Improved PULSE: (Photo Upsampling via Latent Space Exploration of Generative Models) Using Anti-spoofing System

Yubo Wang, Aaron Wang, Weilan Chen, Yuting Jiang, Shaoan Wang, Yuzhe Jiang, Yuzheng Wang

Department of Computer & Information Science & Engineering, University of Florida, Gainesville 32608, USA
yubo.wang@ufl.edu

Abstract. Image super-resolution is typically explained as converting a low-resolution (LR) image into a high-resolution (HR) output. However, previous approaches could be blurry and unrealistic, especially for regions with many details. This paper took advantage of Latent Space Exploration from PULSE and a facial liveliness detection algorithm to receive a more accurate and rational outcome. (Related code can be found on https://github.com/cisgroupb/PULSE-with-Anti-Spoofing-Detection)

1. Introduction
Due to modern technology's rapid growth, it has become prevalent to authorize a transaction using our biometric information, including fingerprints and facial features [1]. Along with the convenience that it has brought, a couple of questions arise: is it safe enough? Could a hacker fake a person's face? [2] To prevent potential fraud, researchers have made promising progress on improving the accuracy of face recognition through various aspects; and measuring the liveliness of input is one of the hottest approaches among them [3].

However, a detection system cannot receive high-resolution input in real life because of damaged cameras or even terrible weather. Hence, we would like to propose an updated model for solving the face-spoofing problem in this paper. Specifically, algorithm presented in [4] was modified and combined with the liveliness-detection model argued in https://github.com/zeusees/HyperFAS and [5]. To show the improvement achieved by our method, LPIPS standard suggested in [6] was used, and our model received a higher score than PULSE as a result.

2. Dataset
To obtain more information and a greater variety of facial features, the dataset FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age [7] was chosen to be used throughout the whole project for it containing 108,501 images with faces of different races, gender, and ages [7], which could assure the consistency of accuracy across all groups when pre-train our model and obtain a more optimal result in the end.

3. Method
In this section, the logic and implementation details of the model constructed will be explained.
3.1. Input
To obtain the ideal format of our inputs, align_face.py was first ran in https://github.com/cisgroupb/pulse suggested in [4] by inputting HR images from the dataset FairFace [7]. In particular, the positions of faces in different images could be distinct from each other, which would increase the difficulty and potential inaccuracy when extracting information during training.

Nevertheless, through align_face.py [4], images would be scaled into the same size with faces locating in relatively the same position before they are downsampled and outputted, which provided the perfect simulation of LR pictures for our model to use.

3.2. PULSE (Self-supervised photo Upsampling via Latent Space Exploration of generative models)
Traditionally, a super-resolution algorithm for a low-resolution image focuses on taking the average distance between the super-resolved image and the high-resolution image pixel-wise during a supervised training [8]. However, the results received tend to be blurry, and the evaluation metrics' variance could be significant [4]. To improve this, PULSE was implemented as an algorithm based on a latent space and obtaining the final output by traversing the HR image manifold and searching for the original LR input [4].

Although there are still some limitations and improvements for this particular model [4, 9], we believe PULSE would provide us with a reliable methodology for collecting prescribed outcomes for the following procedure.

3.3. Modified number of outputs
Initially, in the PULSE algorithm, a set of feasible pictures would be produced with a given critical value (referred to as epsilon in [4]), and then the final output image would be chosen from the intersection between the set and the natural image manifold [4]. However, the result created in this way could be random, biased, and sometimes not realistic enough. Therefore, in our model, after inputting an LR image, the parameter of PULSE was changed so that the algorithm could generate HR images five times rather than once before the liveness evaluation algorithm was applied in order to find a more sensible output.

3.4. Liveness evaluation
To pick the most “real” face produced from the five outputs suggested in the previous section, the model argued in [10] and https://github.com/zeusees/HyperFAS was used. Based on the improved state-of-art technique discussed in [5], face detection and alignments have been boosted by using the Cascaded convolutional networks [5]. The authors in [10] then proposed a faster detection model by combining the existed model with a designed novel architecture [10]. Mainly, the MobileNetV3 they created was based on the hardware network architecture search (NAS), and two models were created targeting different resources [5].

4. RESULTS AND EVALUATION

4.1. Evaluation standard: LPIPS
Evaluating the similarity between two pictures can be quite perspective and extremely difficult when dealing with a large data set. In this work, instead of using classic pixel measures, such as Peak Signal-to-Noise Ratio (PSNR) and mean squared error (MSE), which would not provide us proper
assessment due to their pixel-independent property [5], we decided to use the Learned Perceptual Image Patch Similarity (LPIPS) suggested in [11] and https://github.com/cisgroupb/PerceptualSimilarityMaster as our measuring metrics.

This particular standard was chosen because LPIPS illustrates a stable and consistent assessment for trained CNN even when facing some distorting types suggested in [11]. Besides, one should also notice that LPIPS measures the distance between images, which means that the higher the score, the more different the pictures.

4.2. Results

Overall, a slight improvement was shown by computing the average LPIPS scores for all outputs from PULSE: 0.19213 compared with our model, scoring 0.19175.

Although the development does not seem vast, the reliability of production has enhanced because of our model. According to Table 1, after inputting FairFace, the rate of receiving an outcome has risen from 19.13% to 25.94%, by nearly 36%. The result shows that the outputted pictures would be more realistic, but the probability of receiving a rational outcome would increase by using our model.

Table 1. Compares the production rate between PULSE and our model based on the number of outputs

| Expected number of outputs | PULSE | Our model |
|----------------------------|-------|-----------|
| The actual number of outputs | 7595  | 1519      |
| Production rate            | 19.13% | 25.94%    |

To compare the performances even further, the distributions of LPIPS scores of the two models was compared. Figure 2 shows the data for both models showing the tendency to approach a normal distribution, which implies our improved model's soundness. It is also worth mentioning that the proportion of having scores closer to our model's corresponding means is higher than that of PULSE, which again proves the authenticity of the model that was constructed. (Individual distribution histograms are referred in Figure 3)

![Figure 2. A histogram illustrates the comparison between the distribution of LPIPS scores of the two models](image)
5. CONCLUSION AND FUTURE WORK

As a result, this work has provided a methodology of improving PULSE’s performance by combining a liveliness-detection model, and the effects could be validated by statistics provided in this paper. However, it was noticeable that some of the side portraits and portraits with head garments such as hats were not recognized when modifying them using align_face.py [4] to receive the designated input before using PULSE.

The amount of useable data would then be decreased and might inaccuracy for certain groups if there is more illegal data in a specific group than the others. Therefore, to move forward, a better method of processing those images might reduce the inaccuracy.

Due to time restraints, only a relatively simple liveness detection model [5] that is more widely used in mobile networks was used, which could be a reason for having limited development for the performance. Hence, in the future, one can argue that the results would be better if a more complicated model were to be used.

Some other reasonable suggestions for upgrading the PULSE’s quality and accuracy would be extending the dataset by including more different facial portraits or training the model under a supervised algorithm.

Acknowledgments

This work was partially supported by Professor David P. Woodruff [12] and the Cathaypath Institute of Science (CIS) [13].

References

[1] Amanjot Singh and Jagroop Singh Sidhu. Super resolution applications in modern digital image processing. International Journal of Computer Applications, 150(2):0975–8887, 2016.
[2] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. International Journal of Computer Vision, 106(1):9–30, 2014.
[3] Aleksandr Parkin and Oleg Grinchuk. Recognizing Multi-Modal Face Spoofing with Face Recognition Networks. CPVR, 2020
[4] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. arXiv:2003.03808v2 [cs.CV] 24 Jun 2020
[5] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, Senior Member, IEEE, and Yu Qiao, Senior Member, IEEE. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks.
[6] Richard Zhang, Phillip Isola, Alexei A.Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv:1801.03924v2 [cs.CV] 10 Apr 2018
[7] Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race,
Gender, and Age. arXiv:1908.04913 [cs.CV] 14 Aug 2019

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In Proceedings of the European Conference on Computer Vision (ECCV) 2014.

[9] Tianyu Zhao, Changqing Zhang, Wenqi Ren, Dongwei Ren, and Qinghua Hu. Unsupervised degradation learning for single image super-resolution. arXiv preprint arXiv:1812.04240, 2018.

[10] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V.Le and Hartwig Adam. Searching for MobileNetV3. arXiv:1905.02244v2 [cs.CV] 14 May 2019

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. TIP, 2004.

[12] David P. Woodruff: Associate Professor of Computer Science at Carnegie Mellon University. Contact: dwoodruf@cs.cmu.edu http://www.cs.cmu.edu/~dwoodruf/

[13] Cathaypath Institute of Science: https://www.cathaypath.com/ or https://www.torhea.com/