Poster: UnGANable: Defending Against GAN-based Face Manipulation

Zheng Li¹ Ning Yu² Ahmed Salem³ Michael Backes¹ Mario Fritz¹ Yang Zhang¹
¹CISPA Helmholtz Center for Information Security
²Salesforce Research ³Microsoft 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


II. Link

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

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.

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.

References:
