Adversarial Color Enhancement: Generating Unrestricted Adversarial Images by Optimizing a Color Filter

Zhengyu Zhao
z.zhao@cs.ru.nl
Zhuoran Liu
z.liu@cs.ru.nl
Martha Larson
m.larson@cs.ru.nl

Data Science Group,
Institute for Computing and Information Sciences,
Radboud University,
Nijmegen, Netherlands

Abstract

We introduce an approach that enhances images using a color filter in order to create adversarial effects, which fool neural networks into misclassification. Our approach, Adversarial Color Enhancement (ACE), generates unrestricted adversarial images by optimizing the color filter via gradient descent. The novelty of ACE is its incorporation of established practice for image enhancement in a transparent manner. Experimental results validate the white-box adversarial strength and black-box transferability of ACE. A range of examples demonstrates the perceptual quality of images that ACE produces. ACE makes an important contribution to recent work that moves beyond L_p imperceptibility and focuses on unrestricted adversarial modifications that yield large perceptible perturbations, but remain non-suspicious, to the human eye. The future potential of filter-based adversaries is also explored in two directions: guiding ACE with common enhancement practices (e.g., Instagram filters) towards specific attractive image styles and adapting ACE to image semantics. Code is available at https://github.com/ZhengyuZhao/ACE.

1 Introduction

Despite the exceptional success of the Deep Neural Networks (DNNs), recent research has shown that they are remarkably susceptible to *adversarial examples* [44], which are crafted to induce incorrect model predictions. Adversarial image examples have been extensively studied in image classification [6, 16, 22, 33, 34, 35, 49], and also explored in object detection [8, 53], semantic segmentation [2, 50] and image retrieval [31, 46].

A key property of adversarial images that makes them dangerous is that they cause decision conflicts between the model and human annotated labels in a way that is hardly recognizable to human [15, 40]. Most conventional work on adversarial examples has focused on imperceptible additive perturbations, whereby imperceptibility is conventionally measured with the L_p distance between the adversarial images and their clean versions [6, 34, 35]. Later studies proposed to leverage more perception-aligned measurements [10, 32, 48, 52, 54] to

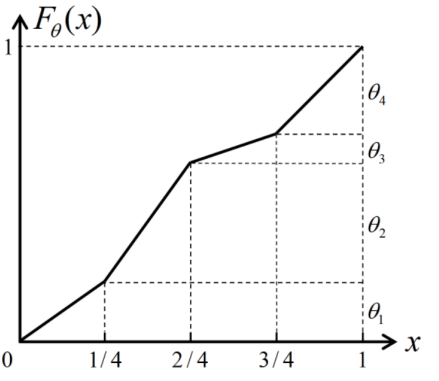

Figure 1: **Left:** Original Images (top) and their adversarial versions (bottom) generated by our Adversarial Color Enhancement (ACE). Additional examples can be found in our GitHub repository. **Right:** Illustration of the color filter adopted in ACE (here $K = 4$ in Equation 3).

address the well known insufficiency of naive L_p norms as perceptual similarity metric [47], but have still focused exclusively on imperceptible perturbations.

Recently, it has been pointed out that when small, imperceptible perturbations were originally introduced by [16], they were intended only to be an abstract, toy example for easy evaluation, and that actually it is hard to find a compelling example that requires imperceptibility in realistic security scenarios [15]. In other words, imposing similarity with respect to an original, clean image is not necessary in real-world threat models. For this reason, recent work has moved beyond small imperceptible perturbations, and started exploiting "unrestricted adversarial examples" [4] that have natural looks even with large, visible perturbations, but remain non-suspicious to the human eye [3, 13, 22]. In general, exploring new types of threat models beyond conventional imperceptible perturbations will provide a more comprehensive understanding of adversarial robustness of the DNNs [49]. And more importantly, relaxing the tight bound on perturbations has been shown to yield practically interesting properties, such as cross-model transferability for black-box adversaries applied in real-world scenarios [3, 39].

Building on these recent developments, we propose a new approach to generating unrestricted adversarial images using a color filter. The approach, called Adversarial Color Enhancement (ACE), introduces non-suspicious perturbations, with minimal impact on image quality, as shown in Figure 1 (left). Although previous work [4] has pointed out that common enhancement practices (e.g., Instagram filters) can degrade the performance of the automatic geo-location estimation, until now, no research has focused on the optimization aspect of exploiting image filters to create adversarial images. Our approach makes use of recent advances in automatic image retouching based on differentiable approximation of commonly-used image filters [11, 19]. In sum, this paper makes the following contributions:

- We explore the vulnerability of the DNNs to commonly-used image filters, and specifically propose Adversarial Color Enhancement (ACE), an approach to generating unrestricted adversarial images by optimizing a differentiable color filter.
- Experimental results demonstrate ACE achieves a better trade-off between the adversarial strength and perceptual quality of the filtered images than other state-of-the-art

methods, implying a stronger black-box adversary for real-world applications.
- We explore two potential ways to further improve ACE on image quality: 1) using widely-used enhancement practices (e.g., Instagram filters) as guidance to specified attractive image styles, and 2) leveraging regional semantic information.

2 Related Work

Differentiable Image Filters. The state of the art for automatic photo retouching mainly uses supervised learning to determine editing parameters via gradient descent, in order to achieve specific image appearances. Most approaches [2, 14, 21, 51, 56] utilize DNNs for the parameterization of the editing process, but inevitably they suffer from high computational cost, fixed image resolution, and more importantly, a lack of interpretability. For this reason, some recent work [11, 19] has proposed to rely on intuitively meaningful edits that are represented by conventional post-processing operations, i.e., image filters, to make the automatic process more understandable to users. Moreover, such methods have much fewer parameters to optimize, and can be applied resolution-independently.

Problem Formulation. A neural network can be denoted as a function $F(x) = y$ that outputs $y \in \mathbb{R}^m$ for an image $x \in \mathbb{R}^n$. Here we focus on the widely-used DNN classifier with a softmax function, which expresses the output y as a probability distribution, i.e., $0 \leq y_i \leq 1$ and $y_1 + \cdots + y_m = 1$. The final predicted label l for x is accordingly obtained by $l = \text{argmax}_i\, y_i$. An adversary aims to induce a misclassification of a DNN classifier $F(x)$ through modifying the original image x into x' such that $F(x') \neq y$.

Restricted Adversary with Imperceptible Perturbations. As mentioned in Section 1, in order to make the modification unrecognizable, most existing work forces the adversarial image x' to be visually close to its original image x with respect to specific distance measurements. The conventional solution is L_p distance (typically L_∞ [5, 16, 26, 33] and L_2 [5, 34, 37, 44], but also L_1 [6] and L_0 [35, 43]). The earliest work in this direction [44] proposed to jointly optimize misclassification with cross-entropy loss and the L_2 distance by solving a box-constrained optimization with the L-BFGS method [30]. The C&W method [5] followed a similar idea, but replaced the cross-entropy loss with another specially designed loss function, namely, the differences between the pre-softmax logits. Moreover, a new variable was introduced to eliminate the box constraint. The method can be expressed as:

$$\begin{aligned} &\underset{w}{\text{minimize}} \;\; \|x' - x\|_2^2 + \lambda f(x'), \\ \text{where} \;\; &f(x') = \max(Z(x')_l - \max\{Z(x')_i : i \neq l\}, -\kappa), \\ \text{and} \;\; &x' = \frac{1}{2}(\tanh(\text{arctanh}(x) + w) + 1), \end{aligned} \tag{1}$$

where $f(\cdot)$ is the new loss function, w is the new variable, and $Z(x')_i$ is the logit with respect to the i-th class given the intermediate modified image x'. The parameter κ is applied to control the confidence level of the misclassification.

This joint optimization is straightforward but suffers from high computational cost due to the need for line search to optimize λ. For this reason, other methods [12, 16, 26, 33, 37] instead rely on Projected Gradient Descent (PGD) to restrict the perturbations with a small L_p norm bound, ε. Specifically, the fast gradient sign method (FGSM) [16] was designed to succeed within only one step and was extended by [12, 26, 33, 37] to exploit finer gradient

information with multiple iterations. The iterative approach can be formulated as:

$$x'_0 = x, \;\; x'_k = x'_{k-1} + \alpha \cdot \; \text{sign}(\nabla_x J(x'_{k-1}, l)), \tag{2}$$

where α denotes the step size in each iteration. The generated adversarial perturbations will be clipped to satisfy the L_∞ bound. A generalization of this formulation to the L_2 norm can be achieved by replacing the $\text{sign}(\cdot)$ with a normalization operation [34, 37]. L_0 and L_1-bounded adversarial images were also studied [6, 35, 43], but not widely adopted since the resulting sparse perturbations are not stable in practice.

Recently, there have also been several attempts to address the limitations of naive L_p by using more perception-aligned solutions for measuring similarity. A straightforward way is by incorporating existing metrics, such as Structural SIMilarity (SSIM) [38], Wasserstein distance [43], and the perceptual color metric CIEDE2000 [24]. Other methods [10, 32, 52] adapted the L_p measurements to the textural properties of the image, i.e., hiding perturbations in image regions with high visual variation. Local pixel displacement was also explored [1, 49]. In general, these solutions yield a better trade-off between adversarial strength and imperceptibility than conventional L_p methods.

Unrestricted Adversaries with Large yet Non-Suspicious Modifications. Due to the assumption of imperceptible perturbations, now considered unrealistic, as mentioned in Section 1, recent work has started to pursue non-suspicious adversarial images with large perturbations, which make more sense in practical use scenarios. Common approaches to creating such unrestricted adversarial images can be divided into three categories: geometric transformation, semantic manipulation, and color modification. The geometric transformation method penalizes image differences with respect to small rotations and translations of the image [13]. Semantic manipulation has been so far mainly studied in the domain of face recognition, where the perturbation is optimized with respect to specific semantic attribute(s), such as colors of skin and extent of makeup [22, 36, 41].

Existing colorization-based work has explored uniform color transformation [18, 28, 39] and automatic colorization [3]. Specifically, the early method [18] randomly adjusts the hue values of each image pixel to search for possible adversarial images. The ColorFool method [39] improves on [18] by imposing semantic-aware norm constraints for better image quality, but still relies on costly random search. The ReColorAdv method [28] optimizes color transformation over a discretely parameterized color space with post-interpolation and regularization on local uniformity, and impose L_∞ bounds on the perturbations. The cAdv method [3] takes a different route, optimizing a pre-trained automatic colorization model to re-colorize the gray-scale version of the original image. It increases the computational overhead due to the huge number of parameters in the deep colorization model, and also has been shown to cause abnormal color stains (see examples in [3] and our Figure 3).

Our ACE falls into the colorization category but is markedly different from existing approaches. Specifically, ACE creates adversarial images by optimizing with gradient information, and, in this way, is fundamentally different from the random search-based approaches in [18, 39]. In Section 4, we also show that our gradient-based ACE outperforms its alternative with random search. Our color filter is simpler and more transparent than the deep colorization model in [3]. Compared with [28], our ACE enjoys a more elegant and continuous formulation. Experimental results in Section 4 demonstrate that our ACE outperforms these approaches in both adversarial strength and image quality.

3 Adversarial Color Enhancement (ACE)

This section describes our proposed Adversarial Color Enhancement (ACE), which generates visually realistic adversarial filtered images based on a commonly-used color filter. Specifically, we adopt the differentiable approximation in [19] to parameterize the color filter by a monotonic piecewise-linear mapping function with totally K pieces:

$$F_\theta(x_k) = \sum_{i=1}^{k-1} \theta_i + (K \cdot x_k - (k-1)) \cdot \theta_k, \quad \text{s.t. } 0 \leq \theta_i \leq 1 \text{ and } \sum_i \theta_i = 1, \tag{3}$$

where x_k denotes any image pixel whose value falls into the k-th piece of the mapping function, and $F_\theta(x_k)$ is its corresponding output after filtering. An example of this function with four pieces ($K = 4$) is illustrated in Figure 1 (right).

Note that we are not optimizing in the pixel space but in the latent space of filter parameters, and the three RGB channels are operated on in parallel. The parameters θ (K in total) can be optimized via gradient descent to achieve a specific objective. Obviously, an image will remain unchanged ($F_\theta(x) = x$) when all the parameters are equal to $1/K$. As a result, we propose to control over the adjustment by imposing constraints on the distance between each parameter and its initial value $1/K$. The misclassification objective and the proposed constraints on the parameters will be jointly optimized with a balance factor λ, expressed as:

$$\underset{\theta}{\text{minimize}}\ f(F_\theta(x)) + \lambda \cdot \sum_i (\theta_i - 1/K)^2, \tag{4}$$

where $f(\cdot)$ is the C&W loss on logit differences in Equation 1.

4 Experiments

We evaluate our ACE in two different tasks: object classification and scene recognition, and consider the following two datasets. **ImageNet-Compatible Dataset** consists of 6000 images associated with ImageNet class labels, and has been used in the NIPS 2017 Competition on Adversarial Attacks and Defenses [27]. Here we use its development set containing 1000 images. **Private Scene Dataset** was introduced by the MediaEval Pixel Privacy task [29],

	→Alex	→R50	→V19	→D121	→Inc3
Alex	99.9	6.26	7.10	6.85	2.08
R50	48.50	98.3	15.83	13.21	5.96
V19	39.52	11.29	98.5	10.65	10.30
D121	46.50	18.51	15.16	98.4	5.61
Inc3	41.12	16.42	14.87	12.35	93.2

	→Alex	→R18	→R50
Alex	100.0	16.40	12.30
R18	48.52	99.2	22.43
R50	48.98	30.47	99.0

Table 1: White-box success rates (diagonal) and black-box transferability of our ACE in ImageNet classification (left) and private scene recognition (right). Considered models: AlexNet (Alex), ResNet18 (R18), ResNet50 (R50), VGG19 (V19), DenseNet121 (D121), and Inception-V3 (Inc3). The success rates are measured with respect to the models in the columns when applying ACE on the models in the rows.

which aims to develop image modification techniques that help to protect users against automatic inference of privacy-sensitive scene information. It contains 600 images with 60 privacy-sensitive scene categories, selected from the Places365 dataset [55]. For the ImageNet task, we consider five distinct classifiers that are pre-trained on ImageNet: AlexNet [25], ResNet50 [17], VGG19 [42], DenseNet121 [20], and Inception-V3 [45]. For the scene task, we consider AlexNet, ResNet18, and ResNet50 pre-trained on the Places365 dataset.

ACE is optimized using Adam [23] with a learning rate of 0.01, under a maximum budget of 500 iterations. Early stopping is triggered when the optimization is no longer making progress as implemented in [6, 37]. If not mentioned specifically, ACE is implemented with the optimal settings, $K = 64$ and $\lambda = 5$. When executed on a single NVIDIA Tesla P100 GPU with 12 GB of memory and with 40 batches of 25 image samples, the implementation in ImageNet takes about 2 seconds per image. Table 1 shows that our ACE can achieve high white-box success rates and have good cross-model transferability. It can also be observed that models with more sophisticated architecture are generally harder to fool in the white-box case, and transferring from a sophisticated architecture to a simple one is easier than the other way around. Note that the transferability is calculated on images for which the prediction of both the models involved is the same.

4.1 Comparisons on Adversarial Strength and Image Quality

We further compare ACE with the following gradient-based baseline methods in terms of adversarial strength and image quality, in the ImageNet task:
FGSM [16] with a L_∞ norm bound $\varepsilon = 2/255$ for ensuring imperceptibility.
BIM [26] with a L_∞ norm bound $\varepsilon = 2/255$, and 10 iterations of gradient descent.
C&W [6] optimized on L_2 with fewer iterations and higher confidence level (iters=3×100 and $\kappa = 40$) than usual to yield larger perturbations for stronger adversarial effects.
ReColorAdv [28] (Unrestricted) with $\varepsilon = 16/255$ and lr=0.001 as in [28], and another version allowing larger perturbations ($\varepsilon = 51/255$ and lr=0.005), denoted as ReColorAdv$^+$.
cAdv [8] (Unrestricted) with the settings leading to optimal color realism ($k = 8$). Note that cAdv can only produce adversarial images sized 224×224 due to the fixed output resolution of its pre-trained deep colorization model.

We adopt Inception-V3 as the white-box model because it is the official model used in the NIPS 2017 Competition. As shown in Table 2, ACE can consistently achieve better

	L_p Norm			Success Rate				
	L_0 (%)	L_2	L_∞	Inc3	→Alex	→R50	→V19	→D121
FGSM [16]	49.34	4.05	2.00	78.10	7.84	5.40	5.74	5.50
BIM [26]	39.23	3.09	2.00	99.1	8.16	4.95	6.44	4.71
C&W [6]	29.06	3.00	15.66	99.6	8.16	4.72	6.79	4.38
ReColorAdv [28]	70.81	18.87	64.00	79.3	9.76	4.50	3.40	2.58
ReColorAdv$^+$ [28]	82.50	47.53	97.21	89.2	31.20	15.64	13.58	10.77
cAdv [8]	41.42	20.54	116.15	91.8	30.08	11.25	11.01	**13.47**
Our ACE	42.99	40.61	45.98	93.2	**41.12**	**16.42**	**14.87**	12.35

Table 2: White-box success rates and black-box transferability of our ACE compared with other baselines. L_0 is the proportion of the perturbed pixels and L_∞ is shown in [0,255]. Here the Inception-V3 is used as the target white-box model.

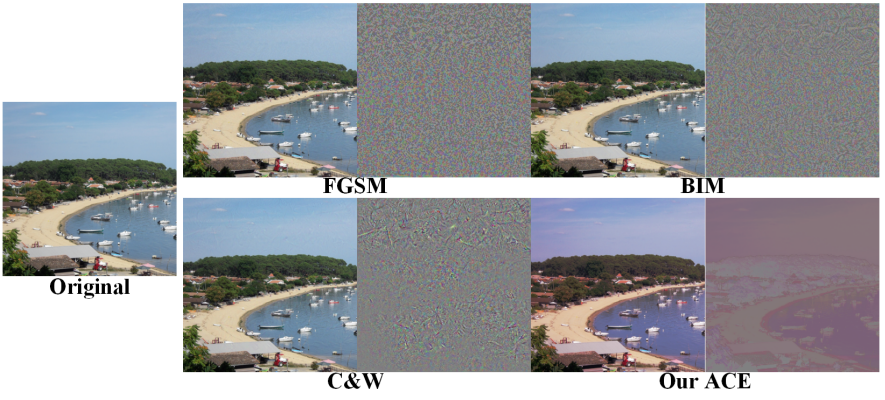

Figure 2: Adversarial images and perturbations by our ACE and L_p methods FGSM [16], BIM [26] and C&W [5]. ACE yields more natural appearances without abnormal patterns.

Figure 3: Examples from three different unrestricted methods. From left to right: original image, adversarial images generated by cAdv [3], ReColorAdv$^+$ [28], and our ACE. ACE yields more acceptable images with smooth appearances and fewer artifacts.

transferability than conventional L_p methods, while not introducing visually suspicious noisy patterns (see Figure 2). Iterative L_p methods (BIM and C&W) could achieve the strongest white-box adversarial effects by fully leveraging the gradient information, but these effects are less generalizable to other unseen models, i.e., worse transferability. Among the unrestricted methods, our ACE achieves the highest white-box success rates and overall best transferability, while yielding smooth adjustment without abnormal colorization artifacts (see Figure 3). Such smoothness is also reflected in L_∞ norms that are lower than other unrestricted methods, meaning that ACE tends to avoid excessive local color changes.

4.2 Ablation Study

Hyperparameters. Figure 4 (left) shows the success rates of ACE with a different number of pieces K under different factor λ values used for balancing the two loss terms in the joint optimization. We can observe that increasing K slightly improves the performance

Figure 4: **Left:** White-box success rates of ACE when varying the number of pieces K with four different λ values. **Right:** Adversarial images with varied λ. For each example, from left to right: original image, adversarial images with λ =10, 5 and 0. Too large a λ makes ACE not reach its full potential, while setting $\lambda = 0$ may cause unrealistic colorization.

by expanding the action space of the adversary. Moreover, increasing K allows more fine-grained color adjustment in the images that have rich colors. It should also be noted that using more pieces means more computational cost during the optimization.

On the other hand, relaxing the constraints by decreasing λ gives the adversary larger action space, leading to higher success rates. However, completely removing the constraints ($\lambda = 0$) will lead to unrealistic image appearances, as can be observed in Figure 4 (right). Specifically, in this paper, we use $K = 64$ and $\lambda = 5$ as optimal settings for a good trade-off.

Gradient Descent vs. Random Search. We compare our gradient-based ACE with a random search-based implementation using the same color filter. In this case, the parameters will be updated with gradient information in our ACE, while being uniformly sampled from the valid range [0,1] for random search.

Figure 5 shows their white-box success rates as a function of iterations. For random search, we repeat several times and got almost the same results. We can observe that our ACE consistently outperforms random search, even with far fewer iterations. Moreover, ACE gradually improves as the number of parameters K increases, indicating that it can benefit from the expanded action space for more fine-grained color adjustment. In contrast, random

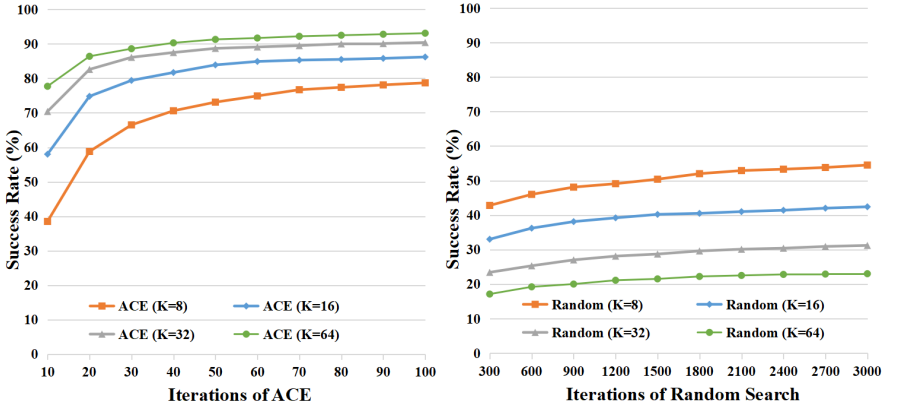

Figure 5: White-box success rates for our ACE (left) and random search (right) as a function of maximum allowed iterations. Note that the scales on the two x-axes are not the same.

Figure 6: Adversarial filtered images by ACE with guidance of specific image filters. Two Instagram filters are considered: Hefe (top) and Gotham (bottom). For each, from left to right: original image, Instagram filtered image and adversarial filtered image by ACE.

search becomes increasingly worse since the potentially successful adversarial samples can no longer be feasibly found in the exponentially expanded parameter space. For comparison we mention that, the grid search-based ColorFool [39] has a success rate of 64.6% (65.4%) with 1000 (1500) iterations when testing on our dataset with the Inception-V3 as the white-box model.

5 ACE Extensions

Adaptation on Image Styles. Previous work [9] has pointed out that popular image enhancement practices can potentially degrade automatic inference. Despite our ACE in Equation 4 achieving visually acceptable results, it is not directly optimized towards enhancing image quality. Therefore, we explore the possibility to guide ACE towards achieving quality enhancement in addition to the adversarial effects. Specifically, we propose to optimize the adversarial image towards specific attractive styles that were obtained by using Instagram filters. Accordingly, the optimization objective is adjusted to:

$$\underset{\theta}{\text{minimize}}\ f(F_\theta(x)) + \lambda \cdot \|F_\theta(x) - x_t\|_2^2, \tag{5}$$

where x_t denotes the target Instagram filtered image, and the final adversarial filtered image F_θ is therefore guided to have similar appearances with x_t by minimizing their distance. As shown in Figure 6, this adaptation of ACE can successfully enhance the image by mimicking the effects of Instagram filters.

Adaptation on Semantics. ACE treats all the image pixels that have the same values in the same way. Inspired by previous work [3, 39], we show that semantically adapting ACE could better maintain image quality by hiding large perturbations in the semantic regions

Figure 7: Adapting ACE on different semantics. Original image (left), successful adversary with stronger constraints for cat but weaker for blanket (middle), or vice versa (right). The middle case yields more realistic results since blanket naturally occurs with various colors.

that remain realistic with various colors. Specifically, Equation 4 is adapted to:

$$\min_{\theta} \sum_{n} [J(F_{\theta_n}(x \cdot M_n)) + \lambda \sum_{i} w_n \cdot (\theta_{n_i} - 1/K)^2], \text{ s.t. } \sum_{n} w_n = 1, \tag{6}$$

where w_n is the weight for the n-th filter, $F^n_\theta(\cdot)$, which is optimized independently for a specific semantic region given its mask M_n obtained by a semantic segmentation method [24]. As shown in Figure 7, this adaptation avoids raising the sense of unrealistic colorization, leading to improved image quality.

6 Conclusion

We have proposed Adversarial Color Enhancement (ACE), an approach to generating unrestricted adversarial images by optimizing a color filter via gradient descent. ACE has been shown to produce realistic filtered images with good transferability, which results in strong real-world black-box adversaries. We also present two potential ways to improve ACE in terms of image quality by guiding it with specific attractive image styles or adapting it to regional semantics.

In the current ACE, the single hyperparameter of the filter (K in Equation 3) is per-fixed for all images without considering their individual properties. Since natural images would differ in the range and complexity of their contained colors, adaptive strategies would be worth exploring in order to yield more suitable modifications. For example, images with most pixels concentrated in a certain color range should have larger action space in that range than those with more uniform color distribution. It would also be interesting to carry out user study on the visual quality of the images generated by ACE. On the other side, developing defenses against the proposed adversarial color filtering is necessary to make current neural networks more robust, based on either adversarial training or algorithms for detecting adversarial modifications by ACE.

Acknowledgement

This work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

References

[1] Rima Alaifari, Giovanni S Alberti, and Tandri Gauksson. ADef: an iterative algorithm to construct adversarial deformations. In *ICLR*, 2019.

[2] Anurag Arnab, Ondrej Miksik, and Philip HS Torr. On the robustness of semantic segmentation models to adversarial attacks. In *CVPR*, 2018.

[3] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. In *ICLR*, 2020.

[4] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. In *arXiv preprint*, 2018.

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE S&P*, 2017.

[6] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In *AAAI*, 2018.

[7] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *ICCV*, 2017.

[8] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Polo Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *ECML PKDD*, 2018.

[9] Jaeyoung Choi, Martha Larson, Xinchao Li, Kevin Li, Gerald Friedland, and Alan Hanjalic. The geo-privacy bonus of popular photo enhancements. In *ICMR*, 2017.

[10] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *ICCV*, 2019.

[11] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Aesthetic-driven image enhancement by adversarial learning. In *ACM MM*, 2018.

[12] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *CVPR*, 2018.

[13] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *ICML*, 2019.

[14] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, 36(4):1–12, 2017.

[15] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. In *arXiv preprint*, 2018.

[16] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[18] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *CVPR Workshops*, 2018.

[19] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics*, 37(2): 26, 2018.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[22] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *ICCV*, 2019.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[24] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019.

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[26] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR*, 2017.

[27] Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*, 2018.

[28] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *NeurIPS*, 2019.

[29] Martha Larson, Zhuoran Liu, Simon Brugman, and Zhengyu Zhao. Pixel privacy: Increasing image appeal while blocking automatic inference of sensitive scene information. In *MediaEval Multimedia Benchmark Workshop*, 2018.

[30] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3):503–528, 1989.

[31] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. Who's afraid of adversarial queries? The impact of image modifications on content-based image retrieval. In *ICMR*, 2019.

[32] Bo Luo, Yannan Liu, Lingxiao Wei, and Qiang Xu. Towards imperceptible and robust adversarial example attacks against neural networks. In *AAAI*, 2018.

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[34] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016.

[35] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE EuroS&P*, 2016.

[36] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditional image editing. In *ECCV*, 2020.

[37] Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l_2 adversarial attacks and defenses. In *CVPR*, 2019.

[38] Andras Rozsa, Ethan M Rudd, and Terrance E Boult. Adversarial diversity and hard positive generation. In *CVPR Workshops*, 2016.

[39] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. ColorFool: Semantic adversarial colorization. In *CVPR*, 2020.

[40] Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. On the suitability of L_p-norms for creating and preventing adversarial examples. In *CVPR Workshops*, 2018.

[41] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security*, 2019.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[43] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23 (5):828–841, 2019.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[45] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016.

[46] Giorgos Tolias, Filip Radenovic, and Ondrej Chum. Targeted mismatch adversarial attack: Query with a flower to retrieve the tower. In *ICCV*, 2019.

[47] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[48] Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *ICML*, 2019.

[49] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *ICLR*, 2018.

[50] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *ICCV*, 2017.

[51] Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics*, 35(2): 1–15, 2016.

[52] Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. Smooth adversarial examples. *EURASIP Journal on Information Security*, 2020.

[53] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *ACM CCS*, 2019.

[54] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *CVPR*, 2020.

[55] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.