X-Ray Image with Prohibited Items Synthesis Based on Generative Adversarial Network

Tengfei Zhao[1,2], Haigang Zhang[1,2], Yutao Zhang[1,2], and Jinfeng Yang[1,2(✉)]

[1] Tianjin Key Lab for Advanced Signal Processing, Civil Aviation University of China, Tianjin, China
[2] Shenzhen Polytechnic, Shenzhen 518055, China
jfyang@szpt.edu.cn

Abstract. Using deep learning to assist people in recognizing prohibited items in X-Ray images is crucial to improve the quality of security inspections. However, these methods require lots of data and the data collection usually takes much time and efforts. In this paper, we propose a method to synthesize X-ray image to support the training of prohibited items detectors. The proposed framework is built on the Generative Adversarial Networks (GAN) with multiple discriminators, trying to synthesize realistic X-Ray prohibited items and learn the background context simultaneously. In the other hand, a guided filter is introduced for detail preserving. The experimental results show that our model can smoothly synthesize prohibited items on background images. To quantitatively evaluate our approach, we add the generated samples into training data of the Single Shot MultiBox Detector (SSD) and show the synthetic images are able to improve the detectors' performance.

Keywords: Image synthesis · Generative Adversarial Network · X-ray baggage security

1 Introduction

Baggage inspection with X-ray machines is a priority task, which can reduces the risk of crime and terrorist attacks [1]. Security and safety screening with X-ray scanners has become an important process in the transportation industry and at border checkpoints [2]. However, inspection is a complex task and the detection for prohibited items relies mainly on the human. Missed inspection is an unavoidable mistake, when the security inspector has worked for a long time. This will cause security risks. Therefore, this type of task is more suitable for computer processing, freeing human from this heavy work.

With the advances of Convolutional Neural Networks(CNN), the realization of intelligent security check is no longer out of reach [3]. However, most prohibited items detection models require lots of images and manually collecting images usually takes much time and efforts. There are currently almost no public data sets containing prohibited items on the web. Therefore, it is very important

Z. Sun et al. (Eds.): CCBR 2019, LNCS 11818, pp. 379–387, 2019.
https://doi.org/10.1007/978-3-030-31456-9_42

to design approaches that automatically synthesize images for extending new datasets. Motivated by recent promising success of GANs [4] in several applications [5–7], we propose to build a GAN-based model to synthesize realistic prohibited items images in real scene and utilize them as the augmented data to train the CNN-based prohibited items detector. We denominate it as X-ray image-Synthesis-GAN(XS-GAN). Compared with adopting the regular GAN, the XS-GAN synthetic images are more realistic and retain more details.

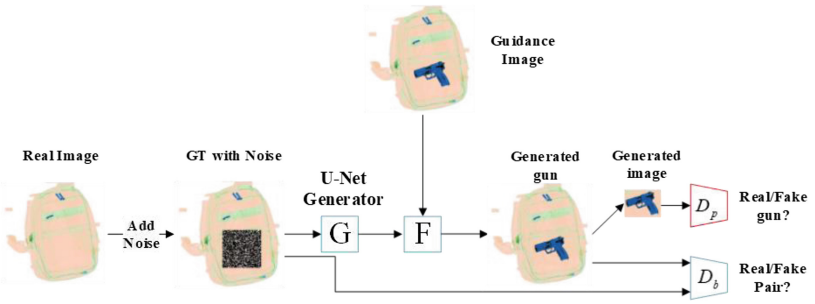

Fig. 1. The XS-GAN model.

XS-GAN adopts the adversarial learning recipe and contains multiple discriminators: D_b for background context learning and D_p for prohibited items discriminating (the gun as an example), as shown in Fig. 1. We replace the prohibited items with the bounding boxes with random noise and train the generator G to synthesize new prohibited items within the noise region. The discriminator D_b, learns to discriminate between real and synthesized pair. Meanwhile, the discriminator D_p learns to judge whether the synthetic prohibited item cropped from the bounding boxes is real or fake. D_b aims to force G to learn the background information. It leads to smooth connection between the background and the synthetic prohibited items. In order to makes G to generate real prohibited items with more realistic shape and details, we introduce guided filters into the proposed XS-GAN. After training, the generator G can learn to generate photo-realistic prohibited items in the noise box regions.

2 Related Work

2.1 Generative Adversarial Network

GANs [4] have achieved great success in generating realistic new images from either existing images or random noises. The main idea is to have continuing adversarial learning between a generator and a discriminator, where the generator tries to generate more realistic images while the discriminator aims to distinguish the newly generated images from real images. It is like a game, and will reach a state of balance. The generate image is consistent with the original image.

2.2 Image Synthesis with GAN

The work of image synthesis using GAN is generally based on the image-to-image translation work. The Pix2pix-GAN [5] is the earliest image-to-image translation model based on the condition GAN [8]. CycleGAN [6], DiscoGAN [9], and DualGAN [10] are similarly in principles. CycleGAN replaces the traditional one-way generated GAN with a loop-generated ring network and changes the traditional input method of paired images. Therefore, the input to the model becomes available for any two images. GAWWN [11] introduced a new synthesis method, which can synthesizes higher resolution images given instructions describing what content to draw in which location. PS-GAN [7] proposed an algorithm that can smoothly synthesize pedestrians on background images of varying and different levels of detail.

2.3 Guided Filter

Guided Filters [12,13] use one image as a guide for filtering another image, which exhibits superior performance in detail preserving filtering. The filtered output is a linear transformation of the guided image, where the guided image can be the input image itself or another different image. Guided filtering has been used for a variety of computer vision tasks. [14] uses guided filter for weighted averaging and image fusion. [15] uses a rolling guidance to fully control the detail smoothing in an iterative manner. [16] uses guided filtering to suppress heavy noise and structural inconsistency. [17] uses guided filtering as a non-convex optimization problem and proposes solutions via majorize-minimization.

Most GANs for image-to-image translation can synthesize high-resolution images, but the appearance transfer usually suppresses image details such as edges and textures. The proposed XS-GAN introduces guided filter into the generator network, which enables both appearance transfer and detail retention.

3 The Proposed Method

Unlike the regular GAN, our method leverages an adversarial process between the generator G and two discriminators: D_b for background context learning and D_p for discriminating prohibited items. In this section, we will give a detailed formulation of the overall objective.

3.1 Model Architecture

U-Net for Generator *G*. The Generator G learns a mapping function $G{:}x \rightarrow y$, where x is the input noise image and y is the ground truth image. In this work, we adopt the enhanced encoder-decoder network (U-Net) [5] for G. It follows the main structure of the encoder-decoder architecture, where the input image x is passed through a series of convolutional layers as down-sampling layers until the bottleneck layer. Then the bottleneck layer feeds the encoded information

of original inputs to the deconvolutional layers to be up-sampled. U-Net uses the skip connections to connect the down-sampling and up-sampling layers to symmetric locations relative to the bottleneck layer, which can preserve richer local information (Fig. 2).

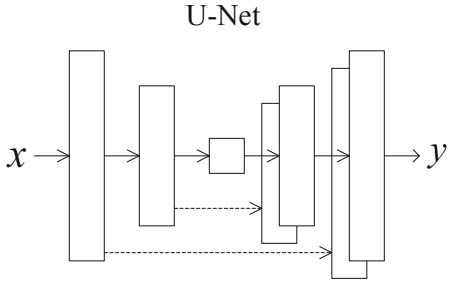

Fig. 2. The U-Net structure of the Generator G.

***D_p* to Discriminate Fake/Real Prohibited Items.** For this discriminator D_p, we crop the synthetic prohibited items from the generated image as a negative sample, while the real prohibited items y_p from the original image y as a positive sample. Therefore, D_p is used to classify whether the generated prohibited item is real or false in the noise area. It forces G to learn the mapping from z to the real prohibited items y_p, where z is the noise region in the noise image x.

***D_b* to Learn Background Context.** The goal of our model is to not only synthesize real prohibited items but also smoothly fill the synthetic prohibited items into the background. Thus our model needs to learn context information. Following the pair-training recipe from Pix2Pix-GAN [5], D_b is used to classify between real and synthetic pairs. The real pair is the noise image x and the ground truth image y, while the synthesized pair is the noise image x and the generated image. The overall framework is shown in Fig. 3.

Guided Filter. Guided filter is designed to perform edge-preserving image smoothing by using the structure in the guidance image. We introduce the guided filter into the proposed XS-GAN and formulate the detail-preserving as a joint up-sampling problem. In particular, the synthetic images (image detail loss) output of G is the input image I to be filtered and the initially input image act as the guidance image R to provide edge and texture details. Therefore, the detail-preserving image T can be derived by minimizing the reconstruction error between I and T, subjects to the linear model:

$$T_i = a_k I_i + b_k, \forall i \in \omega_k \tag{1}$$

where i is the index of the pixel and ω_k is a local square window centered at pixel k .

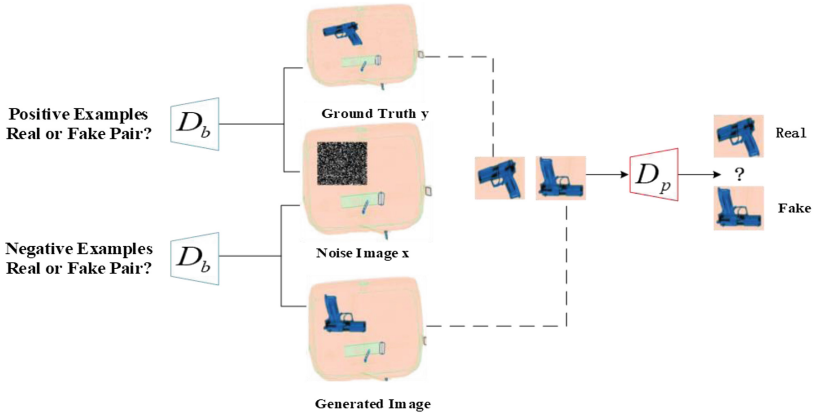

Fig. 3. The overall structure of discriminator.

In order to determine the coefficients of the linear models a_k and b_k, we seek a solution that minimizes the difference between T and filter input R, which can be derived by minimizing the following cost function in the local window:

$$E\left(a_k, b_k\right) = \sum_{i \in \omega_k} \left(\left(a_k I_i + b_k - R_i\right)^2 + \in a_k^2\right) \tag{2}$$

here $a_k I_i + b_k$ represents the output of the filter. Since the output of the filter combines the characteristics of the guidance image and the input image, $\left(a_k I_i + b_k - R_i\right)^2$ is used here to measure the similarity between the output image and the input image. And $\in$ is a regularization parameter that prevents a_k from being too large. It can be solved by linear regression:

$$a_k = \frac{\frac{1}{|\omega|} \sum_{i \in \omega_k} I_i - \mu_k \bar{R}_k}{\bar{\sigma}_k + \in} \tag{3}$$

$$b_k = \bar{R}_k - a_k \mu_k \tag{4}$$

where μ_k and σ_k^2 are the mean and variance of I at ω_k, $|\omega|$ is the number of pixels in ω_k, and $\bar{R}_k = \frac{1}{|\omega|} \sum_{i \in \omega_k}$ is the average of R in ω_k.

By applying a linear model to all ω_k windows on the image and calculating (a_k, b_k), the filter output can be derived by averaging all possible values of T_i:

$$T_i = \frac{1}{|\omega|} \sum_{k: i \in \mu_k} \left(a_k I_i + bk\right) = \bar{a} I_i + \bar{b}_i \tag{5}$$

where $\bar{a}_i = \frac{1}{|\omega|} \sum_{k \in \omega_k} a_k$ and $\bar{b}_i = \frac{1}{|\omega|} \sum_{k \in \omega_i} b_k$. We integrate the guided filter into the generator network structure to implement an end-to-end trainable system.

3.2 Loss Function

As shown in Fig. 1, this model includes two adversarial learning processes $G \Leftrightarrow D_b$ and $G \Leftrightarrow D_p$. The adversarial learning between G and D_b can be formulated as:

$$\begin{aligned}\mathcal{L}_{LSGAN}(G, D_b) = E_{y \sim p_{gt \cdot image}(y)}\left[\left(D_b(y) - 1\right)^2\right] \\ + E_{x,z \sim p_{noise \cdot image}(x,z)}\left[\left(D_b(G(x,z))\right)^2\right]\end{aligned} \tag{6}$$

where x is the image with noise and y is the ground truth image. The original GAN loss is replaced here with the least squared loss of LSGAN.

To encourage G to generate realistic prohibited items within the noise box z in the input image x, another resistance loss is added between G and D_p:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_p) = E_{y_p \sim p_{prohibiteditems}(y_p)}\left[\log D_p(y_p)\right] \\ + E_{z \sim p_{noise}(z)}\left[\log\left(1 - D_p(G(z))\right)\right]\end{aligned} \tag{7}$$

where z is the noise box in x and y_p is the crop prohibited items in the ground truth image y. The negative log-likelihood targets are used to update the parameters of G and D.

GAN training can benefit from traditional losses [5]. In this paper, the L loss is used to control the difference between the generated image and the real image y:

$$\mathcal{L}_{\ell_1}(G) = E_{x,z \sim p_{noise \cdot image}(x,z), y \sim p_{gt \cdot image}(y)}\left[\|y - G(x,z)\|_1\right] \tag{8}$$

Finally, combining the previously defined losses results in a final loss function:

$$\mathcal{L}(G, D_b, D_p) = \mathcal{L}_{LSGAN}(G, D_b) + \mathcal{L}_{GAN}(G, D_p) + \lambda \mathcal{L}_{\ell_1}(G) \tag{9}$$

4 Experimental Results

4.1 Datasets

The datasets used in our experiments are collected from the laboratory. We experiment with several types of prohibited items, such as guns, fruit knives, forks, hammers and scissors.

4.2 Contrast Experiment

In this section, We conducted several synthetic experiments on prohibited items and evaluated the synthesized images. The experimental results are shown in Fig. 4.

As can be seen from Fig. 4, our XS-GAN model with guided filtering has a better effect on the synthesis of security image prohibited items. The pix2pix-GAN model hardly the generated prohibited items. The PS-GAN can generate

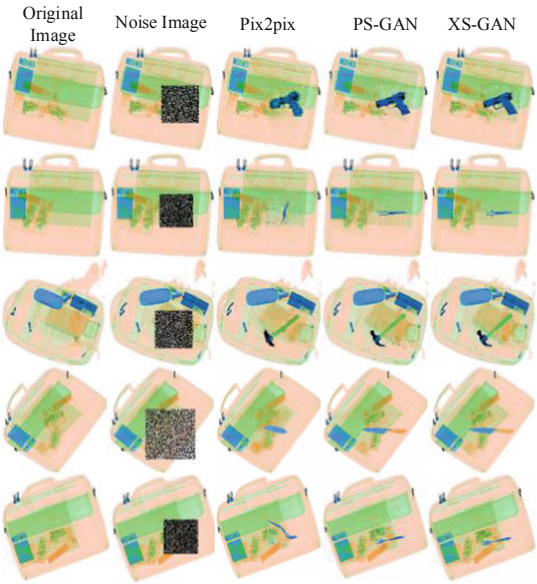

Fig. 4. Columns 1–2 are the input images, columns 3–5 show the images synthesized by pix2pix-GAN, PS-GAN and XS-GAN.

prohibited items, but the synthesis image is not clear enough. The images generated using our improved XS-GAN network model are not only clearer but also can retain more details.

In order to evaluate the quality of synthetic images, we test Fréchet Inception Distance (FID) score. The smaller the value of FID, the closer the synthetic image is to the real image. The test results are shown in Table 1.

Table 1. FID Score Test.

Model	Score
Pix2pix	69.73
PS-GAN	57.44
XS-GAN	47.21

As shown in Table 1, the XS-GAN synthesized images has the lowest FID score, which proves that the images synthesized by our method are closer to the real images.

To analyze the effect of the data augmentation, we combine the real and synthesized data to train the SSD [18] detectors and evaluate the performance. We experimented with images of three prohibited items, pistols, forks, and scissors. In the first experiment we use all the real images for training. In the second

experiment we use all the synthesized images for training. In the third experiment we use half of the real images and half of the synthesized images for training. We use synthetic images from PS-GAN and XS-GAN to train SSD, separately. The results of the evaluation are shown in Table 2.

Table 2. Results of the SSD algorithm evaluation. Experimental evaluations were performed using real data sets, synthetic data sets, and mixed data sets, respectively.

Model	Data	mAP	gun	fork	scissor
	Real images	0.793	0.909	0.891	0.579
PS-GAN	Synthetic images	0.822	0.913	0.901	0.654
	Real + Synthetic	0.845	0.921	0.916	0.697
XS-GAN	Synthetic images	0.877	0.936	0.929	0.767
	Real + Synthetic	0.895	0.949	0.906	0.831

Table 2 shows that the detector is trained with synthetic images from XS-GAN can improve 8% mAP, and the detector is trained with mixed images can improve 10% mAP. However, the detector is trained with synthetic images from PS-GAN can improve 3% mAP, and the detector is trained with mixed images can improve 5% mAP. Thus by adding the synthetic images, the AP rate can be improve, and the image synthesized by our method has better data enhancement effect.

5 Conclusion

This paper introduces the XS-GAN model to synthesizes realistic X-Ray images in certain bounding boxes. The experimental results show that the network model with guided filtering can retain more details when synthesizing images. Our model can generate high quality prohibited items images, and the synthetic images can effectively improve the abilities of CNN based detectors. We use this model to synthesize different prohibited items images, which demonstrates the ability of generalization and transferring knowledge. We will continue to study XS-GAN for prohibited items image synthesis for training better detection models.

References

1. Mery, D., Svec, E., Arias, M., Riffo, V., Saavedra, J.M., Banerjee, S.: Modern computer vision techniques for x-ray testing in baggage inspection. IEEE Trans. Syst. Man Cybern. Syst. **47**(4), 682–692 (2016)
2. Mendes, M., Schwaninger, A., Michel, S.: Does the application of virtually merged images influence the effectiveness of computer-based training in x-ray screening? In: 2011 Carnahan Conference on Security Technology, pp. 1–8. IEEE (2011)

3. Rogers, T.W., Jaccard, N., Griffin, L.D.: A deep learning framework for the automated inspection of complex dual-energy x-ray cargo imagery. In: Anomaly Detection and Imaging with X-Rays (ADIX) II, vol. 10187. International Society for Optics and Photonics, 101870L (2017)
4. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
5. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
6. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
7. Ouyang, X., Cheng, Y., Jiang, Y., Li, C.L., Zhou, P.: Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. arXiv preprint arXiv:1804.02047 (2018)
8. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
9. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 1857–1865. JMLR. org (2017)
10. Yi, Z., Zhang, H., Tan, P., Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2849–2857 (2017)
11. Reed, S.E., Akata, Z., Mohan, S., Tenka, S., Schiele, B., Lee, H.: Learning what and where to draw. In: Advances in Neural Information Processing Systems, pp. 217–225 (2016)
12. He, K., Sun, J., Tang, X.: Guided image filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 1–14. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15549-9_1
13. He, K., Sun, J.: Fast guided filter. arXiv preprint arXiv:1505.00996 (2015)
14. Li, S., Kang, X., Hu, J.: Image fusion with guided filtering. IEEE Trans. Image Process. **22**(7), 2864–2875 (2013)
15. Zhang, Q., Shen, X., Xu, L., Jia, J.: Rolling guidance filter. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 815–830. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10578-9_53
16. Liu, W., Chen, X., Shen, C., Yu, J., Wu, Q., Yang, J.: Robust guided image filtering. arXiv preprint arXiv:1703.09379 (2017)
17. Ham, B., Cho, M., Ponce, J.: Robust guided image filtering using nonconvex potentials. IEEE Trans. Pattern Anal. Mach. Intell. **40**(1), 192–207 (2018)
18. Liu, W., et al.: SSD: single shot multibox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2