# IdProv: Identity-Based Provenance for Synthetic Image Generation (Student Abstract)

**Harshil Bhatia**[1], **Jaisidh Singh** [1], **Gaurav Sangwan** [1], **Aparna Bharati** [2], **Richa Singh** [1], **Mayank Vatsa**[1]

[1] IIT Jodhpur, India
[2] Lehigh University, USA
{bhatia.2, singh.118, sangwan.2, richa, mvatsa}@iitj.ac.in, apb220@lehigh.edu

## Abstract

Recent advancements in Generative Adversarial Networks (GANs) have made it possible to obtain high-quality face images of synthetic identities. These networks see large amounts of real faces in order to learn to generate realistic looking synthetic images. However, the concept of a *synthetic identity* for these images is not very well-defined. In this work, we verify identity leakage from the training set containing real images into the latent space and propose a novel method, IdProv, that uses image composition to trace the source of identity signals in the generated image.

## Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al. 2020) consume a large number of face images in order to learn a latent representation for each image and generate synthetic images. Since the latent space is learnt by observing real faces during its training, it is possible that some identity information from the training set can be "leaked" into this learned space from which vectors are sampled and decoded into faces. Identity leakage, if present, can pose a privacy threat and has been preliminarily explored in existing literature. Tinsley et. al. (Tinsley, Czajka, and Flynn 2021), present a study utilizing face matching for real training image pairs and pairs containing one real and one synthetic face image. The distribution of match scores establishes that identities of real faces used for training, leak into the generated face images. To detect identity leakage using existing matching approaches, we require some detectable identity component of the real face signal in the synthetic face. To investigate these concerns further, we examine the provenance (Moreira et al. 2022) of this leakage for synthetic face images with composite images. Generally, image provenance for complex manipulated images *i.e., query*, aims to analyze the evolution of the query image content from its source images (ones that donated content). This work considers the training images showing leakage to be identity donors for synthetic images.

## Methodology

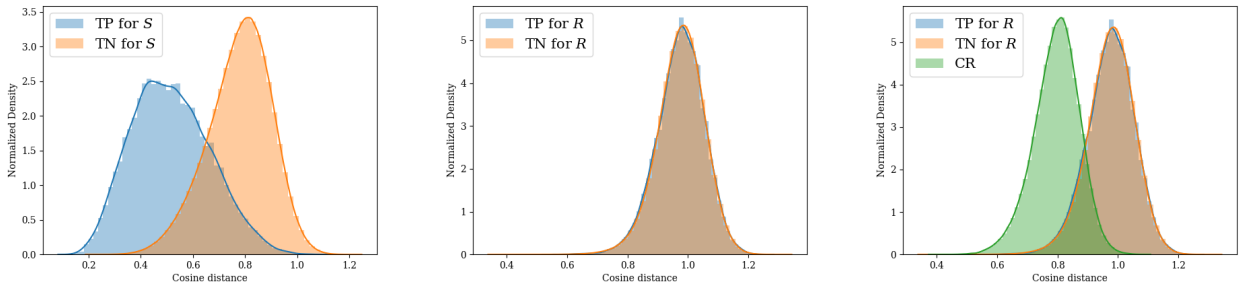
Our method, IdProv, simulates identity composition scenarios by systematically generating synthetic face images. We describe the creation of our synthetic dataset, and the experiments conducted on the same.




Figure 1: Top 3 real images with the most identity leakage into synthetic composites having 2 synthetic parents.

**Data Generation for Experiments** Composite face images are generated using the $\mathcal{W}$ latent space of StyleGAN2 (Karras et al. 2020). From a set of source images, we randomly choose $k$ images (parents), and use their latent vectors $l_1, ...l_k \in \mathcal{W}$ to obtain a composite latent vector given by $l_c = 1/k \cdot \sum_{i=1}^{k} l_i$. Two types of source images are considered - (i) *synthetic*, $\mathcal{S}$ sampled from $\mathcal{W}$, and (ii) *real*, $\mathcal{R}$, selected from Flickr Faces High Quality (FFHQ) dataset. For synthetic sources, $\mathcal{L}_s = \{l^s \in \mathcal{W}\}$ where the latent vectors $l^s$ are randomly sampled. Whereas, Pivotal Tuning Inversion (Roich et al. 2022) is used to obtain the set of latent vectors, $\mathcal{L}_r = \{l^r \in \mathcal{W}\}$ from the real sources. 10K images are generated using both $\mathcal{S}$ and $\mathcal{R}$ for each $k = \{2, ..., 8\}$.

**Retrieving Closest Real Faces** Upon description using a face matcher, the synthetic composites are subsequently matched to the set of synthetic images $\mathcal{S}$. From our experiments, this yields a high number of false positives or matches of composites with non-parents (see fig. 2a). For each image in the set of false positives, the cosine distance is computed to each parent of the composite in the set $\{p_1, ..., p_k\}$ in the face recognition embedding space. The image is subsequently associated with the closest parent in the set. This leads to a set of false positives $\mathcal{F}(p_i) = \{f_1, ..., f_n\}$ for each parent. We constrain the size of $|\mathcal{F}| = 5$, by considering the images with the lowest cosine distance from $p_i$. This diversifies the content that can match with the sources. Finally, the parent $p_i$ and the false positives, $\mathcal{F}(p_i)$ are matched with the set of real images, $\mathcal{R}$. This matching




(a) Synthetic Set, $\mathcal{S}$ (b) Real Set, $\mathcal{R}$ (c) Our method

Figure 2: Cosine distance distributions between the true positives (TP) and negatives (TN) from the composite, made from $\mathcal{S}$ in (a) and from $\mathcal{R}$ in (b). Also, (c) shows the distribution of the cosine distance of the closest real (CR) faces retrieved for a synthetic composite, with the distributions from (b) for comparison. These results are for $k = 2$ source faces.

results in cosine distances (for each $p_i$ and $\mathcal{F}(p_i)$), from the real faces. We then select the 20 closest real faces to each $p_i$ and $f_i \in \mathcal{F}(p_i)$, namely $I_{p_i}$ and $I_{f_i}$, $f_i \in \mathcal{F}(p_i)$. Finally, we find the closest real (CR) faces for the synthetic composite, $I_{cr}$, as intersection of these sets of indices using the following formula.

$$I_{cr} = \bigcap_{i \in [1,...k]} \left( I_{p_i} \cap \left( \bigcap_{f_i \in \mathcal{F}(p_i)} I_{f_i} \right) \right) \quad (1)$$

## Results

Embeddings in the face recognition space are extracted using ArcFace (Deng et al. 2019). In our experiments, the threshold used for obtaining successful matches is 0.68. The distribution of cosine distance between each synthetic composite to the embeddings of $\mathcal{S}$ are shown in Fig. 2a. There is a significant overlap between the distributions of the cosine distance of the true positives (parents) and true negatives (non-parent faces), implying large number of false positives. However, Fig. 2b shows that the match distributions of real composites to their parents and the non-parent real images are nearly identical. Observing both these distributions, we infer that for the large overlap in Fig. 2a to exist, there must be some common identity features being shared among the synthetic faces in $\mathcal{S}$. This inference aligns with the presence of large number of false positives. Thus, we argue that the origin of these common identity features is the training phase of the GAN, where it sees a large number of real identities, which "leak" into the learned space. Further, upon averaging, the leaked identities common across the $k$ parents, aggregate into the composite face. This is not the case when we use real images, which have distinct identities. For real composites, averaging latent representations simply averages their identities, with no semantic aggregation of identity features. To evaluate if this framework can highlight specific identities, for each synthetic composite, we retrieve leaking sources. The cosine distances for closest real images retrieved for the full test set of synthetic composites, is depicted in Fig. 2c. The distributions show that our method retrieves real images which are closer to a synthetic composite than to their real composite. This verifies inherent identity leakage from the FFHQ training images into the $\mathcal{W}$ space of StyleGAN2. Similar behaviour for higher values of $k$ is shown by the distribution means in Table. 1.

| | Number of Parents, $k$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| TP | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| TN | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| CR | 0.79 | 0.78 | 0.77 | 0.76 | 0.75 | 0.74 | 0.73 |

Table 1: We report the means for TP, TN for $\mathcal{R}$ and CR distributions for all $k \in [2, ..8]$. Fig. 2c shows the distribution for $k = 2$.

## Conclusion

Identity leakage in GANs can have an adversarial effect on privacy of identities in the training set. This study successfully demonstrates the threat that StyleGAN2 poses to the privacy of identities and proposes a novel method to retrieve real identities present in a synthetic composite face image.

## References

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Moreira, D.; Theisen, W.; Scheirer, W.; Bharati, A.; Brogan, J.; and Rocha, A. 2022. Image Provenance Analysis. In *Multimedia Forensics*, 389–432. Springer, Singapore.

Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics*.

Tinsley, P.; Czajka, A.; and Flynn, P. 2021. This face does not exist... but it might be yours! identity leakage in generative models. In *IEEE Winter Conf. Appl. Comput. Vis.*