Exploring Quality Camouflage for Social Images

Zhuoran Liu, Zhengyu Zhao, Martha Larson, Laurent Amsaleg

Exploring Quality Camouflage for Social Images

Zhuoran Liu, Zhengyu Zhao, Martha Larson
Radboud University, Netherlands
{z.liu,z.zhao,m.larson}@cs.ru.nl

Laurent Amsaleg
CNRS-IRISA, France
laurent.amsaleg@irisa.fr

ABSTRACT

Social images can be misused in ways not anticipated or intended by the people who share them online. In particular, high-quality images can be driven to unwanted prominence by search engines or used to train unscrupulous AI. The risk of misuse can be reduced if photos can evade quality filtering, which is commonly carried out by automatic Blind Image Quality Assessment (BIQA) algorithms. The Pixel Privacy Task benchmarks privacy-protective approaches that shield images against unethical computer vision algorithms. In the 2020 task, participants are asked to develop quality camouflage methods that can effectively decrease the BIQA score of high-quality images while maintaining image appeal. The camouflage should not damage the image from the point of view of the user: it needs to be either imperceptible, or else to enhance the image visibly, to the human eye.

1 INTRODUCTION

Social images shared online can be misused, causing distress and harm to the people who share them. As computer vision algorithms continue to improve, their potential for misuse grows as well. There is a need for technology that can counter the ability of malicious or heedless actors to easily abuse the power of computer vision.

The Pixel Privacy Task at the MediaEval Multimedia Evaluation Benchmark aims to address this need. The objective of the task is to support the development of approaches that can help protect users' multimedia content. The first two years of the Pixel Privacy Task [15, 17] focused on privacy sensitive information depicted in social images of scenes. Task participants were asked to develop image transformation approaches that could mislead scene recognition algorithms, preventing them from predicting the scene classes for images depicting privacy sensitive scenes.

However, semantic class is not the only aspect of images associated with the threat of computer vision algorithms. This year, the Pixel Privacy Task moves beyond semantics to look at hiding the quality of images. We call this task *quality camouflage*. Specifically the 2020 Pixel Privacy Task asks task participants to create algorithms that cause a Blind Image Quality Assessment (BIQA) algorithm to predict that an image has a lower quality than it actually has. Participants are provided with high-quality images, and must transform them so that they are classified by BIQA to be low-quality images. At the same time, the images must maintain their appeal to the human eye, so that users would still want to share them. The task encourages participants to go beyond maintaining the appeal to actually increasing it by enhancing the image.

MediaEval'20, December 14-15 2020, Online

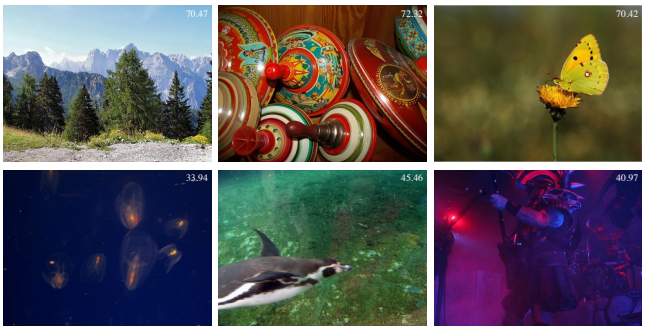

Figure 1: Examples of training images in MediaEval Pixel Privacy task 2020. Images are selected from high-quality class (top row) and low-quality class (bottom row). BIQA score for each image is on the top right corner.

In Fig. 1 we see examples of high-quality and low-quality images as judged by an BIQA algorithm. The images are drawn from the development set of the task, described further in Section 2. Note that the quality of an image is defined as independent of its aesthetic qualities or artistic merit. For example, the picture on the left of the bottom row could be considered beautiful, but would not be considered a high-quality image because of lack of contrast and sharpness. The importance of this difference for the task will be discussed further in Section 4.

Our choice of investigating image quality is motivated by the large part that quality plays in determining the fate of an image within a social media platform or an AI system. Computer vision algorithms that classify image quality are ethically problematic when the quality information they infer is used inconsistently with the intent of the users who originally shared the images. This issue is illustrated by two key examples. First, high-quality images can be recommended to users in image sharing social media [12], or they can also be ranked at the top by the search engine [23]. However, users who generated these images may not intend their contents to be spread widely in the general public. Quality camouflage can reduce the chance that an image would be shared widely beyond their family, friends or colleagues. Second, because the performance of computer vision algorithms can be degraded when the image quality is low [8], image quality classifiers are commonly applied as a pre-processing step to filter out low-quality images in the training or testing [6] phase. However, users may not have anticipated or agreed to having their images used to train certain types of ethically questionable classifiers. Quality camouflage can reduce the chance that their images would be used for training.

1.1 Adversary Model

Quality camouflage is designed to protect images. However, technically, it is considered an *adversary*, and defined according to an

adversary model. We follow the stand dimensions of the adversary model, set out by [22]. First, in the *resources/access* dimension, we assume that we (i.e., the adversary) have general knowledge of the target. We know that we are protecting against a state-of-the-art quality classifier and also the images we are trying to protect will be subject to some minimal pre-processing intended to counter adversarial examples. Second, in the *risk tolerance* dimension, we assume that we are not worried whether it is possible to infer that we have protected our images (with human eyes or a classifier). Third, in the *objectives* dimension, we assume that we are interested in lowering the quality score of high-quality images, thereby raising the probability that the images are passed over. We are not looking for a mathematical guarantee, since our target is open ended.

We are interested in lowering the barrier in task participation, for this initial exploratory investigation of quality camouflage. For this reason, we use a white-box variant of the adversary model, providing participants with the knowledge of the exactly quality classifier (i.e., the BIQA) that is used and the type of pre-processing to expect (90% JPEG compression, which has a negligible effect on the BIQA score). Note that this formulation can also be considered a gray-box, since the exact compression algorithm is not specified.

1.2 Relation to Computer Vision Research

The task is related to directions in which research on adversarial examples are currently expanding. In this section, we sketch this connection. A direct way to achieve quality camouflage is by adding small, imperceptible adversarial perturbations to the image. Such imperceptibility has been addressed conventionally by L_p distance [4, 5, 11, 14, 20], and recently by more visual-perception-aligned measurements [7, 18, 25–27, 29].

However, in real-world scenarios, such small, imperceptible perturbations may not actually impact the target model in the end, since their effect can be erased by practical image pre-processing operations. In addition, it has been recently pointed out that such small, imperceptible perturbations were just introduced as an abstract, toy example to facilitate evaluation [10] and have no truly compelling real-world use scenario. For these reasons, researchers have started to explore *unrestricted adversarial images* [1, 2, 9, 21, 28], which accommodate larger perturbations that are still unsuspicious because they transform groups of pixels along dimensions consistent with human interpretation of images. Participants in past years of the Pixel Privacy Task have also carried out preliminary exploration in such a direction [3, 16].

The task encourages participants to investigate approaches simultaneously fool BIQA and maintain/enhance appeal. We are interested in creating images that people will want to share on social media, in order to inspire them to protect their privacy. For this reason, image appeal is a key consideration. Our position is that as work on unrestricted adversarial examples moves forward it should be guided by adversary spaces that are well aligned with human-interpretable image editing processes, i.e., what users themselves do to enhance their images and ensure that they are appealing.

2 TASK DEFINITION AND DATA

The objective of the Pixel Privacy Quality Camouflage Task is to fool a BIQA algorithm while maintaining or enhancing the appeal of an image to the human eye. Blind image quality assessment is a long-standing research topic in computer vision and image processing [24]. It aims to predict the mean opinion score (MOS) of an image among a sufficiently large sample of human emulators, which is also interpreted as the perceived quality of stimulus [19]. Previously, BIQA algorithms are trained on images with synthetically generated distortions or small scale annotated data. The BIQA we employ here is trained on the KonIQ-10k [13] data set that is the first in-the-wild database with 1.2 million reliable quality ratings from a group of users. We train the model using the same data split of the original paper and achieve PLCC 0.927 on the test set, which is on par with the originally reported results. We release the training code and the pre-trained BIQA model[1,2]. The BIQA model uses an Inception-ResNet-v2 model and predicts a score ranging from 0 to 100. Images with a score of more than 50 are classified as high-quality. We also provide a development set (`MEPP20dev`) containing all the 1000 images from the official validation set of the KonIQ-10k data set. The test set (`MEPP20test`) is composed of 550 high-quality (i.e, BIQA score >= 50) images, and it is a subset of the Pixel Privacy 2019 test set [17], in turn a subset of the Places-365 Standard data set [30].

3 EVALUATION

Task submissions are evaluated in three steps. First, in the pre-processing step, all submitted images are processed by JPEG compression (ratio = 90%). Then, compressed images are classified by the BIQA model as high-quality or low-quality, and the reduction in the accuracy of the predictions is calculated. Finally, approaches that achieve at least 50% reduction of accuracy (i.e., the percentage of low-quality images is more than 50%) are ranked in terms of their appeal by a jury of computer science students. The twenty test images with the highest variance of BIQA scores among all runs will be judged. For each image, the jury receives a set of camouflaged images, and picks the three "most appealing" and two "least appealing" based on their perception of appeal. This year, we have seven annotators in the jury of computer science students. They are provided with a hint, "Appealing could mean that you would like to share the image with your friends on social media". All submitted runs will be ranked by the frequency of "most appealing".

4 DISCUSSION AND OUTLOOK

The 2020 Pixel Privacy Quality Camouflage task explores adversarial images that fool a BIQA model and at the same time remain appealing enough to share. As noted before, image quality does not reduce to aesthetics or artistic merit. The relative independence of quality and appeal makes it sensible to approach the task in two steps. However, if we want users to use quality camouflage images in practice, it is important to systematically explore the whole space of appealing images and understand which parts of this space are best adversarial to image quality classifiers. Moving forward, we would like to better understand, image appeal, quality camouflage coutermeasures, and also the possibility that quality camouflage may transfer to protect against semantic classifiers.

[1] https://github.com/ZhengyuZhao/koniq-PyTorch

[2] 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] Tom Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. 2018. Unrestricted adversarial examples. In *arXiv preprint arXiv:1809.08352*.

[3] Simon Brugman, Maciej Wysokinski, and Martha Larson. 2018. MediaEval 2018 Pixel Privacy Task: Views on image enhancement. In *Working Notes Proceedings of the MediaEval Workshop*.

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

[5] 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*.

[6] Timofey Chernov, Nikita Razumnuy, Alexander Kozharinov, Dmitry Nikolaev, and Vladimir Arlazarov. 2018. Image quality assessment for video stream recognition systems. In *International Conference on Machine Vision*.

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

[8] Samuel Dodge and Lina Karam. 2016. Understanding how image quality affects deep neural networks. In *International Conference on Quality of Multimedia Experience*.

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

[10] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. 2018. Motivating the rules of the game for adversarial example research. In *arXiv preprint arXiv:1807.06732*.

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

[12] Vic Gundotra. 2013. New Google+: Stream, Hangouts, and Photos. https://googleplusproject.blogspot.com/2013/05/new-google-stream-hangouts-and-photos.html, Online; accessed 28-Dec-2020. (2013).

[13] 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 Transactions on Image Processing* 29 (2020).

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

[15] Martha Larson, Zhuoran Liu, Simon Brugman, and Zhengyu Zhao. 2018. Pixel Privacy: Increasing image appeal while blocking automatic inference of sensitive scene information. In *Working Notes Proceedings of the MediaEval Workshop*.

[16] Zhuoran Liu and Zhengyu Zhao. 2018. First steps in Pixel Privacy: Exploring deep learning-based image enhancement against large-scale image inference. In *Working Notes Proceedings of the MediaEval Workshop*.

[17] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. 2019. Pixel Privacy 2019: Protecting sensitive scene information in images. In *Working Notes Proceedings of the MediaEval Workshop*.

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

[19] Anush Moorthy and Alan Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing* 20, 12 (2011).

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

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

[22] Chris Salter, O. Sami Saydjari, Bruce Schneier, and Jim Wallner. 1998. Toward a secure system engineering methodology. In *Proceedings of the 1998 Workshop on New Security Paradigms*.

[23] Harry Shum. 2013. A behind the scenes look at how bing is improving image search quality. https://blogs.bing.com/search-quality-insights/2013/08/23/a-behind-the-scenes-look-at-how-bing-is-improving-image-search-quality, Online; accessed 28-Dec-2020. (2013).

[24] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004).

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

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

[27] Hanwei Zhang, Yannis Avrithis, Teddy Furon, and Laurent Amsaleg. 2020. Smooth adversarial examples. *EURASIP Journal on Information Security* 2020, 1 (2020).

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

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

[30] 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* 40, 6 (2017).