



GLoFool: global enhancements and local perturbations to craft adversarial images

Mirko Agarla^{1,2}  and Andrea Cavallaro^{2,3} 

¹ University of Milano - Bicocca, Italy

m.agarla@campus.unimib.it

² Idiap Research Institute, Martigny, Switzerland

a.cavallaro@idiap.ch

³ École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. Adversarial examples crafted in black-box scenarios are affected by unrealistic colors or spatial artifacts. To prevent these shortcomings, we propose a novel strategy that generates adversarial images with low detectability and high transferability. The proposed black-box strategy, GLoFool, introduces global and local perturbations iteratively. First, a combination of image enhancement filters is applied globally to the clean image. Then, local color perturbations are generated on segmented image regions. These local perturbations are dynamically increased for each region over the iterations by sampling new colors on an expanding disc around the initial global enhancement. We propose a version of the method optimized for quality, GLoFool-Q, and one for transferability, GLoFool-T. Compared to state-of-the-art attacks that perturb colors, GLoFool-Q generates adversarial images with better color fidelity and perceptual quality. GLoFool-T outperforms all the black-box methods in terms of success rate and robustness, with a performance comparable to the best white-box methods.

Keywords: Adversarial images · Black-box attack · Color perturbation

1 Introduction

Adversarial attacks perturb the intensity of image pixels to mislead classifiers. Adversarial images are crafted with imperceptible texture perturbations [2, 17, 33], color perturbations or content manipulations [3, 23]. However, to test the vulnerability of state-of-the-art Deep Neural Networks (DNNs) [23, 30], the perturbations should become stronger thus causing visible artifacts (see Fig. 1). Significant amount of perturbations, which deceive unseen classifiers and are less detectable by defenses, include repeated textural adversarial patterns [5, 16, 31] and heavy modifications of image colors that may result in a shift towards monochromatic colors [30, 32, 34]. Therefore the applicability of these adversarial attacks is restricted to scenarios where the perturbation visibility and the chromatic variety of filters are not critical factors [23]. Methods that introduce imperceptible or sparse textural noise require access to model parameters [15, 33]. Methods



Fig. 1: Adversarial images generated by state-of-the-art methods that perturb colors to fool a ResNet-18 classifier: Natural Color Fool (NCF [30]), Adversarial Color Enhancement (ACE [32]), Adversarial Color Filter (ACF [34]), ColorFool (CF [23]), Semantic Adversarial Examples (SAE [8]), and our method optimized for quality and transferability, GLoFool-Q and GLoFool-T, respectively.

that alter the colors of specific semantic categories [23] lack transferability [30] or may be easily detectable by defense methods [13]. To address these limitations about quality, transferability, and robustness, we propose GLoFool⁴, a black-box attack method that generates visually appealing adversarial images without access to the internal parameters of the targeted classifier. GLoFool achieves color naturalness by sampling modified pixel values, starting from a global enhancement of the clean image and then dynamically expanding the search space to increase the adversarial region perturbations. We evaluate the effectiveness of GLoFool based on image quality [20, 21] against the most robust classifiers in terms of success rate (SR), transferability, and robustness. The source code of GLoFool is publicly available at <https://github.com/idiap/GLoFool>.

2 Related works

We discuss adversarial attacks that introduce color perturbations. Tab. 1 summarizes these attacks based on the type, location of the perturbation, and attacked classifier(s)⁵.

Black-box attacks use only the label of the targeted classifier to generate an adversarial image, and the attacker has no access to the internal parameters of the targeted classifier. Semantic Adversarial Examples (SAE) [8] converts the RGB image into the HSV color space and randomly shifts the hue, thus maintaining a natural appearance while introducing significant perturbations. SAE exploits the fact that hue variations can drastically affect the classifier decisions without significantly altering the image quality as perceived by humans. ColorFool [23] leverages the characteristics of the human visual system to alter colors selectively. This attack introduces perturbations within a predefined natural color range for particular semantic categories (i.e. humans, plants, sky, and water) and alters only the a and b channels of the perceptually uniform Lab color space [22]. Wei et al. [27] generate adversarial examples by manipulating brightness, contrast, sharpness, and chroma. Black-box attacks often generate

⁴ This research was conducted as part of an internship program at Idiap.

⁵ Note that we categorize attacks as white-box if they exploit the model parameters, even if the authors present some of them as black-box, e.g. NCF [30].

Table 1: Adversarial attacks based on colors perturbations: Natural Color Fool (NCF [30]), Adversarial Color Enhancement (ACE [32]), Adversarial Color Filter (ACF [34]), RetouchUAA (RUAA [28]), Semantic Adversarial Examples (SAE [8]) and ColorFool (CF [23]). Key – WB: White-Box attack, BB: Black-Box attack, DN121: DenseNet121, DN201: DenseNet201, R18: ResNet-18, R50: ResNet-50, R152: ResNet-152, V16: VGG-16, V19: VGG-19, IV3: Inception-v3, IV4: Inception-v4, MN2: MobileNetV2, AN: AlexNet, CXL: ConvNeXt-L, and DTB: DeiT-B.

Ref.	Method	Type	Pert. location	Attacked Classifier(s)
[15]	NCF	WB	Global	R18, V19, MN2, IV4
[32]	ACE	WB	Global	IV3, AN, R50, V19, DN121
[34]	ACF	WB	Global	IV3, AN, R50, V19, DN121
[28]	RUAA	WB	Global	IV3, DN121, MN3
[26]	AdvST	WB	Global	R50, AN, DN201, V19
[8]	SAE	BB	Global	V16
[27]	Wei et.al	BB	Global	VGG-16, AN R50, IV3
[23]	CF	BB	Sensitive regions	R50, R18, AN
ours	GLOFool-Q	BB	Global+Regions	R18, IV3, CXL, DTB
	GLOFool-T			

adversarial images by relying on random color changes, leading to a high number of classifier queries. Images crafted in the black-box scenario can result in significant alterations and therefore higher detectability.

White-box attacks use the full knowledge of the targeted classifier (e.g. model parameters and gradients) to target the classifier’s parameters (e.g. the last fully connected layer [30] or an intermediate layer [31]). Most methods like Adversarial Color Enhancement (ACE) [32], Adversarial Color Filter (ACF) [34] and translation-invariant method [5] that target the final fully connected layer exploit the Carlini & Wagner loss [2]. ACE [32] uses color filters optimized through a differentiable approximation [9] via gradient descent to manipulate the image with minimal impact on image quality. Natural Color Fool (NCF) [30] applies color mapping from the color distributions of ADE20K dataset [35] and gradients to generate a set of adversarial variants. ACF [34] extends ACE to produce adversarial images through an explicitly defined color filter space commonly used in photo retouching procedures. RetouchUAA (RUAA) [28] mimics the human retouching style to generate adversarial images. The retouching module is optimized through gradient back-propagation and then constrained by a style guidance module performed using a U-Net network. AdvST [26] is an unrestricted attack that exploits the model’s gradients during the style transfer process from a reference image onto the original image. The attacks above leverage the access to model parameters and perturb the image using filters or color manipulation. These models generate adversarial images with minimal impact on image quality while maximizing the probability of misleading the target classifier [15]. Calculating gradients and optimising perturbations can be computationally expensive, especially for large models. These attacks are often tailored to a specific model, and may not be effective on others architectures or with different training data.

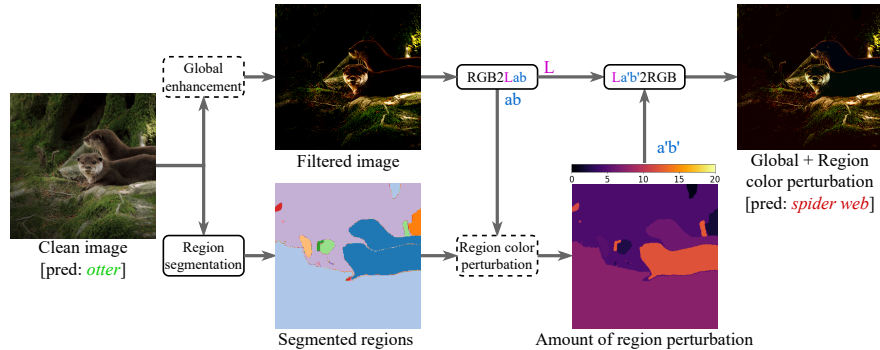


Fig. 2: Main steps of the GLoFool process for adversarial image generation. GLoFool modifies the clean image with a global enhancement using a random combination of enhancement filters (Sec. 3.1), and then it applies a different color perturbation on each region (Sec. 3.2). The global enhancement and the region color perturbation are repeated for up to N iterations, increasing the size of the region perturbation disc until the generated image successfully fools the classifier. Pred: prediction of ResNet-18 classifier.

3 GLoFool

Let $M(X)$ be a DNN classifier that provides the most probable class label for a given RGB image, X . The proposed approach, GLoFool, iteratively perturbs X to create an adversarial image, \dot{X} , until $M(\dot{X}) \neq M(X)$. Fig. 2 shows the main processing steps of GLoFool, which are detailed in the following sections.

3.1 Global enhancement

The global color enhancement filters the clean image to obtain an adversarial image that is visually appealing and more effective against machine learning classifiers. For each iteration, the global color enhancement alters X with a function, $G(\cdot)$, that applies a sequence of image enhancement filters $f_i(\cdot)$, with $i = 1, \dots, e$ to obtain a globally enhanced image E . We set the number of filters to $e = 6$. These filters are contrast, saturation, curve (shadows and mid-tones), sharpening, vibrance, and sepia. Contrast makes images more vivid and enhances distinctions between elements [18]. Saturation intensifies colors for a more appealing image [19]. Curve adjustments control brightness and contrast, preserving the overall color balance. Sharpening makes the image clearer and more defined [10]. Vibrance focuses on less saturated regions, while maintaining the existing balance within the image. Finally, the sepia filter adds warm tones [1]. We present experiments about the number of global enhancement filters in Appendix A.

In each iteration, a globally enhanced image is obtained by applying, in random order, all e filters to the clean image X . This strategy enables the exploration of wider aesthetic enhancements for the final image. Applying the filters

in a non-deterministic sequence generates diverse enhancements for the clean image. Each filter is applied once to the image resulting from the application of the previous filter:

$$G(X, \tau) = (\bigcirc_{i=1}^e f_i(\tau_i))(X) \quad (1)$$

where \bigcirc denotes the sequential composition of the filters and each τ_i is independently sampled from $[-\tau, +\tau]$ that represents the intensity value of each filter. The curve filtering involves adjusting a Bezier curve using four control points: p0, p1, p2, and p3. To maintain the full dynamic range of the image, the points p0 and p3 correspond to input values 0 and 255, respectively. Point p1 is chosen from the shadow range of [75, 150] and is mapped to a random value between [50, 125], ensuring that shadows can be either deepened or slightly brightened, enhancing contrast and detail without over-darkening. Point p2 is selected from the mid-tone range of [150, 225] and mapped to a random value between [175, 250], ensuring that mid-tones can be either brightened or slightly dimmed, improving overall image brightness and mid-tone separation. These points are used to calculate the Bezier curve, and the remaining values are interpolated to create a smooth adjustment curve.

The resulting globally enhanced image, E , is the initial point for the subsequent region perturbation.

3.2 Region color perturbation

We partition the input image into regions to apply perturbations with varying intensities to individual regions of the globally enhanced image.

We operate on the *Lab* color space, a perceptually uniform color space that mimics human vision and separates the color information from brightness, enabling color adjustments without affecting overall brightness. We convert X and E from the *RGB* to the *Lab* color space obtaining X_{Lab} and E_{Lab} , respectively. In the following notations, we selectively access the color components of both variables using the notations X_{ab} and E_{ab} , respectively.

Region segmentation partitions X into K regions, o_1, o_2, \dots, o_K . The union of these regions equals the original image and the intersection of any two regions is the empty set:

$$S(X, s) = \{o_i\}_{i=1}^K, \quad \text{where } X = \bigcup_{i=1}^K o_i \text{ and } o_i \cap o_j = \emptyset \text{ for all } i \neq j. \quad (2)$$

where $j \in \{1, \dots, K\}$. We select as segmentation algorithm $S(\cdot)$ that automatically determines the optimal K based on a stability score threshold s [11]. This threshold controls the number of segmented regions within the image, where a value close to 1 returns only high-confidence regions. A lower threshold will return more regions as the confidence decreases. We set the stability score threshold to 0.97 to avoid obtaining many small regions and only retain regions exceeding this threshold.

We alter only the color components, a and b , of each region E_{o_i} while maintaining the lightness L unaltered, equal to that of the respective region of the

enhanced image E_{o_i} . We perturb the color components of E with the function $\phi(\cdot)$ that shifts the ab color components of each region i , E_{o_i} , by a randomly sampled offset within a disc constrained by d :

$$\phi(E_{ab}, d) = \bigcup_{i=1}^K \left(E_{o_i} + r_i \cdot \begin{bmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{bmatrix} \right), \quad (3)$$

where $\theta_i \in [0, 2\pi]$ is the randomly sampled angle and $r_i \in [0, d]$ is the shift of the ab color components, randomly sampled in a disc constrained by the higher bound radius, d .

By gradually increasing d during each iteration, the method expands the search space for new colors. At each iteration, n , the search space d is increased by an increment step r :

$$\begin{aligned} r &= \frac{c}{N} \\ d &= \begin{cases} r, & \text{if } n = 0, \\ r \cdot n, & \text{otherwise.} \end{cases} \end{aligned} \quad (4)$$

where N is the maximum number of iterations, and c is the upper bound for the maximum color perturbation. c represents the maximum shift for the ab color components, achieved at the last iteration following the definition of Eq. (4). This process allows for exploring different perturbations, resulting in regions with different degrees of color deviation compared to the globally enhanced image.

At each iteration, the region color perturbation modifies (see Eq. (3)) the segmented regions of the globally enhanced image (see Eq. (1)) to generate a perturbed image:

$$\dot{X}^{(n)} = \left[E_L, \lambda \left(\phi(E_{ab}^{(n)}, d^{(n)}) \right) \right] \quad (5)$$

where the region perturbation, $\phi(E_{ab}, d)$, is crafted on the globally enhanced image, $E_{Lab} = G(X, \tau)$. We apply the clipping function $\lambda(\cdot)$ to limit the final perturbed image to the range allowed by the color space. The clipping function limits the pixel values of $\phi(\cdot)$ to the interval of ab channels. Note that in this transformation, the lightness component of the perturbed image is preserved from the enhanced image and concatenated with the perturbed ab values. At the end of this stage we obtain an image \dot{X} , converted into the RGB color space, wherein the colors of each region are perturbed within specific perturbation bounds.

Fig. 3 (c) shows from the first row, the images generated using the global enhancement (a) plus the region color perturbation (e) denoted by r_i (Eq. (3)). To produce an adversarial image, we apply a global color enhancement to the clean image for each iteration, Fig. 3 (a). Then, to subsequently reduce the classifier accuracy, a region color perturbation is added to the globally enhanced image, Fig. 3 (c). These steps are repeated, incrementing the region color perturbation through the iterations, until the resulting image \dot{X} fools the classifier $M(\cdot)$, or the maximum number of iterations $N = 1500$ is reached. The gradual

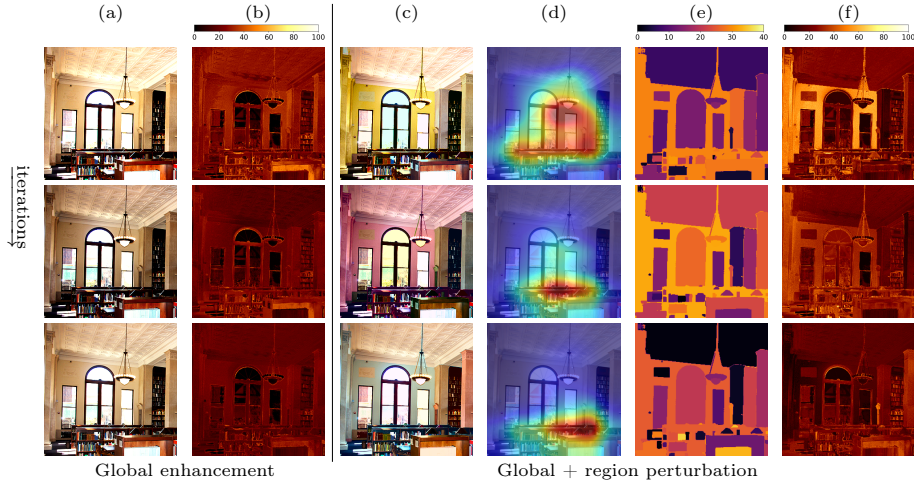


Fig. 3: Global and global+local perturbations by GLoFool attacking a ResNet-18 classifier. The clean image (label: *library*) is shown in Fig. 4 (first row, first column). (a) Global enhancement. (b) ΔE_{76} [21] perceptual color difference between the globally enhanced images and the clean image. (c) Region perturbations lead to the following misclassified labels: *church* (first row), *cinema* (second row), and *pool table* (third row). (d) Activation map [7] of (c). (e) Color shift offset (Eq. (3)). (f): ΔE_{76} [21] perceptual color difference between the globally+locally perturbed images and the clean image.

increase in the region color perturbation across iterations systematically changes the classifier’s activation map [7], Fig. 3 (d), which leads to misclassification and an increase in ΔE_{76} , Fig. 3 (f), in regions with high perturbations. Experiments of GLoFool-T with a limited number of iterations are presented in Appendix B.

We propose two variations of the proposed method, namely GLoFool-T and GLoFool-Q, to meet different requirements regarding transferability and quality, respectively. GLoFool-T (Transferability) generates adversarial images with improved transferability and defensive capabilities (robustness). This improved version is configured with $\tau = 45$, $c = 128$, and amplifies the image generation process by utilising the third generated adversarial image. GLoFool-Q (Quality) is configured with $\tau = 30$ and $c = 64$ to improve the quality score and perceptual color difference compared to GLoFool-T.

4 Evaluation

We compare GLoFool against the most competitive state-of-the-art attacks that offer accessible source code: Natural Color Fool [30], Adversarial Color Enhancement (ACE) [32] and Adversarial Color Filter (ACF) [34] as white-box attacks; and ColorFool [23] and Semantic Adversarial Examples (SAE) [8] as black-box attacks. We use the authors’ implementations for all adversarial attacks. To ensure the comparability of results, we generate adversarial images using the same

framework and software versions as PyTorch and OpenCV. We follow the same evaluation protocol as Natural Color Fool [30], evaluating the adversarial attacks on the ImageNet-compatible Dataset composed of 1,000 images [12].

Classifiers. To assess the behavior of the methods on different perturbation requirements, we evaluate them on two common classifiers, ResNet-18 and Inception-v3, and two robust classifiers, ConvNeXt-L [14] and the Vision Transformer DeiT-B [25], pre-trained on ImageNet [4] for a classification task involving 1,000 classes. The accuracy (top-1) of ResNet-18, Inception-v3, ConvNeXt-L, and DeiT-B on clean images is 83.9%, 81.0%, 96.4%, and 94.1%, respectively.

Quality measures. We evaluate the quality of the generated images with the Haar-based Perceptual Similarity Index (HaarPSI) [20], a full reference metric with a high correlation with human opinion scores. The larger HaarPSI, the higher the perceptual similarity between the adversarial image and its respective clean image. We measure the perceptual color difference between the generated and clean images using ΔE_{76} [21]. Fidelity to the clean image is assessed based on these metrics, where high HaarPSI and low ΔE_{76} values indicate high fidelity.

Effectiveness of an attack. We compare the methods in terms of success rate, transferability, and robustness to defenses. We quantify the ability to fool a classifier as *success rate*, defined as the number of successful adversarial images that fool the classifier divided by the number of dataset images, D :

$$SR = \frac{\sum_{i=1}^D \mathbb{I}[M(\dot{X}_i) \neq M(X_i)]}{D}, \quad (6)$$

where \mathbb{I} is the indicator function that outputs 1 if the condition is true and 0 otherwise. To quantify *transferability*, we evaluate the ability to fool unseen classifiers (i.e. the test classifier differs from the one used to craft the adversarial images). We quantify the *robustness* to defenses as the ratio of adversarial images that fool the classifier into predicting a different class than the clean images, to the total number of images after applying the defense filter. We consider several defense filtering techniques, including re-quantization [29] to 32 and 8 colors, median filtering with a square kernel of size 5 [29], and JPEG compression with quality degradation [6] with quality parameters 75 and 50. Regarding neural-based defense approaches, we adopt High-level representation Guided Denoiser (HGD) [13] and adversarially trained ConvNext-S-CvSt (CNS) [24].

Success Rate and Transferability. Tab. 2 compares the success rate of misleading seen and unseen classifiers. All the adversarial methods achieve high SR with seen classifiers (gray cells). In particular, methods that significantly decrease the color fidelity to the clean image (high ΔE_{76}), such as ACE and ACF, achieve high SR while, reducing also the perceived image quality (low HaarPSI). Fig. 4 shows sample images generated by state-of-the-art methods and our proposed attack. The methods that obtain high SR in transferability introduce noticeable color artifacts. Thanks to the incremental region perturbation applied over the global enhancement, GLoFool produces images with the minimum perturbation required to fool the classifier. All state-of-the-art methods find it challenging to generate adversarial images with high color fidelity when

Table 2: Success rate (SR) of adversarial attacks against ResNet-18 (R18), Inception-v3 (IV3), ConvNeXt-L (CXL) and DeiT-B (DTB). Key – NCF: Natural Color Fool, ACE: Adversarial Color Enhancement, ACF: Adversarial Color Filter, CF: ColorFool, SAE: Semantic Adversarial Examples, AC: Attacked classifier, BB: Black box, WB: White box. The best and second-best results are marked in **boldface** and underlined, respectively. Seen classifiers are highlighted in gray.

AC	Type	Method	HaarPSI \uparrow	$\Delta E_{76} \downarrow$	SR Test Classifier \uparrow			
					R18	IV3	CXL	DTB
R18	WB	NCF [30]	0.44	34.8	<u>92.7</u>	44.1	18.3	32.7
		ACE [32]	0.49	38.2	<u>99.5</u>	31.6	5.6	13.0
		ACF [34]	0.46	43.2	<u>97.5</u>	34.9	7.2	15.2
	BB	CF [23]	<u>0.59</u>	27.3	93.0	<u>16.1</u>	<u>3.3</u>	6.8
		SAE [8]	0.64	35.2	92.3	24.7	<u>11.1</u>	<u>22.3</u>
		GLoFool-Q	0.56	22.4	93.6	31.9	5.0	11.3
	GLoFool-T	0.52	<u>25.1</u>	99.6	<u>36.3</u>	7.7	14.9	
IV3	WB	NCF [30]	0.47	33.4	57.4	84.0	13.1	25.2
		ACE [32]	0.51	34.2	40.3	<u>96.6</u>	6.5	12.7
		ACF [34]	0.49	35.9	43.2	<u>92.9</u>	6.3	14.2
	BB	CF [23]	<u>0.58</u>	28.9	31.3	82.3	<u>4.3</u>	9.2
		SAE [8]	0.64	36.7	44.7	74.9	<u>12.3</u>	<u>23.3</u>
		GLoFool-Q	0.57	21.5	37.0	88.1	5.1	12.1
	GLoFool-T	0.51	<u>26.4</u>	<u>50.1</u>	98.2	7.9	19.5	
CXL	WB	NCF [30]	0.46	34.6	<u>58.3</u>	39.4	56.8	<u>36.4</u>
		ACE [32]	0.48	38.2	49.1	38.2	97.2	23.1
		ACF [34]	<u>0.50</u>	36.8	40.2	30.4	75.8	19.8
	BB	CF [23]	0.45	43.9	52.2	<u>32.1</u>	50.7	25.9
		SAE [8]	0.59	40.6	46.7	31.9	48.7	31.5
		GLoFool-Q	<u>0.50</u>	24.7	42.7	<u>39.4</u>	64.2	24.7
	GLoFool-T	0.41	<u>34.3</u>	65.7	56.0	<u>90.6</u>	48.3	
DTB	WB	NCF [30]	0.45	34.9	<u>60.1</u>	42.6	23.6	72.5
		ACE [32]	0.48	38.6	47.5	35.6	11.6	98.1
		ACF [34]	0.49	38.1	46.9	35.0	10.6	84.6
	BB	CF [23]	0.47	41.4	47.1	<u>31.2</u>	8.5	75.8
		SAE [8]	0.59	41.1	51.5	29.4	14.7	73.3
		GLoFool-Q	<u>0.52</u>	23.7	47.3	<u>42.9</u>	9.7	79.6
	GLoFool-T	0.45	<u>30.7</u>	60.7	52.9	<u>15.7</u>	<u>98.0</u>	

attacking robust networks such as ConvNeXt-L and DeiT-B. Herein, GLoFool-T generates images with low ΔE_{76} , obtaining the highest transferability SR, surpassing the second-best method, NCF, by 13%. In small(er) networks like ResNet-18 and Inception-v3, SAE obtains a high value of HaarPSI and a very low value of ΔE_{76} , maintaining a high level of transferability in unseen classifiers at the cost of achieving the worst SR in all the seen classifiers compared to the other methods. GLoFool-Q generates images with high quality and high color fidelity, sacrificing the transferability SR in small(er) networks. GLoFool-T improves the transferability of GLoFool-Q in terms of SR in seen and unseen classifiers, while slightly increasing the perceptual color difference compared to clean images.

Fig. 5 compares the likelihood \mathcal{L} of the clean images with the SR and the ΔE_{76} of the respective generated adversarial images against the Vision Transformer DeiT-B [25]. Methods like NCF, ACF, SAE and ACE introduce too much perturbation when the required perturbation to fool the classifier is low, i.e. low clean likelihood. While the algorithms based on incremental perturbation, such

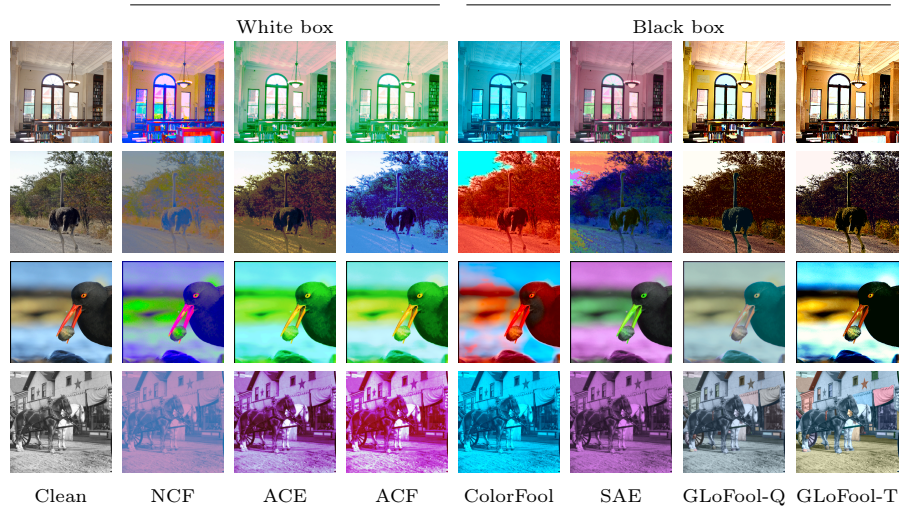


Fig. 4: Sample adversarial images generated for the ResNet-18 classifier. From left to right: clean image, images generated by Natural Color Fool (NCF), Adversarial Color Enhancement (ACE), Adversarial Color Filter (ACF), ColorFool, Semantic Adversarial Examples (SAE), and our methods, GLoFool-Q and GLoFool-T.

as ColorFool, GLoFool-Q, and GLoFool-T generate high-quality images with colors more faithful to the clean image even when no high perturbation is required. However, the quality of images generated by ColorFool is significantly reduced when the images are subject to high perturbations.

Robustness to defenses. Tab. 3 shows the robustness results of state-of-the-art methods. NCF and GLoFool-T are the most robust methods against defenses. The defense performance decreases when attacking a robust classifier like ConvNeXt-L. For example, considering the median filtering (MF5) defense, NCF and ACF drop from a SR of 92.1 and 82.8 for the ResNet-18 to a SR of 75.5 and 55.8 for the ConvNeXt-L, respectively. GLoFool-T obtains a more constant defense SR of 70.2 and 72.7 for the ResNet-18 and ConvNeXt-L. The HGD defense method successfully reduces the most adversarial perturbations within the generated images. In this setup, after attacking ConvNeXt-L and DeiT-B, GLoFool-T improves the defense SR of SAE by 3%. Methods that incrementally introduce perturbation to generate a natural-looking adversarial image, such as GLoFool-Q and CF, demonstrate reduced robustness against defenses.

5 Analysis

In this section we analyze the effect of the global enhancement threshold τ (Eq. (1)), color perturbation threshold c (Eq. (4)), the stability score threshold s (Eq. (2)), and the impact of the main components of the proposed method.

The impact of τ is shown in the first row in Fig. 6. We start from $\tau = 10$ and

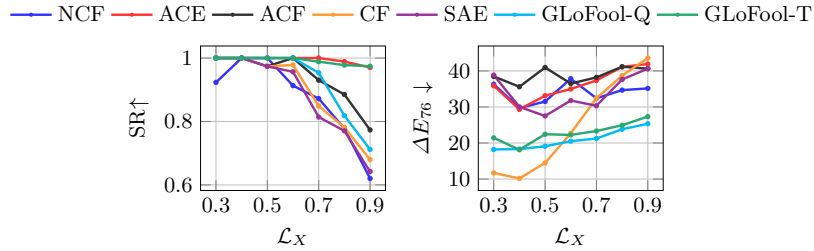


Fig. 5: Success rate (SR) and ΔE_{76} with respect to the likelihood \mathcal{L} of the clean images for state-of-the-art attacks against the Vision Transformer DeiT-B [25].

Table 3: Robustness of adversarial attacks to defenses: quantization to 32 and 8 colors (Q32, Q8), median filtering with a kernel size 5 (MF5), lossy JPEG compression with quality 75 and 50 (J75, J50), High-level representation Guided Denoiser (HGD), and adversarially trained ConvNext-S-CvSt (CNS). Key – R18: ResNet-18, IV3: Inception-v3, CXL: ConvNeXt-L, DTB: DeiT-B, NCF: Natural Color Fool, ACE: Adversarial Color Enhancement, ACF: Adversarial Color Filter, CF: ColorFool, SAE: Semantic Adversarial Examples, AC: Attacked classifier, WB: White box, BB: Black box. The best and second-best results are marked in **boldface** and underline, respectively.

AC	Type	Method	J75	J50	MF5	Q32	Q8	HGD	CNS
R18	WB	NCF [30]	93.0	91.9	92.1	98.5	93.3	29.6	53.6
		ACE [32]	72.0	68.3	79.4	86.9	69.1	8.4	28.2
		ACF [34]	<u>78.8</u>	<u>74.2</u>	<u>82.8</u>	<u>91.8</u>	77.0	<u>13.1</u>	33.0
	BB	CF [23]	55.1	55.5	60.0	69.1	61.3	1.8	22.7
		SAE [8]	70.2	66.6	77.1	84.0	69.7	7.5	29.0
		GLoFool-Q	60.9	55.3	60.0	70.5	60.0	5.5	28.0
		GLoFool-T	73.8	70.5	70.2	84.5	<u>80.7</u>	6.5	<u>37.4</u>
IV3	WB	NCF [30]	80.9	76.4	76.1	95.5	77.0	28.3	44.9
		ACE [32]	56.3	49.9	62.2	79.4	56.4	9.5	23.1
		ACF [34]	67.6	61.7	70.7	<u>88.5</u>	62.9	<u>14.5</u>	28.1
	BB	CF [23]	<u>47.8</u>	<u>45.9</u>	<u>50.8</u>	<u>65.0</u>	51.4	6.1	29.6
		SAE [8]	58.1	57.3	61.5	73.2	57.0	13.8	35.1
		GLoFool-Q	60.9	55.3	60.0	70.5	60.0	5.5	28.0
		GLoFool-T	<u>72.3</u>	<u>70.2</u>	<u>71.3</u>	78.1	<u>71.6</u>	10.7	<u>42.5</u>
CXL	WB	NCF [30]	65.0	64.0	75.5	88.8	73.6	40.2	59.8
		ACE [32]	37.4	39.8	53.7	49.6	30.5	17.1	34.6
		ACF [34]	43.0	43.7	55.8	64.9	44.2	21.5	36.9
	BB	CF [23]	<u>52.1</u>	<u>52.1</u>	<u>64.5</u>	<u>65.9</u>	<u>59.2</u>	<u>23.5</u>	45.0
		SAE [8]	<u>59.5</u>	<u>61.4</u>	71.3	<u>74.5</u>	<u>62.8</u>	28.7	54.0
		GLoFool-Q	41.1	43.2	60.8	51.3	42.5	22.3	49.3
		GLoFool-T	50.6	56.7	<u>72.7</u>	59.7	58.5	<u>37.2</u>	70.3
DTB	WB	NCF [30]	85.3	80.6	83.2	96.3	81.7	37.6	59.8
		ACE [32]	52.7	47.6	56.8	70.4	40.5	14.0	31.8
		ACF [34]	64.8	58.6	66.0	84.0	51.5	18.8	38.2
	BB	CF [23]	60.6	59.2	69.7	<u>75.7</u>	<u>52.2</u>	<u>12.0</u>	53.0
		SAE [8]	67.1	65.1	<u>78.3</u>	<u>82.0</u>	61.1	15.7	45.7
		GLoFool-Q	60.2	58.9	62.3	72.8	53.5	11.4	40.6
		GLoFool-T	<u>69.9</u>	<u>68.9</u>	73.9	74.4	<u>70.6</u>	<u>20.6</u>	<u>57.9</u>

increment by 5 until $\tau = 50$. Lower values of τ lead to better image quality at the cost of a lower SR. High τ values significantly increase the robustness and

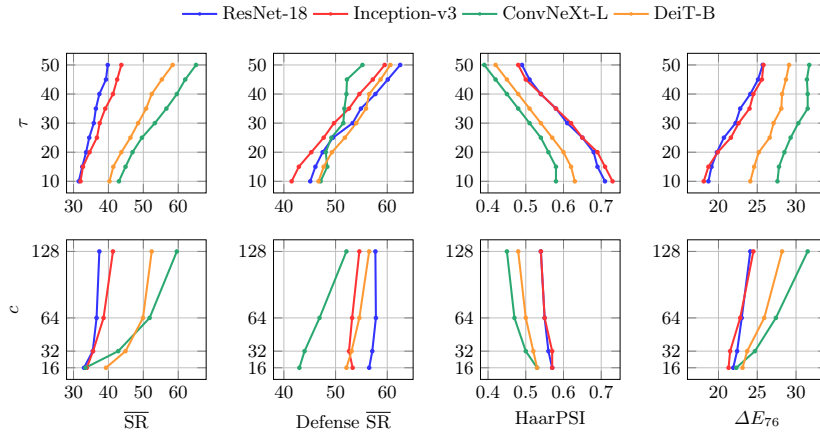


Fig. 6: First row: performance of GLoFool with different global enhancement thresholds τ , when the maximum color perturbation threshold c is set to 128. Second row: performance of GLoFool with different color perturbation thresholds, c , when the global enhancement threshold τ is set to 40. $\overline{\text{SR}}$ is the average SR across the seen and unseen classifiers. The defense $\overline{\text{SR}}$ is the average defense SR across all defenses.

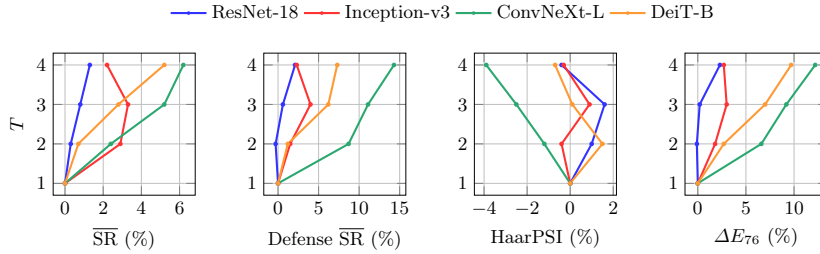


Fig. 7: Performance of GLoFool-T with different transfer-optimized values, T , for the generated adversarial images. E.g., $T = 3$ means that the 3^{rd} generated adversarial image is proposed as an adversarial image. GLoFool-T is configured with $\tau = 45$ and $c = 128$. The SR results are averaged across all the seen and unseen classifiers. The defense SR results are averaged across all the defenses.

the SR of the generated images. Higher τ values are measured as distortions by HaarPSI that is sensitive to low frequencies, whereas they have a lower impact on ΔE_{76} . The impact of $c \in \{16, 32, 64, 128\}$ is shown in Fig. 6, second row. The higher c , the farther away the perturbed ab color components are from the original colors of the clean image. This deviation results in a higher ΔE_{76} and HaarPSI, and an increment of the average SR, especially for the large models such as ConvNeXt-L and DeiT-B.

Motivated by the analysis of Fig. 6, we designed GLoFool-T, a method that optimizes τ and c to improve the transferability and defense capabilities without significantly compromising image quality and color fidelity. In Fig. 7 we show the

Table 4: Performance of GLoFool-T with different stability score threshold s after attacking, ResNet-18 (R18), Inception-v3 (IV3), ConvNeXt-L (CXL) and DeiT-B (DTB). Key – AC: Attacked Classifier, SR: success rate, $\overline{\text{SR}}$ def: average success rate over all the defensive algorithms. Seen classifiers are highlighted in gray.

AC	s	HaarPSI \uparrow	$\Delta E_{76}\downarrow$	SR Test Classifier \uparrow				$\overline{\text{SR}}$ def \uparrow
				R18	IV3	CXL	DTB	
R18	99	0.514	25.0	99.6	36.7	7.0	15.2	61.0
	98.5	0.520	25.2	99.5	36.6	8.3	15.4	59.5
	98	0.516	25.2	99.7	35.8	6.8	16.3	61.3
	97.5	0.518	25.3	99.6	35.8	8.4	15.1	60.1
	97	0.520	25.2	99.6	36.3	7.7	14.9	60.5
IV3	99	0.499	26.5	50.9	98.2	9.3	19.3	59.3
	98.5	0.503	26.4	49.5	98.4	9.3	20.7	58.7
	98	0.510	26.1	49.7	98.3	8.7	16.0	58.7
	97.5	0.511	25.9	46.0	98.2	7.4	19.1	57.7
	97	0.509	26.4	50.1	98.2	7.9	19.5	59.5
CXL	99	0.401	35.0	68.0	57.3	91.2	46.9	58.2
	98.5	0.407	34.6	65.4	55.9	91.0	44.6	57.2
	98	0.412	34.1	66.9	54.7	91.8	45.7	57.2
	97.5	0.407	34.7	66.1	55.4	91.3	45.3	57.1
	97	0.409	34.3	65.7	56.0	90.6	48.3	58.0
DTB	99	0.448	30.3	60.9	54.9	15.5	98.1	62.6
	98.5	0.446	30.7	62.9	52.9	16.2	98.0	62.1
	98	0.449	30.5	61.4	50.9	14.3	97.5	62.1
	97.5	0.446	30.6	61.9	53.0	15.8	97.9	62.4
	97	0.449	30.7	60.7	52.9	15.7	98.0	62.3

performance of GLoFool-T using $\tau = 45$ and $c = 128$ using the n^{th} generated adversarial image as the final proposal. When progressively using subsequent adversarial images as proposals, i.e. increasing the T threshold, our method improves the SR, especially in the defensive setup. A high value of T also reduces the quality of generated images for ConvNeXt-L, whereas for the other models, the quality remains unaffected. We selected the third adversarial image because it provides the best balance between improvements in SR, defensive SR, and image quality.

Tab. 4 shows the impact of $s \in \{99, 98.5, 98, 97.5, 97\}$ on GLoFool-T. The higher s , the fewer segmented regions per image are detected. On average, the number of regions per image is $\{18, 26, 35, 44, 52\}$ for each threshold value, respectively. Increasing the number of segmented regions does not lead to proportional improvements in performance. We use $s = 97$ because, on average, it slightly improves the success rate on both seen and unseen classifiers and leads to better image quality.

In Tab. 5 we present the ablation study for the global and/or region perturbations and the transfer-optimized configuration. The configuration for enhancing the transferability, GLoFool-T, combines global and local perturbations and selects the third generated adversarial image. Compared to the configuration with only the region perturbation, exploiting both the global and region perturbations decreases by 40% the number of required iterations to fool the classifier and significantly increases the transferability and robustness. The configuration with both global and region perturbation increases the SR and slightly reduces

Table 5: Ablation study for the main components (global enhancement, region perturbation, transfer-optimized) of the proposed method, configured with $\tau = 45$ and $c = 128$. Key – R18: ResNet-18, IV3: Inception-v3, CXL: ConvNeXt-L, DTB: DeiT-B, AC: Attacked Classifier, T: transfer-optimized configuration, Iter: average iterations for the successful attacks, SR: success rate, $\overline{\text{SR}}$ def: average success rate over all the defenses. The best and second-best results are marked in **boldface** and underlined, respectively. Seen classifiers are highlighted in gray.

AC	Global	Region	T	$\overline{\text{Iter}}\downarrow$	HaarPSI \uparrow	$\Delta E_{76}\downarrow$	SR Test Classifier \uparrow					$\overline{\text{SR}}$ def \uparrow
							R18	IV3	CXL	DTB	Seen	
R18	✓	-	-	56	0.60	20.3	89.5	27.6	4.3	10.5	55.7	
	-	✓	-	358	0.70	17.2	<u>93.8</u>	18.2	4.0	8.5	40.6	
	✓	✓	-	82	0.51	25.1	99.6	36.8	6.5	<u>14.4</u>	<u>60.1</u>	
	✓	✓	✓	131	0.52	25.1	99.6	<u>36.3</u>	7.7	14.9	60.5	
IV3	✓	-	-	78	<u>0.61</u>	<u>19.2</u>	33.8	<u>80.0</u>	4.8	10.5	53.1	
	-	✓	-	346	0.72	15.9	21.8	<u>87.5</u>	2.9	10.1	36.5	
	✓	✓	-	<u>119</u>	0.50	25.6	<u>49.2</u>	98.2	<u>7.2</u>	<u>15.5</u>	<u>57.2</u>	
	✓	✓	✓	180	0.51	26.4	50.1	98.2	7.9	19.5	59.5	
CXL	✓	-	-	203	0.58	20.1	32.3	27.1	<u>37.2</u>	15.8	49.0	
	-	✓	-	632	<u>0.57</u>	<u>27.8</u>	44.8	36.2	<u>58.1</u>	32.8	48.8	
	✓	✓	-	<u>320</u>	0.42	31.4	<u>63.7</u>	<u>51.4</u>	90.6	<u>42.1</u>	<u>52.2</u>	
	✓	✓	✓	457	0.41	34.3	65.7	56.0	90.6	48.3	58.0	
DTB	✓	-	-	153	<u>0.58</u>	20.7	38.0	29.8	8.3	56.6	53.8	
	-	✓	-	508	0.63	<u>22.6</u>	34.4	28.7	6.3	<u>86.8</u>	44.7	
	✓	✓	-	<u>222</u>	0.45	28.7	<u>59.4</u>	<u>50.4</u>	13.3	98.0	58.7	
	✓	✓	✓	324	0.45	30.7	60.7	52.9	15.7	98.0	62.3	

the quality of the generated image. In this configuration, the global enhancement improves the transferability and robustness against defenses. In contrast, the region perturbation increases the SR in seen classifiers without a significant impact on the image quality. Combining the transfer-optimized configuration with the global and region perturbations significantly improves the SR in seen and unseen classifiers and the average robustness.

6 Conclusion

We presented GLoFool, a black-box method designed to generate adversarial images that are robust against unseen classifiers and exhibit a high degree of undetectability against defense methods. GLoFool perturbs the colors of individual regions within an image, starting from a global enhancement. We proposed two versions of the method to improve the transferability (GLoFool-T) or the quality of the generated images (GLoFool-Q). GLoFool-T outperforms the average SR of seen and unseen classifiers of all state-of-the-art methods by 18% and 30%, respectively. GLoFool-Q obtains a competitive SR against the black-box state-of-the-art methods, outperforming the average HaarPSI and ΔE_{76} of all state-of-the-art methods by 5% and 21%, respectively. In for future work, we will study how to strengthen classifiers by analysing the relationship between color perturbations in specific image regions.

References

1. Bakhshi, S., Shamma, D., Kennedy, L., Gilbert, E.: Why we filter our photos and how it impacts engagement. In: Proceedings of the International AAAI Conference on Web and social media. vol. 9, pp. 12–21 (2015)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57 (2017)
3. Chen, Z., Li, B., Wu, S., Jiang, K., Ding, S., Zhang, W.: Content-based unrestricted adversarial attack. arXiv preprint arXiv:2305.10665 (2023)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on Computer Vision and Pattern Recognition (2009)
5. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
6. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpeg compression on adversarial images. arXiv preprint arXiv:1608.00853 (2016)
7. Fernandez, F.G.: Torchcam: class activation explorer. <https://github.com/frgfm/torch-cam> (March 2020)
8. Hosseini, H., Poovendran, R.: Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2018)
9. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. ACM Transactions on Graphics (TOG) **37**(2), 1–17 (2018)
10. Kaur, S., Kaur, M.: Image sharpening using basic enhancement techniques. Int. J. Res. Eng. Sci. Manag. **1**(12), 122–126 (2018)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
12. Kurakin, A., Goodfellow, I., Bengio, S., Dong, Y., Liao, F., Liang, M., Pang, T., Zhu, J., Hu, X., Xie, C., et al.: Adversarial attacks and defences competition. In: The NIPS’17 Competition: Building Intelligent Systems (2018)
13. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2018)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (2022)
15. Modas, A., Moosavi-Dezfooli, S.M., Frossard, P.: Sparsefool: a few pixels make a big difference. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (2019)
16. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2016)
17. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P) (2016)
18. Peli, E.: Contrast in complex images. JOSA A **7**(10), 2032–2040 (1990)

19. Pratt, W.K.: Digital image processing: PIKS Scientific inside, vol. 4. Wiley Online Library (2007)
20. Reisenhofer, R., Bosse, S., Kutyniok, G., Wiegand, T.: A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication* **61**, 33–43 (2018)
21. Robertson, A.R.: The CIE 1976 color-difference formulae. *Color Research & Application* **2**(1), 7–11 (1977)
22. Ruderman, D.L., Cronin, T.W., Chiao, C.C.: Statistics of cone responses to natural images: implications for visual coding. *JOSA A* **15**(8), 2036–2045 (1998)
23. Shamsabadi, A.S., Sanchez-Matilla, R., Cavallaro, A.: Colorfool: Semantic adversarial colorization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
24. Singh, N.D., Croce, F., Hein, M.: Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. In: *NeurIPS* (2023)
25. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877* (2020)
26. Wang, X., Chen, H., Sun, P., Li, J., Zhang, A., Liu, W., Jiang, N.: Advst: Generating unrestricted adversarial images via style transfer. *IEEE Transactions on Multimedia* (2023)
27. Wei, X., Guo, Y., Li, B.: Black-box adversarial attacks by manipulating image attributes. *Information sciences* **550**, 285–296 (2021)
28. Xie, M., He, Y., Fang, M.: Retouchuaa: Unconstrained adversarial attack via image retouching. *arXiv preprint arXiv:2311.16478* (2023)
29. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017)
30. Yuan, S., Zhang, Q., Gao, L., Cheng, Y., Song, J.: Natural color fool: Towards boosting black-box unrestricted attacks. *Advances in Neural Information Processing Systems* **35**, 7546–7560 (2022)
31. Zhang, Q., Li, X., Chen, Y., Song, J., Gao, L., He, Y., et al.: Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. In: *International Conference on Learning Representations* (2021)
32. Zhao, Z., Liu, Z., Larson, M.: Adversarial color enhancement: Generating unrestricted adversarial images by optimizing a color filter. *arXiv preprint arXiv:2002.01008* (2020)
33. Zhao, Z., Liu, Z., Larson, M.: Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020)
34. Zhao, Z., Liu, Z., Larson, M.: Adversarial image color transformations in explicit color filter space. *IEEE Transactions on Information Forensics and Security* (2023)
35. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2017)