SLX: Similarity Learning for X-Ray Screening and Robust Automated Disassembled Object Detection

SLX: Similarity Learning for X-Ray Screening and Robust Automated Disassembled Object Detection

Nikolaos Dionelis
Electronics and Electrical Engineering
The University of Edinburgh, UK
Contact email: nikolaos.dionelis@ed.ac.uk

Richard Jackson
Counter Terrorism and Security Division
Defence Science and Technology Laboratory (Dstl)
Porton Down, UK

Sotirios A. Tsaftaris
Electronics and Electrical Engineering
The University of Edinburgh
Edinburgh, UK

Mehrdad Yaghoobi
Electronics and Electrical Engineering
The University of Edinburgh
Edinburgh, UK

***Abstract*—Baggage screening is important in security-critical applications in airports for detecting threats, including firearms and parts of them. Existing approaches underperform to recognise prohibited objects that are disassembled, especially when learning from limited data and from images produced by different scanners with multi-view orientations. To address such limitations, in this paper, we develop the Similarity Learning X-ray screening (SLX) model for accurate and robust firearm component detection in cluttered scenes. We evaluate SLX on the X-ray Image Library (XIL) dataset that the UK Government has provided us with, for this research. SLX is based on a contrastive similarity learning approach combined with Out-of-Distribution (OoD) detection/ anomaly detection using a deep discriminative model, ResNet-152, for detecting and classifying forbidden items. The evaluation of SLX on the XIL dataset shows that it is effective, beneficial for detecting firearms and their parts, and outperforms other baseline models, on average, by approximately 12 points in accuracy.**

***Index Terms*—X-ray security imaging; Baggage X-ray screening**

I. Introduction

Security screening. Baggage X-ray inspection is important for protecting public space from safety threatening, including terrorism. The screening of luggage is a *core* standard checking measure in airports [1], where security is of significant concern. Airports strive to automate detection to improve effectiveness and efficiency, even for firearm component detection within passengers' baggage, reducing errors and processing times. The problem is the following. From labelled data, we train a model that infers whether an image contains prohibited items, which are defined as items we would not want inside an airplane or parts of them, e.g. firearms and *their components*. The problem we consider in this paper is disassembled object detection in cluttered environments. The proposed discriminative Similarity Learning X-ray baggage screening (SLX) model aims at improving the accuracy for firearm and gun part detection.

Automation. Deep learning has brought an evolution to computer vision improving the detection, discovery, and recognition of threats [1], [2]. In the real world, threat items may be disassembled and mixed with other items, creating a cluttered scene for recognition. Moreover, screening is susceptible to operator errors due to exhausting work schedules, *occlusion*, clutter, and concealed threats. Discriminative models trained with labelled data in a supervised manner for X-ray security screening should be able to detect and recognise disassembled objects, and the SLX model addresses such challenges [1].

Using different scanners. X-ray images are produced by scanners of different type [1], i.e. with different illumination and variable multi-view orientations, and are stored in different formats. Training and testing using scanners with different instrument characteristics, including variable *multi-view* orientations, is challenging and usually does not lead to good results due to non-quantified differences between the scanners. Our proposed SLX model aims at addressing such limitations and at providing a solution for using different scanner machines.

Novelty. The main contribution of this work is the development of the Similarity Learning X-ray screening (SLX) model to detect and recognise guns and disassembled objects/ firearm components in *cluttered* scenes with occlusion, limited training data, and data from multiple domains, i.e. images produced by different scanners. Specifically, our main contributions are:

- The methodological approach of SLX based on contrastive similarity learning, enhancing similar image representations, combined with OoD detection/ anomaly detection using deep discriminative models. To improve prohibited item and *disassembled object* detection, SLX performs joint contrastive learning and classification cross-entropy minimisation.
- To address threat recognition limitations and mitigate overfitting, for training with *limited data* in particular, we minimise a multi-task objective in Sec. III. The SLX model achieves good performance, outperforming other baseline models.
- SLX achieves improved generalisation performance in cluttered scenes, including for data with *multi-view* orientations and data from multiple domains, from different scanners.

The proposed SLX model. SLX is based on multi-view similarity learning of representations, on contrastive learning and the cosine similarity measure, as well as on the cross-entropy loss using the probability of the labels. SLX is trained

and evaluated on a real-world dataset, i.e. X-ray Image Library (XIL), which contains labelled data, including *parts* of firearms, as shown in Fig. 1(a), in cluttered scenes in baggage and parcels. Our experiments show that SLX is effective, its ablation study is a success, and our model outperforms other baseline models. SLX achieves good *generalisation performance*, and this work's methodological and model development values, as well as its application implementation value, are high. The obtained results can be useful for researchers and practitioners. A contribution of this paper is the value for applications, so that researchers studying this real-world problem [1], [2] can take advantage of the evaluation results and the attained good performance.

II. Related Work and Main Challenges

The general problem setting. The general methodologies that have been applied to the baggage problem in the related work are: (i) The Out-of-Distribution (OoD) approach, where the model trains on benign images only, and infers a threat if the image is OoD (i.e., an anomaly) with respect to the learned benign image distribution. Recent work on this topic is based on *generative* models, where a categorisation is between Generative Adversarial Network (GAN) based approaches [2], [3] and Autoencoders (AE) [21], [20]. (ii) The discriminative model approach: The model trains on *labelled* threat and benign data, where the labels are benign, firearms, etc. The model infers a threat if the data is classified as a known threat class.

The general methodologies that could be applied to the X-ray security screening problem [2], [34] are: (a) OoD detection using deep generative models [2], [3], (b) Discriminative models with labelled data [6], [7], (c) Open-Set, or even *Open-World*, classification, combining OoD detection and discriminative models, and (d) Contrastive learning to achieve improved separation of the classes where similar image representations are *attracted* [18], [7], while different image latent feature representations are repelled and pulled apart [27], [18].

Contrastive learning combined with OoD detection using discriminative models. Existing methods for baggage screening [1], [21] are lacking for joint contrastive similarity learning, enhancing similar image representations, and classification cross-entropy minimisation [28], [2]. The accurate detection of components of firearms, for limited training data, using different scanners is also lacking. Generalisation, e.g. data from multiple scanners with different *multi-view* orientations, needs improvement. The OoD detection capability of discriminative classifier models for X-ray screening is also lacking [12].

Baggage screening and data scarcity. One of the obstacles to the development of deep learning screening technology is limited data. While for recognition problems with RGB images, large amounts of data can usually be collected, X-ray baggage images are *difficult* to obtain. In this research work, to train and evaluate SLX, we use the XIL dataset which contains airport security data [28], [22], and we examine the performance of our model for detecting and recognising threats, firearms, and components of guns when *limited data* are provided.

Data augmentation. Because of limited available data, when aiming at learning good representations, data augmentation techniques are needed for effective training for threat item detection, in aviation security. No single transformation suffices to learn good data and class feature representations. One type of data augmentation involves geometric transformations, such as rotation, horizontal and vertical flipping [1], [2], and cropping followed by resizing. Another type of augmentation involves appearance transformations, such as color distortion, including brightness, contrast, color dropping, and blur [18]. Aiming at recognizing firearms and their disassembled components from images of *different* scanners, data augmentation, e.g. cropping and color distortion, is beneficial to improve performance.

Using different multi-view X-ray scanners. Transferring models from one scanner to another is challenging. According to the European Commission, the transfer of models between domains and different scanners is difficult and might lead to bad results [1]. Multi-view classification focuses on improving classification accuracy using information from different views, integrating these views into a good feature representation. Classification based on multi-view image data is well-suited for objects characterised by *intra-class* and inter-class similarity, when different views of same items provide complementary information [1], [17]. In this work, SLX performs normalisation of data from different scanners, and we use the XIL dataset to evaluate our model using *cross-scanner* test samples. The algorithm is based on a similarity learning loss term which enhances the multi-view representations of the image data.

Working with unseen X-ray scanners. Our setting significantly differs from and is more realistic than the examined problem setting in [16], where a combined dataset of scanners is used. In real-world scenarios, samples from a specific scanner may not be possible to access during training, *but only* during inference and testing [1]. Our problem setting and SLX model *differ* from those in [14], where X-ray screening is performed using an encoder-decoder architecture to recognise threats.

III. The Proposed SLX Model

We develop the proposed firearm threat and gun component detector model, which is presented in Fig. 1(b). Our aim is to perform accurate firearm threat and disassembled gun detection. SLX is based on a discriminative classifier, ResNet-152 [11], [7], developed for multi-GPU training for efficient learning with large batch sizes for faster and improved convergence, for detecting and classifying forbidden items. With SLX, successful training is achieved using a large batch sample size, for faster and improved convergence [11], [18]. ResNet-152 achieves an accuracy of 0.87 on ImageNet [19], and outperforms VGG-16 with 0.63, and Inception with 0.78. The training of SLX is based on a contrastive learning loss function which enhances multi-view similarity representations, as well as on the cross-entropy objective loss using the probability of the class labels, also adding model parameter regularisation. SLX is evaluated on the XIL dataset; here, the main challenging and interesting features of XIL, which has been verified, amongst others, by the Turing Institute, are dismembered weapons and firearms, limited data, multi-view images, and the two *different* scanners.

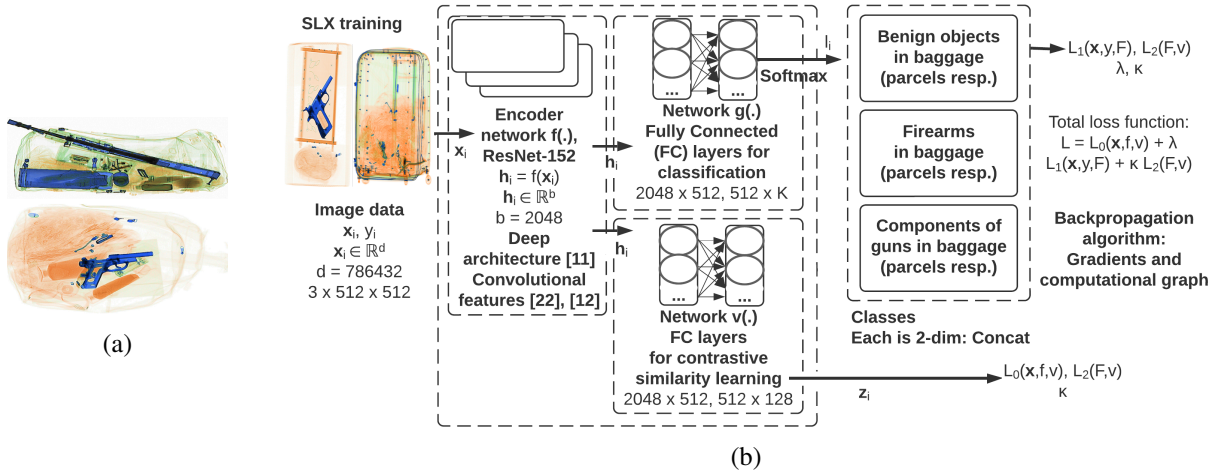


Fig. 1: (a): Disassembled items. (b): Flowchart of SLX for joint contrastive similarity learning and classification cross-entropy minimisation. The data, at the upper-left corner, are reshaped to $\mathbb{R}^d$ and are processed by the backbone encoder $f(\cdot)$ to obtain $\mathbf{h}_i$ [11], [32]. The samples $\mathbf{h}_i$, for i between 1 to N, are processed by the classification head, $g(\cdot)$, and the *projection* head, $v(\cdot)$, at the middle top and bottom, respectively. The outputs of $g(\cdot)$ are passed through the softmax [6], [22] to obtain the classes, 1 to K, [33], [36]. The labels are passed through the *loss* terms $L_1(\mathbf{x},y,F)$ and $L_2(F,v)$, in the loss in (1). The outputs of $v(\cdot)$ are passed through the loss terms $L_0(\mathbf{x},f,v)$ and $L_2(F,v)$, in (1)-(3). Here, dashed-line boxes are used to indicate same modules.

Also, XIL is challenging because of containing *occlusion* and clutter, both baggage and parcels, and different backgrounds.

A. Mathematical formulation of the proposed approach

Formulation of the problem and our method. We denote the data by $\mathbf{x}$ where $\mathbf{x}_i$ are the labelled image data with labels y_i between 1 and K, where K is the number of classes. Let $\mathcal{X} \in \mathbb{R}^d$ be the set of inputs, modelling objects of interest such as images [8], [9], and let $\mathscr{L}$ be the set of class labels. Let $F : \mathcal{X} \to \mathscr{L}$ be a classifier [6], [7] that assigns a unique label from the set $\mathscr{L}$ to an element from $\mathcal{X}$. Now, we denote the latent feature representation of our deep learning model by $\mathbf{h}$, where $\mathbf{h}_i = f(\mathbf{x}_i)$ [18], [24]. Let $F = g \circ f$, where $f : \mathcal{X} \to \mathscr{H}$ and $g : \mathscr{H} \to \mathscr{L}$. Here, without loss of generality, we denote the latent feature representation space by $\mathscr{H} \in \mathbb{R}^b$, where $b < d$. The multi-class classifier, F, assigns a label, i.e. $l_i = g(\mathbf{h}_i)$, to a data input, $\mathbf{x}_i$, which has the corresponding *latent* feature representation $\mathbf{h}_i$. Regarding the data, samples $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathscr{L}$ are drawn from the underlying data distribution, $P(\mathbf{x},y)$.

Backbone encoder network and projection heads. The model $f(\cdot)$ can, for example, be ResNet [11], [23]. The model $g(\cdot)$ is followed by a softmax output layer, i.e. the normalised exponential, softmax(l_i), to obtain the probability over the K classes [6], [25]. For ResNet, $\mathbf{h}_i$ is the output after the final average pooling layer. In addition, without loss of generality, the model $g(\cdot)$ can, for example, be a small neural network classification head that maps the learned features to the output K nodes, which are for the K classes [7], [26]. Here, this *classification head* network can be linear, i.e. $W\mathbf{h}_i$ with matrix dimensions $b \times K$, or a Multi-Layer Perceptron (MLP) with one hidden layer to obtain $l_i = g(\mathbf{h}_i) = \mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{h}_i)$ where σ is a nonlinear activation function, e.g. *ReLU*, and $\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights/ model parameters of the fully-connected layers with dimensions 2048×512 and $512 \times K$, respectively. These dimensions are for when the 50-layer, 101-layer, or 152-layer ResNet is used. When the 18-layer or 34-layer ResNet is used, then $\mathbf{W}_1$ and $\mathbf{W}_2$ have dimensions 512×256 and $256 \times K$.

To alleviate overfitting, to perform effective learning with limited data, and to take *full advantage* of the data without their labels [18], we perform joint contrastive similarity learning and classification cross-entropy minimisation using a neural network projection head $v(\cdot)$. This head maps the feature representations to a space where a contrastive similarity objective is minimised. We denote this mapping by $v : \mathscr{H} \to \mathscr{Z}$, where $\mathscr{Z} \in \mathbb{R}^c$ and $c < b$. Here, $v(\cdot)$ can be a MLP with one hidden layer, to obtain $\mathbf{z}_i = v(\mathbf{h}_i) = \mathbf{W}_3\sigma(\mathbf{W}_1\mathbf{h}_i)$, where σ is a nonlinear activation, e.g. ReLU, and $\mathbf{W}_1$ and $\mathbf{W}_3$ are the weights of the fully-connected layers with dimensions 2048×512 and 512×128, respectively, i.e. $c = 128$. These dimensions are for when the 50-layer, 101-layer, or 152-layer ResNet is used. When the 34-layer ResNet is used, $\mathbf{W}_1$ and $\mathbf{W}_3$ have dimensions 512×256 and $256 \times c$.

B. Loss function

For the data, $(\mathbf{x}_i, y_i)_{i=1}^N$, where $\mathbf{x}_i$ is a vector of length, for example, 786432 for data from XIL, N is the number of training samples, and i is a sample index that takes integer values from 1 to N. The objective cost function of the proposed SLX deep classifier, which is minimised during training, is given by

$$\arg\min_{F,v} L(\mathbf{x},y,F,v), \tag{1}$$

$$\text{where } L = L_0(\mathbf{x},f,v) + \lambda\ L_1(\mathbf{x},y,F) + \kappa\ L_2(F,v), \tag{2}$$

$$\text{and where } L_0 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{z}_i,\tilde{\mathbf{z}}_i/\tau)}{\sum_{k=1}^{N}\mathbb{1}_{k\neq i}\exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_k/\tau)}, \tag{3}$$

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(F_{y_i}(\mathbf{x}_i))}{\sum_{k=1}^{K}\exp(F_k(\mathbf{x}_i))}, \quad L_2 = \sum_{m=1}^{M}||\mathbf{W}_m||_1, \tag{4}$$

where for $F(\cdot)$, the model $f(\cdot)$ is the ResNet-152 discriminative model [11], [7] for classification with K classes. The objective function is the similarity learning loss and the cross-entropy loss between the probability of the labels and the probability of the predictions. The loss in (1) uses the cosine similarity (sim) and the temperature, τ. SLX, with its *multi-task* objective, brings close together the data augmented samples and their representations with contrastive similarity learning [18], [7]. The step in (3) has been used also in [18], [27], and the focal point of this paper is the joint similarity learning and cross-entropy loss minimisation in (1). We develop this simultaneous contrastive learning, enhancing similar image representations, and cross-entropy loss minimisation framework for screening and disassembled object detection, while also providing out-of-data-distribution capability, and to the best of our knowledge, this is the first time this has been done. Joint contrastive learning and classification cross-entropy minimisation is beneficial for the detection and classification of dismantled objects because benign items and gun *parts*, as well as firearms and gun parts, are *near* classes. Here, combined contrastive similarity learning and cross-entropy loss minimisation is advantageous for such near classes that contain samples that are *difficult* to classify correctly, as it groups same-class samples and repels different-class samples away from each other. In this way, SLX learns and captures intra-class variations/ variability of the classes.

The SLX model performs joint contrastive learning and classification cross-entropy minimisation for firearm and gun part detection and recognition, minimising the objective cost function in (1) by using the Stochastic Gradient Descent (SGD) algorithm [11], [8]. In this way, SLX detects and classifies threats and *disassembled* prohibited items, such as components of firearms (e.g. grip and handle, upper canopy, and barrel).

The output of SLX, as presented in Fig. 1(b), is the inferred class, e.g. parts of guns, as well as the Threat/ No Threat post-processing decision. SLX aims at reducing *failures* to detect and recognise disassembled items, as well as false alarms of disassembled items. SLX in (3) computes the cosine similarity [18], [27] between the learned representations of: (i) the data, and (ii) the stochastically augmented samples. Here, the cosine similarity for two vectors, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^c$, is the dot product between the l_2-norm normalised vectors: $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/||\mathbf{u}||\,||\mathbf{v}||$. The objective loss in (1) also uses the class label index k, which takes values between 1 and K, the indicator function $\mathbb{1}_{k \neq i}$ which is 0 when $k = i$ and 1 otherwise, the network weights $\mathbf{W}$ for $F(\cdot)$ and $v(\cdot)$, and the hyper-parameters λ and κ.

The multi-loss objective, L, in (1) is a function of: (a) the data and their labels, i.e. $\mathbf{x}$ and y, (b) the classifier, $F(\cdot)$, and (c) the projection head, $v(\cdot)$. Because $F = g \circ f$, the first loss term, L_0, in (2) is a function of $\mathbf{x}$, the feature extractor *encoder* network, $f(\cdot)$, and $v(\cdot)$. The second loss term, L_1, is a function of $\mathbf{x}$, y, and $F(\cdot)$, while the third term, L_2, is a function of $F(\cdot)$ and $v(\cdot)$. In (3), as also shown in Fig. 1(b), L_0 is expressed in terms of $\mathbf{z}_i = v(\mathbf{h}_i)$, where $\mathbf{h}_i = f(\mathbf{x}_i)$, and $\tilde{\mathbf{z}}_i = v(f(\tilde{\mathbf{x}}_i))$, where $\tilde{\mathbf{x}}_i$ is the stochastically data augmented sample of $\mathbf{x}_i$ [18], [27], specifically with random colour distortion and random cropping followed by resizing. In (4), L_1 is expressed in terms of: (i) $\mathbf{x}$, (ii) y in the numerator of the *ratio*, and (iii) $F(\cdot)$. Here, $F_r(\cdot)$, for r between 1 and K, is the output network node for the class label r, where r is either the label y or the class index k, and K is the number of classes. In (4), L_2 is expressed in terms of the neural network weights of $F(\cdot)$ and $v(\cdot)$, i.e. $\mathbf{W}$, where M is the total number of network model parameters.

The multi-loss objective in (1)-(4) is minimised. In Fig. 1(b), the entire architecture is trained in an *end-to-end* manner [6], [33]. The trainable parameters of SLX in (1) are $T = \{F, v\}$, while the model hyper-parameters, i.e. λ and κ in (2), control the trade-off between the loss terms [36], [33]. Here, the first loss term in (2), i.e. L_0, uses $\mathbf{z}_i = v(\mathbf{h}_i)$ where $\mathbf{h}_i = f(\mathbf{x}_i)$, and is thus a function of $\mathbf{x}$, $f(\cdot)$, and $v(\cdot)$ [11], [18]. The second loss term, L_1, uses $l_i = g(\mathbf{h}_i)$, and is a function of $\mathbf{x}$, y, and $F(\cdot)$. SLX obviates the use of more complex structures and network architectures, such as the ones that try to compute, model, and capture data likelihood and the probability density of a mixture combination of the classes [36]. During testing, the inferred label of each image is found: $\hat{l}_i$. For *any* test sample $\check{\mathbf{x}}_i$, then $\hat{l}_i = g(\check{\mathbf{h}}_i)$ where $\check{\mathbf{h}}_i = f(\check{\mathbf{x}}_i)$. Here, because $F(\cdot)$ and $v(\cdot)$, i.e. T, have been learned using labelled training data in (1), for any queried test sample $\check{\mathbf{x}}_i$, the correlations and interdependencies between this $\check{\mathbf{x}}_i$ and the training data and their labels are modelled, captured, and expressed with $\hat{l}_i$.

Benefits of SLX in confidence measures. By minimising (1), SLX computes a measure of uncertainty. Such measures of confidence, certainty, and trust are crucial in real-world applications. They are crucial for operators in security screening [28], [1]. Certainty assessment and confidence assignment of the model's prediction is useful for the X-ray screening and *part detection* problems. SLX's estimation uncertainty is based on the second loss term, i.e. $L_1(\mathbf{x}, y, F)$ in (2). The confidence measure of SLX is entropy-based and is based on the mutual entropy between the model probability of the labels, and the empirical probability of the labels, i.e. the prediction confidence of the cross-entropy loss. The multi-task objective in (1) has *benefits*, as contrastive similarity learning and cross-entropy loss minimisation are performed jointly, rather than sequentially. If similarity learning was performed on its own [18] followed by Nearest Neighbours, the *confidence measure* would be the distance from the cluster center. Probability metrics, including any entropic measure such as that of SLX, have advantages over *geometric* distances as entropy-based confidence measures model and capture probability, the notion of *likelihood*, and the bulk/ high mass of the underlying distribution of the data.

IV. Evaluation of the Proposed SLX Model

A. The XIl dataset

Image dataset of firearms. To demonstrate that the proposed approach is successful, we use the XIL dataset, which has been provided to us by the UK Government. It includes labelled data and images of components of firearms. It contains gun parts, full-weapon firearms, and non-threat items. The baggage and parcel X-ray images of *components* of firearms are suitable for training part detector models. In this work, we develop the SLX detector model to recognise firearms and their components.

TABLE I: Performance of SLX in accuracy on the XIL dataset. The training set samples are from Scanner A *only*, joining full-weapon firearms and gun components in one class.

CLASSES (TRAIN: SC. A)	TEST: SC. A	TEST: SC. B
AVERAGE	99.54%	89.15%
BENIGN PARCELS	95.40%	86.11%
BENIGN BAGGAGE	**99.86**%	82.75%
WEAPON AND PARTS PARCELS	99.65%	96.17%
WEAPON AND PARTS BAGGAGE	98.12%	**98.10**%

TABLE II: Performance of SLX for scanner generalisation in accuracy on XIL. The training is on both Scanners A and B.

CLASSIFICATION	SC. A	SC. B
AVERAGE	94.93%	89.12%
BENIGN PARCELS	95.64%	87.93%
BENIGN BAGGAGE	95.88%	84.32%
FULL-WEAPON PARCELS	**98.02**%	93.41%
FULL-WEAPON BAGGAGE	97.12%	**93.81**%
FIREARM COMPONENTS PARCELS	92.83%	83.24%
FIREARM COMPONENTS BAGGAGE	91.32%	81.13%

Multi-view learning. XIL is suitable for multi-view learning as it includes threat items taken from different viewpoints: four- and two-view orientations. SLX, with *multi-view* data in parcels and baggage, learns and recognises firearm parts in clutter, i.e. Fig. 1(a), where images from Scanners A and B are shown.

B. Data augmentation

We examine the applicability of data augmentation methods for joint threat detection and classification. We apply different methods comprising rotation (5 to 15 degrees), horizontal and vertical flipping, as well as cropping and color distortion, which improve the performance. Inside their luggage, travellers can put their clothes/ items in *any* position they want, and this is why horizontal flipping is beneficial. Using both deterministic and *stochastic* data augmentation algorithms [18], [29], we effectively significantly increase the number of samples. The augmentation strategy we use [27], [30], including stochastic augmentation, e.g. random image cropping followed by resizing, aids disassembled object detection and multi-view learning.

C. Using different X-ray scanners and generalisation

SLX performs normalisation of image data to achieve good cross-scanner generalisation performance across different scanners. To improve cross-scanner generalisation and attain good performance for scanners with less/ *half* multi-viewpoint images, which is challenging [1], SLX performs mean and variance standardisation of images from the two scanners.

Model initialisation. To help SLX recognise the different classes effectively, we use model initialisation and start from a discriminative model trained on the ImageNet dataset [22]. With this initialisation, we: (i) improve our model's *generalisation* performance, (ii) alleviate the problem of overfitting [7], [23], and (iii) achieve improved convergence/ learning behaviour.

D. Examined settings

We evaluate SLX and compute the classification accuracy, both on average and per-class. We train SLX on data from one scanner only, as well as on data from two scanners. We examine the settings where we test SLX on Scanners A and B separately, either for the class Threats, which combines firearms and their components, or separately for the classes Full-weapon firearms and Gun parts. The latter includes either all the parts, or even *some of them*. We define training on data only from one scanner, and testing on data *only from* the other unseen scanner, by cross-scanner generalisation. Achieving good cross-scanner generalisation is challenging due to domain adaptation and different machine intrinsic properties [1], [31], [22]. Cross-scanner generalisation is even more challenging when the one scanner produces four-view data, while the other scanner machine produces *two views*, which is our examined setting. We also compare the SLX model with other baseline models, examining the performance improvements of our model.

Generalisation for the Threat/ No Threat binary case (trained on Scanner A): Same- and cross-scanner tests. We train SLX on data from Scanner A. Here, firearms and their parts are considered one class, i.e. *Threat*. The results of testing on images from Scanner A and Scanner B separately, presented in Table. I, show that SLX achieves 99.65, in accuracy, for the threat class for images from Scanner A. SLX also achieves 98.10 for the threat class, for image data from Scanner B.

Same-scanner evaluation trained on two scanners. When training SLX on data from Scanners A and B, the results of testing on Scanner A or B, as presented in Table II, show that SLX achieves 92.83 in accuracy for firearm components on Scanner A, and 83.24 for firearm parts on Scanner B. In Table II, SLX is trained on Scanners A and B, while in Table. I, our model is trained *only on* Scanner A. In Table. I, SLX is trained on data combining guns and their components in one class, the threat class, while in Table II, SLX is trained on data from the classes of benign items, guns, and parts of guns. SLX achieves improved performance when combining guns and their components in a single class. In Table II, SLX achieves 98.02 for full-weapon firearms on data from Scanner A, and 93.81 for full-weapon guns on Scanner B data. The results in Table. I show that SLX's joint similarity learning and discriminative cross-entropy minimisation, enhanced by the data augmentation strategy we use, is effective and improves the performance for the detection of full-weapon and disassembled firearms.

Same-scanner evaluation. In Table III, which presents our model's results on XIL, SLX is trained on image data only from Scanner A and achieves the per-class accuracy of 98.12 for the class full-weapon firearms, and of 90.91 for the class

TABLE III: Performance of SLX and of the *baseline* model, ResNet, on XIL, in accuracy, using same-scanner data from Scanner A. Training and testing is performed on Scanner A.

CLASSIFICATION	SLX	RESNET
AVERAGE	92.29%	79.67%
BENIGN PARCELS	93.47%	82.17%
BENIGN BAGGAGE	92.53%	88.06%
FULL-WEAPON IN PARCELS	95.99%	81.28%
FULL-WEAPON IN BAGGAGE	**98.12**%	84.22%
FIREARM COMPONENTS PARCELS	89.89%	72.83%
FIREARM COMPONENTS BAGGAGE	90.91%	71.55%

TABLE IV: Accuracy of SLX and of the baseline model, ResNet, for cross-scanner generalisation on *XIL*. The training set data are from Scanner A and the test from Scanner B.

CLASSIFICATION	SLX	RESNET
AVERAGE	76.31%	54.41%
BENIGN PARCELS	76.04%	53.27%
BENIGN BAGGAGE	71.73%	47.58%
FULL-WEAPON IN PARCELS	**86.54**%	61.79%
FULL-WEAPON IN BAGGAGE	85.84%	65.17%
FIREARM COMPONENTS PARCELS	70.14%	45.17%
FIREARM COMPONENTS BAGGAGE	64.72%	49.31%

firearm components. Comparing Table I to Table III, SLX shows improved generalisation performance when combining guns and *their components* in one class, i.e. accuracy of 99.65.

E. Comparison of the proposed SLX model to ResNet

We evaluate the SLX model and compare it to other baseline models. The structure of this and the next parts of the evaluation section is the following. We first compare SLX to the model ResNet [11], [22], [12]. Next, we present the evaluation results of SLX, and we compare its performance to that of the model proposed in [18]. Then, we *compare* SLX to the model Skip-GANomaly [21], [20]. In the next paragraphs, we show that the SLX model outperforms other baseline models on XIL.

Comparing SLX to ResNet for same-scanner generalisation. The evaluation results of SLX and of the base model, ResNet, are presented in Table III. These results are for the two models trained and tested on XIL, on data from Scanner A only. SLX is effective and significantly outperforms ResNet, in accuracy. The *improvement* of SLX upon ResNet, for the class of components of firearms, in absolute value, is 18.08. For the class guns, the improvement is 13.90. For both assembled/ full objects and for parts of items, in this case guns, SLX outperforms ResNet. We observe in Table III that not every ResNet model solves the examined problem, but SLX, which is based on the discriminative classifier ResNet-152 [11], [23], developed for multi-GPU training for efficient learning with large batch sizes for faster and improved convergence [11], [18] for detecting and classifying prohibited items, does solve this problem. Because ResNet uses the classification cross-entropy loss rather than *joint* contrastive similarity learning and cross-entropy minimisation in (1), ResNet is not specifically designed and 100% designated for security screening [1], [37], neither is it specially designated for disassembled object detection.

Further comparison of SLX to ResNet, for cross-scanner generalisation. We evaluate SLX trained on data from Scanner A, and evaluated on Scanner B, in Table IV, and we compare its performance to the model ResNet [11], [12]. For cross-scanner generalisation, the *improvement* of SLX upon ResNet for the class of components of firearms, in absolute value, is 20.83. For the class guns, the improvement is 31.37. SLX outperforms ResNet in terms of same-scanner and cross-scanner generalisation performance in Tables III and IV, respectively.

F. Comparison of the SLX model to other baseline models

Comparing SLX to SimCLR. We compare SLX to the model from [18], i.e. the Simple framework for Contrastive Learning of visual Representations (SimCLR). SLX outperforms SimCLR in Table V as SimCLR achieves the accuracy of 87.33 on data from the XIL dataset from Scanner A, and of 78.72 on data from Scanner B. The performance improvement of SLX in Table V, compared to SimCLR, is approximately 11.0 in accuracy, in absolute value, on data from Scanner A. Also, the *improvement* of SLX is approximately 14.0 on data from Scanner B. These results are for threat detection for the class firearms and their parts, as in Table. I. To compare our model with other baseline models, we implement and run on XIL the models we compare SLX with, e.g. SimCLR [18], [27]; here, we do not use the evaluation results from tables of other publications. We observe in Table V that not every contrastive learning model solves the examined problem, but SLX, which performs combined contrastive similarity learning and classification cross-entropy minimisation, effectively optimising the multi-task loss in (1), solves this problem. Contrastive learning is beneficial for *near* classes that have a small distance between them [9], [8] and contain data samples that are difficult to correctly classify, but not every contrastive similarity learning algorithm solves the examined problem. Because SimCLR first performs contrastive learning and, then, fine-tuning [18] for classification and label assignment, SimCLR is not specifically designed and designated for the X-ray security screening problem [1], neither is it specially designated for disassembled object detection.

Comparing SLX to Skip-GANomaly. We compare SLX to Skip-GANomaly [21] in Table V. Skip-GANomaly outperforms GANomaly by a large margin, i.e. 28 points in AUROC [21] for detecting parts of firearms. Here, we implement and run Skip-GANomaly on the XIL dataset, and according to the results in [21], this model achieves good performance for baggage

TABLE V: Comparison of SLX to the models SimCLR [18], Skip-GANomaly [21], and Compositional Network [36], in accuracy, on *average*, on XIL. The training is on Scanner A.

MODEL (TRAIN: SC. A)	TEST: SC. A	TEST: SC. B
SLX	**99.54**%	**89.15**%
SIMCLR	87.33%	78.72%
SKIP-GANOMALY	85.81%	60.80%
COMPOSITIONAL NET	84.59%	69.43%

X-ray screening. The Threat or No Threat decision capability of Skip-GANomaly aligns *well* with the examined setting. Skip-GANomaly is a state-of-the-art model for X-ray security imaging, because it effectively addresses both the assembled/ full-weapon firearm detection problem and the detection of parts of guns [21]. For the implementation of Skip-GANomaly, we have used data augmentation and high resolution images, *same* as in SLX, to improve performance. We have trained Skip-GANomaly on X-ray data from XIL, on benign objects in baggage and parcels from Scanner A in particular. We have evaluated Skip-GANomaly, which computes distances in the data and latent feature spaces, on image data from *XIL*, i.e. on benign items and on firearms and parts of guns in baggage and parcels from Scanners A and B in particular. In Table V, SLX outperforms Skip-GANomaly for testing on data from Scanners A and B separately, in accuracy. Here, the *improvement* of SLX upon Skip-GANomaly is approximately 13.74 and 37.30 for image data produced by Scanners A and B, respectively.

We have thus compared SLX to Skip-GANomaly [21], [20], SimCLR [18], and ResNet [11]. [21] achieves state-of-the-art performance on a different benchmark dataset, which however has an overlap with the examined image dataset, i.e. XIL, that is, some of the examined images are *the same*. This is the main reason we compare SLX to Skip-GANomaly [28], [21]. The main difference between the datasets is the setting, as SLX deals with images from two scanners, limited data, and both same-scanner and *cross-scanner* generalisation performance.

G. SLX anomaly detection/ OoD detection

OoD detection ability. SLX can also detect OoD data, and its OoD detection functionality is the following. We define anomaly as samples that are outside the support of the normal class data distribution, i.e. residing in the complement of the *normal class* data distribution's support [8]. To accurately and robustly detect OoD data, SLX computes the OoD/ anomaly score which is the prediction confidence [6]. We use the SLX model in Table. I, and we evaluate our model's OoD detection ability. Here, the performance of SLX for OoD detection/ anomaly detection is 96.3% in AUROC and 96.5% in accuracy when the normal class data are from XIL and abnormal data from the CIFAR-10 image dataset. When the *abnormal data are images of ammunition* in baggage or parcels, the performance of SLX for outlier detection is 88.6% in AUROC and 87.8% in accuracy. In this scenario, the model Skip-GANomaly yields an AUROC of 76.3% and an accuracy of 75.5%. Comparing SLX with the baseline model Skip-GANomaly, the performance improvement of SLX in accuracy, in absolute value, is 12.3.

TABLE VI: Ablation study of SLX trained on Scanner A only, in accuracy, and comparison to models that do *not perform* a specific operation of SLX, i.e. without/ w/o contrastive similarity learning with L_0 in (1), classification cross-entropy minimisation with L_1, and data augmentation in (3).

MODEL (TRAIN: SC. A)	TEST: SC. A	TEST: SC. B
SLX	**99.54**%	**89.15**%
SLX W/O L_0	95.42%	83.72%
SLX W/O L_1	94.57%	82.49%
SLX W/O L_0, CROP, COLOR	94.81%	85.26%
SLX W/O CROP, COLOR	95.90%	83.78%
SLX W/O PRETRAINING	86.02%	71.64%

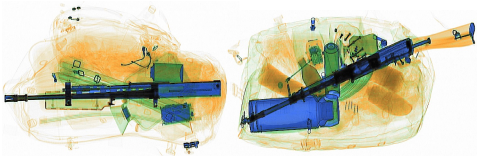
H. Ablation study of SLX

We turn off specific modules of the proposed model SLX in Fig. 1(b), including L_0 or L_1. The mean accuracy of SLX trained on Scanner A, evaluated on Scanner A or *Scanner B*, is shown in Table VI. We present the accuracy on average for the classes in Table. I. L_0 and L_1 in (1) are both beneficial; they are needed for *improved performance* and are equally important. This is based on the performance of (i) SLX, (ii) SLX without (w/o) L_0, and (iii) SLX w/o L_1. Performing (a) contrastive similarity learning with L_0, and (b) stochastic data augmentation as presented in Sec. III is *key*. When ablating L_1, we perform contrastive learning followed by fine-tuning, which outperforms both linear evaluation and Nearest Neighbours.

Stochastic data augmentation, i.e. random cropping followed by resizing, and random colour distortion, is beneficial for cross-scanner generalisation according to Table VI, more specifically for (a) SLX, (b) SLX without random cropping followed by resizing and random colour distortion, and (c) SLX without L_0, random cropping followed by resizing, and random colour distortion. Here, for SLX, for *cross-scanner* generalisation, data augmentation reflects, models, and captures the domain change and improves the performance.

For X-ray, different scanners are usually utilised in practice, producing data with variable characteristics, *multi-view* orientations, pixel intensity colour levels, illumination including glare and lighting, scale, and resolution including *edges* and smooth or rough contour representations. We have examined appropriate data augmentation that reflects, captures, and models the distribution change. The data that *reflect* the distribution change in our model are part of the samples of the data augmentation included in training, improving performance.

Contrastive learning is *beneficial* for cross-scanner generalisation. This is based on the performance of SLX, SLX w/o L_0, and SLX w/o L_1 on XIL's Scanner B in Table VI. Data augmentation using rotation by small angles, up to 15 degrees,



(a) (b)

Fig. 2: Exemplar images of correct classification of threats by SLX. Here, (b) is misclassified by the model SimCLR [18].

(a) (b)

Fig. 3: Images of incorrect classification of threats by SLX and by the model SimCLR [18].

is beneficial. L_2 has small contribution, but consistent positive effect. Model pretraining on ImageNet [22] is beneficial.

Images of incorrect classifications. We examine the correct and wrong classification of threats by SLX. In Figs. 2 and 3, we show the images of correct classifications and misclassifications of threats, respectively. These images have *parts* of guns. Fig. 2(b) was misclassified by the model SimCLR, expanding on our discussion in Section IV-F. The images in Fig. 3, which were also misclassified by SimCLR, have *smaller* parts of guns. For the larger parts of the gun, SLX correctly detects the threat. Figs. 3(a) and (b), which are false negative (Type II) errors, have clutter and occlusion. They do not have the trigger part of the gun, the firing pin (detonation mechanism), the bore/ barrel, the sear, and the hammer part of the gun. Because the spring part of the self-loading pistol in Fig. 3(b), which is also bent and has a tilt, resembles the zipper of coats, dresses, and trousers, an operator may also make the same error [28], [22].

V. Conclusion

We have proposed SLX for security screening and accurate and robust disassembled object detection in cluttered scenes. Existing methods underperform to recognise prohibited items that are disassembled, especially when learning from limited samples and from data originating from multiple domains, i.e. images produced by different scanners. SLX addresses such challenges, and we have trained and evaluated our model on the XIL dataset. Here, SLX detects and recognises threat objects, including components of firearms, achieves good generalisation performance, effectively reducing overfitting, for the problem settings of training with limited image samples and of using multiple scanners with *different* multi-view orientations. The evaluation of SLX on XIL and its ablation study show that SLX is effective and beneficial for detecting threats. SLX achieves an *improvement*, on average, of approximately 12 points in accuracy (in Table V), upon other baseline models for same-scanner generalisation. Finally, we believe that SLX will open the road to disassembled object detection in the X-ray security screening setting, as well as in other *correlated* X-ray screening settings including X-ray waste inspection and classification [35], which are promising nascent research fields, and inspire other researchers to adopt this real-world problem setting.

References

[1] D. Vukadinovic and D. Anderson, "X-ray baggage screening and Artificial Intelligence (AI): A technical review of machine learning techniques for X-ray baggage screening," Joint Research Centre (JRC) Science, 2022.

[2] N. Dionelis, et al., "Few-Shot Adaptive Detection of Items of Concern Using Generative Models with Negative Retraining," in Proc. International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2021.

[3] N. Dionelis, et al., "Tail of Distribution GAN: GAN-Based Boundary of Distribution Formation," in Proc. Sensor Signal Processing for Defence (SSPD), IEEE, 2020.

[4] M. Arjovsky, et al., "Invariant Risk Minimization," arXiv:1907.02893.

[5] K. Xiao, et al., "Noise or signal: The role of image backgrounds in object recognition," in Proc. International Conference on Learning Representations (ICLR), 2021.

[6] N. Dionelis, S. A. Tsaftaris, and M. Yaghoobi, "FROB: Few-shot ROBust Model for Joint Classification and Out-of-Distribution Detection," in Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2022.

[7] N. Dionelis, S. A. Tsaftaris, and M. Yaghoobi, "CTR: Contrastive Training Recognition Classifier for Few-Shot Open-World Recognition," in Proc. International Conference on Pattern Recognition (ICPR), 2022.

[8] N. Dionelis, S. A. Tsaftaris, and M. Yaghoobi, "OMASGAN: Out-of-Distribution Minimum Anomaly Score GAN for Anomaly Detection," in Proc. Sensor Signal Processing for Defence (SSPD), 2022.

[9] N. Dionelis, M. Yaghoobi, and S. A. Tsaftaris, "Boundary of Distribution Support Generator (BDSG): Sample Generation on the Boundary," in Proc. International Conference Image Processing (ICIP), 803-807, 2020.

[10] I. Goodfellow, et al., "Generative Adversarial Nets," in Proc. Advances Neural Information Processing Systems (NeurIPS), p. 2672–2680, 2014.

[11] K. He, et al., "Deep Residual Learning for Image Recognition," in Proc. Conference Computer Vision and Pattern Recognition (CVPR), 2016.

[12] C. Miao, et al., "SIXray: A Large-scale Security Inspection X-ray Benchmark for Prohibited Item Discovery in Overlapping Images," in Proc. Conference Computer Vision Pattern Recognition (CVPR), 2019.

[13] D. Mery, et al., "Object Recognition in X-ray Testing Using Adaptive Sparse Representations," Journal Nondestructive Evaluation, 35(3), 2016.

[14] T. Hassan, et al., "A Novel Incremental Learning Driven Instance Segmentation Framework to Recognize Highly Cluttered Instances of the Contraband Items," IEEE Trans Systems, Man and Cybernetics, 2020.

[15] S. Akçay and T. Breckon, "An Evaluation of Region Based Object Detection Strategies Within X-Ray Baggage Security Imagery," in Proc. IEEE International Conference on Image Processing (ICIP), 2017.

[16] T. Hassan, et al., "Trainable Structure Tensors for Autonomous Baggage Threat Detection Under Extreme Occlusion," in Proc. Asian Conference Computer Vision (ACCV), Lecture Notes in Computer Science, 2020.

[17] Z. Han, et al., "Trusted multi-view classification," in Proc. ICLR, 2021.

[18] T. Chen, et al., "A simple framework for contrastive learning of visual representations," in Proc. Int Conf Machine Learning (ICML), 2020.

[19] M. Wani, et al., "Basics of Supervised Deep Learning," Advances in Deep Learning, Springer, p.13-29, 2020. DOI: 10.1007/978-981-13-6794-6_2.

[20] S. Akçay, et al., "GANomaly: Semi-supervised anomaly detection via adversarial training," in Proc. Asian Conf Comp Vision (ACCV), 2018.

[21] S. Akçay, et al., "Skip-GANomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in Proc. International Joint Conference Neural Networks (IJCNN), 2019.

[22] S. Akçay, et al., "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," IEEE Trans Information Forensics Security, 13(9), 2018.

[23] J. van der Putten and F. Zanjani, "Multi-scale Ensemble of ResNet Variants," Computer-Aided Analysis of Gastrointestinal Videos, 2021.

[24] I. Tyukin, et al., "Demystification of Few-Shot and One-Shot Learning," in Proc. International Joint Conference Neural Networks (IJCNN), 2021.

[25] B. Hariharan and R. Girshick, "Low-shot visual recognition by shrinking and hallucinating features," in Int Conf Computer Vision (ICCV), 2017.
[26] M. Hein, "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem," In CVPR.
[27] J. Tack, et al., "CSI: Novelty detection via contrastive learning on distributionally shifted instances," in Proc. NeurIPS, 2020.
[28] S. Akçay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within X-ray security imaging," Pattern Recognition, 122:108245, 2022.
[29] X. Chen and K. He, "Exploring simple Siamese representation learning," in Proc. CVPR, 2021.
[30] C. Yeh, et al., "Decoupled contrastive learning," arXiv:2110.06848, 2021.
[31] S. Akçay, "Recent Advances of Deep Learning within X-ray Security Imaging," Signal Processing Society (SPS), Webinar, 2022.
[32] T. Dietterich. Anomaly Detection for OoD and Novel Category Detection. Amazon's Annual Machine Learning Conference (AMLC). 2021.
[33] H. Zhang, A. Li, J. Guo, and Y. Guo. Hybrid Models for Open Set Recognition. In Proc. European Conf Computer Vision (ECCV). 2020.
[34] Y. Wei, et al. Occluded prohibited items detection: An X-ray security inspection benchmark and de-occlusion attention module. In Proc. 28th ACM International Conference on Multimedia, p. 138-146. 2020.
[35] L. Qiu, Z. Xiong, X. Wang, K. Liu, et al. ETHSeg: An Amodel Instance Segmentation Network and a Real-world Dataset for X-Ray Waste Inspection. In Proc. IEEE CVF Conf CVPR, p. 2283-2292. 2022.
[36] A. Kortylewski, J. He, Q. Liu, and A. Yuille. Compositional Convolutional Neural Networks: A deep architecture with innate robustness to partial occlusion. In IEEE CVF Conf CVPR, p. 8940-8949. 2020.
[37] T. Hassan, S. Akçay, M. Bennamoun, et al. Tensor Pooling Driven Instance Segmentation Framework for Baggage Threat Recognition. Journal Neural Computing and Applications, 1239-1250 (34). 2022.