ALA: Adversarial Lightness Attack via Naturalness-aware Regularizations

Liangru Sun[1], Felix Juefei-Xu[2], Yihao Huang[3], Qing Guo[3], Jiayi Zhu[1],
Jincao Feng[1], Yang Liu[3], Geguang Pu[1,4]
[1] East China Normal University, China [2] Alibaba Group, USA
[3] Nanyang Technological University, Singapore
[4] Shanghai Industrial Control Safety Innovation Technology Co., Ltd, China

Abstract

Most researchers have tried to reveal the vulnerability of deep neural networks (DNNs) with specialized adversarial examples. *Parts of the attack examples have imperceptible perturbations restricted by $\mathcal{L}_p$ norm, which can be easily defensed. Some works make the perturbations unrestricted for better robustness and transferability. However, these examples usually look unnatural. To generate unrestricted adversarial examples with high image quality and good transferability, 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. We verify the effectiveness of ALA on ImageNet for image classification. The experiments show that the generated adversarial examples have both strong transferability and high image quality.*

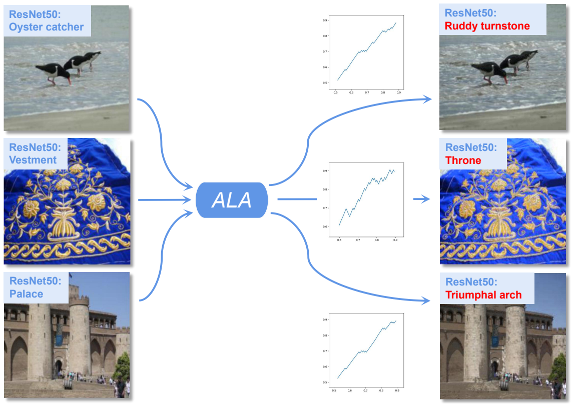


Figure 1. (L) Original images with their labels (successfully classified by ResNet50), (R) ALA attacked images that are incorrectly classified by the same ResNet50 network, with imperceptible lightness shift. Three line charts mean the lightness value shift function generated by using our attack method. Through adjusting the lightness values in the original images with the shift functions, the images become aggressive and natural.

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* [18], which design deceiving inputs to mislead the DNNs into making wrong predictions. A well-known approach to generating adversarial examples for adversarial attacks is crafting restricted noise (*i.e.*, imperceptible to human eyes) to modify original inputs. A majority of researchers use $\mathcal{L}_p$ norm to restrict the noise. However, most restricted perturbations can be defended.

Therefore, studies on exploring *non-suspicious adversarial images* that allow unrestricted but unnoticeable perturbations have been emerging. Geometric attacks [4,9], semantic attacks [16], and color attacks [14, 15, 21] are three main aspects. However, these attack methods are usually contrary to common sense. Among them, the adversarial examples generated by color attacks seem more natural for their uniform transformation. However, even the adversarial examples generated by color attack may not be natural enough to deceive human eyes.

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 little semantic change. In addition, in the real world, it is common to take images of different lightness with respect to the same scene. Thus lightness attack seems promising and 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 [8] that was originally used to adjust the image attribute in image processing. Compared with the color attack, the lightness attack will 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.

To sum up, our work has the following contributions: ❶ To our best knowledge, we are the first to research adversarial lightness attacks by focusing on human-understandable filter. ❷ We propose a crafted unrestricted filter with customized naturalness-aware regularization. We design random initialization and non-stop attack strategies to obtain adversarial images with stronger transferability. ❸ The experiment shows the effectiveness of ALA in generating strong-transferability and high-fidelity attack examples.

2. Related Work

2.1. Restricted Adversarial Attacks

PGD [10] is based on projected gradient descent. It starts at a random point within the allowed $\mathcal{L}_p$ norm boundary and iteratively determines the perturbation with gradient information. Let $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ be the original image, and $l \in \mathbb{1}^K$ is its ground truth for a K-classification problem. For a target model $\mathcal{M}(\cdot)$, $\mathcal{M}(\mathbf{I}) = l$. Adversarial attacks aim to generate an adversarial image $\mathbf{I}'$ according to $\mathbf{I}$ to mislead the model $\mathcal{M}(\cdot)$, *i.e.*, $\mathcal{M}(\mathbf{I}') \neq l$. Carlini and Wagner Attacks (C&W) [1] can be formulated as: $\min_\delta \|\mathbf{I}' - \mathbf{I}\|_p^2 + \lambda \cdot \mathcal{L}_{\text{C\&W}}(\mathbf{I}', l)$, where $\mathcal{L}_{\text{C\&W}}(\mathbf{I}', l) = \max(\mathcal{Z}(\mathbf{I}')_l - \max_i\{\mathcal{Z}(\mathbf{I}')_i : i \neq l\}, -\kappa)$ and $\mathbf{I}' = \frac{1}{2}(\tanh(\operatorname{arctanh}(\mathbf{I}) + \delta) + 1)$. The perturbation $\delta = \mathbf{I}' - \mathbf{I}$, and λ is a constant selected by binary 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

Color attack is a feasible way to obtain non-suspicious examples for its uniformity when modifying images. ColorFool [15] proposes a semantic-guided black-box adversarial attack. It randomly modifies the color of semantic segmentation dividing regions in different pre-defined ranges. The quality is strongly related with the chosen semantic segmentation network. Adversarial Color Enhancement (ACE) [21] piecewise modifies the color of original images by using differentiable parametric filters. FilterFool [14] uses a fully convolutional neural network (FCNN) to hide attacks into traditional image processing filters.

To our best knowledge, the works that focus on lightness attack are EdgeFool [13] and AVA [19]. AVA attacks visual recognition by adding uncommon vignetting, which will reduce the image perception quality. EdgeFool uses an FCNN to generate detail-enhanced adversarial images, which need huge consumption. Compared with the above unrestricted adversarial attacks, ALA is can generate adversarial examples with better naturalness and strong transferability.

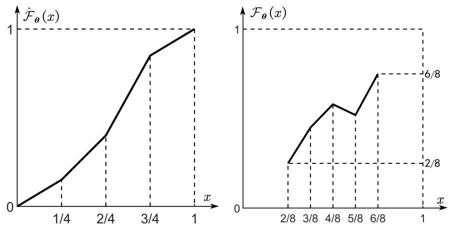
3. Adversarial Lightness Attack (ALA)

3.1. Parametric Filter

We refer to the differentiable parametric filter, which is proposed as a photo enhancement method [8], to realize the ALA. [8] mainly uses filters represented by a set of parameters to realize traditional image processing approaches (*e.g.*, white balance, color curves adjustment) for generating specific styled photos. The linear filter can be formulated as:

$$\hat{\mathcal{F}}_\theta(x_t) = \frac{T}{\theta_{\text{sum}}}\left[\theta_t\left(x_t - \frac{t-1}{T}\right) + \sum_{i=1}^{t-1}\frac{\theta_i}{T}\right], \tag{1}$$

$$\theta_{\text{sum}} = \sum_{t=1}^{T}\theta_t, \theta_t > 0. \tag{2}$$



(a) (b)

Figure 2. (a) monotonic filter $\hat{\mathcal{F}}$. (b) unrestricted filter $\mathcal{F}$ with the valid range [**2/8**, **6/8**]. Both filters are segmented into 4 pieces.

In Eq (1), $\hat{\mathcal{F}}$ is the filter with parameters θ (*i.e.*, $\{\theta_1, \theta_2, \cdots, \theta_T\}$) and T denotes the number of segmented pieces, *e.g.*, $T = 4$ in Fig 2a. The x_t means the pixel values that belong to t-th piece where values are filtered using the parameter θ_t. The θ_t means the gradient of the t-th piece in the mapping function. In [8], the θ_t is restricted to be bigger than zero, making the filter monotonic. The light and shade relationship will not change with monotonic filters.

3.2. Constraint ALA

Given an original image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, its ground truth label l, and a pretrained target model $\mathcal{M}(\cdot)$, We aim to achieving the formula $\mathcal{M}(\text{ALA}(\mathbf{I})) \neq l$. With the monotonic filter $\hat{\mathcal{F}}$, it is easy to construct the ALA by optimizing the objective function:$\arg\max_\theta \mathcal{L}(\mathcal{M}(\hat{\mathcal{F}}_\theta(\mathbf{I})), l)$, where $\mathcal{L}(\cdot)$ denotes the loss function. By directly applying a gradient-based attack we can obtain the adversarial images. Noting that ALA focuses on modifying the image lightness, thus we apply the attack on *Lab* color space [11].

Since θ is restricted to be positive and the pixels values are scaled in range $[0, 1]$ in the original method of parametric filter [8], we call the attack method *constraint ALA*.

3.3. Unrestricted ALA with Constraint Relaxation

By using filter $\hat{\mathcal{F}}$, we can generate adversarial examples while maintaining the light-shade relationship, but the attack success rate is low. Since there are no original images to compare practically, and lightness changing in the real world is much more complex, it is not so necessary to maintain the light-shade relationship. Therefore, we lift the restriction of $\theta_t > 0$ in Eq (2). For lightness value, the valid range $[0, 1]$ is too large. As is shown in Fig 3, the generated adversarial example may look overexposed or underexposed. Considering that the lightness range of each image represents its scene characteristics, it is reasonable to

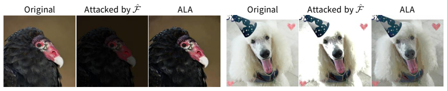


Figure 3. Adversarial examples generated by different filters and their original images.

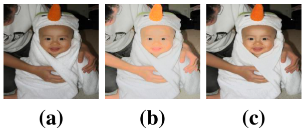


Figure 4. (a) original image, (b) adversarial example filtered by $\mathcal{F}$, (c) regularized example.

set the valid range to $[\mathbf{I}^L_{min}, \mathbf{I}^L_{max}]$. Through modifying the valid range and relaxing the contrast-constraint, we extend the filter $\hat{\mathcal{F}}$ into unrestricted filter $\mathcal{F}$, as shown in Fig 2b.

Algorithm 1 Adversarial Lightness Attack

```
Input: original RGB image $\mathbf{I}_{RGB}$, original label $l$, target model $\mathcal{M}$, number of iterations $N$, learning rate $\alpha$, regularization rate $\beta$, number of segmented pieces $T$
Output: adversarial image $\mathbf{I}'$
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
    $\mathbf{I}^i_L \leftarrow \mathcal{F}_{\theta^i}(\mathbf{I}_L)$
    $\mathbf{I}^i_{Lab} \leftarrow \text{Concatenate}(\mathbf{I}^i_L, \mathbf{I}_a, \mathbf{I}_b)$
    $\mathbf{I}^i_{RGB} \leftarrow \text{LabtoRGB}(\mathbf{I}^i_{Lab})$
    if $\mathcal{M}(\mathbf{I}^i_{RGB}) \neq l$ then
        $\mathbf{I}' \leftarrow \mathbf{I}^i_{RGB}$
    $g \leftarrow \nabla_\theta(\mathcal{L}_{\text{C\&W}}(\mathbf{I}^i_{RGB}, l) + \beta \cdot (-\frac{1}{T}\sum_{j=1}^{T}|\theta_j|))$
    $\theta^{i+1} \leftarrow \theta^i - \alpha \cdot \frac{g}{\|g\|_2}$
return $\mathbf{I}'$
```

3.4. Naturalness-aware Regularization

Simply lifting the restriction of the filter helps to obtain a higher attack success rate. This tends to make the adjacent lightness become the same lightness (*i.e.*, values of parameters θ of these pieces are close to 0). However, the generated images with too much same lightness region are noticeable and suspicious, as shown in Fig 4b. Since the problem shown in the mapping function curve is that some segmented pieces' slopes are close to 0, we just penalize the parameters closing to 0 and the naturalness-aware regularization can be formulated as:$\mathcal{L}_{\text{R}} = -\frac{1}{T}\sum_{j=1}^{T}|\theta_j|$.

3.5. ALA Algorithm with Optimizing Strategy

In Algorithm 1, we show the design of ALA. There are two optimizing strategies: **Random initialization.** By using regularization, the naturalness of adversarial examples has improved (*e.g.*, see Fig 4c). However, the attack success rate dropped. This is mainly due to the repair of lightness distribution by regularization. Thus we randomly initialize the parameters θ into the range $[m, n]$ instead of using parameters of value 1, and finally obtain a high attack success rate with high image-fidelity and strong transferability.

Non-stop attack. Non-stop attack strategy is proposed for better transferability and image fidelity. That is, we don't break the iteration once obtaining a successful example.

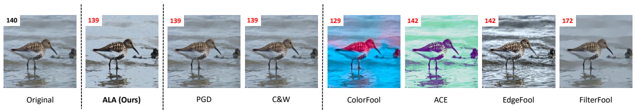


Figure 5. Adversarial examples. The top left corner shows the predicted result (ImageNet index) by MobileNet-v2.

4. Experiment

4.1. Experiment Setup

Datasets. We randomly choose 3 images per class of ImageNet [3] to make up 3,000 images.

Target model. We choose three networks: ResNet50 [7], VGG19 [17], and MobileNet-V2 [12], for our experiments.

Metrics. We assess the image quality from two perspectives with different metrics. To assess the human-perceptual similarity, we use learned perceptual image patch similarity (LPIPS). We use the natural image quality evaluator (NIQE), to partly quantify the quality of images in a non-reference way. Both metrics are better with lower values.

Baseline methods. We choose six adversarial attacks of two types as baselines. For restricted methods, we use PGD [10] with 10 iterations and $\epsilon = 2/255$, and C&W attack [1] with 5×200 iterations and $\kappa = 20$. For unrestricted attack methods, we follow the experiment settings of ACE [21], ColorFool [15], EdgeFool [13], and apply FilterFool [14] with 1,500 iterations and stopping threshold $\tau = 0.006$.

Implementation details. We set $T = 64$ in Eq (1). We set the learning rate $\alpha = 0.5$ and number of iterations $N = 100$. 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 $[m, n] = [-0.2, 0.8]$.

4.2. Image Classification

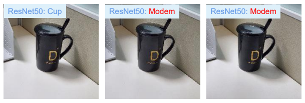
Attack success rate. In Table 1, most attack methods achieve high success rates except for ColorFool, which is a black-box method. EdgeFool and FilterFool obtain just a little higher success rates than ALA, but they cost much more training resources (*e.g.*, 5,000 and 1,500 iterations with both 10.96 GFLOPs) than ACE and ALA (*e.g.*, both are 100 iterations with 8.24 GFLOPs).

Transferability. Among the unrestricted attacks, ACE performs best for transferability. The best two methods in the rest are ALA and FilterFool. Please note 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.

Naturalness. PGD and C&W obtain the highest LPIPS scores because LPIPS assesses the difference between attacked examples and original images, which is hardly influenced by restricted tiny perturbations. As for unrestricted attacks, 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 perfor-

Table 1. Comparison of six attack baselines and our method on ImageNet. It shows the results on three normally trained models: ResNet50, VGG19, and MobileNet-v2. The first three columns are the attack success rates, and the last two columns are image quality metrics LPIPS score, and NIQE scores. We use red, yellow, and blue to mark the first, second, and third best performance among the unrestricted methods.

Target Model	ResNet50					VGG19					MobileNet-v2				
Metrics	ResNet50	VGG19	MobileNet-v2	LPIPS ↓	NIQE ↓	ResNet50	VGG19	MobileNet-v2	LPIPS ↓	NIQE ↓	ResNet50	VGG19	MobileNet-v2	LPIPS ↓	NIQE ↓
PGD	92.87%	4.77%	6.81%	0.005	47.485	4.29%	94.69%	6.51%	0.005	47.432	4.16%	5.11%	98.69%	0.005	47.364
C&W	100.00%	9.73%	10.19%	0.005	47.783	6.26%	100.00%	8.88%	0.004	47.775	7.35%	8.75%	100.00%	0.004	48.022
ColorFool	90.64%	31.45%	36.91%	0.208	44.577	21.43%	91.35%	30.86%	0.205	44.674	18.23%	23.83%	91.98%	0.185	44.885
ACE	96.80%	61.67%	58.04%	0.297	41.603	46.19%	98.92%	54.87%	0.295	41.073	46.73%	58.77%	98.34%	0.297	40.904
EdgeFool	99.27%	34.05%	32.63%	0.127	38.663	23.71%	99.16%	29.25%	0.127	38.267	23.30%	30.12%	99.39%	0.125	38.668
FilterFool	100.00%	43.78%	37.82%	0.111	39.688	24.12%	100.00%	34.44%	0.109	42.071	24.17%	41.18%	100.00%	0.112	38.802
OLF (Ours)	97.53%	44.08%	43.42%	0.124	28.636	24.67%	98.97%	33.84%	0.110	28.472	23.57%	34.10%	99.14%	0.109	28.938



(a) (b) (c)

Figure 6. (a) Original image, (b) ALA image, (c) Physical image.

Table 2. Comparison of the results of the different filters. **Res** denotes restricted filter, and **Unres** denotes unrestricted filter.

T	Filter	Success Rate (%)↑		
		ResNet50	VGG19	MobileNet-v2
64	Res	72.73	82.56	79.83
	Unres	92.46	96.31	95.36

Table 3. Comparison of the results generated on ResNet50 by ALA with/without regularization and initialization. **Reg** means regularization, and **Rd Init.** means random initialization. The **N** and **Y** denotes without/with the operation.

T	Reg/ Rd Init.	Success Rate (%)↑	LPIPS ↓	NIQE ↓
64	N/N	92.46	0.0600	39.857
	Y/N	87.94	**0.0494**	40.048
	Y/Y	**94.56**	0.0667	**36.240**

Table 4. Comparison of the results w/wo random initialization and non-stop attack. **Rd Init.** means random initialization and **Non-stop** means non-stop attack. **N** and **Y** denotes wo/w the operation.

	Rd Init./ Non-stop	ResNet	VGG	MobileNet	LPIPS ↓	NIQE ↓
ResNet	N/N	92.87%	18.97%	18.7%	0.052	40.067
	Y/N	97.53%	39.31%	39.18%	0.102	32.532
	Y/Y	97.53%	44.08%	43.42%	0.124	28.636
VGG	N/N	7.13%	20.97%	12.56%	0.043	41.084
	Y/N	18.04%	99.02%	28.29%	0.087	32.813
	Y/Y	24.64%	98.97%	33.84%	0.110	28.472
MobileNet	N/N	5.21%	21.97%	95.11%	0.038	41.036
	Y/N	16.45%	29.14%	99.14%	0.086	33.114
	Y/Y	23.57%	34.10%	99.14%	0.109	28.938

mance in all cases. In Figure 5, we show the original inputs and the adversarial images generated by different methods. Compared with these unrestricted attack methods, ALA looks like the same scene in different lighting conditions, *e.g*., decreasing the light intensity in the bottom image.

Real-world attack. As is shown in Figure 6, we take a photo of the cup (6a) in real-world and classify this image by ResNet50. Figure 6b shows the ALA-attacked image in digital. To realize the attack in the real world, we simulate the lightness condition reference to ALA and take the photo again. The image (6c) also misleads the ResNet50, which shows the guiding effect of ALA on real-world attacks.

4.3. Ablation Study

Here verify the effectiveness of the proposed unrestricted filter, naturalness-aware regularization, random initialization, and non-stop attack. The dataset follows the Sec 4.1.

Unrestricted filters. We compare the success rate of the adversarial examples generated by different lightness filters, restricted monotonic filter $\hat{\mathcal{F}}$ (Sec 3.1) with bounded parameters θ (*i.e*., $\theta_t \in [1/4, 4]$) and unrestricted filter $\mathcal{F}$ (Sec 3.3). As is shown in Table 2, simply using the restricted filter can not attack these classifiers effectively. Compared with filter $\hat{\mathcal{F}}$, the success rate has increased obviously by about 15% for all settings through using the filter $\mathcal{F}$.

Regularization and random initialization. The first two rows in Table 3 compare the adversarial examples generated by our filter without/with customized regularization (Sec 3.4). We can see that the regularization can obviously improve the image quality (the regularizing rate $\beta = 1$). However, better image quality comes at the cost of success rate. So as mentioned in Sec 3.5, we initiate our filter randomly with the range of $[m, n]$ (the range is $[0, 1]$). Then, as shown in the second and third rows in Table 3, we can both raise the success rate and image quality to some extent.

Random initialization and non-stop attack. As is shown in Table 4, if we apply ALA without the proposed optimizing strategies, the attack performance is not satisfactory. After adding the random initialization, the transferability and some aspects of image fidelity are significantly improved. Furthermore, if we attack the models with the maximum number of iterations, almost all metrics will be better compared to training the filter with the early stop.

5. Acknowledgment

Geguang Pu is supported by National Key Research and Development Program (2020AAA0107800), and Shanghai Collaborative Innovation Center of Trusted Industry Internet Software. This work was supported by the National Research Foundation, Singapore under its AI Singapore Programme (Award No: AISG2-RP-2020-019), Singapore National Cybersecurity R&D Program No. NRF2018NCR-NCR005-0001, National Satellite of Excellence in Trustworthy Software System No. NRF2018NCR-NSOE003-0001, and NRF Investigatorship No. NRF-NRFI06-2020-0001. We gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC).

6. Conclusion

We propose a novel lightness adjustment approach ALA, which generates unrestricted examples with strong transferability and better naturalness compared to existing unrestricted adversarial attacks. In the future, we will combine some other image editing methods [2, 5, 6, 19, 20] with the lightness modification for better attack performance.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 2, 3

[2] Yupeng Cheng, Qing Guo, Felix Juefei-Xu, Shang-Wei Lin, Wei Feng, Weisi Lin, and Yang Liu. Pasadena: Perceptually aware and stealthy adversarial denoise attack. *IEEE Trans. Multim.*, 24:3807–3822, 2022. 4

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3

[4] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning*, pages 1802–1811. PMLR, 2019. 1

[5] Ruijun Gao, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Huazhu Fu, Wei Feng, Yang Liu, and Song Wang. 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 (CVPR)*, pages 2150–2159, June 2022. 4

[6] Qing Guo, Ziyi Cheng, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Yang Liu, and Jianjun Zhao. Learning to adversarially blur visual object tracking. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10819–10828, 2021. 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

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

[9] Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: analysis and improvement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4441–4449, 2018. 1

[10] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2, 3

[11] Daniel L Ruderman, Thomas W Cronin, and Chuan-Chin Chiao. Statistics of cone responses to natural images: implications for visual coding. *JOSA A*, 15(8):2036–2045, 1998. 2

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 3

[13] Ali Shahin Shamsabadi, Changjae Oh, and Andrea Cavallaro. Edgefool: an adversarial image enhancement filter. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1898–1902. IEEE, 2020. 2, 3

[14] Ali Shahin Shamsabadi, Changjae Oh, and Andrea Cavallaro. Semantically adversarial learnable filters. *IEEE Transactions on Image Processing*, 30:8075–8087, 2021. 1, 2, 3

[15] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1160, 2020. 1, 2, 3

[16] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 1

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 3

[18] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1

[19] Binyu Tian, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Xiaohong Li, and Yang Liu. AVA: adversarial vignetting attack against visual recognition. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1046–1053. ijcai.org, 2021. 2, 4

[20] Run Wang, Felix Juefei-Xu, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, and Yang Liu. Amora: Black-box adversarial morphing attack. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1376–1385, New York, NY, USA, 2020. Association for Computing Machinery. 4

[21] Zhengyu Zhao, Zhuoran Liu, and Martha A. Larson. Adversarial robustness against image color transformation within parametric filter space. *CoRR*, abs/2011.06690, 2020. 1, 2, 3