Indonesian Journal of Electrical Engineering and Computer Science
Vol. 27, No. 2, August 2022, pp. 1007~1015
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v27.i2.pp1007-1015

ThreatNet: advanced threat detection, region-based convolutional neural network framework

Anurag Singh[1], Naresh Kumar[1], Seifedine Kadry[2]
[1]Department of Computer Science and Engineering, School of computing science and Engineering, Galgotias University, Greater Noida, India
[2]Faculty of Applied Computing and Technology, Noroff University College, Kristiansand, Norway

Article Info

Article history:

Received Apr 23, 2022
Revised Jun 3, 2022
Accepted Jun 16, 2022

Keywords:

Mask RCNN
Semantic segmentation
Surveillance
Threat pipeline
Transfer learning

ABSTRACT

It is critical for many countries to ensure public safety in detecting and identifying threats in a night, commercial places, border areas and public places. Majority of past research in this area has focused on the use of image-level categorization and object-level detection techniques. As an X-ray and thermal security image analysis strategy, object separation can considerably improve automatic threat detection when used in conjunction with other techniques. In order to detect possible threats, the effects of introducing segmentation deep learning models into the threat detection pipeline of a large imbalanced X-ray and thermal dataset were investigated. With the purpose of boosting the number of true positives discovered, a faster regional convolutional neural network (R-CNN) model was trained on a balanced dataset to identify probable hazard zones in X-ray and thermal security pictures. In order to get the final results, we combined the two models i.e faster R-CNN with Mask RCNN into a single detection pipeline using the transfer learning technique, which outperforms baseline and end-to-end instance segmentation methods using less number of the practical dataset, with mAPs ranging from 94.88 percent to 91.40 percent helps in detecting the person with guns, knives, pliers to avoid cross border threats.

Corresponding Author:

Anurag Singh
Department of Computer Science and Engineering, School of Computing Science and Engineering
Galgotias University
Plot No. 2, Yamuna Expy, Buddha International Circuit, Sector 17A, Greater Noida, Uttar Pradesh, India
Email: anurag.singh485@gmail.com

1. INTRODUCTION

Thermal imaging and X-ray imaging are commonly employed in public and border areas to keep them safe [1]. To complete the time-consuming and complex work of detecting risks in X-ray and thermal security images, it is necessary to create algorithms that aid human inspectors in locating threats in X-ray and thermal security images deployed at border areas. Deep learning has lately outperformed all prior systems for automatic threat recognition in X-ray and thermal security pictures, becoming the industry's most extensively used method with the highest accuracy. The most common way for detecting threats in X-ray and thermal security photographs is to train deep learning models on real-world data. X-ray and thermal security images, on the other hand, lack the fine detail found in nature photography and have a restricted color spectrum, as well as low contrast and roughness. The presence of object overlap in these pictures distinguishes them, posing a challenge to deep learning models since object overlap increases intra-class differences as discussed in [2]. Because each pixel represents a muted radiation intensity that may be viewed in an image, pixels in security images convey information about object overlap. In X-ray and thermal security pictures, attenuation increases

in areas with many overlapping items or a limited number of non-overlapping high-density objects. Lighter areas of the picture, on the other hand, show less attenuation. To deal with items that overlap, X-ray and thermal security images can be analyzed pixel by pixel, as seen in the example below. Pixel-level deep learning, for example, has been a dominating paradigm in medical X-ray and thermal imaging. As a result, pixel-level analysis can provide similar advantages in efficiency and reliability in X-ray and thermal security applications as well as other X-ray imaging industries, such as structural materials inspection [3].

In picture segmentation, semantic segmentation and instance segmentation are two distinct challenges [4]. Using semantic segmentation, all pixels in an image may be allocated to one of the provided object classes. The instance segmentation technique uses an item's position and pixel-by-pixel classification to distinguish each distinct object instance in an image. Case segmentation appears to be the optimum task domain for detecting risks in X-ray and thermal security images since it allows for pixel-level localization of each possible threat. This job cannot be applied directly to X-ray or thermal security photos since each pixel must be classed as a single instance of an object, which is impossible. To cope with the problem of overlap, object separation, a more narrowly defined task domain, is required. This is a more sophisticated version of instance segmentation that employs numerous classes and labels. According to [5], this task domain was initially developed as a method for detecting possibly overlapping things in X-ray and thermal security images and then connecting the appropriate pixel values with each object's estimated atomic number. Disallowed products with non-organic material traits, such as a weapon, can be recognized using the form, texture, and other visual aspects as part of the object separation operation's initial stage. Despite the fact that explosives and illicit chemicals have no apparent features, this component may identify them. This is performed by identifying the constituent elements of the forbidden items. Using the whole object separation task domain, a generic solution for disallowed things in X-ray and thermal security pictures may be found. Deep learning can help with the first half of the object separation problem since annotations are readily available at an early level. There is still a lot of work to be done to finish the second phase of object separation, however, this knowledge is currently unavailable to the general public.

As the public's knowledge of X-ray security photographs grew, a researcher established SIXray, the world's largest collection of X-ray and thermal security images that accurately portray the actual scene. Even if the overlapping of objects is revealed, the issue of imbalance in this dataset is also highlighted, which is an even more critical point. This is due to the fact that X-ray security screening detects threats far less frequently than it does normal objects, hence the data distribution of X-ray security images is substantially skewed towards the majority class or negative samples. Traditional models trained on imbalance datasets are heavily biased toward properly forecasting the majority class because of the overrepresentation of the majority class in the datasets. A substantial issue arises because the cost of misclassifying minorities in the majority or positive samples is significantly greater than the cost of misclassifying minorities in the majority or negative samples [6]. Oversampling the minority class or undersampling the majority class is the simplest method for balancing an unequal dataset. Even though the majority class is enormously more numerous, oversampling the minority group is ineffective and has been proved to be outperformed by other methods several times. Over- or under-sampling of minority groups might result in an increase in false positives since the traditional model tries to match items from unobserved negative samples to any identified threats that have been under-sampled. It is critical to minimize the number of false positives when utilizing a detection algorithm in security systems to assist human inspectors [7]. An imbalanced, huge X-ray and thermal security image dataset, commonly known as a realistic dataset, was used to investigate how well the threat detection pipeline performed when the pixel-level analysis was added to solve the first half of the object separation challenge. In order to get the most accurate results, we used a Faster-RCNN [8] trained on a well-balanced subset of the dataset. A DeepLabV3+ was trained to categorize each pixel in the probable hazard zones to reduce false positives. The two models were combined into a single detection method for the final predictions. To choose these models, we conducted a thorough review of the top object recognition and semantic segmentation models currently on the market.

The following are the most significant contributions made by this paper: i) object separation has been restored as a separate task domain for X-ray security pictures, ii) a realistic X-ray security dataset has been segmented to solve the problem of class imbalance, iii) deep learning models have been evaluated on our own dataset, iv) to make the most of the annotation already available, a two-stage threat detection approach has been designed that separates detection and segmentation, v) Use of data augmentation technique to overcome the dataset issues, and vi) with the use of transfer learning, we can merge faster R-CNN and Mask R-CNN to minimize the complexity of training a model from scratch.

According to the studies, the suggested technique outperformed both earlier baseline methods and an end-to-end instance segmentation method by a wide margin, achieving mean average precision (AP) of 94.88 percent, 91.40 percent, and 89.52 percent across increasing imbalance ratios. The following parts make up the remainder of the paper: A summary of the relevant research is presented in section 2. The dataset that is skewed to one side is discussed in section 3. There are measures that will be used to evaluate the project in section 4.

The approach's methodology is examined in detail in section 5 of the proposed plan. According to the conclusions of the study, the experiments are detailed in section 6. This investigation's results are presented in section 7.

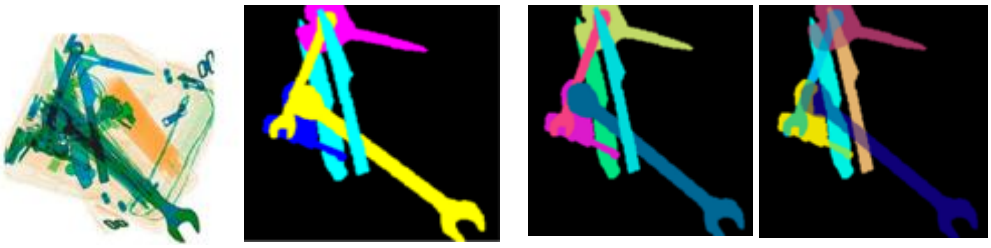
2. RELATED WORK

These findings are related to previous studies that detected threats using pixel-level analysis in X-ray and security threat photos, as well as works that addressed class imbalance in massive security threat images. The use of X-ray and threat images increased log space. Most genuine port security operations, as shown in the images below, do not necessitate multiple photographs of the same target item. Because their method resulted in incorrect segmentation, they shifted to predicting material qualities using atomic numbers. As a result of this limitation, their findings cannot be applied to other threats such as weapons. There has been no attempt since then to partially or completely solve the problem. Pixel-level X-ray and threat security imaging techniques were also evaluated. According to the researchers, machine learning algorithms can distinguish between organic and inorganic X-ray and threat security photos based on colour. To identify anomalies in an item, a two-stage segmentation technique was used. Identifying X-ray threats is similar to segmentation. Because semantic segmentation does not distinguish between duplicates, it cannot separate objects.

Using a convolutional neural network (CNN), Miao *et al.* presented class-imbalance hierarchical refinement (CHR), which uses poorly connected objects from the feature map to improve threat detection in unbalanced X-ray security images. Study show discussed in [9] looked at the impact of employing a generative adversarial network (GAN) to learn how negative samples are distributed under the surface, in order to identify positive cases that differ from the learned distribution as anomalies. As a result of identifying anomalies in the dataset, we may train our classifier on an ideal dataset while simultaneously minimizing the number of false positives it produces. A variety of image-generating GANs was used to execute picture synthesis, and the results of this study [10] looked at the effect on the number of positive samples. When threat objects are separated from backgrounds that differ from positive samples when creating synthetic images by combining isolated threat objects and negative samples, the model is better able to generalise and suppress false positives for object-level threat detection, according to our findings. Early and late feature fusion is performed by concatenating features gathered earlier in the classification model as discussed in [11] which was achieved by taking the weighted total of losses determined at different stages of the classification model. Due to the unequal data, these components are combined into a dual branch network in order to take use of low-level spatial properties and eliminate bias. A combination of approaches, such as the tensor pooling method presented in [12], can be used to detect pixels of risk items utilizing preprocessed inputs. In this method, the image's contours are recovered and represented as a tensor representation at multiple orientations. When it came to threat detection, however, the authors were anxious about danger item separation or isolation since they neglected the problem of picture overlap in X-ray security images. They also have to name each variety of the threat category, which increases the cost of annotation by tenfold.

3. DATASET

More than 8,000 pictures in this collection are expected to contain at least one of the following tools: a handgun, a knife, or one of the following: a wrench, pliers, or scissors. SIXray has about 1 million images with at least one of the following objects labelled: a weapon, a knife, a wrench, pliers, and scissors. Furthermore, the dataset is divided into three subgroups, SIXray10, SIXray100, and SIXray1000, based on rising imbalance ratios between positive and negative values. Regardless, the collection only contains annotations at the picture and object levels, with no information on the source images. To complete the object separation method, pixel-level annotations are necessary. The ground truth masks in Figure 1 address semantic segmentation, instance segmentation, and object separation in Figure 1(a) annotations at the pixel level for various activities, Figure 1(b) images with overlapping items as input, Figure 1(c) ground truth labels for instance segmentation, and Figure 1(d) ground truth labels for object separation. An object's semantic group and object instance can only be defined by occlusion, however, a pixel can be classed under many instances of the same object. Both semantic and instance segmentation can benefit from occlusion. As a result, the object separation task's distinctive characteristic is the employment of ground truth labels in the supervised training process. Each object instance, as shown in Figure 2, has a unique binary mask for instance segmentation and object separation. Consider the instance segmentation task: objects near the bottom of the stack tend to lose more, if not all, of their pixel-wise labels to those above them, even if the complete item is visible in the image. Using these masks to train a model will clearly result in a model that ignores item overlaps and lacks information about the items at the bottom of a stacked hierarchy.



(a) (b) (c) (d)

Figure 1. Annotations at the pixel level for various activities (a) image with overlapping items as input, (b) segmentation by semantics, (c) ground truth labels for instance segmentation, and (d) ground truth labels for object separation

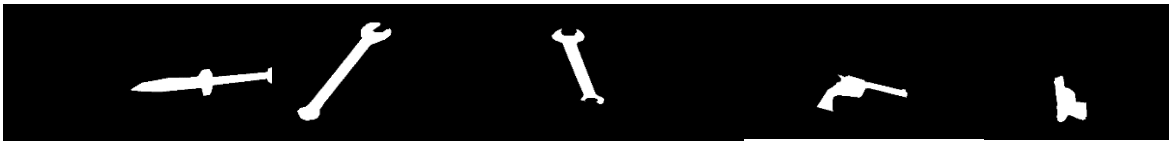


Figure 2. Each object instance's isolated binary ground truth labels for instance segmentation and object separation

As a result, we manually categorised the photographs so that the ground truth mask includes all of the pixels that define each occurrence. This is why we choose to start with a random sampling of 2,500 images from the positive samples used in training subgroups and share our findings in order to inspire future research into this task domain to generate further labelled datasets that include humans with these threat-related elements. In the pixel-level labeled dataset, category 1 has more instances than any other category discussed in Figure 3. Researchers in [13] examine the dataset in great depth. When the data was analyzed, it was discovered that certain samples had been wrongly labeled as negative despite the fact that they clearly included hazards. The subject of machine learning frequently encounters labels that are noisy or broken. According to the findings, the percentage of datasets with incorrect labels might range from 8 to 38.5 percent [14]. An entirely new field of training approaches for models that can withstand noisy labels has grown out of this study's original focus.

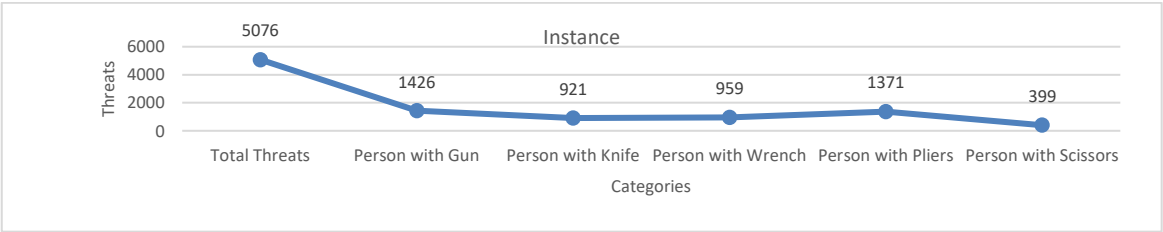


Figure 3. Distribution of the subset's occurrences

4. EVALUATION MATRICES

To select the best detection model, we evaluated performance using average precision [15], as (1).

$$Ap = \sum_{n=0}(r_{n+1} - r_n)pinterp_{(r_n+1)} \quad (1)$$

Where r is the recall and $pinterp$ is the interpolated precision given by $pinterp_{(r_n+1)}$=max $(r_n + 1)$ p(er), wherein p is the precision at er and n includes all of the recall points, and n includes all of the recall points. The precision-recall curve's area under the curve (AP) is another way of defining Ap. The mean of the averages of the APs determined for each type of threat items is the mean of the mAPs. As one of the most widely used metrics for assessing the efficacy of identification and classification systems, this is a good place to start.

Other evaluation criteria, such as intersection-over-union (IoU), dice coefficient (DC), precision, and recall, are used to select the segmentation model, as (2)-(5).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

$$DC = \frac{2|A \cap B|}{|A| + |B|} \quad (3)$$

Where A and B are the segmentation masks for the target and anticipated segments. For each threat type item, intersection over union (IoU) is determined. The mean IoU for each threat item class is recorded, and the average is shown as the mean IoU, (mIoU). Another prominent metric for gauging the performance of modern segmentation algorithms is DC, which measures the amount of overlap between two segments and is similar to the IoU metric in that it measures the amount of overlap between two segments.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

True positives (TP), false positives (FP), and false negatives (FN) are the proportions of pixels in the masks that are assessed to be true positive, false positive, and false negative, respectively. Precision and recall may be calculated independently for each class as well as as a total for all classes taken together. In accordance with the benchmark standard established in [16] we once again used the mAP, described in (1), to compare our suggested method to previously published methods. We calculated the AP for each of the classes by ranking the classification predictions according to their confidence scores, and then calculated the average (mAP) across all of the classes in order to report the overall performance of the models.

5. PROPOSED METHOD

5.1. Model selection

In this section, we preciously discuss research that has used deep learning algorithms to tackle critical computer vision tasks like object identification and proposed methodology that adopted and provides an in-depth discussion of the model selection process as well as the whole threat detection pipeline with an accuracy of different models on the same dataset. Annotate image, object, and pixel data when building an end-to-end object separation model. X-ray security imaging lacks this data. Missing annotation data is discarded when training end-to-end models. Instead, we created a threat detection pipeline that separates object instance localization and segmentation mask prediction into separate models. Every sample in the dataset that only has picture and object-level annotations is used to train the object identification model, and every sample in the dataset is used to train the segmentation model. It is possible to designate the pixels associated with each individual object instance using this separation model while also utilising the annotations that are made available when all of these pipelines are put together in one place at the same time. By using the transfer learning technique and data augmentation technique in conjunction with constrained annotated X-Ray and Thermal image data, the pre-trained model faster R-CNN was merged with mask R-CNN (semantic segmentation). The following two subsections of this section go into further detail on how experiments are used to find appropriate models for these two purposes.

5.2. Detection models

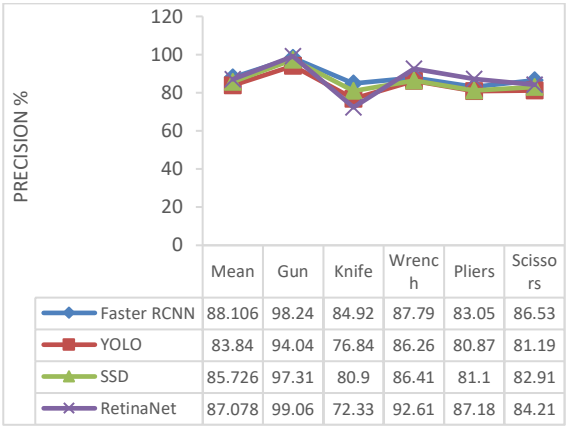
To find the optimum detection model, we examined four of the most extensively utilised cutting-edge technologies. Zhao *et al.* [17] discussed a two-stage object detection model that is part of the region-based CNNs object detection model family. After following RPN operation, it removes areas of an image suspected of holding target items and sends them to the second stage of an algorithm that classifies those things; this algorithm is known as a region-proposal network (RPN). The method then uses the extracted region to perform a classification operation on the extracted objects. The prediction of the bounding boxes can be improved by treating it as a regression problem and accounting for the difference between the actual bounding boxes and the anticipated regions. In comparison to previous models, this model is intended to be the most efficient and precise.

You only look once (YOLOv3) [18], a one-stage object detector, as part of the study, the algorithm does not require a geographical proposal and instead analyses a dense sample of the likely locations. It is divided into grid cells (known as priors), and each cell uses the YOLO method to predict a predefined number of bounding boxes as well as the image's confidence ratings. A single CNN predicts all of these outcomes

simultaneously, making it one of the most efficient real-time algorithms. In this experiment, the third version of the system was used, which was meant to recognise small objects more precisely by utilising shortcut connections. The new version is intended to identify minor details more accurately than the previous two.

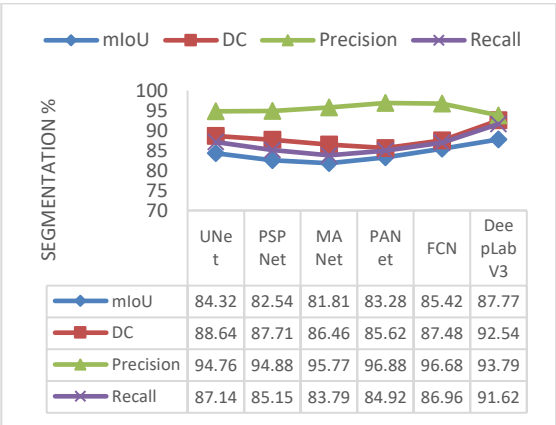
The single-shot multibox detector (SSD) [19] detects various sizes and scales at each pyramidal layer of the CNN using a single-stage approach. Instead of separating the input into rows and columns, anchor boxes in a feature map are utilised to anticipate the offset of the default boxes. Each level of the CNN pyramid has its own set of receptive fields for feature maps. The early layers' feature maps are finer-grained, whereas the latter layers' feature maps are coarser-grained. Because anchor boxes have fixed sizes in relation to their respective cells, predictions at higher layers capture larger objects in the image, while predictions at lower levels capture smaller objects. This is due to the anchor boxes having a fixed size in proportion to the cells to which they are connected.

Using the notion of detection at each level of the CNN's pyramidal layers, RetinaNet is a one-stage approach that is analogous to the concept of detection in SSD. They do add a feature pyramid network (FPN) to build more robust representations, which clubs succeeding feature maps with prior feature maps. "Hard" data, or samples that are regularly misclassified by researchers, are further highlighted by the use of a new loss function called as focused loss. For the evaluation of each detection model's performance in identifying dangers in the original picture, we solely utilized positive samples. Training and a validation dataset were created from the data set. The same backbone network, ResNet-50 [20], was used to train all the models, and a total of 60,000 iterations were conducted on it. On the validation set, Figure 4 displays the detectors' per-category and mean APs. The overall analysis suggests that Faster R-CNN is the most accurate detector, despite the fact that the findings varied by category discussed in Figure 5. RetinaNet had an FPN connected to its backbone network at the time of testing, which none of the other models had. As a result, in terms of total performance, it exceeded earlier one-stage detectors, approaching the faster R-CNN technique. As a result of this discovery, throughout the remainder of the trial, we used faster R-CNN as our object detection model.



	Mean	Gun	Knife	Wrench	Pliers	Scissors
Faster RCNN	88.106	98.24	84.92	87.79	83.05	86.53
YOLO	83.84	94.04	76.84	86.26	80.87	81.19
SSD	85.726	97.31	80.9	86.41	81.1	82.91
RetinaNet	87.078	99.06	72.33	92.61	87.18	84.21

Figure 4. Detection mean average precision (%)



	UNet	PSP Net	MA Net	PAN et	FCN	DeepLab V3
mIoU	84.32	82.54	81.81	83.28	85.42	87.77
DC	88.64	87.71	86.46	85.62	87.48	92.54
Precision	94.76	94.88	95.77	96.88	96.68	93.79
Recall	87.14	85.15	83.79	84.92	86.96	91.62

Figure 5. Segmentation performance

6. EXPERIMENT RESULTS

We used predetermined chunks of the SIXray dataset and ms coco dataset to train our models and then used those same areas to test our predictions. At a ratio of 1 to 10, 1 to 100, and 1 to 1000 in SIXray, negative samples vastly outnumber positive ones. While the detection model was trained using all of the photographs in the training sets, the segmentation model was developed using only the annotated images. Based on what was mentioned in Section 2, we constructed the patches that were utilized to train the segmentation model. A total of 192×192 pixels were added to the size of all patches.

Stochastic gradient descent (SGD) [21], [22] was used to train faster-RCNN, with a batch size of 2 and an SGD rate of base learning rate of 0.001 that progressively decays by an order of magnitude after iteration 30,001 through 50,000. Between iterations 30K and 50K, we used a learning rate of 0.01 that linearly decreased by 0.01 between the 30K and 50K points. There were 250 epochs of DeepLabV3+ training with a 32-batch batch size and a 0.001 learning rate, which linearly declined by 0.1 after 75th, 150th, and 200th iterations. It was used to train DeepLabV3+ for 250 iterations with a batch size of 32 using the Adam optimizer. Mask R-CNN, an advanced end-to-end instance segmentation framework, was also trained using the entire training set with comprehensive annotations. Both faster and the mask R-CNN were trained using the identical backbone

and settings and produced the same results [23]. It is shown in figure 6 that the trial had a qualitative effect. Figures 7 (a)-(e) shows how our method accurately segmented the detected dangerous objects, despite the fact that the knife, a wrench, and a firearm all overlapped. This is seen in Figure 7(c), where our approach was unable to properly separate overlapping objects despite the fact that the majority of them possessed high-density material qualities, as indicated in the preceding section. For the vast majority of non-overlapping items, however, our technique was able to correctly verify the detections. Second-stage techniques significantly decreased the number of false positives we saw during our testing of the scissors class. The scissors class may have had an abnormally high number of false-positive predictions from other methodologies, which may have contributed to its poor performance. After that, the failure scenarios that we encountered while creating our technique are depicted in Figure 7. Faster R-CNN incorrectly forecasted suspicious regions, therefore DeepLabV3+ built segmentation masks for these images.

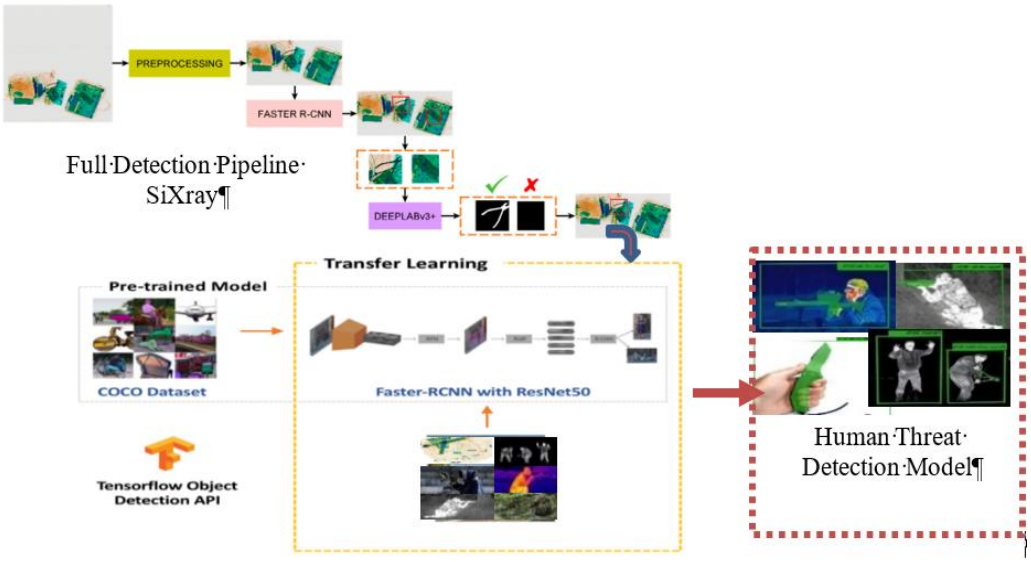


Figure 6. Threat detection pipeline using transfer learning approach

(a) (b) (c) (d) (e)

Figure 7. Exemplar images of verified detections, (a) person with gun, (b) person with knife, (c) person with gun, (d) thermal imaging person with gun, and (e) person with plier

During our investigation, we found that the vast majority of the mistakes were the result of incorrect localization and verification of the suspicious knives and wrenches. Products in these categories may have been mistaken for elongated metals in the suitcase owing to their blandness, according to certain theories [24], [25]. Weaponry, on the other hand, was able to be continuously anticipated with high precision, regardless of how lopsided the overall scenario was.

7. CONCLUSION

We examined each module in our threat detection pipeline to see how they affected the algorithm's overall performance. The performance examined of four distinct instances discussed in above figures, each with its own threat detection pipeline configuration. To begin, we only looked at the detection model without considering the pre-processing and testing of the segmentation model. The detection method was then

combined with a pre-processing technique. Following that, we merged detection and segmentation (Det + Segm), and finally, we combined all modules (Crop, Det, and Segm) to establish a single threat detection pipeline. As previously proven, we may make a small but significant improvement by simply removing unwanted picture regions, such as air gaps/spaces that are typical in X-ray security photos. Cropping photos to expose only key details and extracting additional features improves the detection model's capacity to recognise more items with better reliability. Because the detection model was trained on a balanced training set, it must match a substantial amount of previously discovered data from negative samples to more recognised targets in positive samples. This enhances both true positive and false positive detection. This is identified as the primary bottleneck, which is solved by including the segmentation model to validate the preliminary predictions.

REFERENCES

[1] D. Mery, D. Saavedra and M. Prasad, "X-Ray baggage inspection with computer vision: a survey," *IEEE Access*, vol. 8, pp. 145620-145633, 2020, doi: 10.1109/ACCESS.2020.3015014.

[2] C. Miao *et al*., "SIXray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 15, no. 20, pp. 2114–2123, 2019, doi:10.48550/arXiv.1901.00303.

[3] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, "Image segmentation using deep learning: a survey," I*EEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3523-3542, 2022, doi: 10.1109/TPAMI.2021.3059968.

[4] G. Heitz and G. Chechik, "Object separation in x-ray image sets," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2093-2100, doi: 10.1109/CVPR.2010.5539887.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009, doi: 10.1109/TKDE.2008.239.

[6] M. Chouai, M. Merah, J. L. Sancho-Gómez and M. Mimi. "A machine learning color-based segmentation for object detection within dual X-ray baggage images," *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, 2020, pp. 1-11 doi:10.1145/3386723.3387869.

[7] I. Goodfellow *et al.* "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020, doi:10.1145/3422622

[8] H. Zhang, I. Goodfellow, D. Metaxas and A. Odena, "Self-attention generative adversarial networks," *Proceedings of the 36th International Conference on Machine Learning in Proceedings of Machine Learning Research*, 2019, vol. 97, pp- 7354-7363.

[9] Y. Xu and J. Wei, "Deep feature fusion based dual branch network for x-ray security inspection image classification," *Applied Sciences*, vol. 11, no. 16, pp. 7485-87, 2021, doi: 10.3390/app11167485

[10] T. Hassan, S. Akçay, M. Bennamoun, S. Khan and N. Werghi, "Tensor pooling driven instance segmentation framework for baggage threat recognition," *Neural Computing & Applications*, vol. 34, pp. 1239-1250, 2021, doi: 10.1007/s00521-021-06411-x.

[11] J. K. Dumagpi and Y. J. Jeong, "Evaluating gan-based image augmentation for threat detection in large-scale xray security images," *Applied Sciences*, vol. 11, no. 1, pp. 1-20. 2021, doi: 10.3390/app11010036.

[12] H. Song, M. Kim, D. Park, Y. Shin and J. -G. Lee, "Learning from noisy labels with deep neural networks: a survey," *IEEE Transacations on Neural Networks & Learning Systems*, pp. 1-20, 2022, doi: 10.1109/TNNLS.2022.3152527.

[13] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," *Computer Vision Pattern Recognition—ECCV 2016*, 2016, pp. 1-6.

[14] W. Liu *et al.*, "SSD: single shot multibox detector," *Computer Vision – ECCV 2016*, 2016, pp. 21-37, doi: 10.1007/978-3-319-46448-02.

[15] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions, Pattern Analysis and Machine. Intelligence*, vol. 42, no. 2, pp. 318–327, 2020, doi: 10.1109/TPAMI.2018.2858826.

[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.

[17] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6230–6239.

[18] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *Computer Vision Pattern Recognition*, pp. 1-14, 2017.

[19] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *Computer Vision Pattern Recognition*, pp. 1-13, 2018.

[20] T. Fan, G. Wang, Y. Li, and H. Wang, "MA-Net: a multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020, doi: 10.1109/ACCESS.2020.3025372.

[21] J.K. Dumagpi, W. Jung, and Y. Jeong, "KNN-based automatic cropping for improved threat object recognition in x-ray security images," *Journal of IKEEE*, vol. 23, no. 4, pp. 1134–1139,2019, doi: 10.7471/IKEEE.2019.23.4.1134

[22] C. Shorten, T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data,* vol. 6, no. 1, p. 60,2019, doi:10.1186/s40537-019-0197-0

[23] M. Everingham, L. Van Gool, C. K. I. I. Williams, J. Winn and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, doi: 10.1007/s11263-009-0275-4.

[24] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587, doi: 10.1109/CVPR.2014.81.

[25] R. Girshick, "Fast R-CNN," *In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV*, 2015), 2015, pp. 1440–1448, doi: 10.1109/ICCV.2015.169

BIOGRAPHIES OF AUTHORS

Anurag Singh is working as an Assistant Professor and Research Scholar in School of computing science & Engineering, Galgotias University, Uttar Pradesh, India. His area of research is in computer vision and deep learning. He has published 5 Scopus publications which includes conferences and Journals. He has done his B.Tech, M.Tech and pursuing Ph.D. in Computer science and Engineering. He has total experience of 9 years which includes industry and teaching. He can be contacted at email: anurag.singh485@gmail.com.

Naresh Dhull is working as a Professor and Dean PG & PhD in Galgotias University, Uttar Pradesh, India. His area of research is in Cloud computing, Networking, Computer vision, Deep Learning. He has published 17 Scopus and SCI publications with 54 citations. He has almost 20 years of experience in area of computer science. He has done his Bachelor of Computer Applications, Masters in Computer Applications and Ph.D. in Computer Science and Engineering. He can be contacted at email: naresh.dhull@gmail.com.

Seifedine Kadry is working as a Professor of Data Science at Noroff University College, Norway. Dr. Seifedine Kadry has a Bachelor degree in 1999 from Lebanese University, MS degree in 2002 from Reims University (France) and EPFL (Lausanne), Ph.D. in 2007 from Blaise Pascal University (France), HDR degree in 2017 from Rouen University (France). His research focuses on Data Science, education using technology, system prognostics, stochastic systems, and applied mathematics. He has published more than 20 Scopus and SCI publications with 7490 citations. He can be contacted at email: skadry@gmail.com.