AN EVALUATION OF REGION BASED OBJECT DETECTION STRATEGIES WITHIN X-RAY BAGGAGE SECURITY IMAGERY

Samet Akcay, Toby P. Breckon
Durham University, Durham, UK

ABSTRACT

Here we explore the applicability of traditional sliding window based convolutional neural network (CNN) detection pipeline and region based object detection techniques such as Faster Region-based CNN (R-CNN) and Region-based Fully Convolutional Networks (R-FCN) on the problem of object detection in X-ray security imagery. Within this context, with limited dataset availability, we employ a transfer learning paradigm for network training tackling both single and multiple object detection problems over a number of R-CNN/R-FCN variants. The use of first-stage region proposal within the Faster RCNN and R-FCN provide superior results than traditional sliding window driven CNN (SW-CNN) approach. With the use of Faster RCNN with VGG16, pretrained on the ImageNet dataset, we achieve 88.3 mAP for a six object class X-ray detection problem. The use of R-FCN with ResNet-101, yields 96.3 mAP for the two class firearm detection problem requiring 0.1 second computation per image. Overall we illustrate the comparative performance of these techniques as object localization strategies within cluttered X-ray security imagery.

***Index Terms*—** Object detection, deep learning, X-ray baggage security

1. INTRODUCTION

X-ray baggage security screening is widely used to maintain aviation and transport security, itself posing a significant image-based screening task for human operators reviewing compact, cluttered and highly varying baggage contents within limited time-scales. With both increased passenger throughput in the global travel network and an increasing focus on wider aspects of extended border security (e.g. freight, postal), this poses both a challenging and timely automated image detection task.

Aviation security screening systems including bulk and vapor detection are of interest and have been studied for decades [1]. Computer aided screening (CAS) that performs automated recognition, however, remains an unsolved problem. Previous work [2, 3] has focused on image enhancement [4–6], segmentation [7, 8], classification [9–13] or detection [14–17] tasks in order to further investigate the real time applicability of CAS to automatize aviation security screening.

For the classification of X-ray objects, the majority of the papers in the literature proposes traditional machine learning approaches based on Bag-of-Visual-Words (BoVW) representation scheme using hand crafted features together with a classifier such as Support Vector Machine (SVM) [9–13].

Kundegorski *et al.* [13] exhaustively explores the use of various feature point descriptors as visual word variants within a BoVW model for image classification based threat detection within baggage security X-ray imagery with a FAST-SURF feature detector and descriptor combination giving maximal performance with SVM classification (2 class firearm detection: 94.0% accuracy).

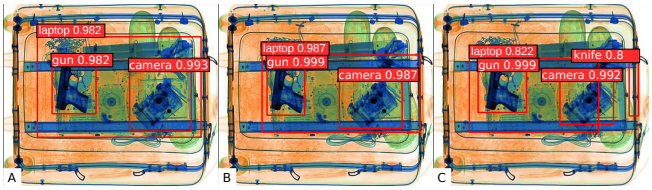

Fig. 1: Exemplar X-ray baggage imagery containing firearms, and detections with bounding boxes.

The study of [18] compares BoVW approach and CNN exploring the use of transfer learning approach by fine tuning weights of varying layers transferred from another networks trained for a different task. CNN outperforms BoVW method even when there is no fine tuning at all, meaning the network are not trained for firearm detection, but transferred directly from the source task. Outperforming the previous state of the art, CNN, based on AlexNet [19] architecture, is shown to perform 98.92% and 97.6% (accuracy) the same 2-class (firearms) and six class detection problems considered here (patch wise evaluation).

The work of [11] proposes a model that performs recognition of objects using multi-view X-ray images achieving $\%93$ recognition accuracy for 12 objects that consist of clips, springs and razor blades.

Object classification is a significant task for the identification of particular object against the others, ie. illicit vs. non-illicit objects. One step further of this task is detection in which objects are localized with a bounding box. Being a challenging task, detection based models within X-ray baggage imagery are rather limited in literature.

In [14], detection of region of interests (ROI) in X-ray images is performed via geometric model of the object by using structure from motion. Potential regions obtained from segmentation step are then tracked based on their similarity, achieving 94.3% true positive and 5.6% false positive rates on a small uncluttered dataset.

Franzel *et al.* [15] also propose a sliding window detection approach with the use of linear SVM classifier and histogram of oriented gradients (HOG). As a next step, called multi-view integration, detections of single view X-ray images are fused to avoid false detections and find the intersection of the true detections. Multi-view detection is shown to provide superior detection performance than single view detection for handguns (mAP: 64.5).

Similarly, [17] explores object detection in X-ray baggage

imagery by evaluating various hand crafted feature detector and descriptor combinations with the use of branch and bound algorithm and structural SVM classifier (mAP: 88.1 for 6400 images of handguns, laptops and glass bottles).

By contrast, here we look to expand and explore the CNN driven work of [18] investigating both the use of a sliding window paradigm (akin to [15,17]] and evaluating contemporary approaches to learn object localization via R-CNN and R-FCN approaches. As shown in prior work [18] the challenging and cluttered nature of object detection in X-ray security imagery often poses additional challenges for established classification and detection approaches like RCNN/R-FCN [20,21].

2. OBJECT DETECTION

Convolutional neural networks (CNN) have emerged as the state of the art for classification in the field of computer vision especially after Krizhevsky *et al.* [19] successfully proposed their network. Since then, many models have been proposed achieving promising results in classification [22–24], detection [20, 21, 25–31] and segmentation [24, 26] tasks in challenging competitions such as ILSVRC [32] and COCO [33]. These methods have also been employed in several other tasks with small dataset availability [34]. [3,18,35] use CNN in the task of object classification in X-ray security imagery, and achieve encouraging results especially with the use of transfer learning paradigm. This approach provides promising performance especially for single and non-occluded objects in a given image. When it comes to classifying multiple objects, however, more sophisticated approaches are needed in order for localization and detection.

Sermanet *et al.* (OverFeat) [25] uses sliding window approach to generate the region proposals, which is then fed into a convolutional neural network for the classification. The key idea here is that bounding box regression is performed with an extra regression layer which shares the weights with main network. [26] proposes a detection algorithm (RCNN), based on three main stages: region proposal generation, feature extraction and classification. First stage employs an external region proposal generator, followed by a fine tuned CNN in the next stage for feature extraction. The last stage performs classification with an SVM classifier. Even though outperforming the previous work by a large margin, model is not considered to be real time applicable due to runtime and memory issues. SPPNet [27] contains spatial pooling layer between convolutional and fully connected layers, which allows network being fed with images with varying scales and aspect ratios. With this design, image representations can be computed once in SPPNet, which makes the network significantly faster than RCNN. Like RCNN, however, the network has several separate stage, which is computationally expensive, and requires memory. Fast RCNN by Girshick [28] combines feature extraction, classification and bounding box regression stages by designing a partially end to end CNN network, significantly outperforming [26, 27] in terms of speed and accuracy. The novelty of the work is to employ a region of interest pooling layer (RoI) before fully connected layer (fc) to have fixed size region proposals. Besides, using a multi-task loss function class probabilities and bounding box regressions are computed in a single network. The limitation, however, is that the network still needs an external region proposal algorithm such as selective search [36]. Inspired by the strong and weak points of [26–28], Ren *et al.* [20] propose a model performing all the aforementioned stages in an end to end network. In so doing not only reduces time complexity and required memory but also significantly boosts the accuracy. [21] propose fully convolutional detection framework (RFCN), which yields faster training and testing performance with competitive accuracy compared to Faster RCNN [20].

Providing a significant boost in accuracy, Fast RCNN and RFCN models are adapted in this work to use within X-ray baggage object detection context, and to compare to previous object detection approaches mainly based on sliding window alike detection frameworks [15, 17]

2.1. Detection Strategies

Within this work we consider a number of such detection frameworks and explore their applicability and performance for object detection in X-ray baggage imagery.

Sliding Window Based Approaches generate region proposals by sliding a fixed size window through the image. With a combination with image pyramids, this method can also generate proposals with varying sizes and scales. Figure 2A demonstrates that sliding window with image pyramids generates region proposals with varying scales, which are then fed into a CNN for the classification step.

Faster RCNN is based on a two subnetworks, containing a unique region proposal network (RPN) and Fast RCNN network together. Instead of utilizing an external region proposal algorithm as done previously in [26, 28], this model has its own region proposal network, which is the main difference than Fast RCNN. The RPN consists of convolutional layers that generate object proposals and two fully connected layers that predict coordinates of bounding boxes and the associated probability of it being a target object (belonging to one of the classes upon which the network is trained). Convolutional layers in RPN are shared with Fast RCNN network, which makes region proposal step cost efficient compared to using an external region proposal algorithm. Features of object candidates are generated by sliding a window over the feature map of the last convolutional layer. Using these features, the last two fully connected layers (bounding box regression and classification) then generate region proposals. RoI pooling layer resizes the proposals to have fixed sized width and length. fc layers then create feature vector to be used by bounding box regression and softmax layers.

R-FCN is proposed by Dai et al. [21] by pointing out the main limitation of Faster RCNN that each region proposal within RoI pooling layer is computed hundreds of time due

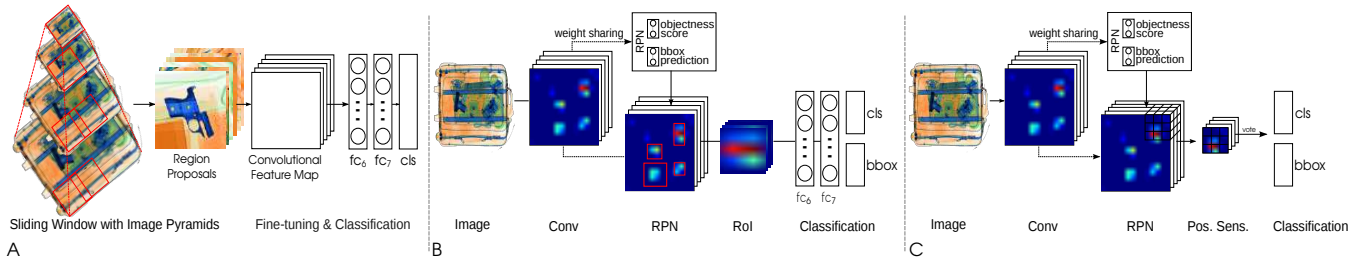

Fig. 2: Schematics of the CNN driven detection strategies evaluated (right to left: Sliding Window based CNN (SW-CNN) [13, 18], Faster RCNN (FRCNN) [20], R-FCN [21]).

to the two subsequent fully connected layers, which is computationally expensive (Figure 2B). They propose a new approach removing fully connected layers after RoI pooling, and employing a new approach called *"position sensitive score map"* [21], which handles translation variance issue in detection task (Figure 2C). Since no fully connected subnetwork is used, the proposed model shares weights within almost entire network. This leads to much faster convergence both in training and test stages, while achieving similar results to Fast RCNN.

2.2. Detection Architectures

We compare three localization architectures for our object detection task within X-ray security imagery:- a traditional sliding window approach [15] coupled with CNN classification [25], the Faster RCNN (FRCNN) approach of [20] (a contemporary leading architecture within recent object recognition challenge results [32, 33]) and the R-FCN approach of [21] (comparable to FRCNN in performance yet offering significant computational efficiency gains over the former).

Dataset: To investigate the performance on object detection within in X-ray baggage imagery, we address binary class (firearms vs background), and multiple class X-ray object detection problems (6 classes: firearm, firearm-components, knives, ceramic knives, camera and laptop). We use a dataset constructed using multi-view conventional X-ray imagery with associated false color materials mapping (from dual energy, [2] see Figure 1). The dataset contains 11,627 samples (5,867 training, 2,880 validation and 2,880 test samples).

Sliding Window (SW-CNN): We employ 800x800 input image, 50x50 fixed size window with a step size of 32 to generate region proposals. We also use image pyramids to fit the window to varying sized of objects. For the classification of the proposed regions we use AlexNet [19], $\text{VGG}_{\{\text{M, 16}\}}$ [22], and ResNet-{50, 101} [24] networks. Although [25] employs an extra bounding box regression layer within SW-CNN approach, we do not perform regression as none of the prior work within this domain does so.

Faster RCNN: Using the original implementation with a few modifications, we train Faster RCNN with AlexNet [19], $\text{VGG}_{\{\text{M, 16}\}}$ [22], and ResNet-{50, 101} [24] architectures.

R-FCN: Since R-FCN is fully convolutional by design, we use ResNet-{50, 101} [24] networks for the feature extraction step.

For the training of the models, we employ transfer learning approach, and use networks pre-trained on ImageNet dataset. In so doing not only increases the performance but also reduces training time significantly. We use stochastic gradient descent (SGD) with momentum and weight decay of 0.9 and 0.0005, respectively. The initial learning rate of 0.001 is divided by 10 with step down method in every 10,000 iteration. For FRCNN/R-FCN, batch size is set to 256 for the RPN. All of networks are trained using dual core Intel Xeon E5-2630 v4 processor and Nvidia GeForce GTX Titan X GPU.

3. EVALUATION

Performance of the models is evaluated by PASCAL VOC metrics [37], using average precision metric (AP) for the evaluation of each class, and mean average precision (mAP) for the overall performance.

Table 1 and 2 show binary and multi-class detection results for SW-CNN, Faster-RCNN and R-FCN with varying networks, and with a fixed sized number of region proposals of 300. For completeness, we additionally present the comparative results for Fast R-CNN (RCNN) [28] (detection architecture pre-dating that of Faster R-CNN [20] and R-FCN [20]).

As a general trend we observe that performance increases with overall network complexity such that superior performance is obtained with VGG16 and ResNet-101 for all of the approaches. This observation holds for both the 2-class and 6-class problems considered here.

For firearm detection, Table 1 shows that SW-CNN even with a complex network such as VGG16 and ResNet-101 performs poorer than any other detection approaches. This is mainly due to not employing a bounding box regression layer (Figure 2), a significant performance booster as shown in [25, 26]. Likewise, best performance of RCNN with VGG16 (mAP: 85.4) is worse than any FRCNN or R-FCN. This is because the RPN within FRCNN and R-FCN provides superior object proposals than selective search used in RCNN. For the overall performance of firearm detection, R-FCN with ResNet-101 yields the highest mAP of 96.3.

For the multi-class detection task (Table 2) we see a similar pattern to firearm detection problem: SW-CNN performs worse than any network trained for Faster RCNN or R-FCN.

Model	Network	mAP	Firearm
SW-CNN	AlexNet	75.3	75.3
	VGGM	77.2	77.2
	VGG16	80.6	80.6
	ResNet-50	83.6	83.6
	ResNet-101	84.7	84.7
RCNN	AlexNet	82.3	83.2
	VGGM	83.6	83.6
	VGG16	85.4	85.4
FRCNN	AlexNet	94.5	94.5
	VGGM	94.8	94.8
	VGG16	96.0	96.0
	ResNet-50	95.1	95.1
	ResNet-101	96.0	96.0
R-FCN	ResNet-50	94.9	94.9
	ResNet-101	96.3	96.3

Table 1: Detection results of SW-CNN, Fast-RCNN (RCNN) [28], Faster RCNN (FRCNN) [20] and R-FCN [21] for firearm detection problem (300 region proposals).

By the same token, overall mAP of RCNNs is lower than any RFCN and R-FCN. For the comparison of FRCNN and R-FCN, we observe that Faster RCNN achieves the highest peak using VGG16, with higher mAP than ResNet-50 and ResNet101. R-FCN with ResNet-50 and ResNet-101 yields slightly worse performance, (mAP: 84.6, 85.6) , than that of the best of Faster-RCNN. For the overall performance comparison, Faster RCNN with VGG16 shows superior performance (mAP: 88.3). Figure 1 and 3 show parallel results with Table 2. For a moderate sample (Figure 3 A-C) all of the approaches are able to detect the objects. For an entirely different sample (assault rifle) not within the training set (Figure 3 D-F), FRCNN has higher confidence in localizing the objects. There are, of course, isolated samples (Figure 3 G-I) such that neither of the models manages to detect.

Model	Network	mAP	camera	laptop	gun	gun component	knife	ceramic knife
SWCNN	AlexNet	60.8	68.2	60.9	74.8	71.4	21.2	68.3
	VGGM	63.4	70.7	63.7	76.3	73.1	24.6	71.9
	VGG16	64.9	70.1	72.4	75.2	75.7	22.3	73.4
	ResNet-50	67.1	69.2	80.1	74.7	76.1	31.4	71.3
	ResNet-101	77.6	88.1	90.2	83.1	84.8	39.2	80.3
RCNN	AlexNet	64.7	79.1	81.5	85.3	58.2	18.8	65.8
	VGGM	68.6	79.9	85.5	86.9	65.8	21.0	72.3
	VGG16	77.9	88.8	95.4	87.6	83.2	30.4	81.9
FRCNN	AlexNet	78.8	89.3	75.6	91.4	87.4	46.7	82.3
	VGGM	82.3	90.0	83.4	91.8	87.5	54.2	86.9
	VGG16	**88.3**	88.1	91.8	92.7	93.8	72.1	91.2
	ResNet-50	85.1	84.4	87.9	91.6	90.1	67.7	88.9
	ResNet-101	87.4	85.7	90.4	93.1	91.1	73.2	90.7
RFCN	ResNet-50	84.6	89.4	92.8	93.2	91.8	50.6	89.6
	ResNet-101	85.6	88.7	90.6	94.2	92.5	55.6	92.0

Table 2: Detection results of SW-CNN, Fast-RCNN (RCNN) [28], Faster RCNN (FRCNN) [20] and R-FCN [21] for multi-class problem (300 region proposals).

Figure 4a/b illustrates the impact on the number of region proposals on both detection performance and runtime. Figure 4b shows mean runtime per image where we can see R-FCN with ResNet-101 requires only 0.1s per image, while FRCNN and SW-CNN with ResNet-101 take 0.5s and 4.1s per image, respectively.

4. CONCLUSION

This work examines the relative performance of contemporary object detection strategies for region based object detection within the challenging domain of X-ray security imagery. We examine the relative performance of traditional a sliding window driven CNN model [15, 25] against contemporary region-based CNN variants [20, 21, 28]. We show that the use of CNN classification within an exhaustive search sliding window object detection localization framework (SW-CNN), already empirically shown to outperform hand crafted features [13, 18], are outperformed by Faster RCNN and R-FCN approaches in terms of both speed and accuracy. Faster RCNN (with VGG16) yields 88.3 mAP over a 6-class object detection problem whilst R-FCN (with ResNet-101) achieves 96.3 mAP for firearm detection requiring only 0.1 s per image. This clearly illustrates the real-time applicability and superiority of such integrated region based detection models within this X-ray security imagery context. Future work will consider the use of multi-view X-ray security imagery in an end to end design and a broader comparison between region based object detection [20, 21] and single network based detection methods [29–31].

Acknowledgment: This work is partially supported by HM Government (UK Home Office - Centre for Applied Science and Technology (CAST)).

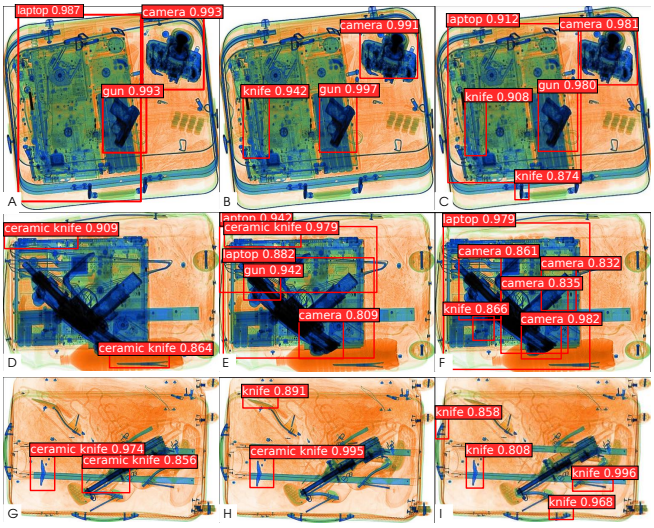

Fig. 3: Detection examples using ResNet-101. Columns: SW-CNN, Faster RCNN and R-FCN. Rows: Moderate to challenging samples.

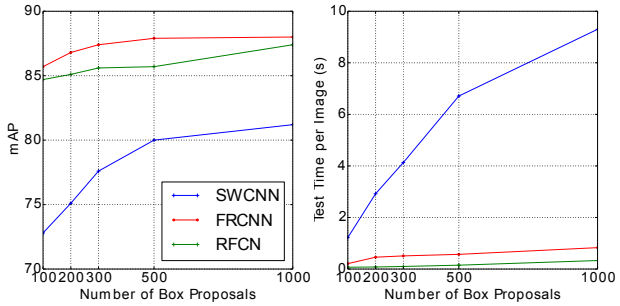

Fig. 4: Impact of number of box proposals on performance (a) and runtime (b). Models are trained using ResNet-101

5. REFERENCES

[1] S. Singh and M. Singh, "Explosives detection systems (eds) for aviation security," *Signal Processing*, vol. 83, no. 1, pp. 31–55, 2003. 1

[2] A. Mouton and T.P. Breckon, "A review of automated image understanding within 3d baggage computed tomography security screening," *Journal of X-Ray Science and Technology*, vol. 23, no. 5, pp. 531–555, September 2015. 1, 3

[3] T. W. Rogers, N. Jaccard, E. J. Morton, and L. D. Griffin, "Automated x-ray image analysis for cargo security: critical review and future promise," *Journal of X-ray science and technology*, , no. Preprint, pp. 1–24, 2016. 1, 2

[4] Z. Chen, Y. Zheng, B. R. Abidi, D. L. Page, and M. A. Abidi, "A combinational approach to the fusion, de-noising and enhancement of dual-energy x-ray luggage images," in *2005 IEEE CVPR'05 - Workshops*, June 2005, pp. 2–2. 1

[5] B. R. Abidi, Y. Zheng, A. V. Gribok, and M. A. Abidi, "Improving weapon detection in single energy x-ray images through pseudocoloring," *IEEE Tran on Sys, Man, and Cyber, Part C*, vol. 36, no. 6, pp. 784–796, Nov 2006. 1

[6] Q. Lu and R. W. Conners, "Using image processing methods to improve the explosive detection accuracy," *IEEE Tran on Sys, Man, and Cyber, Part C*, vol. 36, no. 6, pp. 750–760, Nov 2006. 1

[7] M. Singh and S. Singh, "Image segmentation optimisation for x-ray images of airline luggage," in *Comp. Intel. for Homeland Sec. and Pers. Safety, 2004. CIHSPS 2004.*, July 2004, pp. 10–17. 1

[8] G. Heitz and G. Chechik, "Object separation in x-ray image sets," in *CVPR, 2010 IEEE*, June 2010, pp. 2093–2100. 1

[9] M. Baştan, M. R. Yousefi, and T. M. Breuel, *Visual Words on Baggage X-Ray Images*, pp. 360–368, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. 1

[10] D. Turcsany, A. Mouton, and T.P. Breckon, "Improving feature-based object recognition for x-ray baggage security screening using primed visualwords," in *Industrial Tech., Int. Conf. on*. IEEE, 2013, pp. 1140–1145. 1

[11] D. Mery, V. Riffo, I. Zuccar, and C. Pieringer, "Automated x-ray object recognition using an efficient search algorithm in multiple views," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 368–374. 1

[12] D. Mery, E. Svec, and M. Arias, *Object Recognition in Baggage Inspection Using Adaptive Sparse Representations of X-ray Images*, pp. 709–720, Springer International Publishing, Cham, 2016. 1

[13] M.E. Kundegorski, S. Akcay, M. Devereux, A. Mouton, and T.P. Breckon, "On using feature descriptors as visual words for object detection within x-ray baggage security screening," in *Proc. Int Conf on Img for Crime Det and Prev*. November 2016, IET. 1, 3, 4

[14] D. Mery, "Automated detection in complex objects using a tracking algorithm in multiple x-ray views," in *CVPR 2011 Workshops*, June 2011, pp. 41–48. 1

[15] T. Franzel, U. Schmidt, and S. Roth, *Object Detection in Multi-view X-Ray Images*, pp. 144–154, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. 1, 2, 3, 4

[16] L. Schmidt-Hackenberg, M. R. Yousefi, and T. M. Breuel, "Visual cortex inspired features for object detection in x-ray images," in *ICPR, 2012 21st International Conference on*, Nov 2012, pp. 2573–2576. 1

[17] M. Baştan, "Multi-view object detection in dual-energy x-ray images," *Mach. Vis. and App.*, vol. 26, no. 7-8, pp. 1045–1060, 2015. 1, 2

[18] S. Akcay, 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 *Proc. ICIP*. September 2016, IEEE. 1, 2, 3, 4

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Eds., pp. 1097–1105. Curran Associates, Inc., 2012. 1, 2, 3

[20] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. 2, 3, 4

[21] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: object detection via region-based fully convolutional networks," *CoRR*, vol. abs/1605.06409, 2016. 2, 3, 4

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. 2, 3

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. 2

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. 2, 3

[25] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013. 2, 3, 4

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*. IEEE, 2014, pp. 580–587. 2, 3

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *CoRR*, vol. abs/1406.4729, 2014. 2

[28] R. B. Girshick, "Fast R-CNN," *CoRR*, vol. abs/1504.08083, 2015. 2, 3, 4

[29] W. Liu, D.Anguelov, D. Erhan, C. Szegedy, S.E. Reed, C.-Y. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. 2, 4

[30] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015. 2, 4

[31] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *ArXiv e-prints*, Dec. 2016. 2, 4

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015. 2, 3

[33] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft COCO: Common Objects in Context*, pp. 740–755, Springer International Publishing, Cham, 2014. 2, 3

[34] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*. IEEE, 2014, pp. 1717–1724. 2

[35] 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 Tran on Sys, Man, and Cyber: Systems*, 2016. 2

[36] J. RR Uijlings, K EA van de Sande, T. Gevers, and A WM Smeulders, "Selective search for object recognition," *IJCV*, vol. 104, no. 2, pp. 154–171, 2013. 2

[37] M. Everingham, L. Van Gool, C KI Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010. 3