Poster: UnGANable: Defending Against GAN-based Face Manipulation

Zheng Li[1] Ning Yu[2] Ahmed Salem[3] Michael Backes[1] Mario Fritz[1] Yang Zhang[1]
[1]CISPA Helmholtz Center for Information Security
[2]Salesforce Research [3]Microsoft Research
{zheng.li, director, fritz, zhang}@cispa.de
ning.yu@salesforce.com t-salemahmed@microsoft.com

Abstract

Deepfakes pose severe threats of visual misinformation to our society. One representative deepfake application is face manipulation that modifies a victim's facial attributes in an image, e.g., changing her age or hair color. The state-of-the-art face manipulation techniques rely on Generative Adversarial Networks (GANs). In this paper, we propose the first defense system, namely `UnGANable`, against GAN-inversion-based face manipulation. In specific, `UnGANable` focuses on defending GAN inversion, an essential step for face manipulation. Its core technique is to search for alternative images (called cloaked images) around the original images (called target images) in image space. When posted online, these cloaked images can jeopardize the GAN inversion process. We consider two state-of-the-art inversion techniques, including optimization-based inversion and hybrid inversion, and design five different defenses under five scenarios depending on the defender's background knowledge. Extensive experiments on four popular GAN models trained on two benchmark face datasets show that `UnGANable` achieves remarkable effectiveness and utility performance, and outperforms multiple baseline methods. We further investigate four adaptive adversaries to bypass `UnGANable` and show that some of them are slightly effective.

I. Bibliography

Zheng Li, Ning Yu, Ahmed Salem, Michael Backes, Mario Fritz, and Yang Zhang. UnGANable: Defending Against GAN-based Face Manipulation. In proceedings of 32nd USENIX Security Symposium (USENIX Security). USENIX, August 9–11, 2023, Anaheim, CA, USA.

II. Link

https://www.usenix.org/conference/usenixsecurity23/presentation/lizheng

Poster: UnGANable: Defending Against GAN-based Face Manipulation

Zheng Li[1] Ning Yu[2] Ahmed Salem[3] Michael Backes[1] Mario Fritz[1] Yang Zhang[1]

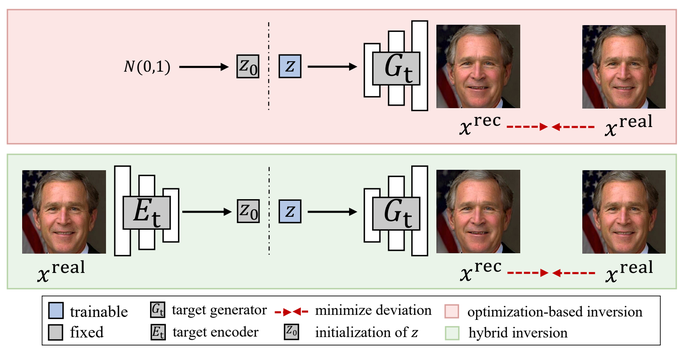
[1] *CISPA Helmholtz Center for Information Security* [2] *Salesforce Research* [3] *Microsoft Research*

1. Introduction

Deepfakes pose severe threats of visual misinformation to our society. One representative deepfake application is face manipulation that modifies a victim's facial attributes in an image, e.g., changing her age or hair color. The state-of-the-art face manipulation techniques rely on Generative Adversarial Networks (GANs), which consist of a two-step operation, i.e., GAN inversion that inverts a victim's facial image to a latent code and then performs latent code manipulation. In this paper, we propose the first defense system, namely `UnGANable`, against such GAN-inversion-based face manipulation. In specific, `UnGANable` focuses on defending GAN inversion, an essential step for face manipulation which consists of a two-step operation, i.e., GAN inversion and latent code manipulation. Its core technique is to search for alternative images (called cloaked images) around the original images (called target images) in image space. When posted online, these cloaked images can jeopardize the GAN inversion process.

2. GAN Inversions

We consider two representative and most widely-used techniques of GAN inversion, i.e., optimization inversion [1] and hybrid inversion [2].



3. Intuition of UnGANable

We derive the intuition behind our `UnGANable` from the basic pipeline of how inversion works. Since the optimization inversion is part of the hybrid inversion, here we focus only on the former. The optimization inversion employs a loss function that is a weighted combination of the perceptual loss and the pixel-wise MSE loss, to guide the optimization into the correct region of the latent space. This methodology leads to the following observations.

- The pixel-wise MSE loss works in the pixel space, i.e., the image space.
- The perceptual loss measures the similarity of features extracted from different images using a pre-trained model, which works in the feature space.
- The optimization aims to search for the optimal latent code, which works in the latent space.

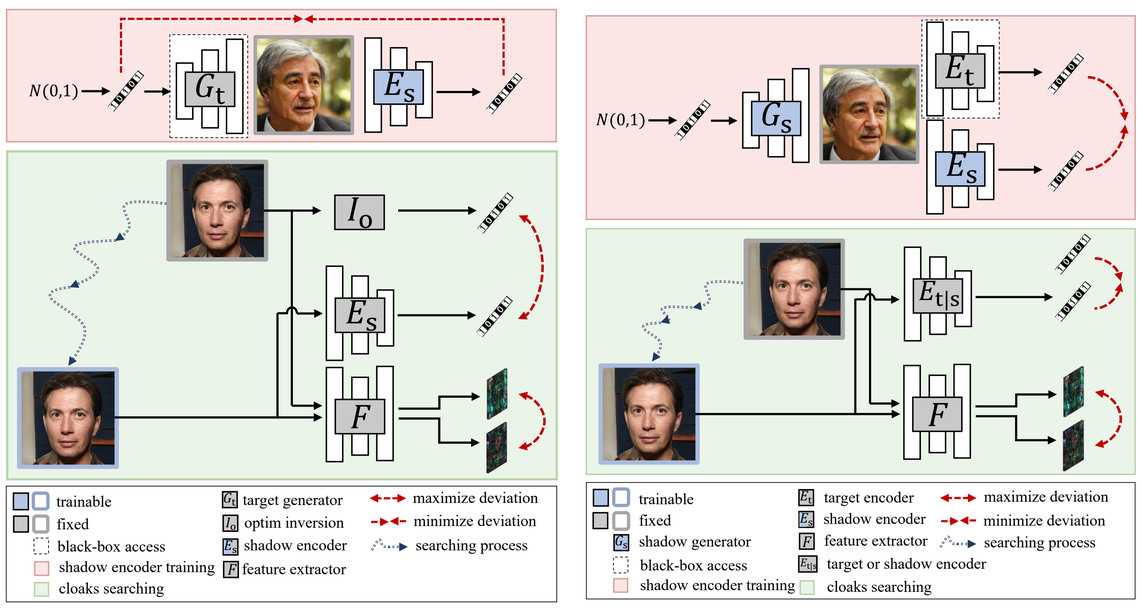
These observations motivate our `UnGANable`, which aims to maximize deviations in both latent and feature spaces with the cloaked images, meanwhile maintaining the image indistinguishable in the image space.

4. Design of UnGANable

Any user (also called defender) can use `UnGANable` to search for cloaked images, which are around the target images in the image space. The design goals for these cloaks are:

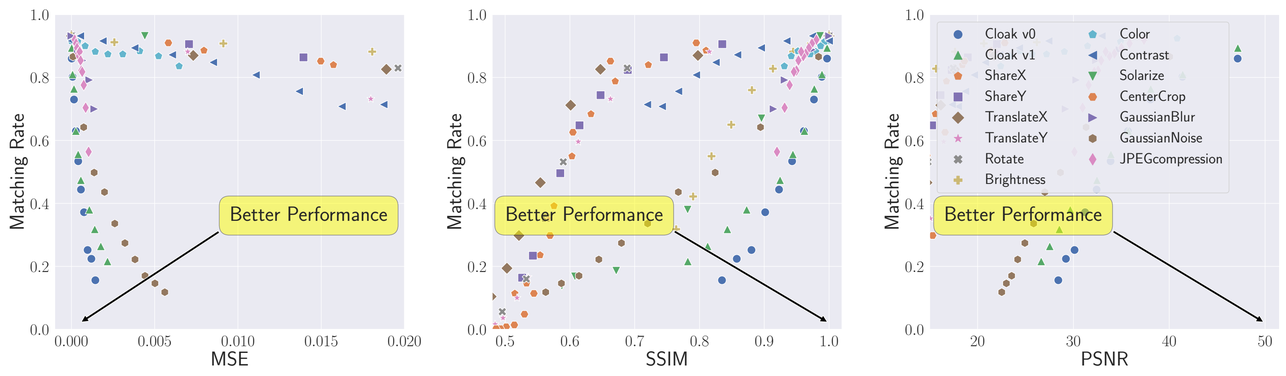
- cloaked images should be indistinguishable from the target images;
- when inverting the cloaked image, the adversary can only get a misleading latent code, which is far from its accurate one in the latent space.

Generally, `UnGANable` aims to maximize the deviations in the latent space and feature space, while keeping the images indistinguishable in image space.

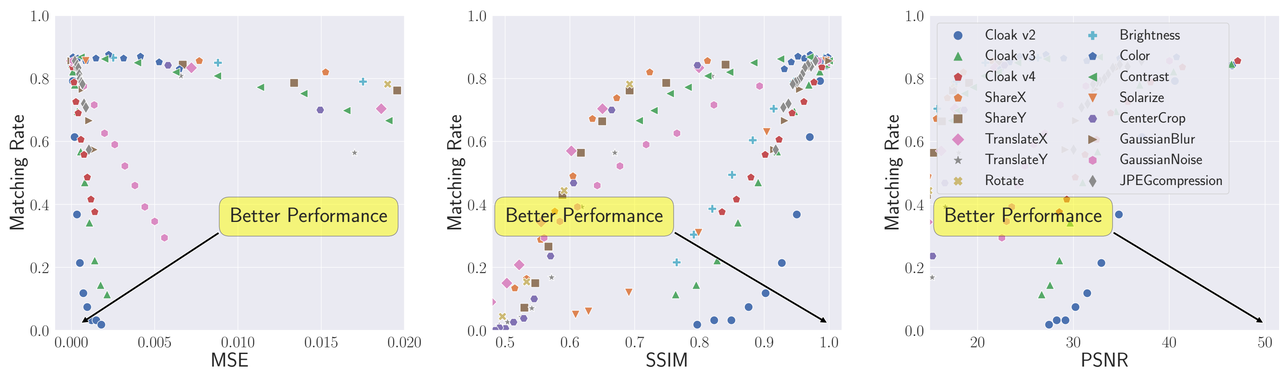


5. Experimental Results

We compare `UnGANable` extensively with thirteen baseline image distortion methods. The MSE/SSIM/PSNR measures the distance between cloaked and original images, while the matching rate measures the ratio of reconstructed images with changed identities.
The top three show the comparison between `UnGANable`(Cloak v0/v1) and all baseline methods against optimization inversion on StyleGANv2.



The bottom three show the comparison between `UnGANable`(Cloak v2/v3/v4) and all baseline methods against hybrid inversion on StyleGANv2.



`UnGANable` achieves better effectiveness (lower matching rate) and utility (lower MSE, higher SSIM, and PSNR) performance than all baselines.

7. References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? In *IEEE International Conference on Computer Vision (ICCV)*, pages 4431–4440. IEEE, 2019.

[2] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-Domain GAN Inversion for Real Image Editing. In *European Conference on Computer Vision (ECCV)*, pages 592–608. Springer, 2020.