Visual inspection for illicit items in X-ray images using Deep Learning

1st Ioannis Mademlis Department of Informatics and Telematics Harokopio University of Athens Athens, Greece imademlis@hua.gr

3rd Adamantia Anna Rebolledo Chrysochoou Department of Informatics and Telematics Harokopio University of Athens Athens, Greece adamantia.reb@hua.gr 2nd Georgios Batsis Department of Informatics and Telematics Harokopio University of Athens Athens, Greece gbatsis@hua.gr

4th Georgios Th. Papadopoulos Department of Informatics and Telematics Harokopio University of Athens Athens, Greece g.th.papadopoulos@hua.gr

Abstract-Automated detection of contraband items in X-ray images can significantly increase public safety, by enhancing the productivity and alleviating the mental load of security officers in airports, subways, customs/post offices, etc. The large volume and high throughput of passengers, mailed parcels, etc., during rush hours practically make it a Big Data problem. Modern computer vision algorithms relying on Deep Neural Networks (DNNs) have proven capable of undertaking this task even under resourceconstrained and embedded execution scenarios, e.g., as is the case with fast, single-stage object detectors. However, no comparative experimental assessment of the various relevant DNN components/methods has been performed under a common evaluation protocol, which means that reliable cross-method comparisons are missing. This paper presents exactly such a comparative assessment, utilizing a public relevant dataset and a well-defined methodology for selecting the specific DNN components/modules that are being evaluated. The results indicate the superiority of Transformer detectors, the obsolete nature of auxiliary neural modules that have been developed in the past few years for security applications and the efficiency of the CSP-DarkNet backbone CNN.

Index Terms—Deep Neural Networks, Object Detection, Xrays, Security, Convolutional Neural Networks, Transformers

I. INTRODUCTION

Detecting contraband items using X-ray scanning of luggage, parcels, etc. is a crucial requirement for ensuring public security (e.g. preventing terrorist attacks, fighting smuggling of illegal goods, etc.) [1] [2]. X-rays are electromagnetic waves with wavelengths shorter than that of visible light, able to penetrate most materials; X-ray scanners exploit this fundamental property to screen items, such as luggage or packages (e.g., in airports, post/customs offices, etc.). Human operators are able to detect a wide range of potential threats, such as explosives, weapons, or sharp objects, using high-resolution images generated by scanning machines [3]. However, fully

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 (Ceasefire).

manual screening has important shortcomings: the quality of the scan image can be influenced by several factors, such as occluded objects, cluttered environment or certain material properties of the scanned items [4], while heavy traffic during rush hours may mentally overload human security officers. Thus, illicit items may be missed, due to the need for "the line to keep moving" or because of perceptual limitations. The high volume and high throughput of X-ray scans in such scenarios render manual screening ineffective and demand automated Big Data analysis solutions.

Efficient automated X-ray image analysis/screening for illicit item detection is nowadays possible thanks to the advances of computer vision and machine learning. Such methods are ideal for large-scale information processing and, therefore, hold the potential to facilitate the detection of illicit item trafficking activities, suspected terrorist attacks, etc. Deep Neural Networks (DNNs) have proven to be remarkably capable in supporting human operators for similar tasks, thus greatly increasing their productivity and reducing the possibility of mistakes. Both whole-image recognition and object detection methods have been proposed for illicit/contraband item detection in X-ray images. While the former ones simply classify an entire image and assign it an overall class label, algorithms of the latter type identify Regions-of-Interest (RoIs), i.e., bounding boxes that localize (in 2D pixel coordinates) specific objects visible in an input image. While there have been significant advancements in object detection algorithms through the use of DNNs, achieving sufficient performance in real-world scenarios continues to be a challenge [5] [6] [7] [8].

The typical goal is for a deployed DNN to automatically detect illicit goods, such as drugs or weapons, in passengers, luggage or mailed parcels. The dominant trends are similar to those of the RGB image analysis, but obviously different training datasets are utilized. Additionally, special/auxiliary neural modules are commonly employed as part of the overall DNN architecture, so that accuracy is improved in the face of typically encountered issues such as high occlusions, very cluttered backgrounds and large class imbalance. These mechanisms are designed to handle similar application domainspecific aspects.

Given the practical importance of the task, recent literature surveys have overviewed detection of illicit items in X-ray scans for security applications [9] [10]. Yet, none of them has assessed the various state-of-the-art methods using a common experimental evaluation protocol, thus rendering cross-method performance comparisons difficult. In an attempt to remedy the situation, this paper contributes a thorough quantitative assessment, using the most common relevant public dataset (SIXray [11]) and comparing various combinations of DNN backbones, auxiliary modules and detection heads. The results indicate:

- the superiority of Transformer detectors,
- the obsolete nature of auxiliary neural modules that have been developed in the past few years for security applications,
- the high efficiency of the CSP-DarkNet backbone CNN.

The remainder of the paper is organized as follows. Section II overviews the recent literature on illicit item detection in X-ray scan images using DNNs. Section III briefly presents the specific DNN backbones, auxiliary modules and detection heads that are being quantitatively assessed. Section IV outlines the experimental evaluation process, which was conducted on a well-known public dataset, and discusses the obtained results. Section V concludes the preceding discussion by identifying the implications of these findings and directions for future research.

II. RELATED WORK

Various approaches have been employed over the years for illicit items detection in X-ray scan images. The method of [12] addressed the issue of limited training data by employing a pretrained CNN and fine-tuning it in the X-ray domain. This is an important issue in automated X-ray screening, since negative images (where no illicit item is present) are typically significantly more than the positive ones, with this fact reflected in the relevant available datasets.

Common DNNs for object detection have also been evaluated with regard to their discrimination capacity and transferability between different X-ray scanners [13]; examples include Faster R-CNN [14], Mask R-CNN [15] and RetinaNet [16]. However, modifying fast, anchor-based, singlestage object detectors such as Single Shot MultiBox Detector (SSD) [17] or You Only Look Once (YOLO) [18] is the most common approach, due to their ability to operate in real-time even in embedded computer hardware. Such modifications may have various forms. For instance, a Cascaded Structure Tensor (CST) is proposed in [19] which takes advantage of contour-based information to extract object proposals; the latter ones are then classified using a CNN. An alternative lightweight object detector, called LightRay, is introduced in [20] as a modified version of the YOLOv4 model for small illicit item detection in complex backgrounds. It consists of a fast MobileNetV3 [21] backbone CNN and a feature enhancement network that includes a Lightweight Feature Pyramid Network (LFPN) [22], to obtain information of objects at different scales, and a Convolutional Block Attention Module (CBAM) [23], for refining feature maps through a spatial attention mechanism.

A different approach is followed in [24], where a novel mechanism called Foreground and Background Separation (FBS) is proposed for separating illicit items from complex/cluttered backgrounds. This is achieved by using a feature extraction DNN combined with Spatial Pyramid Pooling (SPP) and a Path Aggregation Network, which extracts highlevel features. These feature maps serve as an input to two neural decoders, which reconstruct the background and the foreground simultaneously. Then, an attention module directs the overall model's focus on the foreground objects.

Focusing on real-time performance, YOLOv5 is modified in [25] using the Stem [26] and CGhost [27] modules, resulting in a model with reduced number of parameters that still achieves competitive results in comparison with the baseline method.

III. EMPLOYED METHODS FOR ILLICIT ITEM DETECTION IN X-RAY SCAN IMAGES

This section briefly illustrates the different deep neural modules/architectures that have been selected for comparative experimental assessment. First, the relevant one-stage object detection heads are described. Then, the various backbone networks and the auxiliary modules that have been included are being presented. Finally, the specific combinations of the above-mentioned components that are experimentally compared in Section IV are discussed and justified.

A. Detection Heads

Most single-stage object detectors utilize reference anchor boxes of different sizes and aspect ratios, which are placed at various positions across the input image. The goal of these anchor boxes is to capture the variation in object shapes and sizes present in the dataset. Typically, they are predefined (e.g., calculated based on prior knowledge of the sizes, aspect ratios, and distributions of ground-truth objects in the COCO dataset [28]). In many implementations the match between these predefined anchor boxes and the training dataset is verified before training commences, by computing the achievable recall rate if the object detector using these anchors has access to the ground-truth for all objects in the dataset. If this recall rate is too low, the predefined anchors are assumed to be unfit and a new set of dataset-specific anchor boxes is estimated (e.g., via clustering). The detection head essentially outputs the offset (in pixel space) of each predicted bounding box from a known anchor box. After a set of raw detections has been generated, a typically handcrafted Non-Maximum Suppression (NMS) algorithm refines them by merging/filtering any spatially overlapping detected RoIs which correspond to a single visible object [1] [29] [30].

You Only Look Once (YOLO) [18] is a series of fast anchorbased, single-stage object detectors, where object localization and classification are performed using a single CNN. This architecture can, however, be divided into a backbone network, a succeeding neck network and a final prediction head. YOLOv5 [31], which is an update of YOLOv4 [32], is inspired by EfficientNet [33] and, thus, can be easily reconfigured for different network complexity profiles. Out of the common variants (YOLOv5s, YOLOv5m, YOLOv51, YOLOv5x) the one employed in this paper is YOLOv51. The overall YOLOv5 architecture is presented in Fig. 1.



Fig. 1. YOLOv5 overall architecture.

While all detectors of the YOLO family rely on preset anchors, Fully Convolutional One-stage Object Detection (FCOS) [34] is one of the first successful anchor-free onestage CNN detectors that outputs per-pixel predictions. Thus, it avoids the initial computational load for setting-up the anchors before the main training process, as well as all relevant hyperparameters that are difficult to tune. FCOS requires a neck network based on Feature Pyramid Network (FPN) [22], which aggregates different backbone-derived feature maps corresponding to different image scales. Features from the downsampling path are fed to the upsampling one through lateral synapses. Thus, objects of different sizes can be detected at different levels of the feature pyramid. Detection is conducted by the shared head, which analyzes the outputs of the FPN levels and is composed of three branches: one for classification, one for centerness and one for regression. All three of them output per-pixel predictions: the first one predicts the object's class, the second one how far a pixel deviates from the center of its associated bounding box, while the third one outputs the dinstance (in pixels) of the pixel in question and the corners of its bounding box. One disadvantage of FCOS is that it requires higher input image resolutions to operate correctly, due to the per-pixel nature of its predictions; this creates an execution time overhead during both the training and the inference stage. A high-level diagram of its architecture is depicted in Fig. 2.

The anchor-free direction is also followed by YOLOv8 [35], a recent successor to YOLOv5 that directly predicts the centers of bounding boxes. Along with various minor improvements in the CNN architecture and an enhanced data augmentation strategy during training, YOLOv8 achieves an outstanding balance between inference speed and prediction accuracy. The YOLOv8 variant that is utilized in this paper is YOLOv81.

Despite the early dominance of CNNs as detection heads, top-performing Vision Transformer DNNs have emerged during the past few years. One of the first such approaches was Detection Transformer (DETR) [36]: it is an Encoder-Decoder Transformer DNN [37], placed after a CNN backbone, which treats image blocks as tokens. DETR handles object detection as a set prediction task and assigns labels by bipartite graph matching. Learned positional encodings, the so-called "object queries", essentially look for a particular object in the image. The method is not only anchor-free, but also NMSfree; DETR does not need any handcrafted algorithmic components. A state-of-the-art improvement of DETR is DINO [38], which accumulates various minor enhancements over baseline DETR and reinstates the use of anchor boxes in a Transformer-compatible manner. Moreover, it exploits an additional contrastive loss term during training [39], by adding two different types of noise to the same ground-truth RoI; the resulting bounding box with a smaller/larger amount of noise is considered a positive/negative sample, respectively. The goal is to push the DNN towards avoiding duplicate bounding box outputs that correspond to a single ground-truth object. A highlevel diagram of the DINO architecture is depicted in Fig. 3.

B. Backbone Networks

ResNet-101 [40] is a well-known CNN backbone, very commonly employed for almost any image analysis task. It is one of the first CNNs that was able to be trained with large network depth without being negatively impacted by the gradient vanishing problem, mainly thanks to its introduction of the "skip synapses". The continuing popularity of ResNet for almost a decade showcases its value for the wider computer vision community.

The default backbone CNN of YOLOv5 is CSP-Darknet53 [32], a modified version of Darknet53 [41] combined with a Cross Stage Partial Network (CSPNet) strategy [42], which is specifically designed for assisting object detection. As presented in Fig. 4, the main convolutional block of CSP-Darknet53 consists of convolutional layers, residuals and the SiLU activation function, while the final feature maps are refined using a Spatial Pyramid Pooling-Fast (SPPF) module [43]. The neck network consists of a Feature Pyramid Network (FPN) and a Path Aggregation Network (PAN) [44]. These modules repeatedly fuse feature maps from different scales and depth levels, thus leading to final image representations, which are simultaneously characterized by accurate spatial localization details, rich semantics and high invariance regarding object detection. Finally, the prediction head outputs the candidate detected RoIs through a set of convolutional operations.



Fig. 2. The architecture of FCOS [34].



Fig. 3. The architecture of DINO [38].

Due to the rather low inference speed of very deep ResNet variants, fast generic CNN backbones appeared over the years, targeting execution on embedded computers with limited processing power. One of the most important relevant architectures is MobileNet, which accelerates inference by incorporating "separable convolutions" [45]. Additionally, the widespread use of 1×1 convolutional kernels allows their optimized implementation through generalized matrix multiplication, while test accuracy and training are aided by the utilization of batch normalization and ReLU activation functions. MobileNetV2 improves this architecture by periodically decimating the number of convolutional channels along the depth dimension (similarly to SqueezeNet [46]), adding skip synapses and reducing the need for greater number of channels per convolutional layer in later layers. The next iteration, i.e., MobileNetV3 [21], further enhances the architecture by introducing a channel-wise attention module within each separable convolution and optimizing architectural details at the network design phase, through the use of Neural Architecture Search (NAS). Overall, MobileNets achieve a very good balance between speed and accuracy: in most applications, they lag only slightly compared to non-lightweight deep CNN backbones, while being significantly faster.

An alternative lightweight fast CNN backbone is Efficient-Net [33]. As in the case of MobileNetV3, it is designed by employing NAS based on reinforcement learning; however, the reward function prefers a low total number of computational operations during the forward pass instead of a low required inference runtime. In general, however, the individual neural layers/modules are similar to the ones utilized by MobileNetV3. EfficientNet variants of various complexities are available, so that the speed-accuracy trade-off can be adjusted based on the desired application and the computational power which is available at the inference stage. More complex variants are typically deeper (more convolutional layers), wider (more channels per layer) and process input images of higher resolution. The architecture family has been improved with EfficientNetV2 [47], which makes NAS to also reward higher training efficiency and incorporates enhancements in the regularization scheme utilized during training. The EfficientNet variant utilized in this paper is EfficientNetV2-S.

C. Auxiliary Modules

Due to the peculiarities of illicit item detection in Xray scan images of luggage, parcels, etc., various additional domain-specific, plug-in neural modules have been proposed over the years. For instance, the method of [11] introduces a module called Class-balanced Hierarchical Refinement (CHR), to enhance the prediction capacity of the CNN under extreme class imbalance. CHR can be placed as a neck module on top of any CNN backbone.

In an orthogonal direction, the De-occlusion Attention Module (DOAM) [48] is a neural module designed to overcome occlusions in X-ray images; this is important because occlusions are common, due to the absorption of X-rays by certain materials, such as metals, and the visual overlap of multiple objects within densely packed parcels. The latter phenomenon implies that a single pixel may correspond to multiple semantic classes, of objects located at different vertical distances from the sensor, due to the penetrative nature of X-rays. Thus, the overall X-ray image can be considered a superposition of various sub-images. DOAM consists of two sub-modules, named Edge Guidance (EG) and Material Awareness (MA), which identify edge and material cues for all visible objects. An alternative domain-specific module is Lateral Inhibition Module (LIM) [49], which includes two components: Bidirectional Propagation (BP) and Boundary Activation (BA). The former one minimizes the impact of neighboring regions, by isolating irrelevant information and the latter one captures object boundaries. Both DOAM and LIM have shown promising results in overcoming object occlusion issues in X-ray scan images.

In a subsequent attempt to overcome the issues induced by the typically high visual overlap of objects within a densely packed luggage/parcel, the method in [50] introduces the socalled Dense De-overlap Module (DDoM). It operates by assigning learned weights to each channel of a convolutional feature tensor, indicating how relevant it is to the object class in question. This operates under the assumption that different convolutional channels are responses to different sub-images, including irrelevant background ones. Finally, the integrated Prohibited Object Detection (POD) method [51] for X-ray image analysis combines a learnable Gabor layer for edge information retrieval, a spatial attention module for directing focus on low-level features, a Global Context Feature Extraction (GCFE) module and a Dual Scale Feature Aggregation (DSFA) module to enhance semantic information from highlevel features.

D. Methodology for comparative assessment

The literature of DNNs for illicit item detection in Xray scan images mostly employs common neural architectures/building blocks (detectors, backbones, necks), typically preferring fast and proven ones. Thus, most of the specific neural components reviewed in the previous subsections were chosen to be included in this comparative experimental assessment because they are commonly found in recent relevant papers (e.g., YOLOv5, FCOS, ResNet-101, EfficientNet, MobileNet). However, the final selection of individual components is influenced by other considerations as well, such as state-ofthe-art status (e.g., YOLOv8, DINO). In particular, one the goals of this work is to identify how relevant domain-specific neural modules, such as CHR, LIM, DOAM or DDoM, remain in the face of the advancements offered by modern generic detectors.

Thus, given that it would be very impractical to quantitatively evaluate all potential combinations of the selected neural building blocks, the following process has been followed:

- First, commonly employed CNN detectors are evaluated in combination with the selected CNN backbones.
- Second, the state-of-the-art one-stage CNN detector, i.e., YOLOv8, is evaluated in combination with the bestperforming CNN backbone.
- Third, the various auxiliary modules (serving as neck subnetworks) are evaluated in combination with the overall best CNN detection head.
- Four, the best performing CNN backbone is evaluated in combination with DINO; a representative of state-of-the-art Transformer-based detection heads.

The details and the results of this incremental experimental assessment are presented in Section IV.

IV. EXPERIMENTAL EVALUATION

This section overviews the common experimental setup used for evaluating and comparing the components presented in Section III. Subsequently, the assessment results are reviewed and discussed.

A. Experimental Dataset

SIXray [11] is employed for conducting the experimental method assessment. It is a well-known publicly available Xray security dataset consisting of 1,059,231 X-ray images from subway stations. The 6 classes of illicit objects contained in these images are "gun", "knife", "wrench", "pliers" and "scissors". Additionally, a "negative" class includes all images without any illicit item. Three different dataset subsets are typically utilized in different experimental setups, namely SIXray10, SIXray100 and SIXray1000, where the number indicates the ratio of negative against positive samples. SIXray contains ground-truth whole-image class label annotations manually set by human security inspectors, while their groundtruth object RoIs/bounding boxes are available only for the test set. This paper uses the revised object detection annotations for the training subset provided by [52]. Despite the fact that only images containing at least one contraband item were utilized,



Fig. 4. The main CSP-Darknet53 components.

the official training-test set split was adopted. Fig. 5 depicts examples of detections on SIXray test set images.

B. Evaluation Metrics

The effectiveness of the proposed method is measured using the mean Average Precision (mAP) metric. In object detection tasks, IoU is commonly used to measure the overlap between the predicted and the corresponding ground-truth RoI. In addition, a threshold value is defined in order to decide whether the prediction is actually correct. True Positives (TP), False Positives (FP), and False Negatives (FN) depend on the IoU, the predicted label and the ground-truth label. These elementary metrics are utilized to calculate Precision and Recall:

$$Precision = \frac{TP}{TP + FP}.$$
 (1)

$$Recall = \frac{TP}{TP + FN}.$$
 (2)

The Precision-Recall (PR) curve depicts the trade-off between precision and recall for different discrimination thresholds. Average Precision (AP) is the area under the PR curve and its range is between 0 to 1. AP is defined as:

$$AP = \int_0^1 p(r) \, dr. \tag{3}$$

mAP is calculated as the mean of AP over all classes:

$$mAP = \frac{1}{N} \sum_{i}^{N} AP_i.$$
 (4)

C. Experimental Evaluation

Evaluation of all competing method combinations in the SIXray dataset was conducted using the mAP metric at a 0.5 IoU threshold.

Table I summarizes the mAP of the evaluated method combinations, selected under the rationale described in Subsection III-D. As it can be seen, the Transformer-based DINO outperforms all CNN-based detectors, but the CSPDarkNet-53 CNN backbone, which has been designed specifically for object detection, surpasses all competing approaches. Finally, as it can be deduced from the quantitative results, the domainspecific auxiliary modules that have been evaluated as neck subnetworks in combination with YOLOv8/CSPDarkNet-53 are essentially useless in combination with such an advanced CNN detector; they significantly degrade its accuracy. One potential reason may be that they are not really generic plug-in modules able to augment any CNN backbone/detector combination, but can only cooperate effectively with specific such combinations. Exploring this aspect is a fertile future research avenue.

It must be noted that the above results contradict those of the survey in [10], which concludes that Transformer-based DNNs do not work equally well on X-ray images because they emphasize contours, while CNNs emphasize texture. In practice, the comparative assessment results presented in this paper indicate that this is not in fact an issue, at least when a CNN backbone is utilized in combination with a state-ofthe-art Transformer detector. This is in-line with the recent findings of [53], where Transformer-based detection heads are shown to outperform all competitors.

Detector	Backbone Architecture	mAP
YOLOv5	CSPDarkNet-53	0.82
	ResNet-101	0.81
	MobileNetV3	0.76
	EfficientNetV2-S	0.81
FCOS	ResNet-101	0.78
	MobileNetV3	0.73
	EfficientNetV2-S	0.73
YOLOv8	CSPDarkNet-53	0.84
DINO	CSPDarkNet-53	0.89
Detector	Auxiliary Module	mAP
YOLOv8	CHR	0.68
	LIM	0.82
	DOAM	0.78
	DDoM	0.81
TABLE I		

Results of the quantitative assessment of the various selected method combinations, under the chosen experimental protocol. MAP@0.5 is the employed evaluation metric (higher is better).



Fig. 5. Predictions on the SIXray test subset.

V. CONCLUSIONS

The automated detection of contraband items in X-ray images obtained in airports, subways or post/customs offices is a task critical for public safety. Due to the large volume and high throughput of passengers, mailed parcels, etc., this is a Big Data analysis problem that requires fast algorithms. Existing one-stage DNNs for object detection have indeed been adapted and trained for this application domain, but so far they have not been compared under a common evaluation protocol. This paper presented exactly such a comparative assessment of various commonly employed or state-of-the-art deep neural components for object detection (detection heads, backbones, auxiliary domain-specific necks), using a wellknown, large-scale public relevant dataset. The results indicate the superiority of Transformer detectors, the obsolete nature of auxiliary neural modules that have been developed in the past few years for security applications and the high efficiency of the CSP-DarkNet backbone CNN. Future research directions include an investigation of whether domain-specific auxiliary modules can be effectively utilized in combination with advanced modern object detectors to further improve accuracy, as well as how an end-to-end Transformer solution would perform in comparison to the winning CSP-DarkNet+DINO combination.

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 (Ceasefire). 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

 G. Batsis, I. Mademlis, and G. T. 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)*, 2023.

- [2] I. Mademlis, M. Mancuso, C. Paternoster, S. Evangelatos, E. Finlay, J. Hughes, P. Radoglou-Grammatikis, P. Sarigiannidis, G. Stavropoulos, K. Votis, and G. T. Papadopoulos, "The invisible arms race: digital trends in illicit goods trafficking and AI-enabled responses," *techRxiv preprint*, 2023.
- [3] D. Mery, D. Saavedra, and M. Prasad, "X-ray baggage inspection with computer vision: A survey," *IEEE Access*, vol. 8, pp. 145 620–145 633, 2020.
- [4] S. Akcay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within X-ray security imaging," *Pattern Recognition*, vol. 122, p. 108245, 2022.
- [5] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos, "Deep affordance-grounded sensorimotor object recognition," in *Proceedings* of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [6] K. Gkountakos, A. Dimou, G. T. Papadopoulos, and P. Daras, "Incorporating textual similarity in video captioning schemes," in *Proceedings* of the IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), 2019.
- [7] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, pp. 261–318, 2020.
- [8] I. Mademlis, C. Symeonidis, A. Tefas, and I. Pitas, "Vision-based drone control for autonomous UAV cinematography," *Multimedia Tools and Applications*, pp. 1–29, 2023.
- [9] 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.
- [10] J. Wu, X. Xu, and J. Yang, "Object detection and X-ray security imaging: A survey," *IEEE Access*, 2023.
- [11] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. 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)*, 2019.
- [12] S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. 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)*, 2016.
- [13] Y. F. A. Gaus, N. Bhowmik, S. Akcay, and T. Breckon, "Evaluating the transferability and adversarial discrimination of convolutional neural networks for threat object detection and classification within X-ray security imagery," in *Proceedings of the IEEE International Conference On Machine Learning And Applications (ICMLA)*, 2019.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," *Proceedings of* the Advances in Neural Information Processing Systems (NIPS), 2015.
- [15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
- [16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single-shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016.
- [18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [19] T. Hassan, S. Akcay, M. Bennamoun, S. Khan, and N. Werghi, "Cascaded structure tensor framework for robust identification of heavily occluded baggage items from X-ray scans," *arXiv preprint arXiv:2004.06780*, 2020.
- [20] Y. Ren, H. Zhang, H. Sun, G. Ma, J. Ren, and J. Yang, "LightRay: Lightweight network for prohibited items detection in X-ray images during security inspection," *Computers and Electrical Engineering*, vol. 103, p. 108283, 2022.
- [21] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan et al., "Searching for MobileNetV3," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
- [22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

- [23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [24] F. Shao, J. Liu, P. Wu, Z. Yang, and Z. Wu, "Exploiting foreground and background separation for prohibited item detection in overlapping X-ray images," *Pattern Recognition*, vol. 122, p. 108261, 2022.
- [25] B. Song, R. Li, X. Pan, X. Liu, and Y. Xu, "Improved YOLOv5 detection algorithm of contraband in x-ray security inspection image," in *Proceedings of the International Conference on Pattern Recognition* and Artificial Intelligence (PRAI). IEEE, 2022.
- [26] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2018.
- [27] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [28] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014.
- [29] A. J. Shepley, G. Falzon, P. Kwan, and L. Brankovic, "Confluence: A robust non-IoU alternative to non-maxima suppression in object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [30] C. Symeonidis, I. Mademlis, I. Pitas, and N. Nikolaidis, "Neural attention-driven Non-Maximum Suppression for person detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 2454–2467, 2023.
- [31] G. Jocher, "YOLOv5 by Ultralytics." [Online]. Available: https: //github.com/ultralytics/yolov5
- [32] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.
- [33] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for Convolutional Neural Networks," in *Proceedings of the International Conference* on Machine Learning (ICML). PMLR, 2019.
- [34] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional onestage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [35] G. Jocher, "YOLOv8 by Ultralytics." [Online]. Available: https: //github.com/ultralytics/ultralytics
- [36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings* of the Advances in Neural Information Processing Systems (NIPS), 2017.
- [38] 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.
- [39] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition (CVPR), 2016.
- [41] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
- [42] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF Conference on Computer Vision* and Pattern Recognition Workshops (CVPRW), 2020.
- [43] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904– 1916, 2015.
- [44] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [45] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

- [46] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and; 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.
- [47] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proceedings of the International Conference on Machine Learning* (*ICML*). PMLR, 2021.
- [48] 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 ACM International Conference on Multimedia (ACM MM)*, 2020.
- [49] 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 (ICCV)*, 2021.
- [50] B. Ma, T. Jia, M. Su, X. Jia, D. Chen, and Y. Zhang, "Automated segmentation of prohibited items in x-ray baggage images using dense de-overlap attention snake," *IEEE Transactions on Multimedia*, 2022.
- [51] C. Ma, L. Zhuo, J. Li, Y. Zhang, and J. Zhang, "Occluded prohibited object detection in X-ray images with global context-aware multi-scale feature aggregation," *Neurocomputing*, vol. 519, pp. 1–16, 2023.
- [52] H. D. Nguyen, R. Cai, H. Zhao, A. C. Kot, and B. Wen, "Towards more efficient security inspection via deep learning: A task-driven Xray image cropping scheme," *Micromachines*, vol. 13, no. 4, p. 565, 2022.
- [53] B. K. Isaac-Medina, S. Yucer, N. Bhowmik, and T. P. Breckon, "Seeing through the data: A statistical evaluation of prohibited item detection benchmark datasets for X-ray security screening," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2023.