# Fooling Blind Image Quality Assessment by Optimizing a Human-Understandable Color Filter

Zhengyu Zhao
Radboud University, Netherlands
z.zhao@cs.ru.nl

## ABSTRACT

This paper presents the submission of our RU-DS team to the Pixel Privacy Task 2020. We propose to fool the blind image quality assessment model by transforming images based on optimizing a human-understandable color filter. In contrast to the common work that relies on small, $L_p$-bounded additive pixel perturbations, our approach yields large yet smooth perturbations. Experimental results demonstrate that in the specific context of this task, our approach is able to achieve strong adversarial effects, but has to sacrifice the image appeal.

## 1 INTRODUCTION

High-quality images shared online can be misappropriated for promotional goals. The Pixel Privacy Task [15] this year is focused on developing adversarial techniques to decrease the predicted quality scores of an automatic Blind Image Quality Assessment (BIQA) model [10], which effectively camouflages images from being promoted. A key requirement of such adversaries is that the adversarial image should remain its original quality or become more appealing to the human eye. Conventional work on generating adversarial images has been focused on small additive perturbations, mostly bounded by $L_p$ distance [2, 3, 9, 16], or other more visual-perception-aligned metrics [4, 18, 19, 21]. In this way, the adversarial image is only designed to maintain its original appearance as much as possible, instead of enhancing the image appeal.

In contrast, recent studies [1, 6, 7, 13, 14, 17, 20] have started to explore *non-suspicious adversarial images* that accommodate larger perturbations without arousing suspicion because they transform groups of pixels along dimensions consistent with human interpretation of images. Among them, the Adversarial Color Enhancement (ACE) [20] can simultaneously achieve the adversarial effects and image enhancement by optimizing a human-understandable parametric color filter. Its effectiveness has been originally validated in the domain of image classification and segmentation.

One may argue that it is easier to separately conduct the optimization for adversarial effects and image enhancement. However, we note that the joint optimization can yield larger perturbations that enjoy two important practical properties: robustness against common image processing operations and transferability to a black-box target model [1, 17, 20]. In this paper, specifically, we will explore the usefulness of ACE in this Pixel Privacy Task for decreasing the BIQA score while enhancing the image appeal.


*MediaEval'20, December 14-15 2020, Online*

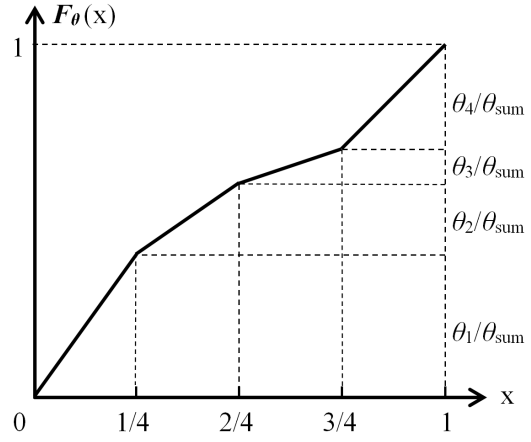


**Figure 1: A 4-piece color filter in ACE ( from [20]).**

## 2 APPROACH

In this section, we firstly recall the general formulation of Adversarial Color Enhancement (ACE) as proposed by [20], and then present the modifications for applying it in our specific Pixel Privacy Task.

### 2.1 Parametric Image Enhancement

Most advanced automatic photo enhancement algorithms have proposed to parameterize the image editing process by the DNNs, which however suffers from high computational cost and low interpretability [8, 12, 22]. In contrast, recent work [5, 11] has proposed to parameterize the process as human-understandable image filters. Such methods have far fewer parameters to optimize, and can be applied independently of the image resolution.

Specifically, ACE adopts the approximation of the color filter in [11], which is formulated as a simple monotonic piecewise-linear mapping function:

$$F_{\boldsymbol{\theta}}(x_k) = \sum_{i=1}^{k-1} \frac{\theta_i}{\theta_{\text{sum}}} + (K \cdot x_k - (k-1)) \cdot \frac{\theta_k}{\theta_{\text{sum}}}, \qquad \theta_{\text{sum}} = \sum_{k=1}^{K} \theta_k, \tag{1}$$

where $K$ demotes the total number of pieces. In this case, an input image pixel $x_k$ falling in the $k$-th piece will be filtered using the parameter $\theta_k$, and $F_{\boldsymbol{\theta}}(x_k)$ is its corresponding output. By doing this, pixels with similar colors will be filtered with the same parameter, leading to smooth color transformation. Specifically, the three RGB channels are processed independently. An example of this function with four pieces ($K = 4$) is illustrated in Fig. 1.

**Table 1: Detailed settings of our five runs.**

| Runs | Methods | Parameters |
|---|---|---|
| 1 | ACE-PGD | $K = 64$, $\epsilon = 16$, and iters. = 20 |
| 2 | ACE-PGD | $K = 64$, $\epsilon = 32$, and iters. = 20 |
| 3 | ACE-PGD | $K = 256$, $\epsilon = 16$, and iters. = 20 |
| 4 | ACE-PGD | $K = 256$, $\epsilon = 64$, and iters. = 20 |
| 5 | ACE-Ins | $K = 64$, $\lambda = 0.01$, and iters. = 100 |

## 2.2 Adversarial Color Enhancement

ACE generates non-suspicious adversarial images by iteratively updating the parameters of the color filter defined in Eq. 1, in contrast to the conventional attacks that are operated in the raw pixel space.

There are two methods to constrain the color transformation strength. The first method imposes adjustable bounds on the filter parameters, formulated as:

$$\min_{\boldsymbol{\theta}} L_{adv}(F_{\boldsymbol{\theta}}(\boldsymbol{x})), \text{s.t. } 1 \leq \| \frac{\boldsymbol{\theta}}{\boldsymbol{\theta}_0} \|_\infty \leq \epsilon, \tag{2}$$

where $\boldsymbol{\theta}_0$ denotes the initial parameters, equaling to $\mathbb{1}^K/K$. The adversarial loss, $L_{adv}$, adopts the specific logit loss from the the well-known C&W method [2]. Note that this parameter bound is not necessarily to tight as in the $L_p$ methods, since the color filtering can inherently guarantee the uniformity of the image transformation even when the perturbations are large. This bounded variant of ACE is referred to as ACE-PGD.

The second method guides the transformation towards specific appealing color styles, in addition to achieving the adversarial effects. To this end, additional guidance from common enhancement practices is incorporated into the adversarial optimization. Specifically, the targeted appealing color styles are obtained by using Instagram filters, and the optimization can be formulated as:

$$\min_{\boldsymbol{\theta}} L_{adv}(F_{\boldsymbol{\theta}}(\boldsymbol{x})) + \lambda \cdot \|F_{\boldsymbol{\theta}}(\boldsymbol{x}) - \boldsymbol{x}_{\text{ins}}\|_2^2, \tag{3}$$

where $\boldsymbol{x}_{\text{ins}}$ denotes the targeted Instagram filtered image with a specific color style. This variant of ACE is referred to as ACE-Ins. One popular Instagram filter style, Nashville, is considered in our submitted runs, and the implementation is automated using the GIMP toolkit with the Instagram Effects Plugins[1].

In the context of fooling BIQA, the $L_{adv}$ is formulated as:

$$L_{adv} = \max\{\text{BIQA}(F_{\boldsymbol{\theta}}(\boldsymbol{x})) - C, 0\}, \tag{4}$$

where the target score can be set by adjusting $C$. Specifically, we set $C$ a bit lower than the standard target, 50, to make sure the adversarial effects could remain after the JPEG compression.

# 3 RESULTS AND ANALYSIS

In total, we submitted five runs. We tried different parameters of ACE-PGD for the first four runs, and used ACE-Ins for the last run.

As can be seen from Table 1, all the five runs effectively decrease the model accuracy to a level below 50%. Specifically, as expected, higher $K = 4$ and $\epsilon$ lead to stronger adversarial effects. In addition, we find that the results before and after the JPEG compression remain similar, suggesting that our approach is stale against compression.

[1]https://www.marcocrippa.it/page/gimp_instagram.php.

**Table 2: Evaluation results of our five runs. The accuracy (%) is calculated over all the 550 test images, which are compressed with JPEG 90 before evaluation. The number of times selected as "Top-3" most appealing among the total 13 qualified runs is evaluated by user study with 7 people on 20 representative images that have the largest BIQA score variance. The maximum number is 140.**

| Runs | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Acc before JPEG | 48.00 | 33.27 | 50.00 | 21.82 | 35.09 |
| Acc after JPEG | 45.27 | 33.45 | 47.45 | 22.55 | 44.91 |
| Number of Top-3 | 2 | 7 | 6 | 4 | 7 |

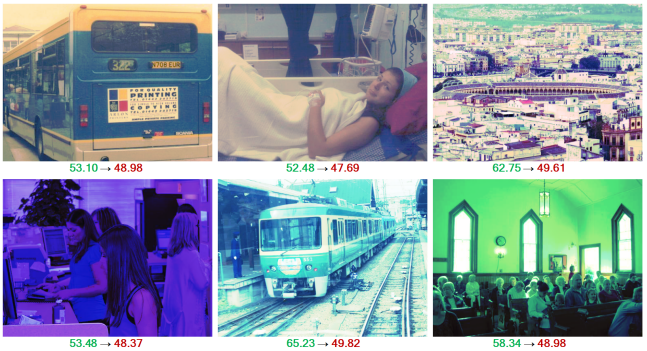


**Figure 2: Adversarial images achieved by our approach with the original and decreased scores. The top row shows the examples with relatively high appeal and the bottom row shows the failed examples with low appeal.**

However, the human evaluation results on the 20 selected images are not satisfying. It implies that the BIQA model is more stable against the interference of smooth modifications, such as ACE, than the classification models. Specifically, we notice that ACE-Ins fails to drive the image into a target appealing style since the optimization has to be focused on lowering the score. This may be because the quality assessment model tends to rely on high-frequency features but the ImageNet classifier learns both low-frequency (e.g. shape) and high-frequency (e.g. textures) features. This makes the quality assessment model more robust against the low-frequency perturbations by our ACE. We will explore this in more depth for the future work.

Figure 2 visualizes the successful adversarial examples with high and low appeal. We can observe that ACE can yield good image examples with filtering-like styles, but the bad examples suffer from over-colorization effects.

# ACKNOWLEDGMENTS

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

## REFERENCES


[1] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. 2020. Unrestricted Adversarial Examples via Semantic Manipulation. In *ICLR*.

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

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

[4] Francesco Croce and Matthias Hein. 2019. Sparse and Imperceivable Adversarial Attacks. In *ICCV*.

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

[6] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. 2019. Exploring the Landscape of Spatial Robustness. In *ICML*.

[7] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning models. In *CVPR*.

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

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

[10] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE TIP* 29 (2020), 4041–4056.

[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* 37, 2 (2018), 26.

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

[13] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. 2019. Semantic Adversarial Attacks: Parametric Transformations That Fool Deep Classifiers. In *ICCV*.

[14] Cassidy Laidlaw and Soheil Feizi. 2019. Functional Adversarial Attacks. In *NeurIPS*.

[15] Zhuoran Liu, Zhengyu Zhao, Martha Larson, and Laurent Amsaleg. 2020. Exploring Quality Camouflage for Social Images. In *Working Notes Proceedings of the MediaEval Workshop*.

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

[17] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. 2020. ColorFool: Semantic Adversarial Colorization. In *CVPR*.

[18] Eric Wong, Frank Schmidt, and Zico Kolter. 2019. Wasserstein Adversarial Examples via Projected Sinkhorn Iterations. In *ICML*.

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

[20] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. 2020. Adversarial Robustness Against Image Color Transformation within Parametric Filter Space. In *arXiv preprint arXiv:2011.06690*.

[21] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. 2020. Towards Large yet Imperceptible Adversarial Image Perturbations with Perceptual Color Distance. In *CVPR*.

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