

# Illicit object detection in X-ray images using Vision Transformers

Jorgen Cani

*Department of Informatics and Telematics  
Harokopio University of Athens  
Athens, Greece  
cani@hua.gr*

Ioannis Mademlis

*Department of Informatics and Telematics  
Harokopio University of Athens  
Athens, Greece  
imademlis@hua.gr*

Adamantia Anna Rebolledo Chrysochoou

*Department of Informatics and Telematics  
Harokopio University of Athens  
Athens, Greece  
adamantia.reb@hua.gr*

Georgios Th. Papadopoulos

*Department of Informatics and Telematics  
Harokopio University of Athens  
Athens, Greece  
g.th.papadopoulos@hua.gr*

**Abstract**—Illicit object detection is a critical task performed at various high-security locations, including airports, train stations, subways, and ports. The continuous and tedious work of examining thousands of X-ray images per hour can be mentally taxing. Thus, Deep Neural Networks (DNNs) can be used to automate the X-ray image analysis process, improve efficiency and alleviate the security officers’ inspection burden. The neural architectures typically utilized in relevant literature are Convolutional Neural Networks (CNNs), with Vision Transformers (ViTs) rarely employed. In order to address this gap, this paper conducts a comprehensive evaluation of relevant ViT architectures on illicit item detection in X-ray images. This study utilizes both Transformer and hybrid backbones, such as SWIN and NextViT, and detectors, such as DINO and RT-DETR. The results demonstrate the remarkable accuracy of the DINO Transformer detector in the low-data regime, the impressive real-time performance of YOLOv8, and the effectiveness of the hybrid NextViT backbone.

**Index Terms**—Object Detection, X-rays, Vision Transformers, Deep Neural Networks

## I. INTRODUCTION

The detection of concealed illicit items using X-ray images is a common security procedure [1], [2]. X-ray scanners are a non-invasive way to examine suspicious containers, such as mailed parcels or luggage, in places where security is of utmost importance, such as airport terminals, customs offices, post offices, etc. [3]. The detection process involves X-ray machines directing beams of high-energy radiation towards the object being scanned. The varying densities of materials within the object cause the X-rays to be absorbed at different rates. The resulting high resolution X-ray images, where denser materials appear lighter and less dense materials appear darker, are manually inspected for the presence of illicit items (e.g., firearms) by security officers in real-time [4].

However, X-ray scans present certain shortcomings that perpetrators can exploit for concealing contraband, such as the occlusion of layered objects, a cluttered environment, similarity of different objects, as well as certain material properties which may impact the image appearance [5]. Additionally, heavy traffic during rush hours may mentally overload the security officers, causing poor decisions during manual inspection. Hence, efficient and well-performing automated solutions are necessary for overcoming such issues.

Recent advances in object detection and in the exploitation of multiple modalities [6] have led to promising X-ray image analysis algorithms [7]. Deep Neural Network (DNN) architectures such as Convolutional Neural Networks (CNNs) [8] and Vision Transformers [9] are the state-of-the-art approaches in general object detection, where the DNN analyzes its input image and outputs the bounding boxes/Regions-of-Interests (RoIs) of all objects visible on it, in pixel coordinates, along with the corresponding class labels. However, when it comes to detecting illicit items in X-ray scan images, the more recent and powerful Transformer architectures, which rely on the attention mechanism [10], have not been extensively employed. Relevant literature is still mostly dominated by CNN-based approaches [11] [12] [13], similarly to visual firearms recognition in the RGB domain [14] [15].

In order to address these gaps in the literature, this paper attempts to systematically assess the capabilities of Vision Transformers at the task of illicit item detection in X-ray images. A hybrid architecture that internally combines convolutions and the attention mechanism is also being evaluated. Thus, extensive experimental comparisons are conducted using the SWIN [16] and Next-ViT [17] deep neural backbones, as well as the DINO [18] and RT-DETR [19] detection heads. Additionally, YOLOv8 [20] is included as a baseline CNN-based one-stage object detector. Evaluation is conducted by pretraining on SIXRay [21], i.e., a relevant, large-scale public dataset, and then finetuning the object detectors on CFray, i.e.,

Research was possible due to the funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101073876 (Ceasefire)

a smaller, custom dataset containing X-ray images of parcels containing firearms or related objects.

The remainder of this paper is organized in the following manner. Section II briefly surveys the state-of-the-art regarding illicit item detection in X-ray images using DNNs. Section III presents the specific deep neural architectures that are being evaluated in this paper. Subsequently, Section IV discusses the evaluation datasets, process and results. Finally, Section V draws conclusions from the conducted study.

## II. RELATED WORK

A wide range of the methods have been employed over the years for detecting illicit items in X-ray scan images. The method in [22] adopts a complex CNN pretrained in natural RGB images and finetunes it at the X-ray domain, in an attempt to address the common issue of DNN overfitting to limited data. This is a significant problem in relevant literature, since it is difficult to construct large-scale X-ray image datasets due to the high cost and low availability of X-ray scanners. In addition, the percentage of “positive” objects (meaning parcels or luggage that indeed contain illicit items) is typically very low in such datasets, rendering effective DNN training even more difficult.

Common DNNs for general object detection have repeatedly been evaluated in this domain, including Faster R-CNN [23], Mask R-CNN [24], and RetinaNet [25]. However, the standard approach is to use real-time one-stage object detectors such as YOLO [26] or SSD [27], due to their significantly higher inference speed. LightRay [13], an alternative lightweight object detector, is a modification of YOLOv4 designed to detect small illicit items detection in complex environments. It consists of a MobileNetV3 [28] backbone CNN, a feature enhancement network and a Lightweight Feature Pyramid Network (LFPN) [29], that obtains information about objects at different scales, and it also includes a Convolutional Block Attention Module (CBAM) [30] for refining feature maps using a spatial attention mechanism.

A different approach is taken in [31], where a new mechanism called Foreground and Background Separation (FBS) is proposed to separate illicit items from complex or cluttered backgrounds. This is accomplished by using a feature extraction DNN combined with Spatial Pyramid Pooling (SPP) and a Path Aggregation Network to extract high-level features. These feature maps serve as input to two neural decoders that simultaneously reconstruct the background and foreground. An attention module then directs the model’s focus to the foreground objects.

YOLO-CID [11] is a modified YOLOv5 architecture that focuses on real-time performance. The architecture is enhanced with the Stem [32] and CGhost [33] neural modules. Thus, it achieves accuracy competitive with respect to the baseline, but with a smaller model complexity and higher inference speed.

## III. OBJECT DETECTION METHODS

This section details the selected deep neural architectures, starting from the backbone networks and then moving on to the object detection heads.

### A. Backbone Networks

SWIN [16] is a commonly used hierarchical Vision Transformer that attempts to enhance scale-invariance in image analysis. It is composed of the Swin Transformer block, which replaces the standard multi-head self attention (MSA) block. SWIN is a strong backbone network option for object-detection methods, as seen in [18] and [34]. Four variants of the SWIN Transformer network exist: tiny, small, base and large. This paper utilizes the base configuration of SWIN, consisting of 88 million parameters.

NextViT [17] is a backbone network that aims for a latency/accuracy trade-off. ViTs have great results, but are slow to infer. In contrast, CNNs are fast to infer but inferior to ViTs in terms of performance. NextViT introduces the Next Convolution Block (NCB) and Next Transformer Block (NTB), which, respectively, capture both local and global information. The architecture is primarily designed for industrial/embedded use-cases in mind, such as security inspection devices in airport terminals, subways, etc. It relies on the Next Hybrid Strategy (NHS), which creates four stages of various NCB and NTB blocks, where each stage’s number of blocks depends on the configuration of the model. The available configurations are small, base and large.

This paper utilizes the small variant of NextViT, which is depicted in Figure 1. The figure shows that NextViT-s consists of four stages, with the final layers being the closest to the final output. Overall, there are 24 layers:

- 1) Layers 0-4 are stem layers.
- 2) Layers 4-6 are NCB components, named Stage 1.
- 3) Layers 7-10 contain three NCB components, followed by an NTB component, named Stage 2.
- 4) Layers 11-20 contain four NCB component, followed by an NTB component, which in turn, is followed by four other NCB components and one final NTB component, named Stage 3.
- 5) Layers 21-23 contain two NCB components and one NTB component, named Stage 4, which is the final stage.

### B. Detection Head Networks

Detection Transformer (DETR) [35] is a Transformer-based detector, which is placed after a CNN backbone. It handles object detection as a set prediction task. DETR is not only anchor-free, but also eliminates the need of handcrafted components (e.g., the Non-Maximum Suppression algorithm). DETR with Improved deNoising anchor boxes (DINO) [18] is a DETR-like method that boasts impressive performance gains by accumulating various minor enhancements over the base DETR and by reintroducing the usage of anchor boxes in a Transformer setting. The goal of DINO is to avoid duplicate bounding box outputs that correspond to a single ground-truth object, which is achieved by adding two different types of noise to a single ground-truth RoI during training. Figure 2 depicts the DINO architecture.

Real-Time DETection TRansformer (RT-DETR) [19] is the first real-time end-to-end DETR-like object detector. The

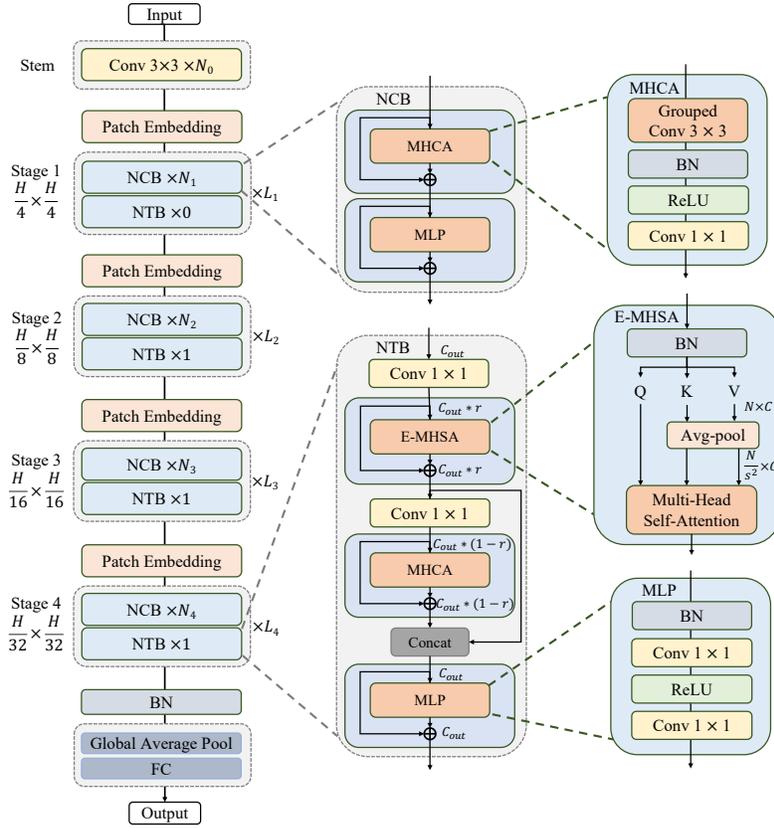


Fig. 1. The architecture of the NextViT-s backbone. Image from [17].

architecture consists of a backbone, the Efficient Hybrid Encoder, an IoU-aware Query selector, and a Decoder & Head final component. The object queries in DETR-like models are a set of learnable embeddings, which are optimized by the decoder component and are mapped to the classification values and bounding boxes by the prediction head. These queries however are hard to interpret and optimize. RT-DETR introduces the IoU score into the objective loss function of the classification branch, which leads to higher quality encoder features. In common implementations, RT-DETR is typically combined with the HGNetV2 CNN backbone network, with residual connections at various stages linking the backbone with the detection head.

**You Only Look Once (YOLO)** [26] is a fast, anchor-based, single-stage object detector that performs object localization and classification using a single CNN. YOLO is an efficient algorithm that has gained popularity due to its speed and accuracy. Its architecture consists of a backbone network, a neck network, and a final prediction head. The latest iteration of the architecture, YOLOv8 [20], is used in this paper.

Many variants of the model exist (nano, small, medium, large, extra-large; n, s, m, l, x); this paper exploits YOLOv8-l. Vanilla YOLOv8 implementations cooperate with a variant of the YOLO-specific DarkNet CNN backbone. Residual connections at various stages are exploited to connect this default CNN backbone to the detection neck and head. There are a few

exceptions in the literature to the privileged use of DarkNet with YOLO. For instance, AG-YOLO [36] leverages NextViT for optimal performance in a citrus fruit detection task, thus better addressing occlusion problems.

#### IV. EXPERIMENTAL EVALUATION

This section details the experimental evaluation datasets and metrics, the exact experimental setup and algorithmic configuration, as well as the obtained results.

##### A. Experimental Datasets

This paper has utilized the public SIXray [21] dataset for pretraining the competing DNNs, before subsequently finetuning them on the smaller, custom CFray dataset.

The SIXray [21] dataset is a publicly available dataset that serves as an object detection benchmark for identifying illicit goods on X-ray scan images of travel luggage. It consists of 1,059,231 X-ray images taken at subway stations, where 8,929 illicit items are annotated. There are 6 classes contained in the dataset: “gun”, “knife”, “wrench”, “pliers” and “scissors”. The final class indicates the absence of illicit objects, named “negative”. The authors split the dataset into three subsets, named SIXray10, SIXray100 and SIXray1000, where the number indicates the ratio of negative to positive images. SIXRay-D [37] is a modified version of the dataset, utilizing a cropping scheme that preserves only the task-related

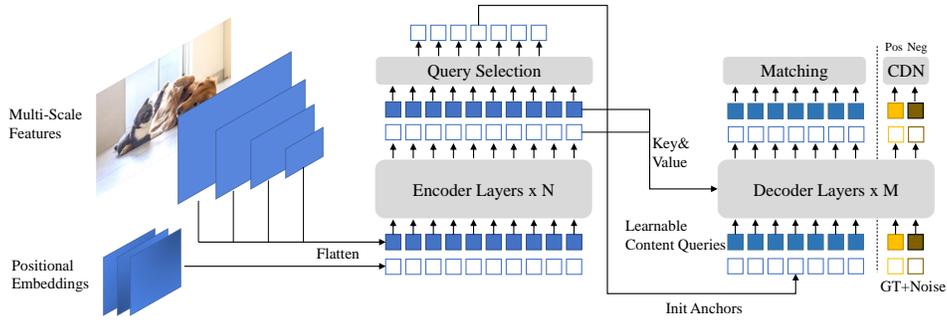


Fig. 2. The DINO architecture. Image from [18].



Fig. 3. A sample from the CFray dataset.

regions of each X-ray image. This paper exploits the SIXRay-D version of SIXray10.

CFray is a custom dataset that contains X-ray scan images of parcels containing firearms and firearm components. The images have been captured using an X-ray scanning machine and annotated using the VoTT<sup>1</sup> annotation tool. Parcels have been filled with both illicit and non-illicit objects to emulate, as closely as possible, real-life illicit parcels. The dataset contains 1368 annotated images, with the supported classes being: “SMGs”, “Metal Pistols”, “Components”, “Revolvers” and “Plastic Pistols”. Figure 3 depicts indicative reference images.

### B. Evaluation Metrics

This paper presents evaluation results regarding both the achieved object detection accuracy and the achieved inference speed.

1) **Detection Accuracy Metrics:** Object detectors typically output the bounding boxes/ROIs of the detected objects, the predicted classes, and the corresponding confidence value for each predicted class. Each input image may contain multiple

objects (e.g., gun components, gun ammunition, and a gun in one parcel), therefore each object should be separately located and then recognized.

The Intersection over Union (IoU) ratio is calculated by comparing the predicted bounding box with the ground-truth RoI. Ideally, they should be totally overlapping:

$$\text{IoU} = \frac{\text{size of intersection}}{\text{size of union}}. \quad (1)$$

However, the calculated IoU output is typically not 1. Therefore, a threshold is used to determine whether the predicted RoI is a sufficiently accurate result. Thresholds of 0.5 and 0.5-0.95 are used in this paper.

The IoU is used as a threshold ratio in the mean Average Precision (mAP) metric, which is one of the evaluation metrics used. To calculate the mAP, true positives (TP), false positives (FP) and false negatives (FN) are also required. Precision measures the percentage of correct predictions, while Recall measures the correct predictions relative to the ground-truth. These metrics are calculated using the following equations:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (3)$$

The Precision-Recall (PR) curve illustrates the balance between precision and recall for a given class. Average Precision (AP) is calculated as the area under the PR curve, ranging from 0 to 1. The formula for AP is as follows:

$$\text{AP} = \int_0^1 p(r)dr. \quad (4)$$

Finally, the mAP is calculated as the mean value of AP for each class over all the N classes:

$$\text{mAP} = \frac{1}{N} \sum_k^N \text{AP}_k. \quad (5)$$

<sup>1</sup><https://github.com/microsoft/VoTT>

2) *Speed Metric*: Inference time in milliseconds (ms) is a metric that calculates the time required for an object detection method to output the bounding boxes, predicted classes, and confidence value of one input image. It is typically calculated as an average over multiple images.

### C. Experimental Setup

The methods presented in Section III have been adopted and compared in the chosen datasets. Various combinations of detectors and backbones have been exploited, focusing as much as possible on comparing Transformer/hybrid components. Thus, both DINO and even the YOLOv8 detection head itself have been combined with the NextViT backbone (although YOLO itself is a CNN).

Regarding the connection of RT-DETR and YOLOv8 detectors with the NextViT backbone, i.e., combinations which have been engineered from scratch in this paper, a validation process has been utilized to identify the optimal residual connections. Experiments have shown that the best option for YOLOv8 is to be connected to backbone layers 7 and 17, while the respective result for RT-DETR are layers 9 and 19. However, the final performance differences are very small compared to different connectivity configurations.

All experiments have been conducted on a PC with Ubuntu Linux 22.04, utilizing a 13th Gen Intel(R) Core(TM) i9-13900K CPU and two NVIDIA GeForce RTX 4070 Ti GPUs. Development has taken place in Python, PyTorch and Ultralytics. The training hyperparameters used with the NextViT backbone are the following ones: AdamW [38] optimizer with 0.01 learning rate, weight decay of 0.0005 and finally trained for 100 epochs using early-stopping. For SWIN and DINO, the hyperparameter values given in [18] have been adopted.

### D. Evaluation Results

Table I shows the results of pretraining the chosen methods on SIXray10-D and finetuning on CFray. All methods are comparable in terms of mAP. However, both methods using the NextViT backbone require about a tenth of the inference time of the SWIN-b and DINO methods, while generating results of similar accuracy. In terms of mAP@50 performance, SWIN+DINO is the most powerful method. In X-rays, it is important to have good mAP scores at all thresholds, hence the more stringent mAP@50-95 threshold evaluation. For this threshold, YOLOv8 [20] combined with NextViT [17] is slightly ahead of SWIN+DINO.

Thus, if real-time performance is needed, the combination of NextViT+YOLOv8 is without question the best choice. On the other hand, if the most important priority is to avoid false negatives (as in many security inspection use-cases) and inference speed is not equally significant, then the SWIN+DINO combination should be considered.

### E. Ablation Studies

Besides the pretraining+downstream finetuning arrangement, additional experiments have been conducted with single-pass training only on SIXray10-D or on CFray. The respective results are shown in Tables III and II, respectively.

TABLE I  
EXPERIMENTAL EVALUATION RESULTS FOR THE SELECTED METHOD COMBINATIONS (PRETRAINING ON SIXRAY10-D, DOWNSTREAM FINETUNING ON CFRAY).

Configuration		Metrics		
Backbone	Detector	mAP@50	mAP@50-95	Inf. Time
SWIN-b	DINO	<b>0.993</b>	0.816	175.4
NextViT-s	RT-DETR-l	0.943	0.765	16
NextViT-s	YOLOv8-l	0.986	<b>0.82</b>	<b>11.6</b>

TABLE II  
EXPERIMENTAL RESULTS OF METHODS TRAINED ONLY ON CFRAY.

Configuration		Metrics		
Backbone	Detector	mAP@50	mAP@50-95	Inf. Time
SWIN-b	DINO	0.942	<b>0.837</b>	175.4
NextViT-s	RT-DETR-l	0.94	0.748	16
NextViT-s	YOLOv8-l	<b>0.984</b>	0.76	<b>11.7</b>

TABLE III  
EXPERIMENTAL EVALUATION RESULTS FOR THE SELECTED METHOD COMBINATIONS ON SIXRAY10-D.

Configuration		Metrics		
Backbone	Detector	mAP@50	mAP@50-95	Inf. Time
SWIN-b	DINO	0.902	0.765	161.3
NextViT-s	RT-DETR-l	0.889	0.762	16
NextViT-s	YOLOv8-l	<b>0.906</b>	<b>0.793</b>	<b>12.8</b>

Tables I and II demonstrate that pretraining on the SIXray10-D dataset and then finetuning on CFray can significantly and clearly improve the achieved accuracy only for NextViT-s+YOLOv8-l, and in fact only when the stricter IoU thresholds are considered.

On the other hand, Table III displays the performance of the employed methods on the SIXray10-D dataset. The NextViT-s+YOLOv8 method outperforms all others on all metrics, especially on the mAP@50-95 metric. The SWIN-b+DINO combination, despite its higher model complexity (and, therefore, lower inference speed) lags in comparison, accuracy-wise. Given that SIXray10-D is a more reasonably sized dataset compared to the very small CFray dataset, this implies that an end-to-end Transformer architecture faces issues in the highly specific X-ray domain, compared to hybrid solutions that also employ convolutions, but is surprisingly accurate in the low-data regime as shown by Table II.

## V. CONCLUSIONS

Automated illicit object detection in X-ray images taken at critical terminals, such as airports, train stations, subways, and ports, is a crucial task for public safety. This paper evaluates efficient and performant methods for handling the large volume and throughput of parcels, luggage, and passengers. The results demonstrate the remarkable accuracy of the DINO Transformer detector in the low-data regime, the impressive real-time performance of YOLOv8, and the effectiveness of

the NextViT backbone. Future research directions may involve combining X-ray-specific auxiliary neural modules with the methods evaluated in this paper to further improve accuracy.

## VI. ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101073876 (Ceasfire). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

## REFERENCES

- [1] Ioannis Mademlis, Georgios Batsis, Adamantia Anna Rebolledo Chrysochoou, and Georgios Th. Papadopoulos. Visual inspection for illicit items in X-ray images using deep learning. In *Proceedings of the IEEE International Conference on Big Data (BigData)*, pages 4081–4089. IEEE, 2023.
- [2] Georgios Batsis, Ioannis Mademlis, and Georgios Th. Papadopoulos. Illicit item detection in X-ray images for security applications. In *Proceedings of the IEEE International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 63–70. IEEE, 2023.
- [3] Taimur Hassan, Samet Akcay, Mohammed Bennamoun, Salman Khan, and Naoufel Werghi. Cascaded Structure Tensor Framework for robust identification of heavily occluded baggage items from X-ray scans. *arXiv preprint arXiv:2004.06780*, 2020.
- [4] Domingo Mery, Daniel Saavedra, and Mukesh Prasad. X-ray baggage inspection with computer vision: A survey. *IEEE Access*, 8:145620–145633, 2020.
- [5] 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.
- [6] Nikolaos Rodis, Christos Sardanios, Georgios Th Papadopoulos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, and Iraklis Varlamis. Multimodal eXplainable Artificial Intelligence: A comprehensive review of methodological advances and future research directions. *arXiv preprint arXiv:2306.05731*, 2023.
- [7] Mehdi Rafiei, Jenni Raitoharju, and Alexandros Iosifidis. Computer vision on X-ray data in industrial production and security applications: A comprehensive survey. *IEEE Access*, 11:2445–2477, 2023.
- [8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 25, 2012.
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.
- [11] Ning Gan, Fang Wan, Guangbo Lei, Li Xu, Chengzhi Xu, Ying Xiong, and Wen Zhou. YOLO-CID: Improved YOLOv7 for X-ray contraband image detection. *Electronics*, 12(17):3636, 2023.
- [12] Chunjie Ma, Li Zhuo, Jiafeng Li, Yutong Zhang, and Jing Zhang. Occluded prohibited object detection in X-ray images with global context-aware multi-scale feature aggregation. *Neurocomputing*, 519:1–16, 2023.
- [13] Yu Ren, Haigang Zhang, Hongxing Sun, Guanglin Ma, Jin Ren, and Jinfeng Yang. LightRay: Lightweight network for prohibited items detection in X-ray images during security inspection. *Computers and Electrical Engineering*, 103(C), 2022.
- [14] Sotirios Konstantakos, Despina Chalkiadaki, Ioannis Mademlis, Adamantia Chrysochoou, and Georgios Papadopoulos. Self-supervised visual learning for analyzing firearms trafficking activities on the Web. pages 4071–4080, 2023.
- [15] Ioannis Mademlis, Marina Mancuso, Caterina Paternoster, Spyridon Evangelatos, Emma Finlay, Joshua Hughes, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Georgios Stavropoulos, Konstantinos Votis, et al. The invisible arms race: digital trends in illicit goods trafficking and AI-enabled responses. *Authorea Preprints*, 2023.
- [16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. SWIN transformer: Hierarchical Vision Transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.
- [17] Jiashi Li, Xin Xia, Wei Li, Huixia Li, Xing Wang, Xuefeng Xiao, Rui Wang, Min Zheng, and Xin Pan. Next-ViT: Next Generation Vision Transformer for efficient deployment in realistic industrial scenarios. *arXiv preprint arXiv:2207.05501*, 2022.
- [18] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.
- [19] Wenyu Lv, Shangliang Xu, Yian Zhao, Guanzhong Wang, Jinman Wei, Cheng Cui, Yuning Du, Qingqing Dang, and Yi Liu. DETRs beat YOLOs on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023.
- [20] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. Ultralytics, 2023.
- [21] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. SIXray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2119–2128, 2019.
- [22] Samet Akcay, Mikolaj Kundegorski, Michael Devereux, and Toby Breckon. Transfer learning using Convolutional Neural Networks for object classification within X-ray baggage security imagery. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1057–1061. IEEE, 2016.
- [23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with Region Proposal Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, volume 28, 2015.
- [24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.
- [26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
- [27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single-Shot MultiBox Detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37. Springer, 2016.
- [28] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, 2019.
- [29] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017.
- [30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.
- [31] Fangtao Shao, Jing Liu, Peng Wu, Zhiwei Yang, and Zhaoyang Wu. Exploiting foreground and background separation for prohibited item detection in overlapping X-Ray images. *Pattern Recognition*, 122:108261, 2022.
- [32] Robert J Wang, Xiang Li, and Charles X Ling. Pelee: A real-time object detection system on mobile devices. volume 31, 2018.
- [33] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. GhostNet: More features from cheap operations. In *Proceedings of*

*the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1580–1589, 2020.

- [34] Zhuofan Zong, Guanglu Song, and Yu Liu. DETRs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023.
- [35] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020.
- [36] Yishen Lin, Zifan Huang, Yun Liang, Yunfan Liu, and Weipeng Jiang. AG-YOLO: A rapid citrus fruit detection algorithm with Global Context Fusion. *Agriculture*, 14(1):114, 2024.
- [37] Hong Duc Nguyen, Rizhao Cai, Heng Zhao, Alex C. Kot, and Bihan Wen. Towards more efficient security inspection via deep learning: A task-driven X-ray image cropping scheme. *Micromachines*, 13(4):565, 2022.
- [38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.