

HCMUS at Pixel Privacy 2020: Quality Camouflage with Back Propagation and Image Enhancement

Minh-Khoi Pham^{*1,3}, Hai-Tuan Ho-Nguyen^{*1,3}, Trong-Thang Pham^{1,3}, Hung Vinh Tran^{1,3},
Hai-Dang Nguyen^{1,3}, Minh-Triet Tran^{1,2,3}

¹University of Science, VNU-HCM

²John von Neumann Institute, VNU-HCM

³Vietnam National University, Ho Chi Minh city, Vietnam

{pmkhoi, hnh Tuan, ptthang, tvhung, nh dang}@selab.hcmus.edu.vn, tmtriet@fit.hcmus.edu.vn

ABSTRACT

As our needs to share moments evolve, the more high-quality photos appear on the Internet. Hence, it is more likely that shared photos will be used for purposes that the owner does not want by someone else. If the target photos are high-quality, the attacker may use some criteria to assess the quality of images, such as the Blind Image Quality Assessment (BIQA) classifier. Pixel Privacy 2020 aims to tackle this problem. In this challenge, we have implemented methods of combining image enhancement and an end to end attack. The final results show that all of our approaches successfully fool the BIQA. In particular, our best run results in 84 photos being chosen as top-3 most attractive while maintaining 100% attack's accuracy on the assessment model, and out-performs all other submissions.

1 INTRODUCTION

With the recent rapid development of social networks, the need for sharing images also increases. Thus, smartphones have more high-end cameras, which leads to increment image quality. These images, which target to share with your friends, could be exploited by attackers for your private data. For example, high-quality images could be used as a filter for your honey-moon images. Therefore, we consider the Pixel Privacy task is vital for this age of connection.

In Pixel Privacy 2020[5], we are given a set of images that was evaluated as high quality by BIQA[9] model. Our target is to fool the BIQA model so that the model will consider the modified image as low quality, and the image remains attractive under human eyes. To be more specific, this BIQA model was trained on KonIQ-10k dataset[1], and the given images were from Places365 validation dataset. The output image would be processed by JPEG-compression (ratio = 90%) before being evaluated.

We propose three approaches. One is vanilla end-to-end with an image-to-image based. In this approach, we aim to learn a single network that could enhance image quality and protect the quality from being evaluated by the BIQA model. To be flexible in changing the image enhancement method, we also propose a two-stage approach. The first stage is enhancing image quality, in which we experiment with multiple methods to improve image quality. The second stage is to camouflage the enhanced image's quality. Three of our runs, namely Pillow, Cartoonization, and Retouch follow this

approach. For the last approach, we assume that if the enhancement model is good enough, it will keep the image attributes, which in this case is protected from the BIQA model. The End-to-End with I-FGSM applies this approach.

2 APPROACH

2.1 Vanilla End-to-End

In this run, we use an Image-to-Image network to reconstruct the input image; then, we forward the reconstructed result to the BIQA regressor.

We simply choose the U-Net[6] model as our main network because it is one of the most popular baseline models for image to image problem and simple enough to implement.

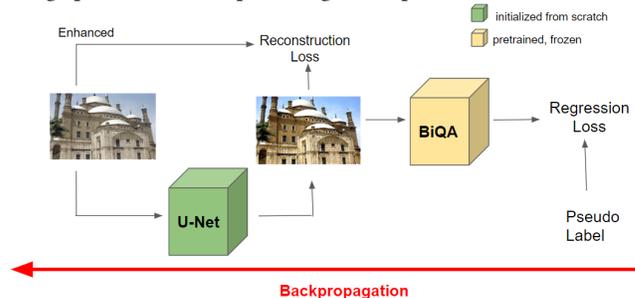


Figure 1: Vanilla End-to-End method

In Fig 1, the image x is first taken as input to U-Net, and model outputs the image y . Simultaneously, x is also enhanced to x' by using simple transformations from available computer vision libraries (same as 2.2.2). After that, we use the trained frozen BIQA to predict a score for y . We then generate a pseudo target score to attack the true score of y . Here, we have two objective functions for the network to minimize: the reconstruction loss between x' and y and the regression loss between pseudo score and true score.

Reconstruction loss: We experiment on both L2 loss, and SSIM Loss [10] and find that model trained with L2 gives out more visually appealing images than other objective functions.

Regression loss: We choose L2 loss to compute the distance between two scores. The pseudo score is generated by subtracting A from the original score B . We experiment A with the values of 30, 50, and B . We choose A equals 30 to submit in this run.

We add both losses and then back-propagate it to U-Net for the model to be able to learn. We train the network from end to end on the *pp2020_dev* dataset with U-Net being initiated from scratch.

*These authors contributed equally.

2.2 Two-stages approaches

2.2.1 Attack Algorithm.

In these approaches, we utilize Iterative Fast Gradient Sign Method (I-FGSM)[4] to perform a white-box attack on the BIQA model after the images are enhanced. Since the BIQA is a regression model, we use L2 loss function instead of Cross-Entropy loss, same as in 2.1.

Our modified I-FGSM is described as follows, with X the input image, y the BIQA score of X_N^{adv} , y' and attacking score. $J(X, y, y')$ the L2 cost function of the neural network, given image X , score y and attacking score y' , measuring the distance between y and y' :

$$X_0^{adv} = X \quad (1)$$

$$X_{N+1}^{adv} = \text{Clip}_{X, \epsilon} \{X_N^{adv} + \alpha \text{sign}(\nabla_x J(X_N^{adv}, y, y'))\} \quad (2)$$

Given y the predicted score on X_N^{adv} , we iteratively add the perturbation to X until y becomes smaller than score y' . For all the runs below, we find that setting $\alpha = 0.05$, $\epsilon = 0.05$ and $y' = 30$ gives desirable results in most cases.

2.2.2 Image Enhancement algorithm.

Pillow: We use several image enhancement operations which is provided by Pillow, such as adjusting color balance by 1.5^1 , sharpness by 3.0^1 , brightness by 1.0^1 and contrast by 1.5^1 . We apply the same configuration for all images in the data set.

Cartoonization: For this run, we apply a GAN-based White-box Cartoonization method[8] to convert input images to cartoon images with styles from Shinkai Makoto, Miyazaki Hayao, and Hosoda Mamoru films.

Retouch: In this run, we also want to compare one deep learning "white box" approach[2] with natural enhancement and "black-box" method. This method applies deep reinforcement learning and GAN Model to produce parameters for traditional image processing methods to improve image quality.

2.3 End-to-End with I-FGSM

However, different from other previous approaches, we will integrate I-FGSM with enhancement model. For each iteration, we will first feed forward image to a deep learning model and BIQA model then we will backward from L2 loss to calculate gradient and apply I-FGSM on input image. For this particular experiment, we choose EnlightenGAN [3] as enhancement model.

3 EXPERIMENTS AND RESULTS

As can be seen in table 1, *accuracy (after JPEG 90)* is the accuracy of model on dataset after being compressed 90% (lower is better). *Number of times selected as "Best"* (Max. 140) base on human experts evaluation. To be more specific, 20 images with largest BIQA variance will be selected for 7 human experts to choose best three runs out of all runs.

Table 1: Official evaluation result (provided by organizers)

Methods	Accuracy (after JPEG 90)	Number of times selected as "Best"
End-to-End (UNet)	13.27	11
Retouch + I-FGSM	0.18	34
Cartoonization + I-FGSM	1.27	34
Pillow + I-FGSM	48.18	60
End-to-End (EnlightenGAN) + I-FGSM	0.00	84

With the above result, we have successfully fooled BIQA by more than 50% in all runs. EnlightenGAN proves to be the best method in our runs, followed by enhancement performance based on Pillow's traditional image processing approach. This is an interesting result compare to our result in the previous Pixel Privacy[7]. Last year, traditional methods still outperformed GAN-based approaches, but this year, our proposed approach combine with a GAN method has proven to be better than a traditional image processing method. This could be explained as GAN could perform a more flexible image enhancement based on image, case by case, comparing to traditional image processing algorithms with hard-coded parameters.

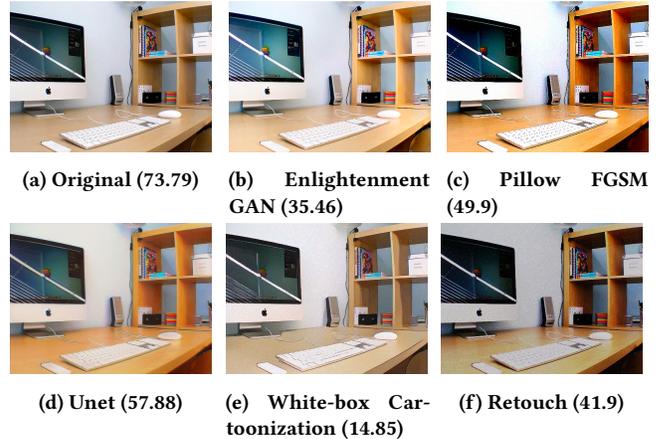


Figure 2: Sample outputs with BIQA scores

4 CONCLUSION

All of our approaches are simple but effective enough to fool BIQA model while maintaining the high quality of images.

The Image-to-Image based method benefits that it does not require pair-to-pair images to train and it also performs attacks on feature-level not on raw images, in comparison with other methods. Although it gives out the worst result among others, we still believe that it can be further investigated and improved.

The two-stage approaches, whose results are better, still have clearly visible noise over the images. I-FGSM also shows its strength in white-box attacks, both classification or even regression manner.

Our new proposed approach, which is the combination of End-to-end and I-FGSM, shows both efficiency and effectiveness in camouflaging and visually enhancing the images.

ACKNOWLEDGMENTS

Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19.

¹See Pillow documents for explanation of these numbers

REFERENCES

- [1] 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), 4041–4056. <https://doi.org/10.1109/tip.2020.2967829>
- [2] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. 2017. Exposure: A White-Box Photo Post-Processing Framework. *CoRR* abs/1709.09602 (2017). arXiv:1709.09602 <http://arxiv.org/abs/1709.09602>
- [3] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. 2019. Enlighten: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972* (2019).
- [4] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial examples in the physical world. (2017). arXiv:cs.CV/1607.02533
- [5] Zhuoran Liu, Zhengyu Zhao, Martha Larson, and Laurent Amsaleg. 2020. Exploring Quality Camouflage for Social Images. In *Working Notes Proceedings of the MediaEval Workshop*.
- [6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. (2015). arXiv:cs.CV/1505.04597
- [7] Hung Vinh Tran, Trong-Thang Pham, Hai-Tuan Ho-Nguyen, Hoai-Lam Nguyen-Hy, Xuan-Vy Nguyen, Thang-Long Nguyen-Ho, and Minh-Triet Tran. 2019. HCMUS at Pixel Privacy 2019: Scene Category Protection with Back Propagation and Image Enhancement. (2019).
- [8] Xinrui Wang and Jinze Yu. 2020. Learning to Cartoonize Using White-Box Cartoon Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [9] Xin Li. 2002. Blind image quality assessment. In *Proceedings. International Conference on Image Processing*, Vol. 1. I–I. <https://doi.org/10.1109/ICIP.2002.1038057>
- [10] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. 2016. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging* 3, 1 (2016), 47–57.