X-Adv: Physical Adversarial Object Attacks against X-ray Prohibited Item Detection

USENIX Security 2023

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]

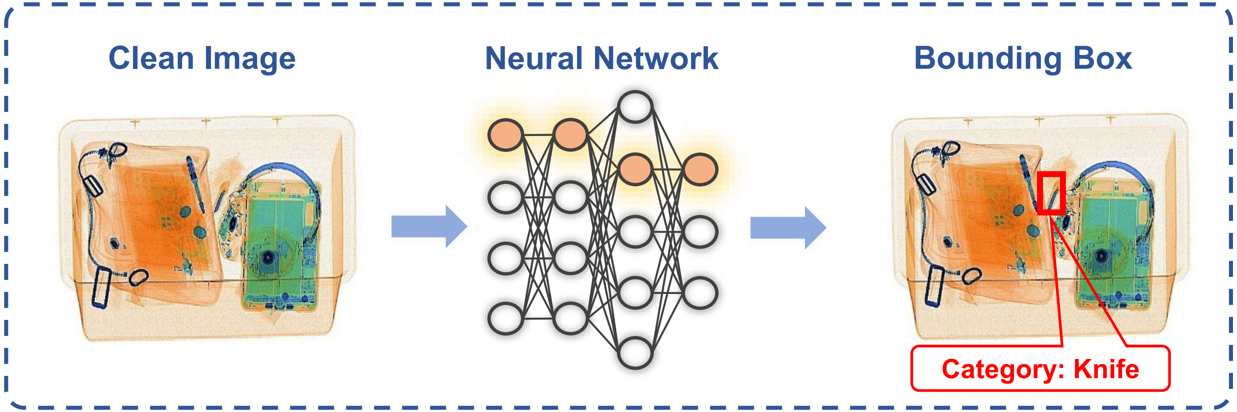
[1]*Beihang University*, [2]*Zhongguancun Laboratory*, [3]*Chinese Academy of Sciences*, [4]*University of Science and Technology of China*, [5]*iFLYTEK*, [6]*Hefei Comprehensive National Science Center*, [7]*JD Explore Academy*

Introduction

X-ray Prohibited Item Detection

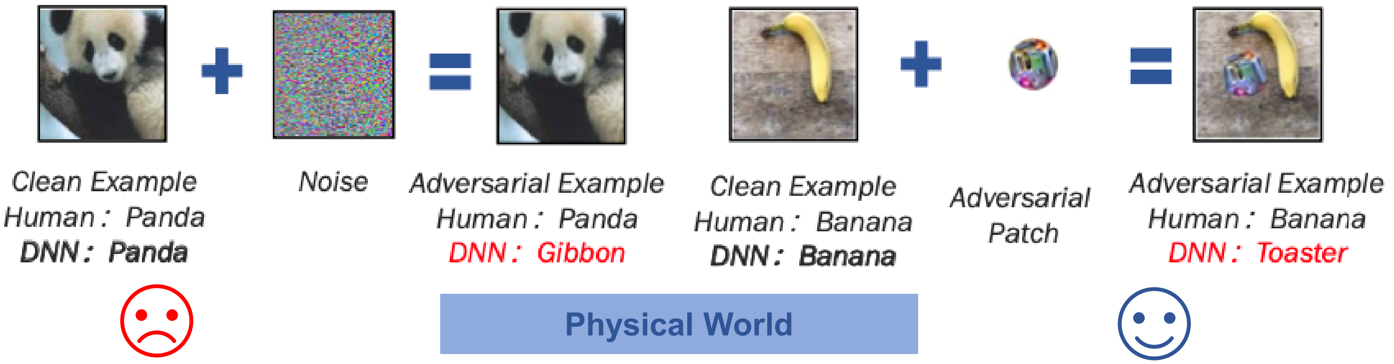
The object detector f_Θ takes an X-ray image $\mathbf{I}$ as input and outputs K detection boxes with locations $\boldsymbol{b}_k$ and confidences c_k. The optimization objective could be written as follows:

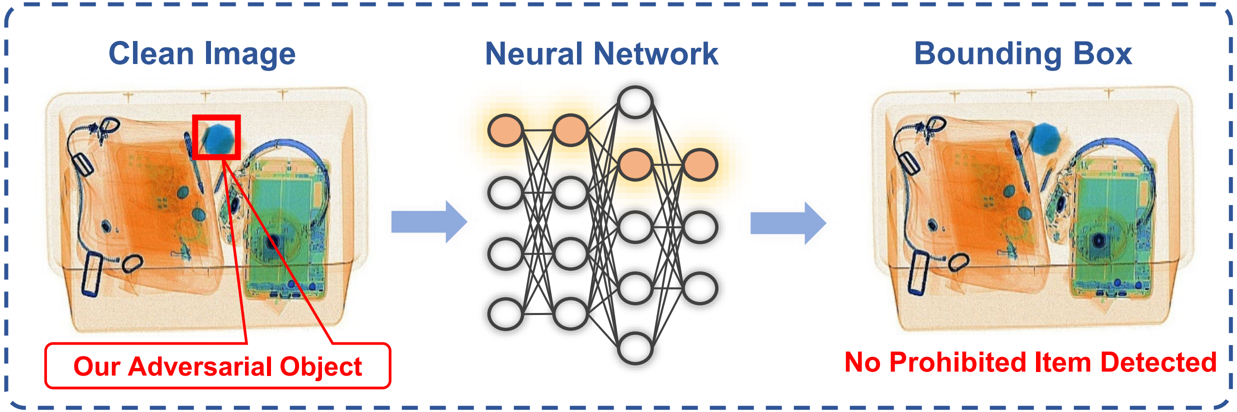
$$\min_{\Theta} \mathbb{E}_{(\mathbf{I},\{\mathbf{y}_k,\mathbf{b}_k\})\sim\mathbb{D}} \mathcal{L}(f_\Theta(\mathbf{I}), \{\mathbf{y}_k, \mathbf{b}_k\})$$



Introduction

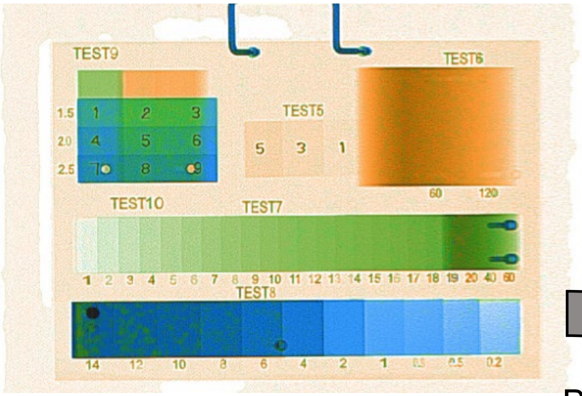
Adversarial Attack





Challenges

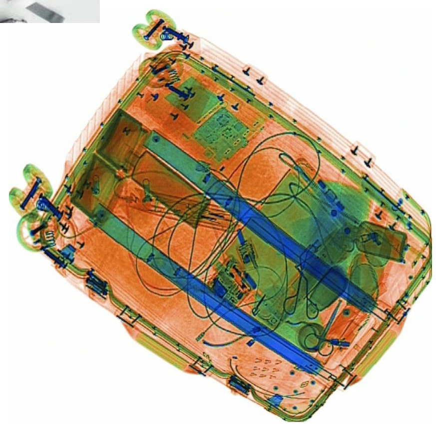
Imaging Principles



Black-box

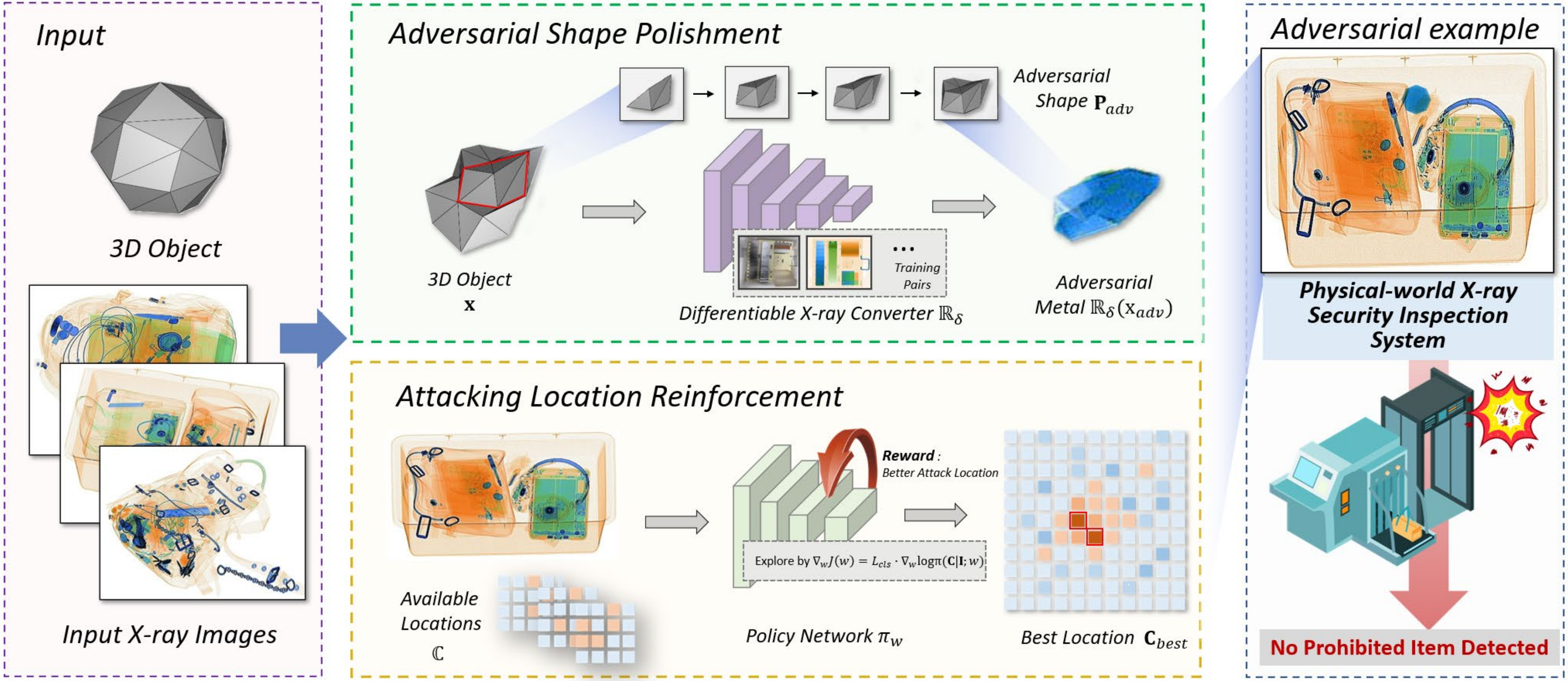
Complex Overlap

Diversity of sampling scenarios
Massive number of luggage items



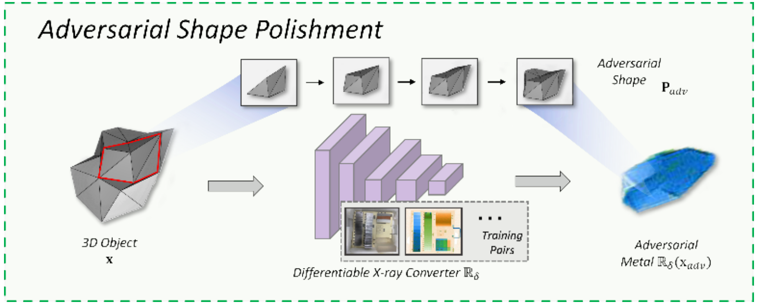
Approach

X-adv Adversarial Object Generation Framework



Approach

Adversarial Shape Polishment



X-ray photon beam intensity attenuation

$$I = I_0 \cdot exp(-\mu(\rho, Z)x)$$

Differentiable X-ray Converter (a, b, q are coefficient)

$$g_m(d) = a \cdot exp(-b \cdot d) + q$$

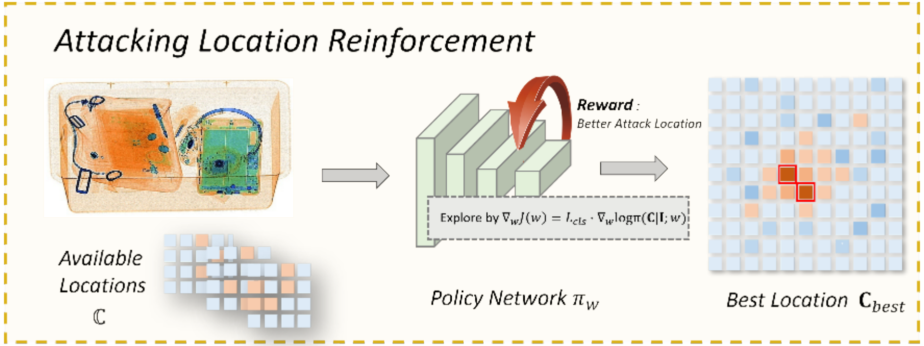
Attacking Location Reinforcement

The gradient of objective $J(\mathbf{w})$
(from REINFORCE algorithm)

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

Definition of reward G

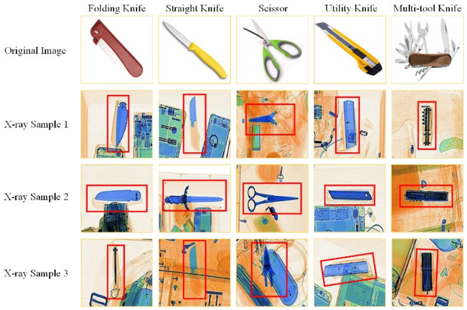
$$G = \mathcal{L}_{cls}(f_\Theta(\mathbb{R}_\delta(\mathbf{X}, \mathbf{x}_{adv}^{\mathbf{P}_{ori}, \mathbf{C}})), \{\mathbf{y}_k, \mathbf{b}_k\}) + \alpha \cdot \sigma_{\mathbf{C}}$$



Experiments

Datasets

We conduct experiments on high-resolution X-ray prohibited item detection datasets.

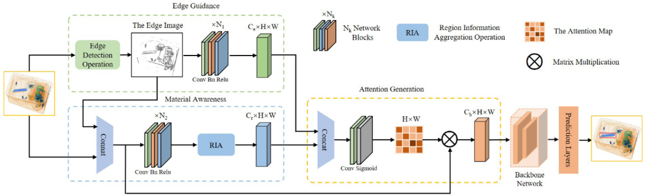


OPIXray

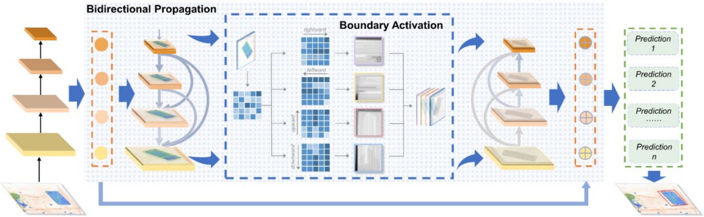
HiXray

Models

We apply Faster R-CNN and two SSD models designed for X-ray prohibited item detection.



DOAM



LIM

Experiments

Digital-World White-Box Attacks

mAP (mean average precision): lower mAP indicates better attack performance.

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
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
X-Adv	**23.33**	**26.62**	**3.44**	**62.91**	**15.33**	**8.36**

(c) DOAM

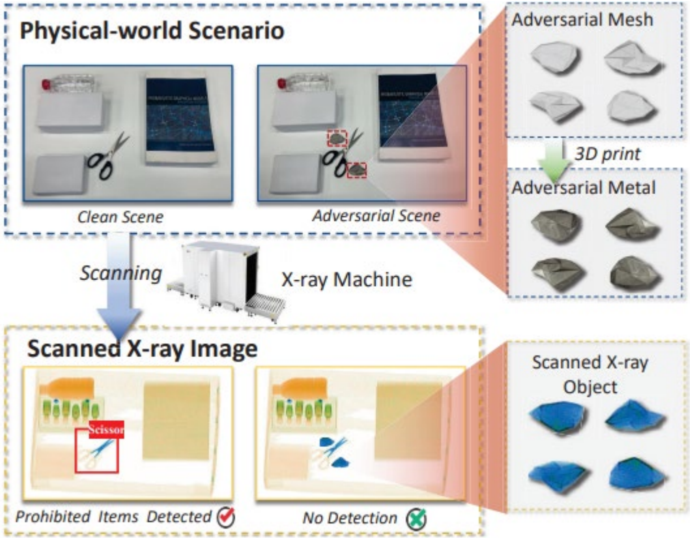
Setting	mAP	Categories				
		FO	ST	SC	UT	MU
Clean	74.02	78.92	40.88	95.65	74.08	80.55
Vanilla	67.79	74.26	32.57	91.37	63.41	77.34
MeshAdv	56.36	60.09	23.04	86.87	47.11	64.68
AdvPatch	42.04	45.57	9.41	81.19	26.44	47.60
X-Adv	**23.05**	**18.40**	**4.05**	**64.80**	**18.57**	**9.45**

(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
X-Adv	**22.46**	**31.64**	**4.28**	**52.59**	**16.65**	**7.13**

Experiments

Physical-World Black-Box Attacks



(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

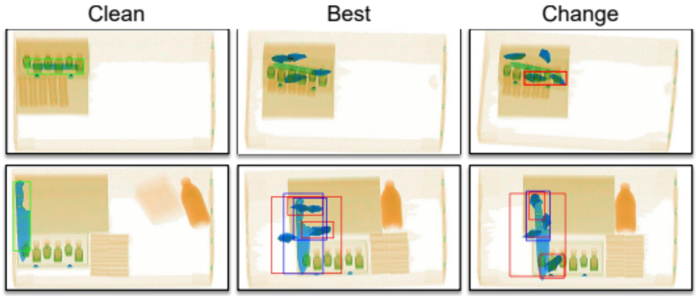


Figure 4: Detection results of some X-ray images in our physical-world experiments (we choose images with fewer items for better visualization). Green boxes indicate correct classes and suitable locations; blue boxes represent correct classes in incorrect locations; red boxes indicate incorrect classes.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

(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

Experiments

Ablation Studies & Analysis

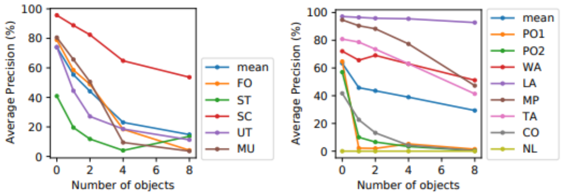
Table 3: Ablation studies on different attack locations. Our strategy achieves the best attack performance.

(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**

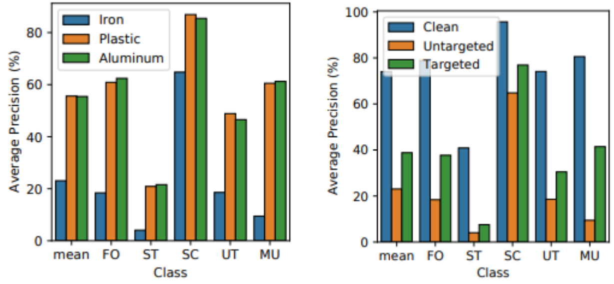
(b) HiXray

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**



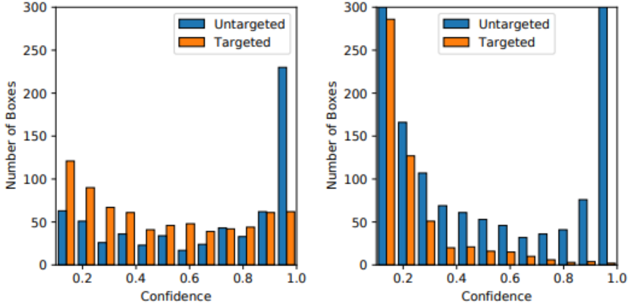
(a) OPIXray (b) HiXray

Figure 5: Ablations on the numbers of adversarial objects.



(a) Materials (b) Targeted/untargeted

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



(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.

Countermeasures

Data augmentation can improve the robustness against *X-adv*.

Adversarial detection classifiers cannot reach high accuracy.

AT with PGD cannot defend against X-adv due to the difference between perturbations and patch/object attacks.

AT with X-adv can mitigate the X-adv to a certain extent.

(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

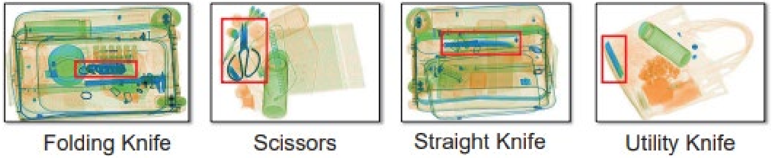
(c) Adversarial Training

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

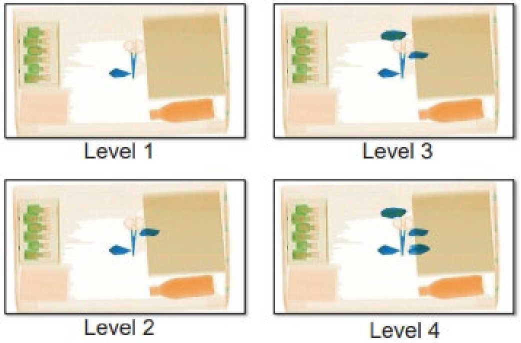
Dataset

Physical World XAD Dataset

The first physical-world attack dataset in X-ray scenario.



(a) Prohibited item category



(b) Adversarial objects **(c) Severity levels**

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

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

Thanks!

Email: junguo@buaa.edu.cn
Code: https://github.com/DIG-Beihang/X-adv