SynthProv: Interpretable Framework for Profiling Identity Leakage

Jaisidh Singh[†], Harshil Bhatia[†], Mayank Vatsa[†], Richa Singh[†], Aparna Bharati[◇]
[†] IIT Jodhpur, India [◇] Lehigh University, PA, USA
{singh.118, bhatia.2, mvatsa, richa}@iitj.ac.in apb220@lehigh.edu

Abstract

Generative Adversarial Networks (GANs) can generate hyperrealistic face images of synthetic identities based on a latent understanding of real images from a large training set. Despite their proficiency, the term "synthetic identity" remains ambiguous, and the uniqueness of the faces GANs produce is rarely assessed. Recent studies have found that identities from the training data can unintentionally appear in the faces generated by StyleGAN2, but the cause of this phenomenon is unclear. In this work, we propose a novel framework, SynthProv, that utilizes the improved interpolation ability of StyleGAN2 latent space and employs image composition to analyze leakage. This is the first method that goes beyond detection and traces the source or provenance of constituent identity signals in the generated image. Experiments show that SynthProv succeeds in both detection and provenance tasks using multiple matching strategies. We identify identities from FFHQ and CelebA-HQ training datasets with the highest leakage into the latent space as "leaking reals". Analyzing latent space behavior to evaluate generative model privacy via leakage is an important research direction, as undetected leaking reals pose a significant threat to training data privacy. Our code is available at https://github.com/jaisidhsingh/SynthProv.

1. Introduction

Synthetic image generation is desirable for creating and augmenting datasets for sensitive tasks [38] for which obtaining a large and diverse set of real-world data is challenging, such as robust facial analytics [39, 53, 57]. Generative Adversarial Networks (GANs) are one of the most widely used generative models due to their ability to generate high-fidelity images from random noise [15], upon learning a representative latent distribution given a large set of training samples. Improved understanding of the latent distribution and its editability has led to more controlled and higher quality image generation [45, 55, 62]. Images generated using state-of-the-art GAN architectures such as StyleGAN2 [23] not only resemble the photorealism of real images but are considered better than real images in aspects such as human trustworthiness [27, 36]. This suggests that these generated images could potentially substitute actual training samples in facial recognition tasks [20, 39, 57]. Nevertheless, the reliance of the image generation process on the genuine training examples raises privacy issues for user content within the training dataset [28], even when it is not explicitly utilized for downstream tasks.

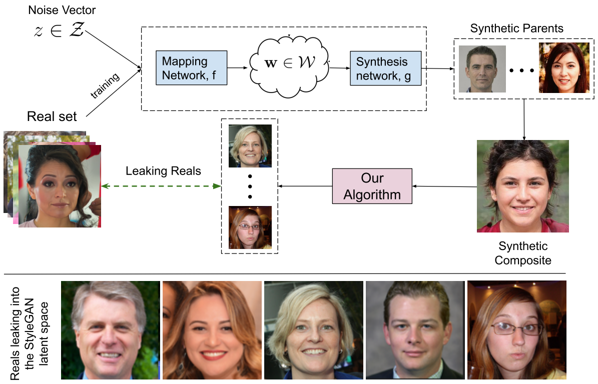


Figure 1. Overview of *SynthProv*: the first of its kind framework to trace the *leakage of real identities into the learned StyleGAN2 latent space*. We exploit the properties of semantically meaningful interpolation in the latent space to construct composite images, which are used to trace the presence of identity-related information in synthetic images to contributing real images.

Deeper understanding of the relationship between the generated images and the training data has demonstrated that information from the training data leaks into the synthetic images through associations between samples in the latent space [49, 56]. In most cases, GAN latent space may not enable direct matching of real and synthetic image pairs due to its non-Euclidean nature [30], but it does not rule out detectable associations between training images of real subjects and synthetically generated images. Generative models differ from other predictive models, as they generate images as predictions, which may share visual features with training images on various latent manifolds learned for different image understanding tasks. Leaked information from training images can be easily detected through the released synthetic image data using image matching techniques [49]. Identifying instances of information leakage is essential for assessing the privacy of models, yet detection alone does not ensure privacy. To effectively reduce leakage and improve the privacy of training data, it is critical to examine both the origin and characteristics of this leakage.

For improved generation and editability, latent manifolds learned by GANs provide the ability to interpolate and generate semantically relevant images [45, 62] while imposing very few limits on the density with which the latent distribution can be sampled [7]. This has helped discover semantically relevant directions in the StyleGAN2 latent space [45, 55, 62] and enabled controlled image editing via operations on latent vectors along those directions. The proposed approach utilizes these properties and employs strategic sampling with aggregation of representations in the latent space to highlight shared information between the support set, i.e., training images, and the generated samples. Due to certain directions encoding specific facial attributes [45], *composite images* constructed in the latent space from samples along identity-invariant latent directions, act as pseudo-generated images and can help highlight information sharing behavior essential for detecting and tracing identity leakage. The difference between this behavior for synthetic and real image representations in the latent space, is utilized to detect leakage and retrieve the images leaking most information, termed as *leaking reals*. The proposed method (see Figure 1) traces the identity information present in synthetic faces back to the real training set. This extra step provides output in a *human interpretable* format, where synthetic images can be visually compared to a group of real images to explain leakage.

To the best of our knowledge, this is the first method to conduct provenance of identity leakage within a widely-used GAN model such as StyleGAN2 [23]. In other words, the proposed framework leverages the latent space dynamics of powerful and easily controllable GANs to detect and trace back, or ascertain the source of, identity information that has inadvertently leaked from the training images. Experiments profile identity information leaking from the commonly used training sets of StyleGAN2 and highlight the real identities which are at risk in terms of privacy.

2. Related Work

The proposed method associates content between synthetic images and training images to identify sources of shared information. To do so, it builds upon the existing understanding of GAN latent space, prior observations of leakage in the space and image provenance analysis for immensely complex composite images. In this section, we discuss the three groups and highlight how the proposed work builds upon their findings to conduct provenance of leakage.

2.1. Latent Space Understanding

Modern image generation and manipulation methods have tried to understand and edit latent representations to increase the control and fidelity for conditional image generation [24,45,62]. Latent space of GANs have been treated as a Reimannian Manifold [2, 8] and several works have proposed editing images based on the arithmetic properties of the latent space [40, 52]. AttGAN [19] successfully modeled the relation between attributes and the latent space learned by a GAN for constrained conditional manipulation. Other models exchange attribute information in latent codes [58, 67] or disentangle images into identity and non-identity based attributes for editing [25, 47]. Shen *et al.,* [45] in their method that enables face editing in the latent space of GANs, show that certain directions and subspaces can correspond to representation of different attributes. The understanding developed by these methods enable the design of our framework, which relies on latent space sampling and traversal to highlight information leakage and the sources. For synthetic image editing, a sampled vector from the GAN latent space is directly changed whereas for manipulating real images, most methods employ an inversion approach [1, 10, 21, 42]. State-of-the-art inverters use a two-step process. The first stage produces a latent vector, which is then modified by the generator. Two popular recent approaches are Pivotal Tuning Inversion (PTI) [43] and HyperInverter [13]. PTI finetunes the generator for every latent vector to improve reconstruction, leading to expensive inference time. In contrast, HyperInverter uses hypernetworks to predict the residual weights (information lost when mapping input image to the $\mathcal{W}$ space) and finetunes the generator using this to reconstruct the final image. This reduces fine-tuning required during inference and for better efficiency. Our method requires combining information in latent space to create composite identities from real ones, which are evaluated for leakage. In our experiments, we use HyperInverter to infer the representations of the real images in an already learnt GAN latent space, as it balances the quality-efficiency tradeoff.

2.2. Information leakage in GANs

The adversarial training in GANs encourages a distribution that is centered around training samples. This puts the current generation GANs at a significant risk of disclosing private information from training images through the generated samples, which is termed information leakage [61]. Goodfellow *et al,* [15] hypothesize that given infinite time, GANs can recreate the samples found in the training data. Further, [14] showed that given access to model architecture and weights, it is possible to reconstruct a person's facial image using facial recognition confidence scores. The vulnerabilities of GANs and other generative models are further highlighted using *membership inference attacks* [46] where the goal is to detect whether a query image was used to train the model [9,41,44,51]. The success of such attacks have motivated the design of GANs architectures with privacy guarantees [20,50,59,61]. In the literature, techniques for mitigation have received more attention than detection, but the recent study by Tinsley *et al.* [49] specifically discusses leakage of identity information in StyleGAN2 space. The study hypothesizes that identity information from real face images seen during the training of a GAN leak into the latent space. Upon observing comparison score distributions for a given set of *Real - Real* (R-R) and *Real - Generated* (R-G) image pairs, they infer that the presence of identity leakage is detected using only specific face matchers such as ArcFace. Similar to their work, we perform distribution analysis using the properties of the latent space to study identity leakage. However, our construction of evaluation pairs differs significantly from their approach and provides better identity relevant matching. Additionally, our approach also associates identity leakage with the source images and performs provenance analysis.

2.3. Image Provenance Analysis

Owing to the large scale online availability of manipulated images, *i.e.,* containing content from multiple donor images [5], Image Provenance Analysis [32] aims to identify the origin and intent of such content. Actively tagging all online images for ease of tracking its processing and usage [54,60] requires standardization of the procedure and may not always be feasible. Provenance analysis approaches in the literature that solely rely on image content are a two-stage process [31]. The first step is query, i.e. retrieving related images to any given image and the second is creating a directed acyclic graph where the edges denote pairwise forensic relationship [66] or content contribution from one image to another. This step-by-step analysis [3] is intuitive for manipulated content in online media, but the general framework may not directly apply to manipulated content generated using GANs. Provenance for GAN-generated deepfakes has also been proposed to trace possible sources of content [34,35]. The possible set of pairwise relationships for this use case is smaller than the general provenance framework, as it mostly considers deepfakes as a composite of two original source images. This setup is different from the proposed approach, in which we associate sampled synthetic images from a latent distribution to the true observed samples of that distribution.

3. Methodology

StyleGAN2-style architectures utilize learned latent space from training images to represent face images on manifolds, which may vary in terms of the richness of encoding semantic information. Previous works [13,43] learn to map realistic face images to these latent representation spaces. The latent manifold is guided by the information of real face images, and thus identity information encoded in this manifold is a derivative, or a function of the identities of real faces in the training set. The highly complex and variable function for each face representation in the latent space is learned during training. This serves as the basis of identity leakage, which we analyze by using the semantically rich latent space. The shared semantic information between pairs of features is utilized to construct composite face images (in Sec. 3.1) which are then used to associate real data with synthetic samples and detect identity leakage. Subsequently, Sec. 3.2 describes the extraction and usage of a specific direction in the StyleGAN2 latent space that encodes identity-invariance. Lastly, we propose SynthProv, a method that characterizes the latent space using composite images and incorporates the identity-invariant direction to perform provenance of identity information (Sec. 3.3).

3.1. Composites for Identity Leakage

Given a random noise vector $z \sim \mathcal{N}(0,1)$, StyleGAN2 first maps this noise vector to a latent vector $l \in \mathcal{W}$, by $l = F_{map}(z)$, where F_{map} is the mapping network. This latent vector l is then utilized by G, the generator, to produce a face image I, given by $I = G(l)$. The mapping of the latent vector l by the generator network leads to the encoding of rich representations of face images in the $\mathcal{W}$ space. This enables semantic editing tasks such as facial interpolation and editing [29,42,43,64] and style disentanglement [19,47,52]. We utilize this semantically meaningful $\mathcal{W}$ space to construct two types of image composites, *synthetic composites* and *real composites* which are obtained in the following manner.

Here, a $\mathcal{Q}$-set notation is used as a placeholder for our two settings of synthetic and real latent vectors and face images. We create a $\mathcal{Q}$-set composite, where $\mathcal{Q} = \{l, l \in \mathcal{W}\}$ and randomly sample k vectors from $\mathcal{Q}$, to obtain a *$\mathcal{Q}$-set composite vector* l_c^q using

$$l_c^q = \frac{1}{k}\sum_{j=1}^{k} l_{p_j}^q \tag{1}$$

where $\{l_{p_1}^q, ...l_{p_k}^q\}$ are the *$\mathcal{Q}$-set parent vectors* of l_c^q. Particularly, $k <<< |\mathcal{Q}|$, and the remaining non-parent synthetic latent vectors are denoted by $\{l_n^q\} = \mathcal{Q} - \{l_p^q\}$. Subsequently, we use the generator G to obtain images for each vector in $\mathcal{Q}$ denoted by I^q. The face image corresponding to a composite latent vector l_c^q is denoted by I_c^q. Similarly, the images obtained by decoding the parent vectors and non-parent vectors of l_c^q are given by $I_p^q = G(l_p^q)$ (parent images) and $I_n^q = G(l_n^q)$ (non-parent images) respectively.

We utilize the aforementioned formulation for $Q \in \{\mathcal{S}, \mathcal{R}\}$. Here $\mathcal{S}$ is the set of randomly sampled synthetic latent vectors, whereas $\mathcal{R}$ is the set of latent vectors representing real face images from the training set in the $\mathcal{W}$

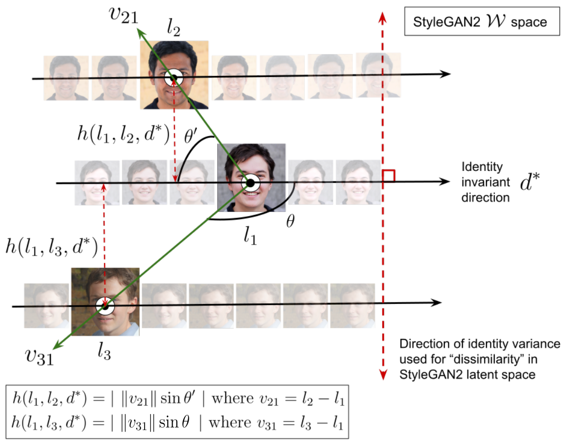

v_{21}
l_2
StyleGAN2 $\mathcal{W}$ space
$h(l_1, l_2, d^*)$
θ'
Identity invariant direction
d^*
θ
l_1
$h(l_1, l_3, d^*)$
v_{31}
l_3
Direction of identity variance used for "dissimilarity" in StyleGAN2 latent space
$h(l_1, l_2, d^*) = | \ \|v_{21}\| \sin\theta' \ |$ where $v_{21} = l_2 - l_1$
$h(l_1, l_3, d^*) = | \ \|v_{31}\| \sin\theta \ |$ where $v_{31} = l_3 - l_1$

Figure 2. Computation of latent identity distance (refer Sec. 3.2) in the GAN latent space. The identity dissimilarity is measured perpendicular to the global identity invariant direction, d^*.

space, obtained using HyperInverter [13]. The notations regarding the latent representations and their corresponding images are described in the rest of the section (see supplementary material for a table with all notations).

3.2. Identity in GAN Latent Space

Prior research [24, 45, 62] has demonstrated high quality manipulation of facial attributes through the projection of latent vectors along disentangled directions that encode specific style or attributes. Building upon this line of research, we first compute identity-invariant directions in the $\mathcal{W}$ space and then use those directions to compute an identity distance metric. Note that this is different from a face-matcher, which encodes identity-invariance in a different image-derived space. Face matchers generally try to minimize distance between representations of images of the same identity, while maximizing distance between different identity representations. The learned space only encodes identity-invariant information, with very few other image-level details. Whereas, StyleGAN's $\mathcal{W}$ space is known to encode different levels of face details, with directions corresponding to specific attribute information and some encoding higher identity invariance than others. Using a direction to compute dissimilarity in the identity features encoded in the latent vectors of the $\mathcal{W}$ space, allows us to use its other properties, such as GAN inversion, to relate latent behavior with image space changes. This facilitates interpretability in the leakage analysis.

A face-matcher m operates on a face image I to produce an embedding given by $e = m(I)$. For any two face images I_1 and I_2, these identity embeddings are given by $e_1 = m(I_1)$ and $e_2 = m(I_2)$. A higher similarity in identities is represented when two embeddings are closer in the face-matcher space, which is quantified by a *dissimilarity-based match score*, $\phi(e_1, e_2)$, between the two embeddings. This section describes how we compute identity dissimilarity in the $\mathcal{W}$ space.

Identity-invariant direction: We utilize the relation of identity information and latent representations in the $\mathcal{W}$ space, and find a direction in the $\mathcal{W}$ space which encodes identity-salient features. To do this, we solve the optimization problem,

$$d^* = \underset{d \in \mathcal{W}}{\arg\min} \quad \phi(e_0^s, m(G(l_0^s + \alpha d))) \tag{2}$$

where d is the parameterized identity-invariant direction. Here, l_0^s is a fixed, latent vector, which is decoded by G to produce a face image I_0^s. The face-matcher embedding of I_0^s is given by $e_0^s = m(I_0^s)$.

The identity-invariant direction is optimized for each image. However, experimental results (see supplement for results of this step) show that an overall globally consistent and instance independent optimal direction d^* is obtained. This optimal identity-invariant direction d^* is used in constructing the identity distance.

Latent identity distance: This metric quantifies the dissimilarity between two latent vectors by utilizing the global identity invariant direction, d^*. By incorporating d^* into our framework, we aim to encompass identity representations within the semantically rich latent space. Consequently, we define $h(l_1, l_2, d^*)$ as the measure of identity distance between two latent vectors, $l_1, l_2 \in \mathcal{W}$ given by:

$$h(l_1, l_2, d^*) = | \ \|v\| sin\theta \ | \tag{3}$$

where $v = l_2 - l_1$ and $\theta = cos^{-1}(\frac{v \cdot d^*}{||v|| \cdot ||d^*||})$, which is the angle between the latent vector v and d^*. We observe that the identity doesn't change as we decode points while traversing along d^* (results are shown in the supplementary material). Consequently, evaluating the dissimilarity in identity between two latent vectors can be more accurately assessed by measuring the distance perpendicular to the direction invariant to identity shown in Figure 2.

3.3. SynthProv Framework

The proposed, *SynthProv*, framework utilizes the semantic information gathered in the synthetic composites and the properties of the face matcher space and the $\mathcal{W}$ space for retrieving real faces showing identity leakage. Figure 3 presents an overview of the framework and the steps are explained as follows.

3.3.1 Selection of assistants

Our algorithm begins with a synthetic composite image I_c^s, and its synthetic parent images I_p^s. Firstly, for each synthetic composite image, a threshold t is computed, given by

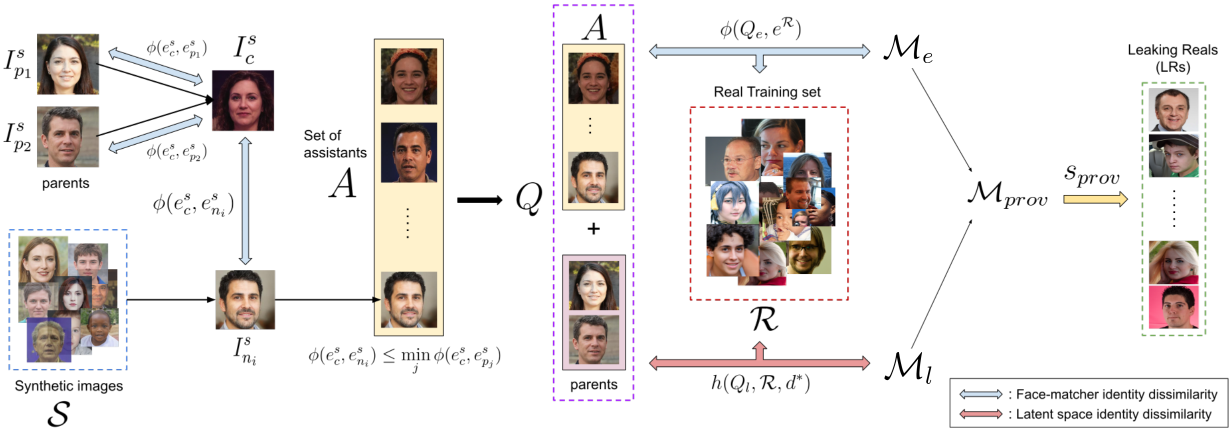


Figure 3. The overview of the proposed SynthProv framework considering only two parents ($k = 2$). We first construct the composite image I^s_c using which a set of assistants is chosen from non-parent synthetic images I^s_n. Next, we compute identity dissimilarity of assistants and parents to the real training set, in both the face-matcher space and the $\mathcal{W}$ space. Finally, these mixed scores of identity dissimilarity are ranked to trace the real faces leaking identity into synthetic samples.

$$t = \min_{j \in \{1,\ldots,k\}} \phi(e^s_c, e^s_{p_j}) \tag{4}$$

where $e^s_{p_j}$ is the face-matcher embedding of the j^{th} synthetic parent of the synthetic composite. We then select *assistants*, A, a set of synthetic non-parent images in I^s_n, which appear closer to the synthetic composite in the face-matcher space than any of its parents. Formally, given by

$$A = \{i \mid \phi(e^s_c, e^s_{n_i}) \leq t\} \tag{5}$$

In instances where the set A is large, we constrain the cardinality of A to a, restricting the analysis solely to the indices closest to the parents. Identity signals of the synthetic composite are pooled across its parents and the assistants, to use them as the best proxies for the identity of the synthetic composite image. These identity proxies are then used for the provenance of identity leakage, as described below.

3.3.2 Mixed score-based retrieval

The set of latent vectors corresponding to identity proxies is defined as $Q_l = \{l^s_{p_1}, \ldots, l^s_{p_k}, l^s_{A_1}, \ldots, l^s_{A_q}\}$. We compute the latent identity distance of each member of Q_l from each real latent vector in $\mathcal{R}$. This results in a distance-matrix $\mathcal{M}_l$ given by

$$\mathcal{M}_l = h(Q_l, \mathcal{R}, d^*). \tag{6}$$

Additionally, we construct a proxy embedding set Q_e, such that $Q_e = \{e^s_{p_1}, \ldots, e^s_{p_k}, e^s_{A_1}, \ldots, e^s_{A_q}\}$. Each face-matcher embedding in the proxy embedding set Q_e, is matched to the face-matcher embeddings of the real training set, $e^{\mathcal{R}}$. and the scores are denoted by $\mathcal{M}_e$, where

$$\mathcal{M}_e = \phi(Q_e, e^{\mathcal{R}}). \tag{7}$$

Using these, we compute $\mathcal{M}_{prov} = \mathcal{M}_e \circ \mathcal{M}_l$. The matrix $\mathcal{M}_{prov}$ contains the combined dissimilarity scores (both matcher and latent space based) of each identity proxy from the query image. Lastly, the row-wise mean of the matrix $\mathcal{M}_{prov}$ is used to obtain *provenance scores*, given by s_{prov}. This score signifies the extent of the identity contribution of each training face, in the synthetic composite face. The lower the value of the provenance score, the greater is the extent of identity leakage. The images having the lowest provenance scores are termed as *leaking reals* (LRs) for a given synthetic composite image query.

4. Implementation Details

Dataset Generation - We construct synthetic image set $\mathcal{S}$, synthetic composite images and real composite images from a given set of real images $\mathcal{R}$ in the following manner. For each dataset that StyleGAN2 was trained on, $|\mathcal{S}| = |\mathcal{R}| =$ number of training samples of the dataset. Thus, for the Flickr Faces High Quality dataset, or FFHQ, $|\mathcal{S}| = |\mathcal{R}| = 70,000$ while for the CelebAHQ [24] dataset $|\mathcal{S}| = |\mathcal{R}| = 30,000$. Further, for each dataset we vary k from 2 to 6, and construct $10,000$ composites for each value of k. This is done for both synthetic and real composites.

Face Matchers - Our experiments employ two face matchers to obtain $\mathcal{M}_e$, namely ElasticFace [4] and ArcFace [12]. ArcFace, uses an additive angular loss with a CNN backbone such as ResNet-100 [18] to learn highly discriminative features for face recognition, and is trained on the MS-Celeb-1Mv2 [17]. ElasticFace employs flexible margin values in existing spherical loss functions to outperform previous approaches. The CASIA-WebFace [63] dataset is used for training ElasticArcFace+ with a ResNet-50 backbone. In particular, we use the ElasticArcFace+ version of ElasticFace. Cosine distance is used for matching using Arc-

Face, while ElasticFace uses Euclidean distance to compute match scores. We utilize normalized face-matcher embeddings in all our experiments.

Provenance - In our algorithm, for each synthetic composite query, we first select a assistants to act as identity proxies along with its parents. The value of a is set to be 10 in our experiments, for all values of k. This is done so that the identity proxies consist of highly similar identities w.r.t. that of the query face image.

To minimize eq. (2), we sample points in the $\mathcal{W}$ space at steps of size α from l_0^s, which are decoded into face images $G(l_0^s + \alpha d)$. This is done by iteratively incrementing α by a fixed value $\delta = 0.1$, hence the update in α is given by $\alpha \leftarrow \alpha + \delta$. We use PyTorch [37] as the framework for our experiments, where d is directly set as a learnable parameter. The Adam [26] optimizer, initialized with a learning rate of $1e - 4$, is used for the backpropagation of the mean squared error. Additionally, both d and d^* are normalized by their L2 norm.

Lastly, the number of synthetic composite queries in our algorithm is set as $10,000$. All experiments are run on an NVIDIA V100 GPU. We pre-compute $h(\mathcal{S}, \mathcal{R}, d^*)$ and $\phi(e^{\mathcal{S}}, e^{\mathcal{R}})$ for efficiency. Finally, our framework takes approximately 2.5 hours to retrieve leaking reals and approximately 14.5 hours is needed for the pre-computation of h.

5. Experiments

Training StyleGAN2 on the FFHQ and CelebAHQ datasets separately guides the latent representation space differently. Different identities present in the training set, along with different dataset sizes and training iterations, can affect how faces and identities are encoded in the $\mathcal{W}$ space. Hence, we evaluate our methodology for each variant of StyleGAN2. To investigate identity leakage, the following experiments were devised.

5.1. Embedding Space Density

Composite images created using latent space interpolation capture more information shared among samples in the latent space than individual synthetic samples. Due to this, synthetic composites can be more useful for identity leakage analysis than randomly sampled synthetic images themselves. This can be verified by evaluating the embedding space density [33]. Density has been used as a metric to evaluate the diversity of the synthetic samples of generative models. Given a feature space containing real samples and synthetic samples (given by X and Y respectively), density evaluates the expected number of real-sample neighborhoods which contain a synthetic sample Y_j. Formally, density $D(\cdot, \cdot)$ is defined as

$$D(X, Y) = \frac{1}{kM} \sum_{j=1}^{M} \sum_{i=1}^{N} 1_{Y_j \in B(X_i, NND_u(X_i))} \tag{8}$$

where $N = |X|$ and $M = |Y|$. $B(x, r)$ is the sphere in the feature space around x with radius r and $NND_u(X_i)$ is the distance from X_i to the u^{th} nearest neighbor among $X \setminus \{X_i\}$. In this experiment, we show how density in the face-matcher embedding space can be associated with identity leakage. More real embedding neighborhoods containing a synthetic embedding imply more real identities being similar to the synthetic identity features, hence greater identity leakage. Therefore, we evaluate if $D(e^{\mathcal{R}}, e_c^{\mathcal{S}}) > D(e^{\mathcal{R}}, e^{\mathcal{S}})$ for all considered datasets and matchers, and the results are presented and analyzed in Sec. 6.1.

5.2. Match Scores for Identity Leakage Detection

To detect identity leakage, we perform distribution analysis on synthetic composite images and real composite images. We compute the distribution of $\phi(e_c^{\mathcal{S}}, e_p^{\mathcal{S}})$, the match scores between the identity embeddings of synthetic composite images and their synthetic parent images. We juxtapose this distribution with the distribution of $\phi(e_c^{\mathcal{R}}, e_p^{\mathcal{R}})$, the match scores between the identity embeddings of real composite images, and real images which are strictly their parents. Similar distributions of the match score of both types of composite images, but with strictly non-parent synthetic and real images, i.e. the distributions of $\phi(e_c^{\mathcal{S}}, e_n^{\mathcal{S}})$ and $\phi(e_c^{\mathcal{R}}, e_n^{\mathcal{R}})$ are also computed. The match score distributions with parents and non-parents are compared to establish of the presence of identity leakage in StyleGAN2.

5.3. Leaking Reals Retrieval

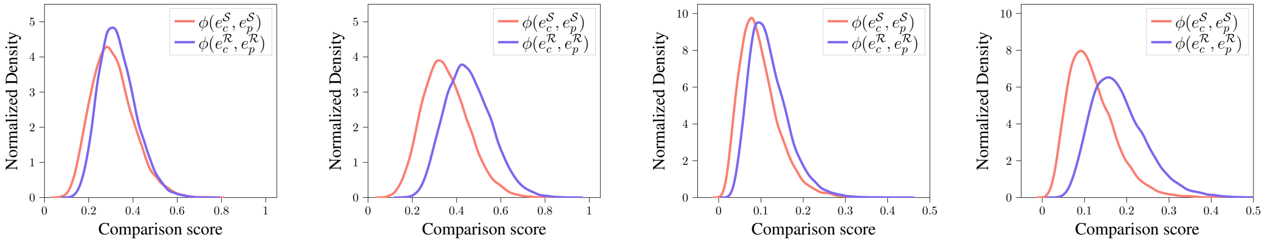
For this experiment, the algorithm defined in Sec. 3.3 is used to retrieve the leaking reals queried for $10,000$ synthetic composite queries, for each value of k. The obtained provenance scores s_{prov} are used to rank the real face images in the training set to obtain the top-5 leaking reals (LR) for each synthetic composite query. Additionally, the match scores of these leaking reals to their respective queries are analyzed in order to evaluate the presence of identity leakage in the GAN's latent space. The match scores for leaking reals being significantly lesser than those of real parents to real composites indicates that LR images contribute their identities more to the synthetic composite query, than to their real composite. This is performed for all face-matchers and datasets, with the results reported in Sec. 6.3.

6. Results

The results of the experiments described in the previous section are discussed below.

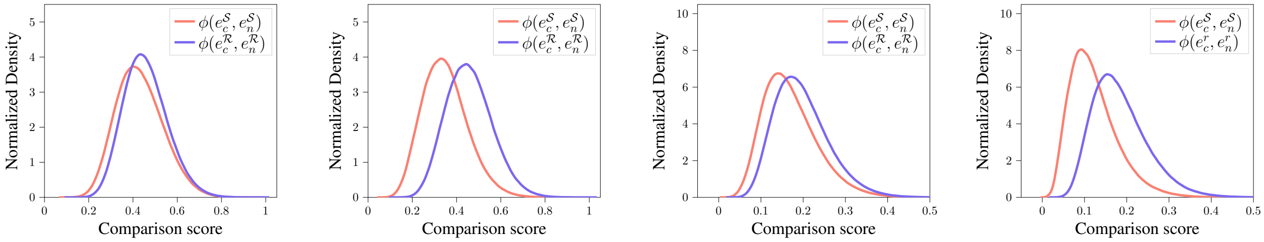
6.1. Embedding Space Density

We compute density as devised in Sec. 5.1 for the given face matchers in the embedding space. Table 1 shows that the density of synthetic composites w.r.t to the set of real



(a) ElasticFace on FFHQ (b) ElasticFace on CelebAHQ (c) ArcFace on FFHQ (d) ArcFace on CelebAHQ

Figure 4. Match score distributions between real and synthetic composite image and their parent images for FFHQ and CelebAHQ



(a) ElasticFace on FFHQ (b) ElasticFace on CelebAHQ (c) ArcFace on FFHQ (d) ArcFace on CelebAHQ

Figure 5. Match score distributions between real and synthetic composite image and their non-parent images for FFHQ and CelebAHQ.

images D($e^{\mathcal{R}}, e_c^{\mathcal{S}}$) is greater than that of synthetic images $D(e^{\mathcal{R}}, e^{\mathcal{S}})$ for all our datasets and face-matchers. Thus, on average, the embedding of a synthetic composite image has more real embeddings as its u nearest neighbors in comparison to the embedding of a randomly sampled synthetic image. This signifies that a set of synthetic composite faces shows greater association with the identities of real faces, in the face-matcher space, than that shown by non-composite synthetic faces. The transitive aggregation of real identity information leaking into its parents leads to greater density, and hence, greater identity leakage in synthetic composite faces. Thus, synthetic composite images serve as better samples for the analysis of identity leakage.

Table 1. Density computation in the matcher embedding space for FFHQ and CelebAHQ.

Dataset	Face-matcher	$D(e^{\mathcal{R}}, e_c^{\mathcal{S}})$	$D(e^{\mathcal{R}}, e^{\mathcal{S}})$
FFHQ	ArcFace	5.20	3.13
	ElasticFace	3.65	2.36
CelebAHQ	ArcFace	6.97	5.11
	ElasticFace	5.30	3.76

6.2. Match Scores for Identity Leakage Detection

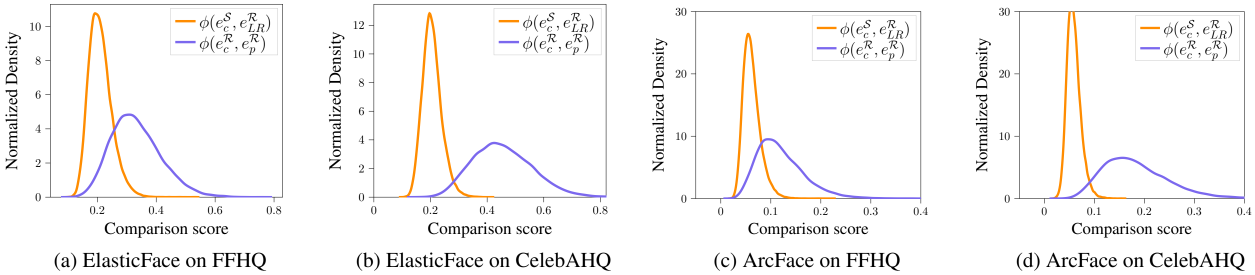
In Figure 4, in the embedding space of the face matcher, synthetic parent images are closer to their composite image than real parents are to their composite image. This exhibits that synthetic composites have more common identity features with their respective synthetic parents, than real composites have with their real parents. Additionally, synthetic composites have an overall lower dissimilarity score w.r.t reals. The real training set images are considered to have unique identities. Therefore, when a real composite is constructed, due to the semantic attributes of the latent space, the identity features of the real images are subdued. This composite can have a highly entangled identity which is perceived by the face matchers to be different from the distinct identities of the parents. We also analyze match scores of synthetic composites with the remaining synthetic samples given in Figure 5. Similar to Figure 4, synthetic composites are on average closer. This implies that there exist identity signals that are not only common between the composite and its parents, but they are also shared in the entire $\mathcal{W}$ space, confirming the presence of identity leakage.

6.3. Retrieving Leaking Reals

We present the results of provenance to obtain leaking reals (LRs), or real face images showing most identity leakage into synthetic composite queries. We report the match scores of synthetic composite images to their LRs in Figure 6. The dissimilarity scores of the LRs are less to a synthetic composite, than comparison scores of real parent images to their own composite images. This shows that a synthetic composite is more similar to a leaking real image than a real composite is to its parents. This provides further empirical proof regarding identity leakage of real samples. In Figure 7, we show qualitative results of our algorithm in retrieving the LRs for the given queries in the leftmost column of the figure. We see retrieval of real faces showing similar and coherent identity features w.r.t to the queries.

6.4. Ablation Studies

Varying the number of parents considered to represent identity signals for a synthetic image can affect detection



(a) ElasticFace on FFHQ (b) ElasticFace on CelebAHQ (c) ArcFace on FFHQ (d) ArcFace on CelebAHQ

Figure 6. Match scores of LRs computed using SynthProv with synthetic composite queries shown with match scores of real composite image with their respective real parents. These scores are computed using all face-matchers on both datasets, for $k = 2$.

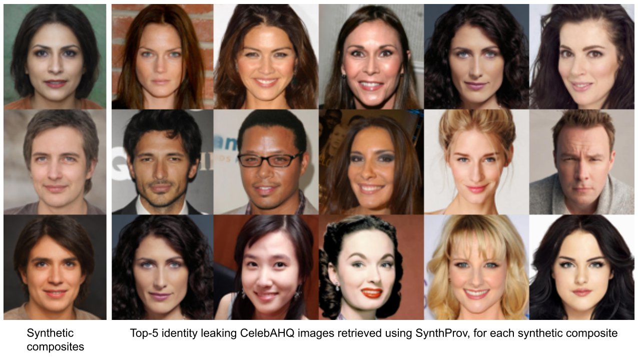

Synthetic composites
Top-5 identity leaking CelebAHQ images retrieved using SynthProv, for each synthetic composite

Figure 7. Qualitative results of SynthProv using ElasticFace on the CelebAHQ dataset. These are the set of real images showing the highest identity leakage, given synthetic composites as query images. The synthetic composites have $k = 2$ synthetic parents.

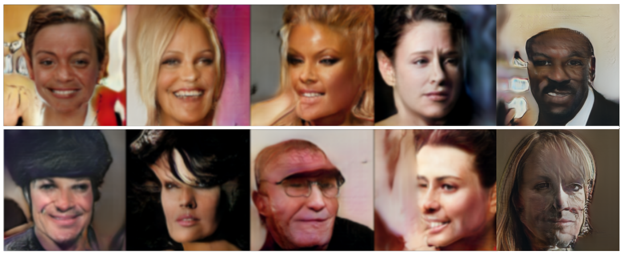


Figure 8. Example low-quality images frequently obtained from PGGAN when its latent space is sampled at a large scale.

and provenance of leakage. Thus, we conduct an ablation experiment to understand the effect of varying the number of parents when creating a composite. For $k \in [2, 6]$, we compute the match scores of the synthetic and real composite with their respective parents and present the mean and standard deviation of the obtained distributions (Table 2 shown in supplementary material). Different trends were observed with respect to k within experiments with different face-matchers and datasets. However, the overlap between the distributions of the real and synthetic parents does not increase, implying that the detection of identity leakage is independent of the choice of k. We further ablate the distributions of match scores of LRs from their queries with respect to k, and find that these distributions also do not converge, showing our algorithm is independent of choice of k (plots shown in Figure 2 of supplementary material).

6.5. Other Generative Architectures

Besides StyleGAN2, PGGAN [22] has also been explored for face image generation and inversion [16, 65]. However, evaluating our method on PGGAN, we find a significant decrease in face image quality as compared to StyleGAN2, shown in Figure 8. This can be due to the less semantically dense latent space learned by PGGAN in comparison to StyleGAN2, which hampers the realism of face images when the latent space is sampled at a large-scale for our experimental setting. Matching unrealistic looking faces is also not well-suited for face-matchers, making PGGAN unfit for quantitative evaluation. While there are other generative models, our experiments focus on GANs, as they are believed to be significantly more private than other state-of-the-art generative models such as diffusion models [6, 48]. A study by Carlini *et al.* [6] showed that the entire training data for diffusion models [11] is extractable through their generate-and-filter pipeline. Our work shows that identity leakage is traceable in StyleGAN2, implying that privacy risks exist in both GANs and diffusion models.

7. Conclusion

Detectable identities in generative models pose a threat to the privacy of individuals present in the training data. This paper introduces the first framework to profile leakage of identity information from training data to synthetic data. The rich $\mathcal{W}$ latent space allows us to (a) create identity aggregated composites and (b) find a globally identity-invariant direction. SynthProv uses this information with a face-matcher to trace identity. Our work successfully highlights the privacy threat posed by identity traceability in a popular model, StyleGAN2. However, identity leakage may be present in various generative models, varying with the specific architecture used. We hope this work inspires further research on generalizable identity leakage analysis for robust and private image generation.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent

space? In *IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 2

[2] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017. 2

[3] Aparna Bharati, Daniel Moreira, Patrick J. Flynn, Anderson de Rezende Rocha, Kevin W. Bowyer, and Walter J. Scheirer. Transformation-aware embeddings for image provenance. *IEEE Transactions on Information Forensics and Security*, 2021. 3

[4] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5

[5] Joel Brogan, Aparna Bharati, Daniel Moreira, Anderson Rocha, Kevin W Bowyer, Patrick J Flynn, and Walter J Scheirer. Fast local spatial verification for feature-agnostic large-scale image retrieval. *IEEE Transactions on Image Processing*, 30:6892–6905, 2021. 3

[6] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023. 8

[7] Dongjie Chen, Sen-ching Samson Cheung, Chen-Nee Chuah, and Sally Ozonoff. Differentially private generative adversarial networks with model inversion. In *IEEE International Workshop on Information Forensics and Security*, 2021. 2

[8] Nutan Chen, Alexej Klushyn, Richard Kurle, Xueyan Jiang, Justin Bayer, and Patrick Smagt. Metrics for deep generative models. In *International Conference on Artificial Intelligence and Statistics*, 2018. 2

[9] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International Conference on Machine Learning*. PMLR, 2021. 3

[10] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE Transactions on Neural Networks and Learning Systems*, 30(7):1967–1974, 2018. 2

[11] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 8

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 5

[13] Tan M Dinh, Anh Tuan Tran, Rang Nguyen, and Binh-Son Hua. Hyperinverter: Improving stylegan inversion via hypernetwork. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2, 3, 4

[14] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security*, 2015. 2

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[16] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3012–3021, 2020. 8

[17] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016. 5

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 5

[19] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 2019. 2, 3

[20] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018. 1, 3

[21] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 8

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 4, 5

[25] Amena Khatun, Simon Denman, Sridha Sridharan, and Clinton Fookes. Semantic consistency and identity mapping multi-component generative adversarial network for person re-identification. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020. 2

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[27] Federica Lago, Cecilia Pasquini, Rainer Böhme, Hélène Dumont, Valérie Goffaux, and Giulia Boato. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Processing Magazine*, 39(1):109–116, 2021. 1

[28] Zinan Lin, Vyas Sekar, and Giulia Fanti. On the privacy properties of gan-generated samples. In *International Conference on Artificial Intelligence and Statistics*, 2021. 1

[29] Hongyu Liu, Yibing Song, and Qifeng Chen. Delving stylegan inversion for image editing: A foundation latent space

viewpoint. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10072–10082, 2023. 3

[30] Mike Yan Michelis and Quentin Becker. On linear interpolation in the latent space of deep generative models. In *ICLR Workshop on Geometrical and Topological Representation Learning*, 2021. 1

[31] Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin W. Bowyer, Patrick J. Flynn, Anderson Rocha, and Walter J. Scheirer. Image provenance analysis at scale. *IEEE Transactions on Image Processing*, 2018. 3

[32] Daniel Moreira, William Theisen, Walter Scheirer, Aparna Bharati, Joel Brogan, and Anderson Rocha. Image provenance analysis. In *Multimedia Forensics*, pages 389–432. Springer, Singapore, 2022. 3

[33] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 6

[34] Kartik Narayan, Harsh Agarwal, Surbhi Mittal, Kartik Thakral, Suman Kundu, Mayank Vatsa, and Richa Singh. Desi: Deepfake source identifier for social media. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 2022. 3

[35] Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, and Richa Singh. Deephy: On deepfake phylogeny. *arXiv preprint arXiv:2209.09111*, 2022. 3

[36] Sophie J Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *National Academy of Sciences*, 119(8), 2022. 1

[37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019. 6

[38] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. 1

[39] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *IEEE/CVF International Conference on Computer Vision*, 2021. 1

[40] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[41] Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[42] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[43] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*, 42(1), 2022. 2, 3

[44] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018. 3

[45] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 4

[46] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, 2017. 2

[47] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2, 3

[48] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. 8

[49] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 1, 2, 3

[50] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 3

[51] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019. 3

[52] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3

[53] Richard Van Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834):354–359, 2020. 1

[54] Run Wang, Felix Juefei-Xu, Qing Guo, Yihao Huang, Lei Ma, Yang Liu, and Lina Wang. Deeptag: Robust image tagging for deepfake provenance. *arXiv preprint arXiv:2009.09869*, 2020. 3

[55] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[56] Ryan Webster, Julien Rabin, Loic Simon, and Frédéric Jurie. Detecting overfitting of deep generative networks via latent

recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[57] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *IEEE/CVF International Conference on Computer Vision*, 2021. 1

[58] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *European Conference on Computer Vision*, 2018. 2

[59] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018. 3

[60] Mingyang Xie, Manav Kulshrestha, Shaojie Wang, Jinghan Yang, Ayan Chakrabarti, Ning Zhang, and Yevgeniy Vorobeychik. Proves: Establishing image provenance using semantic signatures. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 3

[61] Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. Ganobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE Transactions on Information Forensics and Security*, 14(9):2358–2371, 2019. 2, 3

[62] Yangyang Xu, Yong Du, Wenpeng Xiao, Xuemiao Xu, and Shengfeng He. From continuity to editability: Inverting gans with consecutive images. In *IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2, 4

[63] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 5

[64] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European Conference on Computer Vision*, pages 85–101. Springer, 2022. 3

[65] Cheng Yu and Wenmin Wang. Diverse similarity encoder for deep gan inversion. *arXiv preprint arXiv:2108.10201*, 2021. 8

[66] Xu Zhang, Zhaohui H Sun, Svebor Karaman, and Shih-Fu Chang. Discovering image manipulation history by pairwise relation and forensics tools. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1012–1023, 2020. 3

[67] Shuchang Zhou, Taihong Xiao, Yi Yang, Dieqiao Feng, Qinyao He, and Weiran He. Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv preprint arXiv:1705.04932*, 2017. 2