RWSC-Fusion: Region-Wise Style-Controlled Fusion Network for the Prohibited X-ray Security Image Synthesis

Luwen Duan[1, 2] Min Wu[1] Lijian Mao[1] Jun Yin[1] Jianping Xiong[1*] Xi Li[2*]
[1]Zhejiang Dahua Technology Co., Ltd., [2]Zhejiang University

Abstract

Automatic prohibited item detection in security inspection X-ray images is necessary for transportation. The abundance and diversity of the X-ray security images with prohibited item, termed as prohibited X-ray security images, are essential for training the detection model. In order to solve the data insufficiency, we propose a Region-Wise Style-Controlled Fusion (RWSC-Fusion) network, which superimposes the prohibited items onto the normal X-ray security images, to synthesize the prohibited X-ray security images. The proposed RWSC-Fusion innovates both network structure and loss functions to generate more realistic X-ray security images. Specifically, a RWSC-Fusion module is designed to enable the region-wise fusion by controlling the appearance of the overlapping region with novel modulation parameters. In addition, an Edge-Attention (EA) module is proposed to effectively improve the sharpness of the synthetic images. As for the unsupervised loss function, we propose the Luminance loss in Logarithmic form (LL) and Correlation loss of Saturation Difference (CSD), to optimize the fused X-ray security images in terms of luminance and saturation. We evaluate the authenticity and the training effect of the synthetic X-ray security images on private and public SIXray dataset. The results confirm that our synthetic images are reliable enough to augment the prohibited X-ray security images.

1. Introduction

X-ray imagery security inspection is a fundamental part in station/airport, for detecting prohibited items in baggage or suitcase images.

Recently, computer vision methods particularly deep learning [15, 33] have brought benefits to prohibited item detection [1, 16, 18, 25, 37]. The performance of these automated detection models relies heavily on a mass of annotated images. However, real X-ray security images usually are arduous and time-consuming to collect, and the occurrence rate of prohibited items is very low. Moreover, very few of public X-ray security image datasets contain large amounts of prohibited items. Existing datasets are mainly: (1) GDXray: The GDXray dataset [24] contains only three kinds of prohibited items: guns, shurikens, and razor blades. Besides, GDXray only involves grayscale images where backgrounds are too simple to conform with the real color X-ray security images; (2) OPIXray: The OPIXray dataset [36], which is especially designed for the cutter detection, contains 8885 synthetic X-ray security images with five kinds of cutters, and most images contain only one cutter; (3) SIXray: The SIXray dataset [26] contains 1, 059, 231 X-ray security images, but only 8929 images include prohibited items: guns, knives, wrenches, pliers, scissors, and hammers. Given the above, existing public datasets haven't been up to the standard of training.

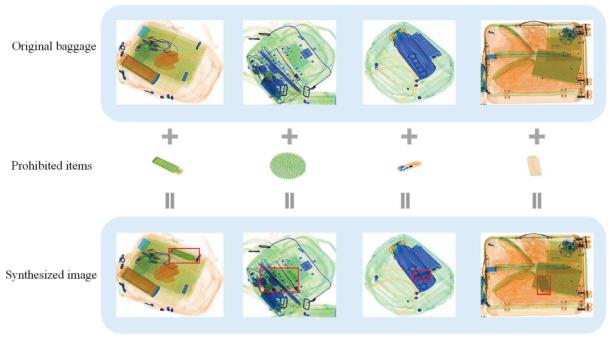

Figure 1. Our RWSC-Fusion allows to superimpose prohibited items onto baggage images, to synthesize prohibited X-ray images. The composite prohibited items are marked with red box.

To overcome the lack of training samples, traditional offline enhancement strategies such as rotation, re-scaling and mixing are applied to the augmentation of training samples [32, 42]. However, unlike natural images and other X-ray scans, X-ray security images usually involve randomly stacked objects [6, 11, 12, 22]. In addition, according to the imaging principle of X-ray security, objects overlap with each other in a translucent state, and appear differently for different materials and thickness. These traditional augmentation methods cannot improve the diversity and complexity for the inter-occlusion between prohibited items [31]. Therefore, it is very necessary to synthesize realistic X-ray security images, so as to enrich the prohibited items in pose, scale and position.

A few researchers studied to directly generate prohibited

*Corresponding author: xilizju@zju.edu.cn, xiongjp362204@163.com

X-ray security images by learning deep synthesis models. Inspired by the success of GANs in image synthesis [10, 34], Zhao *et al*. proposed X-ray Image- Synthesis-GAN [41] to generate the prohibited items from a noise region on the background images. Li *et al*. [17] synthesized the X-ray security images from the semantic label maps based on GANs. Yang *et al*. [40] enhanced the training of GANs to learn higher-quality X-ray security images. However, they synthesized only one prohibited item rather than stacked and overlapped ones in one image, which thus are not realistic enough for the complex X-ray security images in a real-world scenario. Isaac-Medina *et al*. [14] used the paired X-ray energy maps (high, low, effective-Z maps) to synthesize the pseudo-color images which however only served as a small amount of testing samples. Bhowmik *et al*. [3] developed a Synthetically Composited (SC) data augmentation strategy based on the Threat Image Projection (TIP) method [8], to fuse the prohibited items with the baggage images for generating images with stacked and cluttered prohibited items. However, the SC need adjust the parameters for each image to meet the various colors of different materials, and thus lacks automation, robustness and versatility.

In order to synthesize the prohibited X-ray security images automatically, we propose a color X-ray security images fusion model, to superimpose the prohibited items onto baggage or suitcase images, as shown in Figure 1. In this way, we synthesize prohibited X-ray security images and obtain annotation automatically, thus avoiding the collection of the annotated prohibited X-ray security images for training the prohibited item detection model.

The experimental results prove the advantage of our fusion model over other fusion methods in the field of X-ray security image. In addition, we also compare the prohibited item detection model trained with real and synthetic images. The results verify that our synthetic images are efficient to supplement the prohibited X-ray security images in downstream detection task.

The main contributions of our work are as follows:

1. We propose an unsupervised color X-ray security image fusion model. Due to the imaging particularity, existing fusion loss functions are inapplicable to X-ray security images. We design Luminance loss in Logarithmic form (LL) and Correlation loss of Saturation Difference (CSD) based on the principle of X-ray imaging and Threat image projection (TIP). The LL and CSD optimize the comprehensive luminance-saturation fusion between the foreground item and background image. Thus, we extend TIP to composite color X-ray security images.

2. We propose a Region-Wise Style-Controlled Fusion module, to control the fused appearance by learning the shifting and scaling modulation parameters pertinently. Moreover, the RWSC-Fusion module can adaptively modulate the local region of interest, i.e., the overlapping region of the prohibited items and baggage images.

3. We develop an Edge-Attention module, which inhibits irrelevant information and enhances texture information, to improve the sharpness of generated images.

4. Our RWSC-Fusion and Edge-Attention modules are both universal and plug-and-play to other image fusion or synthesis models for generating high-quality images.

2. Related work

2.1. Threat Image Projection

Plenty of prohibited X-ray security images are the precondition for automatic prohibited item detection models [27, 35]. However, prohibited items are a very small minority on X-ray security images in a real-world scenario. Therefore, threat image projection (TIP) method has been proposed to increase the occurrence of prohibited items [29]. TIP is a process that superimposes isolated prohibited items onto normal images, thus generating composite, yet realistic prohibited X-ray security images.

The X-ray penetrates through objects to form X-ray image. The image intensity is related to the X-ray energy, object material, object thickness [21], simply expressed as:

$$I = I_0 e^{-\mu h} \tag{1}$$

where I_0 is the X-ray beam intensity, μ is the absorption coefficient and h is the thickness of the objects.

Accordingly, the prohibited item X-ray image I_f, and background image I_b, *e.g*. a baggage, can be expressed as:

$$I_f = I_0 e^{-\mu_f h_f}, \quad I_b = I_0 e^{-\mu_b h_b} \tag{2}$$

where μ_f and μ_b are the respective absorption coefficients, h_f and h_b are the respective thickness.

When superimposing the prohibited items I_f onto the baggage image I_b, the fused X-ray image I_{fb} could be:

$$I_{fb} = I_0 e^{-\mu_f h_f - \mu_b h_b} = I_0 e^{-\mu_f h_f} e^{-\mu_b h_b} \tag{3}$$

Thus, we can get:

$$I_{fb} = \frac{I_f \cdot I_b}{I_0} \tag{4}$$

Therefore, the prohibited item can be projected into baggage images through multiplication, which is just the principle of TIP. Rogers *et al*. [30] applied TIP and exploit the approximate multiplicative property of X-ray imagery, to project threat items onto cargo transmission images. Mery *et al*. [23] replaced the multiplication with the addition of logarithmic, giving access to linear strategy for superimposing the prohibited items onto images. The TIP is almost exclusively applicable to grayscale X- ray images. However, X-ray security images in a real-world scenario are mostly rendered with pseudo-color. In this work, we will extend TIP principle to color X-ray images, to develop a prohibited X-ray security images synthesis model.

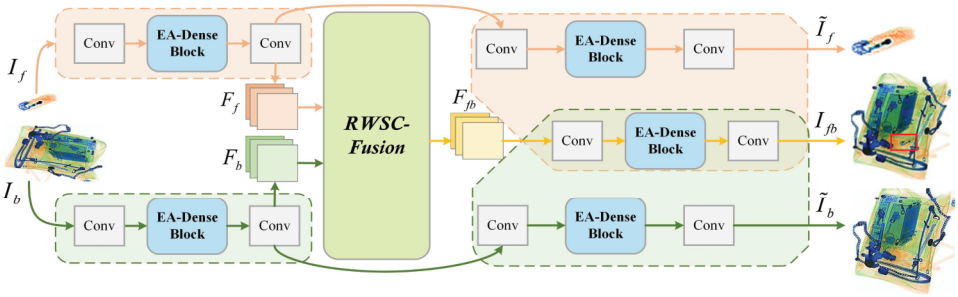

Figure 2. Overview of the proposed RWSC-Fusion.

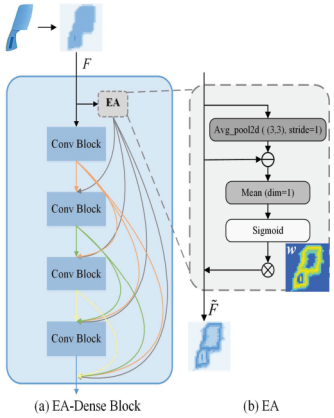

Figure 3. EA-Dense Block with Edge-Attention module.

2.2. Color X-ray Security Images Rendering

As mentioned above, real X-ray security images are rendered with pseudo-colors according to certain rules [5]. The most common rule is based on the three components: Hue (H), Saturation (S) and Value (V) in the HSV color space, which approximates the way human perceives and interprets color. First, the H component of pseudo-color is assigned according to the material category of the object, which depends on the dual energy responses. e.g., organic materials are usually shown in red, inorganic materials are in blue, and mixtures are in green. Next, the V component of the three is a normalized nonlinear mapping, while the S component is mostly correlated linearly with the single energy or dual energy responses. As a consequence, the colors of objects in X-ray security image (see examples in the supplementary material), vary differently according to the materials, thickness, viewpoint and background.

3. Related work

The overview of RWSC-Fusion is shown in Figure 2. The prohibited item image (foreground: I_f) and baggage image (background: I_b) are inputted separately into two branches, and are reconstructed through convolution layers and Edge-Attention Dense (EA-Dense) blocks to be $\tilde{I}_f$ and $\tilde{I}_b$ respectively. They share weights in parallel to encode respective features F_f and F_b. The two encoding features are inputted together into a RWSC-Fusion module, which fuses the two in a region-wise style-controlled manner, and then are reconstructed to be the composite image I_{fb}.

3.1. Edge-Attention Dense Block

The Edge-Attention Dense (EA-Dense) block in our RWSC-Fusion is shown in Figure 3(a). The dense block is composed of four convolutional blocks, each taking the feature maps of all preceding blocks as input. To reinforce the information representation, we design and add an Edge-Attention (EA) module (shown in Figure 3(b)) into the top convolutional block before flowing into subsequent blocks.

Given the low-level feature F of the top block, we first employ 3×3 average pooling layer with the stride of 1 to obtain the local mean feature. We subtract the local mean feature maps from the original feature maps to explore the structural texture, similar to Laplacian sharpening which is robust to noise for edge information extraction. Next, we calculate the mean responses along the channel dimension to reduce the channel number to 1, and obtain the edge-aware activation map w by using the sigmoid function. At last, the edge-enhanced feature $\tilde{F}$ are acquired through the multiplication between the weight w and feature F. The whole process of EA module can be expressed as:

$$w = \sigma(M(|AP2D(F) - F|)) \tag{5}$$

$$\tilde{F} = w \odot F \tag{6}$$

where $AP2D$ denotes 2D average pooling, M denotes the mean value along the channel dimension, σ denotes the sigmoid activation, $\odot$ denotes pixel-wise multiplication.

Since the weight w can activate and enhance the edge and texture information through sharpening operation, the resulting feature $\tilde{F}$ take advantage of both global and local texture information of the feature F. Therefore, in our EA-Dense block, each convolutional block takes as input the feature of all preceding block and the edge-enhanced feature $\tilde{F}$, and thus sharpen the image quality.

3.2. Region-Wise Style-Controlled Fusion Module

Common image fusion methods usually fuse the features of two source images in a simple manner, such as addition,

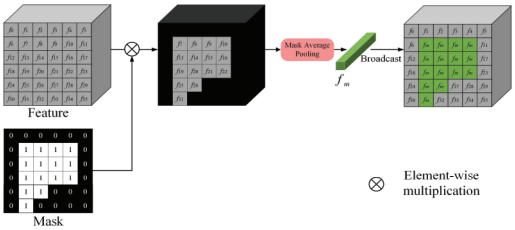

Figure 4. Details of mask average pooling layer.

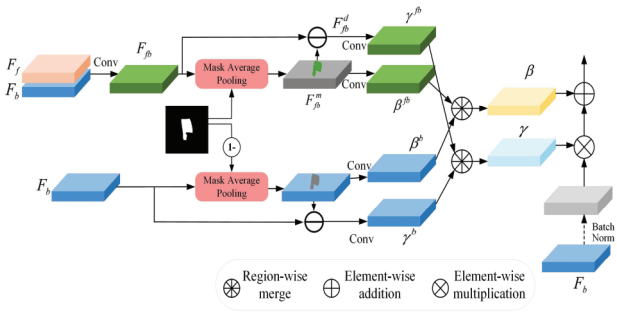

Figure 5. Details of RWSC-Fusion module.

maximum or concatenation. However, for the fusion of the prohibited item and the baggage image, we focus more on the appearance fusion of the overlapping region rather than the global information fusion. Therefore, we propose a RWSC-Fusion module to fuse the features of prohibited item F_f and the baggage image F_b in a region-wise style-controlled manner. The RWSC-Fusion module modulates the appearance of the overlapping region between the two images by applying region-wise normalization.

The traditional region-wise normalization [28] is:

$$\gamma_{c,h,w}\left(\frac{b_{n,c,h,w}-\mu_c(x)}{\sigma_c(x)}\right)+\beta_{c,h,w} \tag{7}$$

where $\gamma_{c,h,w}$ and $\beta_{c,h,w}$ represent pixel-wise scaling and shifting parameters. Some researchers [7] learn $\gamma_{c,h,w}$ and $\beta_{c,h,w}$ from the style maps by using mask average pooling, which is shown in Figure 4. However, such parameters have less potential for realizing rich and fine style-control.

By contrast, the region-wise normalization in RWSC-Fusion module innovate in two respects: (1) It learns the shifting and scaling parameters separately based on the mask average pooling and the deviation maps to modulate the mean and standard deviation; (2) It normalizes the local region adaptively with pixel-wise modulation parameters, to control the appearance of the region of interest.

The RWSC-Fusion module is detailed in Figure 5. On one branch, the features F_f and F_b of two images are first sent jointly to one convolution layer. Then, the resulting features F_{fb} with auxiliary mask of prohibited item are fed into the mask average pooling layer, where the average feature vectors are extracted and broadcast to create the mean maps F_{fb}^{m}. Next, we subtract the mean maps F_{fb}^{m} from the initial features F_{fb} to get the local standard deviation maps F_{fb}^{d}. Lastly, the deviation maps F_{fb}^{d} and the mean maps F_{fb}^{m} pass through two separate convolution layers to generate the pixel-wise scaling and shifting parameters $\gamma_{c,h,w}^{fb}$ and $\beta_{c,h,w}^{fb}$. Meanwhile, on the other branch, we obtain parameters $\gamma_{c,h,w}$ and $\beta_{c,h,w}$ by doing the same operations on the feature F_b of the single baggage image.

We design the region-specific modulation parameters through the region-wise mergence, and thus the final modulation parameters are expressed as:

$$\gamma_{c,h,w}=M^{+}\odot\gamma_{c,h,w}^{fb}+(1-M^{+})\odot\gamma_{c,h,w}^{b} \tag{8}$$

$$\beta_{c,h,w}=M^{+}\odot\beta_{c,h,w}^{fb}+(1-M^{+})\odot\beta_{c,h,w}^{b} \tag{9}$$

where M^{+} denotes the mask of prohibited item, while 1-M^{+} denotes the complementary mask. The appearance of the overlapping region within the mask is modulated jointly by the prohibited item and baggage image, while the non-overlapping region is modulated by the single baggage image, avoiding the interference of prohibited item image.

Our region-adaptive normalization learns appropriate shifting and scaling parameters separately based on the mean maps and the deviation maps instead of single mask mean maps, so as to modulate the mean and standard deviation specifically and pointedly. Thus, it allows for fine control to generate rich stylization by considering the local information from the deviation maps. Hence, the RWSC-Fusion module enables the region-wise normal-appearing fusion of the prohibited item and baggage image to synthesize the prohibited X-ray security image.

3.3. Loss Function

Existing unsupervised image fusion models usually maximize the similarity of fused image with source images instead of ground-truth, by using loss function as follows:

$$\mathcal{L}=1-\left[w_1\bullet\ell(x_1,y)+w_2\bullet\ell(x_2,y)\right] \tag{10}$$

where $\ell(x,y)$ denotes the similarity metrics, such as SSD or SSIM. The weights w_1 and w_2 represent the information preservation degree of two source images. However, these functions are inapplicable to X-ray security image, for two reasons: (1) they produce global fusion, including the non-overlapping region, which instead should remain intact; (2) the addition strategy of these functions does not conform to the attenuation character of X-ray security imagery.

As mentioned in Sec.2.1, when fusing the prohibited items I_f with the baggage image I_b by using the weights w_1 and w_2, the fused X-ray image I_{fb} could be:

$$I_{fb}=I_0e^{-w_1\mu_f h_f-w_2\mu_b h_b}=I_0e^{-w_1\mu_f h_f}e^{-w_2\mu_b h_b} \tag{11}$$

$$I_{fb}=\frac{I_f^{w_1}\cdot I_b^{w_2}}{I_0^{w_1+w_2-1}} \tag{12}$$

Table 1. Quantitative evaluations for image quality of the models with and without the EA modules.

Model	without EA module	with EA module
MAE	2.091	**1.938**
PSNR	38.402	**39.084**

Therefore, the fused X-ray image I_{fb} is approximatively correlated with the multiplication rather than addition of the prohibited item I_f and the baggage image I_b.

In this case, we adjust Eq. (10) by using the logarithmic form, giving access to multiplication relationship in linear form. Finally, we develop a Luminance loss in Logarithmic form (LL) $\mathcal{L}_{LL}$ as follows:

$$\mathcal{L}_{LL} = 1 - \left[w_1 \bullet \log \ell(I_f, I_{fb}) + w_2 \bullet \log \ell(I_b, I_{fb}) \right] \quad (13)$$

We put w_1 and w_2 inside the metric formula to develop a new similarity metric $\ell(x,y,w)$, and thus $\mathcal{L}_{LL}$ turns into:

$$\mathcal{L}_{LL} = 1 - \left[\log \ell(I_f, I_{fb}, w_1) + \log \ell(I_b, I_{fb}, w_2) \right] \quad (14)$$

where $\ell(x,y,w) = \dfrac{2\mu_x\mu_y + \varepsilon}{\mu_x^{4w} + \mu_y^{2} + \varepsilon}$, and thus the local mean of fused image μ_{fb} is expected to converge to $\mu_f^{w_1} \cdot \mu_b^{w_2}$. The details and deduce are in the supplementary materials.

We automatically estimate the weights w_1 and w_2 from the feature of R, G and B component of the prohibited item and the baggage image respectively, to get different power weights w_1 and w_2 for R, G and B component, and the sum of w_1 and w_2 is adjusted to 2.

On the other hand, considering that the saturation (S) component of color X-ray security image is mostly correlated linearly with the energy responses, we assume that the S component of the fused image is correlated linearly with the multiplication of the S component of the two images. Therefore, the difference between the S component of the prohibited item (S_f) and the fused image (S_{fb}) comes almost from the S component of the baggage image (S_b). Similarly, the difference between the S_b and S_{fb} comes almost from S_f. We develop an another loss term Correlation loss of Saturation Difference (CSD) $\mathcal{L}_{CSD}$:

$$D_b = \left(1 - S_{fb}\right) / \left(1 - S_f\right)$$
$$D_f = \left(1 - S_{fb}\right) / \left(1 - S_b\right) \quad (15)$$
$$\mathcal{L}_{CSD} = 1 - \left[CC(D_f, 1 - S_f) + CC(D_b, 1 - S_b) \right] / 2 \quad (16)$$

where CC is the normalized correlation coefficient.

In addition, a third loss term is the reconstruction loss $\mathcal{L}_{recon}$ of two source images $\tilde{I}_f$ and $\tilde{I}_b$ as follows:

$$\mathcal{L}_{recon} = \left\| I_f, \tilde{I}_f \right\| + \left\| I_b, \tilde{I}_b \right\| \quad (17)$$

In conclusion, the total loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_{LL} + \mathcal{L}_{CSD} + \mathcal{L}_{recon} \quad (18)$$

4. Experiment Results

4.1. Datasets

Private data: We have collected millions of X-ray security images from urban rail transit, railway, highway and other real-world daily security inspection scenario. There are more than 40 kinds of manually-annotated prohibited items. **Public data:** The SIXray dataset [26] and OPIXray dataset [36], which have been introduced in Sec.1.

4.2. Implementation Details

The training data are 12000 groups randomly composed by 12000 baggage images and 60 individual prohibited items of 15 categories. During training, all baggage images are resized to a resolution of 256 × 256, and all prohibited item images are randomly padded with [255,255,255] to share the same size. It is important to note that the random padding of the prohibited item images is regulated by utilizing the activation maps of the baggage images, in order to enable realistic and appropriate location of the prohibited item in the internal region of the baggage.

4.3. Ablation Study

4.3.1. Edge-Attention module

Due to the sparse texture of X-ray security images and the information loss in deep neural network, the generated images often suffer from low sharpness. We design EA module in the dense blocks to extract and recover the texture information for the fused image. In this section, we evaluate the efficacy of the proposed EA module.

We quantitatively calculate the reconstruction loss of the non-overlapping regions between 6000 pairs of the fused images and the source images in terms of the Mean Absolute Error (MAE) and Peak Signal to-Noise Ratio (PSNR). The lower the MAE and the higher the PSNR, the closer the non-overlapping regions is to the source image and the higher the quality of fused images. The results are shown in Table 1. The model with EA modules produces lower MAE and higher PSNR than that without the EA modules. It indicates that the EA modules can effectively preserve the source information more completely in the resulting non-overlapping regions and improve the definition and sharpness of the synthesized images.

We visualize the synthesized images from the models with and without the EA modules in Figure 6. Comparing the regions within the purple ROI which are magnified into view at the top-right corner, the X-ray security images generated by the model without EA modules are blurred and miss details, such as the zipper in the second row and the shoe sole texture in the fourth row. On the contrast, the model with EA modules can activate and enhance texture and edge information by using the attention map shown in Figure 7. Thus it can alleviate the shortage of texture and

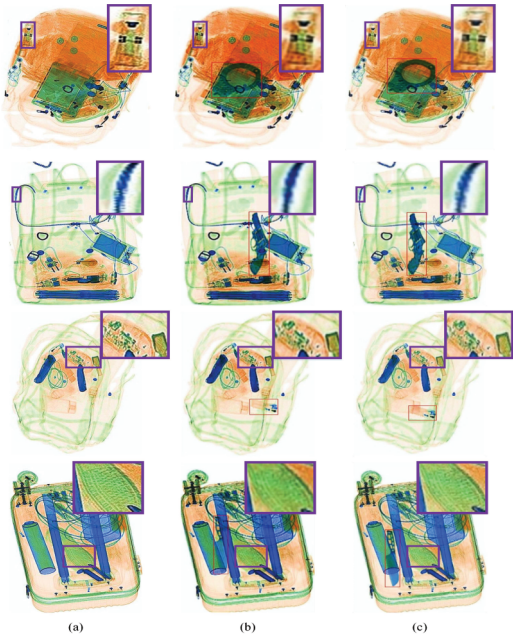

Figure 6. The generated images from the models with and without the EA modules. From left to right are: (a) original baggage image; (b) the synthesized images from the model without EA modules; (c) the synthesized images from the model with EA modules.

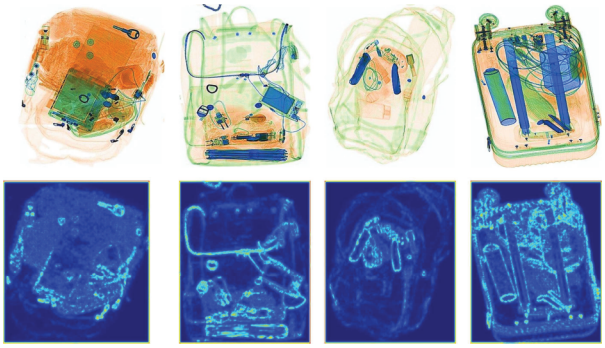

Figure 7. The edge attention maps of the X-ray security images in Figure 6. The first row shows the original images, and the second row shows the edge attention maps extracted by the EA module.

offer clearer appearance. It further follows that the EA module can indeed provide high-quality images with sharp textures via the Edge Attention mechanism.

4.3.2. Region-wise normalization

Existing region-adaptive normalization methods learn the shifting and scaling parameters both based on the mask mean maps [7, 28]. On the contrast, we learn the shifting and scaling parameters separately based on the mask mean maps and deviation maps, so as to modulate the mean and standard deviation pointedly. In this section, we evaluate the efficacy of our normalization mechanism.

We visualize the generated X-ray security images from the fusion module with only mask average pooling layer

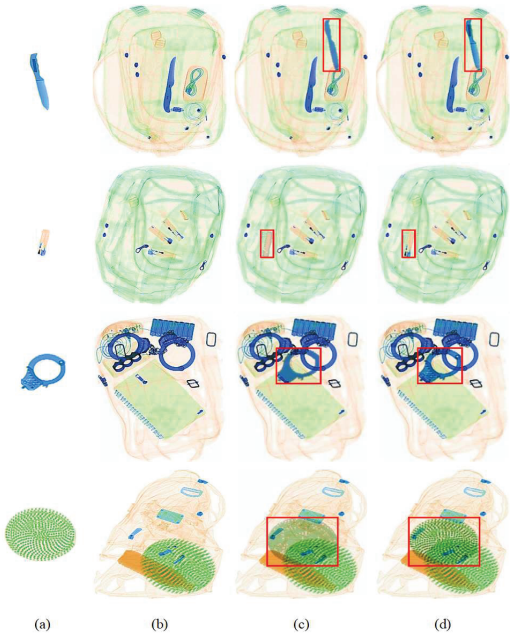

Figure 8. The generated images from the fusion module with only mask average pooling layer and our RWSC-Fusion module. From left to right are: (a) prohibited items; (b) original baggage image; (c) the synthesized images from the model with only mask average pooling layer; (d) the synthesized images from our model.

and our RWSC-Fusion module in Figure 8. By comparing the prohibited items within the red ROI, the prohibited items synthesized by the model with only mask average pooling layer are fuzzy due to a lack of texture details, while the prohibited items from our model are more clear. It proves that our region-adaptive normalization method allows for fine control to provide rich stylization.

4.4. Comparison with State-of-the-art Methods

In this section, we compare our RWSC-Fusion with four state-of-the-art image fusion methods: U2Fusion [38], FusionGAN [19], DeepFuse [13] and MEFNet [20], to evaluate the performance of X-ray security image fusion for the prohibited X-ray security image synthesis.

The synthesized images from different methods are shown in Figure 9. It can be seen that compared with U2Fusion, FusionGAN and DeepFuse, RWSC-Fusion alleviates the problem of global fusion and successfully preserves the original appearance of the non-overlapping regions in the baggage images. The overlapping regions marked by the red box from the competitors, especially MEFNet, noticeably deviate from our desired appearance that can match the real X-ray security images in terms of hue, saturation and sharpness. Overall, our model achieves superior visual performance than these four competitors. These well demonstrate that our fusion model is better suited to the X-ray security image, and can synthesize more ideal and realistic prohibited X-ray security images.

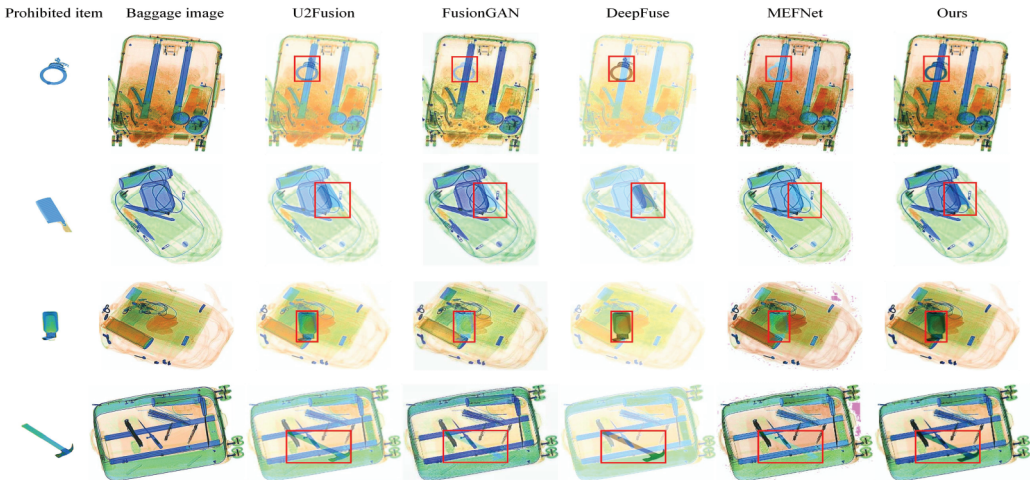

Figure 9. The X-ray baggage images fused with the prohibited item by four other fusion methods and the proposed RWSC-Fusion.

Table 2. Quantitative comparison with state-of-the-art methods.

Metric	QE	Qabf	EI	PSNR
U2Fusion [38]	0.5635	0.3467	48.8190	16.7155
FusionGAN [19]	0.5640	0.5924	78.0934	13.8471
DeepFuse [13]	0.6157	0.1814	41.6266	16.8173
MEFNet [20]	0.7509	0.6090	**85.0403**	17.7364
Ours	**0.8339**	**0.6472**	81.3807	**18.6455**

The quantitative comparisons are performed on the 6000 image pairs. Four metrics, namely, QE [9], Qabf [39], Edge Intensity (EI) [2], and PSNR are used for evaluation. For all four metrics, larger values indicate better performance. The comparative results of four metrics are shown in Table 2. Our RWSC-Fusion ranks first on QE, Qabf and PSNR, proving that our method can synthesize images with richer textures and details. Although our RWSC-Fusion ranks second on EI, it also achieves comparable results. The promising outcomes show that our model keeps high fidelity with less distortion and noise.

4.5. Supplement to Private Dataset

In this section, we synthesize the prohibited X-ray security images on our private dataset.

First, we conduct subjective evaluation experiment, which is detailed in the supplementary material. This result demonstrates that our synthetic prohibited X-ray security images are realistic enough to fool the inspectors.

To analyze the efficacy of supplementing the training samples, we compare the performance of prohibited item detection model under the training of real and synthetic images. We use the YOLOv4 as the detector [4] to detect the five kinds of prohibited items: plastic bottles, cutters, scissors, thermos cups and glass bottles.

The real training dataset contains 40000 samples (DataR: with 26048 plastic bottles, 34856 cutters, 10391 scissors, 4968 thermos cups and 17526 glass bottles). Besides, we train the detector with the following dataset:

(1) DataR1/2: We remove 20000 samples from DataR, resulting in 20000 training samples with 13127 plastic bottles, 17055 cutters, 5117 scissors, 2518 thermos cups and 8710 glass bottles.

(2) DataRS: We synthesize and supplement the prohibited items into DataR1/2, resulting in 20000 samples with 26048 plastic bottles, 34856 cutters, 10391 scissors, 4968 thermos cups and 17526 glass bottles, with the same number of prohibited items as the DataR. (Note: since there are no extra 20000 negative samples excluding all five kinds of prohibited items, we cannot achieve the pure synthesis of 20000 samples, to be perfectly aligned with DataR)

(3) DataRS+: We synthesize prohibited items into other 20000 samples, and merge new synthetic samples with 20000 samples in DataR, resulting in 40000 samples with total 38983 plastic bottles, 51956 cutters, 15597 scissors, 7522 thermos cups and 26034 glass bottles.

The detection results on the 9954 testing samples of YOLOv4 trained with above datasets are listed in Table 3.

As compared with DataR1/2, the DataRS provides a considerable increase of more than 3% in mAP after we synthesize prohibited items in DataR1/2. This proves that our synthetic prohibited items are indeed beneficial to training the detection model. Moreover, since the results of DataRS can be comparable with or even better than those of DataR, it demonstrates the authenticity and reliability of our synthetic prohibited items for replacing the real ones. Accordingly, with the addition and the contribution of synthetic prohibited items, DataRS+ also has a remarkable advantage over DataR.

4.6. Supplement to SIXray and OPIXray Dataset

In this section, we evaluate on public SIXray [26] and OPIXray dataset [36]. We fuse our private prohibited items with the negative samples in SIXray and OPIXray dataset. Moreover, we directly apply the model trained by our private dataset without any additional parameter tuning.

Table 3 Detection results of YOLOv4 trained with different data on private dataset (recall and precision at confidence level of 0.5).

Training Data		DataR	DataR1/2	DataRS	DataRS+
Plastic bottle	Recall	69.49	64.82	70.46	74.15
	Precision	85.36	83.29	88.14	85.13
	AP	69.50	71.16	71.00	76.97
Cutter	Recall	86.17	74.04	81.70	84.50
	Precision	92.62	87.60	91.81	93.45
	AP	86.96	79.98	83.49	87.70
Scissor	Recall	71.22	58.03	70.87	72.97
	Precision	90.60	89.34	88.37	88.23
	AP	74.28	66.81	72.83	78.29
Therm os cup	Recall	88.86	88.86	91.59	92.76
	Precision	93.52	90.13	91.38	92.54
	AP	85.52	86.17	88.98	90.25
Glass bottle	Recall	56.49	48.74	58.71	58.82
	Precision	86.79	80.97	84.02	88.15
	AP	56.31	53.76	58.13	61.87
mAP		74.52	71.58	74.88	79.02

Table 4 Detection results of YOLOv4 trained with different data on SIXray dataset.

Training Data		SIXR	SIXR1/2	SIXRS	SIXRS+
Gun	Recall	79.73	73.02	83.23	83.69
	Precision	96.32	91.76	97.33	97.17
	AP	78.71	75.21	82.19	82.74
Knife	Recall	58.88	45.62	64.17	66.04
	Precision	92.65	86.39	92.79	93.39
	AP	62.84	47.88	66.64	69.17
Wrench	Recall	55.97	24.25	59.33	66.04
	Precision	87.72	60.75	83.68	83.49
	AP	65.93	26.30	66.82	68.88
Plier	Recall	66.35	50.07	71.45	76.14
	Precision	85.49	72.34	86.81	85.16
	AP	74.20	57.63	75.84	79.53
Scissor	Recall	63.62	25.70	52.80	67.76
	Precision	91.16	66.27	91.87	85.80
	AP	72.11	39.08	62.94	73.60
mAP		70.76	49.22	70.89	74.78

Table 5 Detection results of different training data on OPIXray.

Training data	OPIXray dataset		
	Recall	Precision	AP
OPIR	71.08	75.85	67.47
OPIRS	71.14	76.60	72.20
OPIRS+	**80.46**	**90.13**	**84.80**

For SIXray dataset, it contains six kinds of prohibited items. We don't use the hammer class since it involves merely 60 samples. The SIXray10 dataset contains 13412 testing samples, and 74960 training samples (SIXR:7496 positive samples with 4322 guns, 2758 knives, 2816 wrenches, 4624 pliers and 918 scissors). Moreover, we also train YOLOv4 by using following 74960 training samples:

(1) SIXR1/2: We remove half the (3748) positive samples from SIXR, and supplement the other 3748 negative samples into SIXR.

(2) SIXRS: We synthesize the prohibited items into the newly-added 3748 negative samples in SIXR1/2, to supplement the positive samples, resulting in mixed 7496 true/pseudo positive samples with the same number of prohibited items as SIXR.

(3) SIXRS+: We synthesize prohibited items into other new 3748 negative samples, and replace 3748 negative samples in the SIXR, resulting in mixed 11244 true/pseudo positive samples, with 6469 guns, 4154 knives, 4195 wrenches, 6933 pliers and 1398 scissors.

For OPIXray dataset, it contains 7109 training and 1776 testing images for cutter detection. The training datasets are: (1) OPIR: All the original 7109 images; (2) OPIRS: 3555 images supplemented with synthesized cutter, with the same number of cutter as OPIR; (3) OPIRS+: 7109 images supplemented with half the synthesized cutter.

The detection results on SIXray and OPIXray dataset are respectively listed in Table 4 and Table 5.

By comparing the models trained by SIXRS and SIXR1/2, SIXRS recover the degraded performance of SIXR1/2 by supplementing the synthetic positive samples effectively. Moreover, compared to SIXR, SIXRS can achieve higher mAP, and overall higher AP except for the scissor class. The comparison of OPIRS and OPIR also well verifies that our synthetic cutters are real enough to supplement the missing samples. The SIXRS achieves higher precision but lower recall for the scissor class, because our private scissor item database does not cover all the scissor type of SIXray dataset, such as nail clipper set, to simulate the SIXray samples. The recall and AP of the scissor class could be improved furthermore if the prohibited items that are more similar to SIXray dataset are used for fusion. The comparison results between the SIXR and SIXRS+, OPIR and OPIRS+ also both can indicate that our synthetic positive images are beneficial in augmenting the training dataset to further improve the performance of the detection model.

5. Conclusion

We propose a RWSC-Fusion model, which fuses the prohibited items with baggage image in a Region-Wise Style-Controlled manner, for synthesizing the prohibited X-ray security images. The RWSC-Fusion learns the fused X-ray security images unsupervisedly based on novel loss functions: LL and CSD. Moreover, to generate more high-quality realistic and images, we innovate style-controlled mechanism and Edge-Attention module. The ablation studies verify the efficacy of the RWSC-Fusion and Edge-Attention module in achieving superior perceptual quality. The comparative experimental results illustrate that our method outperforms the state-of-the-arts in the field of X-ray security image. The performance comparison of the detector training demonstrates the authenticity and reliability of the synthetic images in promoting down-stream tasks. In future work, we will generalize RWSC-Fusion to solve multiple fusion problems.

Acknowledgements. This work is supported by Zhejiang Dahua Technology Co., Ltd. and Zhejiang University.

References

[1] Samet Akcay, Toby P. Breckon. Towards Automatic Threat Detection: A Survey of Advances of Deep Learning within X-ray Security Imaging. *Pattern Recognition*, 2022.

[2] Rajalingam Balakrishnan, Rainy Priya, Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis, *Int. J. Eng. Sci. Invent*. 2 (Special issue), 52-60, 2018.

[3] Neelanjan Bhowmik, Qian Wang, Yona Falinie A. Gaus. The Good, the Bad and the Ugly: Evaluating Convolutional Neural Networks for Prohibited Item Detection Using Real and Synthetically Composited X-ray Imagery, *British Machine Vision Conference Workshop*, 2019.

[4] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv2004*.10934, 2020.

[5] Jer Chan, Paul Evans, Xun Wang. Enhanced color coding scheme for kinetic depth effect X-ray (KDEX) imaging. In *IEEE International Carnahan Conference on Security Technology*, pages 155-160, 2010.

[6] An Chang, Yu Zhang, Shunli Zhang, Leisheng Zhong, Li Zhang. Detecting prohibited objects with physical size constraint from cluttered X-ray baggage images. *Knowledge-Based Systems*, 237, 107916, 2022.

[7] Jaehyeong Cho, Wataru Shimoda, Keiji Yanai. Mask-based Style-Controlled Image Synthesis Using a Mask Style Encoder. In *International Conference on Pattern Recognition (ICPR)*, 2020

[8] Voria Cutler and Susan Paddock. Use of threat image projection (TIP) to enhance security performance. In *International Carnahan Conference on Security Technology*, 2009.

[9] Piella Gemma and Heijmans Henk. A new quality metric for image fusion. In *Proceedings of IEEE International Conference on Image Processing*, pages 173-76, 2003.

[10] Ian J. Goodfellow, *et al*. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2672-2680, 2014.

[11] Taimur Hassan, Samet Akcay, Mohammed Bennamoun, Salman Khan, and Naoufel Werghi. A Novel Incremental Learning Driven Instance Segmentation Framework to Recognize Highly Cluttered Instances of the Contraband Items. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 2021.

[12] Taimur Hassan, Salman H. Khan, Samet Akcay, Mohammed Bennamoun, and Naoufel Werghi. Deep CMST Framework for the Autonomous Recognition of Heavily Occluded and Cluttered Baggage Items from Multivendor Security Radiographs. *arXiv preprint arXiv*1912.04251, 2019.

[13] Prabhakar K R, Srikar V S, Babu R V. DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs. In *Proc. IEEE Int. Conf. Comput. Vis*, 2017.

[14] Brian K. S. Isaac-Medina, Neelanjan Bhowmik, Chris G. Willcocks, Toby P. Breckon. Cross-modal Image Synthesis within Dual-Energy X-ray Security Imagery. In *CVPR*, 2022.

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[16] Kevin. J. Liang, Chris. Gregory, Souleymane. O. Diallo, Kris. Roe. Automatic threat recognition of prohibited items at aviation checkpoint with x-ray imaging: a deep learning approach. *Anomal*. *Detect*. *Imag*, X-Rays III, 2018.

[17] Dashuang Li, Xiaobing Hu, Haigang Zhang, and Jinfeng Yang. A GAN based method for multiple prohibited items synthesis of X-ray security image. *Optoelectronics Letters*, 17:112-117, 2021.

[18] Jinyi Liu, Xiaxu Leng, Ying Liu. Deep convolutional neural network based object detector for X-ray baggage security imagery. In *IEEE International Conference on Tools with Artificial Intelligence*, pages 1757-1761, 2019.

[19] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11-26, 2019.

[20] Kede Ma, Zhengfang Duanmu, Hanwei Zhu, Yuming Fang. Deep guided learning for fast multi-exposure image fusion. *IEEE Transactions on Image Processing*. 29:2808-2819, 2020.

[21] Harry.E Martz, Clint.M Logan, Daniel.J Schneberk, Peter.J Shull. X-ray Imaging: Fundamentals. *Industrial Techniques and Applications*, CRC Press, 2016.

[22] Domingo Mery. Computer vision for X-Ray testing. *Springer International Publishing*, 973-978, 2015.

[23] Domingo Mery, Aggelos Katsaggelos. A logarithmic x-ray imaging model for baggage inspection: Simulation and object detection. In *CVPR*, pages 57-65, 2017.

[24] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of X-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34(4):42, 2015.

[25] Domingo Mery, Erick Svec, Marco Arias, Vladimir Riffo, Jose M Saavedra, and Sandipan Banerjee. Modern computer vision techniques for x-ray testing in baggage inspection. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*: *Systems*, 47(4):682-692, 2017.

[26] Caijing Miao, Lingxi Xie, Fang Wan, Chi Su, Hongye Liu, Jianbin Jiao, and Qixiang Ye. Sixray: A large-scale security inspection X-ray benchmark for prohibited item discovery in overlapping images. In *CVPR*, pages 2119-2128, 2019.

[27] Stefan Michel, Saskia M. Koller, Jaap C. de Ruiter, Robert Moerland, Maarten Hogervorst, Adrian Schwaninger. Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners. In *Annual IEEE International Carnahan Conference on Security Technology*, pages 201-206, 2007.

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

[29] Robin Riz à Porta, Yanik Sterchi, and Adrian Schwaninger. How Realistic Is Threat Image Projection for X-ray Baggage Screening?. *Sensors*, 22(6), 2022.

[30] Thomas William Rogers, Nicolas Jaccard, Emmanouil. Protonotarios, James. Ollier. Threat Image Projection (TIP) into X-ray images of cargo containers for training humans and machines. In *IEEE International Carnahan Conference on Security Technology (ICCST)*, pages 1-7, 2016.

[31] Fangtao Shao, Jing Liu, Peng Wu, Zhiwei Yang, and Zhaoyang Wu. Exploiting foreground and background separation for prohibited item detection in overlapping X-Ray images. *Pattern Recognition*, 2022.

[32] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 2019.
[33] Karen Simonyan, and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556, 2014.
[34] Youssef Skandarani, Pierre-Marc Jodoin, Alain Lalande. GANs for Medical Image Synthesis: An Empirical Study. *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, 2021.
[35] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei. Exploring Endogenous Shift for Cross-domain Detection: A Large-scale Benchmark and Perturbation Suppression Network. In *CVPR*, CCF-A, 2022.
[36] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded Prohibited Items Detection: an X-ray Security Inspection Benchmark and De-occlusion Attention Module. In *Proceedings of ACM International Conference on Multimedia*, pages 138-146, 2020.
[37] Jianping Xiong, Dong Hu, Lijian Mao, Min Wu, and Jingsong Zhu. DoubleRYOLO: Rotated Prohibited Item Detection for X-ray Security Inspection System. In *International Conference on Graphics and Signal Processing*, 2021.
[38] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, Haibin Ling, U2Fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell*, 2020.
[39] C. Xydeas, and Vladimir. Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308-309, 2000.
[40] Jinfeng Yang, Zihao Zhao, Haigang Zhang, and Yihua Shi. Data augmentation for X-ray prohibited item images using generative adversarial networks. *IEEE Access*, 7:28894-28902, 2019.
[41] Tengfei Zhao, Haigang Zhang, Yutao Zhang and Jinfeng Yang. X-Ray Image with Prohibited Items Synthesis Based on Generative Adversarial Network. *Chinese Conference on Biometric Recognition*, 2019.
[42] Barret Zoph, Ekin D. Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. Learning Data Augmentation Strategies for Object Detection. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.