ColorFool: Semantic Adversarial Colorization

Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, Andrea Cavallaro
Centre for Intelligent Sensing
Queen Mary University of London, UK
{a.shahinshamsabadi,ricardo.sanchezmatilla,a.cavallaro}@qmul.ac.uk

Abstract

Adversarial attacks that generate small L_p-norm perturbations to mislead classifiers have limited success in black-box settings and with unseen classifiers. These attacks are also not robust to defenses that use denoising filters and to adversarial training procedures. Instead, adversarial attacks that generate unrestricted perturbations are more robust to defenses, are generally more successful in black-box settings and are more transferable to unseen classifiers. However, unrestricted perturbations may be noticeable to humans. In this paper, we propose a content-based black-box adversarial attack that generates unrestricted perturbations by exploiting image semantics to selectively modify colors within chosen ranges that are perceived as natural by humans. We show that the proposed approach, ColorFool, outperforms in terms of success rate, robustness to defense frameworks and transferability, five state-of-the-art adversarial attacks on two different tasks, scene and object classification, when attacking three state-of-the-art deep neural networks using three standard datasets. The source code is available at https://github.com/smartcameras/ColorFool.

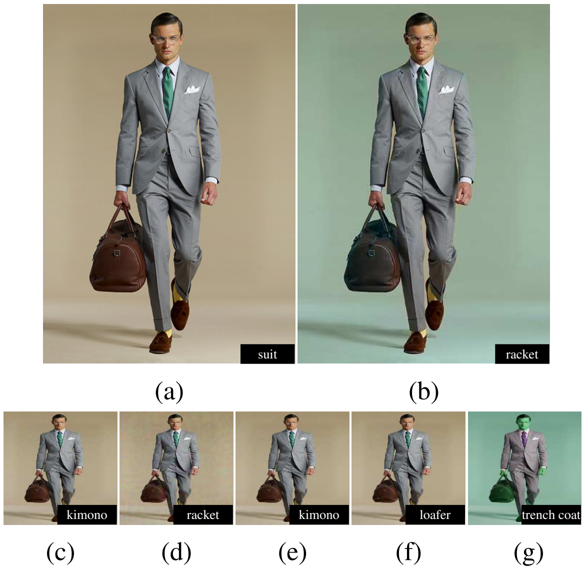
1. Introduction

Adversarial attacks perturb the intensity values of a *clean* image to mislead machine learning classifiers, such as Deep Neural Networks (DNNs). These perturbations can be restricted [7, 16, 19, 22, 23, 26] or unrestricted [1, 13] with respect to the intensity values in the clean image. Restricted perturbations, which are generated by controlling an L_p-norm, may restrain the maximum change for each pixel (L_∞-norm [7, 16, 19]), the maximum number of perturbed pixels (L_0-norm [22, 26]), or the maximum energy change (L_2-norm [23]); whereas unrestricted perturbations span a wider range, as determined by different colorization approaches [1, 13].

Defenses against adversarial attacks apply re-quantization [29], median filtering [29] and JPEG compression [4, 8] to remove adversarial perturbations prior to classification, or improve the robustness of the classifier through adversarial training [10], or by changing the loss functions [25]. The property of *robustness* of an adversarial attack is the success rate of misleading a classifier in the presence of defense frameworks. Most adversarial attacks assume white-box settings, i.e. the attacker has full knowledge of the architecture and parameters (and hence gradients) of the classifier [7, 16, 19, 22, 23, 26]. However, real-world scenarios may prevent access to the classifier (unseen classifier) or limit the exposed information to only the output of the classifier (black-box settings). The property of *transferability* is the success rate of adversarial images in misleading an unseen classifier [5]. Finally, the perturbation in an adversarial image should be *unnoticeable*, i.e. the shape and spatial arrangement of objects in the adversarial image should be perceived as in the clean image and the colors should look natural.

Restricted perturbations [2, 16, 19, 23] often have high spatial frequencies that can be detected by defenses [4, 8, 21, 29]. Moreover, restricted perturbations that are sparse and with large changes in intensity are noticeable [22, 26]. Instead, unrestricted attacks arbitrarily perturb intensity values through a colorization process [1], which is based on an expensive training phase, followed by per-image adversarial fine-tuning. Alternatively, attacks can arbitrarily change the hue and saturation components in the HSV color space [13]. However, even small variations can result in large and perceivable distortions caused by unnatural colors (see Fig. 1(g)).

In this paper, we propose a black-box, unrestricted, content-based adversarial attack that exploits the characteristics of the human visual system to selectively alter colors. The proposed approach, *ColorFool*, operates only on the de-correlated a and b channels of the perceptually uniform Lab color space [27], without changing the lightness, L. Moreover, ColorFool introduces perturbations only within a chosen natural-color range for specific semantic categories [30]. Unlike other adversarial attacks, the proposed adversarial perturbation can be generated for images



(a) (b)

(c) (d) (e) (f) (g)

Figure 1. Adversarial image generated for a sample (a) clean image by (b) ColorFool, (c) Basic Iterative Method (BIM) [16], (d) Translation-Invariant BIM [7], (e) DeepFool [23], (f) SparseFool [22] and (g) SemanticAdv [13]. BIM and DeepFool generate unnoticeable adversarial images with restricted perturbations. ColorFool generates any-size, natural-color adversarial images by considering the semantic information and preserving the colors of regions within an image that the human vision system is more sensitive to (in this case the person). The text in the bottom right of each image indicates the predicted class.

of any size (see Fig. 1(b)). We validate ColorFool in attacking three state-of-the-art DNNs (ResNet50, ResNet18 [11] and AlexNet [15]) that have been trained for scene and object classification tasks using three datasets (ImageNet [6], CIFAR-10 [14] and Private Places365 (P-Places365) [31]). We show that ColorFool generates natural-color adversarial images that are effective in terms of success rate in seen and unseen classifiers, robustness to defense frameworks on an extensive comparison with five state-of-the-art attacks.

2. Adversarial attacks

Let $\mathbf{X} \in \mathbb{Z}^{w,h,c}$ be an RGB clean image with width w, height h and $c = 3$ color channels. Let $M(\cdot)$ be a DNN classifier that predicts for a given image the most probable class, $y = M(\mathbf{X})$. An adversarial attack perturbs $\mathbf{X}$ to generate an adversarial image, $\dot{\mathbf{X}}$, such that $M(\dot{\mathbf{X}}) \neq M(\mathbf{X})$.

Adversarial attacks can be grouped based on their perturbations into two categories, namely restricted and unrestricted. An adversarial image can be generated with a perturbation controlled by L_0, L_1, L_2 or L_∞-norms. Adversarial attacks that use restricted perturbations are Basic Iterative Method (BIM) [16], Translation-Invariant BIM (TI-BIM) [7], DeepFool [23] and SparseFool [22]. Alternatively, an adversarial image can be generated with an unrestricted perturbation on the colors considering the preservation of the shape of the objects, as in SemanticAdv [13] or BigAdv [1].

BIM [16] constrains the maximum perturbation of each pixel by imposing an L_∞-norm constraint. BIM searches for adversarial images by linearising the cost function, $J_M(\cdot)$, in the input space. The search starts from $\dot{\mathbf{X}}_0 = \mathbf{X}$ and iteratively moves in the direction of the gradient of the cost of predicting y with respect to the input image, $\nabla_{\mathbf{X}} J_M(\cdot)$, with step size δ in each iteration:

$$\dot{\mathbf{X}}_N = C_{\mathbf{X},\epsilon}\left(\dot{\mathbf{X}}_{N-1} + \delta \operatorname{sign}\left(\nabla_{\mathbf{X}} J_M(\theta, \dot{\mathbf{X}}_{N-1}, y)\right)\right), \tag{1}$$

until $M(\dot{\mathbf{X}}_N) \neq M(\mathbf{X})$ or a maximum number of N iterations, where θ are the parameters of $M(\cdot)$, sign$(\cdot)$ is the sign function that determines the direction of the gradient of the cost function. $C_{\mathbf{X},\epsilon}(\cdot)$ is a clipping function that maintains the adversarial images within the ϵ-neighborhood of the clean image as well as $[0, 255]$:

$$C_{\mathbf{X},\epsilon}(\dot{\mathbf{X}}) = \min\left\{\mathbf{255}, \mathbf{X} + \epsilon, \max\left\{\mathbf{0}, \mathbf{X} - \epsilon, \dot{\mathbf{X}}\right\}\right\}, \tag{2}$$

where $\mathbf{0}$ and $\mathbf{255}$ are images whose pixel intensities are all 0 and 255, respectively, and min$(\cdot)$/max$(\cdot)$ are the per-pixel min/max operation.

TI-BIM [7] generates BIM adversarial perturbations over an ensemble of translated images to improve the transferability to unseen classifiers. As the gradients of a translated image correspond to translating the gradient of the original image [7], TI-BIM convolves the gradient with a pre-defined kernel $\mathbf{W}$ at each iteration:

$$\dot{\mathbf{X}}_N = C_{\mathbf{X},\epsilon}\left(\dot{\mathbf{X}}_{N-1} + \delta \operatorname{sign}\left(\mathbf{W} * \nabla_{\mathbf{X}} J_M(\theta, \dot{\mathbf{X}}_{N-1}, y)\right)\right), \tag{3}$$

where $\mathbf{W}$ can be a uniform, linear or Gaussian kernel.

DeepFool [23] finds the minimal L_2-norm adversarial perturbation by finding the direction towards the closest decision boundary. For example in the case of the binary classifier, adversarial images can iteratively be generated by projecting the adversarial image of each iteration onto the closest linearized decision boundary of $M(\cdot)$:

$$\dot{\mathbf{X}}_N = \mathbf{X} + (1 + \eta) \sum_{n=1}^{N} -\frac{M(\dot{\mathbf{X}}_n)}{\|\nabla M(\dot{\mathbf{X}}_n)\|_2^2} \nabla M(\dot{\mathbf{X}}_n), \tag{4}$$

where $\eta \ll 1$ is a constant that is multiplied by the accumulative adversarial perturbations to reach the other side of the decision boundary. Note that DeepFool does not impose constraints on pixel values, which, as a result, may lie outside the permissible dynamic range.

SparseFool [22] uses the L_1-norm between the clean and adversarial images to minimize the number of perturbed

Table 1. Comparison of adversarial attacks. An adversarial image is generated with a perturbation restricted by L_0, L_1, L_2 and L_∞ or unrestricted perturbation on the **C**olors considering two **Type**s: **W**hite- or **B**lack-attack for two **Task**s: **O**bject and **S**cene classification. **Datasets** with the number of chosen classes for the attack are reported for ImageNet, CIFAR-10 and Private-Places365 (P-Places365) with 1000, 10 and 60 classes as well (**U**nknown, if number of classes are not written in their papers). JSMA is tested on the MNIST dataset (10 classes). KEY– BIM: Basic Iterative Method; TI-BIM: Translation-Invariant BIM; P-BIM: Private BIM; CW: Carlini-Wagner; JSMA: Jacobian-based Saliency Map Attack; and SemAdv: Semantic Adversarial.

Ref	Attack	Perturbation	Type	Attacked classifier	Datasets			Task
					ImageNet	CIFAR-10	P-Places365	
[16]	BIM	L_∞	W	Inc-v3	1000			O
[7]	TI-BIM ("TI-FGSM")	L_∞	W	Inc-v3, Inc-v4, ResNet152	60			O
[19]	P-BIM ("P-FGSM")	L_∞	W	ResNet50			60	S
[23]	DeepFool	L_2	W	LeNet, CaffeNet, GoogleNet	1000	10		O
[22]	SparseFool	L_1	W	LeNet, ResNet18, Inc-v3, DenseNet,VGG16	U	10		O
[26]	JSMA	L_0	W	LeNet				O
[2]	CW	$L_{0,2,\infty}$	W	Inc-v3	1000	10		O
[1]	BigAdv	C	W	ResNet50, DenseNet121, VGG19	10			O
[13]	SemAdv	C	B	VGG16		10		O
ours	ColorFool	C	B	ResNet50, ResNet18, AlexNet	1000	10	60	O and S

pixels. SparseFool leverages the fact that DNNs have a low mean curvature in the neighborhood of each image [9] and generates sparse perturbations based on this curvature and adversarial images on the closest L_2 decision boundary. SparseFool approximates the decision boundary near the clean image $\mathbf{X}$ by a hyperplane, $\mathbf{v}^T\dot{\mathbf{X}}_{\text{DF}}$, passing the minimal L_2-norm DeepFool adversarial image (i.e. Eq 4), $\dot{\mathbf{X}}_{\text{DF}}$, and a normal vector $\mathbf{v}$. Then, SparseFool iteratively finds the minimal L_1-norm projection of the clean image onto the approximated decision boundary:

$$\dot{\mathbf{X}}_N = D(\dot{\mathbf{X}}_{N-1} + \boldsymbol{\delta}^*), \tag{5}$$

where $\boldsymbol{\delta}^*$ is the sparse adversarial perturbation and $D(\cdot)$ is a clipping function that maintains the pixel values between $[0, 255]$:

$$D(\dot{\mathbf{X}}) = \min\left\{\mathbf{255}, \max\{\mathbf{0}, \dot{\mathbf{X}}\}\right\}. \tag{6}$$

Each d-th value of SparseFool perturbation, δ^*_d, is iteratively computed as

$$\delta^*_d = \frac{|\mathbf{v}^T(\dot{\mathbf{X}}_n - \dot{\mathbf{X}}_{\text{DF}})|}{|v_d|}\,\text{sign}(v_d), \tag{7}$$

where T is the transpose operator.

SemanticAdv [13] unrestrictedly changes colors in the HSV color space by shifting the hue, $\mathbf{X}_H$, and saturation, $\mathbf{X}_S$, of the clean image while preserving the value channel, $\mathbf{X}_V$, in order to not affect the shapes of objects:

$$\dot{\mathbf{X}}_N = \beta\left(\left[\mathbf{X}_H + [\delta_H]^{w,h}, \mathbf{X}_S + [\delta_S]^{w,h}, \mathbf{X}_V\right]\right), \tag{8}$$

where δ_S, $\delta_H \in [0, 1]$ are scalar random values, and $\beta(\cdot)$ is a function that converts the intensities from the HSV to the RGB color space. Eq. 8 is repeated until $M(\dot{\mathbf{X}}_N) \neq M(\mathbf{X})$ or a maximum number of trials (1000) is reached.

BigAdv [1] aims to generate natural-color perturbations by fine-tuning, for each $\mathbf{X}$, a trained colorization model [30], $F(\cdot)$, parameterized by $\hat{\theta}$ with a cross-entropy adversarial loss J_{adv} [2]:

$$\dot{\mathbf{X}} = \arg\min_{\hat{\theta}} J_{adv}\left(M\left(F(\mathbf{X}_L, \mathbf{C}_h, \mathbf{L}_h; \hat{\theta})\right), y\right), \tag{9}$$

where $\mathbf{X}_L$ is the L value of the image in the Lab color space and $\mathbf{C}_h$ is the ground-truth color for the locations that are indicated by the binary location hint, $\mathbf{L}_h$. BigAdv decolorizes the whole image and again colorizes it. This process may severely distort the colors if $\mathbf{C}_h$ and $\mathbf{L}_h$ are not carefully set.

Finally, ColorFool, the proposed approach (see Sec. 3), is an unrestricted, black-box attack like SemanticAdv. However, SemanticAdv perturbs pixel intensities without considering the content in an image thus often producing unnatural colors. ColorFool instead perturbs colors only in specific semantic regions and within a chosen range so they can be still perceived as natural. The other state-of-the-art unrestricted attack, BigAdv, is a white-box attack that trains a colorization model to learn image statistics from large datasets and to fine-tune the model for each image. Tab. 1 summarizes the adversarial attacks for object or scene classification tasks.

3. ColorFool

We aim to design a black-box adversarial attack that generates adversarial images with natural colors through generating low-frequency perturbations that are highly transferable to unseen classifiers and robust to defenses. Moreover, the attack shall operate on the native size of the images.

First, we identify image regions whose color is important for a human observer as the appearance of these sensitive regions (e.g. human skin) is typically within a specific range. Other (non-sensitive) image regions (e.g. wall and

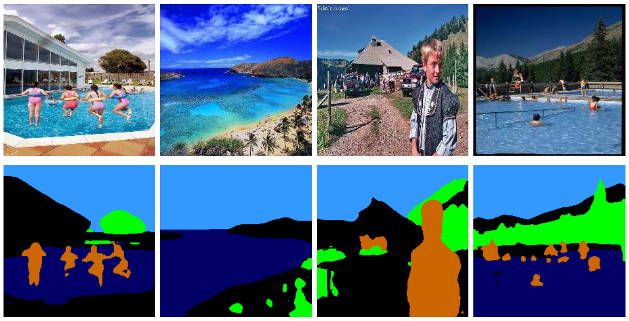


Figure 2. Sample results of image semantic segmentation [32]. ColorFool identifies non-sensitive regions (in black) and color-sensitive semantic regions, namely person (in orange), vegetation (in green), sky (in light blue) and water (in dark blue).

Table 2. Adversarial color perturbation considered by ColorFool to modify the colors of sensitive semantic regions. The natural-color ranges are chosen based on the color recommendation of people to gray-scale objects [3, 18] that are also used as ground-truth colors in colorization methods [30]. The adversarial color perturbation of the k-th semantic region considers the extreme values of the semantic class as $l_k^a = \min(\mathbf{S}_k)$ and $u_k^a = \max(\mathbf{S}_k)$. The adversarial perturbation is chosen randomly within each natural-color range and applied as in Eq. 11. Note that no color changes are applied to image regions classified as person.

Semantic region	a channel	b channel
$\mathcal{S}_1$: Person	$\mathcal{N}_1^a = \{0\}$	$\mathcal{N}_1^b = \{0\}$
$\mathcal{S}_2$: Vegetation	$\mathcal{N}_2^a = \{\text{-}128\text{-}l_2^a, \ldots, \text{-}u_2^a\}$	$\mathcal{N}_2^b = \{\text{-}l_2^b, \ldots, 127\text{-}u_2^b\}$
$\mathcal{S}_3$: Water	$\mathcal{N}_3^a = \{\text{-}128\text{-}l_3^a, \ldots, \text{-}u_3^a\}$	$\mathcal{N}_3^b = \{\text{-}128\text{-}l_3^b, \ldots, \text{-}u_3^b\}$
$\mathcal{S}_4$: Sky	$\mathcal{N}_4^a = \{\text{-}128\text{-}l_4^a, \ldots, \text{-}u_4^a\}$	$\mathcal{N}_4^b = \{\text{-}128\text{-}l_4^b, \ldots, \text{-}u_4^b\}$

curtains in Figs. 1 and 5, first row), instead, may have their colors modified within an arbitrary range and still look natural [3]. We consider four categories of sensitive regions, whose unusual colors would attract the attention of a human observer [3, 18, 30]: person, sky, vegetation (e.g. grass and trees), and water (e.g. sea, river, waterfall, swimming pool and lake).

Let us decompose an image $\mathbf{X}$ into K semantic regions

$$\mathcal{S} = \{\mathbf{S}_k : \mathbf{S}_k = \mathbf{X} \cdot \mathbf{M}_k\}_{k=1}^{K}, \tag{10}$$

where $\mathbf{M}_k \in \{0,1\}^{w,h}$ is a binary mask that specifies the location of pixels belonging to region $\mathbf{S}_k$ and "$\cdot$" denotes a pixel-wise multiplication. Binary masks are outputted by a pyramid Pooling R50-Dilated architecture of Cascade Segmentation Module segmentation [33], trained on the MIT ADE20K dataset [32] on 150 semantic region types. Fig. 2 shows examples of the considered semantic regions.

We separate the sensitive regions, $\mathbb{S} = \{\mathbf{S}_k\}_{k=1}^{S}$, from the non-sensitive regions, $\overline{\mathbb{S}} = \{\overline{\mathbf{S}}_k\}_{k=1}^{\overline{S}}$, where $\mathcal{S} = \mathbb{S} \cup \overline{\mathbb{S}}$ and $\cup$ is the union operator. After identifying these two sets, we appropriately modify the colors of the region in the perceptually uniform Lab color space [27], which separates color information from brightness: a ranges from green (-128) to red (+127), b ranges from blue (-128) to yellow (+127), and L ranges from black (0) to white (100).

We then modify the color of the sensitive regions, $\mathbb{S}$, to generate the adversarial set $\dot{\mathbb{S}}$ as

$$\dot{\mathbb{S}} = \{\dot{\mathbf{S}}_k : \dot{\mathbf{S}}_k = \gamma(\mathbf{S}_k) + \alpha[0, N_k^a, N_k^b]^T\}_{k=1}^{S}, \tag{11}$$

where $\gamma(\cdot)$ converts the intensities of an image from the RGB to the Lab color space, $N_k^a \in \mathcal{N}_k^a$ and $N_k^b \in \mathcal{N}_k^b$ are the adversarial perturbations in the channels a and b that are chosen randomly from the set of natural-color ranges [30], $\mathcal{N}_k^a$ and $\mathcal{N}_k^b$, in the a and b channels. These ranges are defined based on the actual colors, region semantics and prior knowledge about color perception in that region type (see Tab. 2). We allow multiple trials, until a perturbation misleads the classifier. Let n be the index of the trial and N be the maximum number of trials. To avoid large color changes in the first trials, we progressively scale the randomly chosen perturbation by $\alpha = \frac{n}{N}$.

We modify the color of the non-sensitive regions, $\overline{\mathbb{S}}$, to produce the set $\dot{\overline{\mathbb{S}}}$ as

$$\dot{\overline{\mathbb{S}}} = \{\dot{\overline{\mathbf{S}}}_k : \dot{\overline{\mathbf{S}}}_k = \gamma(\overline{\mathbf{S}}_k) + \alpha[0, \overline{N}^a, \overline{N}^b]^T\}_{k=1}^{\overline{S}}, \tag{12}$$

where $\overline{N}^a \in \{-127, \ldots, 128\}$ and $\overline{N}^b \in \{-127, \ldots, 128\}$ are chosen randomly inside the whole range of a and b, as the regions can undergo larger intensity changes.

Finally, the adversarial image $\dot{\mathbf{X}}$ generated by ColorFool combines the modified sensitive and non-sensitive image regions as

$$\dot{\mathbf{X}} = Q\left(\gamma^{-1}\left(\sum_{k=1}^{S} \dot{\mathbf{S}}_k + \sum_{k=1}^{\overline{S}} \dot{\overline{\mathbf{S}}}_k\right)\right), \tag{13}$$

where $Q(\cdot)$ is the quantization function which ensures that the generated adversarial image is in the dynamic range of the pixel values, $\dot{\mathbf{X}} \in \mathbb{Z}^{w,h,c}$, and $\gamma^{-1}(\cdot)$ is the inverse function that converts the intensities of an image from the Lab to the RGB color space.

4. Validation

Algorithms under comparison. We compare the proposed attack, ColorFool, against the state-of-the-art adversarial attacks discussed in Section 2: Basic Iterative Method (BIM) [16], Translation-Invariant BIM (TI-BIM) [7], DeepFool [23], SparseFool [22] and SemanticAdv [13] (we excluded BigAdv [1] as no code was available at the time of submission). These attacks include restricted and unrestricted perturbations and generate adversarial images that are transferable (TI-BIM), unnoticeable (DeepFool) and robust to defenses (SemanticAdv). We also compare against the simple yet successful BIM attack and SparseFool, a sparse attack. Furthermore, we consider a modification of the proposed attack, named ColorFool-r, where no priors are considered for the semantic regions. We use the authors' implementations for all adversarial attacks apart from SemanticAdv that we re-implemented in PyTorch. All adversarial images are generated using the same read/write framework, image filters and software version in PyTorch and OpenCV to make the results comparable.

Datasets. We use three datasets Private-Places365 (P-Places365) [31], a scene classification dataset; CIFAR-10 [14], an object classification dataset; and ImageNet [6], another object classification dataset. For P-Places365, we employ a subset of classes that were defined as sensitive in the MediaEval 2018 Pixel Privacy Challenge [17]. P-Places365 includes 50 images for each of the 60 private scene classes. For CIFAR-10, we use the whole test set, which is composed of 10K images of 10 different object classes. For ImageNet, we consider the 1000 classes and 3 random images per class from the validation set. All the images are RGB with varying resolution except for the images from CIFAR-10 whose $w = h = 32$.

Classifiers under attack. We conduct the attacks on two different architectures: a deep residual neural network (ResNet [11], 18 layers (R18) and 50 layers (R50)) and AlexNet (AN) [15]. We choose these three classifiers to study the transferability comparing both homogeneous (i.e. ResNet classifiers) and heterogeneous architectures (i.e. AlexNet).

Table 3. Success rate on Private-Places365 (P-Places365), CIFAR-10 and ImageNet datasets against ResNet50 (R50), ResNet18 (R18) and AlexNet (AN). The performance of these classifiers on the clean images is presented in the third row. The higher the success rate, the most successful the attack. KEY– AC: attacked classifier; TC: test classifier; Acc: accuracy. A gray (white) cell denotes a seen (unseen) classifier. ColorFool is more transferable than other adversarial attacks, except SemanticAdv, which however severely distort the colors of all regions (see Fig. 5).

Attack	Dataset	P-Places365			CIFAR-10			ImageNet		
	TC \ AC	R50	R18	AN	R50	R18	AN	R50	R18	AN
Acc. on *clean* images		.554	.527	.466	.944	.935	.722	.726	.649	.517
BIM	R50	1.00	.284	.073	.999	.095	.021	.873	.123	.087
	R18	.231	1.00	.081	.078	.999	.022	.143	.945	.099
	AN	.061	.081	1.00	.014	.013	.999	.088	.092	.944
TI-BIM	R50	.995	.339	.186	.843	.153	.173	.992	.235	.176
	R18	.268	.996	.198	.083	.943	.138	.173	.997	.183
	AN	.157	.193	.995	.315	.349	.889	.121	.163	.994
DF	R50	.957	.107	.030	.829	.226	.064	.983	.071	.018
	R18	.009	.969	.030	.234	.875	.076	.055	.991	.017
	AN	.021	.028	.956	.020	.024	.637	.017	.019	.993
SF	R50	.998	.151	.127	.999	.408	.186	.987	.167	.176
	R18	.101	.999	.120	.353	.999	.216	.086	.997	.134
	AN	.070	.066	1.00	.130	.151	.999	.062	.079	.999
SA	R50	.936	.563	.713	.863	.429	.704	.889	.540	.769
	R18	.480	.954	.714	.339	.898	.705	.422	.931	.757
	AN	.424	.466	.990	.155	.191	.993	.359	.431	.994
CF-r	R50	.963	.336	.514	.956	.255	.635	.948	.362	.608
	R18	.275	.970	.501	.431	.954	.689	.235	.954	.580
	AN	.157	.171	.999	.065	.058	.999	.104	.137	.998
CF	R50	.959	.334	.491	.975	.254	.641	.917	.348	.592
	R18	.267	.971	.475	.415	.971	.696	.223	.934	.543
	AN	.171	.157	.998	.059	.055	1.00	.114	.147	.995

Performance measures. We quantify the *success rate* in misleading a classifier, the *robustness* to defenses and the *image quality* of the adversarial images. The *success rate* (SR) is quantified as the ratio between the number of adversarial images that mislead the classifier on its most-likely predicted class and the total number of images. For the transferability, we compute the SR of adversarial images generated for a seen classifier in misleading unseen classifiers. The *robustness* to defenses is measured as follows. Firstly, we quantify the SR in seen classifiers of adversarial images after filtering. As filters we use re-quantization [29] with 1 to 7 bits, in steps of 1; median filtering [29], with squared kernel of dimension 2, 3 and 5; and lossy JPEG compression [8, 4], with quality parameters 25, 50, 75 and 100. We report the results on retrieving the class that was predicted on the clean images with the most effective filter (i.e. the one that obtains the lowest SR). Secondly, we report the undetectability as the ratio between adversarial images not identified as adversarials and the total number of images using the previously mentioned image filters [29]. Specifically, for each classifier and parameter of each image filter, we compute a threshold that determines if an image is adversarial or clean by comparing the L_1-norm of the difference between the prediction probability vectors of the given image and the same image after the image filtering. Each threshold is calculated as the value that allows for a 5% false-positive rate in detecting clean images on a training dataset. Then, images with L_1-norm difference larger than the threshold are considered to be adversarials. Thirdly, we evaluate the SR when attacking a seen classifier trained with Prototype Conformity Loss (PCL) [25] and adversarial training [10]. Finally, we quantify the image quality of the adversarial image with a non-reference perceptual image quality measure named neural image assessment (NIMA) [28] trained on the AVA dataset [24]. NIMA estimates the perceived image quality and was shown to predict human preferences [17].

Success rate. Tab. 3 shows the SR on a seen classifier (on-diagonal elements) and transferability to unseen classifiers (off-diagonal elements). All adversarial attacks achieve high SR in a seen classifier for most of the classifiers and datasets. Restricted attacks never achieve SRs higher than 0.41 in unseen classifiers, while unrestricted attacks achieve a SR of up to 0.77. ColorFool achieves a high SR on both seen and unseen classifiers with, for example, 0.97 when both attacking and testing in R18 in CIFAR-10 and 0.69 and 0.41 when evaluated with AN and R50, respectively. However, other attacks only achieve SRs of 0.02 (BIM), 0.14 (TI-BIM), 0.07 (DeepFool), 0.21 (SparseFool). A possible reason is that restricted attacks such as BIM, iteratively overfit to the parameters of the specific classifier, which means that the adversarial images rarely mislead other classifiers, while the randomness in changing the color in ColorFool prevents this overfitting. TI-BIM overcomes the overfitting of BIM and achieves higher transferability than BIM, while its SR in seen classifiers decreases. Unrestricted attacks obtain high transferability rates. For instance, in the CIFAR-10 dataset, SemanticAdv, ColorFool-r and ColorFool obtain SRs of 0.71, 0.69 and 0.70 when attacking R18 and evaluating in AN. While ColorFool outperforms SemanticAdv with seen classifiers, SemanticAdv obtains higher transferability rates. This is due to the large color changes that SemanticAdv introduces in the whole image, including regions that are more informative for the classifier (higher transferability) but also regions that are sensitive for the human vision system, thus generating unnatural colors (see Fig. 5). Further insights are discussed in the image quality analysis later in this section. As previously studied [20], adversarial images generated on stronger classifiers (e.g. R50) have a higher transferability rate when tested on weaker classifiers (e.g. AN). This behavior can be observed, for instance, when looking at the results of ColorFool in P-Places365. Adversarial images crafted with R50 obtain a SR of 0.96, which decreases to 0.49 when tested in AN. However, when adversarial images are crafted with AN the SR is 0.99, but when tested in R50 (a stronger classifier) the SR is only 0.17.

Table 4. Success rate of Basic Iterative Method (BIM), Translation-Invariant BIM (TI-BIM), DeepFool (DF), SparseFool (SF), SemanticAdv (SA), ColorFool-r (CF-r) and ColorFool (CF) against ResNet110 trained with softmax, on Prototype Conformity Loss (PCL) [25] and its combination with adversarial training (AdvT) [10] on CIFAR-10. The higher the success rate, the more robust the attack. In bold, the best performing attack.

Training	BIM	TI-BIM	DF	SF	SA	CF-r	CF
	[16]	[7]	[23]	[22]	[13]	ours	ours
Softmax	.969	.963	.855	**.994**	.867	.992	**.994**
PCL	.560	.619	.784	.801	.896	**1.00**	**1.00**
PCL+AdvT	.500	.577	.665	.691	.966	.998	**.999**

Robustness to defenses. The SR of adversarial attacks after applying any of the three image filters is depicted in Fig. 3. Restricted attacks such as DeepFool and SparseFool are the least robust to image filtering, as these filters can remove restricted adversarial noises (especially L_0 sparse adversarial perturbation) prior to the classification and correctly classify around 70% of them. BIM and TI-BIM obtain higher SR than other restricted attacks in P-Places365 and ImageNet but similar in CIFAR-10. The most robust attacks are the unrestricted ones where SemanticAdv, ColorFool-r and ColorFool consistently obtain a SR above 60% across datasets and classifiers. The undetectability results (Fig. 4) show that restricted attacks are more detectable than unrestricted ones when considering all image filters across all classifiers and datasets. For instance, when attacking R50 in P-Places365, BIM, TI-BIM, DeepFool and SparseFool obtain undetectability rates of 5%, 19%, 1% and 11% with re-quantization, median filtering and JPEG compression. Unrestricted attacks such as SemanticAdv, ColorFool-r and ColorFool obtain 73%, 72% and 75%. We believe that one reason for this is related to the spatial frequency of the generated adversarial perturbations. Restricted attacks generate high-frequency adversarial perturbations, whereas unrestricted attacks generate low-frequency perturbations (see Fig. 5). Low-frequency perturbations (those generated by unrestricted attacks) are more robust to re-quantization, median filtering and JPEG compression. In general, JPEG compression is the most effective detection framework. When we consider all of the filters applied, as an example, to P-Places365, the restricted attacks BIM, TI-BIM, DeepFool and SparseFool are detectable in 95%, 81%, 99% and 89% of the cases. However, unrestricted attacks such as SemanticAdv and ColorFool-r are detectable in only 27% of the cases and ColorFool is the least detectable (25%).

Another observation is that the robustness of the adversarial images is proportional to the accuracy of the classifier used for their generation (see Figs. 3, 4). For example, misleading a highly-accurate DNN such as R50, which obtains an accuracy of almost 0.95 on CIFAR-10, needs larger perturbations, which increase the robustness but also the detectability.

Tab. 4 shows the SR of adversarial attacks in misleading ResNet110 [11] trained on CIFAR-10 as well as the robustness to an improved training procedure based on PCL [25] and its combination with adversarial training [10]. For the adversarial training, ResNet110 is trained on the clean and adversarial images generated by BIM (i.e. the strongest defense [12]). Tab. 4 shows that ColorFool is robust as its SR remains above 99% when misleading ResNet110 equipped with both PCL and adversarial training defenses. Instead, the SR of restricted adversarial attacks drops considerably.

Quality. Sample adversarial images are shown in Fig. 5. For instance, even though the restricted attacks such as TI-BIM or SparseFool generate adversarial images with minimal perturbations, they are noticeable. SemanticAdv and ColorFool-r generate unrealistic colors. However, even if ColorFool generates adversarial images that are largely different (in an L_p-norm sense) from the clean images, they

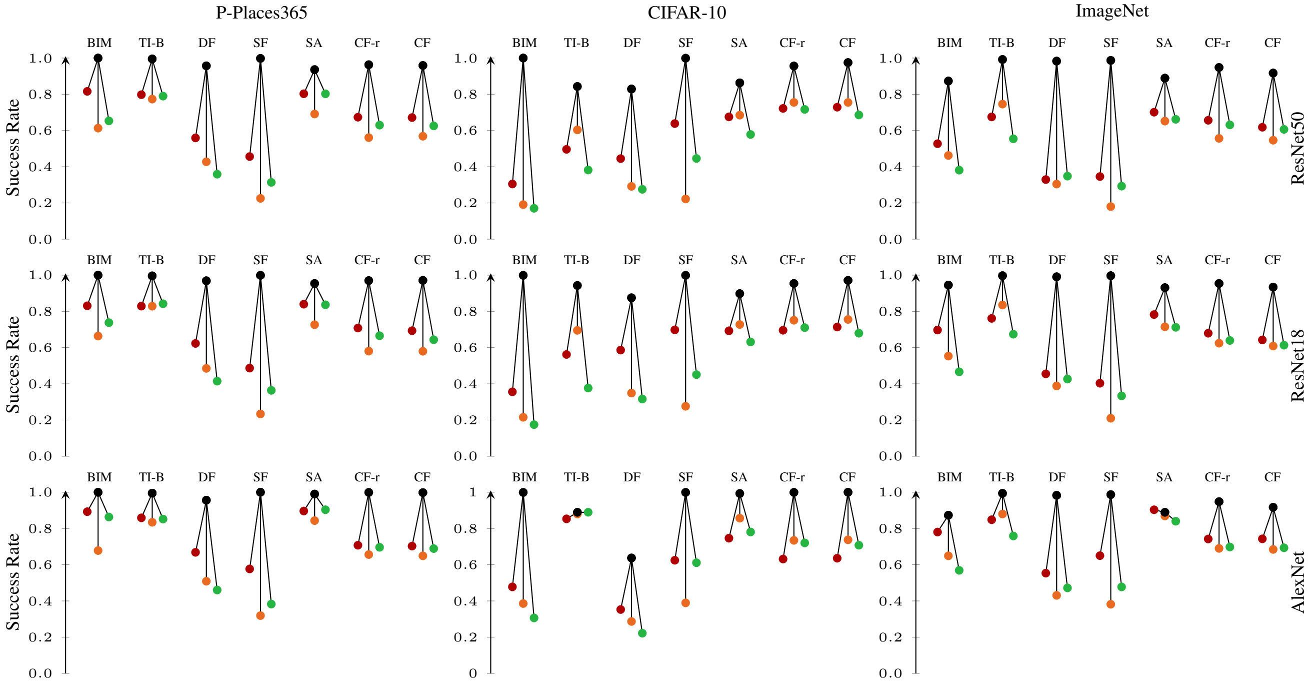


Figure 3. Robustness of Basic Iterative Method (BIM), Translation-Invariant BIM (TI-B), DeepFool (DF), SparseFool (SF), SemanticAdv (SA), ColorFool-r (CF-r) and ColorFool (CF) on ResNet50, ResNet18 and AlexNet against re-quantization (●), median filtering (●) and JPEG compression (●) on the Private subset of Places365 (P-Places365), CIFAR-10 and ImageNet.

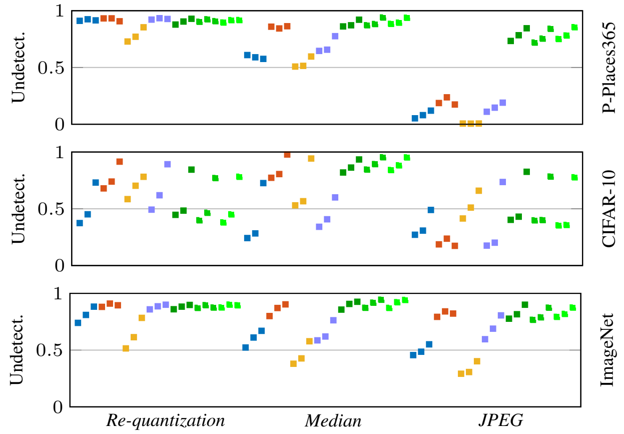


Figure 4. Undetectability (Undetec.) of ■ Basic Iterative Method (BIM), ■ Translation-Invariant BIM (TI-BIM), ■ DeepFool, ■ SparseFool, ■ SemanticAdv, ■ ColorFool-r and ■ ColorFool, when attacking ResNet50, ResNet18 and AlexNet classifiers (first, second and third square of each color, respectively) using re-quantization, median filtering and JPEG compression. The higher the undetectability, the higher the robustness to defenses.

look natural. Moreover, ColorFool generates images with the same dimensions as the clean images. The results of the image quality evaluation are shown in Tab. 5. Unrestricted attacks obtain the highest NIMA scores across all attacks, classifiers and datasets. Specifically, in P-Places365 and ImageNet, ColorFool-r and ColorFool obtain the highest scores (over 5.19). For CIFAR-10, SemanticAdv, ColorFool-r and ColorFool obtain similar results with scores over 4.96. This implies that adversarial images generated by ColorFool do not deteriorate the perceived image quality while restricted attacks such as DeepFool or SparseFool obtain slightly lower results. ColorFool obtains equal or higher NIMA scores than the clean images considering all datasets and classifiers.

Randomness analysis. As ColorFool generates random perturbations, we analyze what is the effect of this randomness on the SR, the number of trials to converge and whether the predicted class of the generated adversarial image varies. We execute ColorFool 500 times with thirteen random images from ImageNet that belong to different classes for attacking R50. We select R50 for this analysis as it is the most accurate classifier among the considered ones. Fig. 6 shows the SR, the statistics (median, min, max, 25 percentile and 75 percentile) of the number of trials to converge and the number of classes that the executions converge to. Results for different images are shown on the x-axis. We can observe that the number of trials that ColorFool requires to converge remains with a low median and standard deviation for images that always succeed in misleading the classifier (see the first and second plot in Fig. 6). Finally, most of the executions for a given image converge to the same class (see median value in the third

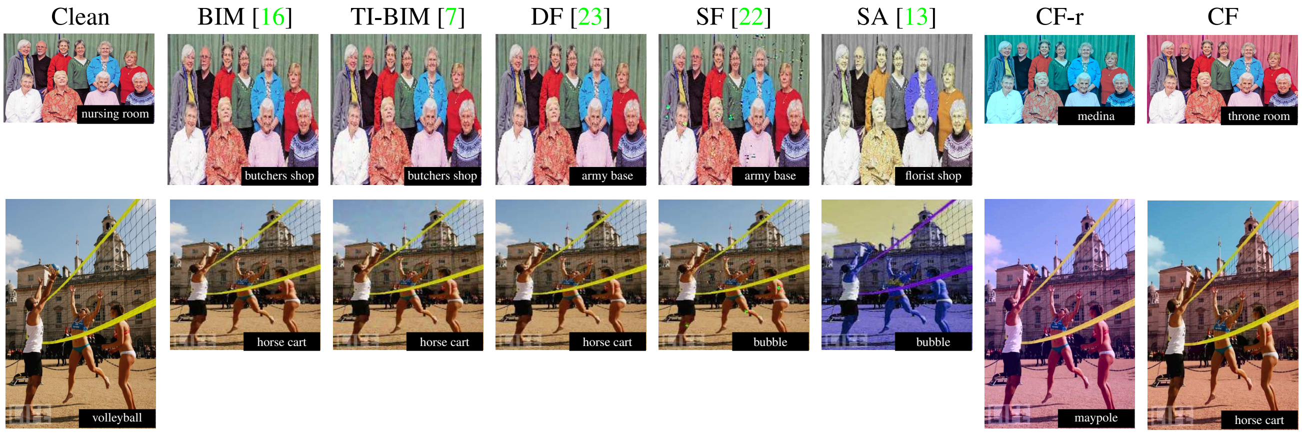


Figure 5. Adversarial image samples from Private-Places365 (first row) and ImageNet (second row) datasets generated by Basic Iterative Method (BIM), Translation-Invariant BIM (TI-BIM), DeepFool (DF), SparseFool (SF), SemanticAdv (SA), ColorFool-r (CF-r) and the proposed ColorFool (CF). Please note that CF-r and CF generate examples at the native image resolution. The predicted class is shown on the bottom right of each image.

Table 5. Image quality (NIMA, the higher the better) of adversarial images from the Private-Places365 dataset, CIFAR-10 and ImageNet datasets for all adversarial attacks against ResNet50 (R50), ResNet18 (R18) and AlexNet (AN). We report only the mean value as the standard deviations are similar across all attacks with typical values of 4.4. KEY – AC: attacked classifier. In bold the best performing attack per classifier and dataset.

Dataset	P-Places365			CIFAR-10			ImageNet		
AC / Attack	R50	R18	AN	R50	R18	AN	R50	R18	AN
Clean	5.02	5.02	5.02	4.91	4.91	4.91	5.23	5.23	5.23
BIM	4.88	4.88	4.85	4.90	4.90	4.92	4.88	4.89	4.87
TI-BIM	4.92	4.92	4.86	4.92	4.92	4.95	4.83	4.83	4.77
DF	4.95	4.94	4.94	4.88	4.88	4.92	4.93	4.93	4.92
SF	4.99	4.99	4.97	4.86	4.86	4.87	4.97	4.96	4.94
SA	5.05	5.05	5.06	5.01	5.01	**4.98**	4.80	4.79	4.80
CF-r	**5.24**	**5.22**	**5.20**	**5.05**	**5.04**	4.96	**5.24**	**5.25**	**5.23**
CF	5.22	**5.22**	5.19	5.04	5.03	4.96	**5.24**	5.24	**5.23**

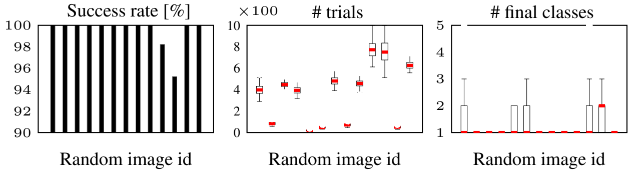


Figure 6. Influence of the randomness in the generation of ColorFool adversarial images on the success rate, the number of trials to converge and the number of different final classes at convergence with thirteen random images (500 random initializations) from ImageNet when attacking ResNet50.

plot in Fig. 6), regardless of the randomness.

5. Conclusion

We proposed a novel black-box adversarial attack, ColorFool, that modifies the color of semantic regions in an image based on priors on color perception. ColorFool achieves state-of-the-art results regarding success rate in misleading seen and unseen classifiers, robustness to defenses that employ filters, adversarial training or improved training loss function, as well as being less detectable than restricted attacks, especially JPEG compression. Furthermore, ColorFool generates adversarial images with the same size as the clean images. We hope that our work will encourage studies on adversarial attacks that simultaneously consider the human visual system and the semantic information of the objects in the image, and new defenses against colorization to make DNNs robust to color changes.

As future work, we will evaluate adversarial attacks under a larger set of defenses and explore the behavior of adversarial attacks based on colorization in tasks such as object detection and semantic segmentation.

Acknowledgements

This work is supported in part by the CHIST-ERA program through the project CORSMAL, under UK EPSRC grant EP/S031715/1. Andrea Cavallaro wishes to thank the Alan Turing Institute (EP/N510129/1), which is funded by the EPSRC, for its support through the project PRIMULA.

References

[1] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Big but imperceptible adversarial perturbations via semantic manipulation. *arXiv preprint arXiv:1904.06347*, April 2019.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of the Symposium on Security and Privacy (S&P)*, San Jose, California, USA, May 2017.

[3] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Marseille, France, October 2008.

[4] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with JPEG compression. *arXiv preprint arXiv:1705.02900*, May 2017.

[5] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru, and Fabio Roli. Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In *Proceedings of the USENIX Security Symposium*, Santa Clara, California, USA, August 2019.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami Beach, Florida, USA, June 2009.

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019.

[8] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. *arXiv preprint arXiv:1608.00853*, August 2016.

[9] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, Utah, USA, June 2018.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, California, USA, May 2015.

[11] 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 (CVPR)*, Las Vegas, Nevada, USA, June 2016.

[12] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019.

[13] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, Salt Lake City, Utah, USA, June 2018.

[14] Alex Krizhevsky and Geoffrey Hinton. *Learning multiple layers of features from tiny images*. Master's thesis, University of Toronto, April 2009.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, Nevada, USA, December 2012.

[16] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations Workshop (ICLRW)*, Toulon, France, April 2017.

[17] Martha Larson, Zhuoran Liu, Simon Brugman, and Zhengyu Zhao. Pixel Privacy: Increasing image appeal while blocking automatic inference of sensitive scene information. In *Working Notes Proceedings of the MediaEval Workshop*, Sophia Antipolis, France, October 2018.

[18] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization using optimization. *ACM Transactions on Graphics (TOG)*, 23(3):689–694, August 2004.

[19] Chau Yi Li, Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, Riccardo Mazzon, and Andrea Cavallaro. Scene privacy protection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019.

[20] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Toulon, France, April 2017.

[21] Zihao Liu, Qi Liu, Tao Liu, Yanzhi Wang, and Wujie Wen. Feature distillation: DNN-oriented JPEG compression against adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019.

[22] Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. SparseFool: a few pixels make a big difference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, USA, June 2019.

[23] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, June 2016.

[24] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, USA, June 2012.

[25] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianbing Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Seoul, Korea, October 2019.

[26] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The

limitations of deep learning in adversarial settings. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, Saarbrcken, Germany, March 2016.

[27] Daniel L Ruderman, Thomas W Cronin, and Chuan-Chin Chiao. Statistics of cone responses to natural images: implications for visual coding. *Journal of the Optical Society of America (JOSA) A*, 15(8):2036–2045, August 1998.

[28] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing (TIP)*, 27(8):3998–4011, April 2018.

[29] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings of the Network and Distributed Systems Security Symposium (NDSS)*, San Diego, California, USA, February 2018.

[30] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, October 2016.

[31] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(6):1452–1464, June 2018.

[32] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, USA, July 2017.

[33] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20k dataset. *International Journal of Computer Vision (IJCV)*, 127(3):302–321, March 2019.