Decentralized Attribution of Generative Models

An Anonymous Preprint

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 group of models derived from the same dataset and published by different users, attributability is achieved when a public verification service associated with each model (a linear classifier) returns positive only for outputs of that model. Attribution allows tracing of machine-generated contents back to its source model, thus facilitating IP-protection and content regulation. Decentralized attribution prevents forgery of source models by only allowing users to have access to their own classifiers, which are parameterized by keys distributed by a registry. Our major contribution is the development of design rules for the keys, which are derived from first-order sufficient conditions for decentralized attribution. Through validation on MNIST, CelebA and Cityscapes, we show that keys need to be (1) orthogonal or opposite to each other and (2) belonging to a subspace dependent on the data distribution and the architecture of the generative model. We also empirically examine the trade-off between generation quality and robust attributability against adversarial post-processes of model outputs.

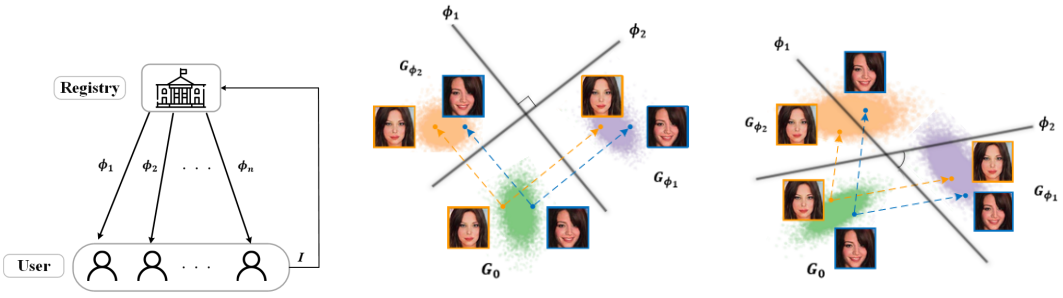
1 Introduction

Recent advances in generative models [1] have enabled the creation of synthetic contents that are indistinguishable even by naked eyes [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. Such successes raised serious concerns [13, 14] regarding adversarial applications of generative models, e.g., for the fabrication of user profiles [15], articles [16], images [17], audios [18], and videos [19, 20, 21]. Necessary measures have been called for the filtering, analysis, tracking, and prevention of malicious applications of generative models before they create catastrophic sociotechnical damages [14].

Existing studies primarily focused on the *detection* of machine-generated contents. Marra et al. [22] 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. [23] 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 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. [24] 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 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

Preprint. Under review.



(a) protocol (b) orthogonal keys (c) acute keys

Figure 1: (a) Protocol of decentralized attribution: Keys are distributed by the registry and used to produce key-dependent generators for individual users. (b) Orthogonal keys (ϕ_1 and ϕ_2) achieve distinguishability and attributability. (c) Acute keys achieve distinguishability but not attributability.

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 are 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 focus on linear classifiers: $f_\phi(x) = \text{sign}(\phi^T x)$. **Example:** The registry (a company) developes a new GAN model for photo post-processing. Individuals download the app that consists of a GAN model and a key. The installation modifies the GAN according to the keys so that the resulting model can be verified. The keys are then deleted from the users' end. All outputs from the user-end models can now be traced back to the users (Fig. 1).

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

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

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

$$A(\mathcal{G}) := \sum_{i=1}^{N} \frac{1}{2N}\left(\mathbb{E}_{x\sim P_{G_{\phi_i}}}\left[\mathbb{1}(f_{\phi_i}(x) = 1)\right] + \frac{1}{N-1}\sum_{j\in\{1,...,N\}\setminus i} \mathbb{E}_{x'\sim P_{G_{\phi_j}}}\left[\mathbb{1}(f_{\phi_i}(x') = -1)\right]\right). \tag{2}$$

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 measure the (lack of) *generation quality* of G_ϕ through both the FID score [25] 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 answers the following question: *What are the rules for designing keys, so that the resultant generative models can achieve distinguishability individually and attributability collectively?*

Contributions We claim the following contributions:

1. We develop first-order sufficient conditions for distinguishability and attributability, to connect these metrics with the geometry of the data distribution, the sensitivity of the generative model, the angles between keys, and the generation quality.
2. The sufficient conditions yield simple design rules for the keys, which should be (1) data compliant, i.e., $f_\phi(x) = -1$ for $x \sim P_{\mathcal{D}}$, (2) orthogonal or opposite to each other, and (3) within a model- and data-dependent subspace to maintain generation quality.

3. We empirically validate the design rules and study the capacity of keys using DCGAN [26], PGAN [5], and CycleGAN [3] on MNIST [27], CelebA [28], and the Cityscape [29] datasets.
4. We empirically test tradeoffs between generation quality and robust attributability under post-processes including image blurring, cropping, noising, JPEG conversion, and a combination of all, and show that robust attributability can be achieved, with degraded yet acceptable generation quality.

Notations Throughout the paper, we denote by $a_{(i)}$ the ith element of vector a, and $A_{(i,j)}$ the (i,j)th element of matrix A. $||a||_H^2 = a^T H a$ for vector a and matrix H. $\nabla_x y|_{x_0}$ is the gradient of y with respect to x, evaluated at $x = x_0$. We use $\text{supp}(P)$ to denote the support of distribution P.

2 Key Design for Distinguishability, Attributability, and Generation Quality

2.1 A toy case

The connections among distinguishability, attributability, and generation quality are illustrated through a toy case with the following settings: (1) *One-hot orthogonal keys*: Let $\phi_i \in \Phi$ be one-hot and $\phi^T \phi' = 0$ for all $\phi \neq \phi'$. (2) *Data compliance*: Let $x \sim P_{\mathcal{D}}$ have negative elements so that $f_\phi(x) = -1$ for all x, i.e., the authentic data is correctly attributed by all verifiers as not belonging to their associated generators. (3) *Distinguishability through output perturbation*: A key-dependent generative model G_ϕ achieves distinguishable output distribution P_{G_ϕ} by adding a fixed and bounded perturbation δ to the output of the root model G_0:

$$\min_{||\delta|| \leq \varepsilon} \quad \mathbb{E}_{x \sim P_{\mathcal{D}}} \left[\max\{1 - (x+\delta)^T \phi, 0\} \right], \tag{4}$$

where $\varepsilon > 0$. The solution to Eq. (4) is $\delta^*(\phi) = \varepsilon \text{sign}(\phi) = \varepsilon \phi$, which yields $||\Delta x|| = ||\delta^*|| = \varepsilon$. With these settings, we have the following proposition (proof in Appendix A):

Proposition 1. *(Toy case) If* $||\Delta x|| > \max_{x \sim P_{\mathcal{D}}} \{||x||_\infty\}$, $D(G_\phi) = 1 \; \forall \; \phi \in \Phi$ *and* $A(\mathcal{G}) = 1$.

While simplistic, Proposition 1 reveals that (1) the lower bound on the degradation of generation quality to suffice distinguishability is dependent on the data geometry, and (2) orthogonality of the keys ensures attributability. These properties are preserved for a more realistic case discussed below.

2.2 A more realistic case

A few modifications are made to the toy case: (1) *Normalized keys*: We consider data-compliant keys $\phi \in \mathbb{R}^{d_x}$ in a convex cone, and constrain $||\phi|| = 1$ for identifiability. (2) *Distinguishability through model parameter perturbation*: The output perturbation in the toy case can be reverse engineered and removed when generative models are white-box to end users. Therefore, we propose to perturb model parameters instead through the following problem:

$$\min_{||\theta - \theta_0|| \leq \varepsilon} \quad \mathbb{E}_{z \sim P_z} \left[\max\{1 - \phi^T G_\phi(z;\theta), 0\} \right]. \tag{5}$$

Distinguishability We start by a first-order analysis, where we assume that for a small ε, Eq. (5) is solved by a gradient descent step: $\Delta\theta = \gamma \mathbb{E}_{x \sim P_{G_0}} \left[\nabla_\theta x^T |_{\theta_0} \right] \phi$ with $\gamma > 0$, and a linear approximation can capture the perturbation from $x_0 = G(z;\theta_0)$ to $x = G(z;\theta)$ for latent z: $x = x_0 + \nabla_\theta x_0 |_{\theta_0} \Delta\theta$. Here we used the data-compliance condition: $1 - \phi^T x > 0$ for $x \sim P_{G_0}$ for the approximation of $\Delta\theta$. To reduce notational burden, we denote by $J(x) := \nabla_\theta x^T |_{\theta_0}$ the Jacobian of G_0 with respect to its parameters, and let $M = \mathbb{E}_{x \sim P_{G_0}}[J(x)] \mathbb{E}_{x \sim P_{G_0}}[J(x)]^T \in \mathbb{R}^{d_x \times d_x}$. The following conjectures about $J(x)$ and M are empirically tested (Appendices C and D):

Conjecture 1. *Let the* (i,j)*th element of* $\Sigma(x) = J(x) \mathbb{E}_{x \sim P_{G_0}}[J(x)]^T - M$ *be* $\Sigma_{(i,j)}$ *with variance* σ_{ij}^2. *Then* $\Sigma_{(i,j)}$ *is approximately drawn independently from* $\mathcal{N}(0, \sigma_{ij}^2)$.

Conjecture 2. *Denote by* $\Lambda = \{\lambda_1, ..., \lambda_{d_x}\}$ *the eigenvalues of* M. *For existing deep generative models, there exists a large subset of similarly small eigenvalues.*

Remarks. $\{\sigma_{ij}^2\}$ *reflects the difficulty of controlling generative models:* Let $J_i(x)^T$ be the ith row of $J(x)$ and $\bar{J}_i = \mathbb{E}_{x \sim P_{G_0}}[J_i(x)]$. $J_i(x)$ represents the sensitivity of the ith element of $x \sim P_{G_0}$

(a) (b)

Figure 2: (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 $G_0 - G_{eigenvectors}$

with respect to θ_0. Let $\Delta J_i = J_i - \bar{J}_i$, then $H_i = \mathbb{E}_{x\sim P_{G_0}}[\Delta J_i \Delta J_i^T]$ is the variance-covariance matrix of $J_i(x)$. Let $\Delta_i(x) = J_i^T(x)\Delta\theta$ be the perturbation along the ith element of x due to $\Delta\theta$, and $\bar{\Delta}_i = \bar{J}_i^T\Delta\theta$ the expected perturbation. Lastly, let $Var(\Delta_i) = ||\Delta\theta||^2_{H_i}$ be the variance of the perturbation. For $\Delta\theta$ with unit norm, we can show that $Var(\Delta_i) = \sigma_{ij}^2/||\bar{J}_j||^2$ when $\Delta\theta$ is chosen to maximize $\bar{\Delta}_j$ ($\Delta\theta = \bar{J}_j/||\bar{J}_j||$). Therefore, σ_{ij}^2 reflects the difficulty of controlling supp(P_{G_ϕ}) through $\Delta\theta$. $\{\sigma_{ij}^2\}$ concentrates at zero for DCGANs on MNIST and CelebA (Appendix C).

The first-order sufficient conditions for model distinguishability is as follows (proof in Appendix E):

Theorem 1. *(Realistic case) Let* $d_{max}(\phi) = \max_{x\sim P_{\mathcal{D}}} |\phi^T x|$, $\sigma^2(\phi) = \sum_{i,j} \sigma_{ij}^2 \phi_{(i)}^2 \phi_{(j)}^2$, *and* δ_d *be a positive number greater than* $\exp\left(-\frac{1}{2}\left(\frac{\phi^T M\phi}{\sigma(\phi)}\right)^2\right)$ *for a data-compliant key* $\phi \in \Phi$. *If*

$$||\Delta x(\phi)|| \geq \frac{d_{max}(\phi)\sqrt{\phi^T M^2 \phi}}{\phi^T M\phi - \sigma(\phi)\sqrt{\log\left(\frac{1}{\delta_d^2}\right)}}, \tag{6}$$

then $D(G_\phi) \geq 1 - \delta_d/2$.

Remarks. Theorem 1 reveals the connection between distinguishability and generation quality: In addition to the data geometry ($d_{\max}$) as in the toy case, the lower bound of the generation quality also depends on model-related properties (M and σ). It should be noted that the lower bound is over approximated when $\phi^T M\phi$ is small: Specifically, Appendix E shows empirically that distinguishability can be achieved even when $\phi^T M\phi$ is small. We hypothesize that this is due to the nonlinear change of $\sigma(\phi)$ along the gradient descent process.

Generation quality Note that the mean perturbation following the first-order analysis is $\Delta x = \mathbb{E}_{x_0\sim P_{G_0}}[x - x_0] = \mathbb{E}_{x\sim P_{G_0}}[\gamma J(x)\mathbb{E}_{x\sim P_{G_0}}[J(x)]\phi] = \gamma M\phi$. We verify through experiments that for ϕ that are eigenvectors of M, $\Delta x \propto \phi$ (Fig. 2b). These together with Theorem 1 lead to the following conjecture consistent with intuition, again tested through experiments (Appendix F):

Conjecture 3. $||\Delta x|| \leq \tau d_{max}$, *where* τ *is finite and dependent on the condition number of* M.

There are two aspects of generation quality that we care about: First, for $||\Delta x||$ to be small, Conjecture 3 suggests that we should pick ϕ with small $d_{\max}$. Second, Spectral analysis of M for MNIST and CelebA shows that ϕs corresponding to large eigenvalues have more structured patterns, while those for small eigenvalues resemble white noise. As a result, keys in the eigenspace of small eigenvalues of M achieve better FID scores and are preferred for maintaining the salient contents of the authentic data. Fig. 2a shows the eigenvectors of the largest and smallest eigenvalues of M for DCGANs on MNIST and CelebA. Fig. 2b are the outputs of the corresponding models that achieve distinguishability.

Attributability The first-order sufficient conditions for attributability are as follows (proof in Appendix H):

Theorem 2. *Let* $d^*_{min} = \min_{\phi\in\Phi, x\sim P_{\mathcal{D}}} |\phi^T x|$, $\bar{\sigma}^2(\phi) = \sqrt{\phi^T V^T V \phi - (\phi^T V \phi)^2}$ *where* $V_{(i,j)} = \sigma^2_{ij}$. *If* $D(G) \geq 1 - \delta_d$ *for all* $G_\phi \in \mathcal{G}$, $\phi^T \phi' \leq 0$ *for all* $\phi, \phi' \in \Phi$, *and*

$$||\Delta x(\phi)|| \leq \frac{d^*_{min}\sqrt{\phi^T M^2 \phi}}{\sqrt{\phi^T M^2 \phi - (\phi^T M \phi)^2} + \bar{\sigma}(\phi)\sqrt{\log\left(\frac{1}{\delta_a^2}\right)}}, \tag{7}$$

for all $\phi \in \Phi$, *then* $A(\mathcal{G}) \geq 1 - (\delta_d + \delta_a)/2$.

Remarks. (1) *Conflict exists between distinguishability and attributability*: The degradation of generation quality is lower bounded for distinguishability yet upper bounded for attributability. This is because the former requires model distributions to be away from $\mathcal{D}$, while the latter requires G_ϕ to stay away from the half spaces $\{x \in \mathbb{R}^{d_x} | \phi'^T x > 0\}$ of all other keys $\phi' \neq \phi$ (see Fig. 1b).

(2) *Attributability is inherently limited by the model architecture*: There are two reasons for G_ϕ to enter $\{x \in \mathbb{R}^{d_x} | \phi'^T x > 0\}$ by moving away from $\mathcal{D}$: (i) P_{G_ϕ} diverges as we perturb θ due to non-zero $\bar{\sigma}^2(\phi)$; (ii) the center of support(P_{G_ϕ}) moves along $M\phi$ rather than ϕ. In the special case where $\bar{\sigma}^2(\phi) = 0$ and $M\phi \propto \phi$ (when M has a condition number of 1), the upper bound on $||\Delta x||$ becomes $+\infty$.

(3) *Keys need to be strictly data compliance*: When $d^*_{\text{min}} = 0$, support($\mathcal{D}$) is tangent to one of the keys. Attributability cannot be achieved unless $\bar{\sigma}^2(\phi) = 0$ and $M\phi \propto \phi$.

(4) $\phi^T \phi' \leq 0$ *implies orthogonal and opposite keys*: $\phi^T \phi' \leq 0$ requires ϕ and ϕ' to have an orthogonal or obtuse angle. Note that for a given vector space, the capacity of keys to satisfy $\phi^T \phi' \leq 0$ for all $\phi \neq \phi'$ is achieved when all keys are orthogonal or opposite to each other. Therefore, we can focus on computing orthogonal keys (and flipping their signs to get the other half).

3 Implementation

The above analysis suggests the following rules for designing keys: (R1) strict data compliance, (R2) orthogonality, (R3) small d_{max}, and (R4) belonging to the eigenspace of M associated with small eigenvalues.

Key generation. The registry computes a sequence of keys to satisfy (R1) and (R2) for decentralized attribution:

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

The orthogonality penalty is omitted for the first key. Some remarks: (1) For fast computation of keys, we convexify Eq. (8) by removing the unit norm constraint. Each key is normalized right after solving the relaxed problem. (2) $P_{\mathcal{D}}$ and P_{G_0} do not perfectly match in practice, and therefore expectations are taken over both distributions. (3) We use a hinge loss to promote strict data compliance. (4) Computation of d_{max} requires minimax, and M is not always available for deep generative models due to their large parameter space. Therefore, we do not explicitly enforce (R3) or (R4), but will use them for generation quality control (see Sec. 4).

Generative models. To train key-dependent models, Eq. (5) is relaxed by introducing a penalty on the generation quality:

$$\min_{\theta_i} \mathbb{E}_{z\sim P_z}\left[\max\{1 - f_{\phi_i}(G_i(z;\theta_i)), 0\} + C||G_0(z) - G_i(z;\theta_i)||^2/d_x\right]. \tag{9}$$

The hyperparameter C is tuned through a parametric study (see Appendicies K).

Robust training. Lastly, we consider the scenario where outputs are post-processed before being verified. We train a robust version of the generative models against a distribution of post-processes $T : \mathbb{R}^{d_x} \to \mathbb{R}^{d_x} \sim P_T$ through

$$\min_{\theta_i} \mathbb{E}_{z\sim P_z, T\in P_T}\left[\max\{1 - f_{\phi_i}(T(G_i(z;\theta_i))), 0\} + C||G_0(z) - G_i(z;\theta_i)||^2/d_x\right]. \tag{10}$$

Table 1: Empirical averaged distinguishability ($\bar{D}$), attributability ($A(\mathcal{G})$), Δx and FID scores from 20 generative models for each dataset. Standard deviations reported when applicable, or omitted if ≤ 0.05. FID of G_0 (FID_0) is the baseline. FID is not applicable to CycleGAN.

GANs	Angle	Dataset	$\bar{D}$	$A(\mathcal{G})$	$\|\|\Delta x\|\|$	FID_0	FID
DCGAN	Orthogonal	MNIST	0.99	**0.99**	5.20(0.31)	4.98(0.15)	5.68(0.23)
DCGAN	45 degree	MNIST	0.99	0.64	5.63(0.39)	-	5.85(0.32)
DCGAN	Orthogonal	CelebA	0.99	**0.99**	4.19(0.18)	33.95(0.13)	52.09(2.20)
DCGAN	45 degree	CelebA	0.99	0.59	4.75(0.20)	-	59.57(2.56)
PGAN	Orthogonal	CelebA	0.99	**0.99**	9.29(0.95)	13.31(0.07)	21.62(1.73)
PGAN	45 degree	CelebA	0.99	0.71	12.03(1.56)	-	28.84(3.37)
CycleGAN	Orthogonal	Cityscapes	0.99	**0.99**	55.85(3.67)	-	-
CycleGAN	45 degree	Cityscapes	0.99	0.69	54.94(5.20)	-	-

4 Experiments

Settings. We test three widely adopted generative models, DCGAN [26], PGAN [5], and CycleGAN [3]), and three datasets: MNIST [27], CelebA [28] and Cityscape [29]. See Appendix I for details on GAN settings and dataset descriptions. For the root models, we train DCGANs from scratch on MNIST and CelebA, and use pre-trained PGAN [5] and CycleGAN [3].

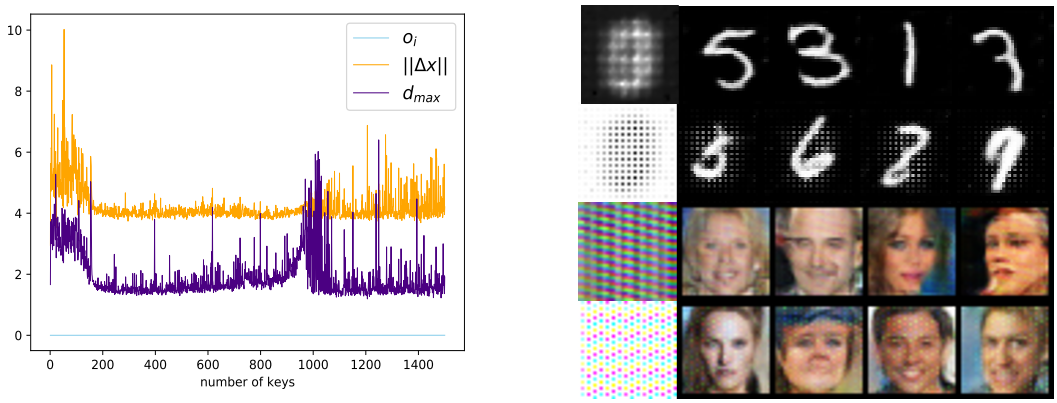
We answer the following questions empirically through experiments.

Can decentralized attributability be achieved through orthogonal keys? For each dataset, we compute twenty keys (Eq. (8)) and their corresponding generative models (Eq. (9)). Table 1 reports the empirical averaged distinguishability and attributability for the collections. For comparison, we randomly sample 20 data-compliant keys by solving an alternative to Eq. (8) where the angle between keys is constrained to $45\,\text{deg}$. The results are presented in the same table. Generation quality metrics ($||\Delta x||$ and FID) are reported in the same table.

Is there a limited capacity of keys? For real-world applications, we would need the capacity of keys to achieve decentralized attribution to be large. From the analysis, the capacity is limited by the availability of orthogonal keys, which is required by attribution, and the generation quality. In Fig. 3a, we report the quantities for 1500 keys generated for MNIST: Orthogonality $o_i = \sum_{j=1}^{i-1} |\phi_j^T \phi_i|/(i-1)$ ($o_1 = 0$), key-perturbation correlation $c_i = \phi_i^T M \phi_i$, $d_{\text{max}}(\phi_i)$, distinguishability $D(G_{\phi_i})$, attributability $A(\{G_{\phi_j}\}_{j=1}^{i})$, and generation quality for $i = 1, ..., 1500$. Some remarks: (1) Nearly orthogonal keys abound due to the high-dimensionality of the output space, for which decentralized attribution is achieved. (2) Larger c_i indicates more involvement of the key in the eigenspace of M with large eigenvalues. There is a positive correlation (0.63) between c and the FID scores, as expected. (3) d_{max} is bounded and so is $||\Delta x||$. Samples from the generator with the largest $||\Delta x||$ are illustrated in Fig. 3a. The results suggest that the registry can use c and d_{max} to monitor the generation quality.

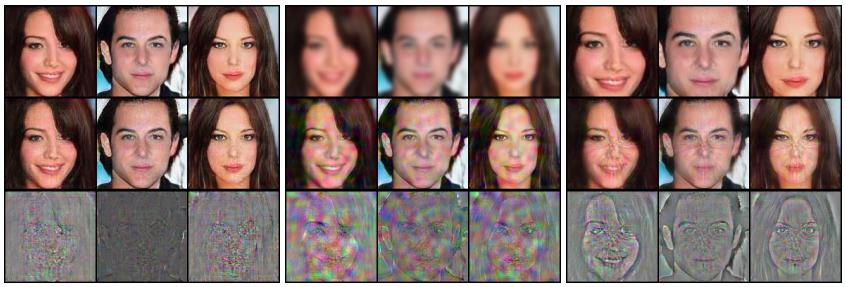
Approximation of M: Since the computation of M (thus c) is expensive for deep generative models with high-dimensional outputs, we seek an empirical approximation of M. Our hypothesis is that the structured patterns associated with eigenvectors of large eigenvalues are mostly associated with in the sensitivities with respect to parameters from the later layers of the generators, and therefore we can approximate M using part of the Jacobian with respect to only those layers. To test the hypothesis, we train relatively shallow DCGANs for MNIST and CelebA, and compute the cosine similarities between the eigenvectors of M with the largest eigenvalue and those from the approximations of M using the last two layers. Results are presented in Fig. 3b, and suggest that it is viable to approximate the largest eigenvectors using the last layers.

How do post-processes affect attributability and generation quality? We consider five types of post-processes: blurring, cropping, noise, JPEG conversion and the combination of these four, and assume that the post-processes are known by the model publishers who then improve the robustness of decentralized attribution by incorporating these processes as differentiable layers and solving Eq. (10). Examples of the post-processed images from non-robust and robust generators are compared in Fig. 4. *Implementation:* `Blurring` uses Gaussian kernel width uniformly drawn from $\frac{1}{3}\{1, 3, 5, 7, 9\}$. `Cropping` crops images with uniformly picked ratio between 80% and 100% and scales the cropped images back to the original size using bilinear interpolation. `Noise` adds iid white noise with



(a) (b)

Figure 3: (a) d_{max} is bounded by $||\Delta_x||$ and o_i are close to 0. (b) Eigenvectors and the corresponding samples from (top to bottom) the largest eigenvector of third layer and last layer of M_{MNIST}, the largest eigenvector of third layer and last layer of M_{CelebA}.



(a) Non-robust (b) Blurring (c) Cropping

(d) Noise (e) JPEG (f) Combination

Figure 4: (a) 1st-2rd row: samples from G_0 and non-robust generator (b-f) 1st-2rd rows: worst-case post-process, samples from robust training against the specific post-processes (prior to the post-processes). 3rd row for all: differences between 2nd row of (a) and 2nd row of each image. As a result, we can reveal the changes in attributions.

standard deviation uniformly drawn in $[0, 0.3]$. `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 [30] and JPEG [31]. For robust training against each post-process, we apply the post-process to mini-batches with 50% probability. *Results:* We report in Table 2 the attributability before and after robust training of distinguishability. `Blurring`, `Cropping` and `Combination` are all effective before robust training. Defense against these random post-processes can be achieved except for `Combination`. Table 3 reports $||\Delta x||$ and FID scores of the robust models, showing the trade-off between attributability and generation quality. Readers are referred to Appendix J for more results of robust training.

5 Related Work

Fingerprints of GANs. Researches have shown that convolutional neural network based generator leaves artifacts [32]. Marra et. al. [22] empirically showed that the artifact can be used as a fingerprint.

Table 2: Distinguishability (top), attributability (btm) before (Bfr) and after (Aft) robust training.

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
PGAN	CelebA	0.50	0.98	0.51	0.99	0.97	0.99	0.96	0.99	0.50	0.76
CycleGAN	Cityscapes	0.49	0.92	0.49	0.87	0.98	0.99	0.55	0.99	0.49	0.67
DCGAN	MNIST	0.49	0.96	0.49	0.97	0.85	0.98	0.53	0.99	0.49	0.65
DCGAN	CelebA	0.50	0.99	0.50	0.99	0.95	0.99	0.51	0.99	0.50	0.85
PGAN	CelebA	0.50	0.97	0.50	0.99	0.96	0.98	0.96	0.99	0.50	0.76
CycleGAN	Cityscapes	0.50	0.92	0.50	0.86	0.97	0.98	0.54	0.99	0.50	0.67

Table 3: $||\Delta x||$ (top) and FID score (btm) w/ and w/o robust training. Standard deviations in parenthesis. DCGAN-M: DCGAN for MNIST, DCGAN-C: DCGAN for CelebA. FID score not applicable to CycleGAN. *Lower is better.*

GANs	Non-robust	Blurring	Cropping	Noise	JPEG	Combination
DCGAN-M	5.20(0.31)	15.96(2.18)	9.17(0.65)	5.93(0.34)	6.48(0.94)	17.08(1.86)
DCGAN-C	4.19(0.18)	11.83(0.65)	9.30(0.31)	4.75(0.17)	6.01(0.29)	13.69(0.59)
PGAN	9.29(0.95)	18.49(2.04)	21.27(0.81)	10.20(0.81)	10.08(1.03)	24.82(2.33)
CycleGAN	55.85(3.67)	68.03(3.62)	80.03(3.59)	55.47(1.60)	57.42(2.00)	83.94(4.66)
DCGAN-M	5.68(0.23)	41.11(20.43)	21.58(2.44)	5.79(0.19)	6.50(1.70)	68.16(24.67)
DCGAN-C	52.09(2.20)	73.62(6.70)	98.86(9.51)	59.51(1.60)	60.35(2.57)	87.29(9.29)
PGAN	21.62(1.73)	28.15(3.43)	47.94(5.71)	25.43(2.19)	22.86(2.06)	45.16(7.87)

However, their method depends on the dissimilarities of the target data. Yu et al. [23] trained external classifier to identify the images from a finite and fixed set of generators, and showed that the classifier can achieve robustness against post-processed images by fine-tuning the classifier using post-processed images. But the result is not guaranteed to have the same performance when the set of generators grows arbitrarily. Albright et al. [33] showed that they can find the origin of images by solving the generator inversion problem. This method requires that the registry save all generators. Furthermore, the registry needs to solve the optimization problem for all generators.

Digital watermarking. Digital watermarking has been used for identifying the ownership of digital signals. Research on watermarking focused on the least significant bits in images [34, 35] and frequency domain [36, 37, 38]. Zhu et al. [31] showed that GANs can be used for watermarking by introducing various operation layers to the training step. Since watermarks are directly added to the outputs, they are similar to the presented toy case. Along the same direction, Fan et al. [39] imposed *passport* to classification networks. Without proper *passport*, the classification accuracy of the network drops. Their approach, however, has not been extended to the decentralized attribution setting.

6 Conclusion

This paper investigated the feasibility of decentralized attribution for generative models. We used a protocol where a registry generates and distributes keys to users, and the user creates a key-dependent generative model for which the outputs can be correctly attributed by the registry. Our investigation led to simple design rules of the keys to achieve correct attribution while maintaining reasonable generation quality. Specifically, correct attribution requires keys to be data compliant and orthogonal; and generation quality can be monitored through data- and model-dependent metrics. With concerns about adversarial post-processes, we empirically show that robust attribution can be achieved with further loss of generation quality. This study defines the design requirements for future protocols for the creation and distribution of attributable generative models.

7 Broader Impact

With recent advances of generative models, researchers focus on the potential misuses and their forensics [24, 40]. Current state-of-the-art models can generate realistic fake images [10, 11, 12], voices [18] and videos [19, 20, 21]. Against these developments, studies of forensic have also been in the spotlight [24, 40]. This paper takes a different perspective than this ongoing competition between the two sides. We are motivated by the requirement of model attribution, i.e., the ability to tell which exact models do the contents come from, in addition to *whether* the contents are machine generated or not.

To this end, the paper focused on a regulation approach in the setting where generative models are white-box to end users, keys are black-box (withheld by the model publishers), and datasets are proprietary. While we focus on the technical feasibility of decentralized attribution of generative models, the applicability of the proposed method would require discussions beyond the scope of the paper. We assume that the protocol, i.e., key distribution by the model publisher and key-dependent training on the user end, can be embraced by all stakeholders involved (e.g., social media platforms and news organizations). While this protocol does not eliminate risks from individual adversaries, it will be a necessary constraint on publishers that have the computational, technological, and data resources to create and distribute high-impact machine-generated contents.

References

[1] 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*, pages 2672–2680, 2014.

[2] 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*, pages 2536–2544, 2016.

[3] 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*, pages 2223–2232, 2017.

[4] 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.

[5] 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.

[6] 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*, pages 8798–8807, 2018.

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

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

[9] 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*, pages 8789–8797, 2018.

[10] 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*, pages 4401–4410, 2019.

[11] 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*, 2019.

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

[13] Makena Kelly. Congress grapples with how to regulate deepfakes. *Congress grapples with how to regulate deepfakes*, Jun 2019.

[14] ALI BRELAND. The bizarre and terrifying case of the "deepfake" video that helped bring an african nation to the brink. *motherjones*, Mar 2019.

[15] Raphael Satter. Experts: Spy used ai-generated face to connect with targets. *Experts: Spy used AI-generated face to connect with targets*, Jun 2019.

[16] 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.

[17] Faceswap. https://faceswap.dev.

[18] 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*, pages 14881–14892, 2019.

[19] Wikipedia contributors. Deepfake — Wikipedia, the free encyclopedia, 2019. [Online; accessed 17-June-2019].

[20] 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*, 2018.

[21] 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.

[22] 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)*, pages 506–511. IEEE, 2019.

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

[24] 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*, 2019.

[25] 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*, pages 6626–6637, 2017.

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

[27] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[29] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[30] E. Riba, D. Mishkin, D. Ponsa, E. Rublee, and G. Bradski. Kornia: an open source differentiable computer vision library for pytorch, 2020.

[31] 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)*, pages 657–672, 2018.

[32] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

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

[34] 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)*, pages 666–673, 1993.

[35] 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, pages 86–90. IEEE, 1994.

[36] 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.

[37] 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.

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

[39] 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*, pages 4716–4725, 2019.

[40] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

A Proof of Proposition 1

Proposition 1. *For the toy case, if* $\varepsilon > \max_{x\sim P_{\mathcal{D}}}\{||x||_\infty\}$, $D(G_\phi) = 1$ *for all* $\phi \in \Phi$ *and* $A(\mathcal{G}) = 1$.

Proof. Let ϕ and ϕ' be any pair of keys such that $\phi^T\phi' = 0$, and let x, x', and x_0 be sampled from P_{G_ϕ}, $P_{G_{\phi'}}$, and $P_{\mathcal{D}}$, respectively. When $\varepsilon > \max_{x\sim P_{\mathcal{D}}}\{||x||_\infty\}$, we have

$$\begin{aligned}\phi^T x &= \phi^T(x_0 + \varepsilon\phi)\\ &= \phi^T x_0 + \varepsilon\\ &> \phi^T x_0 + \max_{x\sim P_{\mathcal{D}}}\{||x||_\infty\}\\ &> \phi^T x_0 - \phi^T x_0 = 0.\end{aligned} \tag{11}$$

Combined with the data-compliant assumption $\phi^T x_0 < 0$, we have $D(G_\phi) = 1$. Further, since

$$\phi^T x' = \phi^T(x_0 + \varepsilon\phi') = \phi^T x_0 < 0, \tag{12}$$

we have $A(\mathcal{G}) = 1$. □

B Empirical test for the linear approximation

For first-order analyses, we approximate the key-dependent generative model to be updated from the root model through $\theta = \theta_0 + \Delta\theta$, where

$$\Delta\theta = \gamma\mathbb{E}_{x\sim P_{G_0}}\left[\nabla_\theta x^T|_{\theta_0}\right]\phi, \tag{13}$$

and

$$x = x_0 + \nabla_\theta x_0|_{\theta_0}\Delta\theta. \tag{14}$$

Let $J(x) = \nabla_\theta x|_{\theta_0}$ and $M = \mathbb{E}_{x\sim P_{G_0}}[J(x)]\mathbb{E}_{x\sim P_{G_0}}[J(x)^T]$. We focus on testing the following result of the linear approximation: For ϕ and G_ϕ with high distinguishability, we should observe that with high probability,

$$\phi^T\tilde{x} = \phi^T\left(x_0 + \gamma J(x_0)\mathbb{E}_{x\sim P_{G_0}}\left[J(x)^T\right]\phi\right) > 0, \tag{15}$$

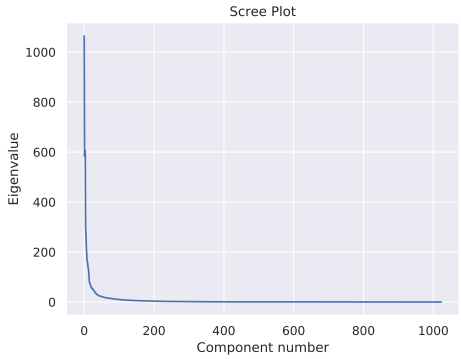
for $x_0 \sim P_{G_0}$. To test this, we use a DCGAN trained on MNIST as G_0. We train 20 keys and update Gs correspondingly following the method detailed in the **Experiments** section. The resulting average distinguishability from the 20 generative models is 0.99.

To compute $\Pr(\phi^T\tilde{x} > 0)$, we calculate $J(x_0)$ and $\mathbb{E}_{x\sim P_{G_0}}[J(x)]$ based on samples from G_0. From Eq. (13), $||\Delta\theta|| = ||\gamma\mathbb{E}_{x\sim P_{G_0}}[J(x)^T]\phi|| = \gamma\sqrt{\phi^T M\phi}$. Therefore $\gamma = ||\Delta\theta||/\sqrt{\phi^T M\phi}$. $||\Delta\theta||$ can be directly computed by comparing θ and θ_0; M can be computed through SVD on $\mathbb{E}_{x\sim P_{G_0}}[J(x)]$ (the tested DCGAN has 1,065,984 parameters, and output dimension of 1024, thus $J \in \mathbb{R}^{1024\times 1{,}065{,}984}$). Empirical test showsro $\frac{1}{20}\sum_{\phi\in\Phi}\Pr(\phi^T\tilde{x} > 0) = 0\%$.

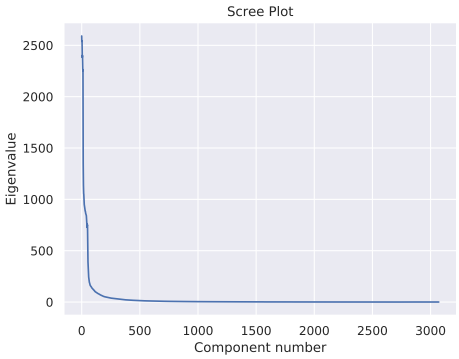
C Empirical test for Conjecture 1

Conjecture 1. *Let the* (i,j)*th element of* $\Sigma(x) = J(x)\mathbb{E}_{x\sim P_{G_0}}[J(x)]^T - M$ *be* $\Sigma_{(i,j)}$ *with variance* σ_{ij}^2. *Then* $\Sigma_{(i,j)}$ *is approximately drawn i.i.d. from* $\mathcal{N}(0, \sigma_{ij}^2)$.

Normality. We use a DCGAN trained on MNIST as G_0 and collect 512 samples of Σ by sampling $x_0 \sim P_{G_0}$. We empirically pick the best distributions for $\Sigma_{(i,j)}$. To do that, we calculate the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for each $\Sigma_{(i,j)}$ (1024^2 calculations in total). Candidate distributions include beta, birnbaumsaunders, exponential, extreme value, gamma, generalized extreme value, generalized pareto, inversegaussian, logistic, loglogistic, lognormal, nakagami, normal, rician, tlocationscale, and weibull distributions. We only report AIC and BIC of normal and extreme value distributions. Among all, the lowest mean AIC and BIC are found from the normal distribution ($AIC = -26.51$ and $BIC = -18.03$). The second best comes from the extreme value distribution ($AIC = 161.42$ and $BIC = 169.90$). From the reported results, we argue that it is reasonable to assume normality for $\Sigma_{(i,j)}$.



(a) Scree plot of MNIST



(b) Scree plot of CelebA

Figure 5: Scree Plots. Most of the eigenvalues are close to 0.

Independence. Due to normality, we test independence through correlations. In theory, this requires a 1024^2-by-1024^2 covariance matrix for all $\Sigma_{(i,j)}$. Without overloading the computational resources, we randomly pick one elements from $\Sigma_{(i,j)}$ and compute correlation coefficient with others (1024^2 calculation). We do such calculation for fifty times without duplication. The resulting average absolute value of the correlations is smaller than 0.1, suggesting that the independence assumption is reasonable. Multiple repetition of calculations did not show notable variations of correlations.

D Empirical test for Conjecture 2

Conjecture 2. *Denote $\Lambda = \{\lambda_1, ..., \lambda_{d_x}\}$ as the eigenvalues of M. For existing deep generative models, there exists and only exists a subset of eigenvalues that are strictly positive.*

We use the same DCGAN trained on MNIST and CelebA as the root models to compute $\mathbb{E}_{x\sim P_{G_0}}[J]$. SVD on the resulting matrix reveals the eigenvalues of M, which are reported in Fig. 5.

E Proof of Theorem 1

Theorem 1. *Let $d_{max}(\phi) = \max_{x\sim P_{\mathcal{D}}} |\phi^T x|$, $\sigma^2(\phi) = \sum_{i,j} \sigma_{ij}^2 \phi_{(i)}^2 \phi_{(j)}^2$, and δ_d be a positive number greater than $\exp\left(-\frac{1}{2}\left(\frac{\phi^T M \phi}{\sigma(\phi)}\right)^2\right)$. For the realistic case and for a given key $\phi \in \Omega$, if*

$$||\Delta x(\phi)|| \geq \frac{d_{max}(\phi)\sqrt{\phi^T M^2 \phi}}{\phi^T M \phi - \sigma(\phi)\sqrt{\log\left(\frac{1}{\delta_d^2}\right)}}, \tag{16}$$

$D(G_\phi) \geq 1 - \delta_d/2$.

Proof. We first note that due to data compliance of keys, $\mathbb{E}_{x\sim P_{\mathcal{D}}}\left[\mathbb{1}(\phi^T x < 0)\right] = 1$. Therefore $D(G_\phi) \geq 1 - \delta_d/2$ iff $\mathbb{E}_{x\sim P_{G_\phi}}\left[\mathbb{1}(\phi^T x > 0)\right] \geq 1 - \delta_d$, i.e., $\Pr(\phi^T x > 0) \geq 1 - \delta_d$ for $x \sim P_{G_\phi}$. We now seek a key-dependent lower bound on ε to satisfy this inequality. We first connect generation quality to the step size (learning rate) γ following the linear approximation:

$$||\Delta x(\phi)|| = ||\gamma M \phi|| = \gamma\sqrt{\phi^T M^2 \phi}. \tag{17}$$

Next, given ϕ, we look for a sufficiently large γ, so that $\phi^T x > 0$ with probability at least $1 - \delta_d$. To do so, let x and x_0 be sampled from P_{G_ϕ} and P_{G_0}, respectively. Then with first order approximations we have

$$\begin{aligned} \phi^T x &= \phi^T \left(x_0 + \gamma J(x_0)\mathbb{E}_{x\sim P_{G_0}}\left[J(x)^T\right]\phi\right) \\ &= \phi^T x_0 + \gamma \phi^T M \phi + \gamma \phi^T \Sigma \phi. \end{aligned} \tag{18}$$

For $\Pr(\phi^T x > 0) \geq 1 - \delta_d$, γ should satisfy

$$\Pr\left(\phi^T \Sigma \phi > -\phi^T x_0/\gamma - \phi^T M \phi\right) \geq 1 - \delta_d. \tag{19}$$

Since $d_{\max}(\phi) \geq -\phi^T x_0$, it is sufficient to have

$$\Pr\left(\phi^T \Sigma \phi > d_{\max}(\phi)/\gamma - \phi^T M \phi\right) \geq 1 - \delta_d. \tag{20}$$

From Conjecture 1, $\phi^T \Sigma \phi \sim \mathcal{N}(0, \sigma^2(\phi))$. Due to the symmetry of $p(\phi^T \Sigma \phi)$, the sufficient condition for γ in Eq. (20) can be rewritten as

$$\Pr\left(\phi^T \Sigma \phi \leq \phi^T M \phi - d_{\max}(\phi)/\gamma\right) \geq 1 - \delta_d. \tag{21}$$

Recall the following tail bound of $x \sim \mathcal{N}(0, \sigma^2)$ for $y \geq 0$:

$$\Pr(x \leq \sigma y) \geq 1 - \exp(-y^2/2). \tag{22}$$

Compare Eq. (22) with Eq. (21), the sufficient condition becomes

$$\begin{aligned} & \phi^T M \phi - d_{\max}(\phi)/\gamma \geq \sigma(\phi)\sqrt{\log\left(\frac{1}{\delta_d^2}\right)} \\ \Rightarrow \quad & \gamma \geq \frac{d_{\max}(\phi)}{\phi^T M \phi - \sigma(\phi)\sqrt{\log\left(\frac{1}{\delta_d^2}\right)}}. \end{aligned} \tag{23}$$

Using Eq. (17), we have

$$||\Delta x(\phi)|| \geq \frac{d_{\max}(\phi)\sqrt{\phi^T M^2 \phi}}{\phi^T M \phi - \sigma(\phi)\sqrt{\log\left(\frac{1}{\delta_d^2}\right)}}, \tag{24}$$

provided that $\phi^T M \phi - \sigma(\phi)\sqrt{\log\left(\frac{1}{\delta_d^2}\right)} > 0$ or

$$\delta_d > \exp\left(-\frac{1}{2}\left(\frac{\phi^T M \phi}{\sigma(\phi)}\right)^2\right). \tag{25}$$

□

F Empirical test for Conjecture 3

Conjecture 3. $||\Delta x|| \leq \tau d_{max}$.

The conjecture comes from the following approximations: First, from Conjecture 1, we observe that $\{\sigma_{ij}\}^2$ are small. Using the proof of Theorem 1, a sufficient degradation of generation quality can be approximated by

$$||\Delta x(\phi)|| \approx \frac{d_{\max}(\phi)\sqrt{\phi^T M^2 \phi}}{\phi^T M \phi} = \frac{d_{\max}\sqrt{c^T \Lambda^2 c}}{c^T \Lambda c}, \tag{26}$$

where $c = P^T \phi$ and $M = P \Lambda P^T$. From Lemma 1,

$$\frac{\sqrt{c^T \Lambda^2 c}}{c^T \Lambda c} \in \left[1, \frac{1 + \lambda_{\max}/\lambda_{\min}}{2\sqrt{\lambda_{\max}/\lambda_{\min}}}\right]. \tag{27}$$

Let

$$\tau = \frac{1 + \lambda_{\max}/\lambda_{\min}}{2\sqrt{\lambda_{\max}/\lambda_{\min}}}, \tag{28}$$

then $||\Delta x|| \leq \tau d_{\max}$.

G Useful lemmas

Lemma 1 is used for Conjecture 3 and Lemmas 2 for the proof of Theorem 2.

Lemma 1. *Let $c \in \mathbb{R}^n$ and $||c|| = 1$, $\Lambda = diag(\lambda_1, ..., \lambda_n)$ be positive definite. Then*

$$\frac{\sqrt{c^T \Lambda^2 c}}{c^T \Lambda c} \in \left[1, \frac{1 + \lambda_{max}/\lambda_{min}}{2\sqrt{\lambda_{max}/\lambda_{min}}}\right]. \tag{29}$$

Proof. Let $x = [c_1^2, ..., c_n^2]$, $a = [\lambda_1^2, ..., \lambda_n^2]$, and $b = [\lambda_1, ..., \lambda_n]$. Then $c^T \Lambda^2 c = a^T x$ and $c^T \Lambda c = b^T x$.

We now consider the following problem:

$$\begin{aligned} \max_x \quad & \frac{1}{2} \log a^T x - \log b^T x \\ \text{s.t.} \quad & 1^T x = 1 \\ & x_i \geq 0, \ \forall i. \end{aligned} \tag{30}$$

The KKT conditions for this problem are

$$\begin{aligned} -\frac{1}{2a^T x} a + \frac{1}{b^T x} b + \lambda 1 - \mu = 0, \\ 1^T x = 1 \\ x_i \geq 0, \ \mu_i \geq 0, \ \forall i \\ \mu^T x = 0, \end{aligned} \tag{31}$$

where λ and μ are the Lagrangian multipliers.

When b has unique elements, there exist two sets of KKT points: x is either one-hot, or x has zero entries except for elements i and j where $x_i = b_j/(b_i + b_j)$ and $x_j = b_i/(b_i + b_j)$, for all (i, j) combinations. If b has repeated elements, then we can combine these elements and reach the same conclusion.

When x is one-hot, the objective is $\log a_i/2 - \log b_i = 0$. For the second type of solutions and let $\tau_{ij} = \lambda_i/\lambda_j$, we have

$$\begin{aligned} \frac{1}{2} \log a^T x - \log b^T x &= \frac{1}{2} \log \frac{a_i b_j + a_j b_i}{b_i + b_j} - \log \frac{2 b_i b_j}{b_i + b_j} \\ &= \frac{1}{2} \log \lambda_i \lambda_j - \log \frac{2\lambda_i \lambda_j}{\lambda_i + \lambda_j} \\ &= \log \frac{1 + \tau_{ij}}{2\sqrt{\tau_{ij}}} \geq 0, \end{aligned} \tag{32}$$

where equality holds when $\tau_{ij} = 1$. Since the objective monotonically increases with respect to $\tau_{ij} > 1$, the maximum is reached when $\tau_{ij} = \lambda_{\max}/\lambda_{\min}$. □

Lemma 2. *Let $a, b \in \mathbb{R}^n$, $||a|| = 1$, $||b|| = 1$, and $a^T b \leq 0$. Let $V \in \mathbb{R}^{n \times n}$. Then $\max_a \{a^T V b\} = \sqrt{b^T V^T V b - (b^T V b)^2}$.*

Proof. Consider the following problem

$$\begin{aligned} \min_a \quad & - a^T V b \\ \text{s.t.} \quad & a^T b \leq 0 \\ & a^T a = 1, \end{aligned} \tag{33}$$

with the following KKT conditions:

$$\begin{aligned} -Vb + \mu b + 2\lambda a = 0 \\ a^T b \leq 0 \\ a^T a = 1. \end{aligned} \tag{34}$$

The solution is

$$\begin{aligned} \lambda &= a^T V b/2 \\ \mu &= b^T V b \\ a &= \frac{(V - b^T V b I)b}{\sqrt{b^T V^T V b - (b^T V b)^2}}. \end{aligned} \tag{35}$$

Note that

$$\begin{aligned} ||(V - b^T V b I)b||^2 &= b^T (V^T - b^T V b I)(V - b^T V b I)b \\ &= b^T V^T V b - (b^T V b)^2, \end{aligned} \tag{36}$$

thus $b^T V^T V b - (b^T V b)^2 \geq 0$.

Since the Hessian of the Lagrangian with respect to a is $2\lambda I$, and from the solution

$$\begin{aligned} \lambda &= a^T V b/2 \\ &= \sqrt{b^T V^T V b - (b^T V b)^2}/2 \geq 0, \end{aligned} \tag{37}$$

therefore the solution is the minimizer, i.e., $\max_a \{a^T V b\} = \sqrt{b^T V^T V b - (b^T V b)^2}$. □

H Proof of Theorem 2

Theorem 2. *Let $d^*_{min} = \min_{\phi \in \Phi, x \sim P_{\mathcal{D}}} |\phi^T x|$, $\bar{\sigma}^2(\phi) = \sqrt{\phi^T V^T V \phi - (\phi^T V \phi)^2}$, and $V_{(i,j)} = \sigma^2_{ij}$. When $D(G) \geq 1 - \delta_d$ for all $G_\phi \in \mathcal{G}$, if the degradation of generation quality for all models in $\mathcal{G}$ satisfies*

$$||\Delta x(\phi)|| \leq \frac{d^*_{min}\sqrt{\phi^T M^2 \phi}}{\sqrt{\phi^T M^2 \phi - (\phi^T M \phi)^2} + \bar{\sigma}(\phi)\sqrt{\log\left(\frac{1}{\delta_a^2}\right)}}, \tag{38}$$

and $\phi^T \phi' \leq 0$ for all $\phi, \phi' \in \Omega$, then $A(\mathcal{G}) \geq 1 - (\delta_d + \delta_a)/2$.

Proof. Let ϕ and ϕ' be any of the two orthogonal keys, and x' and x_0 be sampled from $P_{G_{\phi'}}$ and P_{G_0}, respectively. $A(\mathcal{G}) \geq 1 - (\delta_d + \delta_a)/2$ and $D(G) \geq 1 - \delta_d$ for all $G \in \mathcal{G}$ together imply that $\Pr(\phi^T x' < 0) \geq 1 - \delta_a$. Now we focus on deriving the sufficient conditions for this inequality.

From first order approximations,

$$\begin{aligned} \phi^T x' &= \phi^T \left(x_0 + \gamma(\phi') J(x_0) \mathbb{E}_{x \sim P_{G(\theta_0)}} \left[J(x)^T \right] \phi' \right) \\ &= \phi^T x_0 + \gamma(\phi') \phi^T M \phi' + \gamma(\phi') \phi^T \Sigma \phi'. \end{aligned} \tag{39}$$

Therefore

$$\begin{aligned} \Pr(\phi^T x' < 0) &= \Pr\left(\phi^T \Sigma \phi' < -\phi^T x_0/\gamma(\phi') - \phi^T M \phi' \right) \\ &\geq \Pr\left(\phi^T \Sigma \phi' < d_{\min}(\phi)/\gamma(\phi') - \phi^T M \phi' \right). \end{aligned} \tag{40}$$

Note that the RHS of Eq. (40) suggests that $\gamma(\phi')$ needs to be sufficiently small for $\Pr(\phi^T x' < 0)$ to be large. To see where that upper bound is, we start by noting that $\phi^T \Sigma \phi'$ has zero mean and is normally distributed. To analyze its variance, we use Lemma 2 to show that

$$Var(\phi^T \Sigma \phi') \leq \bar{\sigma}^2(\phi') = \sqrt{\phi'^T V^T V \phi' - (\phi'^T V \phi')^2}, \tag{41}$$

where $V_{(i,j)} = \sigma^2_{ij}$.

Using the same tail bound of normal distribution as in Theorem 1, $\gamma(\phi')$ is sufficiently small if

$$\begin{aligned} & d_{\min}(\phi)/\gamma(\phi') - \phi^T M \phi' \geq \bar{\sigma}(\phi')\sqrt{\log\left(\frac{1}{\delta_a^2}\right)} \\ \Rightarrow \quad & \gamma(\phi') \leq \begin{cases} \frac{d_{\min}(\phi)}{\phi^T M \phi' + \bar{\sigma}(\phi')\sqrt{\log\left(\frac{1}{\delta_a^2}\right)}} & \text{if } \phi^T M \phi' + \bar{\sigma}(\phi')\sqrt{\log\left(\frac{1}{\delta_a^2}\right)} > 0, \\ +\infty & \text{otherwise} \end{cases} \end{aligned} \tag{42}$$

Since $||\Delta x(\phi')|| = \gamma(\phi')\sqrt{\phi'^T M^2 \phi'}$, we have

$$||\Delta x(\phi')|| \leq \begin{cases} \frac{d_{\min}(\phi)\sqrt{\phi'^T M^2 \phi'}}{\phi^T M \phi' + \bar{\sigma}(\phi')\sqrt{\log\left(\frac{1}{\delta_a^2}\right)}} & \text{if } \phi^T M \phi' + \bar{\sigma}(\phi')\sqrt{\log\left(\frac{1}{\delta_a^2}\right)} > 0, \\ +\infty & \text{otherwise} \end{cases} \tag{43}$$

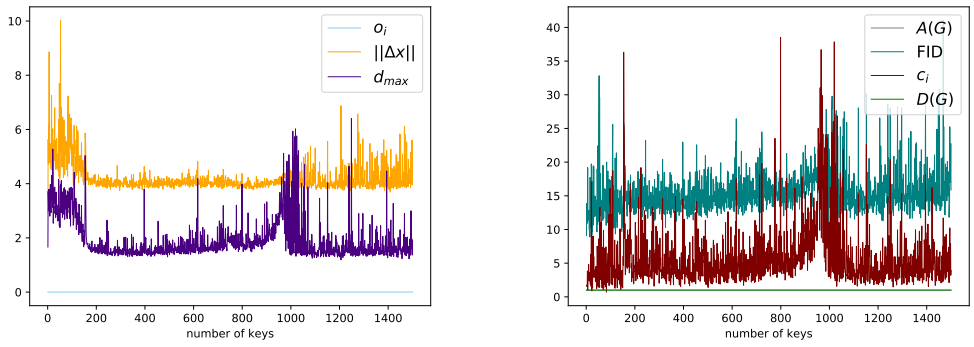
We would like to find a lower bound of the RHS of Eq. (43) that is independent of $\phi \neq \phi'$. To this end, first denote $d^*_{\min} = \min_\phi d_{\min}(\phi)$. Now use Lemma 2 again to derive an upper bound of $\phi^T M \phi'$:

$$\phi^T M \phi' \leq \sqrt{\phi'^T M^2 \phi' - (\phi'^T M \phi')^2}. \tag{44}$$

Replace $\phi^T M \phi'$ in Eq. (43) with its upper bound to reach a ϕ-indepdent sufficient condition for $||\Delta x(\phi')||$:

$$||\Delta x(\phi')|| \leq \frac{d^*_{\min}\sqrt{\phi'^T M^2 \phi'}}{\sqrt{\phi'^T M^2 \phi' - (\phi'^T M \phi')^2} + \bar{\sigma}(\phi')\sqrt{\log\left(\frac{1}{\delta_a^2}\right)}}. \tag{45}$$

□



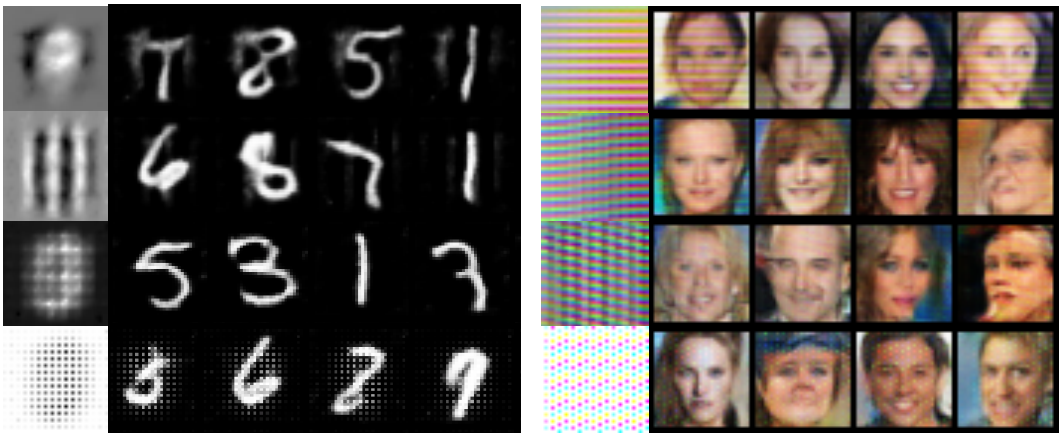
(a) (b)

Figure 6: (a) d_{max} is bounded by $||\Delta_x||$ and o_i is close to 0. (b) c_i and FID show positive correlation (0.63). Also, $D(G_{\phi_i})$ and $A(\{G_{\phi_j}\}_{j=1}^i)$ are close to 1.

I Limited capacity of keys

We generate 1500 keys for MNIST: orthogonality $o_i = \sum_{j=1}^{i-1} |\phi_j^T \phi_i|/(i-1)$ ($o_1 = 0$), key-perturbation correlation $c_i = \phi_i^T M \phi_i$, $d_{\max}(\phi_i)$, distinguishability $D(G_{\phi_i})$, attributability $A(\{G_{\phi_j}\}_{j=1}^i)$, lack of generation quality $||\Delta x||$ and FID for $i = 1, ..., 1500$. Some remarks: (1) $d_{\max}$ is bounded and so is $||\Delta x||$ (Fig. 6a). (2) Larger c_i indicates more involvement of the key in the eigenspace of M with large eigenvalues. There is a positive correlation (0.63) between c and the FID scores, as expected (Fig. 6b). (3) Nearly orthogonal keys abound due to the high-dimensionality of the output space, for which decentralized attribution is achieved (Fig. 6b). Thus, the results suggest that the registry can use c and $d_{\max}$ to monitor the generation quality.

Approximation of M: The hypothesis is that the structured pattern of large eigenvectors is associated with eigenvectors of the later layers of the generators. Therefore, M can be approximated using the Jacobian of these layers. For empirical experiments, we train four-layer DCGANs for MNIST and CelebA, and compute the cosine similarities between the largest eigenvector of M and the largest eigenvecotrs of Jacobian of each of layers. Results are presented in Fig. 7 with visual examples. Also, it is viable to approximate the largest eigenvectors with the last layers.

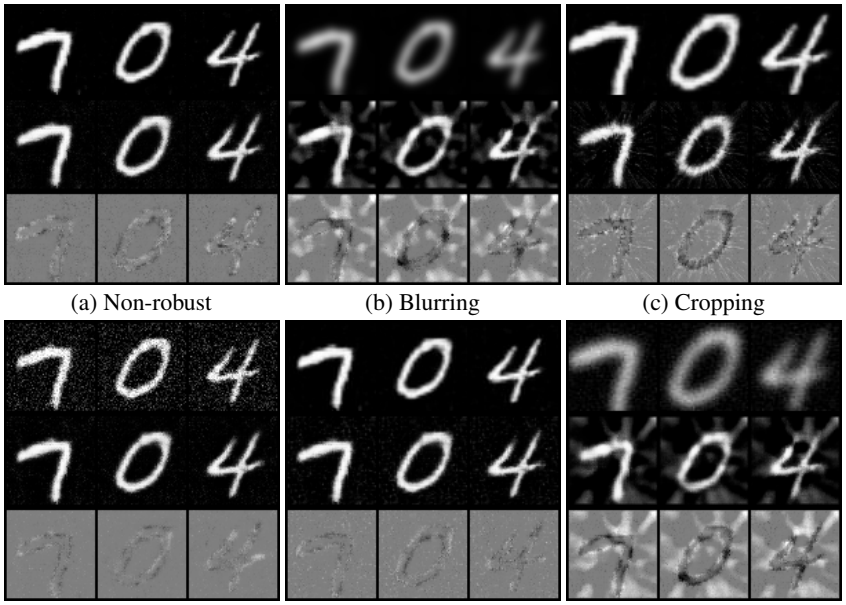


(a) (b)

Figure 7: (a) Largest eigenvectors of the first layer to last layer (top to bottom) and corresponding samples. Cosine similarities with largest eigenvector of M are -0.49, 0.20, -0.98, 0.49. (b) Largest eigenvectors of the first layer to last layer (top to bottom) and corresponding samples. Cosine similarities with largest eigenvector of M are 0, 0.01, -0.05, -0.5.

J Examples of GANs

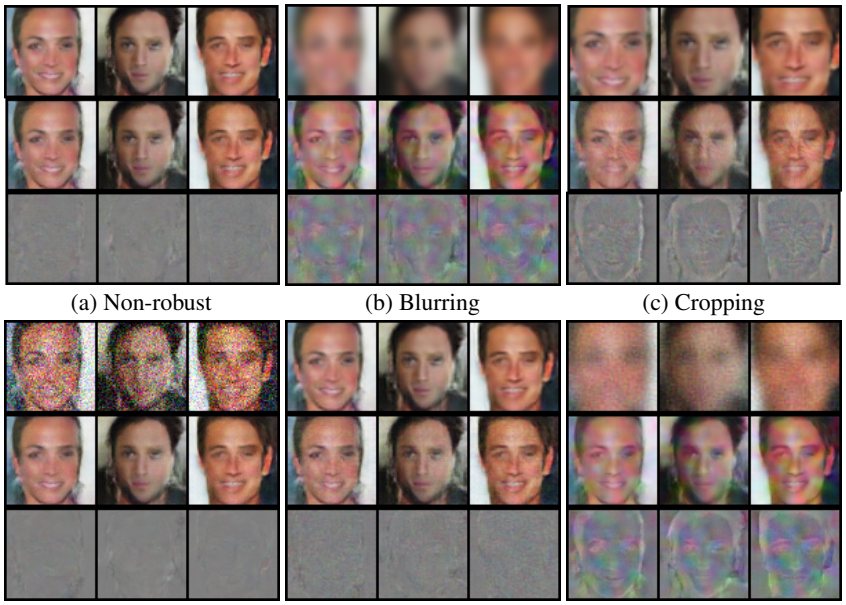
In the paper, we show examples from PGAN with CelebA. Here, we illustrate other GANs examples. For Fig. 8, 9, 10, (a) 1st-2rd row: authentic data, samples from the non-robust generator (b-f) 1st-2rd rows: worst-case post-process, samples from robust training against the specific post-processes (before the post-processes). 3rd row for all: numerical differences between 2nd row of (a) and 2nd row of each case. Thus, the differences show the effect of robust training on attribution.



(a) Non-robust (b) Blurring (c) Cropping

(d) Noise (e) JPEG (f) Combination

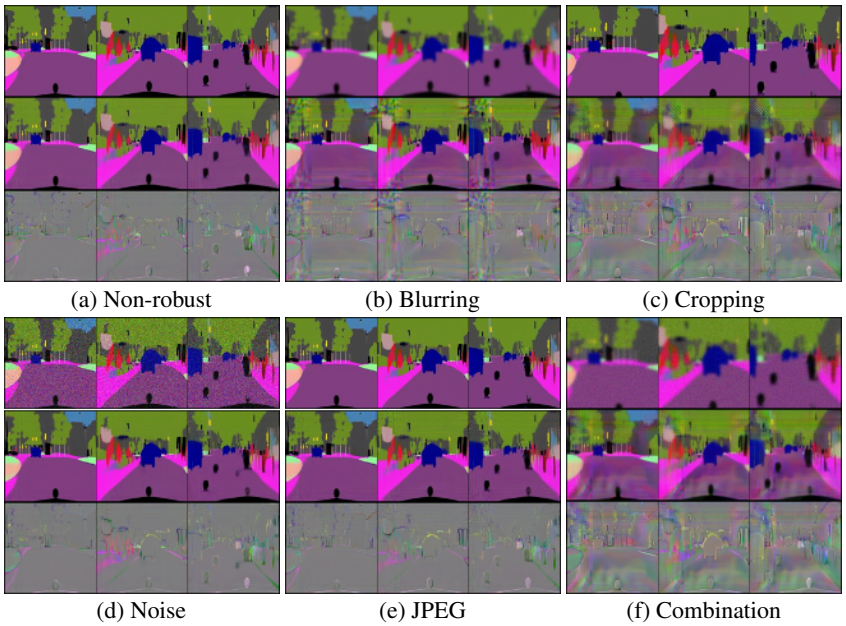
Figure 8: DCGAN-MNIST



(a) Non-robust (b) Blurring (c) Cropping

(d) Noise (e) JPEG (f) Combination

Figure 9: DCGAN-CelebA



(a) Non-robust (b) Blurring (c) Cropping

(d) Noise (e) JPEG (f) Combination

Figure 10: CycleGAN-Cityscapes

K Training Details

K.1 Parameters

We adopt Adam optimizer for gradient descent. We attach other parameters in Table 4. Note that we fix the hyper-parameters when we optimize Eq.(Robust training) in **Implementation**.

K.2 Training Time

For experimental validations, we use V:100 Tesla GPUs. Exact number of GPUs are reported in Table 5.

Table 4: Hyper-parameters for training Eq.(Key generation) (top) and Eq.(Generative models) (btm). Equations are in **Implementation**.

GANs	Dataset	Batch Size	Learning Rate	β_1	β_2	Epochs	C
DCGAN	MNIST	128	0.001	0.5	0.99	10	-
DCGAN	CelebA	64	0.001	0.5	0.99	10	-
PGAN	CelebA	32	0.001	0.5	0.99	10	-
CycleGAN	Cityscapes	4	0.001	0.5	0.99	20	-
DCGAN	MNIST	16	0.0005	0.5	0.99	10	10
DCGAN	CelebA	64	0.0005	0.5	0.99	10	10
PGAN	CelebA	16	0.0005	0.0	0.99	5	100
CycleGAN	Cityscapes	1	0.0005	0.5	0.99	5	1000

Table 5: Training time (in minute) of one key (Eq.(Key generation)) and one generator (Eq.(Generative models)). DCGAN-M: DCGAN for MNIST, DCGAN-C: DCGAN for CelebA. Equations are in **Implementation**.

GANs	GPUs	Key	Naive	Blurring	Cropping	Noise	JPEG	Combination
DCGAN-M	1	1.77	8.52	4.12	3.96	4.19	5.71	5.12
DCGAN-C	1	5.31	9.12	10.33	9.56	10.35	10.25	10.76
PGAN	2	50.89	141.07	140.05	131.90	133.46	132.46	135.07
CycleGAN	1	20.88	16.04	16.26	15.43	15.71	15.98	16.41

L Ablation Study

We attach the table of ablation study of how C affects the result of distinguishability, attributability, $||\Delta x||$ and FID scores in Table 6. C does not affect to the distinguishability and attributability. But C improves $||\Delta x||$ and FID for every generators. Furthermore, we investigate how C term affects the robustness in Table 7 and Table 8. We can observe that, as C increases, robustness decreases but generation quality increases.

Table 6: Empirical averaged distinguishability ($\bar{D}$), attributability ($A(\mathcal{G})$), $||\Delta x||$ and FID scores. Standard deviations reported when applicable, or omitted if ≤ 0.05. FID of G_0 (FID_0) is the baseline. FID is not applicable to CycleGAN.

GANs	Dataset	C	$\bar{D}$	$A(\mathcal{G})$	$\|\|\Delta x\|\|$	FID_0	FID
DCGAN	MNIST	10	0.99	0.99	5.20(0.31)	4.98(0.15)	5.68(0.23)
DCGAN	MNIST	100	0.99	0.99	4.09(0.53)	-	5.32(0.11)
DCGAN	MNIST	1K	0.99	0.99	3.88(0.60)	-	5.23(0.12)
DCGAN	CelebA	10	0.99	0.99	4.19(0.18)	33.95(0.13)	52.09(2.20)
DCGAN	CelebA	100	0.99	0.99	3.08(0.27)	-	45.02(3.37)
DCGAN	CelebA	1K	0.99	0.99	2.55(0.36)	-	40.85(3.41)
PGAN	CelebA	100	0.99	0.99	9.29(0.95)	13.31(0.07)	21.62(1.73)
PGAN	CelebA	1K	0.99	0.99	6.51(1.85)	-	19.05(3.14)
PGAN	CelebA	10K	0.98	0.98	5.05(1.63)	-	16.75(1.87)
CycleGAN	Cityscapes	1K	0.99	0.99	55.85(3.67)	-	-
CycleGAN	Cityscapes	10K	0.99	0.99	49.66(5.01)	-	-

Table 7: Distinguishabilit (top), attributability (btm) before (Bfr) and after (Aft) robust training. DCGAN-M: DCGAN for MNIST, DCGAN-C: DCGAN for CelebA.

GANs	C	`Blurring`		`Cropping`		`Noise`		`JPEG`		`Combination`	
-	-	Bfr	Aft	Bfr	Aft	Bfr	Aft	Bfr	Aft	Bfr	Aft
DCGAN-M	10	0.49	**0.96**	0.52	**0.99**	0.85	**0.99**	0.54	**0.99**	0.50	**0.66**
DCGAN-M	100	0.49	0.61	0.51	0.98	0.76	0.98	0.53	0.99	0.50	0.52
DCGAN-M	1K	0.49	0.50	0.51	0.81	0.69	0.91	0.53	0.97	0.50	0.51
DCGAN-C	10	0.49	**0.99**	0.49	**0.99**	0.95	**0.98**	0.51	**0.99**	0.50	**0.85**
DCGAN-C	100	0.50	0.96	0.49	0.99	0.92	0.93	0.50	0.99	0.49	0.61
DCGAN-C	1K	0.50	0.62	0.49	0.97	0.88	0.91	0.50	0.99	0.49	0.51
PGAN	100	0.50	**0.98**	0.51	**0.99**	0.97	**0.99**	0.96	**0.99**	0.50	**0.76**
PGAN	1K	0.50	0.89	0.49	0.95	0.94	0.95	0.88	0.99	0.50	0.60
PGAN	10K	0.50	0.61	0.50	0.76	0.89	0.90	0.76	0.98	0.50	0.51
CycleGAN	1K	0.49	**0.92**	0.49	**0.87**	0.98	**0.99**	0.55	**0.99**	0.49	**0.67**
CycleGAN	10K	0.49	0.70	0.50	0.66	0.94	0.96	0.52	0.98	0.50	0.51
DCGAN-M	10	0.49	**0.96**	0.49	**0.97**	0.85	**0.98**	0.53	**0.99**	0.49	**0.65**
DCGAN-M	100	0.50	0.54	0.49	0.97	0.74	0.96	0.52	0.94	0.49	0.52
DCGAN-M	1K	0.50	0.50	0.49	0.80	0.68	0.89	0.52	0.89	0.49	0.50
DCGAN-C	10	0.50	**0.99**	0.50	**0.99**	0.95	**0.99**	0.51	**0.99**	0.50	**0.85**
DCGAN-C	100	0.50	0.96	0.49	0.99	0.92	0.93	0.50	0.99	0.49	0.61
DCGAN-C	1K	0.49	0.61	0.50	0.98	0.87	0.89	0.50	0.99	0.50	0.51
PGAN	100	0.50	**0.97**	0.50	**0.99**	0.96	**0.98**	0.96	**0.99**	0.50	**0.76**
PGAN	1K	0.50	0.87	0.50	0.95	0.93	0.94	0.86	0.99	0.49	0.59
PGAN	10K	0.50	0.60	0.50	0.77	0.88	0.89	0.76	0.97	0.50	0.51
CycleGAN	1K	0.50	**0.92**	0.50	**0.86**	0.97	**0.98**	0.54	**0.99**	0.50	**0.67**
CycleGAN	10K	0.50	0.70	0.50	0.66	0.92	0.94	0.52	0.98	0.49	0.51

Table 8: $||\Delta x||$ (top) and FID score (btm). Standard deviations in parenthesis. DCGAN-M: DCGAN for MNIST, DCGAN-C: DCGAN for CelebA. FID score not applicable to CycleGAN. $||\Delta x||$ and FID score in Table 6 are baseline. *Lower is better.*

GANs	C	Baseline	`Blurring`	`Cropping`	`Noise`	`JPEG`	`Combination`
DCGAN-M	10	5.20(0.31)	15.96(2.18)	9.17(0.65)	5.93(0.34)	6.48(0.94)	17.08(1.86)
DCGAN-M	100	4.09(0.53)	12.95(4.47)	7.62(1.55)	4.57(0.78)	4.70(1.02)	12.70(3.37)
DCGAN-M	1K	**3.88(0.60)**	**7.17(2.10)**	**7.43(1.37)**	**4.22(0.77)**	**5.12(1.94)**	**7.56(1.41)**
DCGAN-C	10	4.19(0.18)	11.83(0.65)	9.30(0.31)	4.75(0.17)	6.01(0.29)	13.69(0.59)
DCGAN-C	100	3.08(0.27)	10.00(1.61)	7.80(0.58)	3.20(0.45)	4.26(0.59)	11.65(1.48)
DCGAN-C	1K	**2.55(0.36)**	**7.68(1.53)**	**7.13(0.47)**	**2.65(0.24)**	**3.39(0.58)**	**9.23(1.22)**
PGAN	100	9.29(0.95)	18.49(2.04)	21.27(0.81)	10.20(0.81)	10.08(1.03)	24.82(2.33)
PGAN	1K	6.52(1.85)	14.79(4.15)	18.88(1.96)	6.40(1.48)	7.09(1.62)	22.09(2.12)
PGAN	10K	**5.04(1.63)**	**10.19(2.87)**	**18.23(0.94)**	**5.13(1.14)**	**5.67(1.62)**	**17.26(1.39)**
CycleGAN	1K	55.85(3.67)	68.03(3.62)	80.03(3.59)	55.47(1.60)	57.42(2.00)	83.94(4.66)
CycleGAN	10K	**49.66(5.01)**	**58.64(3.70)**	**66.05(3.47)**	**53.14(0.44)**	**54.52(2.30)**	**66.24(5.29)**
DCGAN-M	10	5.68(0.23)	41.11(20.43)	21.58(2.44)	5.79(0.19)	6.50(1.70)	68.16(24.67)
DCGAN-M	100	5.32(0.11)	23.83(14.29)	18.39(3.70)	5.41(0.18)	5.46(0.11)	36.05(16.20)
DCGAN-M	1K	**5.23(0.12)**	**10.85(4.28)**	**18.08(1.77)**	**5.37(0.14)**	**5.30(0.96)**	**21.86(4.16)**
DCGAN-C	10	52.09(2.20)	73.62(6.70)	98.86(9.51)	59.51(1.60)	60.35(2.57)	87.29(9.29)
DCGAN-C	100	45.02(3.37)	73.12(11.03)	85.50(12.25)	47.60(2.57)	50.48(4.58)	78.11(12.95)
DCGAN-C	1K	**40.85(3.41)**	**55.63(7.97)**	**72.11(13.81)**	**40.87(3.03)**	**45.46(5.03)**	**57.13(7.20)**
PGAN	100	21.62(1.73)	28.15(3.43)	47.94(5.71)	25.43(2.19)	22.86(2.06)	45.16(7.87)
PGAN	1K	19.05(3.14)	25.19(5.26)	43.48(12.24)	19.20(2.96)	19.05(2.82)	35.07(8.72)
PGAN	10K	**16.75(1.87)**	**18.96(2.65)**	**37.01(8.74)**	**16.94(1.89)**	**17.39(2.33)**	**26.63(4.44)**