Illicit item detection in X-ray images for security applications

1st Georgios Batsis Department of Informatics and Telematics Harokopio University of Athens Athens, Greece gbatsis@hua.gr 2nd Ioannis Mademlis Department of Informatics and Telematics Harokopio University of Athens Athens, Greece imademlis@hua.gr

3rd 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 make it a Big Data analysis task. 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, anchor-based object detectors. This paper proposes a two-fold improvement of such algorithms for the Xray analysis domain, introducing two complementary novelties. Firstly, more efficient anchors are obtained by hierarchical clustering the sizes of the ground-truth training set bounding boxes; thus, the resulting anchors follow a natural hierarchy aligned with the semantic structure of the data. Secondly, the default Non-Maximum Suppression (NMS) algorithm at the end of the object detection pipeline is modified to better handle occluded object detection and to reduce the number of false predictions, by inserting the Efficient Intersection over Union (E-IoU) metric into the Weighted Cluster NMS method. E-IoU provides more discriminative geometrical correlations between the candidate bounding boxes/Regions-of-Interest (RoIs). The proposed method is implemented on a common single-stage object detector (YOLOv5) and its experimental evaluation on a relevant public dataset indicates significant accuracy gains over both the baseline and competing approaches. This highlights the potential of Big Data analysis in enhancing public safety.

Index Terms—Illicit item detection, X-ray image analysis, Deep Neural Networks, Object detection, Non-Maximum Suppression

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.). X-rays are electromagnetic waves with wavelengths shorter than thee 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 [1]. However, fully 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 [2], 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 automatic illicit item detection is nowadays possible thanks to the advances of computer vision and machine learning. 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, based on DNNs. 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 over the last few decades, achieving sufficient performance in real-world scenarios continues to be a challenge [3]. The majority of the proposed methods incorporates mechanisms designed to handle domain-specific aspects (e.g., high occlusions, very cluttered backgrounds, large class imbalance, etc.). Additionally, due to the typically cluttered background of X-ray scan images of luggage, mailed parcels, etc., Non-Maximum Suppression (NMS) is also particularly important for security applications. NMS is the final refinement step incorporated to almost every visual object detection framework, assigned the duty of merging/filtering any spatially overlapping detected RoIs which correspond to a single visible object [4] [5].

Regarding image recognition, the method of [6] addressed the issue of limited training data by employing a pretrained CNN and fine-tuning it in the X-ray domain, while the method of [7] introduced a module named Class-balanced Hierarchical Refinement (CHR) to enhance the prediction capacity of the CNN under extreme class imbalance. 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. CHR is separately evaluated on top of three different CNNs: Res-Net101 [8], Inception-v3 [9] and Dense-Net [10].

Common DNNs for object detection have also been evaluated with regard to their discrimination capacity and transferability between different X-ray scanners [11]; examples include Faster R-CNN [12], Mask R-CNN [13] and RetinaNet [14]. However, modifying fast, anchor-based, singlestage object detectors such as Single Shot MultiBox Detector (SSD) [15] or You Only Look Once (YOLO) [16] 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 [17] which took 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 [18] as a modified version of the YOLOv4 model for small illicit item detection in complex backgrounds. It consisted of a fast MobileNetV3 [19] backbone CNN and a feature enhancement network that includes a Lightweight Feature Pyramid Network (LFPN) [20], to obtain information of objects at different scales, and a Convolutional Block Attention Module (CBAM) [21], for refining feature maps through a spatial attention mechanism.

A different approach was followed in [22], where a novel mechanism called Foreground and Background Separation (FBS) was 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 high-level 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. In an orthogonal direction, the De-occlusion Attention Module (DOAM) [23] is a neural module designed to overcome occlusion 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. 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) [24], 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.

NMS has also been modified in object detectors for X-ray scan image analysis. For instance, the framework of [25] is a modified YOLOv4 detector adopting deformable convolutions [26], the Gradient Harmonizing Mechanism (GHM) loss [27] and an augmented NMS algorithm combining Soft-NMS [28] with the Distance-Intersection-over-Union (DIoU) metric. Focusing on real-time performance, YOLOv5 was modified in [29] using the Stem [30] and CGhost [31] modules, resulting in a model with reduced number of parameters that still achieves competitive results in comparison with the baseline method. Finally, the integrated illicit Object Detection (POD) method [32] 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 high-level features.

However, to the best of the authors' knowledge, no object detector devised for the security domain has attempted to modify one basic building block of most single-stage detection frameworks: the anchor boxes. This is particularly important for X-ray screening of luggage or mailed parcels, because better matching between the anchor boxes and the distribution of object sizes/shapes in the training dataset leads to better detection performance on test images. Additionally, despite certain attempts to improve NMS for security applications, the results remain typically sub-optimal under object occlusions, which are common in this domain. Thus, this paper proposes a two-fold improvement of anchor-based, single-stage object detectors for automatically detecting contraband items in Xray scan images, contributing the following two novelties:

- Anchor box optimization by applying Hierarchical Clustering (HC) on the ground-truth object RoIs of the training set. By clustering the ground-truth bounding boxes based on their similarity in terms of size, shape, and position, the resulting clusters can be used to define a natural hierarchy, with larger clusters representing more general object shapes and smaller clusters capturing finer details and variations. The resulting hierarchy can also provide information about the relationships between different object classes.
- NMS modification to handle occluded object detection and to reduce false predictions, by computing richer geometrical correlations among candidate RoIs before final bounding box prediction. This is achieved by incorporating the Efficient-IoU metric into the Weighted-Cluster NMS method [33].

The remainder of the paper is organized as follows. Section II briefly presents the specific baseline algorithms which is

adopted for implementing the proposed novelties (YOLOv5, Weighted-Cluster NMS). Section III details the proposed method, consisting of an anchor box refinement approach and a modified NMS algorithm. 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, the limitations of this study and directions for future research.

II. PRELIMINARIES

In order to evaluate the proposed two-fold method, YOLOv5 [34] was adopted as a baseline object detector. The reason behind this choice was solely practical; in principle, the proposed method can be used to augment any other variant of the general anchor-based, single-stage object detection framework, as well.

A. YOLOv5 Architecture

You Only Look Once (YOLO) [16] 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 [34], which is an update of YOLOv4 [35], was inspired by EfficientNet [36] 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 backbone CNN of YOLOv5 is CSP-Darknet53, a modified version of Darknet53 [37] combined with Cross Stage Partial (CSP). As presented in Fig. 1, 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 [38]. The neck network consists of a Feature Pyramid Network (FPN) [20] and a Path Aggregation Network (PAN) [39]. 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. Overall YOLOv5 architecture is presented in Fig. 2.

B. Non-Maximum Suppression

Similarly to the majority of object detectors, YOLO generates a large set of overlapping object proposals, in the form of RoIs in pixel coordinates, along with the corresponding class labels and confidence scores. Thus, these candidate RoIs are filtered in a post-processing step based on certain criteria; this is called Non-Maximum Suppression (NMS). The conventional *Greedy NMS* algorithms processes the generated candidate bounding boxes and their corresponding confidence scores for each input image, sorting RoIs in descending



Fig. 1. Main YOLO components.



Fig. 2. YOLOv5 overall architecture.

confidence order. At first, the box with the highest confidence score is selected and the IoU between itself and all other boxes is calculated. All significantly overlapping RoIs, with an IoU greater than a threshold, are removed. This process is repeated until no bounding boxes remain in the sorted list. The NMS algorithm is presented in Fig. 3.

III. PROPOSED METHOD

The proposed method is a two-fold improvement of anchorbased, single-stage object detectors, which is highly suitable for the X-ray security scan image analysis domain, due to the peculiarities of such images (e.g., heavy occlusions, cluttered background, etc.).

A. Anchor boxes refinement

Most anchor-based 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 Algorithm 1 Non - Maximum Suppression Algorithm

Input: A set of bounding boxes $\mathcal{B} = \{B_1, B_2, ..., B_n\}$ with associated confidence scores $C = \{c_1, c_2, ..., c_n\}$. Sort the bounding boxes in \mathcal{B} by their confidence scores in descending order. Initialize an empty list \mathcal{D} to store the selected bounding boxes. **while** $\mathcal{B} \neq \emptyset$ **do** Select the bounding box B_i with the highest confidence score c_i from \mathcal{B} and add it to \mathcal{D} .

for $j \in \{1, 2, ..., n\}, j \neq i$ do if $IoU(B_i, B_j) > \theta$ then Remove B_j and c_j from \mathcal{B} and C. end if end for end while

Output: The list of selected bounding boxes \mathcal{D} with their associated confidence scores.

Fig. 3. The Greedy Non-Maximum Suppression algorithm.

predefined (e.g., in the case of YOLOv5, they have been calculated based on prior knowledge of the sizes, aspect ratios, and distributions of ground-truth objects in the COCO dataset [40]). In many implementations (e.g., YOLOv5) the match between these predefined anchor boxes and the training dataset is verified before training, by computing the achievable recall rate if the object detector using these anchors had 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. This is performed via a run of K-Means to group the ground-truth dataset RoIs into clusters, based on their dimensions in pixel space. The resulting cluster centers are selected as the new anchor boxes, with additional optimizations being possible (e.g., in YOLO a genetic algorithm refines them further). K-Means++ [41] can be utilized instead of classic K-Means [42].

In the current work, Hierarchical Clustering (HC) is used to obtain anchor boxes in a dataset-specific manner, according to Algorithm in Fig. 4. The goal is to generate anchor boxes that both fit the distribution of ground-truth object sizes/shapes and reflect their arrangement into a natural hierarchy, aligned with the spatial interrelations between the dataset's object classes. Thus, the hierarchy of anchor boxes can provide the detector an explicit template of the dataset's semantic structure, as expressed by the spatial relationships between different object classes. For example, in illicit item detection, RoIs corresponding to classes such as "knife" and "wrench" will likely fall under different sub-clusters of a common supercluster, containing all small-sized handheld items. Since a selection of anchor boxes that better match the dataset's object sizes and shapes is known to lead to better object detection accuracy [42], it is reasonable to expect further improvements by obtaining an arrangement of anchors that not only fit the distribution of the dataset's object sizes and shapes, but also reflect the natural hierarchy of the dataset's semantic classes (at least in terms of RoI shape/size).

Algorithm of Fig. 4 adopts an agglomerative HC method

[43] and adapts it to the anchor box refinement task. Its goal is to hierarchically group the training dataset's ground-truth bounding boxes, where each RoI is described as a 2D feature vector: $[w, h]^T$, where w/h is the RoI width/height, respectively. First, the pairwise Euclidean distances between all RoIs in the entire training dataset are computed and a linkage matrix is constructed using Ward's minimum variance [44]. Then, bounding boxes are assigned to clusters using the maximum cluster criterion and the mean of each cluster is calculated to obtain a new set of corresponding anchor boxes (one per cluster). The total number of target clusters is set to 9, according to experimental evaluation. HC generates a tree-like arrangement of clusters and sub-clusters. The leaves and the root of the clustering tree are not included in the final set of formed anchor boxes.

Algorithm 2 Anchor Boxes refinement using Hierarchical Clustering				
Input: Set of dataset bounding boxes wh , number of clusters n				
Output: New set of anchor boxes k				
1. Compute pairwise distances between all boxes using Euclidean distance				

- 2. Construct linkage matrix using Ward's minimum variance method
- 3. Assign boxes to clusters using maximum cluster criterion
- 4. Compute mean of each cluster to obtain new anchor boxes

Fig. 4. Anchor boxes refinement using Hierarchical Clustering.

B. Modified Non-Maximum Suppression

The default Greedy NMS method suffers in the presence of occlusions and gives rise to false positives, triggering various improvements that have been proposed over the years. For instance, Soft-NMS [28] modifies the candidate RoI scores by a Gaussian decay based on the degree of overlap, instead of directly setting the score of all overlapping bounding boxes to zero. This generates more accurate RoIs, even if they are occluded by other objects. Weighted-NMS [45] utilizes the weighted combination of scores and IoU values to define the merged coordinates of the predicted bounding boxes; the result is higher accuracy at the expense of increased time complexity, due to the number of iterations. To mitigate this, Weighted-Cluster NMS (WC-NMS) [33] has been developed: WC-NMS groups the detected candidate bounding boxes according to the IoU values and then selects the final RoIs according to the maximum score within each group. Implementation-wise this is done with SIMD parallelism and by exploiting cache locality, thanks to formulating the process as a series of matrix operations instead of naive iterative loops, resulting in the fast NMS Algorithm presented in Fig. 5. Thus, suppression is implemented by calculating the so-called IoU matrix. The latter is a symmetric matrix **M**, where $m_{i,j}$ is the IoU between the *i*-th and the *j*-th candidate RoI. Exploiting the symmetry of M, Algorithm in Fig. 5 retains only its upper triangular part.

Considering the importance of an efficient NMS method in the X-ray security image analysis domain, due to the densely packed nature of typical luggage and mailed parcels, this

Algorithm 3 Weighted - Cluster NMS pseudocode
N detected boxes $B = [B_1, B_2, \dots, B_N]$ sorted by classification score $s_1 \ge s_2 \ge$
$\ldots \geq s_N$, IoU threshold ϵ
Compute EIoU matrix $\mathbf{X} = EIoU(\mathbf{B}, \mathbf{B}).triu(diagonal = 1)$
NMS result from the previous iteration: \mathbf{b}_{t-1}
while $t \leq T$ do
$\mathbf{A}_t = diag(\mathbf{b}_{t-1})$
$\mathbf{C}_t = \mathbf{A}_t imes \mathbf{X}$
$\mathbf{g} \leftarrow \max_j \mathbf{C}_{t,j}$
$\mathbf{b}_t \leftarrow find(\mathbf{g} < \epsilon)$
$\mathbf{if} \mathbf{b}_t = \mathbf{b}_{t-1} \mathbf{then}$
$t_* = t$
break
end if
t = t + 1
end while
Calculation of the weighted Coordinates:
C' = CxS, S:scores
$\mathbf{B}_{final} = \frac{\mathbf{C'}_{\mathbf{x}}\mathbf{B}}{Repmat_4(\sum \mathbf{C}_i(i,:))}$

Fig. 5. EIoU Weighted-Cluster NMS algorithm.

paper adopts WC-NMS and improves it, by employing the Efficient-IoU (E-IoU) [46] as the overlap metric. E-IoU is an improvement of Complete-IoU (C-IoU) [33] and captures richer geometrical information about overlapping candidate bounding boxes, taking into account their overlapping area, their distance and their aspect ratios. Eqs. (1)-(2) define the overlapping criterion according to the so-called *E-IoU matrix*.

$$\mathbf{X} = \mathbf{M}_{IoU} - \mathbf{R}_{EIoU},\tag{1}$$

where \mathbf{M}_{IoU} is the IoU matrix, while \mathbf{R}_{EIoU} derives from the E-IoU loss [46]. Both matrices are calculated using the predicted candidate RoIs (in top-left, bottom-right pixel coordinates format), as follows:

$$\mathbf{R}_{EIoU} = \frac{\mathbf{D}_{centers}}{(\mathbf{W}^c)^2 + (\mathbf{H}^c)^2} + \frac{\mathbf{D}^w}{(\mathbf{W}^c)^2} + \frac{\mathbf{D}^h}{(\mathbf{H}^c)^2}, \quad (2)$$

where , $\mathbf{D}_{centers} \in \mathbb{R}^{N \times N}$ is a matrix containing the pairwise Euclidean distances between all N candidate RoIs, with $d_{i,j}$ being the Euclidean distance between the centers of the i-th and the j-th predicted bounding boxes. $\mathbf{W}^{c}\in\mathbb{R}^{N\times N}$ and $\mathbf{H}^{c} \in \mathbb{R}^{N \times N}$ contain the width and the height, respectively, of the smallest enclosing box covering each pair of candidate ROIs. Thus, $w_{i,j}^c$ and $h_{i,j}^c$ are the width and height of the smallest bounding box that can contain both the *i*-th and the *j*-th RoI. $\mathbf{D}^{w} \in \mathbb{R}^{N \times N}$ contains the pairwise Euclidean distances between the widths, whereas $\mathbf{D}^{h} \in \mathbb{R}^{N \times N}$ contains the pairwise Euclidean distances between the heights of all Ncandidate RoIs. Similarly to X, all of the matrices mentioned above are symmetric and only their upper triangular part is important for computations and the final result is calculated using element-wise division between the nominator and the denominator, at each term of the sum in Eq. (2). Power operations in the denominators are element-wise as well. To sum up, the first term of in Eq. (2) captures information about the distance between candidate regions, while the remainder about aspect ratio.

Grouping the candidate RoIs based on their E-IoU values means that rich geometrical information, regarding the spatial interrelations and the arrangement of the detected bounding boxes in 2D pixel space, are inherently considered for grouping them more efficiently. In comparison with D-IoU, which only describes the overlapping area and the distance between candidate boxes, E-IoU captures information about the overlapping area, distance and aspect ratios between the compared RoIs. Although C-IoU also takes into account such geometrical factors, it is computed by estimating the simple difference in aspect ratio between the compared bounding boxes. Such a naive approach may not accurately reflect the actual relationship between the RoI shapes [46]. In order to handle these issues, the second and the third term in Eq. (2) capture more sufficiently the similarity between the compared bounding boxes, in terms of their aspect ratio.

IV. EXPERIMENTAL EVALUATION

This Section overviews the experimental setups used for evaluating the proposed method and discusses the evaluation results. As previously described, YOLOv5 was selected as the baseline to be improved using the proposed method.

A. Experimental Dataset

The proposed method was implemented and tested using SIXray [7], a publicly available X-ray 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 groundtruth whole-image class label annotations manually set by human security inspectors, while their ground-truth 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 [47]. Despite the fact that only images containing at least one contraband item were utilized, official training-test set split was adopted.

B. Evaluation Metrics

The effectiveness of the proposed method is measured using the precision, recall and mean Average Precision (mAP) metrics. In object detection tasks, IoU is used to measure the overlap between the predicted and the corresponding ground-truth RoI. In addition, a threshold value was 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 groundtruth label. These elementary metrics are utilized to calculate Precision and Recall:

$$Precision = \frac{TP}{TP + FP}.$$
(3)

$$Recall = \frac{TP}{TP + FN}.$$
(4)

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:

$$AP = \int_0^1 p(r) \, dr. \tag{5}$$

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

$$mAP = \frac{1}{N} \sum_{i}^{N} AP_{i}.$$
 (6)

C. Experimental Evaluation

Evaluation of all competing methods in the SIXray dataset was conducted using the mAP metric at a 0.5 IoU threshold and the average mAP value at a range of different IoU thresholds. Comparisons were made against the baseline detector implementation before it was augmented with the proposed method (default YOLOv5 with anchor boxes obtained from the COCO dataset and Greedy NMS), as well as with variations using K-Means and K-Means++ clustering for obtaining the anchor boxes, or employing basic (IoU-based) WC-NMS. Additionally, a published YOLOv5 result on SIXray is included for completeness.

Table I summarizes the accuracy of the baseline method, which achieved a precision of 92.1%, a recall of 82.3%, a mAP of 87.6% and an average mAP across different IoU thresholds (from 0.5 to 0.95) of 72.3%. Table II compares the proposed method against this baseline and against competing approaches based on YOLOv5. The first method is a competing one published in [32], using YOLOv5 with default anchor boxes, conventional Greedy NMS and a different mini-batch size during training. The next two approaches use Greedy NMS, with one of them employing K-Means-derived and one employing K-Means++-derived anchor boxes. As it can be seen, the proposed method outperforms all other approaches in terms of mAP.

Additionally, Table III presents an ablation study of the proposed method. The first three variants adopt HC-derived anchor boxes in combination with three different IoU metrics for WC-NMS, namely IoU, D-IoU and C-IoU, respectively. Evidently, D-IoU outperforms the other two metrics in terms of mAP. The last line of Table II demonstrates the results of the full proposed method, integrating both HC-based anchors and advanced E-IoU-based WC-NMS into YOLOv5, which outperforms all other (partial) variants in terms of mAP.

Table IV presents the complete evaluation of the proposed method across all classes, using HC-derived anchor boxes and the E-IoU-based WC-NMS. It outperforms the default YOLOv5-large by 2.3% in terms of mAP. Its mAP is 89.9%

 TABLE I

 BASELINE YOLOV5 RESULTS ACROSS ALL CLASSES.

	Precision	Recall	mAP	mAP50-95
Overall	0.921	0.823	0.876	0.723
Class			AP	AP50-95
Gun	0.978	0.916	0.944	0.882
Knife	0.925	0.758	0.813	0.659
Wrench	0.877	0.768	0.835	0.659
Pliers	0.919	0.845	0.916	0.728
Scissors	0.908	0.828	0.874	0.687

TABLE II Comparative Evaluation.

Method	mAP	mAP50-95
Baseline	87.6	72.3
YOLOv5 baseline of [32]	86.7	-
K-Means anchors + default NMS	88	73
K-Means++ anchors + default NMS	88.3	72.9
HC anchors + E-IoU WC-NMS (proposed)	89.9	75.7

at a 0.5 IoU threshold, while the average mAP across a range of different IoU thresholds is 75.7%. The proposed method significantly reduces the number of false predictions and is more accurate in detecting contraband items, especially in cases of occluded object detection. In Figure 6, the precisionrecall curve of the proposed framework is presented, which highlights the performance of the model in the desired task. The curve shows a high precision score for low recall values, indicating that the model is very selective in its predictions. However, as recall increases, the precision score decreases, suggesting that the model struggles with correctly classifying some samples. However, the Area Under the Curve (AUC) value indicates that the model performs robust predictions. The above findings suggest that the model exhibits potentials for use in specific applications where high precision is critical, such as contraband detection. Finally, our model was deployed in the test subset of SIXray dataset and the predictions are presented in Fig. 7. Notably, neither HC-derived anchor boxes nor E-IoU-based WC-NMS have been proposed/investigated before for object detection.

V. CONCLUSIONS

The large volume and high throughput of passengers or mailed parcels during rush hours, in airports, subways or post/customs offices, make the automated detection of contraband items in X-ray images a Big Data analysis task that is critical for public safety. This paper proposed a novel

TABLE III Ablation study.

Method	mAP	mAP50-95
HC anchors + IoU WC-NMS	89.2	75
HC anchors + D-IoU WC-NMS	89.7	75.4
HC anchors + C-IoU WC-NMS	89.5	75.2
HC anchors + E-IoU WC-NMS (full proposed)	89.9	75.7



Fig. 6. Precision-Recall curve of the proposed method.



Fig. 7. Predictions on the SIXray test subset.

 TABLE IV

 ACCURACY OF THE PROPOSED METHOD ACROSS ALL CLASSES.

Class	Precision	Recall	mAP	mAP 50-95
Overall	0.949	0.837	0.899	0.757
			AP	AP 50-95
Gun	0.977	0.945	0.971	0.917
Knife	0.954	0.789	0.841	0.692
Wrench	0.904	0.781	0.86	0.695
Pliers	0.955	0.833	0.924	0.75
Scissors	0.956	0.84	0.899	0.73

approach to improve the performance of single-stage, anchorbased object detectors in the X-ray domain. It incorporated two complementary improvements: dataset-specific hierarchical clustering of ground-truth training RoIs, so that the derived anchor boxes better match the distribution and semantic hierarchy of object sizes/shapes, and a modification of an efficient NMS algorithm, so as to better handle occluded objects and to reduce false predictions. According to a thorough experimental evaluation on a relevant public dataset, the proposed method outperforms both the baseline and various competing approaches.

Future research will focus on addressing other limitations of existing methods, such as low generalization ability under domain shifts (e.g., if the detector has been trained with X-rays from one type of scanner and is deployed in an airport using a different scanner), as well as integration of the proposed approach to more recent object detectors (e.g., YOLOv7).

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

- 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.
- [2] 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.
- [3] 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), July 2017.
- [4] C. Symeonidis, I. Mademlis, N. Nikolaidis, and I. Pitas, "Improving neural Non-Maximum Suppression for object detection by exploiting interest-point detectors," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.
- [5] C. Symeonidis, I. Mademlis, I. Pitas, and N. Nikolaidis, "Neural attention-driven Non-Maximum Suppression for person detection," *IEEE Transactions on Image Processing*, 2023, accepted for publication.
- [6] 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.

- [7] 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.
- [8] 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.
- [9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [11] 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.
- [12] 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.
- [13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
- [14] 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.
- [15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision– ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14.* Springer, 2016, pp. 21–37.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] 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.
- [20] 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*, 2017, pp. 2117–2125.
- [21] 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, pp. 3–19.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] C. Zhou, H. Xu, B. Yi, W. Yu, and C. Zhao, "X-ray security inspection image detection algorithm based on improved YOLOv4," in *Proceedings* of the IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE), 2021.
- [26] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

- [27] B. Li, Y. Liu, and X. Wang, "Gradient harmonized single-stage detector," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.
- [28] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: improving object detection with one line of code," in *Proceedings of* the IEEE International Conference on Computer Vision (ICCV), 2017.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] 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.
- [33] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," 2021.
- [34] G. Jocher, "YOLOv5 by Ultralytics." [Online]. Available: https: //github.com/ultralytics/yolov5
- [35] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.
- [36] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.
- [37] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
- [38] 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.
- [39] 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*, 2018, pp. 8759–8768.
- [40] 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.
- [41] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," Stanford, Tech. Rep., 2006.
- [42] X. Luo, Y. Wu, and F. Wang, "Target detection method of UAV aerial imagery based on improved YOLOv5," *Remote Sensing*, vol. 14, no. 19, p. 5063, 2022.
- [43] S. Landau, M. Leese, D. Stahl, and B. S. Everitt, *Cluster analysis*. John Wiley & Sons, 2011.
- [44] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.
- [45] H. Zhou, Z. Li, C. Ning, and J. Tang, "CAD: Scale-invariant framework for real-time object detection," in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017.
- [46] Y.-F. Zhang, W. Ren, Z. Zhang, Z. Jia, L. Wang, and T. Tan, "Focal and efficient IOU loss for accurate bounding box regression," *Neurocomputing*, vol. 506, pp. 146–157, 2022.
- [47] 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.