$\mathcal{X}$-*Adv*: Physical Adversarial Object Attacks against X-ray Prohibited Item Detection

Aishan Liu[1]*, Jun Guo[1]*, Jiakai Wang[2], Siyuan Liang[3], Renshuai Tao[1],
Wenbo Zhou[4], Cong Liu[5], Xianglong Liu[1,2,6]†, Dacheng Tao[7]
[1]*Beihang University,* [2]*Zhongguancun Laboratory,* [3]*Chinese Academy of Sciences,*
[4]*University of Science and Technology of China,* [5]*iFLYTEK,*
[6]*Hefei Comprehensive National Science Center,* [7]*JD Explore Academy*

Abstract

Adversarial attacks are valuable for evaluating the robustness of deep learning models. Existing attacks are primarily conducted on the visible light spectrum (*e.g.*, pixel-wise texture perturbation). However, attacks targeting texture-free X-ray images remain underexplored, despite the widespread application of X-ray imaging in safety-critical scenarios such as the X-ray detection of prohibited items. In this paper, we take the first step toward the study of adversarial attacks targeted at X-ray prohibited item detection, and reveal the serious threats posed by such attacks in this safety-critical scenario. Specifically, we posit that successful physical adversarial attacks in this scenario should be specially designed to circumvent the challenges posed by color/texture fading and complex overlapping. To this end, we propose $\mathcal{X}$-*Adv* to generate physically printable metals that act as an adversarial agent capable of deceiving X-ray detectors when placed in luggage. To resolve the issues associated with color/texture fading, we develop a differentiable converter that facilitates the generation of 3D-printable objects with adversarial shapes, using the gradients of a surrogate model rather than directly generating adversarial textures. To place the printed 3D adversarial objects in luggage with complex overlapped instances, we design a policy-based reinforcement learning strategy to find locations eliciting strong attack performance in worst-case scenarios whereby the prohibited items are heavily occluded by other items. To verify the effectiveness of the proposed $\mathcal{X}$-*Adv*, we conduct extensive experiments in both the digital and the physical world (employing a commercial X-ray security inspection system for the latter case). Furthermore, we present the physical-world X-ray adversarial attack dataset XAD. We hope this paper will draw more attention to the potential threats targeting safety-critical scenarios. Our codes and XAD dataset are available at https://github.com/DIG-Beihang/X-adv.

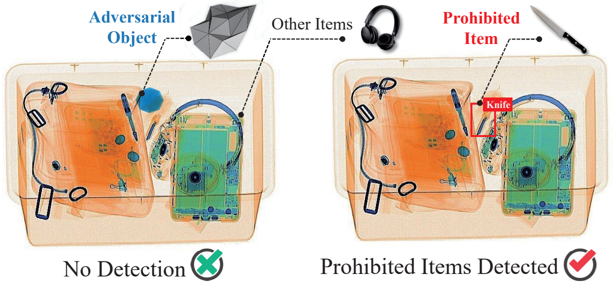


Figure 1: Illustration of physical-world adversarial attacks on X-ray security inspection. This paper proposes $\mathcal{X}$-*Adv* to generate physically realizable 3D adversarial objects. During X-ray scanning, the detector can detect prohibited items in the right image, but our adversarial objects deceive the detector into failing to detect prohibited items in the left image.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable performance across a wide area of applications [1, 19, 21]. Recently, deep learning has been introduced into safety-critical scenarios such as X-ray security inspection in public transportation hubs (*e.g.*, airports). In this scenario [32, 40, 42, 47], deep-learning-based detectors are utilized to assist inspectors in identifying both the presence and location of prohibited items (*e.g.*, pistols and knives) during X-ray scanning. This approach significantly reduces the amount of human labor required and helps to protect the public from severe risks.

Despite their promising performance, DNNs are vulnerable to *adversarial examples* [38]. These elaborately designed perturbations are imperceptible to human vision, but can easily mislead DNNs into making wrong predictions, thus threatening practical deep learning applications [22–24]. By contrast, adversarial examples are also beneficial for evaluating and better understanding the robustness of DNNs [26, 39, 50, 51, 53]. In the past years, extensive research has been conducted into performing adversarial attacks on natural images (visual light); however, the robustness of texture-free X-ray images (such as in the context of X-ray prohibited item detection)

* Equal contribution.
† Corresponding author.

remains underexplored. This sparsity of research presents a severe risk to the safety of the general public, as it increases their vulnerability to attack.

In this paper, we take the first step in physical-world adversarial attacks on the X-ray prohibited item detection scenario, *i.e.*, deceive the detector to wrong predictions by strategically placing adversarial objects around the prohibited item. However, simply extending existing physical attacks that work well on natural images to the context of X-ray images is non-trivial owing to the different imaging principles, *e.g.*, the wavelengths of X-rays (*0.001*$\sim$*10nm*) and visible light (*390*$\sim$*780nm*) have a huge difference. More specifically, X-ray imaging is primarily conducted by utilizing material, thickness, and attenuation coefficients, meaning that the existing physical attacks designed for a visible light context (*e.g.*, interference textures [46] or patches [2]) cannot be effectively applied to X-ray imaging. Thus, X-ray attacks should be considered a new type of attack problem in visually constrained scenarios with different wavelengths. In particular, we identify two key challenges impeding successful and feasible adversarial attacks in this scenario: (1) Color/texture fading. Due to its use of special imaging principles (*i.e.*, beam intensity and attenuation rule), the X-ray scanning process eliminates most of the colors/textures and projects its outputs primarily based on item materials and shapes. Thus, the commonly used perturbations utilizing color disturbances will be removed by the X-ray scanning causing them to be ineffective. (2) Complex overlap. Luggage passed through an X-ray scanner often contains a large number of objects made of different materials, and overlap between these objects can degrade the attack performance; moreover, a successful attack should not rely on the occlusion of the prohibited item. Thus, when designing an adversarial object, it is necessary to consider the worst-case scenario (complex overlapping instances within the luggage), which increases the difficulty of the task.

To address the above problems, this paper proposes an adversarial attack approach called $\mathcal{X}$*-Adv* to generate physically realizable adversarial attacks for X-ray prohibited item detection (as shown in Figure 1). As for the *color/texture fading*, we generate physically realizable 3D objects with adversarial shapes, which enable our attacks to remain effective (since the shape cannot be altered after the X-ray imaging). To guide the design of the shape, we derive a differentiable converter that projects 3D objects into X-ray images so that we could update the shape of the object using the gradients of a surrogate white-box detector. As for the *complex overlap*, we aim to find the locations that achieve strong attack ability even when occluded by other objects; moreover, we ensure that the placed adversarial objects do not overlap with the prohibited item. We thus introduce a policy-based strategy to search for the location that provides optimal attacking performance in the worst-case scenario. In summary, our $\mathcal{X}$*-Adv* can generate adversarial objects by jointly optimizing the shapes and locations for X-ray attacks.

Extensive experiments in both the digital and physical world using multiple benchmarks against several detectors are conducted. Specifically, we first evaluate digital-world attacks on multiple benchmarks against both one-stage and two-stage detectors. We then successfully attack a commercial X-ray security inspection system in the real world by generating adversarial metal objects using a 3D printer. Finally, we present the physical-world X-ray adversarial attack dataset XAD which contains 5,587 images (840 adversarial images). We hope this paper will draw more attention to the potential threats in safety-critical scenarios. Our **contributions** are:

- To the best of our knowledge, this paper is the first work to study the feasibility of physical-world adversarial attacks in the visually-constrained X-ray imaging scenario.
- We propose the $\mathcal{X}$*-Adv* to generate physically realizable adversarial metal objects for X-ray security inspection attacks by addressing the color fading and complex occlusion challenges.
- We conduct extensive experiments on several datasets in both digital- and physical-world settings, and the results demonstrate the effectiveness of our attack.
- We present the physical-world X-ray adversarial attack dataset, XAD, consisting of 5,587 images (840 adversarial images).

2 Backgrounds and Related Work

Prohibited Item Detection in X-ray Images. X-ray imaging has been widely used due to its strong penetrative ability. In the X-ray security inspection scenario, inspectors usually adopt X-ray scanners to check passengers' luggage for the presence of prohibited items. A plethora of studies have been devoted to detecting prohibited items (*e.g.*, pistols) in X-ray scanned luggage images using object detection methods [32, 42, 45, 47] to detection performance.

In addition to the X-ray image detection methods, high-quality X-ray image datasets and benchmarks are also valuable for promoting the development of the research area. Though obtaining colorful X-ray images requires high computational costs, there are still some available open-source specialized datasets for X-ray security inspection. For instance, SIXray [32] is a large-scale X-ray dataset containing millions of X-ray images collected from real-world subway stations. However, the images containing prohibited items are less than 1%, and there is no bounding box annotation provided for object detection. Some high-quality X-ray datasets for object detection have also been made available. Wei *et al.* [47] first released the OPIXray dataset, which contains 8,885 artificially synthesized X-ray images of five categories of cutters. Tao *et al.* [43] proposed the HiXray dataset, comprising 45,364 images containing 102,928 prohibited items. All images from the dataset are collected from X-ray scanners in airports. Recently, Tao *et al.* [41] further proposed the first few-shot object detection dataset in the X-ray security inspection scenario.

Adversarial Attacks. Adversarial examples are inputs with small perturbations, which are imperceptible to humans but can easily mislead DNNs into making incorrect predictions [16,38]. Generally, we can classify them into digital and physical attacks. *Digital attacks* usually generate adversarial perturbation at the pixel level across the whole input image. Szegedy *et al.* [38] first defined adversarial examples and proposed L-BFGS attacks. By leveraging the gradient of the target model, Goodfellow *et al.* [17] proposed FGSM to quickly generate adversarial examples. Since then, many types of adversarial attacks have been proposed, such as PGD [30], DeepFool [33], and JSMA [35]. However, due to their addition of global perturbations to the whole image, these attacks lack physical-world feasibility.

By contrast, *physical attacks* aim to generate adversarial perturbations by perturbing the visual characteristics of real objects in the physical world. To achieve this goal, adversaries often generate adversarial perturbations in the digital world, then perform physical attacks by applying adversarial patches, painting adversarial camouflage, or directly creating adversarial objects in the real world [2,7,12,46]. Brown *et al.* [2] first proposed the adversarial patch by confining the perturbations into a local patch, which could then be printed to deceive the classification models. Eykholt *et al.* [12] then modified the attacking loss function and generated strong adversarial attacks for real-world traffic sign recognition. Chen *et al.* [7] proposed Shapeshifter to attack a Faster R-CNN object detector in the physical world, specifically by attaching it to the STOP signs. In addition to the physical attacks on natural images (visible light domain), there also exist some preliminary studies on other *visually constrained scenarios*. For example, Cao *et al.* [3] investigated attacks in multi-sensor fusion scenarios, making adversarial examples invisible to both cameras and LiDAR. Recently, Zhu *et al.* [54,55] attacked thermal infrared pedestrian detectors using small bulbs and special clothes. Mowery *et al.* [34] attacked a full-body X-ray scanner, while their proposed cyber-physical attacks did not aim at neural networks and are different from adversarial attacks.

In summary, although numerous methods of physical attacks on natural images have been proposed, relatively little is known about the physical-world X-ray security inspection attack. This paper takes the first step to study physical-world adversarial attacks for X-ray security inspection.

3 Threat Model

3.1 Problem Definition

Object detection. An object detector $f_\Theta(\mathbf{I}) \rightarrow \{\mathbf{b},\mathbf{c}\}^K$ with parameters Θ, which takes an image $\mathbf{I} \in [0,255]^n$ as input, outputs K detection boxes with location $\mathbf{b}_k = [s_k, r_k, w_k, h_k]$ and confidence c_k. Moreover, f applies a non-maximum suppression (NMS) operation to remove redundant bounding boxes. The formulation of the training is as follows:

$$\min_{\Theta} \mathbb{E}_{(\mathbf{I},\{\mathbf{y}_k,\mathbf{b}_k\})\sim\mathbb{D}} \mathcal{L}(f_\Theta(\mathbf{I}),\{\mathbf{y}_k,\mathbf{b}_k\}), \tag{1}$$

where $\mathcal{L}(\cdot)$ is the loss function that measures the difference between the output of the detector f and the ground truth. $\mathbf{y}_k$ denotes the true label, and $\mathbf{b}_k$ denotes the true bounding box. In practice, the loss function is a weighted sum of the classification loss $\mathcal{L}_{cls}$ and location loss $\mathcal{L}_{loc}$:

$$\min_{\Theta} \mathbb{E}_{(\mathbf{I},\{\mathbf{y}_k,\mathbf{b}_k\})\sim\mathbb{D}} [\mathcal{L}_{cls}(f_\Theta^{cls}(\mathbf{I}),\mathbf{y}_k) + \lambda \mathcal{L}_{loc}(f_\Theta^{loc}(\mathbf{I}),\mathbf{b}_k)]. \tag{2}$$

Attacks on object detection. Given an object detector f_Θ and an input image $\mathbf{I} \in \mathbb{I}$ with the ground truth label $\{\mathbf{y},\mathbf{b}_k\}$, an adversarial example $\mathbf{I}_{adv}$ satisfies the following:

$$f_\Theta(\mathbf{I}_{adv}) \neq \{\mathbf{y},\mathbf{b}_k\} \quad s.t. \quad \|\mathbf{I}-\mathbf{I}_{adv}\| \leq \varepsilon, \tag{3}$$

where $\|\cdot\|$ is a distance metric and commonly measured via ℓ_p-norm ($p \in \{1,2,\infty\}$). Adversarial examples in visual recognition should also satisfy $\mathbf{I}_{adv} \in [0,255]^n$. In this paper, we focus on deceiving the prediction class labels (*i.e.*, $\mathbf{y}$).

Physical attacks on X-ray prohibited item detection. In this scenario, the items $\mathbf{X} = \{\mathbf{x}_1,...,\mathbf{x}_m\}$ in the luggage are scanned via an X-ray scanner to produce an X-ray image, where $\mathcal{R}$ denotes the process of generating a pseudo-color image depicted in Figure 1 as $\mathbf{I} = \mathcal{R}(\mathbf{X})$). To perform physical attacks, we generate a 3D adversarial object $\mathbf{x}_{adv}$ with adversarial shapes $\mathbf{P}$ and place it at the proper location $\mathbf{C}$ in the luggage; the luggage is then scanned by the X-ray into image $\mathbf{I}_{adv}$, which could deceive the object detector $f_\Theta(\cdot)$, *i.e.*, minimizing $\mathcal{M}$ that measures the performance of the detector:

$$\min_{\mathbf{P},\mathbf{C}} \mathcal{M}\left[f_\Theta(\mathcal{R}(\mathbf{x}_1,...,\mathbf{x}_m,\mathbf{x}_{adv}^{\mathbf{P},\mathbf{C}}),\{\mathbf{y}_k,\mathbf{b}_k\})\right]. \tag{4}$$

3.2 Challenges for X-ray Attacks

Existing attacks mainly aim at the visible light domain by generating adversarial textures. However, it is highly challenging to directly apply these existing attacks to the X-ray domain. Specifically, we observe two main **challenges** as follows.

Challenge ❶: *The significant difference between imaging principles used in the visible light and X-ray contexts (e.g., different wavelengths).* We here first revisit the attenuation rule of X-ray photon beams. According to [31], a narrow beam of X-ray photons with energy E and initial photon intensity I_0, on passing through an absorber of small thickness Δx, will suffer a fractional decrease of intensity $\Delta I/I_0$ given by

$$\frac{\Delta I}{I_0} = -\mu(\rho,Z)\Delta x, \tag{5}$$

where μ is the attenuation coefficient per unit length for an item made of a material of density ρ and atomic composition

Z. When the same photon beam passes through a certain absorber of finite thickness x, the intensity is given by

$$I = I_0 \cdot exp(-\mu(\rho, Z)x). \tag{6}$$

This attenuated intensity then will be received by sensors in X-ray scanners, according to which we can obtain the depth profile of the X-ray images. According to Equation 5 and 6, we can conclude that an X-ray image is constructed primarily with reference to the material, the thickness of the object, and the properties of the light wave itself. Different from the perception of visible light images, X-rays tailor RGB space into a narrow color space, which means that common attacks that change pixel-wise textures will be ineffective for X-rays. To address this challenge, we need to optimize adversarial objectives to use non-color physical properties (*e.g.*, shapes).

Challenge ❷: *Complex overlap due to the diversity of sampling scenarios and a massive number of luggage items in the X-ray security inspection context.* Placing the adversarial object directly on top of prohibited items would appear to be a simple attack method. However, this approach is infeasible in real-world applications, since luggage may be positioned randomly during X-ray scanning, and the overlap rate between adversarial objects and prohibited items under arbitrary sampling conditions is low. Moreover, this violates the definition of adversarial examples. To guarantee a feasible attack, the attacker should consider the worst-case scenario: that is, how to achieve an effective adversarial attack without occluding prohibited items, and with the overlapping of other objects.

3.3 Adversarial Goals

In this paper, we attempt to generate 3D adversarial objects with adversarial shapes to attack physical-world X-ray prohibited item detection models. As illustrated in Section 3.1, given an X-ray prohibited item detector f_Θ that takes an X-ray scanned image $\mathbf{I}$ as input, attackers aim to deceive f_Θ into making wrong predictions. This paper focuses on the more meaningful attack that deceives the detector to predict the wrong class labels rather than the wrong item locations. Specifically, we primarily study the untargeted attack, and the goal is to reduce the detection accuracy of detectors. Meanwhile, we also investigate the possibilities of the more difficult targeted attack, where we aim to force the detector predictions to the `Background` and make these prohibited items "invisible" (Section 5.5). For the untargeted attack, the detector predicts any other labels that are different from the ground truth should be marked as a successful attack; while for the targeted attack, the prediction must match the assigned label.

3.4 Possible Attack Pathways

Regarding adversarial attacks, one of the most important questions that should be answered is whether they are practical. For our $\mathcal{X}$*-Adv* objects, they could be applicable to multiple X-ray image detection-related scenarios, *e.g.*, security inspections in public hubs, and health examinations in hospitals. Using the $\mathcal{X}$*-Adv* approach, adversaries could perform real-world attacks simply by generating an adversarial metal object by using 3D printers, then placing the item into their luggage or bags. The proposed attacks could make detectors yield wrong class predictions with low detection accuracy. Meanwhile, it is also possible for adversaries to conceal a prohibited item and make it "invisible" to the detectors, which can be achieved by simply modifying our attacking loss.

3.5 Adversary Constraints and Capabilities

In considering the real-world X-ray security inspection scenario, we take comprehensive conditions into account and conduct both white-box and black-box attacks. In the white-box attack setting, the adversary has full access to the target model (*e.g.*, architectures, weights), and is able to generate adversarial attacks directly based on its gradients. By contrast, the black-box attack setting is more practical; here, the adversary possesses only a little knowledge about the target model. For this setting, we assume that the target model and the source model are dealing with the same task and that the adversary performs transfer-based attacks. Specifically, the adversary first generates adversarial objects based on a white-box source model from a certain dataset; the adversary then prints the adversarial objects via a 3D printer in the real world; finally, adversaries could simply place adversarial objects in the luggage and attack the deployed X-ray security inspection model. Based on this, we could guarantee that all information of target models is unavailable to the attackers in black-box settings, which helps us to implement the strictest measures for simulating the physical scenarios. Moreover, to ensure our approach is more practical, the size of our adversarial objects should be small; thus the adversarial metal generated in this paper only takes up 1.78% of the X-ray image.

4 $\mathcal{X}$*-Adv* Approach

Selective Search [44] proposes a heuristic strategy to discover potential objects from four perspectives: color similarity, texture similarity, shape similarity, and overlap similarity. Since it is based on the methods by which humans judge objects, this object discovery mechanism has been adopted in current deep-learning-based object detectors. Thus, to deceive the object detector, we can also optimize adversarial objectives from four perspectives: color, texture, shape, and overlap.

Due to the special nature of X-ray imaging principles and the diverse luggage sampling, physical-world adversarial attacks on X-ray security inspection should take the adversarial object's color/texture fading and complex overlapping problems into consideration. Therefore, this paper proposes $\mathcal{X}$*-Adv* to generate physically realizable adversarial objects for X-ray security inspection (as illustrated in Figure 2). This approach

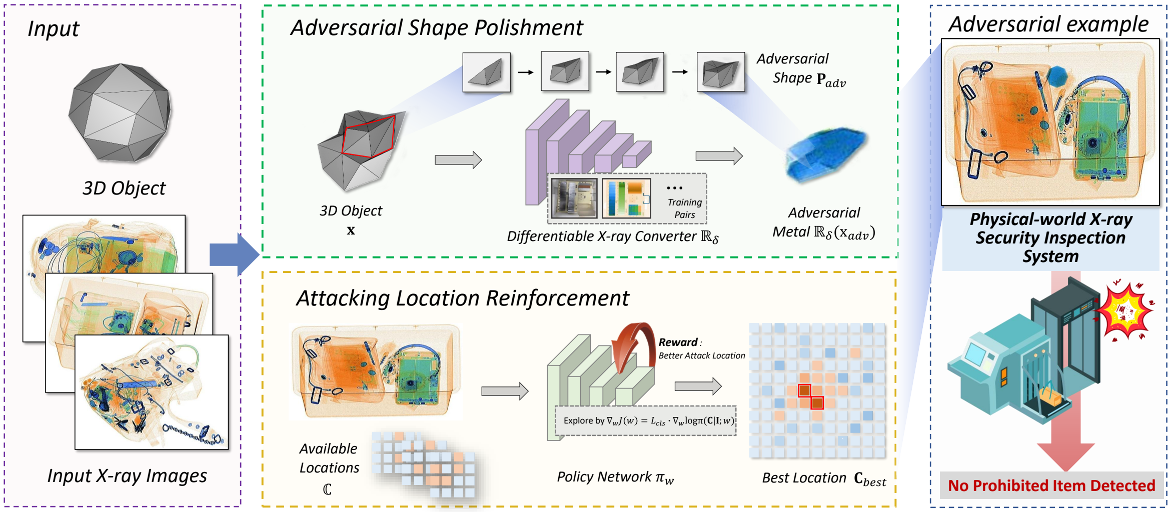

Input
3D Object
Input X-ray Images
Adversarial Shape Polishment
Adversarial Shape $\mathbf{P}_{adv}$
3D Object $\mathbf{x}$
Training Pairs
Differentiable X-ray Converter $\mathbb{R}_\delta$
Adversarial Metal $\mathbb{R}_\delta(\mathrm{x}_{adv})$
Attacking Location Reinforcement
Reward : Better Attack Location
Explore by $\nabla_w J(w) = L_{cls} \cdot \nabla_w \log\pi(\mathbf{C}|\mathbf{I}; w)$
Available Locations $\mathbb{C}$
Policy Network π_w
Best Location $\mathbf{C}_{best}$
Adversarial example
Physical-world X-ray Security Inspection System
No Prohibited Item Detected

Figure 2: Illustration of $\mathcal{X}$-*Adv* approach. For the color/texture fading problem, we derive a differentiable converter that projects 3D objects into X-ray images; this allows us to generate 3D printable objects with adversarial shapes, which is X-ray projection invariant to X-ray imaging. We then introduce a policy-based algorithm to search for the optimal attacking location, which also shows high physical-world feasibility to address the complex occlusion problem. By jointly optimizing the combination of attack locations and shapes, our $\mathcal{X}$-*Adv* can generate physically realizable adversarial attacks for X-ray security inspection.

simultaneously polishes adversarial shapes and reinforces attacking locations in the worst-case scenario (no overlapping) because the color space (*e.g.*, color, texture) is not available.

4.1 Adversarial Shape Polishment

X-ray images, which are generated by X-ray security inspection machines from natural images, emphasize the shape and the material while neglecting the original color/textures of items. Thus, to successfully attack X-ray detectors, we generate objects with *adversarial shapes* rather than *adversarial textures* (since the adversarial colors/textures would be simply eliminated by the X-ray imaging pipeline, making such attacks ineffective). Accordingly, given a 3D object $\mathbf{x}$, we refine its visual characteristics $\mathbf{P}$ (*i.e.*, shape) into adversarial shape $\mathbf{P}_{adv}$, such that the generated adversarial 3D object $\mathbf{x}_{adv}$ can attack the detector $f_\Theta(\mathcal{R}(\mathbf{x}_{adv}))$ after X-ray projection.

However, the X-ray imaging pipeline is highly complex and confidential and also varies significantly across different types of X-ray machines. Meanwhile, it is rather difficult to directly perform black-box query attacks since inspectors will not allow adversaries to query the system several times. Therefore, we propose a possible attack pathway in which we derive a differentiable X-ray converter to simulate the X-ray projection pipeline from 3D objects to 2D X-ray images and then perform gradient-based transfer attacks.

However, the transformation from the depth d of a scanned object to color images g remains unknown. Based on the knowledge of X-ray machine vendors, the transformation process ($d \rightarrow I \rightarrow g$) can be simply represented by exponential functions, where the attenuated intensity I of X-ray beams has an exponential relationship with the object depth d (*c.f.* Eqn. 6) and the intensity I to the color g of X-ray images can be converted using a linear transformation. Therefore, we use the following exponential function to formulate the process:

$$g_m(d) = a \cdot exp(-b \cdot d) + q, \tag{7}$$

where d indicates object depth, m is the material, $g_m(d)$ represents the pixel value of color in a certain depth and material, and a, b, and q are undetermined coefficients correlated to m, which will be calculated from real image sampling and regression fitting. We did not use DNNs for the transformation since it is too costly or even infeasible to collect sufficient data (*i.e.*, different materials with diverse thicknesses) for DNN training (*c.f.* Section 5.1). We use `HSV` color space rather than `RGB` because we found that regression in `HSV` space performs better in reducing the regression error (see Appendix A.2).

However, a depth image cannot represent a unique 3D object. Therefore, we use meshes as the format of our 3D adversarial object, given that meshes have been extensively used to parameterize 3D objects. Given an original mesh $\mathbf{x}_{ori}$, the coordinates on the `XY`-plane represent the shape of the image projected onto the 2D domain, while the coordinates in the `Z`-axis denote the depth (pixel value) of the image. Thus, we can optimize the shape of the 3D object by manipulating the coordinates in the mesh, then project the 3D mesh to a 2D depth image, and finally convert it to an X-ray image (the whole process is differentiable). The adversarial attack loss can be formalized by maximizing the classification loss $\mathcal{L}_{cls}$ of the target model, as follows:

$$\mathcal{L}_{adv}(\mathbf{X},\mathbf{x}_{adv};f_\Theta,\mathbb{R}_\delta) = \arg\max_{\mathbf{P}} \mathcal{L}_{cls}(f_\Theta(\mathbb{R}_\delta(\mathbf{X},\mathbf{x}^{\mathbf{P}}_{adv})),\{\mathbf{y}_k,\mathbf{b}_k\}), \tag{8}$$

where $\mathbb{R}_\delta$ denotes the differentiable converter in Eqn. 7 that can simulate the black-box X-ray scanning process $\mathcal{R}$.

In practice, we overlap the original X-ray image $\mathbf{I}$ and the converted output $g(\cdot)$ by $\mathbf{I} \odot g_m(d(\mathbf{x}^{\mathbf{P}}_{adv}))$, where $d(\mathbf{x}^{\mathbf{P}}_{adv})$ refers to the $\mathbb{Z}$-axis depth of adversarial mesh $\mathbf{x}^{\mathbf{P}}_{adv}$, and $\odot$ indicates pixel-wise multiplication.

4.2 Attack Location Reinforcement

In the X-ray security inspection scenario, there are often a vast number of items in the luggage. The simplest attack method is to attack the target detector by forcing a high overlap of adversarial objects and prohibited items. However, because the angle at which the bag will be scanned is uncontrollable in an X-ray detection scenario, the overlapping probability of adversarial targets and prohibited items are low, while adversarial objects are often occluded by other goods. This complex occlusion problem brings challenges to adversarial attacks. Therefore, we need to study effective attack algorithms in the *worst-case* scenario, whereby the adversarial object does not cover prohibited items and is heavily occluded by other objects. Moreover, an appropriate location would increase the effectiveness of attacks, since our *$\mathcal{X}$-Adv* can only modify shapes while DNNs are more sensitive to texture [8, 20, 37].

Thus, to perform attacks under such a constrained scenario, we make full use of the location of the adversarial object and further improve the efficiency of our attacks by searching for the optimal attack locations. Accordingly, to achieve strong attacks, it is necessary to jointly consider the combination of attacking location and shape $(\mathbf{C}_{best},\mathbf{P}_{best})$:

$$\mathcal{L}_{adv} = \arg\max_{\mathbf{C},\mathbf{P}} \mathcal{L}_{cls}(f_\Theta(\mathbb{R}_\delta(\mathbf{X},\mathbf{x}^{\mathbf{P},\mathbf{C}}_{adv})),\{\mathbf{y}_k,\mathbf{b}_k\}). \tag{9}$$

Meanwhile, these two variables are mutually interactive, and the shape $\mathbf{P}$ is often influenced by the attack locations $\mathbf{C}$. If we determine the attack location $\mathbf{C}_{best}$, the adversarial shape $\mathbf{P}$ can be optimized by the gradient descent algorithm introduced in the previous subsection using the differentiable converter. However, there is no gradient information available for the attack location, which prevents us from optimizing the coordinates of adversarial objects. Also, calculating all possible conditions would result in unacceptably high computational costs. To tackle this problem, we apply a policy-based algorithm to search for the optimal attack location.

Inspired by [56], we use the REINFORCE algorithm [48] to introduce gradients between attack locations and the cost function. We define $\mathbb{C}$ as a finite area surrounding the prohibited item, or the "available attacking area", based on the ground truth bounding box, where we can place our adversarial objects. We simulate the common suitcases scanning scenario in the fixed top-down orientation and divide $\mathbb{C}$ into N evenly spaced grids in the 2D space. In this scenario, the searching problem is relatively simple and the trajectory (state→input, action→location, reward→loss) has only one timestep, and we use N discrete actions to substitute location-choosing operations in a continuous area. We define the policy network $\pi_{\mathbf{w}}$ with parameters $\mathbf{w}$, which receives the original image $\mathbf{I}$ as input and outputs the attacking location $\mathbf{C}$. The gradient of the objective function $J(\mathbf{w})$ with respect to $\mathbf{w}$ is shown as:

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = G \cdot \nabla_{\mathbf{w}} \log \pi(\mathbf{C}|\mathbf{I};\mathbf{w}), \tag{10}$$

where G refers to the reward of the policy. To enhance the feasibility in the physical world, we expect the locations of adversarial objects to vary, which can be quantified by the standard variance $\sigma_{\mathbf{C}}$. Therefore, G consists of two components, attack capability, and location diversity, which can be written as:

$$G = \mathcal{L}_{cls}(f_\Theta(\mathbb{R}_\delta(\mathbf{X},\mathbf{x}^{\mathbf{P}_{ori},\mathbf{C}}_{adv})),\{\mathbf{y}_k,\mathbf{b}_k\}) + \alpha \cdot \sigma_{\mathbf{C}}, \tag{11}$$

where $\mathbf{P}_{ori}$ is the initial shape of the adversarial object, and α balances the two terms. With our policy-based searching algorithm, we can jointly optimize the location and shape of our adversarial objects, enabling us to perform efficient and effective physical-world attacks.

4.3 Overall Optimization

Based on the above discussions, the overall optimization function of our attacks $\mathcal{L}$ consists of the attack loss $\mathcal{L}_{adv}$ and perceptual loss $\mathcal{L}_{per}$, which can be written as follows:

$$\mathcal{L}(\mathbf{X},\mathbf{x}_{adv};f_\Theta,\mathbb{R}_\delta) = \mathcal{L}_{adv}(\mathbf{X},\mathbf{x}_{adv};f_\Theta,\mathbb{R}_\delta) + \beta\mathcal{L}_{per}(\mathbf{x}_{adv},\mathbf{x}_{ori}), \tag{12}$$

where we append the adversarial attack loss with a perceptual loss $\mathcal{L}_{per}$ to ensure the physical feasibility of adversarial meshes, while β is a coefficient to balance the two loss functions. Inspired by [4, 49], we further introduce a total variation loss into our perceptual loss to restrict the shape change as

$$\mathcal{L}_{perc}(\mathbf{x}_{adv},\mathbf{x}_{ori}) = \frac{1}{|\mathbf{x}|} \sum_{V \in \mathbf{x}_{adv}} \sum_{v_i \in V} \sum_{v_q \in N(v_i)} ||\Delta v_i - \Delta v_q||_2^2, \tag{13}$$

where V is the vertex set of 3D adversarial meshes, Δv_i indicates the perturbation distance of a certain vertex v_i between $\mathbf{x}_{adv}$ and the original object $\mathbf{x}_{ori}$, and $N(v_i)$ refers to the vertices adjacent to v_i. The perceptual loss expects that a vertex will have similar perturbations to its neighbors, which avoids severe distortion of adversarial meshes. The pseudo-algorithm code of our *$\mathcal{X}$-Adv* can be found in Appendix A.1.

5 Experiments

5.1 Experimental Setup

Datasets. We choose the commonly-used OPIXray [47] and HiXray [43] datasets. Specifically, the OPIXray dataset has 7,109 images in the training set and 1,776 images in the test set, with five prohibited item categories (*e.g.*, Folding Knife, Straight Knife). All data are scanned from an X-ray security inspection machine to reproduce the real-world scenario in public transportation hubs. The HiXray dataset has 36,295 images in the training set and 9,069 images in the test set, with eight prohibited item categories such as lithium battery, liquid, lighter, *etc*. Images from these datasets are captured and gathered from realistic sources, which contain diverse suitcases of different sizes and shapes (*e.g.*, open trays, bags, luggage), and the prohibited items are surrounded randomly by other items of different materials (*e.g.*, clothes, phones, laptops). Thus, experiments on them could better verify the effectiveness of our attack.

Target models. To verify the effectiveness of our attacks, we train both one-stage SSD [27] and two-stage Faster R-CNN [15] to attack; we also attack the state-of-the-art and commonly used detectors in X-ray prohibited item detection scenario (DOAM [47] and LIM [43]), where we achieve similar results on clean images compared to their original papers.

Compared baselines. As discussed above, we are the first to study adversarial attacks for X-ray prohibited item detection, especially in the physical world. However, to better illustrate the superiority of our attacks, we transfer some adversarial attacks from prior works into the X-ray image scenario and compare our $\mathcal{X}$-*Adv* with them. Specifically, we use the original adversarial patch [2] (denoted as "AdvPatch") combined with our differentiable converter to generate 2D patches, which have no physical feasibility. As for 3D meshes, we apply meshAdv [49] with a certain color of the adversarial patch (denoted as "meshAdv") as a baseline. We also apply vanilla adversarial objects without shape polishment and location reinforcement (denoted as "Vanilla") to examine the capability of the attacks above. Considering the cross-task domain gap, it is reasonable to expect that these comparison methods will not perform as well as their source works on the task at hand.

Evaluation metrics. We select the most widely used metric, *i.e.*, mAP, as the main evaluation metric. The mAP value depicts the overall performance according to precision and recall values, *i.e.*, the area integral to the prediction precision ($\frac{TP}{TP+FP}$) and the prediction recall ($\frac{TP}{TP+FN}$) of object detection. Note that we set the IoU value (the intersection rate of the predicted border and the real border) as 50%. In particular, the lower mAP values indicate better attack performance. For the untargeted attack, we use mAP to evaluate the attacking performance; for the targeted attack, besides mAP, we also report the False Negative (FN) values with confidence as 0.8 (the higher the better).

Implementation details. We define the size of the adversarial object as 20×20 square pixel and the number of objects as 4, which takes around 2% of the whole image. *More details are shown in Appendix A.3.* All the codes are implemented with PyTorch. For all experiments, we conduct the training and testing on an NVIDIA GeForce RTX 2080Ti GPU cluster.

X-ray converter. We obtain the coefficient of the X-ray converter using a commercial AT6550 X-ray scanner. In practice, we have scanned 8 thicknesses (*0.2∼8mm*) of iron objects, 22 thicknesses (*1∼60mm*) of aluminum objects, and 6 thicknesses (*60∼120mm*) of plastic objects using our X-ray machine. Then we sampled their color under X-ray images. We use Eqn. 7 as the convert function, the coefficients of which are acquired from regression fitting.

5.2 Digital-world Attacks

In this part, we evaluate our $\mathcal{X}$-*Adv* in the digital world under both white-box and black-box settings. Specifically, for the white-box setting, we generate the adversarial object based on the model, then test its attacking ability on the same model; for the black-box setting, we first optimize the adversarial object on one model, then test its attack performance on other models via transfer-based attack. In more detail, we employ 4 models in the digital-world experiments including both one-stage and two-stage detectors, and the white-box attack results on OPIXray are shown in Table 1. *More results on HiXray and black-box attacks can be found in Appendix B*. From the results, we can **identify**:

❶ Despite having eliminated most of the colors and textures in the X-ray images, the adversarial attacks still pose challenges in the X-ray prohibited item detection scenario. For example, on the OPIXray dataset against DOAM, the clean mAP is 74.02%, while the mAP value drops significantly to **23.05%** after being attacked by our $\mathcal{X}$-*Adv*. It should be noted that this observation can be made for all employed models: the observed average mAP degeneration is about **50%** on OPIXray and **30%** on HiXray. Moreover, our $\mathcal{X}$-*Adv* outperforms other baselines by large margins.

❷ It should be noted that $\mathcal{X}$-*Adv* seems to fail in some categories of HiXray, *e.g.*, laptops. We hypothesize that the reason lies in the characteristic of the target object. Laptops usually occupy large proportions of an image, while our patch is much smaller than these objects. Therefore, detector models can obtain much more information about objects like laptops, which supports correct classification.

❸ Moreover, it is important to note that the vanilla patch could not successfully attack the detector, which indicates that the observed vulnerabilities of these models are not the result of poorly trained detectors.

The results for the black-box setting in Appendix B show consistent phenomena. In summary, the digital-world evaluations demonstrate that our $\mathcal{X}$-*Adv* could successfully attack the X-ray prohibited item detectors and outperform other

Table 1: Digital-world white-box attacks on OPIXray. “FO”, “ST”, “SC”, “UT”, and “MU” represent Folding Knife, Straight Knife, Scissor, Utility Knife, and Multi-tool Knife.

(a) SSD

Setting	mAP	Categories				
		FO	ST	SC	UT	MU
Clean	72.23	78.37	37.82	92.49	69.58	82.87
Vanilla	61.46	71.51	17.86	90.20	52.45	75.29
MeshAdv	52.77	61.82	10.20	83.72	40.54	67.59
AdvPatch	40.91	47.19	5.86	74.83	25.48	51.21
$\mathcal{X}$-Adv	**19.20**	**24.11**	**1.46**	**44.48**	**12.59**	**13.37**

(b) Faster R-CNN

Setting	mAP	Categories				
		FO	ST	SC	UT	MU
Clean	64.92	60.90	37.19	89.74	66.82	69.96
Vanilla	53.05	53.13	20.75	85.69	49.76	55.93
MeshAdv	49.49	44.26	17.48	81.70	44.03	59.99
AdvPatch	50.19	52.67	15.88	84.03	42.26	56.13
$\mathcal{X}$-Adv	**23.33**	**26.62**	**3.44**	**62.91**	**15.33**	**8.36**

(c) DOAM

Setting	mAP	Categories				
		FO	ST	SC	UT	MU
Clean	74.02	78.92	40.88	95.65	74.08	80.55
Vanilla	67.79	74.26	32.57	91.37	63.41	77.34
MeshAdv	56.36	60.09	23.04	86.87	47.11	64.68
AdvPatch	42.04	45.57	9.41	81.19	26.44	47.60
$\mathcal{X}$-Adv	**23.05**	**18.40**	**4.05**	**64.80**	**18.57**	**9.45**

(d) LIM

Setting	mAP	Categories				
		FO	ST	SC	UT	MU
Clean	73.07	79.01	36.04	94.73	72.94	82.62
Vanilla	66.44	73.58	22.78	93.08	65.17	77.62
MeshAdv	59.60	65.56	19.70	87.27	52.26	73.20
AdvPatch	49.69	54.16	14.66	80.35	35.72	63.55
$\mathcal{X}$-Adv	**22.46**	**31.64**	**4.28**	**52.59**	**16.65**	**7.13**

baselines by large margins.

5.3 Physical-world Attacks

In this part, we further investigate the X-ray prohibited item detection model robustness in the physical-world setting.

We first illustrate the **attack pipeline** for our physical-world attacks (see Fig 3). In detail: ❶ we first generate adversarial objects using our *$\mathcal{X}$-Adv* based on a white-box pre-trained DOAM target model; ❷ we then transform the adversarial objects from 3D mesh format into STL format so that we can use a third-party 3D printer to print these 3D objects in metal; ❸ we then put our adversarial objects into the fabric/plastic box with other items and employ several workers to scan them into X-ray images using a commercial AT150180B X-ray scanner (which is commonly used in the train station and airport security checkpoints); ❹ finally, we test our physical-world adversarial examples (X-ray images) on black-box X-ray prohibited item detection models, specifically, DOAM, LIM, and Faster R-CNN, which are trained on the physical-world dataset proposed in Section 7. Note that, we use the commercial X-ray scanners but cannot use their detection backend because these models/strategies are business secrets, however, we adopt a similar black-box Faster R-CNN. During the experiments, we have no access to and prior knowledge of these target detectors and X-ray scanners.

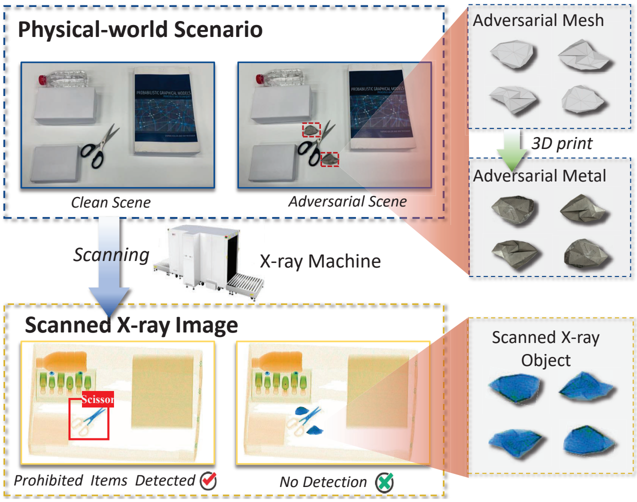


Figure 3: Illustration of the physical-world attacking pipeline. $\mathcal{X}$-adv first generates adversarial meshes (3D objects); we then print these meshes into metal objects using 3D printers; when scanned by X-ray scanners, these metal objects will become adversarial patches in the resulting X-ray images.

Specifically, we use *$\mathcal{X}$-Adv* to generate 16 adversarial metal objects and then print them in `iron` using a 3D printer. We collect items (*e.g.*, laptops, headphones, bags) from our staff and students under their grants. In total, we collected 80 adversarial X-ray images as the test set; some physical-world clean and adversarial X-ray images can be found in Figure 4. Note that all X-ray images are collected without personal information to avoid privacy leakage. To assess the real-world feasibility of our attacks, in addition to having our X-Adv search for the best possible attack location (denoted as “Physical best”), we also impose two attack settings to better simulate the physical-world dynamic environment of items movement in luggage: (1) slight transformations and (2) random placement. Specifically, slight transformations add shift (random 10 pixels in each direction) and rotation (-30°∼+30°) to adversarial objects (denoted as “Physical change”), while random placement randomly places the adversarial objects in the entire suitcase (denoted as “Physical random”). For better comparison, we also provide the results of the 80 images using digital-world attacks (denoted as “Digital attack”) and the physical-world results on clean examples (denoted as “Clean”).

From Table 2, we can conclude that the physical attacking

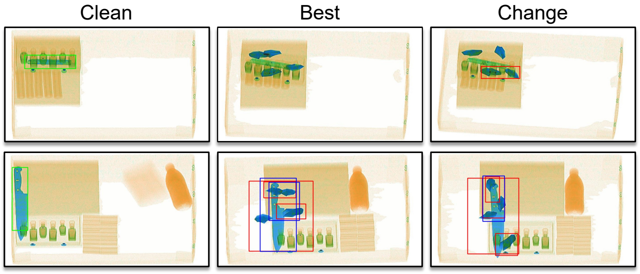


Figure 4: Detection results of some X-ray images in our physical-world experiments (we choose images with fewer items for better visualization). Green boxes indicate correct classes and suitable locations; blue boxes represent correct classes in incorrect locations; red boxes indicate incorrect classes. We only show detection boxes with confidence >10%.

Table 2: Physical-world attack experiments on different detection models. More results are shown in Appendix B.

(a) DOAM

Setting	mAP	Categories			
		SC	FO	ST	UT
Clean	91.35	84.17	98.05	100.00	83.18
Digital attack	30.28	67.54	2.15	50.73	0.69
Physical best	33.16	66.33	18.35	44.48	3.46
Physical change	50.97	74.13	42.19	55.92	31.63
Physical random	76.17	76.06	79.19	85.33	64.10

(b) Faster R-CNN

Setting	mAP	Categories			
		SC	FO	ST	UT
Clean	95.35	94.00	100.00	92.66	94.75
Digital attack	27.18	44.77	0.31	50.63	13.00
Physical best	24.67	62.88	2.26	23.03	10.53
Physical change	57.38	85.84	35.45	72.16	36.07
Physical random	75.57	93.00	56.03	88.95	64.29

ability of the proposed $\mathcal{X}$-*Adv* has a significant impact on detection accuracy, *i.e.*, the mAP value of DOAM on physical clean samples is 91.35%, while the mAP values on the sampled adversarial samples are 33.16% on "Physical best", 50.97% on "Physical change", and 76.17% on "Physical random", which are lower than the results on the clean counterparts. This observation also indicates that the safety problem of X-ray prohibited item detection is worth studying from a practical perspective. Moreover, we also observe that the attacking ability of "Physical best" is stronger than that of "Physical change" (lower mAP), which supports our motivation to search for the critical attack position. Furthermore, compared to the digital-world attack results, the physical-world attack results are weaker; we speculate this is because of the digital-physical domain gap [13, 46].

5.4 Ablation Studies

In this section, we investigate the key factors that might impact the attack ability of our $\mathcal{X}$-*Adv* , thereby providing comprehensive insights and promoting a deeper understanding of our strategy. In brief, we conduct thorough ablations on several factors. All the experiments conducted in this part use the DOAM target model on OPIXray and HiXray datasets.

Attack locations. Here, we investigate three additional location-searching strategies on the attack performance, *i.e.*, fixed position (denoted as "Fix"), random positions (denoted as "Random"), and greedy-search-based positions (denoted as "Greedy"). Our proposed attacking location search strategy is denoted as "Reinforce". For Fix, we place the adversarial objects on the corners of the prohibited items; for Random, we place the adversarial objects randomly around the prohibited items; for Greedy, we first greedy-search the strongest attack locations that maximize $\mathcal{L}_{cls}$ by placing one original object at each location, then optimize the adversarial objects at the corresponding locations. The experimental results on OPIXray and HiXray can be found in Table 3, where we can observe that among all 4 attacking location searching strategies, the result under our "Reinforce" setting shows the strongest adversarial attacking performance.

Moreover, we study a more limited setting where the adversary could stick an adversarial object on the prohibited item. In particular, we add experiments on OPIXray against DOAM, where we put a 32×32 iron rectangle or a 40×40 adversarial object generated by $\mathcal{X}$-*Adv* on top of the target object. The attack performance (49.48 mAP and 25.28 mAP) is still worse than our original position-searching strategy (23.05 mAP). Note that directly placing the target object into an iron box or hiding it with iron plates could make it disappear, which can be easily identified in practice. Meanwhile, this would also violate the definition of adversarial attacks (cover the salient parts and change its semantics). The goal of this experiment is to illustrate the importance of location-searching.

The above studies demonstrate that avoiding overlap and occlusion can increase the attack capability.

Number of objects. Regarding the number of adversarial objects, we study whether the attack ability of the adversarial objects differs when this number is changed. Thus, we set the number of the adversarial objects to 1, 2, 4, and 8 respectively, while keeping the total area of each setting the same. The results can be found in Figure 5. As the results show, more objects usually result in better attack performance. However, too many adversarial objects will introduce an additional cost in terms of physical feasibility. In practice, we set the number of objects as 4.

5.5 Discussions and Analysis

In this part, we provide more detailed discussions and analysis on the attack ability and physical feasibility of $\mathcal{X}$-*Adv* . All

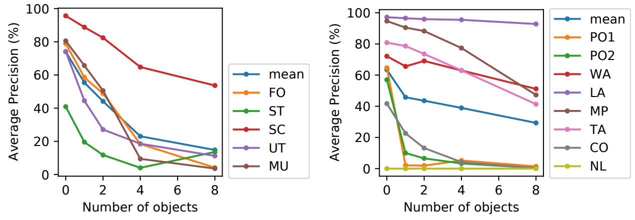
Table 3: Ablation studies on different attack locations. Our strategy achieves the best attack performance.

(a) OPIXray

Setting	mAP	Categories				
		FO	ST	SC	UT	MU
Fix	51.64	55.54	18.22	82.16	39.89	62.38
Random	38.11	40.54	8.39	76.77	26.82	38.01
Greedy	29.38	28.02	5.02	65.46	20.21	28.19
Reinforce	**23.05**	**18.40**	**4.05**	**64.80**	**18.57**	**9.45**

(b) HiXray

Setting	mAP	Categories							
		PO1	PO2	WA	LA	MP	TA	CO	NL
Fix	44.68	10.48	8.95	69.06	96.42	88.76	74.69	9.04	**0.00**
Random	41.98	8.41	6.37	66.05	95.74	82.74	68.63	7.93	**0.00**
Greedy	40.19	5.77	4.14	64.88	95.47	80.44	65.76	5.06	**0.00**
Reinforce	**38.96**	**5.21**	**3.33**	**63.00**	**95.49**	**77.38**	**63.05**	**4.22**	**0.00**



(a) OPIXray (b) HiXray

Figure 5: Ablations on the numbers of adversarial objects.

the experiments conducted in this part are using the DOAM target model based on OPIXray and HiXray datasets.

Object materials. Different materials are rendered on X-ray images in different colors; therefore, it is necessary to investigate their possible influence on the final attack ability of the generated adversarial objects. To this end, we select three kinds of colors (materials), namely blue, green, and orange, which respectively correspond to three materials that commonly appear in the luggage, *i.e.*, iron, aluminum, and plastic. The results on OPIXray can be found in Figure 7(a). It is clear that the generated adversarial objects with blue colors (iron) show stronger attack ability, *i.e.*, lower mAP values. For instance, on the OPIXray dataset, the mAP value of the iron adversarial objects is **23.05%**, while that of the green (aluminum) ones is 55.44%, and that of the orange (plastic) ones is 55.61%. We believe that this observation is reasonable since prohibited items (such as knives and guns) tend to be made of metal, thus adversarial objects made of similar materials (and rendered in similar colors) will have a higher attack ability. On the OPIXray dataset, the prohibited items are knives, meaning that blue (iron) outperforms other colors significantly. *Results on HiXray can be found in Appendix B.*

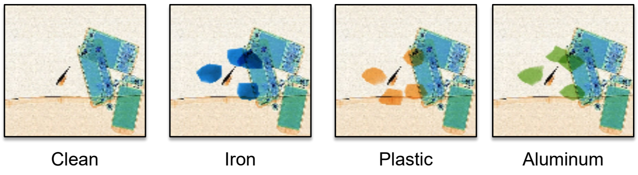
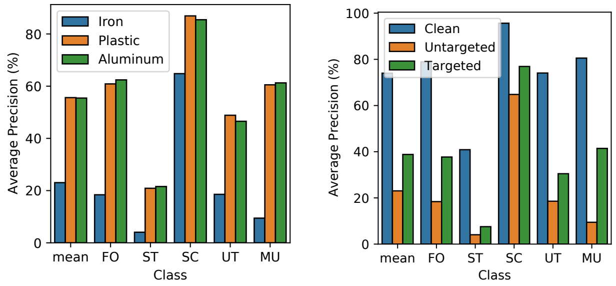


Figure 6: Visualization of adversarial objects with different materials/colors (*i.e.*, blue, orange, and green).



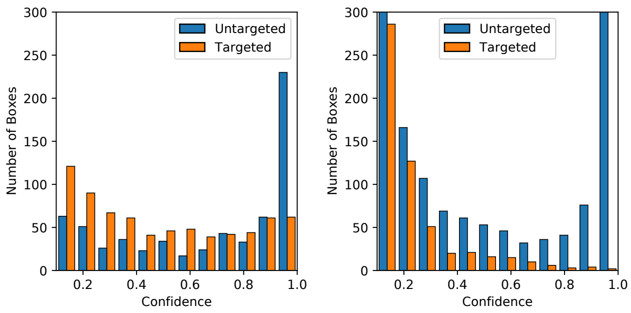
(a) Materials (b) Targeted/untargeted

Figure 7: Results using DOAM on OPIXray: (a) different materials, (b) targeted and untargeted adversarial attacks.

Targeted adversarial attacks. As shown in Eqn. 12, our $\mathcal{X}$-*Adv* maximizes the classification loss $\mathcal{L}_{cls}$ of the detector to output wrong classes, which is the *untargeted attacks*. In addition to the untargeted attack, we here further design another reasonable attack strategy, *i.e.*, perform attacks to evade the detector by confusing the predictor to classify the object of interest as `Background`. Thus, we apply *targeted attacks* and set the target label of attacks to `Background` (one of the classes for detection). Specifically, we substitute $\mathcal{L}_{cls}$ with a cross-entropy loss between the confidence of all predicted boxes and the background class. Since the background is the 0-th class of object detectors, performing attacks that mislead all boxes into the background class can also reduce the number of predicted boxes.

As shown in Figure 7(b), in terms of mAP, the performance of targeted attacks is weaker than that of untargeted attacks (38.82 v.s. 23.05). However, in terms of FN bounding box numbers, the performance of targeted attacks outperforms untargeted attacks largely (1632 v.s. 1274). These results demonstrate the different adversarial goals for targeted and untargeted attacks, and the targeted attack in the X-ray security inspection scenario might be more meaningful. We conjecture the main reasons for the above observations are as follows: ❶ It is more difficult for targeted attacks to reduce the mAP values in general object detection [28]. ❷ As the distribution of all bounding boxes shown in Figure 8, the untargeted attacks produce more false bounding boxes (FP) with high confidence, which could help to reduce the precision of detectors. However, targeted attacks result in fewer FP boxes, and this will help to prevent the detection of prohibited items, but the overall precision will not be too low.

Unseen prohibited items. Moreover, we are interested in discovering the potential of $\mathcal{X}$-*Adv* for attacks on other prohibited items with unseen materials/shapes. In other words,



(a) TP (b) FP

Figure 8: The distribution of TP and FP bounding boxes under different targeted and untargeted adversarial attacks. "TP" represents True Positive, while "FP" denotes False Positive.

adversarial objects are first generated on specific types of prohibited items, and then we use them to directly attack other unseen prohibited item types without re-training. Here, we first conduct experiments in the digital world, where we train a group of adversarial objects using DOAM on HiXray (unseen prohibited items: lighter, liquids, *etc*) and then directly test them against another DOAM on OPIXray (prohibited items are knives). Overall, our attack achieves 29.40 mAP which is slightly lower than the original $\mathcal{X}$-*Adv* attack (23.05 mAP) that is trained on OPIXray. We then verify this in the physical world, where we place the adversarial objects around their unseen prohibited items, and we achieve 45.52 mAP, which is also slightly lower than the original $\mathcal{X}$-*Adv* attack (33.16 mAP). The above results indicate that our attack could still work for other unseen materials/shapes without re-training.

6 Countermeasures against $\mathcal{X}$-*Adv*

In this section, we propose three possible defenses and evaluate our $\mathcal{X}$-*Adv* against them.

Data Augmentation. Data augmentation has been identified as a popular approach for improving model robustness [36]. In light of this, we introduce the data augmentation strategy as the first countermeasure to mitigate our adversarial attacks. Given the special feature space of X-ray image recognition (*i.e.*, limited colors and textures), we believe that introducing additional adversarial-object-like patches might be beneficial for improving the robustness of the X-ray prohibited item detection models. Specifically, for each image, we randomly add 1-4 blue or orange patches and mix the clean examples with the additional examples during training using a ratio of 1 : 1. The results can be found in Table 3(a). Here, "V+C" denotes that the detector is trained **without** additional examples and tested **on** clean examples, "V+A" denotes that the detector is trained **without** additional examples and tested **on** adversarial examples, "D+C" denotes that the detector is trained **with** additional examples and tested **on** clean examples and "D+A" denotes that the detector is trained **with** additional examples and tested **on** adversarial examples. It can be observed that the data augmentation can to a certain extent effectively defend the proposed $\mathcal{X}-$adv method for X-ray prohibited item detection.

Table 4: Countermeasure studies. (a) "V" and "D" denote vanilla training or data augmentation; "C" and "A" refer to testing on clean or adversarial examples; (b) We first generate adversarial examples by meshAdv on DOAM/LIM and train the classifier; we then test the detection performance on $\mathcal{X}$-*Adv* generated on DOAM. ACC denotes classification accuracy, and AUC is the area under the ROC curve; (c) We adversarially train a prohibited item detector using RobustDet.

(a) Data augmentation

Setting	mAP	Categories				
		FO	ST	SC	UT	MU
V+C	74.06	78.75	40.90	95.66	73.56	81.42
V+A	23.05	18.40	4.05	64.80	18.57	9.45
D+C	73.94	79.44	40.52	93.82	73.40	82.54
D+A	46.69	49.06	17.05	81.21	39.68	46.46

(b) Adversarial detection

	DOAM→DOAM		LIM→DOAM	
	ACC	AUC	ACC	AUC
OPIXray	62.66	97.99	56.66	96.53
HiXray	76.73	97.95	74.72	98.91

(c) Adversarial Training

AT Setting	Attack	mAP	Categories				
			FO	ST	SC	UT	MU
PGD	Clean	73.74	77.06	37.86	94.39	72.78	86.61
	$\mathcal{X}$-*Adv*	22.09	20.19	1.36	66.17	17.39	5.32
$\mathcal{X}$-*Adv*	Clean	73.49	78.21	40.77	93.23	73.58	81.64
	$\mathcal{X}$-*Adv*	53.47	55.82	20.26	84.43	49.02	57.82

Adversarial Detection. Another prevailing approach to improving model robustness is adversarial detection. Rather than correctly detecting the target item under the adversarial scenario, adversarial detection aims to detect the existence of adversarial examples [5, 9, 14, 29]. Here, we build a neural classifier capable of distinguishing images containing adversarial objects from clean X-ray images. Specifically, we use a ResNet50 model as a classifier and trained on adversarial examples generated by meshAdv on different models (*i.e.*, DOAM and LIM). We then test the classifier on adversarial examples generated by $\mathcal{X}$-*Adv* on a different DOAM model. The training set and test set of the classifier do not overlap. The results in Table 4(b) indicate that the adversarial examples generated by different methods are quite different and the neural classifier fails to generalize.

Adversarial Training. We choose adversarial training (AT) as the last countermeasure for $\mathcal{X}$-*Adv*. Although AT for image classification has been widely studied [25, 30], only some preliminary studies have been devoted to object detection [6, 10, 52]. Here, we adopt RobustDet [10] as the AT method and use an SSD detector with a backbone of VGG-16. Specifically, we adversarially train two detectors using adversarial examples generated by (1) PGD attacks or (2) our $\mathcal{X}$-*Adv*. Note that the generated adversarial objects by $\mathcal{X}$-*Adv* during AT are different from those for testing. From Table 3(b), we could observe that ❶ AT trained on PGD attacks show limited defense against $\mathcal{X}$-*Adv* mainly due to the differences between perturbations and patch/object attacks; ❷ AT trained on $\mathcal{X}$-*Adv* could mitigate our $\mathcal{X}$-*Adv* attacks to a certain extent; and ❸ compared to classification, it is still comparatively difficult to adopt AT on object detection tasks.

Summary and Discussion. Despite facing significant challenges launched by our attack, the proposed defenses could still mitigate its negative influence to some extent. Specifically, for the strongest countermeasure (AT trained on our $\mathcal{X}$-*Adv*), we could significantly improve the model robustness and achieve 53.47 mAP against $\mathcal{X}$-*Adv* attacks. Meanwhile, we could observe that the proposed countermeasures have rather high practical feasibility for defenders mainly due to ❶ the adversarial-object-like patches of data augmentation and adversarial attacks of adversarial detection can be easily generated and obtained for model training; ❷ $\mathcal{X}$-*Adv* adversarial objects of adversarial training can be generated either based on our open-sourced codes or the provided adversarial objects in the XAD dataset. For the physical-world implementation, defenders could simply employ a 3D printer to print out the 3D objects based on our guidelines. Moreover, defenders could combine adversarial detection with an AT model, which could further mitigate the negative impacts of $\mathcal{X}$-*Adv* attacks. *The feasibility of data augmentation and* $\mathcal{X}$-*Adv AT are verified under the physical-world setting in Appendix B.*

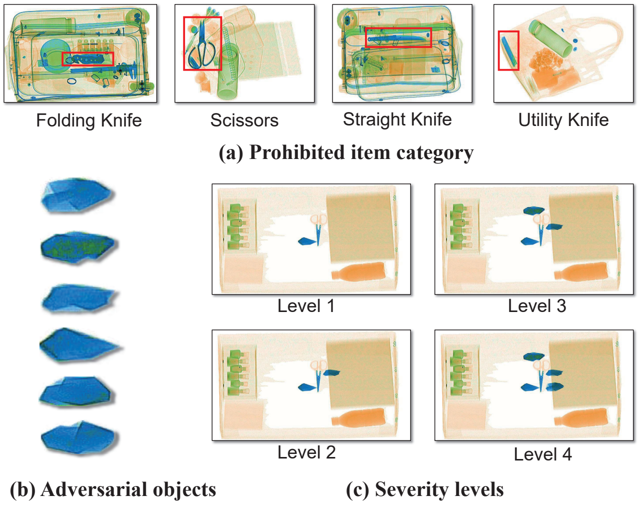
7 Physical-world X-ray Attack Dataset

A dataset is significantly beneficial for boosting research, especially for areas where professional benchmarks are lacking or the data collection is expensive. As we have observed in Section 5.3, physical-world X-ray detectors are vulnerable to our attacks. Therefore, we further present a physical-world X-ray inspection security robustness evaluation dataset to promote the design of robust X-ray prohibited item detectors.

7.1 Construction Details

We first introduce the construction process of our Physical-world X-ray Attack Dataset (XAD), including the data collection, category selection, and quality control.

Data collection. We exploit one advanced X-ray security inspection machine, AT150180B, to generate the X-ray images in our dataset. We first randomly place the objects in the plastic/fabric box to mimic the similar environment in the real-world scenario; we then send these boxes through the security inspection machine, after which the machine outputs the X-ray images. To prevent privacy leakage, all images are collected legally: items are borrowed from our staff members and students, and do not contain personal information.



(a) Prohibited item category

(b) Adversarial objects

(c) Severity levels

Figure 9: Illustration of the images in our XAD. (a) illustrates the prohibited item categories; (b) denotes the X-ray images of physical-world adversarial objects; (c) denotes the different severity levels in the testing set.

Category selection. As shown in Figure 9, we select 4 categories of prohibited items (cutters, scissors, folding knives, straight knives, and utility knives) that frequently appear in daily life. The 4 categories of cutters have different shapes and scales, which meets the category selection diversity requirements. The sufficient numbers of instances can provide a more credible evaluation for various models.

Quality control procedure. We followed a similar annotation quality control procedure to the famous vision dataset Pascal VOC [11]. All annotators followed the same annotation guidelines, including what to annotate, how to annotate bounding, how to treat occlusion, *etc*. Moreover, to ensure the accuracy of annotation, we divided the annotators into 3 groups. All images were randomly assigned to 2 out of the 3 groups for annotation, after which a final group was specially organized for confirmation.

7.2 Data Properties

Subset division. Our XAD contains two subsets, *i.e.*, a training set with clean X-ray images and a testing set with physical-world adversarial attacks. The 4,537 images in the training set simulate the real-world scenario to help models achieve satisfactory generalization performance. For the testing set, we follow [18] and generate adversarial images from 210 clean X-ray images with 4 different severity levels (0 to 4, where "0" denotes clean images). Specifically, for each item layout, we first place all the items in the box and X-ray scanned them to obtain a clean image; we then place 1∼4

Table 5: Detailed data properties of our XAD dataset.

(a) Quality distribution

Category	Scissor	Folding knife	Straight knife	Utility knife
Training	1,048	1,300	1,300	926
Testing	54	54	52	50
Total	2,002	1,354	1,352	976

(b) Object materials and X-ray image colors

Colors	Materials	Typical examples
Orange	Organic Substances	Plastics, Clothes
Blue	Inorganic Substances	Irons, Coppers
Green	Mixtures	Edge of phones

Table 6: Results on different levels of XAD.

Setting	mAP	Categories			
		SC	FO	ST	UT
Level 0	91.74	96.29	86.98	84.86	98.84
Level 1	72.98	79.25	61.32	69.30	82.04
Level 2	50.10	66.47	33.79	60.84	39.29
Level 3	30.83	55.76	18.59	41.15	7.82
Level 4	27.50	53.63	15.19	35.17	6.00

adversarial metals in the box respectively and scanned them to collect 4 versions of adversarial images with 4 severity levels. Thus, our testing set contains 1,050 samples which are all scanned from a real X-ray machine. See Fig 9 for samples.

Category distribution. Our XAD dataset contains 4,537 images and 4 categories of 4,830 instances with bounding-box annotations of prohibited cutters.

Color Information. The colors of objects under X-ray are determined by their chemical composition, mainly reflected in the material, which is introduced in Table 5(b).

Instances per image. In the training set of our XAD, each image contains at least one prohibited object. In particular, the image numbers containing 1, 2, 3, and 4 prohibited objects are 4069, 234, 23, and 1, respectively.

7.3 Preliminary Experiments on XAD

After introducing our XAD, we further conduct experiments on XAD to demonstrate the difficulties and practicability of maintaining robust detectors. Specifically, we use the DOAM model for detection. We train the model on the training set of XAD and then evaluate it on the test set. The implementation details are the same as our main experiment. From Table 6, we can make several observations: ❶ The detector shows weak performance on our XAD dataset with the model's performance on prohibited item recognition reducing by as much as 60% in terms of mAP. ❷ Increasing the number of adversarial objects improves the attack and therefore increases the recognition difficulties in this scenario. We encourage researchers to design stronger training strategies or defense modules and evaluate their robustness on this benchmark.

8 Conclusion and Future Work

This paper takes the first step to study physical-world adversarial attacks for X-ray prohibited item detection. Specifically, we propose $\mathcal{X}$-*Adv* , which generates physically realizable adversarial objects to circumvent the color fading and complex occlusion problems in this scenario. Although the results presented here are promising, there are several research directions that we are interested in exploring in the future. ❶ We hope $\mathcal{X}$-*Adv* can be used as a tool to better debug and understand the nature of object detectors' robustness. ❷ We would like to generate attacks in other soft materials which are more stealthy. ❸ We are interested in attacks against more types of prohibited items. ❹ Our $\mathcal{X}$-*Adv* can be regarded as a general attacking framework for visually constrained scenarios. In this paper, we focus on attacking X-ray inspection scenarios; we will further extend our attacks to other complex scenarios.

9 Ethics Statement

As an effective way to discover safety problems, adversarial attacks will encourage researchers to pay more attention to model robustness. Based on this, this paper proposes $\mathcal{X}$-*Adv* to attack X-ray prohibited item detectors. Our large-scale experiments demonstrated that existing X-ray prohibited item detection models (even commercial systems) are not infallible and can still be easily deceived. All experiments are conducted on public-available datasets, and all images for physical-world attacks are collected legally from our staff and students without personal information under their grants.

To mitigate potential real-world impacts of the attacks, this paper ❶ proposes three countermeasures and discusses their practical feasibility for defending $\mathcal{X}$-*Adv* ; ❷ presents the physical-world X-ray attack dataset XAD to promote the design and re-training of stronger detectors; and ❸ disclose the results, countermeasures, and resources to two relevant X-ray security inspection service providers and a stakeholder user at the airport checkpoint. Based on our easy-to-use countermeasures, we help them to recognize this critical security issue and move the first step to improve the robustness of their detection backend with adversarial training. Moreover, these service providers are also suggested to utilize the white-box $\mathcal{X}$-*Adv* to help reveal the vulnerabilities of their detectors and further design stronger models. Despite the threats identified in this paper, we should note that a real-world adversary would still find it difficult to pass X-ray security inspection systems carrying prohibited items without detection, as human inspectors still help with checking luggage. We thus further suggested the airport checkpoint pay attention to the employment of human inspectors, which can relieve the concerns on the potential negative abuses of *X-Adv*.

Acknowledgement

The authors would like to thank the shepherd and all the anonymous reviewers for their tremendous efforts and insightful comments during reviewing. This work was supported by the National Key Research and Development Plan of China (2021ZD0110601), the National Natural Science Foundation of China (No. 62022009 and 62206009), the State Key Laboratory of Software Development Environment (SKLSDE-2022ZX-23), and the grants from SenseTime.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[3] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *IEEE SP*, pages 176–194. IEEE, 2021.

[4] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *ACM CCS*, pages 2267–2281, 2019.

[5] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

[6] Pin-Chun Chen, Bo-Han Kung, and Jun-Cheng Chen. Class-aware robust adversarial training for object detection. In *CVPR*, pages 10420–10429, 2021.

[7] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018.

[8] Kenneth T Co, Luis Muñoz-González, Leslie Kanthan, Ben Glocker, and Emil C Lupu. Universal adversarial robustness of texture and shape-biased models. In *ICIP*, pages 799–803. IEEE, 2021.

[9] Zhijie Deng, Xiao Yang, Shizhen Xu, Hang Su, and Jun Zhu. Libre: A practical bayesian approach to adversarial detection. In *CVPR*, pages 972–982, 2021.

[10] Ziyi Dong, Pengxu Wei, and Liang Lin. Adversarially-aware robust object detector. *arXiv preprint arXiv:2207.06202*, 2022.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, June 2018.

[13] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *CVPR*, pages 1625–1634, 2018.

[14] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

[15] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[18] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[19] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, and Tara N. Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 2012.

[20] Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. Assessing shape bias property of convolutional neural networks. In *CVPR Workshops*, pages 1923–1931, 2018.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[22] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[23] Aishan Liu, Tairan Huang, Xianglong Liu, Yitao Xu, Yuqing Ma, Xinyun Chen, Stephen Maybank, and Dacheng Tao. Spatiotemporal attacks for embodied agents. In *ECCV*, 2020.

[24] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *AAAI*, volume 33, pages 1028–1035, 2019.

[25] Aishan Liu, Xianglong Liu, Chongzhi Zhang, Hang Yu, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE Transactions on Image Processing*, 2021.

[26] Shunchang Liu, Jiakai Wang, Aishan Liu, Yingwei Li, Yijie Gao, Xianglong Liu, and Dacheng Tao. Harnessing perceptual adversarial patches for crowd counting. In *ACM CCS*, 2022.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016.

[28] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018.

[29] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[31] Edwin C McCullough, Hillier L Baker Jr, O Wayne Houser, and David F Reese. An evaluation of the quantitative and radiation features of a scanning x-ray transverse axial tomograph: the emi scanner. *Radiology*, 111(3):709–715, 1974.

[32] 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.

[33] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016.

[34] Keaton Mowery, Eric Wustrow, Tom Wypych, Corey Singleton, Chris Comfort, Eric Rescorla, J Alex Halderman, Hovav Shacham, and Stephen Checkoway. Security analysis of a full-body scanner. In $\{$*USENIX*$\}$ *Security Symposium*, pages 369–384, 2014.

[35] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE EuroS&P*, pages 372–387. IEEE, 2016.

[36] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

[37] Baifeng Shi, Dinghuai Zhang, Qi Dai, Zhanxing Zhu, Yadong Mu, and Jingdong Wang. Informative dropout for robust representation learning: A shape-bias perspective. In *ICML*, pages 8828–8839. PMLR, 2020.

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[39] Shiyu Tang, Ruihao Gong, Yan Wang, Aishan Liu, Jiakai Wang, Xinyun Chen, Fengwei Yu, Xianglong Liu, Dawn Song, Alan Yuille, Philip H.S. Torr, and Dacheng Tao. Robustart: Benchmarking robustness on architecture design and training techniques. *https://arxiv.org/pdf/2109.05211.pdf*, 2021.

[40] Renshuai Tao, Hainan Li, Tianbo Wang, Yanlu Wei, Yifu Ding, Hongping Zhi Bowei Jin and, Xianglong Liu, and Aishan Liu. Exploring endogenous shift for cross-domain detection: A large-scale benchmark and perturbation suppression network. In *CVPR*, 2022.

[41] Renshuai Tao, Tianbo Wang, Ziyang Wu, Cong Liu, Aishan Liu, and Xianglong Liu. Few-shot x-ray prohibited item detection: A benchmark and weak-feature enhancement network. In *ACM International Conference on Multimedia*, 2022.

[42] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu*. Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In *ICCV*, 2021.

[43] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu. Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In *CVPR*, pages 10923–10932, 2021.

[44] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[45] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. *arXiv preprint arXiv:2108.07020*, 2021.

[46] Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *CVPR*, pages 8565–8574, 2021.

[47] 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 *ACM International Conference on Multimedia*, pages 138–146, 2020.

[48] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[49] Chaowei Xiao, Dawei Yang, Bo Li, Jia Deng, and Mingyan Liu. Meshadv: Adversarial meshes for visual recognition. In *CVPR*, pages 6898–6907, 2019.

[50] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *CVPR*, 2020.

[51] Chongzhi Zhang, Aishan Liu, Xianglong Liu, Yitao Xu, Hang Yu, Yuqing Ma, and Tianlin Li. Interpreting and improving adversarial robustness with neuron sensitivity. *IEEE Transactions on Image Processing*, 2020.

[52] Haichao Zhang and Jianyu Wang. Towards adversarially robust object detection. In *CVPR*, pages 421–430, 2019.

[53] Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. *arXiv preprint arXiv:1905.09797*, 2019.

[54] Xiaopei Zhu, Zhanhao Hu, Siyuan Huang, Jianmin Li, and Xiaolin Hu. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In *CVPR*, pages 13317–13326, 2022.

[55] Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, and Xiaolin Hu. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *AAAI*, volume 35, pages 3616–3624, 2021.

[56] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

Appendix

A Implementation Details

A.1 Pseudo Code of $\mathcal{X}$-*Adv* Algorithm

Algorithm 1 $\mathcal{X}$-*Adv* Algorithm

```
Input: X-ray image $\mathbf{I}$, label $\mathbf{y}$ and bounding box $\mathbf{b}$ of the prohibited item, initial state of meshes $\mathbf{x}_{ori}$, maximum location reinforcement iterations $N_{\mathbf{C}}$, and maximum shape polishment iterations $N_{\mathbf{P}}$.
Output: Adversarial example $\mathbf{I}_{adv}$, adversarial meshes $\mathbf{x}_{adv}$.
Initialize policy network $\pi$.
for $i$ in $N_{\mathbf{C}}$ do
    Sample a location $\mathbf{C}$ in $\pi(\mathbf{I};\mathbf{w})$.
    Calculate reward $G$ according to Eqn. 11.
    Update policy network $\pi$ according to Eqn. 10.
end for
Initialize $\mathbf{x}_{adv}$ as $\mathbf{x}_{ori}$.
for $i$ in $N_{\mathbf{P}}$ do
    Generate $\mathbf{I}_{adv} = \mathbf{I} \odot g_m(d(\mathbf{x}_{adv}^{\mathbf{P},\mathbf{C}}))$.
    Calculate cost function $\mathcal{L}$ according to Eqn. 12.
    Update $\mathbf{x}_{adv}$ by Adam optimizer.
end for
```

A.2 X-ray Converter

To obtain the coefficient of the X-ray converter, we have scanned a group of objects with different types of materials and thicknesses. The photo of the scanned objects and their X-ray images is provided in Figure A.1. For every thickness and material, we sample an area of 20×20 pixels and calculate the average pixel value as the color for this specific thickness and material. We use Eqn. 7 as the conversion function, fitting the coefficient a, b, c to the depth and color pairs. The illustration of regression curves for `iron` is shown in Figure A.2.

The goodness of fit R^2 for Hue, Saturate, and Value is 0 (Hue is a constant value), 0.993, and 0.984, which demonstrates that the proposed conversion function fits well.

A.3 Detailed Settings of Experiments

Training of target models. All models in our paper (*i.e.*, SSD, DOAM, LIM, and Faster R-CNN) use VGG-16 as back-

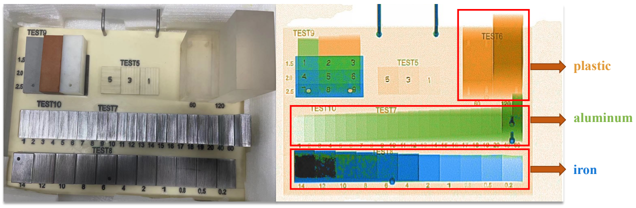


Figure A.1: Physical-world photos and X-ray images of scanned objects with different materials and thicknesses. The number under the objects denotes the thickness of the object in millimeters.

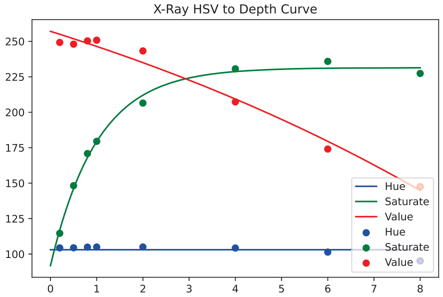


Figure A.2: Regression curves of depths to HSV values for `iron` material. We sample 8 different thicknesses from 0.2mm to 8mm.

bones. We use pre-trained weights on the VOC0712 dataset and fine-tune our model on them, which is applied by previous works [43, 47] to reduce the training time. We use the SGD optimizer with momentum 0.9 and weight decay 5×10^{-4}. For SSD, DOAM, and LIM, we train them with batch size 24 and a learning rate of 0.0001. For Faster R-CNN, we train them with batch size 1 and a learning rate of 0.001. All models are trained with a maximum of 100 epochs, and we select the checkpoints with the highest mAP as our target models.

***$\mathcal{X}$-Adv* .** We adopt an Adam optimizer with a learning rate of 0.1 and a maximum of 24 iterations to optimize the adversarial loss. To accelerate the speed of attacks, we set the batch size of the attack as 10, which means that every 10 images share the same group of adversarial objects. Experimental results have proven the viability of this approach. The initial shape of the 3D object is a sphere with 26 vertices and 48 faces. During optimization, the coordinates of vertices are updated with the guidance of the gradients, while the adjacent relation of vertices remains unchanged. The coefficient of location variance α is 0.05; the coefficient of perceptual loss β is 0.1 in SSD, DOAM, and LIM, and 0.01 in Faster R-CNN. We train the REINFORCE policy for 200 iterations for every batch. For the available attack area $\mathbb{C}$, we define $(x_{min}, y_{min}), (x_{max}, y_{max})$ as the coordinates of a ground truth box, and w, h is the width and height of the box. Adversarial objects should have no overlap with the center area of the ground truth box, *i.e.*, the area of $(x_{min} + 0.25w, y_{min} + 0.25h), (x_{max} - 0.25w, y_{max} - 0.25h)$.

Vanilla adversarial objects. We simply set the optimizing epoch to 0 to get vanilla spheres as adversarial objects.

Table A.1: Time consumption of our *$\mathcal{X}$-Adv* on OPIXray dataset for the SSD detector. We perform 24 iterations of shape polishment and 200 iterations of location reinforcement, which is consistent with other experiments. We show the time consumption of one iteration and one batch.

Time Cost (s)	Shape Polishment	Location Reinforcement
Iteration	0.51	0.28
Batch	12.15	55.93

Table B.1: Performance comparison between different materials on HiXray dataset.

Setting	mAP	Categories							
		PO1	PO2	WA	LA	MP	TA	CO	NL
Plastic	55.22	43.10	39.12	64.01	96.48	91.79	77.15	30.09	**0.00**
Aluminum	43.79	23.11	24.27	**51.41**	**94.44**	83.11	72.67	**1.33**	**0.00**
Iron	**38.96**	**5.21**	**3.33**	63.00	95.49	**77.38**	**63.05**	4.22	**0.00**

MeshAdv [49]. We apply the same loss with *$\mathcal{X}$-Adv* to perform MeshAdv attacks. MeshAdv does not have the X-ray converter and location reinforcement; thus we fix the color of the converter and set the attack location as the four corners of the available attack areas.

Adversarial Patch [2]. We apply the first term of *$\mathcal{X}$-Adv* loss (*i.e.*, $\mathcal{L}_{adv}$) for Adversarial Patch attacks. The initial shape of the adversarial patch is a 20×20 depth image filled with 0. To utilize a 2D patch, the gradient is calculated and updated on the depth image rather than a 3D object. We set the attack location as the four corners of available attack areas.

A.4 Time Consumption of *$\mathcal{X}$-Adv*

In Table A.1, we show the time consumption of our *$\mathcal{X}$-Adv* attack on OPIXray dataset for the SSD detector. All of our experiments are conducted on 1 GPU of NVIDIA RTX 3080Ti and 8 CPU cores of Intel Xeon Gold 6148 @2.40GHz. The results demonstrate that our *$\mathcal{X}$-Adv* generates adversarial attacks with reasonable time consumption.

Also, we should note that the location reinforcement process takes the most time during the attack generation, and we will accelerate this process in future work.

B Additional Experimental Results

Different from OPIXray dataset, images in HiXray dataset may contain more than one prohibited item. We only attack images with only one prohibited item (about 3,227 images) for simplicity. The performance of different materials on the HiXray dataset is shown in Table B.1; the physical-world results on LIM are shown in Table B.2; the countermeasure validation in the physical world is shown in Table B.3;the black-box attack results on OPIXray and HiXray are shown in Table B.4; the white-box attack results on HiXray are shown in Table B.5. These results demonstrate the effectiveness.

Table B.2: Additional results of physical-world attack experiments on LIM.

Setting	mAP	Categories			
		SC	FO	ST	UT
Clean	96.20	98.85	99.55	95.58	90.82
Digital attack	29.56	76.69	4.08	31.06	6.41
Physical best	29.46	73.88	8.80	24.09	11.04
Physical change	51.38	75.30	47.94	41.46	40.81
Physical random	77.26	88.56	86.33	81.60	52.53

Table B.3: Countermeasure studies in the physical world. We train the defended models on the XAD dataset in the digital world, and evaluate their performance by collecting real images in the physical-world scenario.

(a) Data augmentation

Setting	mAP	Categories			
		SC	FO	ST	UT
V+C	91.35	84.17	98.05	100.00	83.18
V+A	33.16	66.33	18.35	44.48	3.46
D+C	90.57	85.76	97.14	98.42	80.98
D+A	51.16	63.15	31.79	72.27	37.42

(b) Adversarial Training

AT Setting	Attack	mAP	Categories			
			SC	FO	ST	UT
$\mathcal{X}$-*Adv*	Clean	91.12	95.03	98.85	100.00	70.59
	$\mathcal{X}$-*Adv*	59.15	89.18	49.64	80.25	17.54

Table B.4: Mean average precision of digital-world black-box attacks on OPIXray and HiXray datasets. The adversarial examples are generated from the source model and evaluated on the target model. **Bold** results denote the best performance of attacking in each column.

(a) OPIXray

Source Model	Target Model			
	SSD	Faster R-CNN	DOAM	LIM
SSD	**19.20**	30.01	44.37	39.31
Faster R-CNN	45.60	**23.33**	55.12	53.67
DOAM	36.46	28.69	**23.05**	40.00
LIM	26.90	28.37	38.59	**22.46**

(b) HiXray

Source Model	Target Model			
	SSD	Faster R-CNN	DOAM	LIM
SSD	**33.41**	41.28	45.02	40.45
Faster R-CNN	50.32	**39.39**	48.18	48.75
DOAM	45.94	42.09	**38.96**	44.87
LIM	42.27	47.94	44.99	**32.53**

Table B.5: Digital-world white-box attacking results on HiXray dataset. PO1, PO2, WA, LA, MP, TA, CO, and NL denote "Portable charger 1 (lithium-ion prismatic cell)", "Portable charger 2 (lithium-ion cylindrical cell)", "Water", "Laptop", "Mobile Phone", "Tablet", "Cosmetic" and "Non-metallic Lighter".

(a) SSD

Setting	mAP	Categories							
		PO1	PO2	WA	LA	MP	TA	CO	NL
Clean	63.06	58.76	57.61	74.24	97.55	95.18	80.66	40.47	0.02
Vanilla	56.94	54.33	42.65	73.99	96.81	92.43	77.87	17.46	0.00
MeshAdv	48.14	40.90	22.56	63.79	96.45	86.20	66.40	8.82	0.00
AdvPatch	43.71	29.55	6.62	65.95	96.65	77.62	59.98	13.30	0.00
$\mathcal{X}$-*Adv*	**33.41**	**10.96**	**3.52**	**46.91**	**95.11**	**57.17**	**52.89**	**0.67**	**0.00**

(b) Faster R-CNN

Setting	mAP	Categories							
		PO1	PO2	WA	LA	MP	TA	CO	NL
Clean	66.91	74.06	62.51	81.19	98.43	95.60	80.66	42.86	**0.00**
Vanilla	58.60	56.04	38.51	80.58	97.73	94.49	80.10	21.38	**0.00**
MeshAdv	51.12	28.53	17.84	77.00	98.20	92.64	80.19	14.54	**0.00**
AdvPatch	48.37	42.05	**1.40**	77.37	97.91	82.84	74.79	10.56	**0.00**
$\mathcal{X}$-*Adv*	**39.39**	**21.91**	4.54	**62.73**	**94.33**	**70.99**	**56.98**	**3.64**	**0.00**

(c) DOAM

Setting	mAP	Categories							
		PO1	PO2	WA	LA	MP	TA	CO	NL
Clean	63.43	64.71	57.02	72.11	97.14	94.68	80.86	41.70	0.01
Vanilla	54.23	44.33	30.91	74.77	96.65	92.43	80.39	14.32	**0.00**
MeshAdv	45.93	12.24	10.01	73.35	96.41	89.54	75.82	10.09	**0.00**
AdvPatch	42.20	16.27	**0.31**	67.50	96.02	85.05	67.20	5.26	**0.00**
$\mathcal{X}$-*Adv*	**38.96**	**5.21**	3.33	**63.00**	**95.49**	**77.38**	**63.05**	**4.22**	**0.00**

(d) LIM

Setting	mAP	Categories							
		PO1	PO2	WA	LA	MP	TA	CO	NL
Clean	64.84	69.89	57.85	69.98	97.89	95.10	81.71	46.24	0.08
Vanilla	55.34	49.03	34.09	66.35	98.10	91.74	75.41	28.00	0.01
MeshAdv	46.51	29.23	11.38	60.59	97.90	86.77	71.13	15.09	**0.00**
AdvPatch	40.75	25.65	**0.31**	52.11	97.98	76.88	61.20	11.83	**0.00**
$\mathcal{X}$-*Adv*	**32.53**	**2.98**	0.70	**37.94**	**97.18**	**60.61**	**58.82**	**1.99**	**0.00**