Bias-based Universal Adversarial Patch Attack for Automatic Check-out

Aishan Liu[1*], Jiakai Wang[1*], Xianglong Liu[1,2†], Bowen Cao[1], Chongzhi Zhang[1], and Hang Yu[1]

[1] State Key Lab of Software Development Environment, Beihang University, Beijing, China
[2] Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beihang University, Beijing, China

Abstract. Adversarial examples are inputs with imperceptible perturbations that easily misleading deep neural networks (DNNs). Recently, adversarial patch, with noise confined to a small and localized patch, has emerged for its easy feasibility in real-world scenarios. However, existing strategies failed to generate adversarial patches with strong generalization ability. In other words, the adversarial patches were input-specific and failed to attack images from all classes, especially unseen ones during training. To address the problem, this paper proposes a bias-based framework to generate class-agnostic universal adversarial patches with strong generalization ability, which exploits both the perceptual and semantic bias of models. Regarding the perceptual bias, since DNNs are strongly biased towards textures, we exploit the hard examples which convey strong model uncertainties and extract a textural patch prior from them by adopting the style similarities. The patch prior is more close to decision boundaries and would promote attacks. To further alleviate the heavy dependency on large amounts of data in training universal attacks, we further exploit the semantic bias. As the class-wise preference, prototypes are introduced and pursued by maximizing the multi-class margin to help universal training. Taking Automatic Check-out (ACO) as the typical scenario, extensive experiments including white-box/black-box settings in both digital-world (RPC, the largest ACO related dataset) and physical-world scenario (Taobao and JD, the worlds largest online shopping platforms) are conducted. Experimental results demonstrate that our proposed framework outperforms state-of-the-art adversarial patch attack methods. [‡]

Keywords: Universal Adversarial Patch, Automatic Check-out, Bias-based Attack

1 Introduction

Deep learning has demonstrated remarkable performance in a wide spectrum of areas, including computer vision [14], speech recognition [21] and natural lan-

[*] These authors contributed equally to this work.
[†] Corresponding author.
[‡] Our code can be found at https://github.com/liuaishan/ModelBiasedAttack.

Fig. 1. In the real-world scenario like Automatic Check-Out, items (*e.g.*, fruits and chocolates) are often tied with patch-like stickers or tags.

guage processing [27]. Recently, deep learning strategies have been introduced into the check-out scenario in supermarkets and grocery stores to revolutionize the way people shopping (*e.g.*, Amazon Go). Automatic Check-Out (ACO) [30,16,4] is a visual item counting system that takes images of shopping items as input and generates output as a tally of different categories. Customers are not required to put items on the conveyer belt and wait for salesclerks to scan them. Instead, they can simply collect the chosen items and a deep learning based visual recognition system will classify them and automatically process the purchase.

Though showing signi?cant achievements in our daily lives, unfortunately, deep learning is vulnerable to adversarial examples [11,28]. These small perturbations are imperceptible to human but easily misleading DNNs, which creates potential security threats to practical deep learning applications, *e.g.*, auto-driving and face recognition systems [18]. In the past years, different types of techniques have been developed to attack deep learning systems [11,28,9,2,7]. Though challenging deep learning, adversarial examples are also valuable for understanding the behaviors of DNNs, which could provide insights into the blind-spots and help to build robust models [32,31,19].

Besides the well-designed perturbations, the adversarial patch serves as an alternative way to generate adversarial examples and enjoy the advantages of being input-independent and scene-independent. [1,12,18]. In real-world scenarios, patches could be often observed which are quasi-imperceptible to humans. For example, as shown in Fig.1, the tags and brand marks on items in the supermarket. Thus, it is convenient for an adversary to attack a real-world deep learning system by simply generate and stick adversarial patches on the items. However, existing strategies [1,6] generate adversarial patches with weak generalization abilities and are not able to perform universal attacks [22]. In other words, these adversarial patches are input-specific and fail to attack images from all classes, especially unseen ones during training.

To address the problem, this paper proposes a bias-based framework to generate class-agnostic universal adversarial patches with strong generalization ability, which exploits both the perceptual and semantic bias. Regarding the ***perceptual bias***, motivated by the studies [32,10] that DNNs are more perceptually biased towards texture information than shape when making predictions, we first

generate textural priors by extracting textural information from multiple hard examples with style similarities. Our priors contain plenty of textural information from hard examples, thus they are more likely to reveal model uncertainties since DNNs are perceptually biased towards textures. By exploiting perceptual biases of a model, our textural prior is more close to decision boundaries which would promote the universal attack to different classes. As for the ***semantic bias***, since models have semantic preference and bias for different classes, *e.g.*, model prefers wheel for car and fur for dog, we then generate prototypes to help training. A prototype is considered to contain the most representative semantics for a class. Thus, we generate a small number of prototypes by maximizing the corresponding model logits for each class to represent the original images. By exploiting the semantic bias of a model for each class, the prototypes contain more representative features. Thus, training with prototypes will alleviate the heavy dependency on large amounts of data in training universal attacks[24]. Extensive experiments including both the white-box and black-box settings in both the digital-world (RPC, the largest ACO related dataset) and physical-world scenario (Taobao and JD, the worlds largest online shopping platforms) are conducted. Experimental results demonstrate that our proposed framework outperforms state-of-the-art adversarial patch attack methods.

To the best of our knowledge, we are the first to generate class-agnostic universal adversarial patches by exploiting the perceptual and semantic biases of models. With strong generalization ability, our adversarial patches could attack images from unseen classes of the adversarial patch training process or target models. To validate the effectiveness, we choose the automatic check-out scenario and successfully attack the **Taobao** and **JD** platform, which are among the world's largest e-commerce platforms and the ACO-like scenarios.

2 Related work

2.1 Adversarial Attacks

Adversarial examples, which are intentionally designed inputs misleading deep neural networks, have attracted research focus in many scenarios [11,28,15,17]. Szegedy *et al.* [28] first introduced adversarial examples and used the L-BFGS method to generate them. By leveraging the gradients of the target model, Goodfellow *et al.* [11] proposed the Fast Gradient Sign Method (FGSM) which could generate adversarial examples quickly. To improve the generalization ability to different classes, Moosavi *et al.* [22] first proposed an algorithm to compute universal adversarial perturbations for DNNs for object recognition tasks. Mopuri *et al.* [23] proposed a data-free objectives to generate universal adversarial perturbations by maximizing the neuron activation. Further, Reddy *et al.* [24] generated data-free universal adversarial perturbations using class impressions.

Besides, adversarial patch [1], with noise confined to a small and localized patch, emerged for its easy accessibility in real-world scenarios. Karmon *et al.* [12] created adversarial patches using an optimization-based approach with a modified loss function. In contrast to the prior research, they concentrated on

investigating the blind-spots of state-of-the-art image classifiers. Eykholt *et al.* [6] adopted the traditional perturbation techniques to generate attacking noises, which can be mixed into the black and white stickers to attack the recognition of the stop sign. To improve visual fidelity, Liu *et al.* [18] proposed the PS-GAN framework to generate scrawl-like adversarial patches to fool autonomous-driving systems. Recently, adversarial patches have been used to attack person detection systems and fool automated surveillance cameras [29].

2.2 Automatic Check-out

The bedrock of an Automatic Check-out system is visual item counting that takes images of shopping items as input and generates output as a tally of different categories [30]. However, different from other computer vision tasks such as object detection and recognition, the training of deep neural networks for visual item counting faces a special challenge of domain shift. Wei *et al.* [30] first tried to solve the problem using the data argumentation strategy. To improve the realism of the target images, through a CycleGAN framework [33], images of collections of objects are generated by overlaying images of individual objects randomly. Recently, Li *et al.* [16] developed a data priming network by collaborative learning to determine the reliability of testing data.

3 Proposed Framework

In this section, we will first give the definition of the problem and then elaborate on our proposed framework.

3.1 Problem Definition

Assuming $\mathcal{X} \subseteq \mathbb{R}^n$ is the feature space with n the number of features. Supposing $(x_i\ ,y_i)$ is the i-th instance in the data with feature vector $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ the corresponding class label. The deep learning classifier attempts to learn a classification function F: $\mathcal{X} \to \mathcal{Y}$. Specifically, in this paper we consider the visual recognition problem.

An adversarial patch δ is a localized patch that is trained to fool the target model F to wrong predictions. Given an benign image x with its ground truth label y, we form an adversarial example x' which is composed of the original image x, an additive adversarial patch $\delta \in \mathbb{R}^z$ and a location mask $M \in \{0,1\}^n$:

$$x' = (1 - M) \odot x + M \odot \delta, \tag{1}$$

where $\odot$ is the element-wise multiplication.

The prediction result of x'_δ by model F is $y' = F(x'_\delta)$. The adversarial patch makes the model predict the incorrect label, namely $y' \neq y$.

To perform universal attacks, we generate a universal adversarial patch δ that could fool the classifier F on items sampled from distribution μ from *almost all* classes:

$$F(x) \neq F(x + \delta) \quad \text{for } \textit{almost all} \quad x \sim \mu. \tag{2}$$

3.2 The Framework

We propose a bias-based attack framework to generate universal adversarial patches with strong generalization ability. The overall framework can be found in Fig.2.

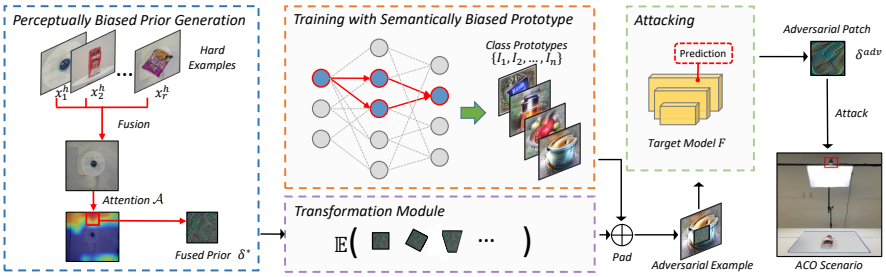


Fig. 2. Our bias-based framework to generate universal adversarial patches. We first generate a perceptually biased prior by fusing textural features from multiple hard examples. Then, we generate semantically biased prototypes to help training universal adversarial patches with a target model F

Recent studies have revealed that DNNs are strongly biased towards texture features when making predictions [10]. Deep learning models are still performing well on patch-shuffled images where local object textures are not destroyed drastically [32].

Thus, we first exploit the perceptual bias of deep models by **generating perceptually biased priors from multiple hard example set** $\mathcal{X}^h = \{x_i^h | i=1, ...r\}$. Textural features are extracted by an attention module $\mathcal{A}$ to fuse a more powerful prior δ^*. We believe the fused prior are more close to decision boundaries of different classes and would boost universal attacks.

Meanwhile, as models have preferences and impressions for different classes, we further exploit the semantic bias of models for each class. To alleviate the heavy dependency on a large amount of data suffered to train universal attacks, we **generate semantically biased prototypes to help training**. As the class-wise preference, prototypes contain the most representative semantics for a class. semantics for each class. Thus, prototypes are generated by maximizing the multi-class margin and used to represent instances from each class. Training with prototypes would reduce the amount of training data required. Thus, we generate prototypes $\{I_1, I_2, ..., I_n\}$ and use them as training data to learn our final adversarial patch δ^{adv} from δ^*.

3.3 Perceptually Biased Prior Generation

Motivated by the fact that deep learning models are strongly biased towards textural features, we first proposed to extract textural features as priors. To

fully exploit the statistic uncertainty of models, we borrow textural features from *hard examples*.

Hard examples appear as instances that are difficult for models to classify correctly. Techniques like hard example mining are used to improve training [8], in which "hard" hence informative examples are spotted and mined. Given a hard example x^h with ground truth label y, assuming that $y^h = F(x^h)$ is the prediction of the model F. The hard example suffices the constraint that $y^h \neq y$ or with relatively low classification confidence. Obviously, a hard example is an instance lying closely to model decision boundaries, and are more likely to cross the prediction surfaces. Thus, using the features from a hard example x^h to train adversarial patches is like "standing on the shoulders of a giant", which would be beneficial to overcome local-optima and gain strong attacking abilities.

To further motivate universal attacks, we extract textural features from multiple hard examples with different labels and fuse them together into a stronger prior. Intuitively, by studying features from multiple hard examples with different labels, our prior would contain more uncertainties for different classes. However, simply learning at pixel-level makes it difficult to extract and fuse textural features. Thus, we introduce the style loss which specifically measures the style differences and encourages the reproduction of texture details:

$$\begin{aligned} \mathcal{L}_s &= \mathbb{E}_k \left[\left|\left| \mathbf{G}(x^*) - \mathbf{G}(x_k^h) \right|\right|_F^2 \right], \\ \mathbf{G}_{ij}(x) &= \sum_k F_{ik}^l(x) \cdot F_{jk}^l(x), \end{aligned} \tag{3}$$

where $\mathbf{G}$ is the Gram matrix of the features extracted from certain layers of the network. $F_{.k}^l$ is the activation of a specific filter at position k in the layer l. x^* is the fused example, and x_k^h is the hard example where k= 1, 2, ..., r.

Besides, entropy has been widely used to depict the uncertainty of a system or distribution. To further improve universal attacks to different classes, we introduce the class-wise uncertainty loss. we increase model prediction uncertainties by minimizing the negative of entropy. Thus, the fused example would be much closer to decision boundaries and obtain low confidence for different classes. It can be written as:

$$\mathcal{L}_u = \mathbb{E}_i \left[\log y^{h,i} \right], \tag{4}$$

where $y^{h,i}$ denotes the model confidence of the i-th class with the fused input x^*.

Thus, to fully exploit the perceptual bias, we optimize the fusion loss function $\mathcal{L}_f$ as follows:

$$\mathcal{L}_f = \mathcal{L}_s + \lambda \cdot \mathcal{L}_u, \tag{5}$$

where λ controls the balance between the two terms.

However, the fused example x^* has a different size with our patches. Thus, an attention module has been introduced to eliminate redundant pixels and generate a textural prior δ^* from the fused example x^*.

$$\delta^* = \mathcal{A}(x^*; F), \tag{6}$$

where $\mathcal{A}(\cdot)$ is a visual attention module that selects a set of suitable visual pixels from the fused sample. These pixels contain the highest stimuli towards model predictions and would be used as textural priors.

Inspired by [25], given a hard example x^h, we compute the gradient of normalized feature maps Z of a specific hidden layer in the model *w.r.t.* y^h. These gradients are global-average-pooled to get the weight matrix which is a weighted combination of feature maps to the hard example x^h :

$$a_{ij} = \sum_{k=1}^{w} \frac{\partial y^h}{\partial Z_{ij}^k} Z_{ij}^k, \tag{7}$$

where a_{ij} represents the weight at position (i, j), Z_{ij}^k is the pixel value in position (i, j) of k-th feature map, and w represents the total feature map number. Note that $i \in [0, u-1]$ and $j \in [0, v-1]$ where u, v are the width and height of Z, respectively. Then, we can combine the pixels with the highest weight to get our textural prior δ^*.

3.4 Training with Semantically Biased Prototypes

With the textural priors generated at the previous stage, we continue to optimize and generate our adversarial patch. To generate universal adversarial perturbations, most of the strategies require a lot of training data [24]. To alleviate the heavy dependency on large amounts of training data, we further exploit the semantic bias.

A *prototype* is an instance that contains the most representative semantics for a class [13]. Prototypes have provided quantitative benefits to interpret and improve deep learning models. Thus, we further exploit the semantic bias of models (*i.e.*, prototypes)for each class. In this stage, we generate class prototypes and use them during training to effectively reduce the amount of training data required.

Thus, inspired by [26], to generate prototypes I representing the semantic preference of a model for each class, we maximize the logits of one specific class. Formally, let $S_t(I)$ denote the logits of class t, computed by the classification layer of the target model. By optimizing the *MultiMarginLoss*, the prototype I_t of class t is obtained:

$$I_t = \underset{x}{\operatorname{argmax}} \frac{1}{\mathrm{C}} \sum_{c \neq t} \max(0, margin - S_t(x) + S_c(x))^p, \tag{8}$$

where x is the input and satisfies the constraint of an RGB image, C denotes the total number of classes and $margin$ is a threshold that controls the multi-class margin. In practice, Adam optimizer is applied to find the optimal prototype of class c with $p = 1$ and $margin = 10$.

To generate adversarial patches misleading to deep models, we introduce the adversarial attack loss. Specifically, we push the prediction label y' of the adversarial example x' (*i.e.*, a clean prototype I appended with the adversarial patch

δ^{adv}) away from its original prediction label y. Therefore, adversarial attack loss can be defined as:

$$\mathcal{L}_t = \mathbb{E}_{I,\delta^{adv}}[P(c = t|I') - \max(P(c \neq t|I'))], \tag{9}$$

where δ^{adv} is the adversarial patch which is initialized as the textural prior δ^*, $P(\cdot)$ is the prediction of the target model to the input, I' is the adversarial example which is composed of the prototype I and adversarial patch δ^{adv}, c means the class, and t denotes the class label of I.

Moreover, recent studies showed that adversarial examples are ineffective to environmental conditions, *e.g.*, different views, illuminations, *etc.* In the ACO scenario, the items are often scanned from different views with different lighting conditions, which would impair the attack ability of our patches. Thus, we further introduce the idea of expectation of transformations to enhance the attack success rate of our adversarial patches in different conditions, as shown in the expectation of conditions $\mathbf{c}$ in Eqn (9).

In conclusion, we first exploit the perceptual bias of models and extract a textural prior from hard examples by adopting the style similarities. To further alleviate the heavy dependency on large amounts of data in training universal attacks, we further exploit the semantic bias to alleviate the heavy dependency on data. As the class-wise preference, prototypes are introduced and pursued by maximizing the multi-class margin. Using the textural prior as initialization, we train our adversarial patches using the prototypes as training data. The illustration of our two-staged adversarial patch attack algorithm can be found in supplementary.

4 Experiments

In this section, we will illustrate the attack effectiveness of our proposed method in different settings in the ACO scenario.

4.1 Dataset and Evaluation Metrics

As for the dataset, we use RPC [30], which is the largest grocery product dataset so far for the retail ACO task. It contains 200 product categories and 83,739 images, including 53,739 single-product exemplary images. Each image is a particular instance of a type of product, which is then divided into 17 sub-categories (*e.g.*, puffed food, instant drink, dessert, gum, milk, personal hygiene, *etc.*).

To evaluate our proposed method, we choose classification accuracy as the metric. Specifically, we further report *top*-1, *top*-3 and *top*-5 accuracy in our experiments. Note that the goal of adversarial attacks is compromising the performance of the model, *i.e.*, leading to worse values of the evaluation metrics above.

4.2 Experimental Settings

The input image is resized to 512×512 and the patch size is fixed at 32×32. The size of patches only accounts for 0.38% of the size of images. To optimize the loss, we use Adam optimizer with a learning rate of 0.01, a weight decay of 10^{-4}, and a maximum of 50 epochs. To get hard examples, we first run the target model over the training set once to get model predictions for each instance. Then, we select 200 images that are misclassified by the model with the lowest confidence as our hard examples. All of our code is implemented in Pytorch. The training and inference processes are finished on an NVIDIA Tesla k80 GPU cluster.

As for the compared methods, we choose the state-of-art adversarial patch attack methods including AdvPatch [1], RP_2 [5], and PSGAN [18]. We follow their implementations and parameter settings. To conduct fair comparisons, we adopt the same backbone models for our method and the compared ones in our experiments. Similar to [22], we use 50 item samples per class (10,000 in total) as the training data for the compared methods. We also extract 15 prototypes for each class (3,000 in total) as the training data for our method. With respect to the models, we follow [16] for the ACO task and use ResNet-152 as the backbone. As for the *margin*, we set the it as 10 since we found the model is insensitive to it (we tested 2, 4, 6, 8, 10, and 12 for *margin* and found similar results). To further improve the attack success rate of adversarial patches against different environments, we introduce transformations as follows:

- **Rotation**. The rotation angle is limited in $[-30°, 30°]$.
- **Distortion**. The distortion rate, *i.e.*, the control argument, moves in $[0, 0.1]$.
- **Affine Transformation**. The affine rate changes between 0 and 4.

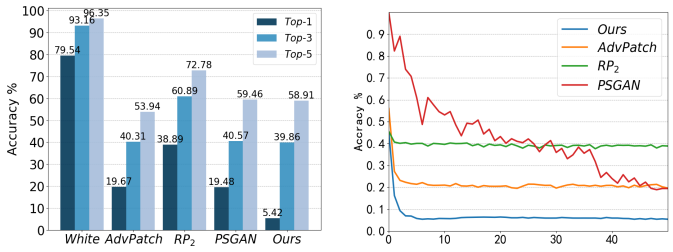
4.3 Digital-world Attack

In this section, we evaluate the performance of our generated adversarial patches on the ACO task in the digital-world in both white-box and black-box settings. We also use a white patch to test the effectiveness of the adversarial attack (denoted as "White").

As for the **white-box** attack, we generate adversarial patches based on a ResNet-152 model and then attack it. As shown in Fig.3(a), our method outperforms other compared strategies with large margins. In other words, our adversarial patches obtain stronger attacking abilities with lower classification accuracy contrast to others, i.e., **5.42%** to 21.10%, 19.78%, and 38.89% in *top*-1 accuracy.

As for the **black-box** attack, we generate adversarial patches based on ResNet-152, then use them to attack other models with different architectures and unknown parameters (*i.e.*, VGG-16, AlexNet, and ResNet-101). As illustrated in Table 1, our generated adversarial patches enjoy stronger attacking abilities in the black-box setting with lower classification accuracy for different models.

Besides the classification accuracy, Fig.3(b) shows the training process of adversarial patches using different methods. After several training epochs, the



(a) White-box Attack (b) Training Process

Fig. 3. (a) shows the White-box attack experiment in the digital-world with ResNet-152. Our method generates the strongest adversarial patches with the lowest classification accuracy. (b) denotes the training process of different methods

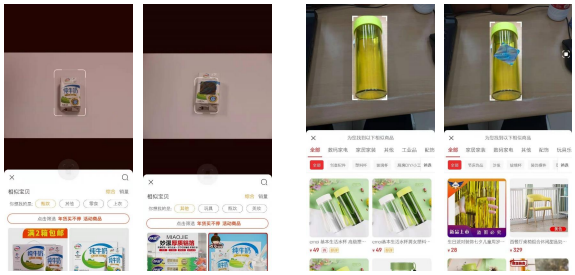
attacking performance of our generated patches becomes stable and keeps the best among all. However, the performance of other methods still vibrates sharply. It is mainly due to the weak generalization ability of other methods. Thus, they achieve different accuracy when attacking different classes showing sharp vibrations.

4.4 Real-world Attack

To further validate the practical effectiveness of our generated adversarial patches, a real-world attack experiment is conducted on several online shopping platforms to simulate the ACO scenario. We use **Taobao** and **JD**, which are among the biggest e-commerce platforms in the world. We take 80 pictures of 4 different real-world products with different environmental conditions (*i.e.*, angles

Table 1. Black-box attack experiment in the digital-world with VGG-16, AlexNet, and ResNet-101. Our method generates adversarial patches with strong transferability among different models

Model	Method	*top*-1	*top*-3	*top*-5
VGG-16	AdvPatch	73.82	**90.73**	**94.99**
	RP_2	81.25	94.65	97.10
	PSGAN	74.69	91.25	96.12
	Ours	**73.72**	**91.53**	**95.43**
AlexNet	AdvPatch	51.11	72.37	79.79
	RP_2	68.27	86.49	91.08
	PSGAN	49.39	72.85	82.94
	Ours	**31.68**	**50.92**	**60.19**
ResNet-101	AdvPatch	56.19	80.99	91.52
	RP_2	73.52	93.75	98.13
	PSGAN	51.26	79.22	90.47
	Ours	**22.24**	**51.32**	**60.28**



(a) Taobao (b) JD

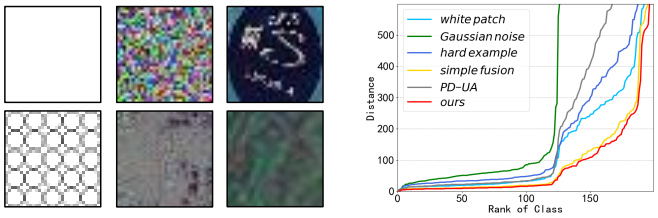
Fig. 4. Attack Taobao and JD platform with our adversarial patches. The `milk` in (a) and the `plastic cup` in (b) are recognised as the `decorations` and the `aluminum foil` when we stick our adversarial patches, respectively

{-30°, -15°, 0°, 15°, 30°} and distances {0.3m, 0.5m, 0.7m, 1m}). The *top*-1 classification accuracy of these images is 100% on Taobao and 95.00% on JD, respectively. Then, we print our adversarial patches by an HP Color LaserJet Pro MFP M281fdw printer, stick them on the products and take photos with the combination of different distances and angles using a Canon D60 camera. A significant drop in accuracy on both platforms can be witnessed with low classification accuracy (*i.e.*, **56.25%** on Taobao, **55.00%** on JD), which is much lower than the compared methods, concretely 66.25%, 61.25%, and 66.25% on Taobao, 72.50%, 68.75%, and 63.75% on JD (the results of compared methods are in following orders: AdvPatch, RP_2, and PSGAN). The results demonstrate the strong attacking ability of our adversarial patches in real-world scenarios on practical applications. Visualizations can be found in Fig.4.

4.5 Generalization Ability

In this section, we further evaluate the generalization ability of adversarial patches on unseen classes. We perform two experiments using the backbone model (ResNet-152), including attacking unseen item classes of adversarial patch training process and target models. For attacking unseen classes of the patch training process, we first train patches on a subset of the dataset, *i.e.*, only images from 100 classes are used *w.r.t.* 200 classes (we use prototypes for our method and item images for compared methods). According to the results in Table 2, our framework generates adversarial patches with strong generalization ability and outperforms other compared methods with huge margins (*i.e.*, **7.23%** to 40.28%, 31.62%, and 60.87%).

Meanwhile, we also tested the generalization ability on classes that have never been "seen" by the target model. Specifically, we train our patches on the RPC dataset and test them on the Taobao platform. We select 4 items and stick adversarial patches on them and take 64 pictures. The categories of the items are unseen to target models (not in the 200 classes for ResNet-152), but



(a) Different Priors (b) Boundary Distance

Fig. 5. (a) shows different priors used to generate adversarial patches. They are white patch, Gaussian noise, hard example, PD-UA, simple fusion, and **textural prior** respectively, from up to down, left to right. (b) is the decision boundary distance analysis, where fused prior achieves the smallest decision boundary distances for each class.

known to the Taobao platform. Interestingly, after attacks, the *top*-1 accuracy on Taobao is **65.63%**. Though our patches are not trained to attack some specified classes, they still generalize well to these unseen classes. Thus, we can draw the conclusion that our framework generates universal adversarial patches with strong generalization abilities to even unseen classes.

4.6 Analysis of Textural Priors

Since textural priors have improved universal attacks, a question emerges: "Why and how the textural priors are beneficial to universal adversarial attacks?" Thus, in this section, we further study the effectiveness of our textural priors.

Training from Different Priors To demonstrate the effectiveness of our textural priors, we begin to study by initializing patches through different priors, *e.g.*, white patch, Gaussian noise, hard example, PD-UA [20], simple fusion, and our textural prior (denoted as "ours"). In contrast to our textual prior, we use the same amount of simple examples to generate the simple version of fused prior (denoted as "simple fusion"). Other experimental settings are the same as the settings of the digital-world attack. The visualization of them can be found in Fig.5(a). We train 6 adversarial patches with all the same experimental settings except for the initialization. The corresponding accuracy after attacking are

Table 2. Attack on unseen classes. Our method generates adversarial patches with the strongest generalization abilities showing lowest accuracy compared with other methods

Method	AdvPatch	RP_2	PSGAN	**Ours**
top-1	40.28	60.87	31.62	**7.23**

17.67% (white patch), 18.96% (Gaussian noise), 16.11% (hard example), 21.10% (PD-UA), 24.09% (simple fusion), **5.42%** (ours). It shows that our fused priors offer adversarial patches the strongest attacking ability.

Decision Boundary Distance Analysis The minimum distance to the decision boundary among the data points reflects the model robustness to small noises [3]. Similarly, the distance to decision boundaries for an instance characterizes the feasibility performing attack from it. Due to the computation difficulty of decision boundary distance for deep models, we calculate the distance of an instance x to specified classes *w.r.t.* the model prediction to represent the decision boundary distance. Given a learnt model F and point x_i with class label y_i $(i = 1, \ldots, N)$, for each direction $(y_j, i \neq j)$ we estimate the smallest step numbers moved as the distance. We use the L_2 norm Projected Gradient Descent (PGD) until the model's prediction changes, i.e., $F(x_i) \neq y_i$.

As shown in Fig.5(b), our textural priors obtain the minimum distance in each direction compared to other initialization strategies. It explains the reason that our textural prior performs stronger adversarial attacks because it is more close to the decision boundaries.

4.7 Ablation Study

In this section, we investigate our method through ablation study. Due to the limitation of paper length, we put more ablations in the supplementary.

The Effectiveness of Class Prototypes We first investigate the amount of data required with our framework to train adversarial patches by solely using prototypes or item images, respectively. Specifically, we first train adversarial patches with 1000 prototypes as Ours_{P1000}. Then, we randomly select 1000, 2000, 4000, 10000 item images from the RPC dataset to train adversarial patches, respectively (denoted by Ours_{I1000}, Ours_{I2000}, Ours_{I4000}, and Ours_{I10000}). The results in Table 3 show that to achieve the approximate attacking ability in Ours_{P1000} setting, twice more items images are required. It indicates the representative ability of class prototypes for different classes. Besides, we also study the time efficiency of generating prototypes. In our practice, it takes 1.8 hours to generate 1000 prototypes, and another 3.3 hours to train the model (**5.1 hours** in total). To achieve similar performance, it takes 6.6 hours to train a model using 2000 original images, which indeed spends more time. Also, it's time-consuming to collect original images in practice, let alone the cost of preprocessing, etc.

Table 3. The *top*-1 accuracy of the adversarial patches obtained using different amount of training data. Our prototypes need only half the data to achieve a similar performance compared to original method.

Settings	Ours_{P1000}	Ours_{I1000}	Ours_{I2000}	Ours_{I4000}	Ours_{I10000}
top-1	**6.51**	12.43	6.57	6.10	5.40

The Effectiveness of Fusion Loss We further analyze the effectiveness of fusion loss by using $\mathcal{L}_u$ only or $\mathcal{L}_s$ only respectively for the fusion loss $\mathcal{L}_f$ in 5. As shown in Table 4, the original $\mathcal{L}_f$ achieves better performance in both white-box (ResNet-152) and black-box (VGG-16, AlexNet, ResNet-101) settings. We believe that the adversarial patch benefits both attacking ability and generalization ability from the $\mathcal{L}_f$ loss. As for λ, we achieve the best performance when it is around 1. The accuracy with different λ values (0.1, 1, and 10) are 5.47%, **5.42%**, and 12.01% respectively.

Table 4. Ablation study on fusion loss. The original $\mathcal{L}_f$ achieves the best results.

Method	White-box	Black-box		
	ResNet-152	VGG-16	AlexNet	ResNet-101
$\mathcal{L}_u$only	5.47%	75.21%	49.42%	61.82%
$\mathcal{L}_s$only	15.73%	77.05%	72.65%	72.55%
$\mathcal{L}_f$	**5.42%**	**73.72%**	**31.68%**	**22.24%**

5 Conclusions

In this paper, we proposed a bias-based attack framework to generate class-agnostic universal adversarial patches, which exploits both the perceptual and semantic bias of models. Regarding the perceptual bias, since DNNs are strongly biased towards textures, we exploit the hard examples which convey strong model uncertainties and extract a textural patch prior from them by adopting the style similarities. The patch prior is more close to decision boundaries and would promote attacks. To further alleviate the heavy dependency on large amounts of data in training universal attacks, we further exploit the semantic bias. As the class-wise preference, prototypes are introduced and pursued by maximizing the multi-class margin to help universal training. Taking ACO as the typical scenario, extensive experiments are conducted which demonstrate that our proposed framework outperforms state-of-the-art adversarial patch attack methods.

Model biases, especially texture-based features, has been used to perform adversarial attacks. In contrast, can we improve model robustness by eliminating the textural features from the training data? We leave it as future work.

Acknowledgement

This work was supported by National Natural Science Foundation of China (61872021, 61690202), Beijing Nova Program of Science and Technology (Z191100001119050), and Fundamental Research Funds for Central Universities (YWF-20-BJ-J-646).

References

1. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
2. Chen, W., Zhang, Z., Hu, X., Wu, B.: Boosting decision-based black-box adversarial attacks with random sign flip. In: Proceedings of the European Conference on Computer Vision (2020)
3. Cortes, C., Vapnik, V.: Support-vector networks. MACH LEARN (1995)
4. Ekanayake, P., Deng, Z., Yang, C., Hong, X., Yang, J.: Naïve approach for bounding box annotation and object detection towards smart retail systems. In: SpaCCS (2019)
5. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning models. arXiv preprint arXiv:1707.08945 (2017)
6. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning models. In: CVPR (2018)
7. Fan, Y., Wu, B., Li, T., Zhang, Y., Li, M., Li, Z., Yang, Y.: Sparse adversarial attack via perturbation factorization. In: European conference on computer vision (2020)
8. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
9. Gao, L., Zhang, Q., Song, j., Liu, X., Shen, H.: Patch-wise attack for fooling deep neural network. In: European Conference on Computer Vision (2020)
10. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
12. Karmon, D., Zoran, D., Goldberg, Y.: Lavan: Localized and visible adversarial noise. arXiv preprint arXiv:1801.02608 (2018)
13. Kim, B., Rudin, C., Shah, J.A.: The bayesian case model: A generative approach for case-based reasoning and prototype classification. In: NeurIPS (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)
15. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
16. Li, C., Du, D., Zhang, L., Luo, T., Wu, Y., Tian, Q., Wen, L., Lyu, S.: Data priming network for automatic check-out. arXiv preprint arXiv:1904.04978 (2019)
17. Liu, A., Huang, T., Liu, X., Xu, Y., Ma, Y., Chen, X., Maybank, S., Tao, D.: Spatiotemporal attacks for embodied agents. In: European Conference on Computer Vision (2020)
18. Liu, A., Liu, X., Fan, J., Ma, Y., Zhang, A., Xie, H., Tao, D.: Perceptual-sensitive GAN for generating adversarial patches. In: AAAI (2019)
19. Liu, A., Liu, X., Zhang, C., Yu, H., Liu, Q., Wu, Q., Tao, D.: Training robust deep neural networks via adversarial noise propagation. arXiv preprint arXiv:1909.09034 (2019)
20. Liu, H., Ji, R., Li, J., Zhang, B., Gao, Y., Wu, Y., Huang, F.: Universal adversarial perturbation via prior driven uncertainty approximation. In: ICCV (2019)

21. Mohamed, A.r., Dahl, G.E., Hinton, G.: Acoustic modeling using deep belief networks. IEEE T AUDIO SPEECH (2011)
22. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017)
23. Mopuri, K.R., Ganeshan, A., Radhakrishnan, V.B.: Generalizable data-free objective for crafting universal adversarial perturbations. IEEE T PATTERN ANAL (2018)
24. Reddy Mopuri, K., Krishna Uppala, P., Venkatesh Babu, R.: Ask, acquire, and attack: Data-free uap generation using class impressions. In: CVPR (2018)
25. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? arXiv preprint arXiv:1611.07450 (2016)
26. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
27. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. NeurIPS (2014)
28. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
29. Thys, S., Van Ranst, W., Goedemé, T.: Fooling automated surveillance cameras: adversarial patches to attack person detection. In: CVPRW (2019)
30. Wei, X.S., Cui, Q., Yang, L., Wang, P., Liu, L.: Rpc: A large-scale retail product checkout dataset. arXiv preprint arXiv:1901.07249 (2019)
31. Zhang, C., Liu, A., Liu, X., Xu, Y., Yu, H., Ma, Y., Li, T.: Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. arXiv preprint arXiv:1909.06978 (2019)
32. Zhang, T., Zhu, Z.: Interpreting adversarially trained convolutional neural networks. arXiv preprint arXiv:1905.09797 (2019)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)