What can Discriminator do? Towards Box-free Ownership Verification of Generative Adversarial Networks

Ziheng Huang[1†], Boheng Li[1†], Yan Cai[1], Run Wang[1*], Shangwei Guo[2],
Liming Fang[3], Jing Chen[1], Lina Wang[1]
[1] Key Laboratory of Aerospace Information Security and Trusted Computing,
Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China
[2] College of Computer Science, Chongqing University, China
[3] College of Computer Science and Technology, Nanjing University of
Aeronautics and Astronautics, China
[†] Equal contribution [*] Corresponding author. E-mail: wangrun@whu.edu.cn

Abstract

*In recent decades, Generative Adversarial Network (GAN) and its variants have achieved unprecedented success in image synthesis. However, well-trained GANs are under the threat of illegal steal or leakage. The prior studies on remote ownership verification assume a black-box setting where the defender can query the suspicious model with specific inputs, which we identify is not enough for generation tasks. To this end, in this paper, we propose a novel IP protection scheme for GANs where ownership verification can be done by checking outputs only, without choosing the inputs (*i.e.*, box-free setting). Specifically, we make use of the unexploited potential of the discriminator to learn a hypersphere that captures the unique distribution learned by the paired generator. Extensive evaluations on two popular GAN tasks and more than 10 GAN architectures demonstrate our proposed scheme to effectively verify the ownership. Our proposed scheme shown to be immune to popular input-based removal attacks and robust against other existing attacks. The source code and models are available at https://github.com/AbstractTeen/gan_ownership_verification.*

1. Introduction

With the rapid development of GANs, we have witnessed fruitful applications of GAN in many fields, such as realistic facial images synthesis [45], fine-grained attribute editing [55], *etc*. Unlike the classification model with specified label prediction, the GANs learn a data distribution and output the synthesized data sample within a certain distribution. In GANs, the discriminator and generator are two essential components, where the discriminator works as a judger to discriminate whether the sample is produced by the generator, the generator learns to generate more realistic samples to confuse the discriminator [16, 7]. Usually, the discriminator is discarded after training since the generator is the core asset for synthesizing high-quality images.

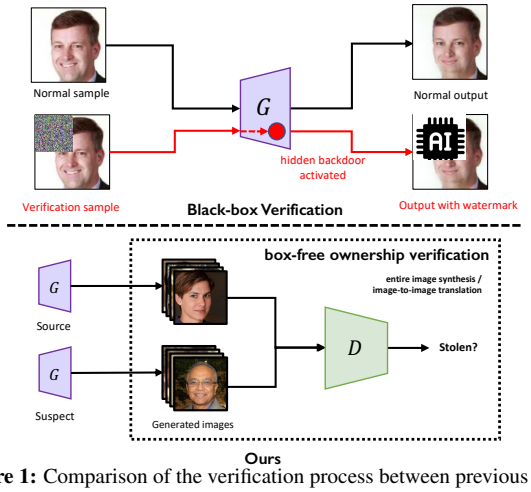

Figure 1: Comparison of the verification process between previous black-box watermark-based verification paradigm [37] and our box-free method. In the black-box setting, carefully-crafted verification samples should be fed to the suspicious model to activate a hidden backdoor in the model (the red circle) and generate watermarked outputs. However, in box-free setting, querying the model with deterministic inputs is not allowed. Ownership verification should be done with only output images.

Training a decent GAN requires a huge investment of resources, such as computing resources, labeled/unlabeled training dataset, time, and human labors [47, 31]. However, well-trained generators are under the threat of unintentional leakage and theft. The adversary may deploy the stolen model on the Internet for profit and the owner (also the *defender*) is only able to verify the ownership remotely by querying the suspicious model [35, 37, 23, 14].

Most existing works on IP protection of DNN models as-