Defending against Universal Adversarial Patches by Clipping Feature Norms

Cheng Yu[1,3], Jiansheng Chen[1,2,3]*, Youze Xue[1], Yuyang Liu[1], Weitao Wan[1], Jiayu Bao[1], and Huimin Ma[3]
[1]Department of Electronic Engineering, Tsinghua University, China
[2]Beijing National Research Center for Information Science and Technology, China
[3]University of Science and Technology Beijing, China

Abstract

Physical-world adversarial attacks based on universal adversarial patches have been proved to be able to mislead deep convolutional neural networks (CNNs), exposing the vulnerability of real-world visual classification systems based on CNNs. In this paper, we empirically reveal and mathematically explain that the universal adversarial patches usually lead to deep feature vectors with very large norms in popular CNNs. Inspired by this, we propose a simple yet effective defending approach using a new feature norm clipping (FNC) layer which is a differentiable module that can be flexibly inserted in different CNNs to adaptively suppress the generation of large norm deep feature vectors. FNC introduces no trainable parameter and only very low computational overhead. However, experiments on multiple datasets validate that it can effectively improve the robustness of different CNNs towards white-box universal patch attacks while maintaining a satisfactory recognition accuracy for clean samples.

1. Introduction

Deep convolutional neural networks (CNNs) have achieved remarkable success on various computer vision tasks. However, researches have shown that most CNNs are vulnerable to adversarial attacks [14, 19], where maliciously crafted perturbations are added to the input image to fool the network. The existence of adversarial examples under different constraints has become a serious concern to visual systems and applications based on CNNs especially in safety-critical domains. Compared to the conventional adversarial attacks which add perturbation with input norm constraints directly to the whole digital image, physical world attack is relatively more challenging. In the physical world, it is hard and expensive to accurately manipulate pixels that may scatter all around the image. A more promising way is to generate a spatially localized patch-like perturbation in which the pixel values can be arbitrarily selected. Usually, such a patch is generated so that it is effective no matter what the image is and where it is placed in the image to achieve real-world attacking robustness. Such an approach is called the universal adversarial patch attack [2], which has remained as the most effective and widely adopted way to attack real-world computer vision systems built upon CNNs including image classifier [2, 8], object detector [4, 24, 7] and face recognizer [17].

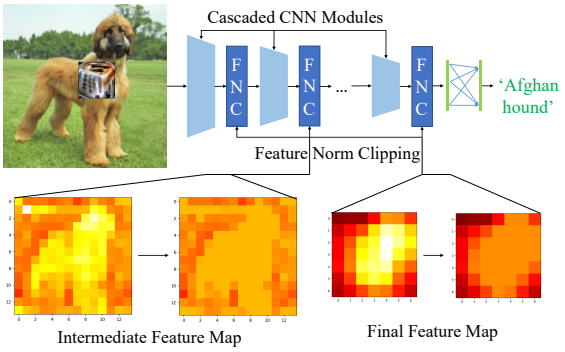

Figure 1. **Illustration of the proposed method based on feature norm clipping (FNC) layers.**

Despite its clear threat to real-world applications, research on defending against the universal adversarial patch attack is still limited. Some previous defending approaches, including Digital Watermarking (DW) [5] and Local Gradient Smoothing (LGS) [13], are based on patch detection following empirical clues. Lacking theoretical foundations, their performance usually drops dramatically when facing white-box or adaptive attacks in which the defending strategy is transparent to attackers [1]. By iteratively generating adversarial patches in the training process, the adversarial training can also be used in defending against patch attacks [22]. However, such an approach requires extra training with extremely high computational overhead and there has been no work to show its feasibility for large-scale datasets like ImageNet [10] by far. Chiang *et al*. [3] proposed a certified defense against the adversarial patch when the output lies in the interval bound. However, this estimated bound gets looser with the increase of the network

*Corresponding author.

depth, leading to the difficulty in its scaling up to commonly used deep networks like ResNet [6] and Inception [18].

In this paper, we propose an effective defending method against universal adversarial patch attacks based on a mathematical analysis of how a universal adversarial patch impact deep feature representations. Generally, a universal adversarial patch only occupies a small image area. However, it can mislead the CNN to predict the same wrong target class no matter where it is placed and what the original image is. How is this achieved? Empirical observations reveal that when an adversarial patched image is passing forward through a CNN, usually the norm of the feature vector spatially located at the position of the adversarial patch is significantly larger than that of other feature vectors. This makes sense considering that under the widely adopted global average pooling strategy, a large norm feature vector can dominate in deciding the direction of the pooling result no matter what other feature vectors are. Such a phenomenon exists only when the adversarial perturbation is localized like adversarial patches. This can be explained by the fact that the effective receptive field (ERF) of a practical CNN is exponentially centered and is obviously smaller than the theoretical receptive field [11]. As such, the impact of the adversarial patch on the feature map has to be spatially concentrated. This indicates that most feature vectors will not be essentially affected, leading to inevitable existence of large norm feature vector at the patch position.

We present a mathematical explanation of the phenomenon, based on which a defending method is proposed. Specifically, we propose to restrict the norm of deep features on different layers through the forward propagation by introducing Feature Norm Clipping (FNC) layers as illustrated in Fig. 1. FNC is a differentiable layer and can be inserted among the cascaded modules in CNNs, *e.g*. the residual block in ResNet. FNC reduces the variance of feature norms by one-side clipping, so as to prevent the generation of feature vectors with extremely large norms. Analysis on ERF demonstrates that the influence of the adversarial patch can be intrinsically weakened by FNC, leading to improved classification accuracy against adversarial patch attacks. We conduct extensive experiments on CIFAR10 [9] and ImageNet [10] with different CNN architectures and attacking methods. Significant improvements on robustness against white-box adversarial patch attacks are achieved. Moreover, in contrast to previous defending methods [5, 13], our proposed approach is implemented in an end-to-end manner with very low computational overhead in both training and inference, leading to its high applicability to real-word visual classification systems based on CNNs.

There are three main contributions in this paper. 1) We analyze the impact of the universal adversarial patch on the deep feature representation, and mathematically explain the existence of large norm feature vectors at the patch location. 2) We propose a simple yet effective defending method using the FNC layer to weaken the effect of the adversarial patch by restricting the generation of large norm feature vectors. 3) We achieve improvements in adversarial accuracy for different networks and datasets. Our method outperforms state-of-the-art patch defending methods.

2. Related Works

Researches on the adversarial patch have raised increasing interest since it was first proposed by Brown *et al*. [2] due to its effectiveness in attacking real-world systems. Karmon *et al*. proposed to generate adversarial patches aiming at inducing targeted misclassification [8]. Later, adversarial patch attacks have been also studied in other vision tasks such as object detection [4, 24, 7], semantic segmentation [16] and person re-identification [21].

In this paper, we focus on defending against universal adversarial patch attacks in visual classification, for which several methods have been proposed. Hayes proposed Digital Watermarking (DW) in which the adversarial patch is localized and masked out based on the saliency map [5]. However, the performance on clean images is severely compromised since the salient parts of the objects are usually falsely masked out. Naseer *et al*. proposed Local Gradient Smoothing (LGS) in which a soft mask is applied to the input image during pre-processing based on the assumption that pixels within the patch areas tend to change dramatically in space [13]. Nevertheless, the performance of LGS drops severely against the white-box attack which takes the pre-processing step into consideration. Wu *et al*. proposed Defense over Occluded Attack (DOA) by combining a new substitute attack approach that represents physically realizable attacks with adversarial training [12] to increase robustness [22]. This method, however, requires extra training with high computational overhead and is difficult to be extended to large-scale datasets such as the ImageNet [10]. Chiang *et al*. proposed a certified defense against the adversarial patch attack [3]. Despite its theoretical contribution, such an optimization-based method is computationally inefficient and difficult to be scaled up to practical deep neural networks. Xiang *et al*. proposed PatchGuard to realize efficient robustness verification for models with small receptive fields [23]. However, a specially designed neural network is required to perform the feature extraction, which means that widely adopted CNN models cannot be directly used for robust classification.

3. Method

In this section, we first provide a mathematical analysis on how the adversarial patch impacts the feature representations of a CNN equipped with global average pooling. Then we elaborate on the proposed defending method.

3.1. Preliminaries

For an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, denote its predicted logit for class y by a CNN as $\mathcal{F}(y|\mathbf{x})$. A universal adversarial patch p is expected to mislead the network to predict a wrong target class in a variety of contexts, including different images and locations placed [2]. Formally, the patched adversarial example $\hat{\mathbf{x}}$ for clean image $\tilde{\mathbf{x}}$ can be formulated as Eq. 1, where $\odot$ is Hadamard product, $\mathbf{M} \in \{0,1\}^{H\times W\times C}$ is the binary mask indicating the position of the patch, and $A(p,t)$ is an operator which applies the geometric transformation t to the patch p so that it is located at the position indicated by $\mathbf{M}$. The adversarial patch p is generated by solving the optimization problem in Eq. 2, where $\hat{y}$ is the target class of the adversarial attack, and $\Pr(y|\mathbf{x})$ is the classification probability calculated by applying softmax operation to the logits.

$$\hat{\mathbf{x}} = (\mathbf{1} - \mathbf{M}) \odot \tilde{\mathbf{x}} + \mathbf{M} \odot A(p,t), \tag{1}$$

$$\max_{p} \ \mathbb{E}_{\tilde{\mathbf{x}}\sim X, t\sim T} \log \Pr(\hat{y}|\hat{\mathbf{x}}), \tag{2}$$

Variants have been proposed to improve the attacking effectiveness. For example in the LaVAN attack [8], $\log \Pr(\hat{y}|\hat{\mathbf{x}})$ in Eq. 2 is replaced by $\mathcal{F}(\hat{y}\,|\,\hat{\mathbf{x}}) - \mathcal{F}(\tilde{y}\,|\,\hat{\mathbf{x}})$, where $\tilde{y}$ is the classifier's predicted class on the benign input $\tilde{\mathbf{x}}$. We will show that our proposal keeps effective as long as the patch is formulated following Eq. 1.

3.2. Impact of Adversarial Patch on Feature Maps

With the help of the effective receptive fields (ERF) theory [11], we present a mathematical explanation for the existence of large norm feature vectors in feature maps under adversarial patch attacks. We start with analysis on the final feature map (FFM), which is the output of the last convolution layer. We then extend the conclusion to the intermediate feature maps of shallow layers.

3.2.1 Properties of Effective Receptive Fields

Luo *et al*. [11] defined the effective receptive field (ERF) to measure how much each input pixel in a receptive field can impact the output. Assume the pixels on each layer are indexed by (i,j), with their center at $(0,0)$. To simplify notation, consider a CNN with N convolutional layers of one single channel for each layer. Denote the (i,j)th pixel on the input to the network and the output of the Nth layer as $x_{i,j}$ and $y_{i,j}$ respectively. ERF measures how much each $x_{i,j}$ contributes to $y_{0,0}$ by calculating the expectations of the partial derivative $\partial y_{0,0}/\partial x_{i,j}$ over input distribution. Luo *et al*. demonstrated that ERF converges to the probability density function of a 2D Gaussian distribution Φ_N subject to Eq. 3, in which Ω_n is a random variable following the distribution of the weights of the nth convolutional layer.

$$\Phi_N \sim \mathcal{N}(0, \sum_{n=1}^{N} \mathrm{Var}[\Omega_n]) \tag{3}$$

Although Eq. 3 requires quite strong assumptions, it agrees well with experiments. It can be concluded that ERF only occupies a small area (with size proportional to $\sqrt{N}$) of the theoretical receptive field (with size proportional to N) and is shaped like a Gaussian function decaying from the center.

3.2.2 Impact on the Final Feature Map

Denote the FFM of input $\mathbf{x}$ by $\mathbf{f}(\mathbf{x})$, where $\mathbf{f} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{h\times w\times c}$. The feature vector at location (i,j) of the FFM and its kth channel are denoted by $\mathbf{f}_{i,j}(\mathbf{x})$ and $f_{i,j,k}(\mathbf{x})$ respectively. To simplify notation, we further denote $\tilde{\mathbf{f}}_{i,j} = \mathbf{f}_{i,j}(\tilde{\mathbf{x}})$, $\hat{\mathbf{f}}_{i,j} = \mathbf{f}_{i,j}(\hat{\mathbf{x}})$ and $\mathbf{f}^*_{i,j} = \mathbf{f}_{i,j}(\hat{\mathbf{x}}) - \mathbf{f}_{i,j}(\tilde{\mathbf{x}})$.

It has been discussed in previous works [25] that intuitively universal perturbations contain dominant features and images behave like noise to them. This means that the difference in FFM caused by the universal adversarial patch is nearly irrelevant to the input image, which can be validated by experiments. Inspired by this, we adopt the theory of ERF to estimate the spatial properties of $\|\mathbf{f}^*_{i,j}\|$ to measure the impact of the adversarial patch on different locations in FFM. Denote the gradient map as $\nabla_{\mathbf{x}} f_{i,j,k}$, we can get Eq. 4, where $\cdot$ denotes inner product, L is an arbitrary curve from $\tilde{\mathbf{x}}$ to $\hat{\mathbf{x}}$, and $\mathcal{U}(0,1)$ is the uniform distribution from 0 to 1. Specifically, Eq. 4a holds because of the property of the conservative field that its line integral along a path depends only on the potential function value of the initial and final points [20]. We then choose a special path $L = \{\tilde{\mathbf{x}} + \gamma(\hat{\mathbf{x}} - \tilde{\mathbf{x}})\}, \gamma : 0 \to 1$ to get Eq. 4b. Consequently, Eq. 4c comes from the definition of expectation.

$$f_{i,j,k}(\hat{\mathbf{x}}) - f_{i,j,k}(\tilde{\mathbf{x}}) = \int_L \nabla_{\mathbf{x}} f_{i,j,k} \cdot d\mathbf{x} \tag{4a}$$

$$= (\int_0^1 \nabla_{\tilde{\mathbf{x}}+\gamma(\hat{\mathbf{x}}-\tilde{\mathbf{x}})} f_{i,j,k} d\gamma) \cdot (\hat{\mathbf{x}} - \tilde{\mathbf{x}}) \tag{4b}$$

$$= \mathbb{E}_{\gamma\sim\mathcal{U}(0,1)}[\nabla_{\tilde{\mathbf{x}}+\gamma(\hat{\mathbf{x}}-\tilde{\mathbf{x}})} f_{i,j,k}] \cdot (\hat{\mathbf{x}} - \tilde{\mathbf{x}}) \tag{4c}$$

Note that ERF is the expectation of the gradient map with output $(0,0)$. Therefore, the expectation term in Eq. 4c approximately equals to $tr_{i,j}(\mathrm{ERF})$, in which $tr_{i,j}$ is the translation to (i,j). Then, $\|\mathbf{f}^*_{i,j}\|$ can be estimated as Eq. 5.

$$\begin{aligned} \|\mathbf{f}^*_{i,j}\| &\approx (c(tr_{i,j}(\mathrm{ERF}) \cdot (\hat{\mathbf{x}} - \tilde{\mathbf{x}}))^2)^{\frac{1}{2}} \\ &\propto \sum tr_{i,j}(\mathrm{ERF}) \odot (\hat{\mathbf{x}} - \tilde{\mathbf{x}}). \end{aligned} \tag{5}$$

Note that $(\hat{\mathbf{x}} - \tilde{\mathbf{x}})$ is zero in the region except the patch. From the properties of ERF, we can find that

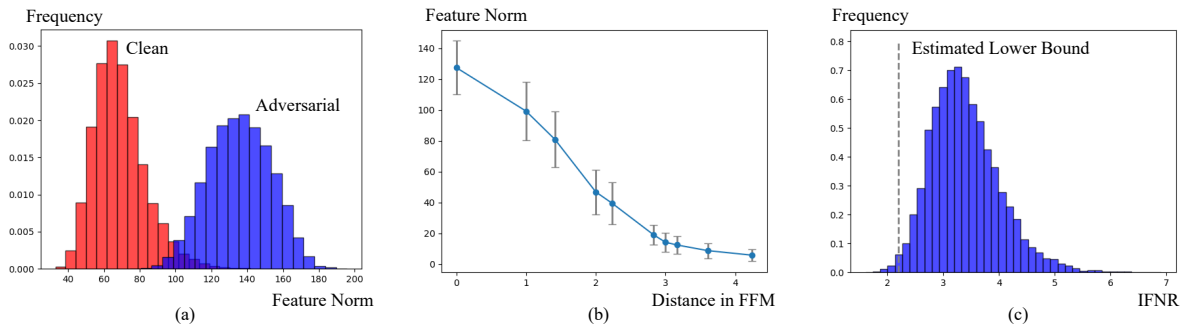

Figure 2. **(a) Histogram of the maximum norm of feature vectors in the FFM for both clean and adversarial examples. (b) Curve of the mean value and standard deviation of $\|\mathbf{f}^*_{i,j}\|$ with the relative location of the adversarial patch. (c) Histogram of $\mathcal{R}$ in FFM as well as the estimated lower bound.** The experiments are conducted on ResNet-50 under LaVAN attack on ImageNet. The adversarial patch covering 5% of the image is placed on random locations over the test images.

$\sum tr_{i,j}(\text{ERF}) \odot (\hat{\mathbf{x}} - \tilde{\mathbf{x}})$ is the weighed sum of values on finite locations for a Gaussian function centered at (i, j). It can be proved that $\|\mathbf{f}^*_{i,j}\|$ is spatially distributed as a weighed sum of Gaussian functions of the same variance but different centers. It therefore has the same squared exponential decay rate as the Gaussian functions, which is further elaborated in the Supplementary. Since the patch is far smaller than the image, $\|\mathbf{f}^*_{i,j}\|$ also has such a squared exponentially centered property in the image.

Then we consider the classifier layer in which the final feature before the fully connected layer is calculated by the global average pooling of FFM. Suppose the weight and bias of the fully connected layer corresponding to class y to be $\mathbf{w}_y \in \mathbb{R}^c$ and b_y. The predicted logit for class y on the adversarial example $\hat{\mathbf{x}}$ is computed by Eq. 6.

$$\mathcal{F}(y|\hat{\mathbf{x}}) = \frac{1}{hw}\sum_{i,j}(\tilde{\mathbf{f}}^\top_{i,j}\mathbf{w}_y + {\mathbf{f}^*_{i,j}}^\top\mathbf{w}_y) + b_y \tag{6}$$

For an ideal universal adversarial patch, $\mathcal{F}(\hat{y}|\hat{\mathbf{x}}) > \mathcal{F}(\tilde{y}|\hat{\mathbf{x}})$ holds for all possible $\tilde{\mathbf{x}}$ belonging to class $\tilde{y}$. We further assume that $\|\mathbf{w}_{\hat{y}}\| \le \|\mathbf{w}_{\tilde{y}}\|$ considering that the target class can actually be freely selected. Since the clean image is classified to class $\tilde{y}$, we have $\tilde{\mathbf{f}}^\top_{i,j}\mathbf{w}_{\hat{y}}/\tilde{\mathbf{f}}^\top_{i,j}\mathbf{w}_{\tilde{y}} \approx 0$. Besides, we ignore b_y since it is generally far smaller than $\mathcal{F}(y|\mathbf{x})$. As such, the condition for a successful patch attack can be approximated by Eq. 7 where $\cos(\mathbf{a}, \mathbf{b})$ denotes the cosine similarity between two vectors $\mathbf{a}$ and $\mathbf{b}$.

$$\begin{aligned}
&0 < \mathcal{F}(\hat{y}|\hat{\mathbf{x}}) - \mathcal{F}(\tilde{y}|\hat{\mathbf{x}}) \\
&= \frac{1}{hw}\sum_{i,j}(\tilde{\mathbf{f}}^\top_{i,j}\mathbf{w}_{\hat{y}} + {\mathbf{f}^*_{i,j}}^\top\mathbf{w}_{\hat{y}} - \tilde{\mathbf{f}}^\top_{i,j}\mathbf{w}_{\tilde{y}} - {\mathbf{f}^*_{i,j}}^\top\mathbf{w}_{\tilde{y}}) + b_{\hat{y}} - b_{\tilde{y}} \\
&\approx \frac{1}{hw}\sum_{i,j}({\mathbf{f}^*_{i,j}}^\top\mathbf{w}_{\hat{y}} - \tilde{\mathbf{f}}^\top_{i,j}\mathbf{w}_{\tilde{y}} - {\mathbf{f}^*_{i,j}}^\top\mathbf{w}_{\tilde{y}}) \\
&\le \frac{1}{hw}\sum_{i,j}\|\mathbf{w}_{\tilde{y}}\|(\|\mathbf{f}^*_{i,j}\|(\cos(\mathbf{f}^*_{i,j}, \mathbf{w}_{\hat{y}}) - \cos(\mathbf{f}^*_{i,j}, \mathbf{w}_{\tilde{y}})) \\
&\quad - \|\tilde{\mathbf{f}}_{i,j}\|\cos(\tilde{\mathbf{f}}_{i,j}, \mathbf{w}_{\tilde{y}})),
\end{aligned} \tag{7}$$

Finally we manage to provide an estimated lower bound of $\max_{i,j}\|\mathbf{f}^*_{i,j}\|/\text{mean}_{i,j}\|\tilde{\mathbf{f}}_{i,j}\|$ according to Eq. 7 and Eq. 5. Intuitively, $\|\mathbf{f}^*_{i,j}\|$ is significantly more centered and decays faster compared to $\|\tilde{\mathbf{f}}_{i,j}\|$, and the cosine similarities are limited. So the ratio is relatively large, leading to the large norm of $\mathbf{f}(\hat{\mathbf{x}})$ on the same location of the patch. Formally, we assume that the chosen $\mathbf{w}_{\hat{y}}$ is orthogonal to $\mathbf{w}_{\tilde{y}}$, which holds approximately in commonly adopted CNNs. Mathematically, the bound depends on the cosine similarities $\cos(\tilde{\mathbf{f}}_{i,j}, \mathbf{w}_{\tilde{y}})$ and $(\cos(\mathbf{f}^*_{i,j}, \mathbf{w}_{\hat{y}}) - \cos(\mathbf{f}^*_{i,j}, \mathbf{w}_{\tilde{y}}))$, which represent the saturation state of model training and patch generation processes respectively and is decided by the CNN. In this paper, we assume that $\cos(\tilde{\mathbf{f}}_{i,j}, \mathbf{w}_{\tilde{y}})$ and $(\cos(\mathbf{f}^*_{i,j}, \mathbf{w}_{\hat{y}}) - \cos(\mathbf{f}^*_{i,j}, \mathbf{w}_{\tilde{y}}))$ can reach their theoretical upper limits (1 and $\sqrt{2}$ respectively) after sufficient training. The limits generally cannot be reached in both processes, but the estimated bound based on the ratio of them agrees well with experiments, which will be demonstrated later. As such, Eq. 7 approximately turns to Eq. 8 according to Eq. 5 and the above assumptions.

$$\begin{aligned}
&0 < \frac{1}{hw}\sum_{i,j}\|\mathbf{w}_{\tilde{y}}\|(\sqrt{2}\|\mathbf{f}^*_{i,j}\| - \|\tilde{\mathbf{f}}_{i,j}\|), \\
&\|\mathbf{f}^*_{i,j}\| = \max_{i,j}\|\mathbf{f}^*_{i,j}\|\frac{\sum tr_{i,j}(\text{ERF}) \odot (\hat{\mathbf{x}} - \tilde{\mathbf{x}})}{\max_{i,j}\sum tr_{i,j}(\text{ERF}) \odot (\hat{\mathbf{x}} - \tilde{\mathbf{x}})}
\end{aligned} \tag{8}$$

Then we can get an estimation of the lower bound as Eq. 9.

$$\frac{\max_{i,j}\|\mathbf{f}^*_{i,j}\|}{\text{mean}_{i,j}\|\tilde{\mathbf{f}}_{i,j}\|} > \frac{hw}{\sqrt{2}}\frac{\max_{i,j}\sum tr_{i,j}(\text{ERF}) \odot (\hat{\mathbf{x}} - \tilde{\mathbf{x}})}{\sum_{i,j}\sum tr_{i,j}(\text{ERF}) \odot (\hat{\mathbf{x}} - \tilde{\mathbf{x}})} \tag{9}$$

We refer to the ratio $\mathcal{R} = \max_{i,j}\|\mathbf{f}^*_{i,j}\|/\text{mean}_{i,j}\|\tilde{\mathbf{f}}_{i,j}\|$ as the Incremental Feature Norm Ratio (IFNR), where the maximum $\|\mathbf{f}^*_{i,j}\|$ is generally on the center of the patch. The lower bound of $\mathcal{R}$ shown in Eq. 9 is verified by the experimental results on ResNet-50 illustrated in Fig. 2, in which

Fig. 2 (a) shows the clear margin of the norm in FFM between clean and adversarial examples, and Fig. 2 (b) indicates that the impact of the adversarial patch in FFM is distributed with a Gaussian-like decay from the center of the patch. We also calculate the estimated lower bound of $\mathcal{R}$ in FFM over the input images according to Eq. 9 with the measured ERF radius 71. The result is 2.2 as shown by the dashed line in Fig. 2 (c), in which the histogram of the empirical ratio among all the tested images is also shown. The calculated bound holds for 99.2% of the tested images, validating the mathematical analysis above.

3.2.3 Impact on the Intermediate Feature Maps

We have derived that the norm of feature vectors is obviously large at the patch location and has an approximate squared exponential decay in FFM under the universal patch attack. Then we extend the conclusion to the intermediate feature maps of shallow layers. We start with the adjacent feature maps, *i.e.* the input and the output feature map of a convolution layer. To simplify notation, we consider one spatial dimension and one channel for the convolution layer without loss of generality, because any 2D convolution with the kernel of rank r can be regarded as a linear combination of r multiplications of two 1D convolutions. Denote $f_{in}[i]$ and $f_{out}[i]$ as the input and output features respectively, where i is the 1D spatial location. The unit impulse response of the convolution layer $o[i]$ is formulated by $o[i] = \sum_m w[m]\delta[i-m]$, where $w[m]$ is the weight of the convolution layer on location m.

We focus on the unit impulse response $o^{-1}[i]$ of the inverse system of the layer. With the help of the z-transformation, $o^{-1}[i]$ can be formulated as Eq. 10, where $O^{-1}(z)$ denotes the z-transformation of $o^{-1}[i]$, and $\{v_s\}_{s=1}^{S}$ are the poles of $O^{-1}(z)$. It can be observed that each $o_s[i]$ has a unilateral exponential decay, and $o^{-1}[i]$ is their linear combination. Furthermore, it is validated by experiments that when the kernel of the convolution layer is randomly chosen, the mean decay radius is approximately equal to the size of the convolution kernel.

$$
\begin{aligned}
o^{-1}[i] &= \sum_{s=1}^{S} q_s o_s[i], \\
o_s[i] &= \begin{cases} {v_s}^i u[i] & |v_s| \le 1 \\ -{v_s}^i u[-i-1] & |v_s| > 1 \end{cases} \\
\text{s.t. } O^{-1}(z) &= \frac{1}{\sum_m w[m] z^{-m}} = \sum_{s=1}^{S} q_s \frac{1}{1 - v_s z^{-1}}
\end{aligned} \tag{10}
$$

Table 1. Average $\mathcal{R}$ of intermediate feature maps in ResNet-50.

Layer	Conv2-3	Conv3-4	Conv4-6	Conv5-3
Ratio	1.05	1.80	2.98	3.39

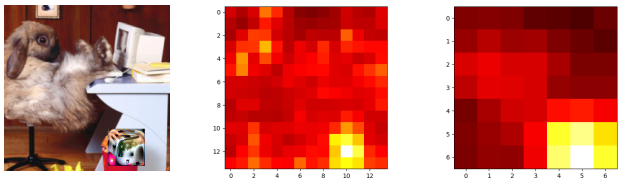

Figure 3. **An example of the feature norm maps on Conv4-6 and Conv5-3 (FFM) of ResNet-50 as well as the input image.**

As such, if $f_{out}[i]$ has large norm feature vectors centered at the location of the patch, $f_{in}[i] = o^{-1}[i] * f_{out}[i]$ ($*$ denotes convolution) may also have large norm feature vectors centered. We generate the adversarial patch on ResNet-50 under LaVAN attack on ImageNet and calculate the average $\mathcal{R}$ for intermediate feature maps on the last residual blocks of each Convλ-ν architectures as is shown in Table 1. Since the feature maps on shallow layers have relatively high resolution and are not smooth enough, we replace $\max_{i,j}\|\mathbf{f}^*_{i,j}\|$ with the mean $\|\mathbf{f}^*_{i,j}\|$ on the center of the patch. It can be observed that $\mathcal{R}$ keeps being greater than 1 and increases with the depth of the layer. Fig. 3 shows an example of the feature norm maps on different layers as well as the adversarial input image, revealing that the impact on the intermediate feature maps keeps centered on the patch and accumulates with the depth of the layer. Detailed deduction and experiments are in the Supplementary.

3.3. Feature Norm Clipping

Based on the above analysis, we propose the Feature Norm Clipping (FNC) layer to restrict the norm of the features on different layers. Define a cascaded CNN architecture with modules $\{D^{(n)}\}_{n=1}^{N}$ and the feature maps $\{\mathbf{f}^{(n)} \in \mathbb{R}^{h^{(n)} \times w^{(n)} \times c^{(n)}}\}_{n=1}^{N}$, where N is the number of cascaded modules. Formally, the forward propagation of the CNN can be formulated as Eq. 11.

$$
\begin{aligned}
&\mathbf{f}^{(1)} = D^{(1)}(\mathbf{x}), \quad \mathbf{f}(\mathbf{x}) = \mathbf{f}^{(N)}, \\
&\mathbf{f}^{(n+1)} = D^{(n+1)}(\mathbf{f}^{(n)}) \text{ for } n = 1, 2, ..., N-1
\end{aligned} \tag{11}
$$

FNC can be regarded as an operator on feature maps that can be inserted among the CNN modules. The output $\mathbf{g}^{(n)}$ of applying FNC on $\mathbf{f}^{(n)}$ is calculated by Eq. 12, where α is the clipping parameter. Then $\mathbf{g}^{(n)}$ is fed to the next CNN module to compute $\mathbf{f}^{(n+1)}$ as in Eq. 11 and Fig. 1.

$$
\begin{aligned}
\mathbf{g}^{(n)}_{i,j} &= \frac{\mathbf{f}^{(n)}_{i,j}}{\|\mathbf{f}^{(n)}_{i,j}\|} \min(\|\mathbf{f}^{(n)}_{i,j}\|, \alpha \overline{\|\mathbf{f}^{(n)}\|}), \\
\overline{\|\mathbf{f}^{(n)}\|} &= \frac{1}{h^{(n)} w^{(n)}} \sum_{i,j} \|\mathbf{f}^{(n)}_{i,j}\|,
\end{aligned} \tag{12}
$$

Intuitively, FNC prevents the generation of feature vectors with extremely large norm by one-sided clipping. Thus the influence of the adversarial patch is weakened. Furthermore, we can measure the effect of FNC with the theory

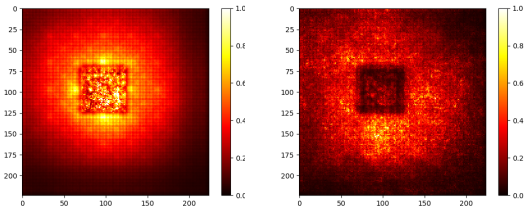

Figure 4. **Normalized ERF of adversarial examples for ResNet-50 with (right) and without (left) FNC.** The adversarial patch is located in the center of the test images for simplification.

of ERF. The Jacobi matrix of input $\mathbf{f}^{(n)}$ and output $\mathbf{g}^{(n)}$ of FNC can be computed as Eq. 13, where $\mathbf{f}_{i,j}^{(n)}$ is the clipped vector. Thus $\alpha\overline{\|\mathbf{f}^{(n)}\|}/\|\mathbf{f}_{i,j}^{(n)}\|$ is less than 1. $\mathbf{I}-\mathbf{P}_{i,j}^{(n)}$ is a projection matrix to the orthogonal direction of $\mathbf{f}_{i,j}^{(n)}$, so that the norm of the gradient will not increase afterwards.

$$\frac{\partial \mathbf{g}_{i,j}^{(n)}}{\partial \mathbf{f}_{i,j}^{(n)}} = \frac{\alpha\overline{\|\mathbf{f}^{(n)}\|}}{\|\mathbf{f}_{i,j}^{(n)}\|}(\mathbf{I}-\mathbf{P}_{i,j}^{(n)}) + \varepsilon_{i,j,i,j}^{(n)},\ \frac{\partial \mathbf{g}_{i,j}^{(n)}}{\partial \mathbf{f}_{i',j'}^{(n)}} = \varepsilon_{i,j,i',j'}^{(n)},$$
$$\mathbf{P}_{i,j}^{(n)} = \frac{\mathbf{f}_{i,j}^{(n)}\mathbf{f}_{i,j}^{(n)^T}}{\|\mathbf{f}_{i,j}^{(n)}\|^2},\ \varepsilon_{i,j,i',j'}^{(n)} = \frac{\alpha}{h^{(n)}w^{(n)}}\frac{\mathbf{f}_{i,j}^{(n)}\mathbf{f}_{i',j'}^{(n)^T}}{\|\mathbf{f}_{i,j}^{(n)}\|\|\mathbf{f}_{i',j'}^{(n)}\|} \approx \mathbf{0}, \tag{13}$$

As a result, the gradient norm of the clipped vector will be decreased with a scale of $\alpha\overline{\|\mathbf{f}^{(n)}\|}/\|\mathbf{f}_{i,j}^{(n)}\|$ at most. It means that as long as the large norm feature vectors on the corresponding location of the patch are clipped by FNC, the gradient norm will be decreased, causing the ERF to be significantly darker on the location of the patch than the benign areas as is shown in Fig. 4. FNC is suitable for most popular CNNs with GAP and cascaded architecture such as ResNet, Inception and MobileNet [15] and has little effects on clean data. After adding FNC to standard trained models, we only need a little extra time going through the training data to adjust the statistics of the BatchNorm layer.

4. Experiments

4.1. Experimental Setup

We use the ResNet-50 [6], Inception-V3 [18], and MobileNet-V2 [15] models pretrained on ImageNet as well as the ResNet-110 [6] pretrained on CIFAR10 as our target model and test the classification accuracy of clean (clean Acc) and patch-attacked adversarial examples (adversarial Acc) before and after applying each defending method. The original Adversarial Patch (AdvP) [2] as well as the LaVAN [8] method is used to generate adversarial examples. The target class is *toaster* for ImageNet and *dog* for CIFAR10. Other target classes lead to similar results as is reported in ablation studies. The FNC layers are operated on feature maps of each cascaded module (*e.g.* residual block for ResNet, inception module for Inception, and inverted residual block for MobileNet) without extra explanations. The defense performances of Digital Watermarking (DW) [5], Local Gradient Smoothing (LGS) [13], PatchGuard [23] and Defense against Occlusion Attacks (DOA) [22] are evaluated as comparisons.

In order to provide a fair platform for comparison, all the experimental results are reported under the pure **white-box** adversarial attack, which takes the defending approaches into consideration when training the adversarial patch. Specifically, for our method, PatchGuard, and DOA which are based on changing the forward propagating function of the CNN model, the adversarial patch is generated to attack the modified model. For LGS which conducts a differential pre-processing step to the input image, we add the step into both forward and backward passes when training the adversarial patch. For other methods like DW with non-differential operators in pre-processing, Backward Pass Differential Approximation (BPDA) [1] is adopted to approximate gradients by forward propagating through the transforming operator as well as ignoring the operator in backward propagation. The reason why we adopt the pure white-box attack instead of the gray-box attack is that it can better reflect the robustness of the defending methods against adversarial patch attacks. Since all the defending methods are convenient to reproduce, the attackers can also design the white-box or adaptive attack according to them easily.

4.2. Experimental Results

The results on ImageNet are presented in Table 2, in which the highest accuracies are shown in bold. Since the adversarial training based method DOA needs high computation overhead and currently cannot be adopted on ImageNet scale, we also report the results on CIFAR10 in Table 3 with the same model architecture ResNet-110 as DOA uses. It can be observed that the proposed method significantly and consistently outperforms previous defending methods on different adversarial examples across different CNNs. Our method keeps the highest adversarial Acc in all the experiments. Besides, compared to other defending methods, the clean Acc of our method is also satisfactory.

Comparison with PatchGuard, DW, and LGS. We start with the comparison with non-adversarial training methods. Despite having a similar clean and adversarial Acc with our method in CIFAR10, PatchGuard cannot work well on large-scale ImageNet. PatchGuard is originally designed for specially designed CNNs with very small receptive field and the size of the window is the largest region on FFM which can be theoretically affected by the patch. Such CNNs generally perform poorly on large-scale datasets. Unlike PatchGuard, FNC layers are operated on multiple feature maps in our method, making the impact of the patch better suppressed. The problem of DW is that it suffers from dramatic clean Acc drops on both datasets.

Table 2. Clean and adversarial Acc for defense methods evaluated on different target models against AdvP and LaVAN on ImageNet. The best results are marked as **bold**.

	ResNet-50			Inception-V3			MobileNet-V2		
Defense	Clean	AdvP	LaVAN	Clean	AdvP	LaVAN	Clean	AdvP	LaVAN
No Defense	76.1%	0.3%	0.3%	77.9%	0.4%	0.3%	71.6%	4.7%	4.8%
PatchGuard Window=4	67.0%	33.8%	31.6%	74.8%	28.8%	23.3%	63.0%	29.7%	27.4%
DW	42.4%	32.3%	32.7%	35.6%	32.4%	30.2%	38.0%	23.4%	23.1%
LGS	69.8%	1.0%	0.6%	**75.0%**	1.2%	1.7%	65.3%	10.5%	11.5%
Ours $\alpha = 1.0$	72.4%	58.6%	58.3%	71.6%	58.7%	58.8%	64.1%	**51.9%**	**52.0%**
Ours $\alpha = 1.1$	**73.3%**	**59.6%**	**59.5%**	74.3%	**59.6%**	**59.0%**	**65.5%**	49.5%	48.9%

Table 3. Clean and adversarial Acc for defense methods evaluated on ResNet110 against AdvP and LaVAN on CIFAR10. The best results are marked as **bold**.

Defense	clean	AdvP	LaVAN
No Defense	94.1%	28.1%	27.2%
PatchGuard Window=4	92.5%	54.6%	54.9%
DW	49.1%	39.2%	33.7%
LGS	91.7%	33.7%	35.0%
DOA	**93.4%**	46.9%	44.8%
Ours $\alpha = 1.2$	92.7%	54.7%	54.9%
Ours+DOA $\alpha = 1.2$	91.9%	**68.0%**	**67.3%**

DW chooses the most salient areas in the image for a network to erase, causing the most important areas in the benign inputs to be erased. Nevertheless, our method can keep a high adversarial Acc without sacrificing the clean Acc by suppressing the impacts of the adversarial patch more reasonably. It can be observed that LGS has almost no effect on defending against the white-box attacks compared to no defense added. LGS assumes that the spatial variation of the pixel value in the patch is more severe than in other regions, which is not necessary for the success of the adversarial patch attack. As a comparison, our method is based on large norms in feature maps causing by the adversarial patch, which can be explained mathematically and regarded as an intrinsic property of the adversarial patch.

Comparison with DOA. Despite adopting simulated attack in the training process to reduce computation, DOA still needs a lot of time in training the robust model and is hard to be implemented on ImageNet scale. Besides, the difference between the simulated attack and the actual attack leads to the drop of adversarial Acc. In comparison, our method is more effective and efficient than DOA with the increase of FLOPs for about 5.0% in ResNet-50 and no extra training parameters. Moreover, it can be observed from Table 3 that our method is suitable to be combined with DOA to achieve an even better defense.

Effect of FNC on Feature Representations. Fig. 5 shows the IFNR on $\mathbf{f}^{(n)}$ of different depths on ResNet-50 without FNC, as well as that on $\mathbf{f}^{(n)}$ and $\mathbf{g}^{(n)}$ on ResNet-50 with FNC. It can be observed that FNC can effectively restrict the IFNR on feature maps in different layers, and the accumulated restriction increases with depth compared to the

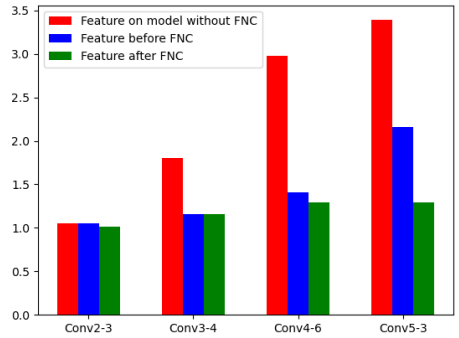

Figure 5. **Comparison of IFNR on feature maps in different layers of ResNet-50.** From left to right are feature in model without FNC, feature before FNC, and feature after FNC.

model without FNC. As such, the influence of the adversarial patch is gradually suppressed by the clipping processes.

Such an effect is also reflected on the norm maps of FFM and the gradient of the classification loss function as is shown in Fig. 6, in which from up to bottom are the input images, the FFM norm maps before and after FNC, and the gradient norm maps. It can be observed from the FFM norm maps that FNCs successfully suppress the large norm feature vectors in both clean and adversarial images and decrease the variance of the norm of feature vectors. Consequently, the features of the benign objects can dominate in classification and the impact of the patch is suppressed as is shown in the gradient norm maps in Fig. 6 (a)(c). The fail case shown in Fig. 6 (b) is probably due to the reason that the texture of the object of the ground truth label (chocolate sauce) is not obvious enough, leading to small norm of the features of the object. So the CNN classifier is still misled by the suppressed feature vectors of the patch.

4.3. Ablation Studies

To demonstrate the effectiveness of FNC, we also perform ablation studies for different experimental settings including the selection of hyper-parameters and target class, non-square patches, and location independent patch for single image. LaVAN attack with the 5% patch for ResNet-50 on ImageNet is employed in all the ablation experiments.

Selection of Hyper-Parameters. We test our method with

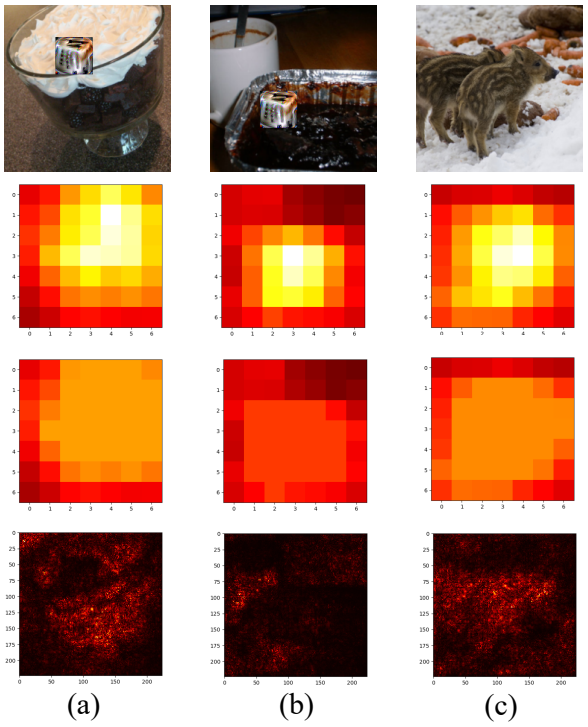

Figure 6. **Visualizations of the FFM norm map before/after FNC and gradient norm map on ResNet-50 with FNC.** (a) and (b) are the success case and fail case for defending against adversarial images, while the input of (c) is a clean image.

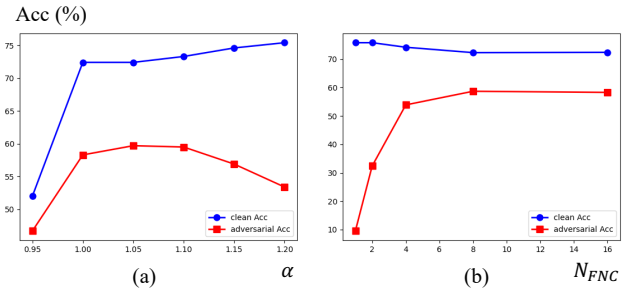

Figure 7. **(a) Curve of clean and adversarial Acc with α. (b) Curve of clean and adversarial Acc with N_{FNC}.**

different clipping parameters α, numbers of FNC layers N_{FNC} as well as target classes. When changing N_{FNC}, we keep the last residual block clipped and other FNC layers uniformly inserted. Fig. 7 illustrates the curve of clean and adversarial Acc w.r.t. the change of α and N_{FNC}. It can be observed that the clean Acc increases with the increase of α and the drop of N_{FNC}. But when α turns too large and the N_{FNC} is too small, the adversarial Acc decreases. A proper selection is to choose a relatively big N_{FNC} and an α between 1.0 and 1.1. The adversarial Acc against attacks for different target classes is shown in Table 4. Generally, our method keeps effective for different target classes.

Patch of Different Aspect Ratios. We generate rectangle patches of different aspect ratios and test the effectiveness of our method against them. As is analyzed before, as long as the size of the patch is far smaller than the entire image in any dimension, $\|\mathbf{f}^*_{i,j}\|$ has the same decay rate as the Gaussian function. When one dimension of the size of the patch gets large enough to be comparable to the image size, the decay rate of $\|\mathbf{f}^*_{i,j}\|$ in this dimension will decrease, leading to the diffusion of the impact of the patch. Theoretically, this will reduce the effectiveness of FNC. As is shown in Table 5, the adversarial Acc drops slightly for the $1:2$ patch, but quite obviously by 2.9% for the $1:4$ patch, in which the patch width is over 44% of that of the image.

Table 4. Adversarial Accs for different target classes.

Target	toaster	black swan	water jug	paper towel
Acc	58.3%	57.5%	63.3%	55.7%

Table 5. Adversarial Acc for patches of different shapes.

Aspect Ratio (height:width)	1:1	1:2	1:4
Acc	58.3%	58.1%	55.4%

Location Independent Patch for a Single Image. The universal assumptions in Section 3.2.2 do not hold in this case. Intuitively, it is more difficult to defend against such an attack for fewer restrictions in generating the patch. Nevertheless, it can be seen from Table 6 that the adversarial Acc only drops for 7.7%. The reason why our method is still effective for such an attack needs further investigation.

Table 6. Adversarial Acc for the universal patch (Universal) and the location independent patch for single image (Single Image).

Attacks	Universal	Single Image
Acc	58.3%	50.6%

5. Conclusion

Empirically, a universal adversarial patch usually leads to feature vectors with very large norms at the patched location in commonly used CNNs with global average pooling. We present a mathematical explanation to such a phenomenon based on which a novel defending method named FNC which adaptively clips deep feature norms is proposed. FNC is effective as long as the adversarial patch is spatially concentrated w.r.t. the original image, and it can be applied to various popular CNN architectures with very low computational overhead. Experimental results validate that FNC is effective against white-box patch attacks on various datasets and models. FNC significantly outperforms previous patch defending methods in terms of adversarial accuracy, and has relatively low influence on the accuracy of clean samples. Moreover, FNC can be readily combined with other defending methods like adversarial training. This work is supported by the National Natural Science Foundation of China (No. 61673234, No. U20B2062), and Beijing Science and Technology Planning Project (No.Z191100007419001).

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018.

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

[3] Ping Yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified defenses for adversarial patches. In *International Conference on Learning Representations*, 2020.

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramèr, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Physical adversarial examples for object detectors. In *Proceedings of the 12th USENIX Conference on Offensive Technologies*, 2018.

[5] Jamie Hayes. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1597–1604, 2018.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 720–729, 2020.

[8] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pages 2507–2515, 2018.

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems*, 2012.

[11] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4905–4913, 2016.

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[13] Muzammal Naseer, Salman Khan, and Fatih Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307. IEEE, 2019.

[14] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[16] Vikash Sehwag, Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Not all pixels are born equal: An analysis of evasion attacks under locality constraints. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 2285–2287, 2018.

[17] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K.Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications security*, pages 1528–1540, 2016.

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[19] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna Estrach, Dumitru Erhan, Ian Goodfellow, and Robert Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

[20] George Brinton Thomas and Ross L Finney. *Calculus*. Addison-Wesley Publishing Company, 1961.

[21] Hongjun Wang, Guangrun Wang, Ya Li, Dongyu Zhang, and Liang Lin. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep misranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 342–351, 2020.

[22] Tong Wu, Liang Tong, and Yevgeniy Vorobeychik. Defending against physically realizable attacks on image classification. In *International Conference on Learning Representations*, 2020.

[23] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. Patchguard: Provable defense against adversarial patches using masks on small receptive fields. *arXiv preprint arXiv:2005.10884*, 2020.

[24] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, pages 665–681. Springer, 2020.

[25] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14521–14530, 2020.