DECENTRALIZED ATTRIBUTION OF GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

There have been growing concerns regarding the fabrication of contents through generative models. This paper investigates the feasibility of decentralized attribution of such models. Given a set of generative models learned from the same dataset, attributability is achieved when a public verification service exists to correctly identify the source models for generated content. Attribution allows tracing of machine-generated content back to its source model, thus facilitating IP-protection and content regulation. Existing attribution methods are non-scalable with respect to the number of models and lack theoretical bounds on attributability. This paper studies decentralized attribution, where provable attributability can be achieved by only requiring each model to be distinguishable from the authentic data. Our major contributions are the derivation of the sufficient conditions for decentralized attribution and the design of keys following these conditions. Specifically, we show that decentralized attribution can be achieved when keys are (1) orthogonal to each other, and (2) belonging to a subspace determined by the data distribution. This result is validated on MNIST and CelebA. Lastly, we use these datasets to examine the trade-off between generation quality and robust attributability against adversarial post-processes.

1 INTRODUCTION

Recent advances in generative models (Goodfellow et al., 2014) have enabled the creation of synthetic contents that are indistinguishable even by naked eyes (Pathak et al., 2016; Zhu et al., 2017; Zhang et al., 2017; Karras et al., 2017; Wang et al., 2018b; Brock et al., 2018; Miyato et al., 2018; Choi et al., 2018; Karras et al., 2019a;b; Choi et al., 2019). Such successes raised serious concerns (Kelly, 2019; BRELAND, 2019) regarding adversarial applications of generative models, e.g., for the fabrication of user profiles (Satter, 2019), articles (Keskar et al., 2019), images (Korshunova et al., 2017), audios (Kumar et al., 2019), and videos (Wang et al., 2018a; Fried et al., 2019). Necessary measures have been called for the filtering, analysis, tracking, and prevention of malicious applications of generative models before they create catastrophic sociotechnical damages.

Existing studies primarily focused on the *detection* of machine-generated contents. Marra et al. (2019) showed empirical evidence that generative adversarial networks (GANs) may come with *data-specific* fingerprints in the form of averaged residual over the generated distribution, yet suggested that generative models trained *on similar datasets* may not be uniquely distinguishable through fingerprints. Yu et al. (2018) showed on the other hand that it is empirically feasible to attribute a *finite* and *fixed* set of GAN models derived from the same dataset, i.e., correctly classifying model outputs by their associated GANs. While encouraging, their study did not

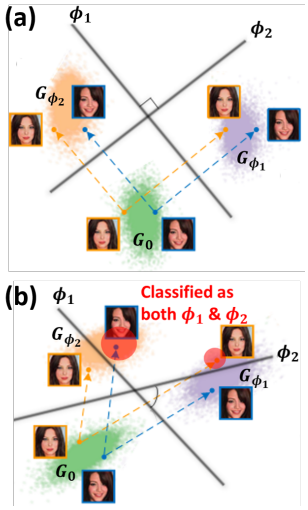


Figure 1: (a) With orthogonal keys (ϕ_1 and ϕ_2), user-end generative models (G_{ϕ_1} and G_{ϕ_2}) achieve attributability as long as they are distinguishable from the authentic dataset (G_0). (b) Keys with acute angles cannot guarantee attributability of G_{ϕ_1} and G_{ϕ_2}.

prove that attribution can be achieved when the model set continues to grow (e.g., when GAN models are distributed to end users in the form of mobile apps). In fact, Wang et al. (2019b) showed that detectors trained on one generative model are transferable to other models trained on the same dataset, indicating that individually trained detectors may perform incorrect attribution, e.g., by attributing images from one model belonging to user A to another model belonging to user B. It should be highlighted that most of the existing detection mechanisms are *centralized*, i.e., the detection relies on a registry that collects all models and/or model outputs and empirically look for collection-wise features that facilitate detection. This fundamentally limits the scalability of detection tools in real-world scenarios where an ever growing number of models are being developed even for the same dataset.

Problem formulation and protocol We are thus motivated to investigate the feasibility of a *decentralized* approach to ensuring the correct attribution of generative models. Specifically, we assume that for a given dataset $\mathcal{D}$, the registry only distributes *keys*, $\Phi := \{\phi_1, \phi_2, ...\}$, to users of generative models without collecting information from the users' models. Each key is held privately by a user, whose key-dependent model is denoted by $G_\phi(\cdot;\theta) : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$ where z and x are the latent and output variables, respectively, and d_z and d_x the corresponding dimensionalities. θ denotes the model parameters. When necessary, we will suppress θ and ϕ to reduce notational burden. The distribution of each key is accompanied by that of a public verification service, which tells whether a query belongs to G_ϕ (labeled as 1) or not (labeled as -1). We call the underlying binary classifier a *verifier* and denote it by $f_\phi : \mathbb{R}^{d_x} \to \{-1, 1\}$. In this paper, we model the *verifier* as linear classifier: $f_\phi(x) = \text{sign}(\phi^T x)$. On multiple GAN benchmarking sets, we show the linear verifier is effective.

Movivating example A company develops a GAN model for image post-processing. A third-party organization (the registry) assigns keys to the company, who is then required to embed a watermark to the GAN models according to the keys for users to download. With the keys, the registry can trace the GAN generated images back to the user-end models.

The following quantities are central to our investigation: the *distinguishability* of G_ϕ is defined as

$$D(G_\phi) := \frac{1}{2}\mathbb{E}_{x\sim P_{G_\phi}, x_0\sim P_{\mathcal{D}}}\left[\mathbb{1}(f_\phi(x) = 1) + \mathbb{1}(f_\phi(x_0) = -1)\right], \tag{1}$$

where $P_{\mathcal{D}}$ is the authentic data distribution, and P_{G_ϕ} the model distribution induced by G_ϕ.

The *attributability* of a collection of generative models $\mathcal{G} := \{G_1, ..., G_N\}$ is defined as:

$$A(\mathcal{G}) := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x\sim G_{\phi_i}}\mathbb{1}(\phi_j^T x < 0, \forall\, j \neq i, \phi_i^T x > 0). \tag{2}$$

Distinguishability of G (attributability of $\mathcal{G}$) is achieved when $D(G) = 1$ ($A(\mathcal{G}) = 1$). Lastly, We denote by $G(\cdot;\theta_0)$ (or shortened as G_0) the root model sent to all users along the key, and assume $P_{G_0} = P_{\mathcal{D}}$. We will measure the (lack of) *generation quality* of G_ϕ by the FID score (Heusel et al., 2017) and the l_2 norm of the mean output perturbation:

$$\Delta x(\phi) = \mathbb{E}_{z\sim P_z}[G_\phi(z;\theta) - G(z;\theta_0)], \tag{3}$$

where P_z is the latent distribution.

This paper addresses the following critical question: *What are the rules for designing keys according to the proposed protocol, so that the resultant generative models can achieve distinguishability individually and attributability collectively, while maintaining their generation quality?*

Contributions We claim the following contributions:

1. We develop sufficient conditions for distinguishability and attributability, to connect these metrics with the geometry of the data distribution, the angles between keys, and the generation quality.
2. The sufficient conditions yield simple design rules for the keys: keys should be (1) data compliant, i.e., $\phi^T x < 0$ for $x \sim P_{\mathcal{D}}$, and (2) orthogonal to each other.
3. We validate the design rules and study the capacity of keys using DCGAN (Radford et al., 2015) on MNIST (LeCun & Cortes, 2010) and CelebA (Liu et al., 2015).

4. We empirically test tradeoffs between generation quality and robust attributability under post-processes including random image blurring, cropping, noising, JPEG conversion, and a combination of all, and show that robust attributability can be achieved, with limited and hard-to-perceive loss of generation quality.

2 SUFFICIENT CONDITIONS FOR ATTRIBUTABILITY

Distinguishability through watermarking We consider *data-compliant* keys $\phi \in \mathbb{R}^{d_x}$ such that $f_\phi(x) = -1$ for all $x \sim P_{\mathcal{D}}$, i.e., the authentic data are correctly attributed as not belonging to any of the generators. All keys are constrained by $||\phi|| = 1$ for identifiability, where $|| \cdot ||$ is the l_2 norm. A key-dependent generative model G_ϕ achieves distinguishable output distribution P_{G_ϕ} by adding a uniform and bounded perturbation Δx to the output of the root model G_0. We can solve the following problem with respect to Δx to achieve distinguishability:

$$\min_{||\Delta x|| \leq \varepsilon} \quad \mathbb{E}_{x \sim P_{\mathcal{D}}} \left[\max\{\phi^T(x + \Delta x), 0\} \right], \tag{4}$$

where $\varepsilon > 0$. We have the following proposition (proof in Appendix A):

Proposition 1. *Let $d_{max}(\phi) := \max_{x \sim P_{\mathcal{D}}} |\phi^T x|$. If $\varepsilon \geq d_{max}(\phi)$, then $\Delta x^* = d_{max}(\phi)\phi$ solves Eq. (4), and $f_\phi(x + \Delta x^*) > 0, \ \forall\, x \sim P_{\mathcal{D}}$.*

Proposition 1 shows that to achieve distinguishability, there is a data geometry-dependent threshold on generation quality (ϵ).

Watermarking user-end models The perturbation Δx^* can potentially be reverse engineered and removed when generative models are white-box to end users. Therefore, we propose to instead retrain the user-end models G_ϕ on the perturbed dataset $\mathcal{D}_{\gamma,\phi} := \{x + \gamma\phi \mid x \sim G_0\}$ with some $\gamma > 0$. Different from Proposition 1, we will show in Theorem 1 that γ needs to be larger than $d_{\text{max}}(\phi)$ in order to guarantee distinguishability of G_ϕ. To explain, note that G_ϕ is trained to match $(z, G_\phi(z; \theta))$ with $(z, G(z; \theta_0) + \gamma(\phi)\phi$ in a supervised manner. Therefore, we introduce the following model:

$$G_\phi(z; \theta) = G(z; \theta_0) + \gamma\phi + \epsilon, \tag{5}$$

where the random error $\epsilon \sim \mathcal{N}(\mu, \Sigma)$. Sec. 3 provides estimates of μ and Σ using sampled ϵs from DCGANs on MNIST and CelebA.

We have the following sufficient condition for model distinguishability (proof in Appendix B):

Theorem 1. *Let $d_{max}(\phi) = \max_{x \in \mathcal{D}} |\phi^T x|$, $\sigma^2(\phi) = \phi^T \Sigma \phi$, $\delta_d \in [0, 1]$, and ϕ be a data-compliant key. $D(G_\phi) \geq 1 - \delta_d/2$ if*

$$\gamma \geq \sigma(\phi)\sqrt{\log\left(\frac{1}{\delta_d^2}\right)} + d_{max}(\phi) - \phi^T \mu. \tag{6}$$

Remarks The computation of $\sigma(\phi)$ requires G_ϕ, which in turn requires γ. Therefore, an iterative search will be needed to determine γ that is small enough to limit the loss of generation quality, yet large enough for distinguishability (see Alg. 1).

Attributability We can now derive the sufficient conditions for attributability of the generative models from a set of N keys (proof in Appendix C):

Theorem 2. *Let $d_{min} = \min_{x \in \mathcal{D}} |\phi^T x|$, $d_{max} = \max_{x \in \mathcal{D}} |\phi^T x|$, $\sigma^2(\phi) = \phi^T \Sigma \phi$, $\delta_a \in [0, 1]$. $A(\mathcal{G}) \geq 1 - N\delta$ if $D(G) \geq 1 - \delta$ for all $G_\phi \in \mathcal{G}$ and for any pair of data-compliant keys ϕ and ϕ':*

$$\phi^T \phi' \leq -1 + \frac{d_{max}(\phi') + d_{min}(\phi') - 2\phi'^T \mu}{\sigma(\phi')\sqrt{\log\left(\frac{1}{\delta_a^2}\right)} + d_{max}(\phi') - \phi'^T \mu}. \tag{7}$$

Remarks When $\sigma(\phi')$ is negligible for all ϕ', RHS of Eq. (7) is approximately $d_{\text{min}}(\phi')/d_{\text{max}}(\phi') > 0$, in which case $\phi^T \phi' \leq 0$ is sufficient for attributability. Later, we empirically show that this is the case for MNIST and CelebA in Section 3.

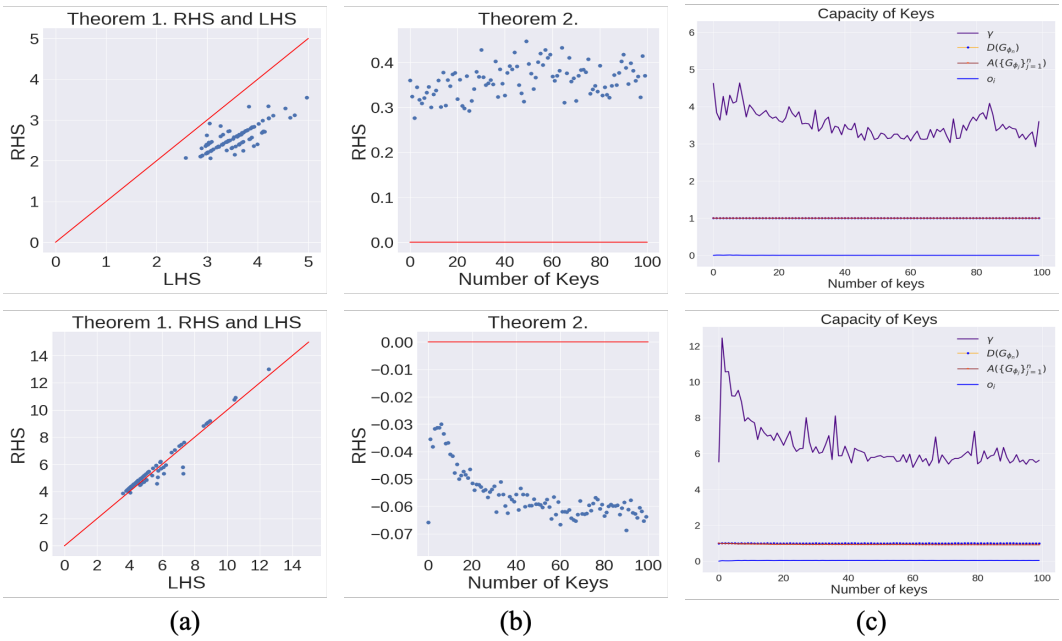


Figure 2: (a) Validation of Theorem 1. RHS=LHS on the red line. (b) Validation of Theorem 2. All RHS of Eq. (7) are positive. (c) For 100 keys, all trained keys are orthogonal. $D(G_{\phi_i})$ and $A(\{G_{\phi_j}\}_{j=1}^{i})$ are close to 1.

3 Experiments and Analysis

The theorems suggest that data-compliant and orthogonal keys guarantee an attributability lower bound. In this section we validate the theorems through experiments.

Key generation We generate keys by iteratively solving the following convex problem starting from an empty key set Φ:

$$\phi_i = \arg\min_{\phi} \; \mathbb{E}_{x\sim P_D, G_0}\left[\max\{1+\phi^T x, 0\}\right] + \sum_{j=1}^{i-1} \max\{\phi_j^T \phi, 0\}. \tag{8}$$

The orthogonality penalty is omitted for the first key. The resultant keys will be normalized before inserted into the next problems. We note that $P_{\mathcal{D}}$ and P_{G_0} do not perfectly match in practice, and therefore expectations are taken over both distributions for the data compliance requirement. G_0s are trained using the standard DCGAN architecture on MNIST and CelebA using ADAM with learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.99$. Training details are in Appendix E.

User-end generative models The training of G_ϕ follows Alg. 1, where we fine-tune γ to balance generation quality and distinguishability. For each γ, we collect a perturbed dataset $\mathcal{D}_{\gamma,\phi} := \{(z, G(z;\theta_0) + \gamma\phi)\}$ with $z \sim P_z$, and solve the following training problem:

$$\min_{\theta} \mathbb{E}_{(z,x)\sim\mathcal{D}_{\gamma,\phi}}\left[||G_\phi(z;\theta) - x||^2\right], \tag{9}$$

starting with $\theta = \theta_0$. If the resultant model does not meet the distinguishability requirement (due to discrepancy between $\mathcal{D}_{\gamma,\phi}$ and G_ϕ), we extend the perturbation by $\gamma = \alpha\gamma$. In experiments, we use a standard normal distribution for P_z, $\delta = 10^{-2}$, and $\alpha = 1.1$.

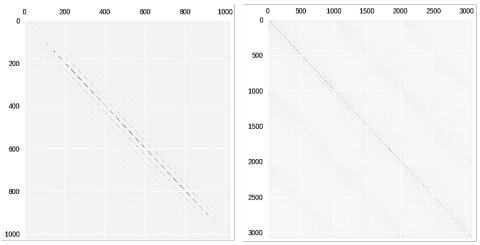
Validation of Theorem 1 We first validate the sufficient condition for distinguishability. Fig. 2(a) compares the LHS and RHS values of Eq. (6) for 100 keys, along with the empirical distinguishability of the corresponding G_ϕs. The mean distinguishability and the standard deviation are $1 - 10^{-4}$ and 3×10^{-4}, respectively. Calculation of the RHS requires an estimation of μ and Σ. To do this, we sample

$$\epsilon(z) = G_\phi(z;\theta) - G(z;\theta_0) - \gamma\phi \tag{10}$$

using 5000 samples $z \sim P_z$ with G_ϕ and γ output from Alg. 1. Σ and μ are then estimated by the sampled ϵs for each ϕ. Figure below demonstrates the estimates of Σ. We note that μs have elements close to zero for the test datasets. Results in Fig. 2(a) show that the sufficient condition (Eq. (6)) is satisfied for all G_ϕs through the training specified in Alg. 1. Lastly, we notice that the LHS values for MNIST are farther away from the equality line than those for CelebA. This is due to the fact that perturbations for samples from MNIST are more likely to exceed the bounds for pixel values due to their binarized values in $\mathcal{D}$. Clamping of these invalid pixel values reduces the effective perturbation length. Therefore to achieve distinguishability, Alg. 1 seeks γs larger than needed. This issue is less observed in CelebA, in which case data points in $\mathcal{D}$ rarely have values close to boundaries.

Algorithm 1: Training of G_ϕ

input : ϕ, G_0
output: G_ϕ, γ
1 set $\gamma = d_{\max}(\phi)$;
2 collect $\mathcal{D}_{\gamma,\phi}$;
3 train G_ϕ by solving Eq. (9) using $\mathcal{D}_{\gamma,\phi}$;
4 compute empirical $D(G_\phi)$;
5 **if** $D(G_\phi) < 1 - \delta$ **then**
6 set $\gamma = \alpha\gamma$;
7 goto step 2 ;
8 **end**



(a) MNIST (b) CelebA

Sample covariance matrix of errors

Validation of Theorem 2 Recall that from Theorem 2, we derive the simple design rule that keys should be orthogonal, based on the observation that the RHS of Eq. (7) is almost always positive. Here we empirically test this assumption, as we plot the RHS values derived from 100 keys in Fig. 2(b), where the values are all positive for MNIST and close to zero for CelebA. The result suggests that orthogonality of keys is a good heuristic for attributability. Note that the condition is sufficient thus could be more stringent than necessary. We conjecture that this is the reason why attributability is empirically achieved with orthogonal keys. Lastly, while it is feasible to further reduce the lower bound on the angles between keys, e.g., some acute angles will still facilitate attribution in MNIST, doing so would require the derivation of new keys to rely on knowledge about existing user-end models (to compute the RHS of Eq. (7)).

Empirical results on distinghuishability, attributability, and generation quality Table 1 reports the metrics of interest empirically measured on 20 user-end models for each dataset. Specifically, we achieve high attributability while all models are trained separately only for distinguishability. As a comparison, we demonstrate cases where keys are $45\deg$ apart ($\phi^T\phi = 0.71$). Note that due to the perturbations, all G_ϕs have worse generation quality than G_0 (i.e., higher FIDs than those of the root models).

Capacity of keys For real-world applications, we hope to maintain attributability for a large set of keys. Our study so far suggests that the capacity of keys is constrained by the data compliance and orthogonality requirements. While sequentially computing keys is convex and affordable (Eq. (8)), finding the maximum number of feasible keys is a problem about optimal sphere packing on a manifold (Fig. 3), which is an open challenge (Cohn et al. (2017); Cohn (2016)). Here we use MNIST on DCGAN to empirically explore the capacity of keys. Results on 100 keys are reported in Fig. 2(c) using the following metrics with respect to an increasing number of keys n: average orthogonality $o_n = \sum_{j=1}^{n-1} |\phi_j^T \phi_n|/(n-1)$ ($o_1 = 0$), perturbation length γ_n, distinguishability $D(G_{\phi_n})$, and attributability $A(\{G_{\phi_j}\}_{j=1}^n)$, for $n = 1, ..., 100$. We show that attributability is almost always 1 while all models are trained only for distinguishability.

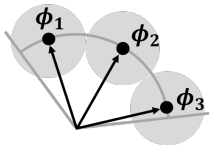


Figure 3: Capacity of keys as a sphere packing problem: The feasible space (arc) is determined by the data compliance and generation quality constraints, and the size of spheres by the minimal angle between keys.

Robust training We now consider the scenario where outputs of the generative models are post-processed (by adversaries) before being attributed. When the post-processes are known, we can take counter measures through robust training, which intuitively, will lead to additional loss of generation quality. To assess this tradeoff between robustness and generation quality, we train G_ϕ against post-

Table 1: Empirical average of distinguishability ($\bar{D}$), $||\Delta x||$, FID scores, and attributability ($A(\mathcal{G})$) from 20 generative models for each dataset. Std in parenthesis (all close to zero for distinguishability). FID_0 is the FID for G_0.

GANs	Angle	Dataset	$\bar{D}$	$A(\mathcal{G})$	$\|\|\Delta x\|\|$	FID_0	FID
DCGAN	Orthogonal	MNIST	0.99	**0.98**	5.05 (0.09)	4.98 (0.15)	5.36 (0.12)
DCGAN	45 degree	MNIST	0.99	0.13	5.39 (0.14)	-	5.42 (0.13)
DCGAN	Orthogonal	CelebA	0.99	**0.98**	5.63 (0.11)	33.95 (0.13)	53.91 (2.20)
DCGAN	45 degree	CelebA	0.99	0.14	5.78 (0.21)	-	53.92 (2.56)

Table 2: 1st-4th rows: distinguishability before (Bfr) and after (Aft) robust training, attributability Bfr and Aft, $||\Delta x||$ after robust training, and FID after robust training. Std in parenthesis. Higer is better for distinguishability and attributability. Lower is better for $||\Delta x||$ and FID.

GANs	Dataset	`Blurring`		`Cropping`		`Noise`		`JPEG`		`Combination`	
-	-	Bfr	Aft	Bfr	Aft	Bfr	Aft	Bfr	Aft	Bfr	Aft
DCGAN	MNIST	0.49	0.96	0.52	0.99	0.85	0.99	0.54	0.99	0.50	0.66
DCGAN	CelebA	0.49	0.99	0.49	0.99	0.95	0.98	0.51	0.99	0.50	0.85
DCGAN	MNIST	0.02	0.94	0.03	0.88	0.77	0.95	0.16	0.98	0.00	0.26
DCGAN	CelebA	0.00	0.98	0.00	0.99	0.89	0.93	0.07	0.98	0.00	0.70
DCGAN	MNIST	15.96(2.18)		9.17(0.65)		5.93(0.34)		6.48(0.94)		17.08(1.86	
DCGAN	CelebA	11.83(0.65)		9.30(0.31)		4.75(0.17)		6.01(0.29)		13.69(0.59)	
DCGAN	MNIST	41.11(20.43)		21.58(2.44)		5.79(0.19)		6.50(1.70)		68.16(24.67)	
DCGAN	CelebA	73.62(6.70)		98.86(9.51)		59.51(1.60)		60.35(2.57)		87.29(9.29)	

processes $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ from a distribution P_T. Due to the potential nonlinearity of T and the lack of theoretical guarantee in this scenario, we resort to the following robust training problem:

$$\min_{\theta_i} \mathbb{E}_{z\sim P_z, T\in P_T}\left[\max\{1 - f_{\phi_i}(T(G_{\phi_i}(z;\theta_i))), 0\} + C||G_0(z) - G_{\phi_i}(z;\theta_i)||^2\right], \tag{11}$$

where $C = 10$ is the hyper-parameter for a penalty on the generation quality. We consider five types of post-processes: blurring, cropping, noise, JPEG conversion and the combination of these four. Examples of the post-processed images are shown in Fig. 4. `Blurring` uses Gaussian kernel widths uniformly drawn from $\frac{1}{3}\{1, 3, 5, 7, 9\}$. `Cropping` crops images with uniformly drawn ratios between 80% and 100%, and scales the cropped images back to the original size using bilinear interpolation. `Noise` adds white noise with standard deviation uniformly drawn from $[0, 0.3]$. `JPEG` applies JPEG compression. `Combination` performs each attack with a 50% chance in the order of `Blurring`, `Cropping`, `Noise` and `JPEG`. We use implementations for differentiable blurring (Riba et al. (2020)) and JPEG (Zhu et al. (2018)). For robust training against each post-process, we apply the post-process to mini-batches with 50% probability.

Table 2 compares the average distinguishability and the attributability with and without robust training of G_ϕ, when post-processes are considered. The results are based on 20 generative models for each architecture-dataset pair, where keys are kept as orthogonal and data compliant. From the results, defense against all post-processes can be achieved, except for `Combination`. Finally, Table 2 demonstrates the tradeoff between robustness and generation quality, using FID as a metric. Fig. 4 demonstrates this tradeoff through examples from DCGANs on CelebA.

Generation Quality Control The loss of generation quality is directly related to the length of perturbation (γ), and from Theorem 1, the value of $d_{\max}$. Fig. 5 compares outputs from two user-end generative models with different $d_{\max}$s. While it is possible to filter ϕs based on their corresponding $d_{\max}$s for generation quality control, here we discuss a potential direction for prescribing a subspace of ϕs within which quality can be controlled. To start, we denote by $J(x)$ the Jacobian of G_0 with respect to its generator parameters θ_0. Notice that elements of $J(x)$ are dependent on the particular sample x. Our finding is specifically related to the matrix $M = \mathbb{E}_{x\sim P_{G_0}}[J(x)]\mathbb{E}_{x\sim P_{G_0}}[J(x)^T]$. A spectral analysis of M reveals that the eigenvectors of M with large eigenvalues are more structured than those with small ones (Fig. 6(a)). This finding is consistent with the definition of M: The largest eigenvectors of M represent the principal axes of all mean sensitivity vectors, where the

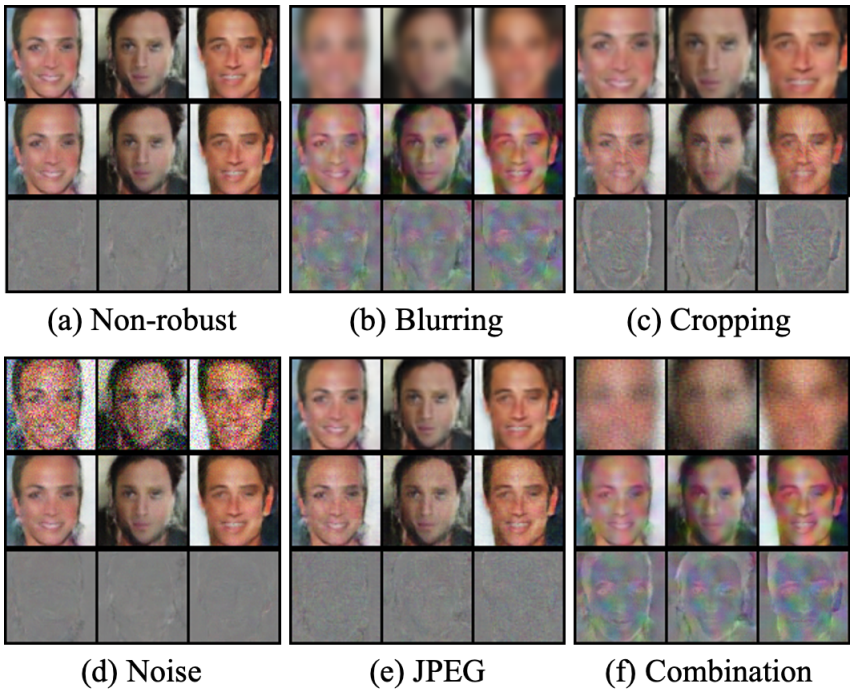


Figure 4: (a) 1st row: samples from G_0. 2nd row: non-robust G_ϕ (b-f) 1st row: worst-case post-process. 2nd row: samples from robustly trained G_ϕ against the post-processes (prior to the post-processes). 3rd row for all: differences between 2nd row of (a) and 2nd rows of others. See Appendix D for MNIST examples.

(a) d_{max}=1.66 (b) d_{max}= 2.92 (c) d_{max}= 3.16 (d) d_{max}=12.59

Figure 5: MNIST and CelebA examples from ϕs with (a-b) small d_{max} and (c-d) large d_{max}.

mean is taken over the latent space. For MNIST, these eigenvectors overlap with the digits; for CelebA, they are structured color patterns. On the other hand, the smallest eigenvectors represent directions rarely covered by the sensitivity vectors, thus resembling random noise. Based on this finding, we test the hypothesis that keys more aligned with the eigenspace of the small eigenvalues will have smaller $d_{\max}$. We compute that the correlations between $d_{\max}$ and $\phi^T M \phi$: the pearson correlations are 0.33 and 0.53 for MNIST and CelebA, respectively. Lastly, we compare outputs of generative models using the largest and the smallest eigenvectors of M as the keys in Fig. 6.

4 Related Work

Detection and attribution of model-generated contents This paper focused on the attribution of contents from generative models rather than the detection of hand-crafted manipulations (Agarwal & Farid (2017); Popescu & Farid (2005); O'brien & Farid (2012); Rao & Ni (2016); Huh et al. (2018)). In this context, it has been shown that generative models based on convolutional neural network leave fingerprints in their outputs (Odena et al. (2016)), allowing them to be detected (Zhang et al. (2019b); Marra et al. (2019); Wang et al. (2019a)). However, generative models trained on similar datasets have similar fingerprints (Marra et al. (2019)). Thus fingerprints cannot be used for attribution. Recent studies show that attribution is achievable through a centralized manner, where multiclass classifiers are trained on ***all*** models (Yu et al. (2018); Albright et al. (2019)). However, such classifiers are not proven to exist, and computing them is not practical when the number of

Figure 6: (a) Eigenvectors for the two largest and two smallest eigenvalues of M for DCGANs on MNIST (top) and CelebA (bottom) (b) Left to right: Samples from G_0 and subtraction of $G_0 - G_{eigenvectors}$.

models grows arbitrarily. Skripniuk et al. (2020) studied decentralized attribution. Instead of using linear classifiers for attribution, they train a watermark encoder-decoder network that embeds (and reads) watermarks into (and from) the content, and compare the decoded watermark with user-specific ones. They do not provide sufficient conditions of the watermarks for distinguishability and attributability. Their empirical attributability is measured on a set of six generative models for CelebA. In comparison, we do so on 100 DCGANs on MNIST and CelebA.

IP protection of digital contents Watermarks have conventionally been used for IP protection of digital contents Tirkel et al. (1993); Van Schyndel et al. (1994); Bi et al. (2007); Hsieh et al. (2001); Pereira & Pun (2000); Zhu et al. (2018); Zhang et al. (2019a). In this context, watermarks are often added to contents as a post-process, and therefore can be removed when the models are white-box to users. Another approach towards content IP protection is through blockchain Hasan & Salah (2019). However, this approach requires meta data to be transferred along with the contents, which may not be realistic in adversarial settings. E.g., one can simply take a picture of a synthetic image to remove any meta data attached to the image file.

IP protection of machine learning models Aside from IP protection of *contents*, mechanisms for protecting IP of discriminative and generative *models* have also been studied Uchida et al. (2017); Nagai et al. (2018); Le Merrer et al. (2019); Adi et al. (2018); Zhang et al. (2018); Fan et al. (2019); Szyller et al. (2019); Zhang et al. (2020). Model watermarking is usually done by adding watermarks into model weights Uchida et al. (2017); Nagai et al. (2018), by embedding unique input-output mapping into the model Le Merrer et al. (2019); Adi et al. (2018); Zhang et al. (2018), or by introducing a passport mechanism so that model accuracy drops if the right passport is not inserted Fan et al. (2019). In the wake of model extraction attacks, where the attacker can reverse engineer the model from its output, existing solutions resort to watermarking of outputs Szyller et al. (2019); Zhang et al. (2020). While closely related, existing work on model IP protection focused on the distinguishability of individual models, rather than the attributability of a model set.

5 Conclusion

This paper investigated the feasibility of decentralized attribution for generative models. We put forward a protocol where a registry generates keys to watermark user-end models, for which the outputs can be correctly attributed by the registry. Our investigation led to simple design rules for the design of the keys to achieve correct attribution while only requiring each user-end model to be distinguishable from the authentic dataset. With concerns about adversarial post-processes, we empirically show that robust attribution can be achieved using the developed design rules, with additional loss of generation quality.

REFERENCES

Yossi Adi, Carsten Baum, Moustapha Cisse, Benny Pinkas, and Joseph Keshet. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pp. 1615–1631, 2018.

Shruti Agarwal and Hany Farid. Photo forensics from jpeg dimples. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2017.

Michael Albright, Scott McCloskey, and ACST Honeywell. Source generator attribution via inversion. *arXiv preprint arXiv:1905.02259*, 2019.

Ning Bi, Qiyu Sun, Daren Huang, Zhihua Yang, and Jiwu Huang. Robust image watermarking based on multiband wavelets and empirical mode decomposition. *IEEE Transactions on Image Processing*, 16(8):1956–1966, 2007.

ALI BRELAND. The bizarre and terrifying case of the "deepfake" video that helped bring an african nation to the brink. *motherjones*, Mar 2019. URL `https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/`.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. *arXiv preprint arXiv:1912.01865*, 2019.

Henry Cohn. A conceptual breakthrough in sphere packing. *arXiv preprint arXiv:1611.01685*, 2016.

Henry Cohn, Abhinav Kumar, Stephen D Miller, Danylo Radchenko, and Maryna Viazovska. The sphere packing problem in dimension 24. *Annals of Mathematics*, pp. 1017–1033, 2017.

Lixin Fan, Kam Woh Ng, and Chee Seng Chan. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. In *Advances in Neural Information Processing Systems*, pp. 4716–4725, 2019.

Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Haya R Hasan and Khaled Salah. Combating deepfake videos using blockchain and smart contracts. *Ieee Access*, 7:41596–41606, 2019.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

Ming-Shing Hsieh, Din-Chang Tseng, and Yong-Huai Huang. Hiding digital watermarks using multiresolution wavelet transform. *IEEE Transactions on industrial electronics*, 48(5):875–882, 2001.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A Efros. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 101–117, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019b.

Makena Kelly. Congress grapples with how to regulate deepfakes. *Congress grapples with how to regulate deepfakes*, Jun 2019. URL https://www.theverge.com/2019/6/13/18677847/deep-fakes-regulation-facebook-adam-schiff-congress-artificial-intelligence.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3677–3685, 2017.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, pp. 14881–14892, 2019.

Erwan Le Merrer, Patrick Perez, and Gilles Trédan. Adversarial frontier stitching for remote neural network watermarking. *Neural Computing and Applications*, pp. 1–12, 2019.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 506–511. IEEE, 2019.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Yuki Nagai, Yusuke Uchida, Shigeyuki Sakazawa, and Shin'ichi Satoh. Digital watermarking for deep neural networks. *International Journal of Multimedia Information Retrieval*, 7(1):3–16, 2018.

James F O'brien and Hany Farid. Exposing photo manipulation with inconsistent reflections. *ACM Trans. Graph.*, 31(1):4–1, 2012.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003. URL http://distill.pub/2016/deconv-checkerboard.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.

Shelby Pereira and Thierry Pun. Robust template matching for affine resistant image watermarks. *IEEE transactions on image Processing*, 9(6):1123–1129, 2000.

Alin C Popescu and Hany Farid. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on signal processing*, 53(2):758–767, 2005.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. URL http://arxiv.org/abs/1511.06434. cite arxiv:1511.06434Comment: Under review as a conference paper at ICLR 2016.

Yuan Rao and Jiangqun Ni. A deep learning approach to detection of splicing and copy-move forgeries in images. In *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2016.

E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch, 2020. URL https://arxiv.org/pdf/1910.02190.pdf.

Raphael Satter. Experts: Spy used ai-generated face to connect with targets. *Experts: Spy used AI-generated face to connect with targets*, Jun 2019. URL https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

Vladislav Skripniuk, Ning Yu, Sahar Abdelnabi, and Mario Fritz. Black-box watermarking for generative adversarial networks. *arXiv preprint arXiv:2007.08457*, 2020.

Sebastian Szyller, Buse Gul Atli, Samuel Marchal, and N Asokan. Dawn: Dynamic adversarial watermarking of neural networks. *arXiv preprint arXiv:1906.00830*, 2019.

Anatol Z Tirkel, GA Rankin, RM Van Schyndel, WJ Ho, NRA Mee, and Charles F Osborne. Electronic watermark. *Digital Image Computing, Technology and Applications (DICTA'93)*, pp. 666–673, 1993.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pp. 269–277, 2017.

Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne. A digital watermark. In *Proceedings of 1st International Conference on Image Processing*, volume 2, pp. 86–90. IEEE, 1994.

Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 10072–10081, 2019a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. *arXiv preprint arXiv:1912.11035*, 2019b.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018b.

Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Analyzing fingerprints in generated images. *arXiv preprint arXiv:1811.08180*, 2018.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pp. 159–172, 2018.

Jie Zhang, Dongdong Chen, Jing Liao, Han Fang, Weiming Zhang, Wenbo Zhou, Hao Cui, and Nenghai Yu. Model watermarking for image processing networks. In *AAAI*, pp. 12805–12812, 2020.

Kevin Alex Zhang, Alfredo Cuesta-Infante, Lei Xu, and Kalyan Veeramachaneni. Steganogan: High capacity image steganography with gans. *arXiv preprint arXiv:1901.03892*, 2019a.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. *arXiv preprint arXiv:1907.06515*, 2019b.

Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 657–672, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.