Toward Human-in-the-Loop Prohibited Item Detection in X-ray Baggage Images

1st Sisi Cao; 3rd Wenwen Song
School of Software Engineering
Xi'an Jiaotong University
Xi'an, China
css2017@stu.xjtu.edu.cn; wa1ter@stu.xjtu.edu.cn

2nd Yuehu Liu; 4th Zhichao Cui
Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University
Xi'an, China
liuyh@mail.xjtu.edu.cn; cui.zhichao@stu.xjtu.edu.cn

5th Xiaojun Lv
Institute of Computing Techonologie
China Academy of Railway Science Corporation Limited
Beijing, China
Lvxiaojun@263.net

6th Jingwei Wan
Department of Criminal Science and Technology
Railway Police College
Zhengzhou, China
wanjingwei@rpc.edu.cn

***Abstract*—X-ray baggage security screening is a demanding task for aviation and rail transit security; automatic prohibited item detection in X-ray baggage images can help reduce the work of inspectors. However, as many items are placed too close to each other in the baggages, it is difficult to fully trust the detection results of intelligent prohibited item detection algorithms. In this paper, a human-in-the-loop baggage inspection framework is proposed. The proposed framework utilizes the deep-learning-based algorithm for prohibited item detection to find suspicious items in X-ray baggage images, and select manual examination when the detection algorithm cannot determine whether the baggage is dangerous or safe. The advantages of proposed inspection process include: online to capture new sample images for training incrementally prohibited item detection model, and augmented prohibited item detection intelligence with human-computer collaboration.**

The preliminary experimental results show, human-in-the-loop process by combining cognitive capabilities of human inspector with the intelligent algorithms capabilities, can greatly improve the efficiency of in-baggage security screening.

***Index Terms*—X-ray baggage security screening, human-in-the-loop baggage inspection, prohibited item detection**

I. Introduction

Security inspection aims at protecting the safety of the passengers and goods, which has been widely applied in the transportation industry such as railways and airplanes [1]. At present, this inspection work is carried out by human inspectors through the X-ray security inspection machine. The manual inspection shows the real disadvantages of high labor consumption and low efficiency, especially in some stations under the situation of large passenger flow [2]. With the advent of advanced deep learning method, many researchers [3]–[5] utilize the object detection technology [6]–[13] to find the prohibited items in X-ray baggage images. Though great improvements have been made by using deep-learning-based detection methods, current prohibited item detection methods [4], [5] can not be applied directly in real-world inspection scenarios, as X-ray baggage images are complicated, with items occluded by each other.

Therefore, this paper focuses on how to apply the auto-detection algorithms in the inspection process, so as to speed up the inspection process while maintaining the accuracy of manual inspection. To fulfill this, the idea of human-in-the-loop (HITL) augmented intelligence [14] is introduced to form the proposed HITL inspection framework. The proposed framework, shown in Fig. 1, integrates human intelligence and prohibited item detection algorithm to achieve an inspection process with human-computer collaboration. Its work pipeline is as follows. The prohibited item detection model takes each X-ray baggage image as input and outputs a set of prohibited item proposals, each with an objectness confidence score. The confidence scores determine whether the baggage is safe or not. When the detection model is not sure, the intervention of human inspectors is required.

The prohibited item detection algorithm plays an important part in our inspection framework. Different from general objects such as pedestrians and vehicles, items in X-ray baggage image are usually part-visible due to item viewpoints variety and item overlapping, which makes current detection methods are not suitable for prohibited item detection. This paper proposes a prohibited item detection algorithm to detect the part-visible prohibited items by introducing a key part detection branch.

The main contributions can be summarized as follows:

1) A human-in-the-loop baggage inspection framework is proposed. The framework integrates human and prohibited item detection algorithm to a trusted inspection process.
2) A prohibited item detection algorithm is proposed. Inspired from the process of human inspectors checking the X-ray baggage images, we add a detection branch to detect the key parts of prohibited items on the basis of current detection frameworks.

978-1-7281-4094-0/19/$31.00 ©2019 IEEE

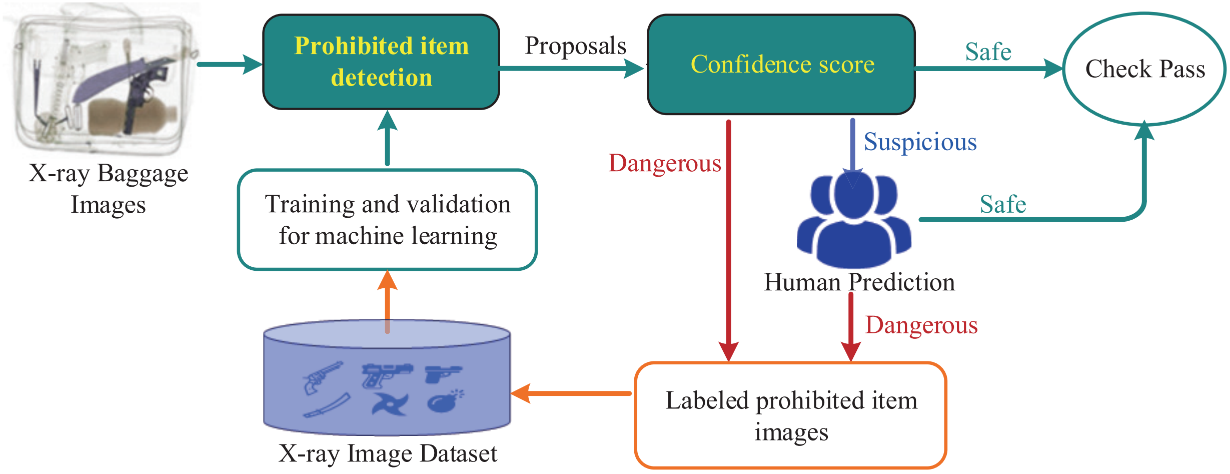

Fig. 1. An implementation framework for man-machine coordination inspection system. This framework integrates the human and prohibited detection algorithm by the introduction of human interaction. It uses the detection algorithm to learn a prohibited detection model from the training dataset and predicts new X-ray baggage images by using the model. When the model is not confident about its predictive result, human inspectors will interpose to make judgments. Then these X-ray baggage images with human's judgments will be used to update the detection model.

II. Related Work

A. Hybrid-augmented Intelligence

In the year of 2017, Zheng et al. [14] proposed the concept of hybrid-augmented intelligence, that is, introducing human cognitive abilities or human-like cognitive models into artificial intelligence (AI) systems to form a new type of AI. Augmented-intelligence-based cognitive computing and HITL hybrid-augmented intelligence with human-computer collaboration are two basic models.

They define the model of HITL hybrid-augmented intelligence as an intelligent system requiring human interaction. In such a system, human always plays an important role and affects the outcome of the system. If the results given by the machine are relative unreliable, the human will make further judgments. With human interaction, the human intelligence is introduced into the loop of an intelligent system. This can achieve the close coupling between the analysis of the fuzzy and uncertain problems and the machine intelligence system with the high-level cognitive mechanism .

The HITL hybrid-augmented intelligence framework usually works as follows: when the intelligence system is abnormal, or the machine is not confident in its judgment, the confidence estimation or the cognitive load state of the machine will determine whether the intervention of human is required, and the intelligence system is updated automatically. The human intervention in the system improve the accuracy and credibility of the intelligence system's predictions.

B. Object Detection Technology

Object detection is a classic and important task in computer vision, of which lots of algorithms have been proposed in the past few decades. The existing object detection methods can be classified into two categories, which are traditional object detection methods [15]–[17], and deep-learning based object detection methods.

Traditional object detection methods are built on the handcrafted features, in which Viola-Jones (VJ) detector [15], Histogram of Oriented Gradients (HoG) detector [16], and Deformable Part-based Model (DPM) [17] are the representative methods. Thanks to the powerful image representation ability of deep convolutional neural network (DCNN), deep-learning based object detection methods [6]–[13] have far surpassed the traditional methods, in terms of detection accuracy. In the deep learning area, object detection methods can also be split into two groups: single-stage detection [10]–[13] and two-stage detection [6]–[9], [18].

In two-stage detection methods, R-CNN [6] is first proposed. It first uses selective search [19] to generate a set of object proposals. Then each proposal is resized to a fixed-size image patch and put into the DCNN to extract features. Finally, linear support vector machines (SVM) is used to classify each proposal feature into its corresponding category. Later, SPPNet [20], Fast RCNN [7] and Faster R-CNN [18] develop the two-stage detection performance. Faster R-CNN has been a popular detection method. It proposes region proposal network (RPN) and trains the detector and a bounding boxes regressor in the end-to-end fashion. Cascade R-CNN [8] further extends the work of Faster R-CNN to form a multi-stage detector with cascade architecture. Mask R-CNN [9] improves the performance of Faster R-CNN by learning with semantic segmentation.

For faster detection, single-stage detection methods have been developing. YOLO [10] and SSD [11] are two typical single-stage methods. RetinaNet [12] addresses the problem of foreground-background class imbalance in single-stage detection methods and propose "focal loss" to achieve comparable accuracy with two-stage detection methods. CornerNet [13] proposes a new insight that an object bounding box can be predicted as a pair of key points.

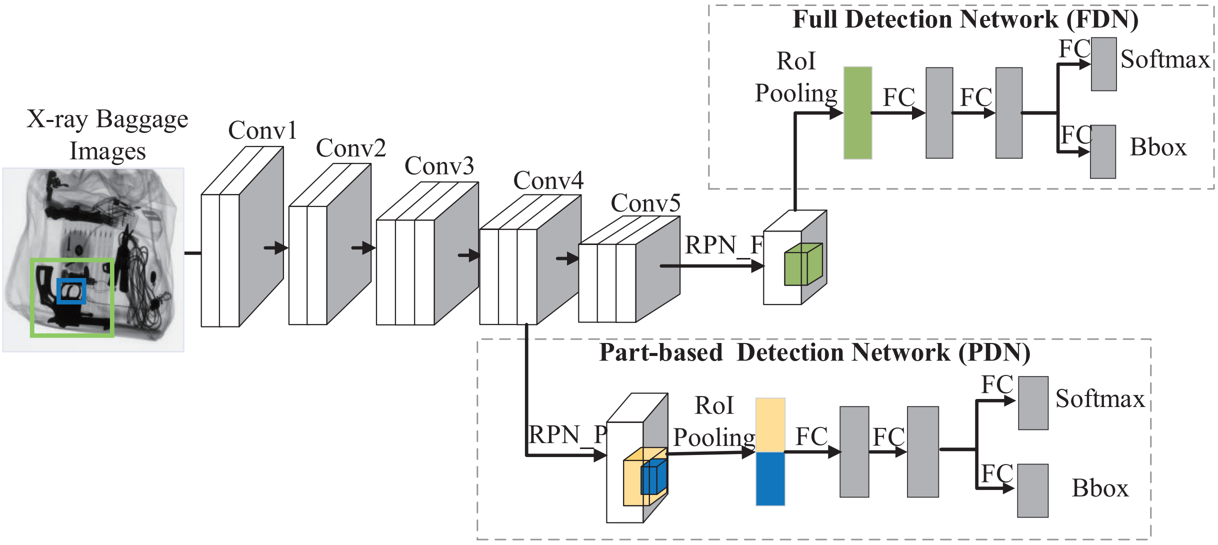

Fig. 2. The architecture of our detection network Xray R-CNN. Based on the framework of Faster R-CNN, Xray R-CNN introduce a new detection branch (PDN) aiming for detecting the key parts of prohibited items. The original branch (FDN) is used to detect the whole body of prohibited items. The two branches share the feature extraction network. Different from the FDN branch, the PDN branch uses the shallow convolution features and context information to detect small parts.

III. Proposed Prohibited Item Detection Method

The challenge of prohibited object detection is that objects in X-ray baggage image are usually part-visible due to object viewpoints variety and object overlapping, which makes the visual appearance of prohibited items varies largely in X-ray baggage images. Current popular detection methods cannot solve this challenge.

To find a solution for this challenge, we observe the process of human inspectors' checking the X-ray baggage images. We find that many human inspectors make their inferences only by the partial appearance of the object exposed to the image.

Inspired by this, we propose a new detection network, Xray R-CNN. Our network judges an object by considering not only its full appearance but also its partial appearance. When the full appearance of an object is not accessible, our network can make an inference by its partial appearance, just like a human inspector.

A. Detection Network Structure

Our method adopts Faster R-CNN [18] as the basis detection framework, which is a popular two-stage detection framework. Faster R-CNN consists of three parts: feature extraction network, region proposal network (RPN), classification and regression network. Among them, the feature extraction network is used to extract the features of the input images, RPN is used to generate object proposals, and the classification regression network to classify the proposals and refine the coordinates of object proposals.

A new detection branch called Part-based Detection Network (PDN) is added to the Faster R-CNN, shown in Fig. 2, which is responsible for detecting the key parts of the prohibited items. The original branch, which is named as Full Detection Network (FDN) in this paper, is responsible for detecting the prohibited items. Each branch of the network has a regional proposal network, a classification and regression network. Two branches of PDN and FDN share the same feature extraction network.

Due to the factor that key parts are smaller compared to the whole bodies, we input the shallow convolution features into the PDN branch. Besides, we follow the work in [21] to introduce the features of the area around the key parts as the context information of these parts. Different from [21], we do not introduce additional convolution layer. Besides, the size of the surrounding area adopted as context information is set twice as large as that of the key part areas. More specifically, as shown in Fig. 2, an X-ray baggage image is first fed into the feature extraction network, and the RPN_P generates the proposals of key parts with the shallow convolution features. Then, we enlarge the width and height of each part proposal by twice. After that, each part proposal and the corresponding enlarged proposal are input to the RoI Pooling layer to output the part feature and its context information. Finally, each part feature and its context information are concatenated together to obtain a final part feature with a size of 7×7 and a feature depth of 1024.

B. Training Method

To train the proposed network, human beings are first needed to label the bounding boxes for the X-ray baggage images to form the labeled training dataset. Then an alternating training method is used to train the detection network with the labeled training dataset. After renewal, the detection network is utilized to infer the unseen X-ray baggage images. Subsequently, the prediction results are fed back to human beings to make further judgments. These unseen images which the detection network predict wrongly are labeled by human beings and used to train the network. The alternating training method consists of three steps.

Step 1: training the feature extraction network (Convolution Layers) and FDN branch, and saving the weight of the detection network;

Step 2: loading the detection network saved in step 1 and training the feature extraction network and PDN branch.

Step 3: keeping the weights of the feature extraction network fixed, and fine-tune other layers of the proposed detection network.

IV. Experiment

A. Data Acquisition and Properties

The proposed method is evaluated on a public X-ray image dataset GDXray. The GDXray data set contains a total of 19,407 X-ray images. The parcel is selected as experimental data, specifically “b0009-b0044” series, and “b0046-b0048” series. In those images, three categories of prohibited items to detect is chosen finally, which are gun, shuriken, and knife, with 966 images in total. All chosen images are split into two subsets: a training set with 352 images and a test set with 614 images. To augment the dataset, the training images and test images are randomly rotated. TABLE I lists the class distribution information of the training and test set after augmentation.

TABLE I
The class distribution of the training set and test set after augmentation

Categories	Gun	Shuriken	Knife	Images
Training set	1048	388	160	1408
Test set	1560	705	246	1842
Total	2608	1093	406	3250

The training set is labeled with not only the full bodies of prohibited items but also their key parts, which are pre-defined. The key part annotations are shown in Fig. 3. For the test set, only the full bodies of prohibited items are labeled.

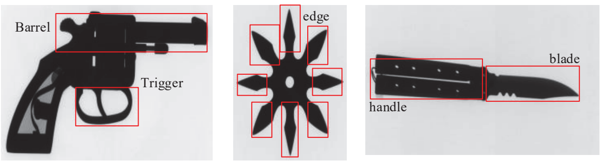

Fig. 3. The part annotations of the training set.

B. Performance of Proposed Method

1) Evaluation of Proposed Detection Method: To verify the effectiveness of the proposed prohibited items detection method, this section compares the performance with the original method of Faster R-CNN. Results are listed in TABLE II. After introducing a part-based detection branch, proposed detection method achieves 78.82% mAP, improving Faster R-CNN method by 13.62pp mAP. This confirms the effectiveness of proposed detection method. The proposed method can also be applied in many other popular detection methods, such as YOLOv3 [22], and SSD [11].

TABLE II
Detection results on GDXray dataset

Methods	Gun	Shuriken	Knife	mAP
Faster RCNN	74.40	77.59	46.30	65.20
Proposed	84.18	89.97	62.32	78.82

2) Evaluation of Proposed Inspection Framework: The framework is first evaluated based on proposed detection method, and then the ideal performance of proposed framework is given. To verify the effectiveness of proposed inspection framework, we choose 608 X-ray baggage images without any prohibited items, and 1842 X-ray baggage images, with each, contains one or more than one prohibited items.

Classification results of proposed framework based on proposed detection model. Each X-ray baggage image is fed into proposed detection model, and a set of prohibited item proposals with confidence scores are output. For each image, if there exists a proposal whose confidence score is higher than 0.9, it will be classified as a dangerous baggage; if all proposals’ confidence scores are lower than 0.5, it will be classified as a safe baggage; otherwise, it will be classified as a suspicious baggage. The predicted results are listed in TABLE III. The qualitative results are shown in Fig. 4 .

TABLE III
Classification results based on the proposed detection model.

Predicted Real	Safe	Suspicious	Dangerous	Total
Legal	282	149	177	608
Illegal	33	188	1621	1842
Total	315	337	1798	2450

Michel et al. [23] stated that the inspection accuracy of human was $80-90\%$ due to the fatigue of inspectors after long inspection work. When using proposed inspection framework, human inspectors only need to check 337 X-ray baggage images instead of 2450 images. This can greatly reduce the work of human inspectors and increase the inspection accuracy of human inspectors.

suppose the inspection accuracy rate of human is 100% after using proposed inspection framework, then the accuracy of the proposed framework can be computed by

$$\frac{282+1621+149\times100\%+188\times100\%}{2450}=91.43\%$$

Similarly, the precision and recall of the proposed framework can be computed by

$$\text{precision}=\frac{1621+188\times100\%}{1621+188\times100\%+177}=91.09\%$$
$$\text{recall}=\frac{1621+188\times100\%}{1842}=98.21\%$$

The proposed detection model needs 1.2 seconds (s) to predict a baggage image. Suppose a human inspector’ checking an X-ray baggage image needs 2 seconds on average. The detection time per image of proposed framework can be approximately estimated by

$$T_{proposed} = (149+188)\times(2+1.2)+(282+33+177+1621)\times 1.2 = 3614s$$

The human inspectors need 4900s to check all these baggages. Even though the results of detection model are not absolute right, proposed framework can still speed up the inspection process and improve the accuracy of inspection.

Ideal classification performance of proposed framework. A relative ideal detection algorithm should classify a legal baggage into the safe or suspicious category and a illegal baggage into the dangerous or suspicious category. When such a relative ideal detection model is used, proposed framework can gain better performance. For example, TABLE IV shows the results by supposing proposed detection algorithm is ideal. Only 547 suspicious baggages need to be checked by human inspectors. And other baggages are classified correctly by the detection model. The classification accuracy of proposed framework achieves 100%. Under this circumstance, the proposed framework can largely release the work of human inspectors, speed up the inspection process, and increase the accuracy of inspection.

⋆ In Summary, from the above experimental results, we can get the conclusion that human-in-the-loop process by combining cognitive capabilities of human inspector with the intelligent algorithms capabilities, can greatly improve the efficiency of in-baggage security screening.

TABLE IV
IDEAL CLASSIFICATION RESULTS OF PROPOSED FRAMEWORK.

Predicted Real	Safe	Suspicious	Dangerous	Total
Legal	282	326	0	608
Illegal	0	221	1621	1842
Total	282	547	1621	2450

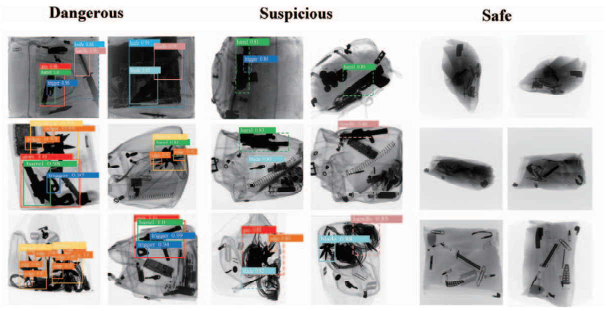

Fig. 4. Simulation results; The images from left to right are dangerous (the first and second columns), suspicious (the third and fourth columns) and safe (the fifth and sixth columns).

V. Conclusion

This paper proposes a human-in-the-loop baggage inspection framework. Besides, an algorithm for prohibited item detection is proposed, which has been proved its effectiveness by the experiments. Based on proposed detection algorithm, the performance of the proposed framework is examined. The results show that human-in-the-loop process by combining cognitive capabilities of human inspector with the intelligent algorithms capabilities, can greatly improve the efficiency of in-baggage security screening.

Acknowledgment

This paper was supported in part by National Key R&D Program of China (2018YFC0823602) and National Natural Science Foundation of China (201805208).

References

[1] M. Bastan, W. Byeon, and T. Breuel, "Object recognition in multi-view dual energy x-ray images," in *British Machine Vision Conference*, 2013.

[2] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for x-ray testing in baggage inspection," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 682–692, April 2017.

[3] T. Franzel, U. Schmidt, and S. Roth, "Object detection in multi-view x-ray images," in *Pattern Recognition*, A. Pinz, T. Pock, H. Bischof, and F. Leberl, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 144–154.

[4] 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 *CVPR*, 2019.

[5] D. M. Kundegorski M E, Akay S, "On using feature descriptors as visual words for object detection within x-ray baggage security screening," in *International Conference on Imaging for Crime Detection & Prevention*, 2016.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 580–587.

[7] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1440–1448.

[8] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6154–6162.

[9] K. He, G. Gkioxari, P. Dollr, and R. Girshick, "Mask r-cnn," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2980–2988.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015.

[12] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2999–3007.

[13] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," *CoRR*, vol. abs/1808.01244, 2018.

[14] N. ning ZHENG, Z. yi LIU, P. ju REN, Y. qiang MA, S. tao CHEN, S. yu YU, and J. ru XUE, "Hybrid-augmented intelligence:collaboration and cognition," *Frontiers of Information Technology & Electronic Engineering*, vol. 18, no. 2, pp. 153–179, 4 2017.

[15] P. VIOLA, "Robust real-time object detection," *International Journal on Computer Vision and Image Understanding*, vol. 57, no. 2, pp. 137–154, 2004.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, June 2005, pp. 886–893 vol. 1.

[17] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[18] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, June 2017.

[19] U. J. R. R., K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] 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, Sep. 2015.

[21] Z. Cai, Q. Fan, R. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *ECCV*, 2016.

[22] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[23] S. Michel, S. M. Koller, J. C. de Ruiter, R. Moerland, M. Hogervorst, and A. Schwaninger, "Computer-based training increases efficiency in x-ray image interpretation by aviation security screeners," in *2007 41st Annual IEEE International Carnahan Conference on Security Technology*, Oct 2007, pp. 201–206.