Analysis and Countermeasure Design on Adversarial Patch Attacks

Yinan Fu(✉), Xiaolong Zheng, Peilun Du, and Liang Liu

Beijing University of Posts and Telecommunications, Beijing, China
fyn@bupt.edu.cn

Abstract. Adversarial patch is an image-independent patch that misleads deep neural networks to output a targeted class. Existing defense strategies mainly rely on patch detection based on the frequency or semantic gaps between the patch and clean image. But we found that they are effective because the gap is huge. This is because existing patch attacks only look for an effective patch instead of the optimized patch that minimizes the gap. We then propose two improved patches, enhanced and smoothed patches, to reduce the gap. Consequently, the decision boundary for adversarial examples of the existing defense means is successfully obscured. To cope with the improved patches, we propose a defense method based on image preprocessing. We leverage multi-scale Gaussian blur to amplify the reduced gap between the patch and clean image. Due to the dense information of patches, for a patch, the dissimilarities of Gaussian blurs with different scales are higher than that of clean images. By enhancing the local multi-scale details and weakening them in another scale set, we maximize its effect on patch with high-frequency information. In this way, our defense method can efficiently distort adversarial patches and cause only a negligible impact on clean images.

Keywords: Adversarial patch · Defense

1 Introduction

The past decade has witnessed the prosperity of Deep Learning. Deep neural networks (DNNs) are widely used in computer vision [9], pattern recognition [10], natural language processing [11], and autonomous driving [15]. However, recent studies have revealed that DNNs are vulnerable to adversarial examples that fool the classifier with subtle modifications. Since adversarial examples modify pixels in the whole image, it is not easy to launch physical attacks. Different from the imperceptible changes of adversarial examples, the adversarial patch is an image-independent patch that misleads the classifier to output a targeted class for any image (Fig. 1). Since it is image-independent, adversarial patches can be printed or placed in the scene to launch physical attacks without any prior knowledge of the scene [12].

L. Cui and X. Xie (Eds.): CWSN 2021, CCIS 1509, pp. 181–195, 2021.
https://doi.org/10.1007/978-981-16-8174-5_14

Fig. 1. Adversarial patches for ResNet-50.

As a targeted attack method, the adversarial patch imitates images of the targeted class to mislead the classifier focusing on the patch and making wrong recognition. Nevertheless, a gap between the patch and clean images is inevitable, leaving room for defense methods to differentiate the patch from clean images. Since a successful patch will be the salient activation source, the defense method can locate the patch and compare its feature with the expected features of the predicted class to detect the inconsistency [6]. Besides, as a universal attack method [13], adversarial patches contain more high-frequency information in a concentrated area than normal images to achieve their versatility. Hence, defense methods can also detect and distort the patch based on this high-frequency information which means the higher gradient of an image.

However, we analyze those defense methods and found that they are effective because the gap between the patch and the clean image is huge. This is because the goal of generation is only finding the patch successfully fool the threat model instead of the patch that is more robust to the defense. Based on this observation, we propose the enhanced and smoothed adversarial patches to respectively obscure the decision boundary of the feature inconsistency detection- and high-frequency information-based defense methods. To enhance the patch with more similar features of the targeted class, we leverage an antagonistic training strategy at the early stage of the patch generation training. To cope with the high-frequency information-based defense method, we propose generating the patch with proactive smoothing. Experimental results demonstrate our improved adversarial patches can significantly decrease the defending ability of existing methods.

To cope with the improved patches, we propose EGP, a new defense method based on image preprocessing that enlarges the reduced gaps of improved patches by amplifying the frequency difference between patches and the original image.

Different from the conventional image preprocessing based defense methods such as blur or JPEG that process the whole image, EGP only processes the key regions of the input image. We leverage multi-scale Gaussian blur [8] to obtain the multi-scale details of the image which will amplify the frequency properties of the original image. We magnify the details and add them to the original image to further enlarge the effect of subtle frequency differences on image processing. With this preprocessing, the clean images will be enhanced with details, while patches will be distorted due to the much higher frequency. To reduce the impact of preprocessing on the clean images, we further propose weakening the local multi-scale details by another Gaussian kernel set. Experimental results show that EGP can significantly improve the classification accuracy of adversarial examples, and cause little impact on clean images.

The main contributions of this paper are as follows: 1) We analyze the effectiveness of existing defense methods against adversarial patches and find that defense methods are effective because the generated adversarial patches are just effective to successfully fool the classifier rather than optimal with the minimized gap to the characteristic of a clean image. 2) Based on the observation, we propose two patch generation methods to obtain the enhanced and smoothed patches that can effectively obscure the decision boundary for adversarial patches and further reduce the effectiveness of current defense methods. 3) As for the countermeasure design, we propose a new defense method based on image preprocessing. The key idea is enlarging the reduced gap between the patch and the clean image by multi-scale Gaussian blur.

2 Related Work

2.1 Adversarial Patch Attacks

Adversarial patch [1], a localized patch, which enjoys strong robustness to position and angle alternation. To further optimize the generation of adversarial patch, Karmon [2] created adversarial patch using optimized loss function and they concentrated on the selection of categorys for targeted attack. Duan [4] adopted style loss and content loss to generate imperceptible patches. Recently, Liu [3] proposed a universal adversarial patch generation framework based on model bias, which can effectively attack the invisible categories in the model training process.

Defenses Against Adversarial Patch Attacks. Naseer [5] proposed a local gradient smoothing scheme to resist adversarial patch attacks. To eliminate the influence of noises, the local high gradient region of the image is detected and smoothed. Hayes [7] put forward a defense strategy based on image inpainting. They discover the location of patches and further leverage image inpainting technology to remove them. To address the lack of versatility and computation of previous methods, Xu [6] was concerned about the feature dissimilarity between input and image of the corresponding category. If the degree of dissimilarity exceeds a threshold, the input is considered to be an adversarial example.

3 Analysis and Improvement on Adversarial Patch

In this section, we mainly analyze the defense mechanism of the existing defense methods. Based on the analysis results, we improve the universal adversarial patch generation method.

3.1 Existing Adversarial Patch

Adversarial Patch [1] generates a universal patch, which can be applied to any image x in the dataset X to mislead the image into the target category regardless of the scale, orientation, or location of the patch.

Given an adversarial patch p, an image x in dataset X, a target class t, a random location in the location space of images $l \in L$, and a random angle transformation over a set of angle transformations $t \in T$, the patch p is then placed in a location l of image x. The algorithm renovates the patch iteratively by optimizing the loss function:

$$\hat{p} = \arg\max_{p} E_{x \in X, t \in T, l \in L}[logPr(\hat{y}|A(p, x, l, t)] \tag{1}$$

3.2 Enhanced Adversarial Patch

Although adversarial patches are effective for digital or physical world attacks, it should be noted that they tend to be abrupt and unreal. The feature similarity between the patch and the image of the predicted class is not high. Therefore, both artificial means and the existing defense method [6] can distinguish them without difficulty. Consequently, we consider proposing an enhanced adversarial patch, which is an improvement based on the original method [1]. Aiming at mining deeper semantic information about the target category, the enhanced adversarial patches can be more realistic and closer to the target category.

Our improvement focus on the early input images in the train set (a pre-training process). For the early inputs, based on the optimization of the objective function, a strategy of antagonistic training is leveraged. The brief ideology of antagonistic training is shown in Fig. 2.

Early generated patches always contain insufficient features. It is due to the over-fitting of the white-box model that an image with an early patch is misclassified successfully. The purpose of patch training is to find an effective patch by fitting the white-box model, rather than to generate a more realistic adversarial patch with details, which leads to premature convergence on the white box. For solving this problem, We feed the adversarial patches generated in the intermediate process to retrain the target model for fooling and generating patch. Specifically, we consider retraining the adversarial image that has been successfully misclassified with the original tag, and then updating the target model for fooling and generating a patch constantly. By constantly feeding the adversarial images with early generated patches to retrain the target model, the target model will be more robust and more difficult to be fooled. While the target model is

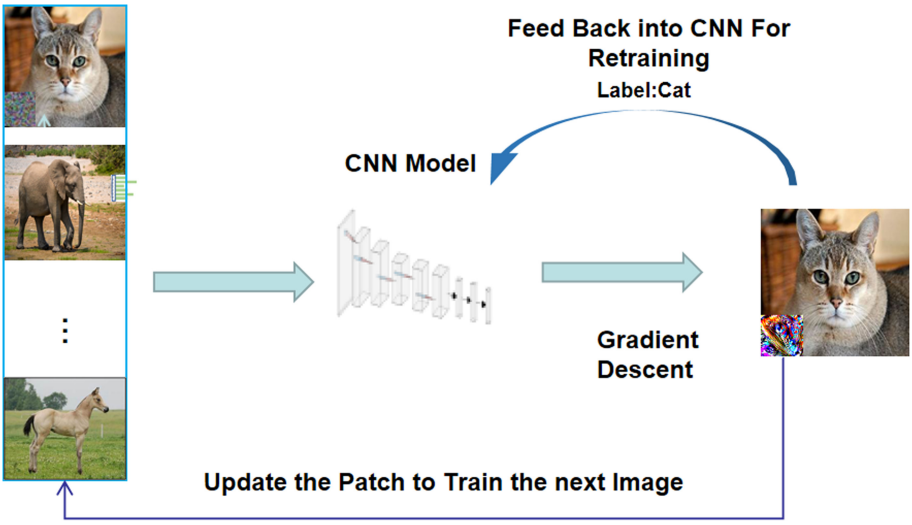

Fig. 2. Antagonistic pretraining of adversarial patch.

more robust, the generated patch is stronger. In short, we strengthen the update of the patch through the continuous update of the target model used for generating patch. In this way, we force the patch training process to continue learning enhanced adversarial features, which reduce the gap between adversarial patch and target image.

3.3 Smoothed Adversarial Patch

Adversarial patches always contain more high-frequency information, which can be the motivation of diverse defense methods. Therefore, by adding smooth processing intermittently during the updating of adversarial patches, we propose generating smoothed patches (Fig. 3).

In Adv Cam [4], adversarial patch for a specific image can generate imperceptible patches by adding style and content loss to the optimization objective function. However, as a targeted universal adversarial patch, we can not add the smoothing loss to the objective loss function to achieve the smoothed patch, because it will cause the direction of smoothing unable to focus on the target category. Therefore, we consider periodically smoothing the patch slightly in the training process to guide the update. Specifically, regarding to slight smoothing, we can get the intermediate smoothed patch p_{sm} as follow:

$$p_{sm} = k * (p_g - p) + p \tag{2}$$

p is the original patch, p_g is a Gaussian blur of p, k is a fuzzy coefficient ($k < 1$) which is used to get images with different blur levels.

Fig. 3. Comparison of three different adversarial patches.

4 Defense Against Adversarial Patch Attacks

In this section, we will describe our defense methodology against adversarial patch attacks in detail.

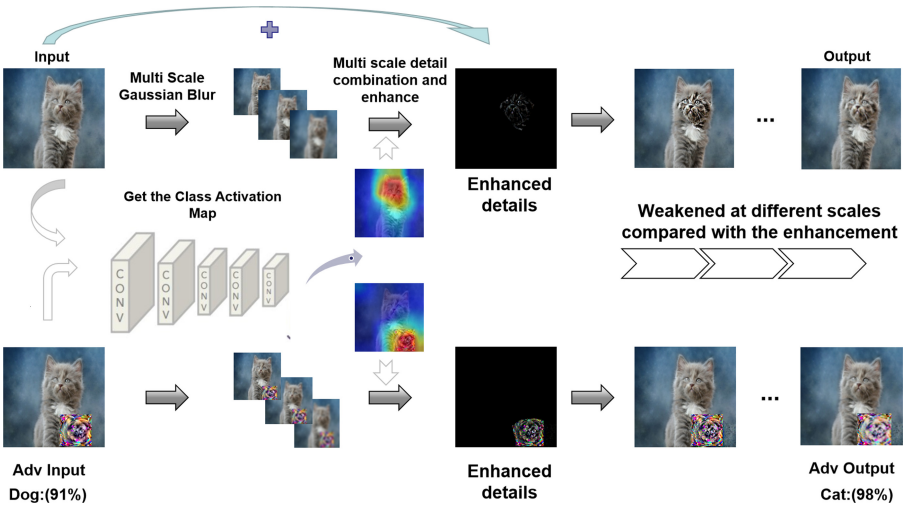

Fig. 4. EGP defense architecture.

4.1 Obtain the Attention Heatmap Matrix

The representative classification models based on convolutional neural networks pay more attention to the local features of images. The adversarial patch will be the salient activation source if the attack successes. Therefore, compared with the global image processing, we aim to process the local areas with strong attention of the models, which causes less impact on the clean image. Consequently, we

need to derive the importance of diverse features in different regions to model decisions, namely, the model's attention heatmap matrix [14]. Specifically, we regard α_k^t as the sensitivity to the $k-th$ channel of the output feature map of the last layer A^k about category t. Then we take α_k^t as weights and combine them linearly. Furthermore, the intermediate result of the weighted combination is fed into the activation function to output the required heatmap matrix M_t.

$$\alpha_k^t = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^t}{\partial A_{ij}^k} \tag{3}$$

$$M_t = RELU \left\{ \sum_k \alpha_k^t A^k \right\} \tag{4}$$

Here Z is a normalizing constant such that $\alpha_k^t \in [-1, 1]$. k is the sequence number of the channel dimension of the feature map, i and j are the sequence number of the width and height dimension respectively, and t is the target category.

4.2 Enhance the Local Multi-scale Details

As an indiscriminate processing method, our method aims to make the distortion of clean image I_c small, but the distortion of adversarial example I_{adv} large.

$$max \left\{ D\left(I_c, I\right) - D\left(I_{adv}, I\right) \right\} \tag{5}$$

D is the measurement of image distance.

Compared with clean images, the feature distribution of patches is demonstrated to be dense and irregular with higher local frequency. Therefore, the dissimilarities of Gaussian blurs with different scales are higher than that of clean images. Specifically, our enhanced image can be obtained by Equation (6). For the original image I, by fusing the Gaussian blur decrease values between different scales, we can obtain the contour details of the target category for the clean image I_c. However, for the adversarial example I_{adv}, details tend to be dense and intensive after the same treatment as clean image I_c.

$$I_{en} = norm \left\{ I + \lambda \cdot \sum_{g_i, g_j \in G} w_{ij} \left(g_i - g_j\right) * M \right\} \tag{6}$$

G is the Gaussian fuzzy set with different Gaussian kernels. λ is the magnification factor ($\lambda > 1$) to enlarge the detailed information. w is the proportional coefficient. M is the mask matrix from the heatmap matrix that limits the processing to the local key region. *norm* is the normalization process. Then, we multiply the multi-scale details by a magnification factor λ to enhance the details of input image. Furthermore, we add the enlarged local details to the original image I and normalize to obtain the enhanced image I_{en}. As shown in Fig. 4, the clean image appears as a regional detail enhancement, while the adversarial example shows high distortion at the location of the patch.

4.3 Weaken the Local Multi-scale Details

To minimize the influence of preprocessing on the clean images and cause further distortion on the adversarial examples, we consider weakening the local multi-scale details in another Gaussian kernel set based on the local multi-scale details enhancement. The enhanced image I_{en} is regarded as the input, we weaken the local multi-scale details in another Gaussian kernel set $G_{de} = \{g_1, g_2, \ldots, g_n\}$. It should be emphasized that although the patch after details enhancement has been distorted and it is difficult to recover after details weakening. To avoid the reduction of distortion on pixel value caused by processing in the same scale set, we think that it is better to weaken the details in another scale set. For a clean image, the details obtained in another scale set are contour details, which are similar to the details obtained during the enhancement process. However, as for a patch, the weakened details in another scale set are not similar to the enhanced details, because the patch is already distorted and more sensitive to different scales.

The output images can be obtained by Eq. (7). We can still obtain the multi-scale details of the enhanced clean image, which is similar to the multi-scale detail information obtained during enhancement. Consequently, the distance between the clean image and the original image is reduced after weakening. On the contrary, for the adversarial example, since the enhanced image I_{en} has been distorted, further detail weakening under the Gaussian blur of another scale set will only aggravate the distortion of the adversarial example. As shown in Fig. 4, both the clean image and the adversarial example are correctly classified as a cat.

$$I_{out} = norm\left\{ I_{en} - \sum_{g_i, g_j \in G_{de}} w_{ij} \left(g_i - g_j\right) * M \right\} \tag{7}$$

I_{en} is the enhanced image, G_{de} is a Gaussian fuzzy set without intersection with G, I_{out} is the output image.

5 Experiments

5.1 Feasibility of Attack Reinforcement

Experimental Setup for Attack. We consider the validation set available with the ImageNet-2012 dataset in our experiments. We choose images of 10 categories comprised of 10000 images to generate our adversarial patches. The pre-training models are used for patch training with a learning rate of 0.0005. The size of patch is 70×70 covering 10% of the image. During the training of adversarial patches, the number of iterations is set to 8, 100 images of each category are randomly selected in each iteration. Furthermore, to generate the enhanced adversarial patch, we leverage our antagonistic pretraining strategy for the top 100 input images of the first iteration (early inputs) and the learning rate is set to 0.001 to retrain the model to be attacked. Besides, to generate smoothed

adversarial patches, we implement a slight Gaussian blur on the updated adversarial patch every 50 inputs during the generation of the ordinary universal adversarial patch. The Gauss kernel is 5 and the fuzzy coefficient is set to 0.2.

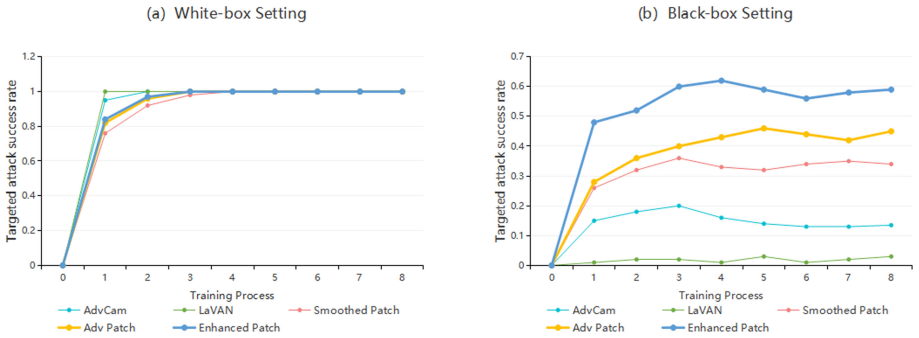

Fig. 5. A comparison of existing methods and our methods for creating adversarial patches. Note that these patches are generated by single model ResNet-50 and the targeted attack success rate refers to the average attack success rate tested in black boxes for the black-box setting.

Attack Performance. We evaluate the performance of our enhanced and smoothed adversarial patches in both white-box and black-box settings. As for the black-box attack, we generate adversarial patches based on ResNet-50, then use them to attack other models with different architectures and unknown parameters (i.e., VGG-16, Inception-V3, and ResNet-152) and record the average target attack success rate.

As indicated in Fig. 5, our generated enhanced adversarial patch enjoys stronger transferability. The enhanced patch avoids the overfitting of the white-box model through the antagonistic pre-training process of the adversarial patch. This process enables the patch to mine the deeper semantic information of the target category rather than to meet the judgment bias of the white-box model, so that it can have better migration ability under the black-box setting. However, regarding our generated smoothed adversarial patch, the attack performance will decline in the black-box setting, but the targeted attack success rate can also reach 100% in the white-box setting. In the process of obtaining the smoothed adversarial patch, some details are discarded. However, our goal is to make the patch smoother on the premise of ensuring the success of the white-box setting attack. Therefore, our smoothed universal adversarial patch may enjoy a better effect in some white-box scenarios with a defense mechanism.

5.2 Evaluation of EGP

Experimental Setup for Defense. Our defense method is evaluated for adversarial patches, enhanced adversarial patches, and smoothed adversarial patches. Patches are generated by ResNet-50, and all defense methods are carried out in white-box settings. For each type of patch, we randomly select 3000 adversarial examples that are successfully misclassified from our test data and then compare the accuracy under the defense of various methods. As for our defense method, we choose three Gauss kernels (5, 9, 19) to get the details with the same proportional coefficient. The magnification factor is set to be 5. We choose another three Gauss kernels (3, 5, 11) to weaken the enhanced images. The selected defense model is ResNet-50.

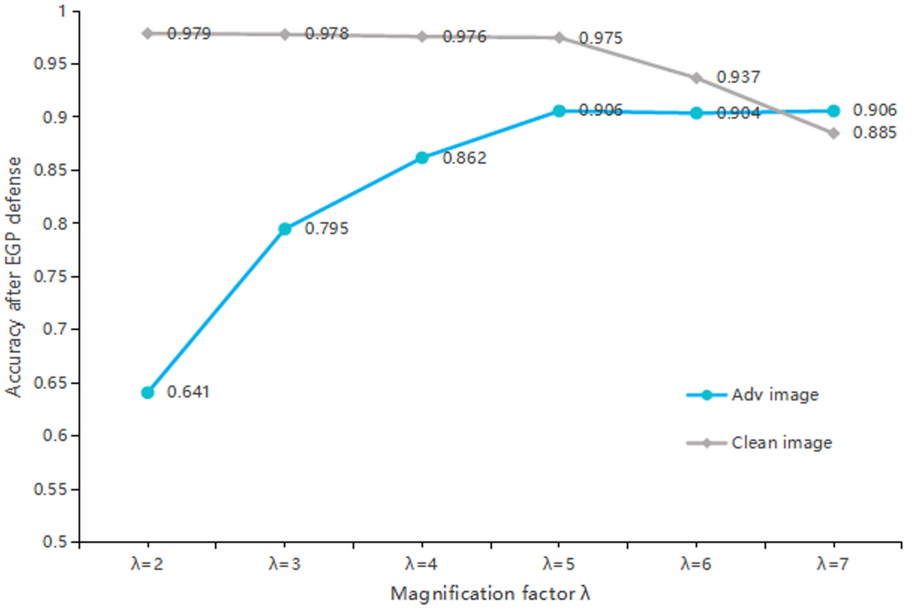

Fig. 6. The effect of amplification factor λ on the experimental results. Note that the accuracy refers to the classification precision after defense on clean images and adversarial images.

The intensity of multi-scale details obtained from the patch is greater than that of a clean image. Therefore, we multiply the multi-scale details by a magnification factor λ to enhance the details of a clean image and limit them to $[-1, 1]$. It should be noted that the values of multi-scale details of adversarial patch are multiple than that of a clean image. Therefore, an appropriate magnification factor λ can be choose to make enhanced details of the patch out of range $[-1, 1]$. As shown in Fig. 6, with the increase of λ, the adversarial images will be distorted due to the excessive enhancement of details, thus increasing

the accuracy of adversarial images. However, the classification accuracy of clean images also decreases due to the processing of our method, but when λ is small, the enhancement does not make the pixel value out of range $[-1, 1]$, and the enhanced images can still be restored through the weakening process. But when λ is too large and exceeds a certain threshold, the clean images will be irreversibly enhanced like the adversarial images which cause a significant decrease. Through the experiment, we found that the experimental result is better when λ is set to 5.

Defense Performance. It should be noted that Table 1 shows the comparison of the efficiency of various defense methods against Adv Patch, enhanced patch, and smoothed patch. Accuracy refers to the accuracy of adversarial examples under the defense. The accuracy of original clean images is 98%. The size of the patch is 70*70 covering 10% of the image (Fig. 7).

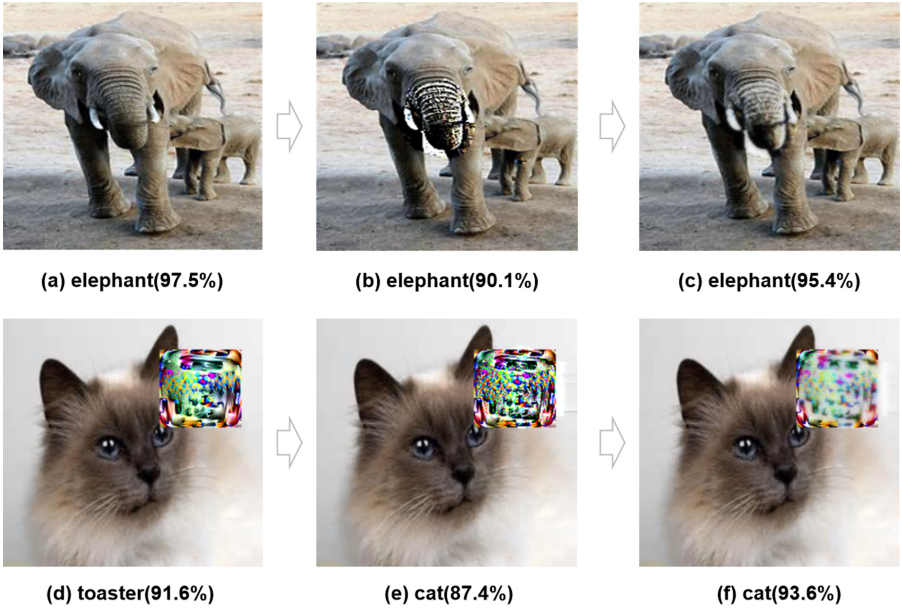

Fig. 7. ResNet-50 confidence scores are shown for example images. (a), (b) and (c) represent the processing of a clean image. (d), (e) and (f) represent the processing of an adversarial example. (b) and (e) represent the enhanced images after local multi-scale details enhancement. (c) and (f) are the output images. As illustrated, EGP restores correct class confidence and causes a negligible impact on clean images.

We directly use the JPEG method to globally compress the images to implement defense. PM constructs a saliency map of the image to detect localized and visible adversarial perturbations. Once a saliency map for the input has been

found, PM uses a combination of erosion and dilation to remove the adversarial perturbations. Lance locates the significant activation sources with CAM [14] and calculates the local input semantic inconsistency with the expected semantic patterns according to the prediction label. Once the inconsistency exceeds a predefined threshold which can be set between 0.1 and 0.18, Lance conducts a recovery process to recover the input image. About LGS, it first estimates the region of interest in an image with the highest probability of adversarial patch and then performs gradient smoothing in only those local regions. Specifically, LGS divides the image into several regions, and then performs gradient smoothing in the region where the image gradient exceeds the threshold value. LGS is used with $\gamma = 2.3$, γ is the smoothing factor for LGS. Note that the accuracy of the detection-based method (i.e., PM, LGS, Lance) is obtained by multiplying the success rate of detection and inpainting.

Table 1 shows the overall defensive performance. Our method EGP outperforms state-of-the-art defense methods for Adv Patch. As for the enhanced adver-

Table 1. Comparison of defense method.

	Defense method	Accuracy (%)
Adv Patch	None	0
	JPEG	45.0
	PM	76.4
	Lance	81.3
	LGS	89.5
	EGP	**90.6**
Enhanced Patch	None	0
	JPEG	45.0
	PM	73.1
	Lance	55.4
	LGS	87.6
	EGP	**90.4**
Smoothed Patch	None	0
	JPEG	42.8
	PM	70.5
	Lance	75.2
	LGS	79.8
	EGP	**89.2**

Table 2. Effect of defense method on clean images.

Method	None	EGP	JPEG	PM	LGS
Accuracy (%)	98.1	97.5	90.5	98.1	98.1

sarial patch, the defense efficiency of the method based on inconsistent features (Lance) [6] will be significantly reduced. Regarding the smoothed patch, our indiscriminate defense method is also excellent. Although other defense methods based on detection also enjoy considerable defensive effect on smoothed patches, it should be emphasized that the defense performance of them decreases more compared with dealing with Adv Patch. Furthermore, as shown in Table 2, the effect of our method on clean images can be ignored compared with other indiscriminate defense methods such as JPEG.

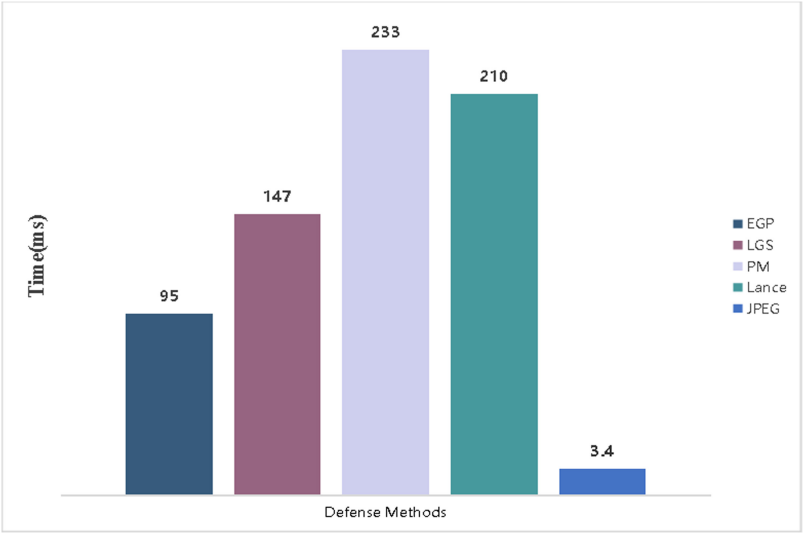

Fig. 8. Average processing time cost comparison of different defense methods. For better display, the average processing time of each image is shown at the top of the histogram.

Moreover, we compare the computational cost of EGP and existing defense methods. The compared methods include both preprocessing based and detection based. Note that our method processes image locally and skip the detection process of adversarial patches. Therefore, our defense strategy costs less computation. As shown in Fig. 8, our defense method only takes 95ms to process per image, which is better than most of the existing defense methods. Although our method is not as good as JPEG in computational cost, our defense efficiency is much better than JPEG.

6 Conclusions

In this paper, we analyze the effectiveness of existing defense methods against adversarial patches. Based on the analysis, we propose two improved patch generation methods to obtain the enhanced and smoothed patches that can effectively obscure the decision boundary for adversarial patches and reduce the

effectiveness of existing defense methods. To generate the enhanced patch, we strengthen the generation of the patch through the continuous update of the target model used for generating patch. Taking ImageNet as the data set, extensive experiments are conducted which demonstrate that our proposed enhanced patch enjoys stronger transferability and be robust to some defense mechanisms. Besides, to generate the smoothed patch, we add smooth processing intermittently during the updating of adversarial patch [1] to guide the update. Experimental results show that our smoothed patches enjoy better attack performance in some white-box scenarios with defense.

As for the countermeasure design, we propose a defense method based on image preprocessing. Leveraging the local multi-scale image processing [8], our method can efficiently interfere with adversarial patches and causes only neglect impact on clean images. Experiments show that our methodology outperforms state-of-the-art defense methods against adversarial patch attacks.

References

1. Brown, T.B., Mane, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. In: Neural Information Processing Systems (NIPS) (2017)
2. Karmon, D., Zoran, D., Goldberg, Y.: Lavan: localized and visible adversarial noise. In: International Conference on Machine Learning (ICML) (2018)
3. Liu, A., Wang, J., Liu, X., Cao, B., Zhang, C., Yu, H.: Bias-based universal adversarial patch attack for automatic check-out. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12358, pp. 395–410. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_24
4. Duan, R., Ma, X., Wang, Y., Bailey, J., Qin, A.K., Yang, Y.: Adversarial Camouflflage: hiding physical-world attacks with natural styles. In: CVPR (2020)
5. Naseer, M., Khan, S., Porikli, F.: Local gradients smoothing: defense against localized adversarial attacks. In: Proceedings Of WACV, pp. 1300–1307 (2019)
6. Xu, Z., Yu, F., Chen, X., LanCe: a comprehensive and lightweight CNN defense methodology against physical adversarial attacks on embedded multimedia applications. In: Asia and South Pacific Design Automation Conference (ASP-DAC) (2020)
7. Hayes, J.: On visible adversarial perturbations and digital watermarking. In: Proceedings of CVPR Workshops, pp. 1597–1604 (2018)
8. Kim, Y., Koh, Y.J., Lee, C., Kim, S., Kim, C.S.: Dark image enhancement based on pairwise target contrast and multi-scale detail boosting. In: IEEE International Conference on Image Processing, pp. 1404–1408 (2015)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS) (2012)
10. Mohamed, A.R., Dahl, G.E., Hinton, G.: Acoustic modeling using deep belief networks. In: IEEE T Audio Speech (2011)
11. Sutskever, I., Vinyals, O., Le, Q.: Sequence to sequence learning with neural networks. In: Neural Information Processing Systems (NIPS) (2014)
12. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv: 1607.02533 (2016)

13. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017)
14. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)
15. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: learning affordance for direct perception in autonomous driving. In: ICCV (2015)