-head self-attention [5] [6]. In particular, DETR [7] formulates the task of object detection as a set prediction problem. Additionally, sparse DETR [8] employs a Swin transformer backbone and adjusts accordingly token usage across various configurations. Moreover, DINO [9] introduces contrastive denoising, mixed query anchors, and dual look-ahead box prediction, along with test time augmentation [10].

More recently, hybrid CNN-transformer architectures have been introduced, which combine the complementary merits of CNNs and ViTs to address various computer vision tasks more efficiently [11]. In particular, when CNNs' excellence in capturing local detail is combined with ViTs' ability to capture long-range relationships across the visual scene, promising object detection performance can be observed [12]. In this context, Next-ViT-S [13] comprises a leading example that alternates specialized convolution blocks with attention-based ones at each stage and outperforms comparable CNNs and ViTs on standard benchmarks. More specifically, Next-ViT-S retains the inductive biases and speed of convolution, while incorporating global receptive fields through the use of self-attention; Moreover, it is explicitly engineered for efficient deployment, thus being an ideal backbone network for object detection tasks in realistic scenarios.

Despite the rise of ViTs and hybrid CNN-transformer architectures in natural image analysis, the X-ray imaging community still focuses on CNN-based approaches [14]. This

is mainly because ViT architectural components often require large training datasets and can be computationally heavy. To this end, the advantageous characteristics of hybrid CNN-transformer architectures have not yet been sufficiently explored in X-ray security imaging [3].

In this paper, the issue of applying hybrid CNN-transformer architectures for the detection of illicit objects [15] in X-ray inspection imaging is investigated, aiming at examining their performance against the typical CNN-only detection baseline. In particular, various hybrid architectures are formed by combining a CNN (HGNetV2 [16]) and a hybrid CNN-transformer (Next-ViT-S [13]) backbone with different CNN/transformer detection heads (namely, YOLOv8 [17] and RT-DETR [18], respectively). The formed architectures are comparatively evaluated against a common and well-performing CNN-only object detection baseline, namely YOLOv8 with its default backbone (CSP-DarkNet53) [17]. All methods are evaluated on three challenging public X-ray inspection datasets, namely EDS [19], HiXray [20], and PIDray [21]. Compared to HiXray and PIDray, EDS introduces a distributional shift in the collected data, due to the use of three different X-ray scanners during the capturing process, simulating in this way real-world deployment challenges and emphasizing on the evaluation of models' robustness. Interestingly, the YOLOv8 detector performs better in the HiXray and PIDray datasets; however, hybrid architectures are shown to be advantageous in the EDS one. Moreover, detailed comparative evaluation results, also including object-level detection performance and object-size error analysis, demonstrate the pros and cons of each architectural combination and suggest guidelines for future research.

The remainder of the paper is organized as follows: Section II presents an overview of previous work in the field of DL-based X-ray object detection. Section III details the hybrid CNN-transformer neural network architectures considered in this work. Section IV discusses the obtained experimental results and highlights key findings. Section V draws conclusions and outlines future research directions in the field.

II. PREVIOUS WORK

Automatic X-ray security inspection systems typically rely on the use of DL-based object detection methods that, so far, have been predominantly based on CNN architectures. In particular, notable early studies, such as DOAM [22] and LIM [20], adapt popular CNN-based object detectors, to suit the characteristics of the X-ray imaging domain. On another direction, detection heads such as SSD, FCOS, YOLOv3 and YOLOv5 have also been explored [14], aiming at addressing real-time analysis requirements. More elaborate approaches follow the respective advancements in generic CNN-based detection schemes. In particular, EM-YOLO [23] enhances YOLOv7 with X-ray specific pre-processing for handling occlusion, low contrast, and class imbalance. SC-YOLOv8 [24] modifies the backbone to adapt to object position and shape. Additionally, Wang et al. [25] propose two YOLOv8 variants with changes to the neck and head networks, maintaining size

while boosting performance. YOLOv8-GEMA [26] adjusts the backbone and neck for improved results. Moreover, TinyRay [27] uses a YOLOv7 variant with the lightweight FasterNet backbone, while Ren et al. [28] apply distillation to train compact YOLOv4 and RetinaNet models.

Following the application of hybrid CNN-transformer architectures to natural RGB images, the same methods have been used in X-ray-based illicit object detection. In particular, the Trans2ray [29] approach comprises a dual-branch hybrid framework for dual-view X-ray imaging, achieving notable performance against other dual-view detectors. Additionally, the EslaXDET [30] method applies self-supervised training [31] on a ViT and introduces a detection head for creating multi-scale feature maps.

Although hybrid CNN-transformer architectures show promise, they remain less adopted in X-ray inspection compared to conventional CNNs. Their limited uptake is mainly due to challenges like increased computational overhead, which is critical for high-throughput systems. In this context, lightweight adaptations, like TinyViT [32] and Efficient Hybrid DETR [18], explore parameter reduction strategies without compromising performance, while leveraging Neural Architecture Search (NAS) for balancing accuracy and latency aspects. On the other hand, the demand of transformer-based components for increased training datasets (compared to the respective CNN case) introduces further concerns and obstacles. In this respect, the lack of sufficiently large X-ray datasets and the presence of biases in the publicly available ones (e.g. due to the inherent scarcity of security threats in publicly annotated benchmarks), require the development of appropriate techniques, like semi-supervised learning and synthetic data augmentation. Towards this direction, the approaches of Lin et al. [33] and Huang et al. [34] demonstrate how hybrid architectures benefit from cross-domain pretraining and attention-based few-shot learning, respectively, reducing in that way the reliance on large labeled datasets.

III. HYBRID CNN-TRANSFORMER-BASED OBJECT DETECTION

This section details the hybrid CNN-transformer architectures investigated in this work, providing the basis for their performance comparison against a conventional CNN-only detection baseline. In particular, various hybrid architectures are formed by combining recent and well-performing CNN (HGNetV2 [16]) and hybrid CNN-transformer (Next-ViT-S [13]) backbones with different CNN/transformer detection heads (namely, YOLOv8 [17] and RT-DETR [18], respectively). Denoting each formed object detector as D(head, backbone), the considered composite architectures in this work are as follows: D(YOLOv8, Next-ViT-S), D(RT-DETR, HGNetV2), and D(RT-DETR, Next-ViT-S). The latter are comparatively evaluated against a common and well-performing CNN-only object detection baseline, namely D(YOLOv8, CSP-DarkNet53), i.e. the YOLOv8 detector with its default CSP-DarkNet53 backbone [17].

In the following sections, the architectural modifications to the key neural components (namely, detectors YOLOv8 and RT-DETR, and backbone Next-ViT-S) that are required for forming the D(YOLOv8, Next-ViT-S) and D(RT-DETR, Next-ViT-S) object detectors are detailed.

A. Next-ViT backbone network

Next-ViT [13] is a hybrid model combining convolutional and transformer network components, designed so as to achieve an optimal balance between latency and accuracy. It adopts a hierarchical pyramidal architecture, which comprises multiple stages that are adaptable for various downstream tasks, including object detection and segmentation. Next-ViT includes an initial stem stage (incorporating standard convolutional layers), followed by four primary ones (labeled S_1 to S_4), where the feature output size is subsequently reduced by half at each stage. Each primary stage incorporates Next Convolutional Blocks (NCB) and Next Transformer Blocks (NTB), in order to model both short- and long-term dependencies within the input visual data. In the current study, the smallest *Next-ViT-S* configuration is considered for ensuring real-time performance, using model weights[1] pre-trained on the ImageNet dataset.

B. Hybrid YOLOv8-based object detector

In order to form the D(YOLOv8, Next-ViT-S) object detector, the CSP-DarkNet53 backbone, included in the standard YOLOv8 architecture, is substituted by the Next-ViT-S one. This requires modifications to the connections of the original YOLOv8 neck network, in order to handle the varying feature map sizes and spatial resolutions of the intermediate layers of the Next-ViT-S backbone.

The default YOLOv8 implementation (with the CSPDarkNet-53 backbone) utilizes multi-scale feature maps that are organized in five stages, labeled P_1 to P_5. In order to effectively utilize these features, YOLOv8 incorporates skip connections (known as residual links) from the backbone to the neck module, a process facilitated by the so called Path Aggregation Network (PAFPN); these skip connections merge and upscale feature maps. By convention, the output of each backbone block is referred to as a 'layer', numbered sequentially throughout the network, in order to differentiate from the P_k feature levels. YOLOv8 utilizes connections to layers 4 and 6, corresponding to the outputs of the layers before P_3 and P_4. In order to integrate the Next-ViT-S backbone to the YOLOv8 detector head, three different combinations of skip connections are considered that are denoted $C(x, y)$, where x is the index of the first skip layer and y the next one, and graphically illustrated in Fig. 1:

- $C(10, 20)$: Layer 10 represents the first transformer block (NTB) in S_2, incorporating both basic visual features and broad contextual information. Layer 20, on the other
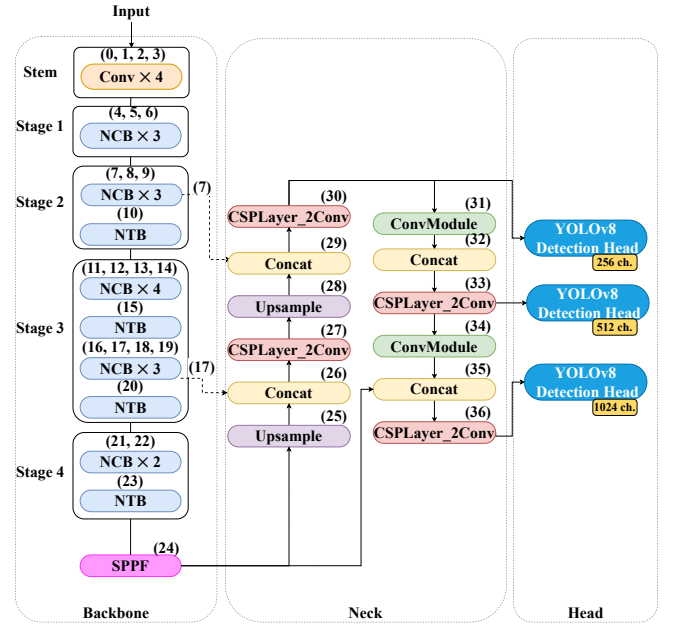
[1]https://github.com/bytedance/Next-ViT?tab=readme-ov-file#image-classification

Figure 1: Architecture of the D(YOLOv8, Next-ViT-S) detector.

hand, is the third NTB block in S_3, which bears more profound features.
- $C(9, 19)$: Layer 9 is the last convolutional block (NCB) in S_2, possessing basic visual information, but lacking comprehensive contextual information. Layer 19 is the last NCB block in S_3, now integrating contextual information from the preceding NTB layers.
- $C(7, 17)$: Layer 7 is the first NCB block in S_2, incorporating features from S_1. Layer 17 is located two layers after the second NTB block, bearing contextual information that has not yet been refined by subsequent NCB layers.

The final output features of the Next-ViT-S backbone are processed by the YOLOv8 SPPF layer (a modified version of the SPP-Network) and then by the original neck network. After extensive experimental evaluation, configuration $C(7, 17)$ was shown to lead to superior detection performance and utilized in all experiments in this study.

C. Hybrid RT-DETR-based object detector

In order to formulate the D(RT-DETR, Next-ViT-S) detector, the default backbone (HGNetV2 [16]) of RT-DETR [18] is replaced by the Next-ViT-S one. The latter requires modifications to the connections of the original RT-DETR neck network towards the backbone for accounting for the different feature map scales and spatial resolutions of Next-ViT-S. In particular, the CNN-only HGNetV2 network comprises a hierarchical architecture with four stages, denoted 'stage 1-4'. Two residual connections are established from the backbone (stages 2 and 3) to the detection head (layers 3 and 7). In order to integrate the Next-ViT-S backbone to the RT-DETR detection head, two skip connections of different feature scales are defined
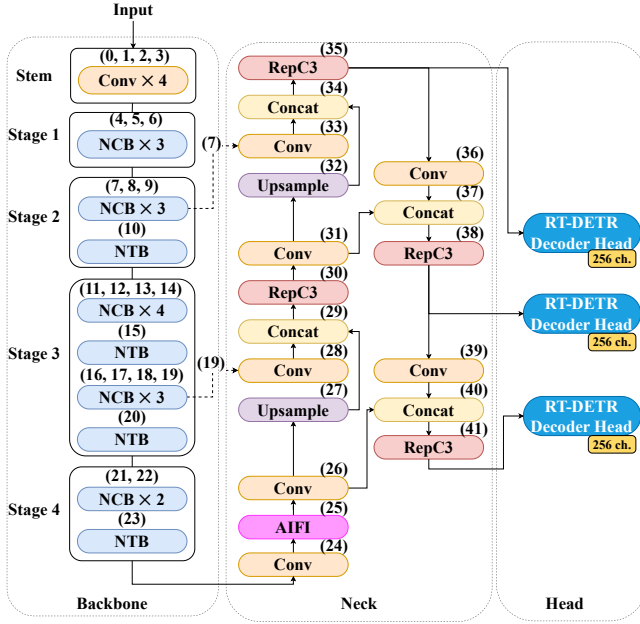
Figure 2: Architecture of the D(RT-DETR, Next-ViT-S) detector.

from the backbone to the neck network. Similarly to III-B, the same skip connections $C(10, 20)$, $C(9, 19)$, and $C(7, 17)$ are again considered. The head network, consisting of three RT-DETR decoder heads, remains unchanged. The overall D(RT-DETR, Next-ViT-S) architecture is illustrated in Fig. 2. Following thorough experimental assessment, configuration $C(9, 19)$ was shown to be the most efficient one and was used in all experiments carried out in this study.

IV. EXPERIMENTAL RESULTS

This section details the defined experimental framework for evaluating the performance of the considered object detectors, while the obtained results, along with their corresponding assessment discussion, are subsequently provided.

A. Experimental framework

1) Datasets: The EDS [19] dataset focuses on the challenge of domain shift that is inherent in X-ray imaging, due to factors like varying parameters across different scanning devices. In particular, three different X-ray scanners are employed, resulting into variations in the captured color, depth and texture information channels, mainly introduced by the different device specs and wear levels. The packages used during the scanning process were artificially prepared. EDS supports ten classes of common daily-life objects, namely *Plastic bottle (DB)*, *Knife (KN)*, *Scissor (SC)*, *Laptop (LA)*, *Umbrella (UM)*, *Lighter (LI)*, *Device (SE)*, *Power bank (PB)*, *Pressure (PR)*, and *Glass bottle (GB)*. The dataset comprises $14,219$ images containing $31,654$ object instances from three domains (X-ray machines), resulting in ~ 2.22 instances per image on average. The defined experimental protocol dictates

the training of a detection model in a single domain and its subsequent evaluation in a different one, resulting in a total of six performed experimental sessions.

The HiXray [20] dataset contains real-world X-ray scans collected from an international airport, where the image annotations were provided by the airport security personnel. The dataset comprises $45,364$ images that include a total of $102,928$ prohibited items, i.e. ~ 2.27 instances per image. The dataset supports eight classes, namely *Portable charger 1 (lithium-ion prismatic cell) (PO1)*, *Portable charger 2 (lithium-ion cylindrical cell) (PO2)*, *Tablet (TA)*, *Mobile phone (MP)*, *Laptop (LA)*, *Cosmetic (CO)*, *Water (WA)*, and *Nonmetallic Lighter (NL)*. HiXray is split into a training (80% of images) and a test (20% of images) set.

The PIDray [21] dataset focuses on deliberately hidden items, mimicking real-world scenarios where prohibited objects are intentionally concealed. The latter fact adds an extra level of complexity to the object detection task, since it is required to identify hidden items (and not 'simply' detecting objects obscured by other items and/or environmental factors). All scan samples are collected under real-world settings, namely at airport, subway, and railway station security checkpoints. PIDray includes twelve classes of prohibited items, namely *Baton (BA)*, *Pliers (PL)*, *Hammer (HA)*, *Power-bank (PB)*, *Scissors (SC)*, *Wrench (WR)*, *Gun (GU)*, *Bullet (BU)*, *Sprayer (SP)*, *Handcuffs (HC)*, *Knife (KN)*, and *Lighter (LI)*. The dataset is split into a training ($29,457$ samples, $\sim 60\%$ of images) and a test ($18,220$ samples, $\sim 40\%$ of images) set. Moreover, the test set is further divided into three subsets, namely an easy (the images contain only one prohibited object), a hard (the images contain more than one illicit items), and a hidden (the images contain deliberately hidden objects) one, with $9,482$, $3,733$, and $5,055$ images, respectively.

2) Performance metrics: Mean Average Precision (mAP) constitutes the most commonly used metric in object detection applications, which estimates a comprehensive and aggregated evaluation of the examined model's performance across different confidence levels and object classes. Among the different variants regarding how mAP is calculated, especially with respect to the selected Intersection over Union (IoU) threshold (that assesses the spatial overlap between the predicted bounding box (generated by a detector model) and the corresponding ground truth one (that defines the actual location of the object) for determining true positive detection, the following ones have been considered in this work: a) mAP^{50}: This refers to the mAP score calculated using an IoU threshold of 0.5. b) $\text{mAP}^{50:95}$: This involves a more rigorous evaluation protocol, which calculates mAP by averaging AP scores over a range of defined IoU thresholds, typically from 0.5 to 0.95 with a step of 0.05, and subsequently averaging the computed results across all object classes. This metric provides a more comprehensive assessment of the model's localization accuracy, by considering its performance at different levels of overlap with the ground truth annotation. In general, higher mAP scores, which receive values ranging from 0 to 1, indicate better performance.

Table I: Object detection results for the EDS, HiXray and PIDray datasets (left column: mAP^{50}, right column: $\text{mAP}^{50:95}$)

Detector	EDS	HiXray	PIDray
D(YOLOv8, CSP-DarkNet53)	0.547 / 0.386	0.845 / **0.564**	0.897 / **0.807**
D(YOLOv8, Next-ViT-S)	0.588 / 0.408	0.841 / 0.551	0.898 / 0.801
D(RT-DETR, HGNetV2)	0.573 / **0.410**	0.839 / 0.510	0.835 / 0.720
D(RT-DETR, Next-ViT-S)	0.504 / 0.322	0.818 / 0.483	0.879 / 0.773

Table II: Object detection results for the various sessions of the EDS dataset (left column: mAP^{50}, right column: $\text{mAP}^{50:95}$)

Detector	$\mathcal{D}_{1\rightarrow2}$	$\mathcal{D}_{1\rightarrow3}$	$\mathcal{D}_{2\rightarrow1}$	$\mathcal{D}_{2\rightarrow3}$	$\mathcal{D}_{3\rightarrow1}$	$\mathcal{D}_{3\rightarrow2}$	Avg.
D(YOLOv8, CSP-DarkNet53)	0.482 / 0.340	0.555 / 0.410	0.454 / 0.295	0.619 / 0.449	0.587 / 0.411	0.590 / 0.411	0.547 / 0.386
D(YOLOv8, Next-ViT-S)	0.512 / 0.347	0.603 / **0.441**	0.515 / 0.341	0.648 / 0.454	0.624 / **0.434**	0.626 / **0.431**	0.588 / 0.408
D(RT-DETR, HGNetV2)	0.506 / **0.352**	0.569 / 0.424	0.506 / **0.350**	0.648 / **0.471**	0.595 / 0.431	0.616 / 0.429	0.573 / **0.410**
D(RT-DETR, Next-ViT-S)	0.446 / 0.292	0.545 / 0.343	0.372 / 0.217	0.450 / 0.286	0.578 / 0.377	0.636 / 0.419	0.504 / 0.322

Table III: Object detection results for the various subsets of the PIDray dataset (left column: mAP^{50}, right column: $\text{mAP}^{50:95}$)

Detector	easy	hard	hidden	overall
D(YOLOv8, CSP-DarkNet53)	0.911 / **0.846**	0.914 / **0.812**	0.797 / 0.682	0.897 / **0.807**
D(YOLOv8, Next-ViT-S)	0.912 / 0.837	0.910 / 0.799	0.803 / **0.685**	0.898 / 0.801
D(RT-DETR, HGNetV2)	0.864 / 0.780	0.864 / 0.724	0.681 / 0.548	0.835 / 0.720
D(RT-DETR, Next-ViT-S)	0.898 / 0.824	0.898 / 0.770	0.779 / 0.646	0.879 / 0.773

3) Implementation details: All reported experiments were conducted using a PC with Ubuntu Linux 22.04 OS, equipped with an Intel Core i9-13900K CPU and two NVIDIA GeForce RTX 4070 Ti GPUs. For the YOLOv8 and RT-DETR detectors the implementations and layer weights available in the Ultralytics framework[2] were used, while for the Next-ViT-S[3] the respective publicly available source code and weights were also utilized. Regarding hyperparameter setting, all detectors were trained using a batch size equal to 18. For the EDS dataset, D(RT-DETR, Next-ViT-S) was trained using Stochastic Gradient Descent (SGD) with a learning rate of 0.01 and a momentum of 0.937, while for the remaining detectors the AdamW optimizer was employed with a learning rate of 0.000714. For the HiXray and PIDray datasets, due to their larger size, all detectors were trained using SGD with a learning rate of 0.01 and a momentum of 0.937. All experiments incorporated an early stopping mechanism to prevent overfitting.

B. Evaluation results and discussion

Table I summarizes the object detection results obtained for all considered detectors for all the datasets employed, where both the mAP^{50} and $\text{mAP}^{50:95}$ metrics are provided. Additionally, Tables II and III present the detailed detection performance for the various experimental sessions and subsets of the EDS and the PIDray datasets, respectively. Object-level performance ($\text{mAP}^{50:95}$ metric) for each dataset is illustrated in Fig. 3. Moreover, Fig. 4 demonstrates object scale-related performance ($\text{mAP}^{50:95}$ metric) for each dataset, where the COCO [35] dataset object scale definitions for 'small',

'medium', and 'large' were considered. Furthermore, indicative detection results of the various detectors are illustrated in Fig. 5.

From the presented results, several critical observations and key insights can be extracted, the most important of which are summarized as follows:

- Overall, the D(YOLOv8, CSP-DarkNet53) detector, i.e. a CNN-only architecture, achieves the best performance (Table I), exhibiting the highest recognition rate in 2 out of the 3 considered datasets (namely, HiXray and PIDray). This demonstrates the increased capability of convolutional operators in modeling appearance patterns in X-ray images. It needs to be highlighted though that both HiXray and PIDray contain data collected from a single scanner each.
- Interestingly, for the EDS dataset, where multiple/different scanners are employed and the defined experimental sessions target the evaluation under domain shifts in the underlying data distributions, most hybrid CNN-transformer detectors outperform the D(YOLOv8, CSP-DarkNet53) one, with D(RT-DETR, HGNetV2) showcasing the highest performance (Table I). The latter suggests that transformer-based components result into increased robustness to data distribution shifts, mainly due to their increased capability in incorporating global contextual information in the extracted features (compared to the respective convolutional-only blocks).
- Across all experiments (Tables I-III), YOLOv8-based detectors with a Next-ViT-S backbone consistently outperform their RT-DETR-based counterparts, demonstrating the superiority of the YOLOv8 detection head over the RT-DETR one. The latter indicates that transformer-based detection heads are not advantageous for the analysis
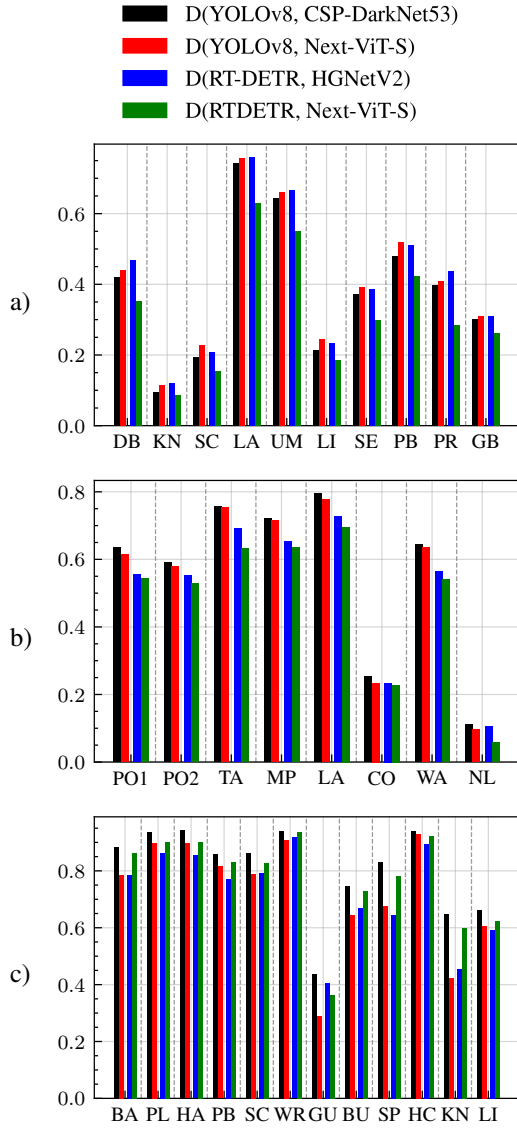
Figure 3: Object-level performance (mAP$^{50:95}$ metric) for datasets: a) EDS, b) HiXray, and c) PIDray.

Figure 4: Object scale-related performance (mAP$^{50:95}$ metric) for datasets: a) EDS, b) HiXray, and c) PIDray.

of X-ray security images, which inherently exhibit no particular spatial structure (i.e. objects in containers are usually positioned without a specific/consistent spatial order).

- Replacing a CNN backbone (either CSP-DarkNet53 and HGNetV2) with a hybrid one (Next-ViT-S) is shown not to always lead to improved performance (Table I), regardless of the employed detection head (either YOLOv8- or RT-DETR-based one). This demonstrates the need for careful design and comprehensive experimentation regarding compatibility aspects, when incorporating a hybrid backbone in an object detection scheme.

- Examining the detection results in the EDS dataset (Table II), it can be seen that the CNN-only D(YOLOv8, CSP-DarkNet53) detector generally leads to inferior perfor-
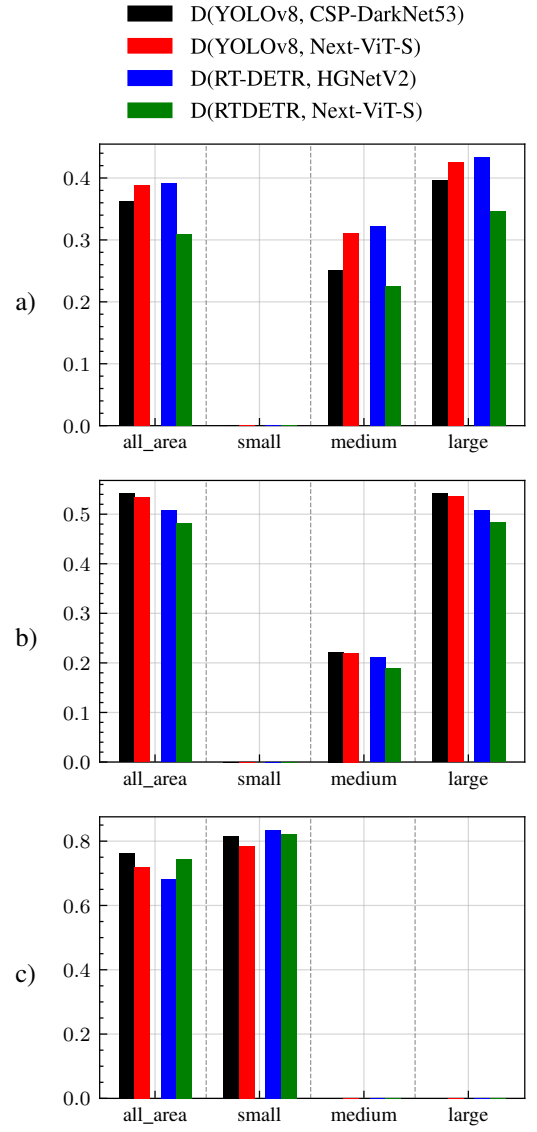
mance in all experimental sessions, compared to most hybrid architectures (with the best D(YOLOv8, Next-ViT-S) and D(RT-DETR, HGNetV2) detectors performing almost equally well), as already discussed and explained above.

- Investigating the performance for the various subsets in the PID dataset (Table III), it can be seen that D(YOLOv8, CSP-DarkNet53) performs best for the 'easy' and 'hard' ones. However, for the 'hidden' partition, where the objects of interest are intentionally concealed and significant occlusions are present, the hybrid D(YOLOv8, Next-ViT-S) detector demonstrates a slightly improved recognition rate.

- Analyzing the performance with respect to individual object types (Fig. 3), it can be observed that the D(YOLOv8,
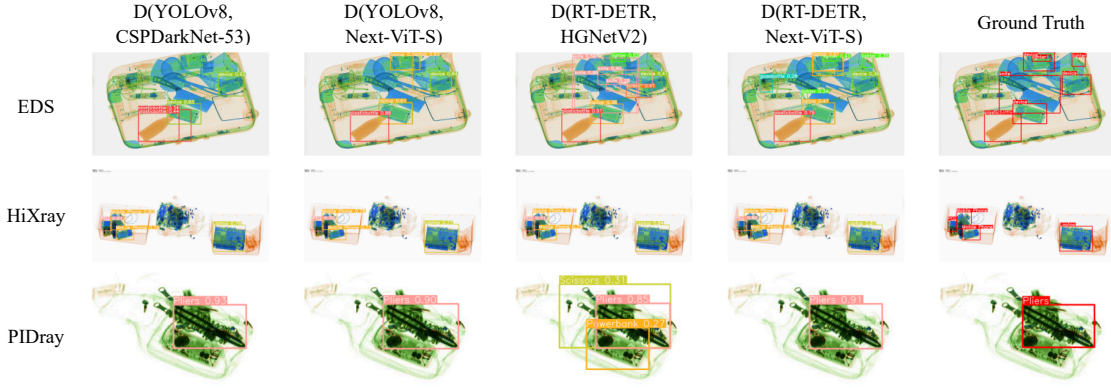
Figure 5: Indicative object detection results in the EDS, HiXray, and PIDray datasets.

CSP-DarkNet53) detector outperforms all remaining ones for all classes in the HiXray and PIDray datasets. However, in the EDS dataset, D(YOLOv8, CSP-DarkNet53) performs inferior for all object types than most hybrid detectors, especially for classes 'Plastic bottle' (DB), 'Scissor' (SC), 'Lighter' (LI), 'Power bank' (PB), and 'Pressure' (PR), i.e. both objects types with fine-grained local patterns (SC, LI, PB, and PR) as well as with broader motifs in their captured chemical composition appearance.

- Examining the impact of the objects' scale (Fig. 4), it can be seen that in the EDS dataset that D(YOLOv8, CSP-DarkNet53) is significantly outperformed by D(YOLOv8, Next-ViT-S) and D(RT-DETR, HGNetV2) for 'medium' size objects, while this difference is decreased for 'large' instances. On the other hand, for the HiXray and PIDray datasets (where D(YOLOv8, CSP-DarkNet53) performs best), the object scale does not seem to significantly affect the performance for all detectors.

V. CONCLUSION

In this paper, various hybrid CNN-transformer architectures were introduced and evaluated against a common CNN object detection baseline, namely YOLOv8. In particular, a CNN (HGNetV2) and a hybrid CNN-transformer (Next-ViT-S) backbone were combined with different CNN/transformer detection heads (YOLOv8 and RT-DETR). The resulting architectures were comparatively evaluated on three challenging public X-ray inspection datasets, namely EDS, HiXray, and PIDray. One of the key observations concerned the fact that while the YOLOv8 detector with its default backbone (CSP-DarkNet53) was shown to be advantageous on the HiXray and PIDray datasets, when a domain distribution shift is incorporated in the X-ray images (EDS datasets), hybrid CNN-transformer architectures demonstrated increased robustness. Future research includes the investigation of additional hybrid CNN-transformer configurations and broader experimental evaluation in additional datasets.

REFERENCES

[1] L. Zhang, L. Jiang, R. Ji, and H. Fan, "Pidray: A large-scale x-ray benchmark for real-world prohibited item detection," *International Journal of Computer Vision*, vol. 131, no. 12, pp. 3170–3192, 2023.

[2] G. Batsis, I. Mademlis, and G. T. Papadopoulos, "Illicit item detection in x-ray images for security applications," in *2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 63–70, IEEE, 2023.

[3] M. Rafiei, J. Raitoharju, and A. Iosifidis, "Computer vision on x-ray data in industrial production and security applications: A comprehensive survey," *Ieee Access*, vol. 11, pp. 2445–2477, 2023.

[4] I. Mademlis, G. Batsis, A. A. R. Chrysochoou, and G. T. Papadopoulos, "Visual inspection for illicit items in x-ray images using deep learning," in *2023 IEEE International Conference on Big Data (BigData)*, pp. 4081–4089, IEEE, 2023.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] J. Cani, I. Mademlis, A. A. R. Chrysochoou, and G. T. Papadopoulos, "Illicit object detection in x-ray images using vision transformers," in *2024 5th International Conference in Electronic Engineering, Information Technology & Education (EEITE)*, pp. 1–6, IEEE, 2024.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, pp. 213–229, Springer, 2020.

[8] B. Roh, J. Shin, W. Shin, and S. Kim, "Sparse detr: Efficient end-to-end object detection with learnable sparsity," *arXiv preprint arXiv:2111.14330*, 2021.

[9] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[10] P. Alimisis, I. Mademlis, P. Radoglou-Grammatikis, P. Sarigiannidis, and G. T. Papadopoulos, "Advances in diffusion models for image data augmentation: A review of methods, models, evaluation metrics and future research directions," *Artificial Intelligence Review*, vol. 58, no. 4, pp. 1–55, 2025.

[11] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq, "A survey of the vision transformers and their cnn-transformer based variants," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 2917–2970, 2023.

[12] W. F. Hendria, Q. T. Phan, F. Adzaka, and C. Jeong, "Combining transformer and cnn for object detection in uav imagery," *ICT Express*, vol. 9, no. 2, pp. 258–263, 2023.

[13] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, and X. Pan, "Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios," *arXiv preprint arXiv:2207.05501*, 2022.

[14] J. Wu, X. Xu, and J. Yang, "Object detection and x-ray security imaging: A survey," *IEEE Access*, vol. 11, pp. 45416–45441, 2023.

[15] I. Mademlis, M. Mancuso, C. Paternoster, S. Evangelatos, E. Finlay, J. Hughes, P. Radoglou-Grammatikis, P. Sarigiannidis, G. Stavropoulos, K. Votis, *et al.*, "The invisible arms race: digital trends in illicit goods trafficking and ai-enabled responses," *IEEE Transactions on Technology and Society*, 2024.

[16] Baidu Paddle Vision Team, *HGNetv2*, 2023.

[17] G. Jocher, J. Qiu, and A. Chaurasia, *Ultralytics YOLO*, Jan. 2023.

[18] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16965–16974, 2024.

[19] R. Tao, H. Li, T. Wang, Y. Wei, Y. Ding, B. Jin, H. Zhi, X. Liu, and A. 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)*, pp. 21157–21167, IEEE, 2022.

[20] R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, and X. 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*, pp. 10923–10932, 2021.

[21] B. Wang, L. Zhang, L. Wen, X. Liu, and Y. Wu, "Towards real-world prohibited item detection: A large-scale x-ray benchmark," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5412–5421, 2021.

[22] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 138–146, 2020.

[23] B. Jing, P. Duan, L. Chen, and Y. Du, "Em-yolo: An x-ray prohibited-item-detection method based on edge and material information fusion," *Sensors*, vol. 23, no. 20, p. 8555, 2023.

[24] L. Han, C. Ma, Y. Liu, J. Jia, and J. Sun, "Sc-yolov8: A security check model for the inspection of prohibited items in x-ray images," *Electronics*, vol. 12, no. 20, p. 4208, 2023.

[25] Z. Wang, X. Wang, Y. Shi, H. Qi, M. Jia, and W. Wang, "Lightweight detection method for x-ray security inspection with occlusion," *Sensors*, vol. 24, no. 3, p. 1002, 2024.

[26] A. Wang, P. Yuan, H. Wu, Y. Iwahori, and Y. Liu, "Improved yolov8 for dangerous goods detection in x-ray security images," *Electronics*, vol. 13, no. 16, p. 3238, 2024.

[27] H. Zhang, W. Teng, X. He, H. Que, and Y. Zhang, "Lightweight prohibited items detection model in x-ray images based on improved yolov7-tiny," *Journal of the Franklin Institute*, vol. 362, no. 1, p. 107421, 2025.

[28] Y. Ren, L. Zhao, Y. Zhang, Y. Liu, J. Yang, H. Zhang, and B. Lei, "Feature knowledge distillation-based model lightweight for prohibited item detection in x-ray security inspection images," *Advanced Engineering Informatics*, vol. 65, p. 103125, 2025.

[29] X. Meng, H. Feng, Y. Ren, H. Zhang, W. Zou, and X. Ouyang, "Transformer-based dual-view x-ray security inspection image analysis," *Engineering Applications of Artificial Intelligence*, vol. 138, p. 109382, 2024.

[30] J. Wu and X. Xu, "Eslaxdet: A new x-ray baggage security detection framework based on self-supervised vision transformers," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107440, 2024.

[31] S. Konstantakos, J. Cani, I. Mademlis, D. I. Chalkiadaki, Y. M. Asano, E. Gavves, and G. T. Papadopoulos, "Self-supervised visual learning in the low-data regime: a comparative evaluation," *Neurocomputing*, vol. 620, p. 129199, 2025.

[32] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *European conference on computer vision*, pp. 68–85, Springer, 2022.

[33] S. Lin, T. Jia, H. Wang, B. Ma, M. Li, and D. Chen, "Detection of novel prohibited item categories for real-world security inspection," *Engineering Applications of Artificial Intelligence*, vol. 144, p. 110110, 2025.

[34] Y. Huang, H. Gao, and X. Li, "Adaptxray: Vision transformer and adapter in x-ray images for prohibited items detection," in *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 402–408, IEEE, 2024.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755, Springer, 2014.