ON THE IMPACT OF USING X-RAY ENERGY RESPONSE IMAGERY FOR OBJECT DETECTION VIA CONVOLUTIONAL NEURAL NETWORKS

Neelanjan Bhowmik
d Yona Falinie A. Gaus
t Toby P. Breckon1,2
Department of {Computer Science1 | Engineering2}, Durham University, UK

ABSTRACT

Automatic detection of prohibited items within complex and cluttered X-ray security imagery is essential to maintaining transport security, where prior work on automatic prohibited item detection focus primarily on pseudo-colour (rgb) X-ray imagery. In this work we study the impact of variant X-ray imagery, i.e., X-ray energy response (high, low) and effective-z compared to rgb, via the use of deep Convolutional Neural Networks (CNN) for the joint object detection and segmentation task posed within X-ray baggage security screening. We evaluate state-of-the-art CNN architectures (Mask R-CNN, YOLACT, CARAFE and Cascade Mask R-CNN) to explore the transferability of models trained with such ‘raw’ variant imagery between the varying X-ray security scanners that exhibits differing imaging geometries, image resolutions and material colour profiles. Overall, we observe maximal detection performance using CARAFE, attributable to training using combination of rgb, high, low, and effective-z X-ray imagery, obtaining 0.7 mean Average Precision (mAP) for a six class object detection problem. Our results also exhibit a remarkable degree of generalisation capability in terms of cross-scanner transferability (AP: 0.835/0.611) for a one class object detection problem by combining rgb, high, low, and effective-z imagery.

Index Terms— x-ray imagery, deep convolutional neural network, object detection, transferability

1. INTRODUCTION

X-ray security screening plays a pivotal role in aviation security. However, manual inspection of potentially prohibited items is challenging due to the clutter and occlusion present within X-ray scanned baggage. A modern X-ray security scanner makes use of multiple X-ray energy levels in order to facilitate effective materials discrimination [1]. Subsequently, a dual-energy X-ray scanner imagery consists of two intensity images acquired at two discrete energy levels (low and high), facilitating the recovery of material properties (effective atomic number, effective-z). The information is fused with the help of a colour transfer function into a single pseudo-colour X-ray image (Figure 1A) to facilitate the interpretation of the baggage contents [2].

The advancement of deep Convolutional Neural Networks (CNN) has brought new insight to the automation of this X-ray imagery screening task [3–5] where the primary task is both to localise and classify the prohibited items. Prior works [4,6] are concentrated on the shaped-based detection of prohibited items achieving both high detection performance with low false positive. The work of [3] uses a pre-trained GoogleNet model for classification task in X-ray baggage scans for detecting potentially prohibited items. Subsequently, the work of [4] compares contemporary region-based and single forward-pass based CNN architectures (Faster R-CNN [7], YoloV2 [8]) achieving 0.88 and 0.97 mean Average Precision (mAP) over six class and two class X-ray baggage security object detection problems respectively. Following these works, [5] proposes a dual-stage CNN architecture for anomaly detection in a six class problem. Semi-supervised adversarial learning is used in the works of [9, 10] for prohibited item detection. The availability of the large-scale X-ray baggage datasets (SIXray [11], and OPIXray [12]) has provided further insight into the transferability and generalisation abilities of the CNN architectures [6] across varying X-ray security scanners which all exhibit varying characteristics in terms of projection geometry common resolution and pseudo-colour mapping. While most prior work [4–6] process each view of multi-view X-ray scanners independently, [13] utilise the corresponding information between the views for a detection task achieving 0.91 mAP.

Almost all of the prior work discussed here only use pseudo-colour/false colour (rgb) X-ray imagery that is itself generated from the ‘raw’ high/low/effective-z imagery ob-
Object Detection

Source X-Ray

I as the target object proceeds through the scanner tunnel, defined 150kV [19] from a X-ray source. Generally, the X-ray images are composed of an X-ray source emitter and detector (Figure 2, left). X-rays are emitted with photon energy ranging up to 2, left). X-ray images (Figure 2, right) in Section 2.2.

2. PROPOSED APPROACH

In this study, we present dual-energy X-ray imaging technique (Figure 2, left) in Section 2.1 and followed by object detection and segmentation strategies (Figure 2, right) in Section 2.2.

2.1. Dual-Energy Projection X-ray Imaging

The primary components of X-ray security scanner system are composed of an X-ray source emitter and detector (Figure 2, left). X-rays are emitted with photon energy ranging up to 150kV [19] from a X-ray source. Generally, the X-ray images are constructed by attenuating the signal on the material as the target object proceeds through the scanner tunnel, defined as \( I(E) = I_0e^{-\mu t} \), where \( I(E) \) is the captured intensity as a function of the thickness \( t \), the emitted intensity \( I_0 \) and the absorption coefficient \( \mu \). The absorption coefficient is defined by \( \mu = \alpha(Z, E)\rho \), where \( Z \) is the atomic number, \( E \) is the energy, \( \rho \) is the density, and \( \alpha(Z, E) \) corresponds to the mass attenuation coefficient in terms of \( Z \) and \( E \) [20].

In the dual-energy source X-ray imaging, two intensity responses captured at two different energy levels, low and high \( (E = \{l, h\}) \) and are subsequently combined to construct low and high energy response images (Figure 2, left). Given the Compton scatter coefficient \( (\mu_c) \) and the photoelectric absorption coefficient \( (\mu_p) \) [21], material identification (approximate atomic number, effective-\( z \); \( Z_{eff} \)) can be calculated as:

\[
Z_{eff} = K' \left( \frac{\mu_c}{\mu_p} \right)^{\frac{1}{n}}
\]

where \( K' \) and \( n \) are constant [21]. In this work, we evaluate the use of the pseudo-colour (\( rgb \)), dual-energy response (\( h, l \)) and effective-\( z \) (Figure 1) as alternative inputs imagery for CNN-based object detection.

2.2. Object Detection and Segmentation Strategy

We consider four contemporary CNN architectures of differing characteristics, spanning both single stage and multi stage detection approaches, and explore their applicability for prohibited item detection within varying configurations of dual-energy X-ray imagery inputs.

Mask R-CNN [15] is a two-stage detector for object instance segmentation, developed on top of Faster R-CNN [7]. Mask R-CNN [15] uses the Faster R-CNN [7] architecture for feature extraction, Region Proposal Network (RPN), and followed by region of interest alignment (RoIAlign) via bilinear boundary interpolation to produce higher resolution feature map boundaries suitable for input into a secondary classifier. The output from the RoIAlign layer is subsequently fed into a series of segmentation processing layers (mask head), that generate an additional image mask indicating pixel membership of a given detected object.

YOLACT [16] is an one-stage detector, based on RetinaNet [22], that directly predicts boxes without a separate region proposal step. YOLACT [16] generates a set of prototype masks, linear combination coefficients for each predicted instance, and associated bounding boxes. It combines the prototype masks using the corresponding predicted mask coefficients followed by cropping with a predicted bounding box to generate the final output.

CARAFE [17] is a two-stage architecture, which proposes effective feature up-sampling operators and integrates it into Feature Pyramid Network to boost the performance. For instance segmentation, a feature map, which represents the object shape accurately, is used to predict the final instance segmentation result.

Cascade Mask R-CNN [18], a multi-stage detector, is a hybrid of Cascade R-CNN and Mask R-CNN [15]. Similar to Mask R-CNN [15], each stage has a segmentation mask.
branch, a label prediction branch, and a bounding box detector branch. The current stage will accept RPN or the bounding box returned by the previous stage as an input. The second stage increases localisation performance accuracy, and subsequently, it further refines the output. This is repeated over multiple stages with increasingly refined criteria for discarding low-quality proposals from the previous stage such that it predicts precise bounding boxes and masks at the final stage.

In this study, we compare these four CNN architectures for object detection (Figure 2, right) using combination of different variants dual-energy X-ray imagery (Section 2.1). To assess the impact of dual-energy X-ray imagery variants on object detection we first use rgb, high (h), low (l), and effective-z (z) imagery individually. Second, h, l and z are combined as three channels (hlz) images. Thirdly, we combine rgb, high, low, and effective-z imagery for joint object detection and segmentation task.

Within the X-ray imagery security domain, imagery may be sourced using varying scanners [19, 23, 24], which have different X-ray energy spectra, spatial resolution and material colour profiles. In prior work [6, 25] on transferability and generalisation ability, [25] focuses on transfer learning between cargo parcel scanning (different scanner equipment due to the differences in scale). The work of [6] shows cross-scanner transferability of CNN architectures (using rgb X-ray imagery) in terms of their generalisation across varying X-ray scanner characteristics. In this study, we further evaluate the effectiveness of using variants of dual-energy X-ray imagery (Section 2.1) on generalisation capabilities of the CNN architectures.

3. EVALUATION

We focus on three datasets that are sourced from different X-ray scanners [19, 23, 24]. The deei6 is created from a Gilardoni X-ray scanner [19], and consists of rgb, high, low, and effective-z imagery. The other two datasets, dbs_laptop and dbr_laptop, are generated by a Smith Detection [23] and Rapiscan X-ray scanner [24] respectively and consist of rgb X-ray imagery. The four CNN architectures (Section 2.2) are trained using rgb and combinations of rgb, high, low, and effective-z X-ray imagery from deei6 dataset. Subsequently, we evaluate the model performance on rgb X-ray imagery of dbs_laptop and dbr_laptop datasets.

deei6: Our dataset (Durham Electrical and Electronics Items) is constructed using a dual-energy Gilardoni FEP ME 640 AMX scanner [19] with associated pseudo-colour materials mapping. This dataset is composed of six-classes of consumer electronics, electrical and other items: {bottle, hairdryer, iron, toaster, phone-tablet, laptop}, totalling 7,022 images (70:30 data split for experiments). We also access the high, low, and effective-z imagery to construct deei6rgb, deei6h, deei6l and deei6z imagery as depicted in Figure 1. To investigate the generalisation capabilities of the CNN architectures, we also use the following two datasets:
dbs_laptop: comprises 488 laptop class rgb X-ray image examples (with associated pseudo-colour materials mapping), which is sourced from a Smith Detection X-ray scanner [23]. dbr_laptop: comprises 107 laptop class X-ray rgb image examples (with associated pseudo-colour materials mapping). This dataset is sourced from Rapiscan 620DV X-ray scanner [24].

The CNN architectures (Section 2.2) are implemented using MM Detection framework [26]. Through the transfer learning paradigm, training (using X-ray imagery variants) of all CNN architectures (Section 2.2) are initialised with ImageNet [27] pretrained weights (which originate from training on colour RGB imagery). Our CNN architectures are trained using ResNet50 [28] backbone with following training configuration: backpropagation optimisation performed via Stochastic Gradient Descent, initial learning rate of $2.5 \times 10^{-4}$ with decay by a factor of 10 at $7$th epoch, and a batch size of $4$. The model performance is evaluated by MS-COCO metrics [29] (IoU of $0.50 : 0.55 : 0.95$), using Average Precision (AP) for class-wise and mAP for overall performance.

3.1. Impact of Dual-energy X-ray Imagery

In the first set of experiments (Table 1), exemplar items in X-ray security imagery are detected using the CNN architectures set out in Section 2.2. We use variants of dual-energy X-ray imagery of the deei6 dataset for training and evaluation denoted as deei6x for $x = \{rgb, h, l, z, h lz\}$. The highlighted mAP signifies the maximal results obtained for overall performance. At first, the CNN architectures are trained and evaluated on rgb X-ray imagery (Table 1, rgb), in line with [4, 5, 11]. The best performance is achieved by Cascade Mask R-CNN (CM R-CNN) [18] producing maximal mAP ($0.693$) and outperforming other three CNN architectures. When we train CNN architectures using high, low, and effective-z imagery individually and together as three channels (hlz), the overall performance (Table 1) does not improve compared to rgb imagery. The lowest performing training set is deei6z imagery achieving only $0.627$ of mAP (with Cascade Mask R-CNN [18]). It is possibly due to the lack of contrast in the pixel intensity in effective-z imagery where the target objects appear similar to the background, leading to inferior detection performance. The impact of dual-energy X-ray imagery can be observed while combining rgb, high, low and effective-z (deei6rgb,hlz) together. The maximal mAP of 0.7 (Table 1, rgb,hlz) is achieved by CARAFE [17] marginally outperforming rgb imagery (mAP: $0.693$). Although YOLACT [16] is the simplest architecture (34.76 million parameters), it outperforms Mask R-CNN [15] while training using rgb,hlz X-ray imagery (mAP: $0.686$ vs $0.680$).

In the confusion matrices (Figure 3) of CARAFE [17], we observe strong true positive (diagonal) and low false positive (off-diagonal) occurrence. The advantage of combining rgb, high, low and effective-z can be seen in the class phone-tablet (0.698 to 0.729, Figure 3(A)$\rightarrow$(B)), with improvement of confidence in localising small objects within cluttered X-
ray security imagery.

<table>
<thead>
<tr>
<th>Model</th>
<th>Bottle</th>
<th>Harder</th>
<th>Iron</th>
<th>lenon</th>
<th>P</th>
<th>tablet</th>
<th>Laptop</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>M RCNN</td>
<td>0.635</td>
<td>0.651</td>
<td>0.688</td>
<td>0.791</td>
<td>0.550</td>
<td>0.741</td>
<td>0.677</td>
<td></td>
</tr>
<tr>
<td>YOLACT</td>
<td>0.640</td>
<td>0.596</td>
<td>0.672</td>
<td>0.784</td>
<td>0.540</td>
<td>0.759</td>
<td>0.688</td>
<td></td>
</tr>
<tr>
<td>CARAFE</td>
<td>0.637</td>
<td>0.638</td>
<td>0.692</td>
<td>0.788</td>
<td>0.543</td>
<td>0.750</td>
<td>0.678</td>
<td></td>
</tr>
<tr>
<td>CM RCNN</td>
<td>0.650</td>
<td>0.659</td>
<td>0.708</td>
<td>0.803</td>
<td>0.560</td>
<td>0.761</td>
<td>0.683</td>
<td></td>
</tr>
<tr>
<td>M RCNN</td>
<td>0.607</td>
<td>0.615</td>
<td>0.665</td>
<td>0.761</td>
<td>0.521</td>
<td>0.749</td>
<td>0.652</td>
<td></td>
</tr>
<tr>
<td>YOLACT</td>
<td>0.641</td>
<td>0.597</td>
<td>0.649</td>
<td>0.756</td>
<td>0.533</td>
<td>0.765</td>
<td>0.657</td>
<td></td>
</tr>
<tr>
<td>CARAFE</td>
<td>0.631</td>
<td>0.624</td>
<td>0.678</td>
<td>0.754</td>
<td>0.522</td>
<td>0.757</td>
<td>0.661</td>
<td></td>
</tr>
<tr>
<td>CM RCNN</td>
<td>0.632</td>
<td>0.638</td>
<td>0.687</td>
<td>0.782</td>
<td>0.539</td>
<td>0.783</td>
<td>0.677</td>
<td></td>
</tr>
<tr>
<td>M RCNN</td>
<td>0.635</td>
<td>0.653</td>
<td>0.703</td>
<td>0.792</td>
<td>0.520</td>
<td>0.773</td>
<td>0.662</td>
<td></td>
</tr>
<tr>
<td>YOLACT</td>
<td>0.619</td>
<td>0.576</td>
<td>0.658</td>
<td>0.771</td>
<td>0.520</td>
<td>0.760</td>
<td>0.651</td>
<td></td>
</tr>
<tr>
<td>CARAFE</td>
<td>0.632</td>
<td>0.636</td>
<td>0.677</td>
<td>0.774</td>
<td>0.530</td>
<td>0.768</td>
<td>0.662</td>
<td></td>
</tr>
<tr>
<td>CM RCNN</td>
<td>0.641</td>
<td>0.627</td>
<td>0.673</td>
<td>0.784</td>
<td>0.541</td>
<td>0.778</td>
<td>0.674</td>
<td></td>
</tr>
<tr>
<td>YOLACT</td>
<td>0.545</td>
<td>0.521</td>
<td>0.629</td>
<td>0.798</td>
<td>0.489</td>
<td>0.718</td>
<td>0.616</td>
<td></td>
</tr>
<tr>
<td>CARAFE</td>
<td>0.550</td>
<td>0.492</td>
<td>0.629</td>
<td>0.786</td>
<td>0.522</td>
<td>0.718</td>
<td>0.616</td>
<td></td>
</tr>
<tr>
<td>CM RCNN</td>
<td>0.560</td>
<td>0.516</td>
<td>0.634</td>
<td>0.796</td>
<td>0.507</td>
<td>0.749</td>
<td>0.627</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Object detection results of CNN architectures using different X-ray imagery from the deei6 dataset.

Fig. 3. Confusion Matrix of the CARAFE [17] trained on rgb (A) and combination of \{rgb,hlz\} (B) X-ray imagery.

3.2. Cross-scanner Transferability

In this set of experiments (Table 2), we assess the CNN architecture performance across the X-ray imagery (\textit{dbs\_laptop} and \textit{dbr\_laptop}) from different scanner sources [23, 24]. The CNN architectures are trained using variants of dual-energy X-ray imagery of deei6 dataset but evaluated on a test set of only rgb pseudo-colour imagery (\textit{dbs\_laptop} and \textit{dbr\_laptop}).

The positive impact of combining \{rgb,hlz\} X-ray imagery is evident with all four CNN architectures (Table 2). For \textit{dbs\_laptop}, CARAFE [17] produces the best performance (AP: 0.835, Table 2, lower) when trained using combination of rgb, high, low and effective-z X-ray imagery, significantly outperforming rgb X-ray imagery (AP: 0.763, Table 2, upper). Similar significant performance improvement is noticeable on \textit{dbr\_laptop} dataset with CARAFE [17] achieving the highest AP of 0.611 (Table 2, lower). A plausible explanation for the performance improvement is that the variation in X-ray imagery by combining rgb, high, low and effective-z imagery during training, leads the CNN architectures to learn meaningful image features, which alleviates to achieve a higher degree of model generalisation in object detection within X-ray imagery. Although CARAFE [17] is a simpler architecture (49.41 million parameters) compared to the Cascade Mask R-CNN [18] (77.04 million parameters), it offers a better generalisation ability by training on a more varied set of multiple X-ray imagery variants. In Figure 4A, the target laptop is missed in both test images when trained solely on rgb imagery, but successfully detected when trained with combined \{rgb,hlz\} X-ray imagery (Figure 4B). Hence, we can deduce that although X-ray images are from differing scanners, the transferability of the trained CNN models is significantly improved by training over a more varied training set that includes both pseudo-colour rgb and variant dual-energy X-ray imagery.

Table 2. Object detection results (AP) on \textit{dbs\_laptop} and \textit{dbr\_laptop} datasets, where CNN architectures are trained using variant X-ray imagery from the deei6 dataset.

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Model</th>
<th>Test-set</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{dbs_rgb}</td>
<td>M RCNN</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>YOLACT</td>
<td>0.633</td>
</tr>
<tr>
<td></td>
<td>CARAFE</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td>CM RCNN</td>
<td>0.763</td>
</tr>
<tr>
<td>\textit{dbs_hlz}</td>
<td>M RCNN</td>
<td>0.807</td>
</tr>
<tr>
<td></td>
<td>YOLACT</td>
<td>0.782</td>
</tr>
<tr>
<td></td>
<td>CARAFE</td>
<td>0.835</td>
</tr>
<tr>
<td></td>
<td>CM RCNN</td>
<td>0.803</td>
</tr>
</tbody>
</table>

Fig. 4. Detection examples from \textit{dbs\_laptop} and \textit{dbr\_laptop} using CARAFE [17] trained on rgb (A) and \{rgb,hlz\} (B) X-ray imagery from deei6 dataset. White dashed box in (A) fails to detect the target.

4. CONCLUSION

This work examines the impact of X-ray imagery variants, i.e., dual-energy X-ray responses (high, low), effective-z and pseudo-colour (rgb), via the use of CNN architectures for the object detection task posed within X-ray baggage security screening. We illustrate that the combination of rgb, high, low and effective-z X-ray imagery produces maximal performance across all four CNN architectures for a six classes object detection problem, with CARAFE [17] achieving the highest mAP of 0.7. Furthermore, our results also demonstrate a remarkable degree of generalisation capability in terms of cross-scanner transferability (AP: 0.835/0.611 with CARAFE [17]) for a one class object detection problem by combining \{rgb, hlz\} X-ray imagery. This clearly illustrates a strong insight into the benefits of using a combination of dual-energy X-ray imagery for object detection and segmentation tasks, which could additionally useful for component-wise anomaly detection analysis. Future work will consider the use of dual-energy variant imagery for combined material discrimination and anomaly detection within cluttered X-ray security imagery.
5. REFERENCES


