Exploring Endogenous Shift for Cross-domain Detection: A Large-scale Benchmark and Perturbation Suppression Network

Renshuai Tao *[1,2], Hainan Li*[1], Tianbo Wang[1], Yanlu Wei[1], Yifu Ding[1],
Bowei Jin[2], Hongping Zhi[2], Xianglong Liu[1†], Aishan Liu[1]
[1]State Key Laboratory of Software Development Environment, Beihang University
[2]iFLYTEK Research
{rstao, hainan, tianbowang, weiyanlu, yifuding}@buaa.edu.cn,
jin1128hf@gmail.com, zhihp@126.com, {xlliu, liuaishan}@buaa.edu.cn

Abstract

Existing cross-domain detection methods mostly study the domain shifts where differences between domains are often caused by external environment and perceivable for humans. However, in real-world scenarios (e.g., MRI medical diagnosis, X-ray security inspection), there still exists another type of shift, named endogenous shift, where the differences between domains are mainly caused by the intrinsic factors (e.g., imaging mechanisms, hardware components, etc.), and usually inconspicuous. This shift can also severely harm the cross-domain detection performance but has been rarely studied. To support this study, we contribute the first Endogenous Domain Shift (EDS) benchmark, X-ray security inspection, where the endogenous shifts among the domains are mainly caused by different X-ray machine types with different hardware parameters, wear degrees, etc. EDS consists of 14,219 images including 31,654 common instances from three domains (X-ray machines), with bounding-box annotations from 10 categories. To handle the endogenous shift, we further introduce the Perturbation Suppression Network (PSN), motivated by the fact that this shift is mainly caused by two types of perturbations: category-dependent and category-independent ones. PSN respectively exploits local prototype alignment and global adversarial learning mechanism to suppress these two types of perturbations. The comprehensive evaluation results show that PSN outperforms SOTA methods, serving a new perspective to the cross-domain research community.

1. Introduction

Traditional CNN-based detection methods [28, 32–34, 38,39,44,57–59] suffer a sharp performance drop when they are applied to a novel scenario. To overcome the quandary of cross-domain detection, one potential approach is to exploit Unsupervised Domain Adaptation (UDA) [14, 20, 45, 51] to transfer essential knowledge from the labeled source domain to the unlabeled target domain. Existing methods primarily study obvious domain shifts, where differences between domains are often caused by external environment and perceivable for human. Relying on the powerful capabilities of CNN, these methods have achieved promising performance.

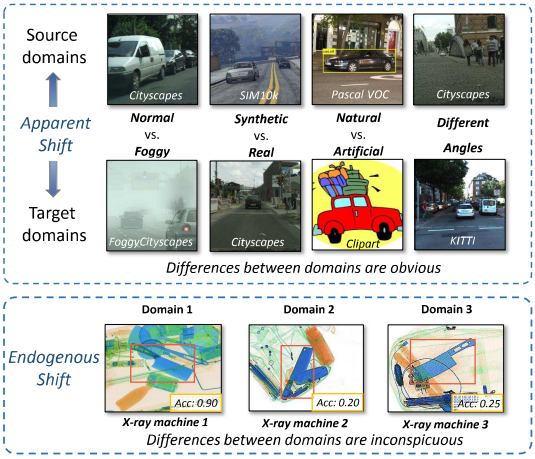

Figure 1. Scenarios of apparent and endogenous shift. The apparent shift is obvious while the endogenous shift is inconspicuous. As illustrated, images from different types of X-ray machines are very similar in colors and subtly different of lightness, texture, *etc*.

However, in industrial scenarios (*e.g.*, MRI medical diagnosis [1, 16, 31, 56], X-ray security inspection [29, 30, 47, 48, 52]), there often exists another type of domain shift, named endogenous domain shift, where differences between domains are mainly caused by the intrinsic factors like imaging mechanisms, hardware components, *etc*., and

*Equal contribution
†Corresponding author

usually inconspicuous. As illustrated in Figure. 1, the shifts between images from different domains (generated by three types of X-ray machines) are barely perceivable to naked eyes. The endogenous shift has been rarely studied in the literature, but severely harm the cross-domain detection performance. Till now, there is no specific datasest and promising models to support this meaningful research.

To support the study of this important issue, in this paper, we contribute the first endogenous domain shift benchmark named Endogenous Domain Shift dataset (EDS) by selecting a typical scenario, X-ray security inspection. Due to the differences in intrinsic mechanisms or hardware components of the imaging systems of different types of X-ray machines, there are subtle perturbations in X-ray images generated by different X-ray machines, which cause the endogenous shift. EDS consists of 14,219 images, including 31,654 instances with bounding-box annotations of 10 common categories, generated by three different types of X-ray security inspection machines. According to our observations, there exist two types of perturbations in these domains. The first type of perturbation is highly sensitive and correlated to the category, and we name it *category-dependent* perturbation. The second type of perturbation mainly refers to the overall imaging qualities, caused by different systems with hardware components containing different parameters, named *category-independent* perturbation.

Due to the fact that traditional CNN-based cross-domain detection methods do not pay attention to the essential cause (*e.g.*, two types of perturbations) of this tiny shift, the performance is not satisfied while applying them directly to handle the endogenous shift. To well deal with endogenous domain shifts, in this work, we further propose the Perturbation Suppression Network (PSN), consisting of two core modules, local prototype alignment (LPA) and global adversarial assimilation (GAA). These two modules suppress category-dependent and category-independent perturbations, receptively. Specifically, in LPA, we aggregate each category of objects into a category prototype through graph-based method for both source and target data. In GAA, we exploit the adversarial learning strategy that the backbone network generates features to confuse the domain classifier, adaptively suppressing perturbation and retaining the salient characteristics of the two domains. We summarize the main **contributions** as follows and hope our study could serve a new perspective to the cross-domain research.

- We first put forward a novel and important type of domain shift in cross-domain detection, the endogenous shift, which may cause severe performance drop but has been rarely studied. We proved the existence and harm of the endogenous shift through experiments.

- To support study of this issue, we contribute a large-scale benchmark, named EDS dataset, by selecting the typical scenario, X-ray security inspection. The shift between domains are mainly caused by two types of perturbations generated by three types of machines.

- To deal with the endogenous shift, we further propose the PSN model, exploiting local prototype alignment and global adversarial learning mechanism to suppress the two types of perturbations in the endogenous shift.

- We comprehensively evaluate PSN in EDS dataset and the simulated dataset. All of the results demonstrate that the proposed method can well deal with the endogenous shift and outperform SOTA methods.

2. Related Work

2.1. Cross-domain Detection Datasets

In cross-domain detection task, previous works [3, 12, 13, 23, 35, 53] are usually four types of scenarios, "different climates", "real vs. composite images", "different camera angles" and "cartoon vs. real images". The common datasets are Cityscapes [5], Foggy Cityscapes [42], KITTI [9], SIM10K [17], Pascal VOC [8] and Clipart [15]. The first scenario usually adopts Cityscapes and Foggy Cityscapes datasets to simulate the weather across sunny to foggy. The second scenario usually adopts Cityscapes and SIM10K to evaluate the adaptation effectiveness from real and simulated. The third scenario usually adopts Cityscapes and KITTI to mimic the object photoed by different camera angles. In the fourth scenario, Pascal VOC and Clipart usually adopted to portray the domain shift between real and cartoon images.

2.2. Cross-domain Detection Methods

Cross-domain detection [2, 11, 18, 36, 55] is more complicated compared to common cross-domain classification because it is necessary to locate and classify all instances of various objects inside images [7, 19, 21, 22, 37]. Recently, several works have been proposed to address the domain shift problem in cross-domain object detection task by various technologies. [4] has made progress in the challenging unsupervised domain adaptative object detection task, which aligns both the image and instance levels in a domain adversarial manner. After that, the following works *Strong-Weak Domain Adaptive Faster R-CNN* [41], *Collaborative Training between Region Proposal Localization and Classification* [60], *Coarse-to-Fine Feature Adaptation* [61], *Graph-induced Prototype Alignment* [54] are proposed one after another to push the direction forward. However, previous works paid less attention to the essential cause (*e.g.*, perturbations) of this endogenous shift, the performance is not satisfied while applying them directly to handle it.

3. Endogenous Domain Shift Dataset

A dataset is significant to boost a research. As Table 1 illustrates, the existing datasets for cross-domain detection task mainly focus on the obvious domain shift, which cannot meet the demands where differences between domains are inconspicuous and mainly caused by intrinsic mechanism or hardware. The differences of the two types of shifts are shown in Figure 1. However, till now, there is no specific dataset to support this meaningful research.

Thus, we contribute the first endogenous domain shift benchmark named Endogenous Domain Shift (EDS) dataset by selecting a typical scenario, X-ray security inspection. Although the imaging mechanisms of the three machines are roughly the same (illustrated in Table 2), there exist large endogenous shifts, such as differences of color depth, texture, which mainly caused by hardware parameters, wear degrees, *etc*., of different X-ray machines.

Scenarios	Datasets	N_d	N_e
Different Climates	Cityscapes vs Foggy	2	2
Real vs composite	SIM10k vs Cityscapes	2	2
Different angles	Cityscapes vs KITTI	2	2
Cartoon vs real	Pascal VOC vs Clipart	2	2
Different machines	**Our EDS Dataset**	**3**	**6**

Table 1. Comparisons between Existing cross-domain detection datasets and EDS dataset. N_d refers to the number of domains and N_e refers to the number of cross-domain experiments supported.

3.1. Construction Details

Data Collection. We exploit three X-ray security inspection machines from different manufactures and with different serving time, which guaranties to generate three domains of images. We randomly put the objects in pre-prepared package to generate images. After sending the package to the security inspection machine, the machine will completely cut out the package by detecting the blank.

Category Selection. As Figure 2 illustrates, EDS dataset contains the 10 categories of common objects, *e.g*., "Plastic Bottle", "Pressure", "Lighter, "Knife", "Device", "Power Bank", "Umbrella", "Glass Bottle", "Scissor" and "Laptop". All of these objects are frequently seen in daily life. Extensive diverse categories and sufficient numbers of instances can provide a more credible evaluation for various cross-domain detection models.

Quality Control. We followed the similar quality control procedure of annotation as the famous Pascal VOC [8]. All annotators followed the same annotation guidelines including what to annotate, how to annotate bounding, how to treat occlusion, *etc*. Besides, to ensure the accuracy of annotation, we divide the annotators into 3 groups and all of the images are randomly designated to 2/3 specific groups to be annotated. Then, the last group is specially organized to confirm the annotation results.

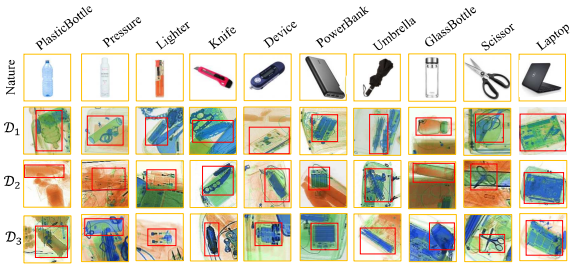

Figure 2. 10 categories of objects from natural scenario and 3 different X-ray machines. "$\mathcal{D}_1$", "$\mathcal{D}_2$" and "$\mathcal{D}_3$" refer to "domain 1", "domain 2" and "domain 3", respectively.

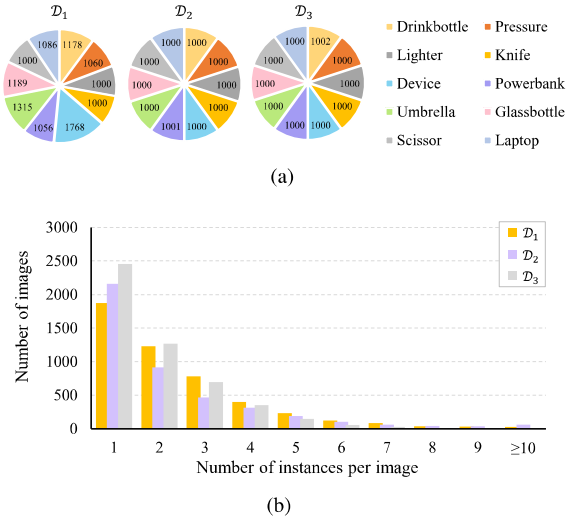

Figure 3. The distributions of the data. (a) illustrates the distribution of numbers of instances per category and (b) illustrates the distribution of numbers of instances per image.

3.2. Data Properties

Instances per category. EDS contains 14,219 X-ray images, 10 categories of 31,655 instances with bounding-box anotations of common objects. Note that we guarantee that the instances in each category is no less than 1000, which is sufficient for the evaluation. The distribution of numbers of instances per category is illustrated in Figure 3 (a).

Instances per image. Each image contains at least one instance and on average there are 2.22 instances per image. The distribution of numbers of images containing different numbers of instances is illustrated in Figure 3 (b).

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 2. The imaging mechanisms of the three machines are roughly the same, subtly different of color depth and texture.

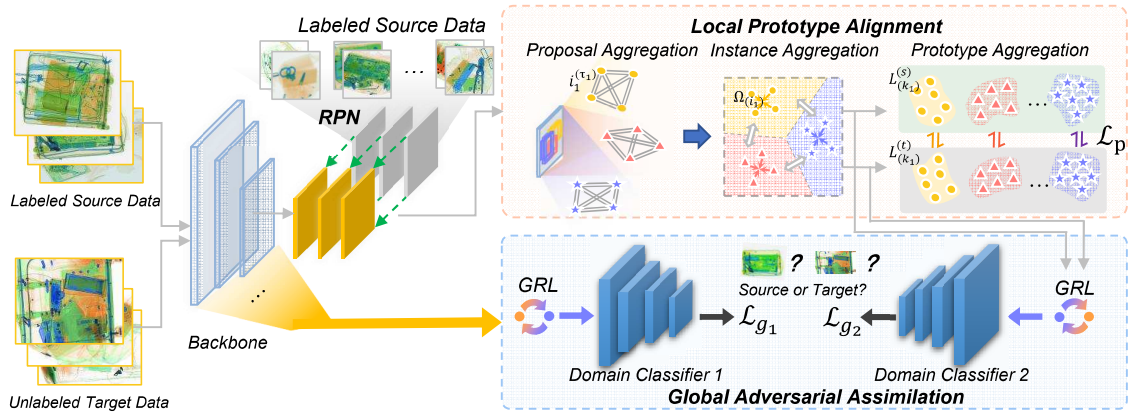

Figure 4. The structure of the Perturbation Suppression Network (PSN). The network consists of three core modules, detection network, local prototype alignment (LPA) and global adversarial assimilation (GAA). GRL refers to the gradient reverse layer.

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

Table 2. The relationship between object material and color in X-ray imaging. Different types of X-ray machines are very similar in imaging colors and subtly different of lightness, texture, *etc*.

4. Perturbation Suppression Network

To deal with the endogenous shift, in this section, we introduce the Perturbation Suppression Network (PSN).

4.1. Motivation

Figure 5 illustrates two typical types of endogenous shifts. First, objects from the same category are primarily composed of similar materials (*e.g.*, metal for knives), which reflect the intrinsic properties of the category. Objects from different categories are often shown to have different imaging qualities (*e.g.*, perturbations with different granularities on knives and plastic bottles). This type of perturbation is highly sensitive and correlated to the category, and we hereby refer to it as the **category-dependent** perturbation. Second, different systems consist of hardware components with different parameters, which will directly influence the overall imaging qualities (*e.g.*, different saturation and hues of the X-ray image, such as background). These perturbations are introduced directly on the global image background while are irrelevant to object categories, which is called the **category-independent** perturbation.

Based on the above observation, we propose the Perturbation Suppression Network (PSN), which exploits a integrated mechanism to suppress the two different types of perturbations mentioned above. Regarding the *category-dependent perturbation*, inspired by the fact that the prototype of each category is the strongest embodiment of the common characteristics of this category prototype, we align the corresponding categorical prototypes for the source and target domains (X-ray machines) to generate domain-invariant features. Considering the *category-independent perturbation*, due to the fact that the perturbation is globally distributed over the entire image, we adopt global adversarial learning to suppress them. In particular, we guide the backbone network to learn features that could confuse the domain classifier to make predictions.

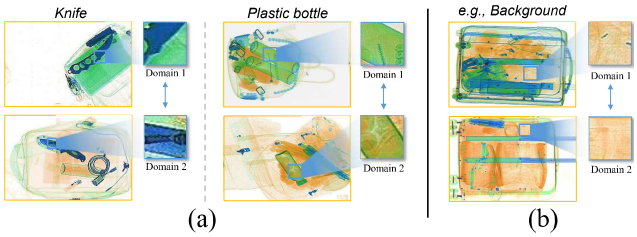

Figure 5. The illustration of the typical endogenous domain shifts. (a) refers to the category-dependent perturbation and (b) refers to the category-independent perturbation.

4.2. Network Architecture

In Figure 4, the framework includes three core modules, the detection network, local prototype alignment module (LPA) and global adversarial assimilation module (GAA). We select the reputed base model, Faster R-CNN, as the detection network. LPA suppress the category-dependent perturbation through aligning the corresponding categorical prototypes. GAA suppress the category-independent perturbation through adopting global adversarial mechanism.

4.2.1 Local Prototype Alignment (LPA)

To deal with the category-dependent perturbation, in this part, we construct a relation graph set $\mathbf{G}^{(i)} = \{\mathrm{G}^{(i_1)}, \mathrm{G}^{(i_2)}, \ldots, \mathrm{G}^{(i_n)}\}$ for each instance i, through structuring the proposals surrounding i generated by the RPN network. Each graph $\mathrm{G}^{(i)} = \{\mathrm{V}^{(i)}, \mathrm{E}^{(i)}\}$, where $\mathrm{V}^{(i)} =$

$\{i^{(\tau_1)}, i^{(\tau_2)}, \ldots, i^{(\tau_n)}\}$ is the proposal set of the instance i and $\mathrm{E}^{(i)}$ is the edge set of each two proposals in $\mathrm{V}^{(i)}$. In the first step, we aggregate all proposals of the same object to generate the most accurate feature map, *i.e.*, the instance prototype $\Omega^{(i)}$, that we consider to represent the instance i. This process is formulated as:

$$\Omega^{(i)} = \left(\sum_{n=1}^{N_{(\tau)}} \sum_{m=1}^{N_{(\tau)}-1} \mathbf{IoU}(i^{(\tau_m)}, i^{(\tau_n)}) \cdot i^{(\tau_n)} \right) / N_{(\tau)} \quad (1)$$

where $i^{(\tau_m)}$ refers to the proposals around of the instance i except $i^{(\tau_n)}$, $N_{(\tau)}$ refers to the number of proposals around the instance i and $\Omega^{(i)}$ refers to the prototype of instance i. Inspired by [54], we choose IoU of two proposals as the metric. Similarly, we exploit the same operation to get category probability prototype $P^{(i)}$ for each instance i.

In the second step, similarly, we construct a relation graph set for each class k in the input image. This step generates the category prototype by the same operation as the first step. This process can be formulated as:

$$\Omega^{(k)} = \left(\sum_{i=1}^{N_{(i)}} \left(P^{(i,k)} \cdot \Omega^{(i)} \right) \right) / \sum_{i=1}^{N_{(i)}} P^{(i,k)} \quad (2)$$

where $P^{(i,k)}$ refers to the probability of $\Omega^{(i)}$ belonging to the category k, $N_{(i)}$ refers to the number of instances of category k and $\Omega^{(k)}$ refers to the prototype of category k.

After the two steps operation, the graph generates a prototype set $\mathbf{\Omega}^{(k)} = \{\Omega^{(k_1)}, \Omega^{(k_2)}, \ldots, \Omega^{(k_n)}\}$. Each elements in $\mathbf{\Omega}^{(k)}$ represents the prototype of one category. In the third step, we construct two prototype libraries, $\mathbf{L}^{(s)}$ and $\mathbf{L}^{(t)}$, for both source data and target data. In each epoch of training, the prototype set $\mathbf{\Omega}$ generated from the input image updates the prototype library. Thus, at the end of each round of training, the LPA module generates two prototype libraries, $\mathbf{L}_{(s)} = \{\mathrm{L}_{(s)}^{(k_1)}, \mathrm{L}_{(s)}^{(k_2)}, \ldots, \mathrm{L}_{(s)}^{(k_n)}\}$ and $\mathbf{L}_{(t)} = \{\mathrm{L}_{(t)}^{(k_1)}, \mathrm{L}_{(t)}^{(k_2)}, \ldots, \mathrm{L}_{(t)}^{(k_n)}\}$. The process of updating can be formulated as follows:

$$\mathrm{L}_{(l)}^{(k)} = \begin{cases} \Omega_{(l)}^{(k)}, & l = 1 \\ \alpha \cdot \Omega_{(l)}^{(k)} + (1-\alpha) \cdot \mathrm{L}_{(l-1)}^{(k)}, & l > 1 \end{cases} \quad (3)$$

where α refers to the cosine distance of two variables, *e.g.*, $\alpha = \cos(\Omega_{(l)}^{(k)}, \mathrm{L}_{(l-1)}^{(k)})$, $\Omega_{(l)}^{(k)}$ refers to the category prototype of k in the l-th training and $\mathrm{L}_{(l-1)}^{(k)}$ refers to the category prototype k in the library after $(l-1)$-th training.

Finally, we try to minimize the distance between the pair of prototypes with the same category and maximize the distance between different categories of the two domains. The process of alignment can be formulated as follow:

$$\mathcal{L}_p = \frac{1}{N_{(k)}} \cdot \sum_{m=1}^{N_{(k)}} \left\| \mathrm{L}_{(s)}^{(m)} - \mathrm{L}_{(t)}^{(m)} \right\|_2 - \Psi \quad (4)$$

where $\mathcal{L}_p$ refers to the loss of the LPA module and Ψ refers to the measure of distances for different categories in both domains, which is illustrated in Supplementary Materials.

4.2.2 Global Adversarial Assimilation (GAA)

To deal with the category-independent perturbation, in this part, we integrate multiple domain classifiers $\mathcal{C}$ into several convolutional blocks in the backbone network $\mathcal{G}$. Thus, the feature map set $\mathbf{\Phi} = \{\Phi^{(1)}, \Phi^{(2)}, \ldots, \Phi^{(n)}\}$ is transported to the domain classifiers $\mathcal{C}$ as "raw material" to generate the adversarial loss.Therefore, the backbone network $\mathcal{G}$ tries to generate collective features of the two domains to confuse the classifier while the classifier tries to distinguish that which domain the input image comes from. In the two-player minimax game, the perturbation information is selectively filtered out while generating the robust collective features. The process of inputting the features outputted by the backbone network into the classifier can be formulated:

$$\begin{aligned} \mathcal{L}_{g_1} = \min_{\theta_{\mathcal{G}}} \max_{\theta_{\mathcal{C}}} \mathbb{E}_{x_s \sim \mathcal{D}_S, x_t \sim \mathcal{D}_T} \{ & \log \mathcal{C}\left(\mathcal{G}\left(x_s\right)\right) \\ & + \log\left(1 - \mathcal{C}\left(\mathcal{G}\left(x_t\right)\right)\right)\} \end{aligned} \quad (5)$$

where x_s and x_t refers to the input from both domains.

In addition, in order to give full play to the ability of GAA for suppressing the global perturbation, we add the prototype set $\mathbf{\Omega}^{(k)} = \{\Omega^{(k_1)}, \Omega^{(k_2)}, \ldots, \Omega^{(k_n)}\}$ generated in each round of training by LPA into the "raw materials" like the feature map set $\mathbf{\Phi}$ above to generate a sufficient number of gradients to adjust the parameters in backbone network $\mathcal{G}$. This process can be formulated as follows:

$$\begin{aligned} \mathcal{L}_{g_2} = \min_{\theta_{\mathcal{G}}} \max_{\theta_{\mathcal{C}}} \mathbb{E}_{\Omega_{(s)}^{(k)} \sim \mathbf{\Omega}_{(s)}, \Omega_{(t)}^{(k)} \sim \mathbf{\Omega}_{(t)}} \{ & \log \mathcal{C}(\Omega_{(s)}^{(k)}) \\ & + \log(1 - \mathcal{C}(\Omega_{(t)}^{(k)}))\} \end{aligned} \quad (6)$$

where $\Omega_{(s)}^{(k)}$ and $\Omega_{(t)}^{(k)}$ refer to the category prototype of k generated by Equation 2 in both domains. Note that the generation of $\Omega_{(s)}^{(k)}$ and $\Omega_{(t)}^{(k)}$ relies on the backbone network $\mathcal{G}$, we omit the backbone network $\mathcal{G}$ for simplicity.

4.2.3 Network Training

The total loss of the perturbation suppression network $\mathcal{L}$ can be calculated as the sum of the loss values of the three core modules, the detection network, LPA and GAA. The calculation of the total loss can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_d + \lambda_p \mathcal{L}_p + \lambda_{g_1} \mathcal{L}_{g_1} + \lambda_{g_2} \mathcal{L}_{g_2} \quad (7)$$

where $\mathcal{L}_d$ refers to the detection loss of the faster R-CNN and λ_p, λ_{g_1}, λ_{g_2} refer to the super parameters, which are discussed in the Supplementary Materials.

In the base training stage, the network are trained with abundant data of source domain, which can easily localize the appropriate proposals and determine their categories. After the base training, a pair of images from two domains are imputed to PSN to calculate the loss values of different modules, respectively. Specifically, the entire training procedure of the whole Perturbation Suppression Network can be viewed as Algorithm 1.

Algorithm 1 Training of Perturbation Suppression Network

Input: A pair of images, the number of categories k.
Output: The loss value $\mathcal{L}$.

```
Generate the feature map set Φ.
Generate the proposal set V^(i) with m proposals.
Calculate the loss value L_d.
for all a = 1, 2, ..., τ_n do
    Calculate each instance prototype Ω^(a).
    Calculate each probability prototype P^(a).
end for
for all b = 1, 2, ..., k_n do
    Calculate each category prototype Ω^(b).
    Update the prototype library L_(b).
end for
Calculate the loss value L_p.
Calculate the loss value L_g1 through Φ.
Calculate the loss value L_g2 through the set of Ω^(k).
Calculate the total loss value L.
```

5. Experiments

In this section, we introduce the comprehensive experiments to evaluate the effectiveness of the proposed method.

5.1. Experimental Settings

5.1.1 Datasets and Baselines

First, we conduct experiments on the proposed EDS dataset. Second, we simulate the endogenous shift in natural images dataset CityScapes [5] by adding adversarial noises, generating two simulated datasets. Third, to evaluate the generalization to common scenario, we conduct experiments on the adaptation from Cityscapes [6] to Foggy-Cityscapes [42]. In both EDS and the simulated dataset, we conduct six groups of experiments, totally in Table 3. Regarding the models for comparison, we select the SOTA models with different mechanisms, SO [40] ("Source Only", *i.e.*, Faster R-CNN model trained on the source domain only, the most commonly used baseline), DA [4] (instance-level and image-level alignment), SWDA [41] (attention mechanism), CST [60] (collaborative learning), CFA [61] (prototype alignment). For fair comparison, we select VGG-16 [43] as the backbone network for all models.

5.1.2 Metrics and parameters

We choose the widely used metric, mean average precision (mAP) for overall performance and average precision (AP) for each category. The parameters of the two domain classifiers are optimized by Adam and other parameters are optimized by the SGD. The initial learning rate is set to 0.001 and becomes 0.0001 after 50000 steps. The momentum of SGD and weight decay are set to 0.9 and 0.0005 respectively. The batch size is set to two (one of the source domain and one of the target) with shuffle strategy while training. Besides, the IoU threshold measuring the accuracy of the predicted bounding box against ground-truth is set to 0.5.

5.2. Comparing with SOTA models

In this section, we illustrate the mAP results and the average result of different categories on the EDS dataset (Section 5.2.1), the simulated dataset (Section 5.2.2), the common natural dataset CityScapes [5] (Section 5.2.3). Note that due to the length limitation of the paper, in Section 5.2.1, we only show the mAP results of six groups of settings on EDS dataset in Table 3 and the average result for different categories in six groups of settings in EDS dataset in Table 4. The entire and specific experimental results of all categories under all groups of settings in Section 5.2.1 are illustrated in the supplementary materials.

5.2.1 Results on the EDS dataset

For the **mAP** results on the EDS dataset, Table 3 illustrates two advantages of our method, stability and effectiveness. Regarding the **stability**, our method outperforms the most commonly used baseline "SO" [40] by a large margin in all six groups settings. Specifically, our method outperforms "Source Only" by **6.0**%, **4.0**%, **9.6**%, **2.4**%, **5.9**% and **1.3**% in the setting of $\mathcal{D}_{1\to2}$, $\mathcal{D}_{1\to3}$, $\mathcal{D}_{2\to1}$, $\mathcal{D}_{2\to3}$, $\mathcal{D}_{3\to1}$ and $\mathcal{D}_{3\to2}$, respectively. In some settings, other methods do not perform better than the baseline "SO". The results demonstrates that our method can achieve a stable performance improvement in handling the endogenous shift in various

Methods	$\mathcal{D}_{1\to2}$	$\mathcal{D}_{1\to3}$	$\mathcal{D}_{2\to1}$	$\mathcal{D}_{2\to3}$	$\mathcal{D}_{3\to1}$	$\mathcal{D}_{3\to2}$
SO [40]	42.3	53.6	41.8	55.4	52.7	53.6
DA [4]	46.3	55.6	45.0	57.5	56.1	54.6
SWDA [41]	46.9	56.5	49.7	56.7	56.6	54.8
CST [60]	46.9	54.3	49.2	55.5	56.5	52.8
CFA [61]	44.3	53.7	51.3	53.6	55.4	51.3
PSN (**ours**)	**48.3**	**57.6**	**51.4**	**57.8**	**58.6**	**54.9**

Table 3. The **mAP** results (%) of various methods on each group of adaptation of **EDS dataset**. $\mathcal{D}_{m\to n}$ refers to the adaptation from domain m to domain n. "SO" refers to "Source only", the Faster R-CNN model trained on the source domain only.

Methods	DB	PR	LI	KN	SE	PB	UM	GB	SC	LA
SO [40]	55.3	44.8	34.1	16.5	43.2	65.8	85.2	37.6	26.5	87.4
DA [4]	54.7	52.7	38.6	15.4	47.7	68.3	86.7	40.2	30.2	90.1
SWDA [41]	55.6	52.6	40.9	17.3	49.5	69.8	86.7	41.1	30.0	90.3
CST [60]	55.1	51.2	39.0	16.0	49.6	69.5	86.5	40.7	25.0	**92.2**
CFA [61]	51.9	52.0	33.7	14.8	49.6	68.9	85.4	41.8	26.8	90.5
PSN (**ours**)	**56.2**	**54.0**	**41.3**	**18.2**	**52.4**	**72.1**	**86.8**	**44.4**	**31.4**	91.4

Table 4. The **AP** results (%) of different categories on on the adaptation of **EDS dataset**. "$\mathcal{D}_{m\to n}$" refers to the model trained on the source domain m and tested on the target domain n. "DB", ..., "LA" refer to "Plastic Bottle", ... "Laptop" in 3.1, respectively.

settings. As for the **effectiveness**, our method outperforms the SOTA methods in different settings, especially by **1.4**%, **1.1**% and **2.0**% in $\mathcal{D}_{1\to2}$, $\mathcal{D}_{1\to3}$ and $\mathcal{D}_{3\to1}$. Note that in the sixth group, *i.e.*, $\mathcal{D}_{3\to2}$, our method has a slight improvement to the baseline by **1.3**%, mainly because the endogenous shift between $\mathcal{D}_3$ and $\mathcal{D}_2$ is more smaller than other pairs of domains. As for $\mathcal{D}_{2\to3}$ with a higher improvement by **2.4**%, it mainly because the baseline achieves a higher performance due to that the backbone network can extract better features, which limits the effectiveness for extracting common features of all cross-domain detection methods.

For the **average results of each category**, Table 4 illustrates that our method outperforms the baseline by a large margin in all categories, especially for SE, PB and GB by **9.2**%, **6.3**% and **6.8**%. Besides, in UM and LA, our method outperforms the baseline by **1.6**% and **4.0**%, while a slighter improvement to other methods. We perform a visual analysis of this difference in results in Figure 6. As Figure 6 illustrates that compared to UM and LA, the endogenous shift, *i.e.*, perturbation difference of SE, PB and GB are obviously heavier, which also proves that the effectiveness of the proposed method to suppress perturbations.

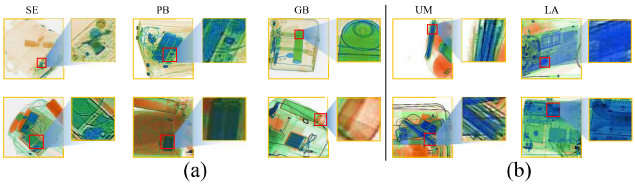

Figure 6. Perturbations around five categories of objects. (a) refers to the heavy endogenous domain shift and (b) refers to the slight endogenous domain shift. Obviously, SE, PB and GB of (a) are heavier than UM and LA of (b).

5.2.2 Results on the simulated dataset

In fact, the endogenous shift also exists in natural images [10, 24–26, 46, 49, 50]. To more comprehensively evaluate the effectiveness of our method in eliminating the endogenous shift, we simulate the endogenous shift in natural images dataset CityScapes [5] by adding adversarial noises. We conduct two groups of experiments to comprehensively evaluate our method on this scenario as follows:

Figure 7. Examples of images generated by two types of adversarial attack methods. Small perturbation caused by adversarial noise makes it difficult to observe the small shift by the naked eyes.

Adaptation from CityScapes to Simulated dataset 1. We select the most commonly-used adversarial attack PGD [27] to generate adversarial noises based on an ImageNet pre-trained ResNet-50 model, so that we could build a new domain of images (the simulated dataset 1). Then, we evaluate the performance of all methods on the adaptation from the original domain to the simulated one. The results on the adaptation from the original Cityscapes dataset [6] to the simulated dataset 1 are shown in Table 5.

Methods	**mAP**	person	rider	car	truck	bus	train	mcyc	bcyc
SO [40]	23.9	31.0	17.6	46.4	16.9	23.8	26.6	14.4	14.8
DA [4]	36.8	48.2	26.1	52.8	31.5	31.6	38.0	35.0	31.4
SWDA [41]	41.4	54.1	30.0	53.5	36.1	33.0	42.2	44.4	37.7
CST [60]	41.1	51.8	31.7	53.5	**39.2**	33.1	42.7	36.8	39.4
CFA [61]	42.8	49.8	30.1	53.8	38.7	**34.1**	46.3	44.1	45.3
PSN (**ours**)	**44.8**	**54.2**	**32.8**	**53.6**	38.8	33.4	**46.9**	**51.2**	**47.5**

Table 5. The **mAP** results (%) and **AP** results (%) on different categories for various methods on the adaptation from the **original Cityscapes dataset** [6] **to the simulated dataset 1**. "SO" refers to the Faster R-CNN model trained on the source domain only.

First, compared with the "SO" model, the performance of domain adaptation methods increases by various margins. Thus, the domain shift, caused by small perturbations in adversarial examples, exists objectively, which is consistent with our hypothesis. Second, for the adaptation from the original Cityscapes dataset to the simulated dataset 1, Table 5 demonstrates that the proposed method PSN is effective. As Table 5 shows, our method outperforms all the baselines and especially achieves a remarkable increase of **20.9**% over the "SO" model. The graph-based CFA [61] achieves the second-best performance and our method outperforms it by **2.0**% and **2.7**%, respectively.

Adaptation from Simulated dataset 1 to 2. We build another new domain images (the simulated dataset 2) by generating adversarial noises based on an ImageNet pre-trained DenseNet model, so that we could build a new adaptation evaluation on the two domains for different adversarial noises. The examples from the two domains (Simulated dataset 1 and 2) are illustrated in Figure 7. Obviously, it is difficult for humans to detect such domain shift with the naked eyes. The results on the adaptation between two simulated datasets are shown in Table 6. For the adaptation between two simulated datasets, Table 6 demonstrates that our method outperforms "SO" by **14.2**% and **15.4**% in two groups of settings, respectively. The attention-based

Methods	Simulated Dataset 1 → Simulated Dataset 2									Simulated Dataset 2 → Simulated Dataset 1								
	mAP	person	rider	car	truck	bus	train	mcyc	bcyc	**mAP**	person	rider	car	truck	bus	train	mcyc	bcyc
SO [40]	35.8	32.5	41.6	52.9	27.2	39.9	30.8	32.9	28.6	35.6	32.7	42.2	42.2	26.6	42.1	25.3	33.4	29.9
DA [4]	47.3	**39.5**	50.4	59.6	46.2	52.9	46.1	45.8	38.3	47.4	38.4	49.8	49.8	46.4	53.3	48.2	44.6	38.3
SWDA [41]	47.9	38.0	49.2	57.9	48.2	56.7	52.5	44.4	36.7	48.3	38.4	49.7	49.7	47.3	57.0	52.7	47.3	38.1
CST [60]	45.7	36.8	48.5	58.2	42.7	52.1	46.0	45.0	36.2	46.5	37.1	49.3	49.3	46.1	53.7	44.1	45.8	37.8
CFA [61]	46.7	39.3	48.6	59.8	46.5	52.5	48.3	43.3	35.4	46.3	38.4	48.8	48.8	47.8	51.6	47.1	43.0	34.1
PSN (**ours**)	**50.0**	39.1	**51.7**	**60.0**	**51.6**	**57.1**	**52.7**	**48.0**	**38.8**	**51.0**	**38.5**	**50.0**	**50.1**	**53.1**	**60.7**	**57.8**	**49.2**	**38.6**

Table 6. The **mAP** results (%) and **AP** results (%) on different categories on the adaptation from **the simulated dataset 1** to **the simulated dataset 2**. Two simulated datasets are based on the Cityscapes dataset [6]. "SO" refers to the "Sourece only" model.

SWDA [41] achieves the second-best performance and our method outperforms it by **2.1**% and **2.7**%, respectively.

5.2.3 Results from CityScapes to Foggy-CityScapes

Our model consists of the prototype adaptation and adversarial learning mechanisms, which can be also exploited to eliminating the apparent shift. In this section, to verify the generalization of our PSN to the common scenario, we also evaluate the performance on the traditional setting, the adapatation from Cityscapes [6] to Foggy-Cityscapes [42].

Methods	**mAP**	person	rider	car	truck	bus	train	mcyc	bcyc
SO [40]	20.8	24.1	29.4	30.6	10.6	25.0	4.6	15.5	26.8
DA [4]	27.6	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1
SWDA [41]	34.3	29.9	42.3	43.5	24.5	36.2	32.6	30.0	35.3
CST [60]	35.9	32.7	44.4	50.1	21.7	45.6	25.4	30.1	36.8
CFA [61]	38.6	34.0	**46.9**	52.1	30.8	43.2	29.9	**34.7**	37.3
PSN (**ours**)	**40.9**	**37.4**	45.2	**53.0**	**31.1**	**48.7**	**38.8**	33.1	**39.2**

Table 7. The **mAP** and **AP** results (%) on different categories on the adaptation from **Cityscapes** [6] to **Foggy-Cityscapes** [42].

The mAP results and AP results on different categories are shown in Table 7. As shown in Table 7, the proposed achieve an remarkable performance of **40.9**% on the weather transfer task, which is the best result among all the counterparts. In particular, we achieve a satisfactory increase of **20.1**% over the "SO" model. Comparing with previous SOTA graph-based adaptation method CFA, our method still improves the mAP by **2.3**%. Although we do not leverage extra attention mechanism, our method still outperforms previous SOTA attention-based SWDA by **6.6**%. As for each category, in the category *truck*, our model achieve a surprising increase of **34.2**%. Overall, the proposed model can achieve stable performance increase on different categories, which verifies the effectiveness of eliminating the apparent shift on common scenarios.

5.3. Ablation Studies

In this section, we conduct several ablation studies to deeply investigate our method on the EDS dataset. We first analysis the effectiveness of GAA module by only integrating the GAA module into the base detection network (the "Source only" model, general practice of the abalation studies for domain adaptation detection methods). Second, we evaluate the effectiveness of the LPA module by only integrating the LPA module into the base detection network. Third, we evaluate the effectiveness of both two modules (*i.e.*, the whole network). The results are shown in Table 8.

Methods	**mAP**	DB	PR	LI	KN	SE	PB	UM	GB	SC	LA
SO [40]	42.3	52.3	30.7	24.8	10.8	30.1	59.3	81.5	29.3	18.7	85.2
+G	44.7	48.2	44.8	25.6	9.7	36.2	55.1	79.9	**37.1**	22.4	87.7
+L	47.0	49.5	47.4	25.4	10.7	**46.2**	60.9	**83.0**	33.7	24.6	88.5
+G+L	**48.3**	**51.9**	**47.5**	**29.9**	**11.3**	43.1	**66.7**	82.9	36.3	**24.5**	**88.7**

Table 8. Average results of ablation studies. "+G" refers to the base model integrated with the GAA and "+L" refers to integrating the LPA module. "+G+L" refers to the whole PSN network.

Table 5 shows that the GAA module helps improving the performance about **2.4**% and the LPA module helps to improve the performance by **4.7**%. Moreover, after the two modules integrated together, the performance of the whole network achieves a remarkable increase of **6.0**%, compared to the widely adopted base "Source only" in other literature.

6. Conclusion

In this paper, we point out that existing cross-domain detection methods mainly study the domain shifts which are usually obvious. We first put forward a novel and inconspicuous types of domain shift in cross-domain detection, endogenous shift. To support study of this issue, we contribute a large-scale benchmark, EDS, by selecting the typical scenario, X-ray security inspection. To deal with the endogenous shift, we further propose the PSN, exploiting local prototype alignment and global adversarial learning mechanism to suppress the two types of perturbations in the endogenous shift. We evaluate the ability of PSN for eliminating the endogenous shift by comprehensive experiments. We release the dataset and code, hoping our study could serve a new perspective to the cross-domain research.

Acknowledge

This work was supported by National Natural Science Foundation of China (62022009, 61872021), Beijing Nova Program of Science and Technology (Z191100001119050), and the Research Foundation of iFLYTEK, P.R. China.

References

[1] Michael Ahdoot, Andrew R Wilbur, Sarah E Reese, Amir H Lebastchi, Sherif Mehralivand, Patrick T Gomella, Jonathan Bloom, Sandeep Gurram, Minhaj Siddiqui, Paul Pinsky, et al. Mri-targeted, systematic, and combined biopsy for prostate cancer diagnosis. *New England Journal of Medicine*, 382(10):917–928, 2020. 1

[2] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4787–4796, 2020. 2

[3] Chaoqi Chen, Zebiao Zheng, Xinghao Ding, Yue Huang, and Qi Dou. Harmonizing transferability and discriminability for adapting object detectors. In *CVPR*, 2020. 2

[4] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018. 2, 6, 7, 8

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 6, 7

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6, 7, 8

[7] Antonio D'Innocente, Francesco Cappio Borlino, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. One-shot unsupervised cross-domain detection. In *European Conference on Computer Vision*, pages 732–748. Springer, 2020. 2

[8] 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. 2, 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 7

[11] Dayan Guan, Jiaxing Huang, Aoran Xiao, Shijian Lu, and Yanpeng Cao. Uncertainty-aware unsupervised domain adaptation in object detection. *IEEE Transactions on Multimedia*, 2021. 2

[12] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019. 2

[13] Zhenwei He and Lei Zhang. Domain adaptive object detection via asymmetric tri-way faster-rcnn. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 309–324. Springer, 2020. 2

[14] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020. 1

[15] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *CVPR*, 2018. 2

[16] Yao Jin, Guang Yang, Ying Fang, Ruipeng Li, Xiaomei Xu, Yongkai Liu, and Xiaobo Lai. 3d pbv-net: an automated prostate mri data segmentation method. *Computers in Biology and Medicine*, 128:104160, 2021. 1

[17] Matthew Johnson-Roberson, Charles Barto, Rounak Mehta, Sharath Nittur Sridhar, Karl Rosaen, and Ram Vasudevan. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983*, 2016. 2

[18] My Kieu, Andrew D Bagdanov, Marco Bertini, and Alberto Del Bimbo. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 546–562. Springer, 2020. 2

[19] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 2

[20] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, 2020. 1

[21] Shuai Li, Jianqiang Huang, Xian-Sheng Hua, and Lei Zhang. Category dictionary guided unsupervised domain adaptation for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1949–1957, 2021. 2

[22] Wanyi Li, Fuyu Li, Yongkang Luo, Peng Wang, et al. Deep domain adaptive object detection: A survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1808–1813. IEEE, 2020. 2

[23] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. *arXiv preprint arXiv:2012.05400*, 2020. 2

[24] 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. 7

[25] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, 2019. 7

[26] Aishan Liu, Jiakai Wang, Xianglong Liu, Bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020. 7

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 7

[28] Xiangxin Meng, Xu Wang, Hongyu Zhang, Hailong Sun, and Xudong Liu. Improving fault localization and program repair with deep semantic features and transferred knowledge. In *ICSE*, 2022. 1

[29] Domingo Mery, Vladimir Riffo, Uwe Zscherpel, German Mondragón, Iván Lillo, Irene Zuccar, Hans Lobel, and Miguel Carrasco. Gdxray: The database of x-ray images for nondestructive testing. *Journal of Nondestructive Evaluation*, 34(4):42, 2015. 1

[30] 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*, 2019. 1

[31] Ahmed Naglah, Fahmi Khalifa, Reem Khaled, Ayman El-Baz, et al. Thyroid cancer computer-aided diagnosis system using mri-based multi-input cnn model. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1691–1694. IEEE, 2021. 1

[32] Yongri Piao, Zhengkun Rong, Miao Zhang, and Huchuan Lu. Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11865–11873, 2020. 1

[33] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *CVPR*, 2020. 1

[34] Binhang Qi, Hailong Sun, Wei Yuan, Hongyu Zhang, and Xiangxin Meng. Dreamloc: A deep relevance matching-based framework for bug localization. *IEEE Transactions on Reliability*, 2021. 1

[35] Haotong Qin, Zhongang Cai, Mingyuan Zhang, Yifu Ding, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Hao Su. Bipointnet: Binary neural network for point clouds. In *International Conference on Learning Representations*, 2020. 2

[36] Haotong Qin, Yifu Ding, Mingyuan Zhang, YAN Qinghua, Aishan Liu, Qingqing Dang, Ziwei Liu, and Xianglong Liu. Bibert: Accurate fully binarized bert. In *International Conference on Learning Representations*, 2021. 2

[37] Haotong Qin, Yifu Ding, Xiangguo Zhang, Aoyu Li, Jiakai Wang, Xianglong Liu, and Jiwen Lu. Diverse sample generation: Pushing the limit of data-free quantization. *arXiv preprint arXiv:2109.00212*, 2021. 2

[38] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. Binary neural networks: A survey. *Pattern Recognition*, 105:107281, 2020. 1

[39] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2250–2259, 2020. 1

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 6, 7, 8

[41] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, 2019. 2, 6, 7, 8

[42] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. 2, 6, 8

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[44] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, 2020. 1

[45] Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *CVPR*, pages 8725–8735, 2020. 1

[46] 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. *arXiv preprint arXiv:2109.05211*, 2021. 7

[47] 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 *IEEE ICCV*, 2021. 1

[48] 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. 1

[49] Jiakai Wang, Aishan Liu, Xiao Bai, and Xianglong Liu. Universal adversarial patch attack for automatic checkout using perceptual and attentional bias. *IEEE Transactions on Image Processing*, 2021. 7

[50] 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*, 2021. 7

[51] Yu Wang, Rui Zhang, Shuo Zhang, Miao Li, YangYang Xia, XiShan Zhang, and ShaoLi Liu. Domain-specific suppression for adaptive object detection. In *CVPR*, 2021. 1

[52] Yanlu Wei, Renshuai Tao, Zhangjie Wu, Yuqing Ma, Libo Zhang, and Xianglong Liu. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 138–146, 2020. 1

[53] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *CVPR*, 2020. 2

[54] Minghao Xu, Hang Wang, Bingbing Ni, Qi Tian, and Wenjun Zhang. Cross-domain detection via graph-induced prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12355–12364, 2020. 2, 5

[55] Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R Qi, and Dragomir Anguelov. Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation. In *ICCV*, 2021. 2

[56] Yan Yang, Na Wang, Heran Yang, Jian Sun, and Zongben Xu. Model-driven deep attention network for ultra-fast compressive sensing mri guided by cross-contrast mr image. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 188–198. Springer, 2020. 1

[57] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 1

[58] Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. Diversifying sample generation for accurate data-free quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15658–15667, 2021. 1

[59] Zhijie Zhang, Yan Liu, Junjie Chen, Li Niu, and Liqing Zhang. Depth privileged object detection in indoor scenes via deformation hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3456–3464, 2021. 1

[60] Ganlong Zhao, Guanbin Li, Ruijia Xu, and Liang Lin. Collaborative training between region proposal localization and classification for domain adaptive object detection. In *ECCV*, 2020. 2, 6, 7, 8

[61] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13766–13775, 2020. 2, 6, 7, 8