HCMUS at Pixel Privacy 2019: Scene Category Protection with Back Propagation and Image Enhancement

Hung Vinh Tran*, Trong-Thang Pham*, Hai-Tuan Ho-Nguyen*, Hoai-Lam Nguyen-Hy*,
Xuan-Vy Nguyen*, Thang-Long Nguyen-Ho*, Minh-Triet Tran
Faculty of Information Technology, University of Science, VNU-HCM, Vietnam
{tvhung,ptthang,nxvy,hnhtuan,nhhlam,nhtlong}@selab.hcmus.edu.vn,tmtriet@fit.hcmus.edu.vn

ABSTRACT

Personal privacy is one of the essential problems in modern society. In some cases, people may not want smart computing systems to automatically identify and reveal their personal information, such as places or habits. This motivates our proposal to protect scene category recognition from photos by back-propagation. To further improve the visual quality and attraction of output photos, we study and propose various strategies for image enhancement, from traditional approaches to novel GAN-based methods. Our solution can successfully fool the Place365 scene classification in 60 categories while achieving the average NIMA score up to 5.36.

1 INTRODUCTION

With the rapid development of computer vision and new machine learning methods, computers can now understand content of images, e.g. which object is in an image, who is in the image, what is happening in the image, or recognize scene's category and attributes, etc. This provides foundation for intelligent interactive systems such as smart homes, self-driving cars, etc. This fact, however, can raise potential risks of personal privacy violation. A person might not want other people to know where he or she is. Important facilities like hospitals or military camps should not be automatically recognized also. This motivates the proposal of the problem of Pixel Privacy: to prevent automatic systems from recognizing scene categories from images [6, 7].

In the Pixel Privacy 2019 [7] task, we are given images in 60 important categories such as hospital or bedroom. Our goal is to modify the given images so that the given automatic system (a ResNet50 [3] trained on the Places365-Standard [10] dataset) can no longer correctly recognize them. This, however, may cause the output images to be degraded significantly. To prevent this, we use Neural Image Assessment (NIMA) to automatically evaluate our output images.

We propose the method to attack the ResNet50 model using back propagation. For each input image, we consider the logit for its target class as a function on the input image's pixels. This allows back propagation to minimize this logit by modifying the input image. We also use several methods to enhance the input images' quality before feeding it through the protection phase, such as natural enhancement with Dynamic Histogram Equalization [1] combined with saliency mask generated from Cascaded Partial Decoder [9], GAN-based approaches like CartoonGAN [2], DPED [5], Retouch [4], and Texture. Our experiment results successfully fool the Places365 scene classification, while achieving an average NIMA score up to 5.36.

The content of our report is as follows. In Section 2, we present our method for protecting scene category and enhancing image with different approaches. Experimental results are in Section 3. Conclusion and future works are discussed in Section 4.

2 METHOD

Our goal is to protect the scene category of the input image, and our output result should be in better quality based on NIMA [8] Score. To tackle these two goals, we desire to enhance the input image first by different methods then we apply our proposed protection method on the output enhanced image. We put the protection method in the latter part to ensure the success of our main objective, which is the privacy of users.

2.1 Protection algorithm

To protect the image's location information, we need to modify the image so that the output class becomes different compared to the ground truth. In other words, we need to reduce the output probability of the ground truth class.

With this intuition in mind, we consider the output probability of ground truth class as the target function f to minimize with gradient descent and the input image as the parameter θ to optimize.

Formally, let θ be the input image, θ' the modified image, f the model we want to fool. Our main goal is to modify input image :

$$\theta' = \theta + \epsilon$$

so that

$$f(\theta') \neq f(\theta)$$

where ϵ is obtained after backwarding through network encoding in Fig.1

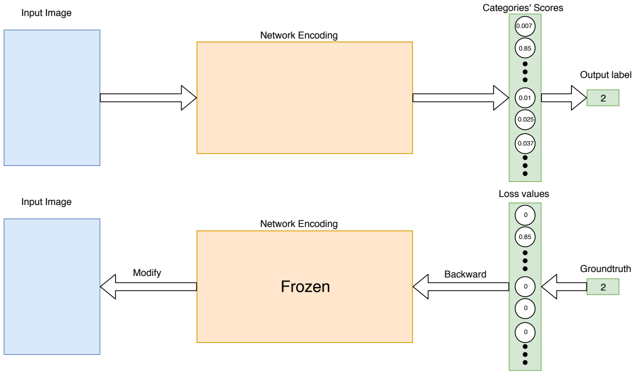


Figure 1: Protection Algorithm

MediaEval'19, 29-31 October 2019, Sophia Antipolis, France

2.2 Image Enhancement algorithm

2.2.1 Natural Enhancement. For the first run, we try to improve image quality with traditional computer graphic methods. We start with Dynamic Histogram Equalization[1] algorithm to enhance each image's contrast and brightness. We also try to configure the saturation of images to emphasize the main objects. We use Cascaded Partial Decoder [9] to generate the saliency mask which describes important objects with high values. Then we feed the saliency mask into Gaussian Blur and get a new mask. Thirdly, we modify saturation value each pixel:

$$y = \begin{cases} \frac{x-\alpha}{\beta} + \gamma, & \text{if } x \geq \alpha \\ \frac{x}{\omega}, & \text{otherwise} \end{cases}$$

In this formula, y is a saturation value and x is a new saliency value in the same pixel.

2.2.2 Style Transfer. For this run, we apply a style transfer-based approach to improve images' appeal. In this particular case, we run CartoonGAN [2] to modify images into some certain styles, such as Hayao, Hosoda and Shinkai.

2.2.3 GAN Enhancement. In this run, we try to improve image's quality by method proposed in [5] , which is a GAN-based algorithm. This model's purpose is to convert a photo taken by phone to be a DSLR-Quality photo, in consequence, improve the color and texture quality.

2.2.4 Retouch. In this run, we assume that the evaluation method with NIMA Score is as natural as the human visual system. We try to enhance the image with [4], which applying deep reinforcement learning and GAN Model, to have the best quality image enhancement created by AI Agent.

2.2.5 Texture. In this run, we hypothesize that any natural image has an average NIMA score between 3 and 4.5. We attempted to validate this hypothesis by computing NIMA score for several randomly created images, which results in fairly high score. With that in mind, we blend noise images to the original images to emphasize details usually ignored by NIMA model, while keeping the scores above 3. We try 2 ways to blend the noise images:

Way 1: $x' = x \odot (\alpha\epsilon)$, where x and x' are the input and ouput images, respectively; ϵ is the crafted noise image; α is the coefficient in which the crafted image is blended into the original.

Way 2: $x' = x + \alpha\epsilon$, where x and y are the V (as in HSV) channels of the original image and the result image, respectively; ϵ is the crafted noise image; α is the coefficient with which the crafted image is blended into the original.

3 EXPERIMENTS AND RESULTS

Most of our experiment is conducted on Google Colab.

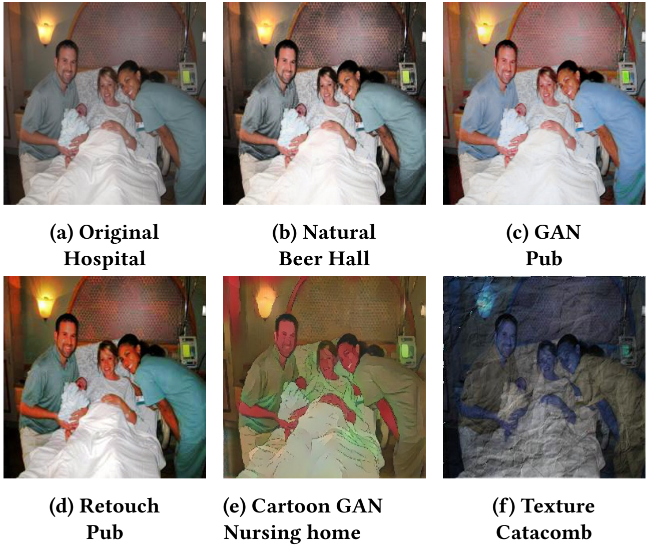
Table 1: Official evaluation result (provided by organizers)

Method	Top-1 Accuracy	NIMA Score
Original Image		4.64
hcmus_naturalenhancement	0	5.14
hcmus_retouch	0	4.71
hcmus_ganenhancement	0	4.84
hcmus_cartoongan	0	4.96
hcmus_texture	0	**5.36**

In the table above, the "Top-1 Accuracy" column shows the prediction accuracy of the attack model (ResNet50 [3] trained on Places365-standard data set), which means lower is better. The "NIMA Score" column represents the mean aesthetics scores of your runs. A higher NIMA score is better.

As our experiment result shows that all of our five runs perfectly protected the scene category of the image from the attack model. And from our observation, most GAN-based method can not achieve the same natural level as traditional computer graphics approach. The texture method currently has the highest score (5.36). Moreover, without the protection method, the texture method also can protect more than 2/3 datasets.

However, sophisticated approaches like Texture or Cartoon Gan transform images excessively, and output images do not have a natural appearance anymore. That is the reason why we propose Traditional Enhancement Methods like Saturation modify or DHE algorithm. Not only does it maintains the original beauty, but Natural Enhancement also achieves 2nd highest score.



(a) Original Hospital **(b) Natural Beer Hall** **(c) GAN Pub**

(d) Retouch Pub **(e) Cartoon GAN Nursing home** **(f) Texture Catacomb**

Figure 2: Sample output

4 CONCLUSION AND FUTURE WORKS

We propose one simple yet effective approach for Pixel Privacy Problem. Our method is using backpropagation to modify certain pixels of the input image while freezing all intermediate modules in the attack model. This method could be expanded for other categories besides scene category, for example, vehicle detection, human tracking, and so on. We also propose and apply five methods to enhance the input image, from new methods, namely, Cartoon GAN (Style Transfer) and Texture effect, to traditional image enhancement methods. All of which provides images with a higher average NIMA Score than original images.

ACKNOWLEDGMENTS

Research is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19. We would like to thank AIOZ Pte Ltd for supporting our team with computing infrastructure.

REFERENCES

[1] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan, and Oksam Chae. 2007. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics* 53, 2 (2007), 593–600.

[2] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. 2018. CartoonGAN: Generative Adversarial Networks for Photo Cartoonization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.

[4] 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

[5] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. 2017. DSLR-Quality Photos on Mobile Devices with Deep Convolutional Networks. In *Proceedings of the IEEE international conference on computer vision.*

[6] 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 2018 Workshop.*

[7] Zhuoran Liu, Zhengyu Zhao, and Martha Larson. 2019. Pixel Privacy 2019: Protecting Sensitive Scene Information in Images. In *Working Notes Proceedings of the MediaEval 2019 Workshop.*

[8] Hossein Talebi and Peyman Milanfar. 2018. Nima: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[9] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. *CoRR* abs/1904.08739 (2019). arXiv:1904.08739 http://arxiv.org/abs/1904.08739

[10] 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* (2017).