ALA: Naturalness-aware Adversarial Lightness Attack

Yihao Huang
Nanyang Technological University
Singapore

Liangru Sun
East China Normal University
China

Qing Guo
IHPC and CFAR, Agency for Science, Technology and Research, Singapore

Felix Juefei-Xu*
Meta AI
USA

Jiayi Zhu
East China Normal University
China

Jincao Feng
East China Normal University & Shanghai Industrial Control Safety Innovation Tech. Co., Ltd, China

Yang Liu†
Zhejiang Sci-Tech University, China
Nanyang Technological University, Singapore

Geguang Pu†
East China Normal University & Shanghai Industrial Control Safety Innovation Tech. Co., Ltd, China

ABSTRACT

Most researchers have tried to enhance the robustness of DNNs by revealing and repairing the vulnerability of DNNs with specialized *adversarial examples*. Parts of the attack examples have imperceptible perturbations restricted by L_p norm. However, due to their high-frequency property, the adversarial examples can be defended by denoising methods and are hard to realize in the physical world. To avoid the defects, some works have proposed unrestricted attacks to gain better robustness and practicality. It is disappointing that these examples usually look unnatural and can alert the guards. In this paper, we propose **A**dversarial **L**ightness **A**ttack (ALA), a white-box unrestricted adversarial attack that focuses on modifying the lightness of the images. The shape and color of the samples, which are crucial to human perception, are barely influenced. To obtain adversarial examples with a high attack success rate, we propose unconstrained enhancement in terms of the light and shade relationship in images. To enhance the naturalness of images, we craft the naturalness-aware regularization according to the range and distribution of light. The effectiveness of ALA is verified on two popular datasets for different tasks (*i.e.*, ImageNet for image classification and Places-365 for scene recognition).

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; • **Security and privacy** → **Software and application security**.

KEYWORDS

Adversarial Attack; Lightness; Naturalness-aware

*Work done prior to joining Meta.

†Yang Liu and Geguang Pu are the corresponding authors. (ggpu@sei.ecnu.edu.cn)

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611914

ACM Reference Format:
Yihao Huang, Liangru Sun, Qing Guo, Felix Juefei-Xu, Jiayi Zhu, Jincao Feng, Yang Liu, Geguang Pu. 2023. ALA: Naturalness-aware Adversarial Lightness Attack. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611914

1 INTRODUCTION

Deep neural networks (DNNs) are widely used in computer vision tasks. However, there are many attack approaches that can do harm to DNNs, especially those called *adversarial attacks* [27], which design deceiving inputs to mislead the DNNs into making wrong predictions. This kind of attack is widely used in various domains.

The well-known approach for generating adversarial examples is crafting L_p norm-restricted human-imperceptible noise on clean images. However, the restricted perturbations can be defended by methods such as adversarial denoising [30]. Furthermore, L_p norm adversarial attacks lack practicality in both the digital and physical worlds. Most imperceptible attacks generate floating-point adversarial examples, which would lose their aggressivity after being saved as integers (*i.e.*, the storage format of images in computers) [1]. It is also difficult to simulate imperceptible noises in the physical world with common carriers such as patches, and paints.

Therefore, studies on exploring *non-suspicious adversarial images* that allow unrestricted but unnoticeable perturbations have been proposed in recent years. Geometric attacks [5, 17], semantic attacks [16, 25] and color attacks [23, 24, 32] are the three main aspects. However, these attack methods are usually contrary to common sense, that is, creating things that seem strange in the real world. Geometric attacks are obvious to human eyes due to the destruction of regular boundaries of objects. It may generate irregularly twisted monitors or desks, *etc.* Semantic attacks modify the target in the images by adding scene-mismatch objects/textures, which usually makes the image become unnatural. For example, [16] generates adversarial face images that look quite different from the original faces and seem strange. Compared with geometric attacks and semantic attacks, the adversarial examples generated by color attacks look more natural. They mainly apply uniform transformation in similar colors [32] or colors of closed pixels [24]. In this way, the

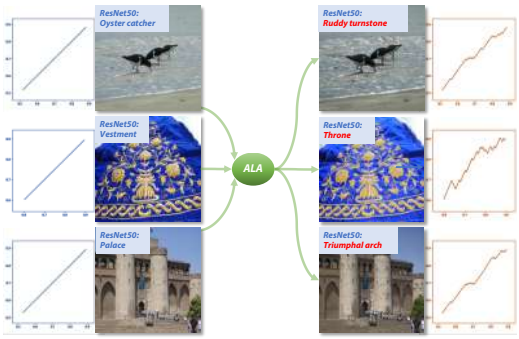

Figure 1: (L) Original images with their labels (successfully classified by ResNet50), (R) ALA attacked images that are incorrectly classified by the same ResNet50 network, with imperceptible lightness shift. The three line charts showcase the lightness value shift function generated by using our attack method.

modified images show the potential to be non-suspicious to human eyes even though the images have large perturbations. Nevertheless, even the adversarial examples generated by color attacks may not be natural enough to deceive human eyes. For example, ColorFool [24] modifies the image color according to semantic segmentation divided regions, thus its performance is strongly related to the result of segmentation, *e.g.*, if a sea is divided into two parts, then in the attacked image it may have two different colors. ACE [32] uses a simple filter to carry out the color attack. However, it may generate images with unusual objects (*e.g.*, purple river, green sky).

Since the previous attacks are semantic interference and arouse suspicion, an attack method that does not easily cause semantic aberration is imperative. There is a simple observation that the variation of lightness (even large variation) in images results in very little semantic change. It basically does not change the shape, texture, and color of the objects in the original images, *i.e.*, it does not generate images containing objects that are contrary to common sense. It essentially just constructs a new light condition for the target scene. In the real world, due to the variation of light intensity and the number of light sources, it is common to take images of different lightness with respect to the same scene and less likely to arouse suspicion. Because of these advantages, the lightness attack seems a promising direction to generate non-suspicious unrestricted adversarial images. Thus we propose **A**dversarial **L**ightness **A**ttack (ALA), a novel lightness adjustment approach to generate natural adversarial images by applying and improving a differentiable filter [11] that was originally used to adjust the image attribute in image processing. Using a filter-based attack has three advantages: ❶ The filter is human-understandable, which can guide the lightness attack in the real world, ❷ The filter is differentiable, which is time-saving than search-based attacks (*e.g.*, ColorFool [24]). ❸ The filter is lightweight and resource-saving (only dozens of parameters). However, the adversarial examples generated by directly using a monotonic lightness filter achieve low attack performance and low image quality. To obtain better attack performance and image quality, we propose two improvements: ❶ unconstrained enhancement, ❷ naturalness-aware regularization. The effectiveness of ALA is verified on two popular datasets for different tasks (*i.e.*, ImageNet for image classification and Places-365 for scene recognition).

To sum up, our work has the following contributions:

- To our best knowledge, we are the first to research adversarial attacks by focusing on adjusting image lightness with a human-understandable filter which is extremely lightweight.
- We propose a specialized unconstrained enhancement to improve the attack success rate by utilizing a non-monotonic filter and random initialization. We also propose naturalness-aware regularization to enhance the image quality of adversarial examples by adding a lightness range constraint and lightness distribution constraint.
- The experiments conducted on two datasets with different tasks shows the effectiveness of ALA in generating excellent attack performance and high-fidelity attack examples.

2 RELATED WORK

2.1 Restricted Adversarial Attacks

Traditional adversarial attacks mainly focus on generating adversarial examples with limited noise. Most researchers use small L_p norm [2, 19] to ensure this. Here are two typical methods: PGD and C&W. PGD [19] is based on projected gradient descent. I-FGSM [18] iteratively determines the perturbation in L_p norm boundary with gradient information. PGD starts the I-FGSM attack with a random point within the L_p norm boundary. Let $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ be the original image, and $\ell \in \mathbb{1}^K$ its ground truth for a K-classification problem. For a target model $\mathcal{M}(\cdot)$, $\mathcal{M}(\mathbf{I}) = \ell$, adversarial attacks aim to generate an adversarial image $\hat{\mathbf{I}}$ according to $\mathbf{I}$ to mislead the model $\mathcal{M}(\cdot)$, *i.e.*, $\mathcal{M}(\hat{\mathbf{I}}) \neq \ell$. Carlini and Wagner Attacks (C&W) [2] can be formulated as: $\min_\delta \|\hat{\mathbf{I}} - \mathbf{I}\|_p^2 + \lambda \cdot \mathcal{L}_{\text{C\&W}}(\hat{\mathbf{I}}, \ell)$, where $\mathcal{L}_{\text{C\&W}}(\hat{\mathbf{I}}, \ell) = \max(\mathcal{Z}(\hat{\mathbf{I}})_\ell - \max\{\mathcal{Z}(\hat{\mathbf{I}})_i : i \neq \ell\}, -\kappa)$ and $\hat{\mathbf{I}} = \frac{1}{2}(\tanh(\text{arctanh}(\mathbf{I}) + \delta) + 1)$. Perturbation $\delta = \hat{\mathbf{I}} - \mathbf{I}$, and λ is a constant selected by search. $\mathcal{Z}(\cdot)_i$ is the i-th class in logit of target model $\mathcal{M}(\cdot)$, and κ controls the confidence level of misclassification.

2.2 Unrestricted Adversarial Attacks

There are mainly three kinds of unrestricted attacks: geometric attacks, semantic attacks, and color attacks. Specifically, geometric attacks [5, 17] implement affine transformation to original images, making the generated images too suspicious. For example, there exist non-image parts (*i.e.*, black border) after rotating the image. This property is also employed to conduct adversarial blur attacks [7, 8]. Semantic attacks don't consider keeping the semantic stable (*e.g.*, [25] using a crafted adversarial eyeglass frame to mislead a face recognition system), leading to human-suspicion. Color attack is a feasible way to obtain non-suspicious examples for its uniformity when modifying images. ColorFool [24] proposes a semantic-guided black-box adversarial attack. Adversarial Color Enhancement (ACE) [32] changes the color of original images by using differentiable parametric filters to piecewise modify the color curve. FilterFool [23] uses an FCNN to approximate traditional image processing filters (*e.g.*, log transformation, Gamma correction, detail enhancement), which can be applied for the color attack.

The attacks that involve lightness are AVA [28], Jedena [6], and ARA [31]. AVA proposes adding vignetting to attack visual recognition. However, vignetting will reduce the image perception quality and is not common in images, which may arouse suspicion. Jedena

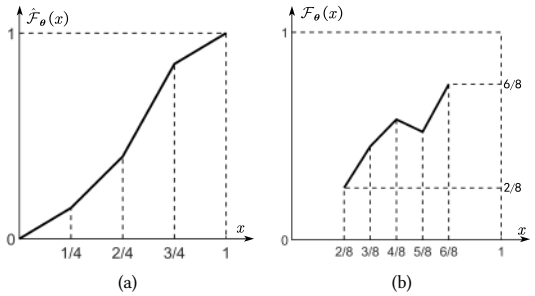

Figure 2: (a) monotonic filter $\hat{\mathcal{F}}_\theta$. (b) scene-adaptive filter $\mathcal{F}_\theta$ with the valid range from 2/8 to 6/8. Both filters are segmented into 4 pieces, *i.e.*, $T = 4$ in Eq. (1).

exploits joint exposure & noise to attack the co-saliency task, which leads to unwieldy multi-factor optimization. ARA proposes a relighting attack on the face recognition task, which is not general enough for other objects. Compared with the above unrestricted adversarial attacks, ALA has two obvious advantages: ❶ ALA, as a gradient-based attack, is more efficient than random search-based attacks. ❷ ALA provides a general and lightweight way to generate adversarial lightness examples with better naturalness.

3 ADVERSARIAL LIGHTNESS ATTACK

3.1 ALA via Parametric Filter

Given a clean image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, ground truth label l, and a pretrained model $\mathcal{M}(\cdot)$, we aim to map $\mathbf{I}$ to a new counterpart $\hat{\mathbf{I}}$ by adjusting its surrounding lightness and make the targeted model predict incorrect category, *i.e.*, $\mathcal{M}(\hat{\mathbf{I}}) \neq \mathcal{M}(\mathbf{I}) = l$. We denote the task as the adversarial lightness attack (ALA) that is desired to output naturally relighted adversarial examples while achieving a high attack success rate.

To this end, we first convert the RGB image to LAB space and separate the lightness and color into three different channels, which ensures the changes only apply to the lightness channel. The value $\mathbf{I}_L$ in the light channel is limited to the range $[0, 1]$. Then, we design the parametric filter-based ALA to formulate the lighting variation. Since piece-wise linear functions are widely used in image enhancement (*e.g.*, stretch the image contrast [29], color/tone adjustments [11]), we prefer to apply the piece-wise function for the parametric filter. Specifically, with the targeted model to guide the optimization of lighting parameters, we can adjust $\mathbf{I}$ by

$$\begin{aligned} &\hat{\mathcal{F}}_\theta(x) = \sum_{t=1}^{T} \tau_t \hat{\mathcal{F}}_{\theta_t}(x), \\ &\text{subject to } \tau_t = 1 \text{ if } x \in \left[\frac{t-1}{T}, \frac{t}{T}\right], \text{otherwise } \tau_t = 0 \end{aligned} \tag{1}$$

where x is the light value (*i.e.*, the light channel of Lab) of one pixel in $\mathbf{I}$, and we can get a new image by replacing the light values of all pixels in $\mathbf{I}$ with the new ones generated by Eq. (1), which is denoted as $\hat{\mathbf{I}} = \hat{\mathcal{F}}_\theta(\mathbf{I})$. Intuitively, the function Eq. (1) is a piece-wise linear function (See Fig. 2(a)) with T linear functions and T parameters, *i.e.*, $\theta = \{\theta_1, \ldots, \theta_T\}$. We define the t-th piece function as

$$\hat{\mathcal{F}}_{\theta_t}(x) = \frac{T}{\sum_{t=1}^{T}\theta_t}\left(\theta_t\left(x - \frac{t-1}{T}\right) + \sum_{i=1}^{t-1}\frac{\theta_i}{T}\right). \tag{2}$$

To allow an adversarial attack like the PGD, we tune the light-aware parameters θ with the guidance of the targeted model $\mathcal{M}$, *i.e.*,

$$\begin{aligned} &\arg\max_{\theta} \mathcal{L}_\theta(\mathcal{M}(\hat{\mathcal{F}}_\theta(\mathbf{I})), \ell), \\ &\text{subject to } \forall\theta_t \in \theta, \theta_t > 0, \mathbf{I}_L \in [0, 1], \end{aligned} \tag{3}$$

where the valid range of value $\mathbf{I}_L$ represents the adjustable range of lightness values during the attack. Note that the restriction $\theta_t > 0$ is generally used in image enhancement to guarantee the monotonically increasing property of the parametric filter, thus maintaining the numerical magnitude relationship between pixels and achieving better image enhancement effects. After optimizing Eq. (3) via gradient descent like PGD, we can get the optimized parameter θ^* and the adversarial image can be obtained by $\hat{\mathbf{I}} = \hat{\mathcal{F}}_{\theta^*}(\mathbf{I})$.

In summary, using the piecewise-based filter in the lightness attack has three advantages: ❶ The filter is differentiable, thus showing higher efficiency than search-based methods (*e.g.*, ColorFool [24]). ❷ The filter is more lightweight and resource-saving than other operations since the filter uses far fewer parameters (64-segment piecewise function only needs 64 parameters). ❸ The filter is human-understandable, which can guide the lightness attack in the real world (see Fig. 9). However, simply applying $\hat{\mathcal{F}}_\theta(\cdot)$ for the lightness attack leads to low image quality and attack success rate.

Figure 3: Complex light and shade relationship in the real world. The sub-images (a), (b), (c) are under various light conditions.

3.2 Unconstrained Enhancement

The attack success rate of vanilla ALA is not satisfactory (about 10% lower than other white-box unrestricted attacks). According to the observation of Eq. (3), we find that the low attack success rate is mainly caused by the monotonically increasing property of the filter. It imposes strict constraints on the light and shade relationship in the image (*i.e.*, the lighter pixels in the original image should also be lighter ones in the adversarial image), which benefits the image enhancement task but is not so necessary for attack tasks. Obviously, strict constraints are indispensable for image properties such as color because the incongruous color relationship that defies common sense may easily cause suspicion. In contrast, in Fig. 3, the lightness variation in the real world is extremely complex (*e.g.*, adding or reducing light sources to the scene of an original image can significantly influence the light and shade relationship), which cultivates people's high acceptance of the light and shade relationship. That is, in human cognition, scenes with a little unusual light and shade relationship are possible and reasonable to appear in the

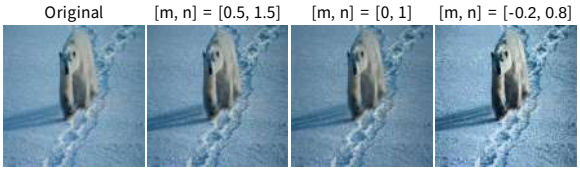

Figure 4: Adversarial examples generated by ALA with different initialization ranges.

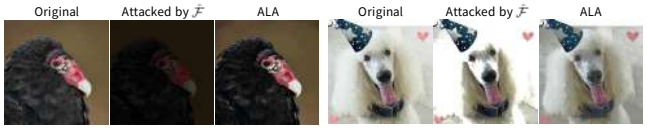

Figure 5: Adversarial images generated by different filters.

real world. Since humans are perception-insensitive to lightness and there are no original images to compare with, it is feasible to make an **unconstrained enhancement**, *i.e.*, relax restrictions of the light and shade relationship in adversarial images.

In specific, we exploit two strategies (**non-monotonic** and **random initialization**) to construct adversarial images with special lightness distribution that are challenging for target neural networks. First, we lift the restriction of $\theta_t > 0$ in Eq. (3) to get the non-strict monotonically increasing filter $\mathcal{F}_\theta(\cdot)$. Second, at the beginning of the attack, we randomly initialize the parameters θ in the range of $[m, n]$ $(m < n)$ instead of using parameters with one as initial values. This means the attack will not take the lightness values of the original image as the starting value for optimization. Thus the attack is more likely to generate a lightness distribution that is unfamiliar to pre-trained models. As is displayed in Fig. 4, the adversarial examples with different initialization ranges look different from the original image to varying degrees. In summary, the Eq. (3) is revised as

$$\arg\max_{\theta} \mathcal{L}_\theta(\mathcal{M}(\mathcal{F}_\theta(\mathbf{I})), \ell), \text{ subject to } \mathbf{I}_L \in [0, 1]. \tag{4}$$

3.3 Naturalness-aware Regularization

Although ALA with unconstrained enhancement significantly improves the attack success rate, the generated images look quite unnatural. In order to improve the naturalness of images, we take a thorough review of the generated images and empirically find that the unnaturalness of images is mainly due to the common lightness-related unnatural phenomenon (*e.g.*, overexposure, underexposure, and over-saturation), as shown in Fig. 5. These unnatural phenomena are mainly caused by abnormal fluctuations in two attributes (*i.e.*, range and distribution) of lightness. Thus the naturalness of images can be enhanced by directly adding lightness-related constraints to reduce the abnormal values of lightness.

In specific, we propose the naturalness-aware ALA with **lightness range constraint** and **lightness distribution constraint**. The lightness range constraint aims to avoid extreme lightness values (*i.e.*, overexposure and underexposure). An intuitive idea is to limit the lightness value from the valid range $[0, 1]$ to a smaller range for that it is the lightness values close to 1 and 0 cause overexposure and underexposure respectively. Considering that the lightness range of each image represents its scene characteristics,

(a) (b) (c)

Figure 6: (a) original image, (b) adversarial example filtered by $\mathcal{F}_\theta$, (c) regularized example.

we propose to adjust the lightness value adaptively according to the image scene. To be specific, for an original image of which the lowest lightness value is $\mathbf{I}_L^{\min}$ and the largest lightness value is $\mathbf{I}_L^{\max}$, it is reasonable to set the valid range to $[\mathbf{I}_L^{\min}, \mathbf{I}_L^{\max}]$ (*e.g.*, $\mathbf{I}_L^{\min} = \frac{2}{8}$ and $\mathbf{I}_L^{\max} = \frac{6}{8}$ in Fig. 2(b)).

Algorithm 1: Adversarial Lightness Attack

Input: original RGB image $\mathbf{I}_{RGB}$, ground truth label l, target model $\mathcal{M}(\cdot)$, number of iterations N, learning rate α, regularization rate β, number of segmented pieces T
Output: adversarial image $\hat{\mathbf{I}}$

Initialize $\theta^1 \leftarrow \text{Random}(T)$
$\mathbf{I}_{LAB} \leftarrow \text{RGBtoLab}(\mathbf{I}_{RGB})$
$\mathbf{I}_L, \mathbf{I}_A, \mathbf{I}_B \leftarrow \text{Split}(\mathbf{I}_{LAB})$
for $i \leftarrow 1$ to N **do**
  $\mathbf{I}_L^i \leftarrow \mathcal{F}_{\theta^i}(\mathbf{I}_L)$
  $\mathbf{I}_{LAB}^i \leftarrow \text{Concatenate}(\mathbf{I}_L^i, \mathbf{I}_A, \mathbf{I}_B)$
  $\mathbf{I}_{RGB}^i \leftarrow \text{LabtoRGB}(\mathbf{I}_{LAB}^i)$
  if $\mathcal{M}(\mathbf{I}_{RGB}^i) \neq \ell$ **then**
    $\hat{\mathbf{I}} \leftarrow \mathbf{I}_{RGB}^i$
  end if
  $g \leftarrow \nabla_\theta(\mathcal{L}_{\text{C\&W}}(\mathbf{I}_{RGB}^i, \ell) + \beta \cdot (-\frac{1}{T}\sum_{j=1}^{T}|\theta_j|))$
  $\theta^{i+1} \leftarrow \theta^i - \alpha \cdot \frac{g}{\|g\|_2}$
end for
return $\hat{\mathbf{I}}$

The lightness distribution constraint aims to avoid the tendency of pixel lightness to be similar values. Fig. 6(b) shows some adversarial images with unnatural lightness distribution. For example, many pixels change to extremely similar high lightness values, which is rare and unnatural. It is the side effect of removing the monotonic constraint (*i.e.*, $\theta_t > 0$), which makes the adjacent lightness become the extremely close lightness (*i.e.*, values of parameters θ of the pieces are close to 0). The generated images with too much "same lightness" region are noticeable and suspicious.

To address this problem, we need to keep the pixels of similar lightness in the original images to be different after the attack. Since the piecewise mapping function is represented by a set of parameters θ, imposing penalties on parameters θ_t that are close to 0 can effectively avoid similar lightness values in adversarial images. Thus we propose the lightness distribution constraint $\mathcal{L}_{\text{R}} = -\frac{1}{T}\sum_{t=1}^{T}|\theta_t|$ and the objective function Eq. (4) can be further reformulated as:

$$\arg\max_{\theta} \left(\mathcal{L}_\theta(\mathcal{M}(\mathcal{F}_\theta(\mathbf{I})), \ell) - \frac{1}{T}\sum_{t=1}^{T}|\theta_t|\right),$$
$$\text{subject to } \mathbf{I}_L \in [\mathbf{I}_L^{\min}, \mathbf{I}_L^{\max}]. \tag{5}$$

3.4 Algorithm of ALA

In Algorithm 1, we specifically show the concrete procedure of **A**dversarial **L**ightness **A**ttack (ALA). We first randomly generate T values in the range of $[m, n]$ as the initial parameters θ^1 (*i.e.*, the

Table 1: Comparison of seven attack baselines and our method on ImageNet. It shows the results of adversarial attacks on four normally trained models: ResNet50, DenseNet121, VGG19 and MobileNet-v2. The first column displays the attack success rates (Succ Rate), and the last three columns are image quality metrics LPIPS score, PIQE score, and NIQE score, where we use red, yellow, and blue to mark the first, second, and third highest image quality for unrestricted attack methods.

Target Model	ResNet50				DenseNet121				VGG19				MobileNet-v2			
Metrics	Succ Rate↑	LPIPS↓	PIQE↓	NIQE↓	Succ Rate↑	LPIPS↓	PIQE↓	NIQE↓	Succ Rate↑	LPIPS↓	PIQE↓	NIQE↓	Succ Rate↑	LPIPS↓	PIQE↓	NIQE↓
PGD	92.87%	0.005	8.059	47.485	95.39%	0.005	8.047	47.628	94.69%	0.005	8.098	47.432	98.69%	0.005	8.032	47.364
C&W	100.00%	0.005	10.449	47.783	100.00%	0.006	10.715	47.769	100.00%	0.004	10.824	47.775	100.00%	0.004	10.086	48.022
MIM	95.71%	0.051	4.302	30.443	99.86%	0.053	4.667	32.380	98.57%	0.047	4.772	32.363	99.29%	0.049	4.474	31.469
ColorFool	90.64%	0.208	13.417	44.577	84.11%	0.229	13.834	44.085	91.35%	0.205	13.520	44.674	91.98%	0.185	13.301	44.885
ACE	96.80%	0.297	12.771	41.603	95.72%	0.288	12.870	41.389	98.92%	0.295	12.698	41.073	98.34%	0.297	12.734	40.904
EdgeFool	99.27%	0.127	12.185	38.663	98.54%	0.126	12.134	38.763	99.16%	0.127	12.097	38.267	99.39%	0.125	12.101	38.668
FilterFool	100.00%	0.111	16.427	39.688	100.00%	0.111	16.754	42.344	100.00%	0.109	16.065	42.071	100.00%	0.112	16.021	38.802
ALA (Ours)	97.53%	0.124	10.406	28.636	96.19%	0.125	10.547	28.657	98.97%	0.110	9.961	28.472	99.14%	0.109	10.093	28.938

first iteration of $\{\theta_1, \theta_2, \cdots, \theta_T\}$) of the mapping filter $\mathcal{F}_\theta$ in line 1, and transform the original image $\mathbf{I}_{RGB}$ to LAB color space $\mathbf{I}_{LAB}$ in line 2. Then we can easily get $\mathbf{I}_L$, the lightness of $\mathbf{I}$, by splitting the LAB channels of the transformed image $\mathbf{I}_{LAB}$. After initialization, we start to generate the adversarial attack example iteratively. In the i-th step, we firstly use the filter $\mathcal{F}_{\theta^i}$ to modify the original lightness channel $\mathbf{I}_L$ and obtain the new lightness channel $\mathbf{I}_L^i$ in line 5. Next, in line 6, we concatenate the new L channel with the original A and B channels to get a new image in LAB color space, $\mathbf{I}_{LAB}^i$. In line 7, we resume it to RGB color space image, $\mathbf{I}_{RGB}^i$. In lines 8 and 9, we update the output $\hat{\mathbf{I}}$ if the $\mathbf{I}_{RGB}^i$ successfully misleads the target model $\mathcal{M}$. In line 11, we combine the $\mathcal{L}_{\text{C\&W}}$ (Sec. 2.1) with the distribution-aware regularization $\mathcal{L}_{\text{R}}$ defined in Eq. (5) as the loss to compute the gradient $\boldsymbol{g}$. The β is a hyper-parameter to control the weight of regularization. Finally, we use gradient $\boldsymbol{g}$ with learning rate α to update the new parameter θ^{i+1} of the filter $\mathcal{F}_\theta$.

4 EXPERIMENT

4.1 Experiment Setup

Datasets. We carry out our experiments on two datasets for different tasks respectively. For image classification, as ImageNet [3] has 1,000 classes, we randomly choose 3 images per class to make up our dataset of 3,000 images with the size of 224×224. We perform the experiment for scene recognition on 1,095 images (randomly selected 3 images per class from Places-365 [33], which has 365 classes) with the size of 256×256.

Target model. We choose four networks: ResNet50 [9], VGG19 [26], DenseNet121 [12], and MobileNet-v2 [21], for image classification. For scene recognition, we choose MobileNet-v2, ResNet50, and DenseNet161 as the target model.

Metrics. To evaluate the naturalness of images, we assess the images from two perspectives. To assess the human-perceptual similarity of the adversarial examples with original images, we use learned perceptual image patch similarity (LPIPS). We also use two non-reference image quality assessment methods, perception-based image quality evaluator (PIQE) and natural image quality evaluator (NIQE), to partly quantify the quality of the adversarial examples. All three metrics are the lower the better.

Baseline methods. We choose three L_p norm-restricted adversarial attacks and four unrestricted attacks as baselines. For restricted methods, we use PGD [19] with 10 iterations and $\epsilon = 2/255$, C&W attack [2] with 5×200 iterations and confidence $\kappa = 20$, and MIM with $\epsilon = 4/255$ for similar image quality with ALA. MIM [4] combines the momentum term with I-FGSM for a better transfer attack success rate. For unrestricted attack methods, we follow the experiment settings of ACE [32], ColorFool [24], EdgeFool [22], and apply FilterFool [23] with 1,500 iterations and stopping threshold $\tau = 0.006$. Although the settings for FilterFool are not the same as in the paper of 3,000 iterations and $\tau = 0.004$, the results are similar.

Implementation details. We divide the valid lightness into 64 pieces, *i.e.*, $T = 64$ in Eq. (1). We set the learning rate $\alpha = 0.5$ and the number of iterations $N = 100$. As mentioned in Sec 3.3, we use $\mathcal{L}_{\text{C\&W}}$ within $\kappa = 0.2$ and the regularization $\mathcal{L}_{\text{R}}$ with $\beta = 0.3$ as the loss function. And the random initialization range is set as $[m, n] = [-0.2, 0.8]$. All the experiments were run on Pytorch 1.8, and CUDA 11.1 with an NVIDIA GeForce RTX 2070 SUPER GPU.

4.2 Image Classification

We compare the effectiveness of ALA with the baselines. In Table 1, most of the attack methods achieve very high success rates except for the black-box method ColorFool. Restricted methods PGD, C&W, and MIM exploit the gradient of every pixel, thus they can successfully attack almost all images. EdgeFool and FilterFool obtain high success rates, but they cost much more training resources (*e.g.*, 5,000 and 1,500 iterations with both 10.96 GFLOPs) than ACE and ALA (*e.g.*, both are 100 iterations with 8.24 GFLOPs).

Next, we compare the naturalness of the adversarial examples generated by different attacks. From Table 1 we can see that the restricted methods obtain quite high LPIPS and PIQE scores. This is because LPIPS assesses the difference between attacked examples and original images, and PIQE mainly uses a special standard deviation to calculate the score, both of them are hardly influenced by restricted tiny perturbations. Even though, ALA obtains better PIQE scores than C&W on more than half of models. For MIM, as we mentioned in Sec. 4.1, we use the MIM with $\epsilon = 4/255$ to generate examples having similar image quality with ALA. In this setting, the examples generated by MIM have worse NIQE and transferability than ALA. Regarding unrestricted attacks, the proposed ALA reaches the best performance of LPIPS and PIQE in almost all cases. FilterFool sometimes gets better LPIPS scores, but it cost more than ten times hours to train compared to ALA. Significantly, ALA clearly obtains the best NIQE performance in all cases.

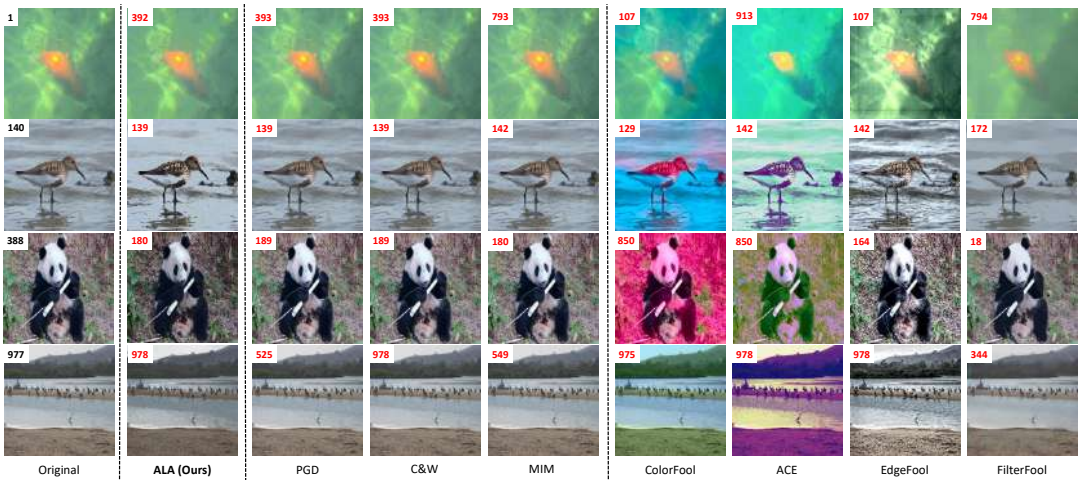

Figure 7: Adversarial examples. The top left corner shows the predicted result (ImageNet index) by MobileNet-v2.

Table 2: Comparison of transferability on baselines and ALA. It shows the success rates of transfer attacks on four standard trained models: ResNet50 (ResNet), DenseNet121 (DenseNet), VGG19 (VGG), and MobileNet-v2 (MobileNet).

Target Model	ResNet50			DenseNet121			VGG19			MobileNet-v2		
Attacked Model	DenseNet	VGG	MobileNet	ResNet	VGG	MobileNet	ResNet	DenseNet	MobileNet	ResNet	DenseNet	VGG
PGD	6.39%	4.77%	6.81%	6.90%	4.82%	6.00%	4.29%	4.37%	6.51%	4.16%	4.37%	5.11%
C&W	10.06%	9.73%	10.19%	15.30%	11.60%	12.66%	6.26%	5.74%	8.88%	7.35%	6.39%	8.75%
MIM	35.17%	29.83%	31.97%	43.03%	37.30%	40.60%	24.07%	24.73%	31.27%	25.54%	26.52%	33.61%
ColorFool	19.42%	31.45%	36.91%	32.89%	36.61%	41.50%	21.43%	15.33%	30.86%	18.23%	13.82%	23.83%
ACE	41.00%	61.67%	58.04%	51.16%	60.29%	57.79%	46.19%	37.33%	54.87%	46.73%	36.58%	58.77%
EdgeFool	23.65%	34.05%	32.63%	25.26%	32.43%	31.06%	23.71%	20.36%	29.25%	23.30%	20.73%	30.12%
FilterFool	28.82%	43.78%	37.82%	28.32%	42.80%	37.07%	24.12%	22.99%	34.44%	24.17%	23.23%	41.18%
ALA (Ours)	28.63%	44.08%	43.42%	33.17%	44.62%	43.77%	24.67%	20.40%	33.84%	23.57%	19.37%	34.10%

In Figure 7, we display the original inputs and the adversarial examples generated by baselines and ALA. The restricted examples of PGD and C&W look highly similar to the original images, as their noises are too small to be noticed by humans, while the noise of MIM can be found when observing the image carefully. As for unrestricted attacks, ColorFool generates unnatural images, *e.g.*, the bird and water in the top picture are of unnatural mixed colors. ACE and FilterFool look obviously have been processed, and EdgeFool looks unnatural around the outlines of the two animals. Compared with these unrestricted attack methods, ALA looks just like the same scene in different light conditions, *e.g.*, decreasing the light intensity in the bottom image. Besides, we use 100 ResNet50 adversarial examples of each unrestricted attack method and their corresponding clean images to calculate the maximum mean discrepancy (MMD), which is often used in transfer learning for measuring the similarity between two different distributions. The lower MMD means the two distributions are more similar. ALA obtains the lowest score of 0.063, the following methods are EdgeFool (0.075), ColorFool (0.11), ACE (0.298), and FilterFool (0.690). The result shows that in unrestricted adversarial attacks, the adversarial images generated by ALA are in the domain that is most similar to the domain of natural (*i.e.*, clean) images. Furthermore, the score of ALA on MMD is even comparable to restricted attacks (MMD of PGD is 0.0439), which fully demonstrates that the ALA is naturalness-aware. This is a very surprising result since the noise generated by PGD is minor and the adjustment by ALA is much larger.

Then we compare the transferability of different attacks in Table 2. Transfer attack means using the adversarial examples generated by the target model to attack other models. As shown in the first two rows of Table 2, although human-imperceptible restricted methods can achieve high success rates in target models, their transferability is too weak. Though ALA doesn't focus on transferability, within similar image quality, it still gains a comparable performance with MIM, which is specially designed for transfer attack. Among the five unrestricted attacks, ACE always performs best for transferability. The best two methods in the rest are ALA and FilterFool. It is noteworthy that, ACE wins the transferability at the price of image quality, and ALA obtains no worse transferability than FilterFool with much better image quality and much fewer training resources (ALA spends about 5 hours, while FilterFool spends more than 100 hours). All in all, among the unrestricted attacks, ALA obtains the best performance of image quality with considerable transferability and few training resources, which fully reflects the practicality of ALA to be applied in attack scenarios.

4.3 Scene Recognition

Scene recognition is a challenging task, as it needs to understand the context that consists of different objects and various relationships.

Table 3: Comparison of attack performance on Places-365. It shows the success rates (%) of adversarial attacks on ResNet50 (ResNet), DenseNet161 (DenseNet), and MobileNet-v2 (MobileNet), in first three columns. In last three columns, it shows scores of three image quality metrics. We use red, yellow, and blue to mark the first, second, and third performances of unrestricted attack methods.

Target Model	ResNet50						DenseNet161						MobileNet-v2					
Model&Metrics	ResNet	DenseNet	MobileNet	LPIPS↓	PIQE↓	NIQE↓	ResNet	DenseNet	MobileNet	LPIPS↓	PIQE↓	NIQE↓	ResNet	DenseNet	MobileNet	LPIPS↓	PIQE↓	NIQE↓
PGD	78.42	9.58	6.81	0.005	7.071	44.462	13.20	79.38	5.41	0.005	7.011	44.430	4.51	2.76	92.50	0.005	7.018	44.135
C&W	100.00	39.77	28.62	0.034	9.149	45.580	46.54	100.00	28.80	0.036	8.948	45.863	17.55	15.10	100.00	0.037	9.699	44.135
MIM	91.30	49.83	45.20	0.098	5.569	29.667	57.00	92.37	45.54	0.106	5.488	29.211	27.53	24.02	97.90	0.085	5.943	28.747
ColorFool	90.50	25.32	37.35	0.156	11.661	41.426	33.98	90.91	41.88	0.170	11.628	41.785	22.38	18.18	95.11	0.137	11.581	41.896
ACE	95.81	60.39	70.33	0.315	11.230	39.597	59.26	95.29	65.10	0.310	11.204	39.395	53.14	49.84	97.91	0.313	11.262	38.961
EdgeFool	99.52	27.76	43.98	0.126	13.036	33.356	32.85	99.84	43.28	0.127	13.125	33.017	29.47	25.49	99.13	0.125	13.164	33.148
FilterFool	89.74	23.37	27.02	0.086	18.070	46.872	41.02	91.42	23.52	0.083	19.277	48.277	28.40	24.17	92.94	0.086	17.721	47.585
ALA(Ours)	98.87	43.02	54.45	0.131	11.193	27.599	44.15	96.43	45.90	0.127	11.113	27.675	31.24	30.52	98.95	0.117	10.667	27.872

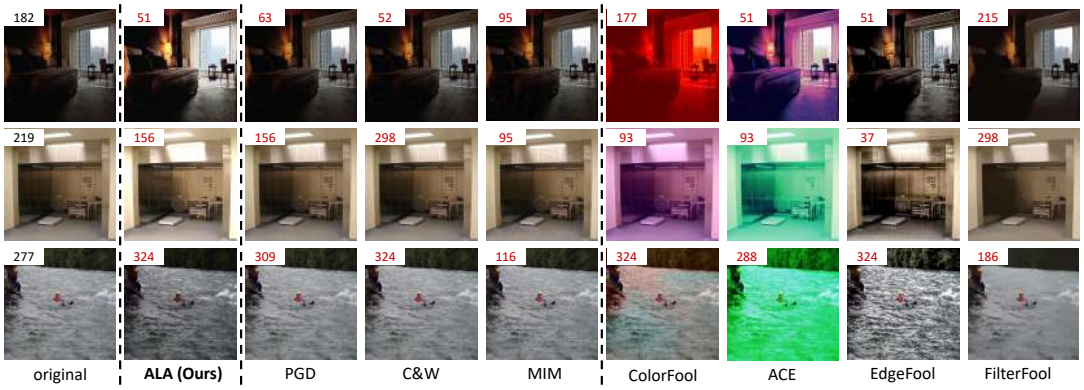

Figure 8: Adversarial examples. The top left corner shows the predicted result (Places-365 index) by MobileNet-v2.

Figure 9: (a) Original image, (b) ALA image, (c) Physical image.

Thus the attack performance of different methods on the Places-365 dataset is quite different from the results of ImageNet. For example, as shown in Table 3, in terms of the restricted attack methods, the attack performance of PGD is not satisfactory. It has a poor attack success rate of lower than 80% on the target model and rarely has attack transferability.

In the unrestricted attack methods, ALA achieves the second-best attack success rate and second-best attack transferability with the best image quality score. Especially, ALA gets the best NIQE scores among all baselines including restricted and unrestricted attack methods. Fig. 8 displays the adversarial examples of baselines on Places-365. Among unrestricted attacks, the adversarial examples generated by ALA show the best naturalness and image quality.

4.4 Real-world Attack

As an attack method that adjusts natural phenomena (*i.e.*, light), ALA has a certain possibility to be implemented in the physical world by manually modifying the number and angle of light sources. To simulate the lightness attack in the real world, we first take a photo of the clean image from the real world. Then we generate an adversarial image by ALA. To conveniently simulate the adversarial examples, we decrease the segment number of ALA, *i.e.*, $T = 16$ in Eq. (1) and don't use random initialization. Third, we take the adversarial example as the reference and construct a light condition similar to it. Finally, we take a photo from the same view as step one. In most cases, such manually-built adversarial examples in the real world can successfully fool the target model.

We carry out this experiment on 50 objects and finally get an attack success rate of 80%. One of the attacked objects is shown in Figure 9, we take a photo of a cup (Fig. 9(a)) in the real world and classify this image by ResNet50. Then we attack this image by ALA (Fig. 9(b)) and successfully fool the model. Next, we simulate the same light condition with the digital adversarial image generated by ALA in the real world and still take a photo of the cup. The new image (Fig. 9(c)) also mislead the ResNet50. The attack effect between Fig. 9(a) and Fig. 9(c) is mainly reflected in the fact that the shadows of the bookshelves and cups become darker in Fig. 9(c).

4.5 Adversarial Training

The generated adversarial examples of ALA can also be used as a kind of data augmentation to enhance the robustness of networks on unseen lightness corruptions. We take the classification task as an example to show this. We combine 6,000 images generated by our method with original images in ImageNet (12,000 images total) as the train set to fine-tune the standard trained ResNet50 model. To verify the effectiveness, we use part of the ImageNet-C [10], which contains the ImageNet validation set with various common corruptions (*e.g.*, brightness corruption, jpeg compression) as the test set. As our method focuses on modifying the lightness, we use the lightness corrupted images. The standard trained ResNet50 achieves 58.93% accuracy in the test set. After fine-tuning with the

Table 4: Ablation study. Range, Dist, N-Mono, and Rand mean lightness range constraint, lightness distribution constraint, non-monotonic, and random initialization respectively.

Method	**Range**	**Dist**	**N-Mono**	**Rand**	ResNet50	DenseNet121	VGG19	MobileNet-V2	LPIPS↓	PIQE↓	NIQE↓
ALA_0					84.33%	11.61%	24.37%	24.10%	0.072	14.714	27.975
ALA_1	√				73.69%	7.48%	14.94%	15.23%	0.041	13.550	25.499
ALA_2		√			83.65%	10.72%	22.21%	22.69%	0.066	14.502	27.714
ALA_3			√		94.24%	15.14%	29.63%	30.61%	0.080	14.030	28.286
ALA_4				√	85.56%	16.78%	30.27%	31.92%	0.097	15.443	29.408
ALA_5	√	√			73.96%	6.91%	13.51%	14.22%	0.039	13.519	25.554
ALA_6			√	√	99.04%	44.80%	63.98%	64.95%	0.195	16.229	38.344
ALA_7	√	√	√	√	97.53%	28.63%	44.08%	43.42%	0.124	10.406	28.636

new training dataset, the accuracy increases to 62.01%, showing the effectiveness of ALA in improving the robustness of DNNs on unseen lightness corruption.

5 ABLATION STUDY

To verify the effectiveness of the proposed lightness range constraint, lightness distribution constraint, non-monotonic, and random initialization. We conduct an ablation study on each of them and their combination. The dataset is the same as the image classification task (Sec. 4.2). ALA_0 is the vanilla version of ALA (implementation of filter in Sec. 3.1), which is of low image quality and fair attack success rate. ALA_7 is the final version of ALA, which exploits all the optimization strategies and finds a trade-off between image quality and attack performance. It achieves a significantly high attack success rate while maintaining the naturalness of images.

Unconstrained Enhancement. As is mentioned in Sec. 3.2, we propose unconstrained enhancement (*i.e.*, non-monotonic filter and random initialization) to improve the attack success rate. From the Table. 4, we can see that the ALA_3 and ALA_4, which respectively use the non-monotonic filter and random initialization, both obtain better attack success rates. Furthermore, the combined attack method, ALA_6, has the best attack performance. However, the images generated by ALA_6 have the poorest image quality on all three metrics and are human-suspicious, as is shown in Fig. 10.

Naturalness-aware Regularization. As is mentioned in Sec. 3.3, we propose the naturalness-aware regularization (*i.e.*, lightness range constraint and lightness distribution constraint) to improve the image quality. In Table. 4, ALA_1 and ALA_2 use the two optimizations respectively, and they both improve the image quality of adversarial examples. We combine them in ALA_5 to gain better image quality. However, only using the naturalness-aware regularization will result in unsatisfactory attack performance. Thus we combine the unconstrained enhancement and naturalness-aware regularization to get ALA_7, the comprehensive optimal ALA with both satisfactory attack performance and image quality.

6 DISCUSSION

We think why the piecewise-based filter works well for the lightness attack can be attributed to three key advantages. ❶ **Efficient**. The piecewise-based filter offers an efficient means of adjusting various image properties, including lightness, in a parameterized manner [1]. Therefore, it is reasonable to leverage the piecewise-based filter to construct an image property-related (e.g., lightness) attack. ❷ **Unlimited**. The piecewise-based filter offers the capability to make significant adjustments to the lightness of images, which is necessary for constructing an unrestricted attack and reduces the difficulty of executing the attack. ❸ **Tidy**. The piecewise-based filter optimizes the lightness value according to their different intensity levels, which avoids the adjustment of brightness to be individually or messy, ensuring the naturalness of the image.

(a) (b) (c)

Figure 10: Different ALA methods examples with their original images. The sub-images (a) is the original image, (b) is generated by ALA_6, and (c) is generated by ALA_7.

7 CONCLUSION

We propose a novel adversarial lightness attack method ALA, which generates unrestricted examples with better naturalness than existing unrestricted adversarial attacks. In the future, we aim to combine other image attributes (*e.g.*, contrast, color curve, and exposure) or degradation (*e.g.*, noise, blur) with lightness for better performance and to explore the real-world attack more systematically. Furthermore, we are interested in exploring adversarial defense methods [13–15, 20] against these unrestricted attacks.

ACKNOWLEDGMENTS

Geguang Pu is supported by the National Key Research and Development Program (2020AAA0107800), and Shanghai Collaborative Innovation Center of Trusted Industry Internet Software. This work is supported by Nanyang Technological University (NTU)-DESAY SV Research Program under Grant 2018-0980, the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-022T). It is also supported by A*STAR Centre for Frontier AI Research, the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), NRF Investigatorship No. NRF-NRFI06-2020-0001.

REFERENCES

[1] Lei Bu, Zhe Zhao, Yuchao Duan, and Fu Song. 2021. Taking care of the discretization problem: A comprehensive study of the discretization problem and a black-box adversarial attack in discrete integer domain. *IEEE Transactions on Dependable and Secure Computing* (2021).

[2] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. IEEE, 39–57.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9185–9193.

[5] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*. PMLR, 1802–1811.

[6] Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Huazhu Fu, Wei Feng, Yang Liu, and Song Wang. 2022. Can you spot the chameleon? adversarially camouflaging images from co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2150–2159.

[7] Qing Guo, Ziyi Cheng, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yang Liu, and Jianjun Zhao. 2021. Learning to Adversarially Blur Visual Object Tracking. In *ICCV*. 10839–10848.

[8] Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, and Yang Liu. 2020. Watch out! Motion is Blurring the Vision of Your Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

[10] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations* (2019).

[11] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2018. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics (TOG)* 37, 2 (2018), 1–17.

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

[13] Yihao Huang, Qing Guo, Felix Juefei-Xu, Lei Ma, Weikai Miao, Yang Liu, and Geguang Pu. 2021. AdvFilter: Predictive Perturbation-aware Filtering against Adversarial Attack via Multi-domain Learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 395–403.

[14] Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. 2022. Prior-guided adversarial initialization for fast adversarial training. In *European Conference on Computer Vision*. Springer, 567–584.

[15] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. 2022. LAS-AT: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13398–13408.

[16] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. 2019. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4773–4783.

[17] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. 2018. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4441–4449.

[18] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. 2016. Adversarial examples in the physical world.

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[20] Hua Qi, Zhijie Wang, Qing Guo, Jianlang Chen, Felix Juefei-Xu, Lei Ma, and Jianjun Zhao. 2021. ArchRepair: Block-level architecture-oriented repairing for deep neural networks. *arXiv preprint arXiv:2111.13330* (2021).

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

[22] Ali Shahin Shamsabadi, Changjae Oh, and Andrea Cavallaro. 2020. Edgefool: an adversarial image enhancement filter. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1898–1902.

[23] Ali Shahin Shamsabadi, Changjae Oh, and Andrea Cavallaro. 2021. Semantically Adversarial Learnable Filters. *IEEE Transactions on Image Processing* 30 (2021), 8075–8087.

[24] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. 2020. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1151–1160.

[25] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. 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*. 1528–1540.

[26] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[28] Binyu Tian, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Xiaohong Li, and Yang Liu. 2021. AVA: Adversarial Vignetting Attack against Visual Recognition. *arXiv preprint arXiv:2105.05558* (2021).

[29] Chun-Ming Tsai and Zong-Mu Yeh. 2008. Contrast enhancement by automatic and parameter-free piecewise linear transformation for color images. *IEEE transactions on Consumer Electronics* 54, 2 (2008), 213–219.

[30] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. 2019. Feature Denoising for Improving Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[31] Qian Zhang, Qing Guo, Ruijun Gao, Felix Juefei-Xu, Hongkai Yu, and Wei Feng. 2021. Adversarial relighting against face recognition. *arXiv preprint arXiv:2108.07920* (2021).

[32] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. 2020. Adversarial robustness against image color transformation within parametric filter space. *arXiv preprint arXiv:2011.06690* (2020).

[33] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).