

Beyond Face Matching: A Facial Traits based Privacy Score for Synthetic Face Datasets

Roberto Leyva^{1,2}
marcial.leyva-fernandez@warwick.ac.uk

Praveen Selvaraj²
pselvaraj@turing.ac.uk

Andrew Elliott^{2,3}
andrew.elliott@glasgow.ac.uk

Gregory Epiphaniou¹
gregory.epiphaniou@warwick.ac.uk

Carsten Maple^{1,2}
cm@warwick.ac.uk

¹ WMG
University of Warwick
Coventry, UK

² The Alan Turing Institute
London, UK

³ Department of Mathematics and
Statistics
University of Glasgow
Scotland, UK

Abstract

Synthetic data is increasingly crucial for training machine learning models, especially in fields where real data is scarce or sensitive. This is particularly true for facial data, given growing privacy concerns and the need for rapid development in face recognition systems. However, synthetic facial data often derives from existing datasets, raising privacy issues as synthesizers may inadvertently expose real training data. Our method is motivated to address this important aspect. In this paper, we develop a model that provides a probabilistic score indicating how likely a synthetic face incorporates elements from the training dataset. We focus on facial traits — eyes, nose, mouth and their fusion — modeling training set membership as a probability. This approach allows us to assess whether a synthesizer captures training set characteristics too closely. In addition to generating whole synthetic faces, we explore the generative models' latent space by creating variations in specific facial traits, to more thoroughly assess whether the synthesizer overly relies on facial features from the training set. This method provides a deeper understanding of the synthesizer's tendency to reproduce learned characteristics. Our findings demonstrate that we can establish boundaries for determining full or partial presence of a sample in the training set, depending on specific facial traits. We also found that combining multiple facial traits in our model improves accuracy. The resulting privacy score indicates how much a synthetic dataset contains identifiable features from its training data, effectively measuring its level of compromise. In summary, our results show that by analyzing individual facial features, we can assess how well a synthetic face dataset preserves privacy, relative to the real dataset used to train its synthesizer.

1 Introduction

The widespread use of face-based applications has raised significant privacy concerns regarding personal data usage. In response, there is a growing trend towards synthetic data generation and profiling [10]. However, state-of-the-art synthesizers can often produce samples with minimal distortion of the training data [19, 20], potentially enabling identification of individuals in the training set – a major privacy setback. Simply removing synthetic samples similar to real individuals in the training set is not a comprehensive solution [10]. While this problem has been studied, there is a lack of methods for detecting the susceptibility of image synthesizers in leaking private information. Existing methods primarily report cases where the training data is almost exactly replicated. In this paper, we argue that replication of identifiable facial features, such as individual facial traits (eyes, nose, mouth), should be considered as potential privacy risks when assessing synthetic facial data. Existing technologies like face swapping allow for generating partially synthetic face images by copying specific traits or regions from one face to another, further emphasizing the need for a trait-based or region-based technique for matching faces. We address this issue in the context of both full and partial face synthesis. In summary, we list our main contributions:

1. We present a fusion strategy to determine the probability of a face coming directly from the training data.
2. We present a model to detect separately which face traits might be synthetic, either partially or fully, thus providing a level of accountability.
3. We evaluate the contribution of each trait to the synthesis detection, and therefore its descriptiveness in the task.

This paper is organized as follows: Section 2 reviews related work, Section 3 presents our approach, Section 4 provides experimental results and finally Section 5 concludes this paper.

2 Related Work

In recent years, we observe a trend in new face synthesis techniques in which enhancing privacy is the primary technology component [27]. Several works address the problem of privacy by - not releasing the training set, hiding/concealing training samples or matching synthetic data with all samples from the training set. One approach to address privacy preservation in synthetic data is **Differential Privacy (DP)** [8]. The key element of **DP** is that it provides a mathematical guarantee that the inclusion or exclusion of a single individual’s data does not “significantly” affect the outcome of any analysis, making it difficult to infer whether a specific person’s data was part of the training set. For example, Zhang *et al.* [20] design a GAN that incorporates **DP** by adding a small amount of noise during gradient propagation. The authors propose a strategy to prevent poor image quality resulting from this perturbation. Whereas, Qingrong *et al.* [5] propose a data augmentation strategy whereby the model is trained in three stages: first with private data, then with public data and in the final stage with public generated data. In the first stage, random noise is injected to the model’s gradients to implement **DP**. Hanyu *et al.* [14] propose a **DP** model by targeting the latent space in order to preserve privacy. The method uses an encoder to extract a feature

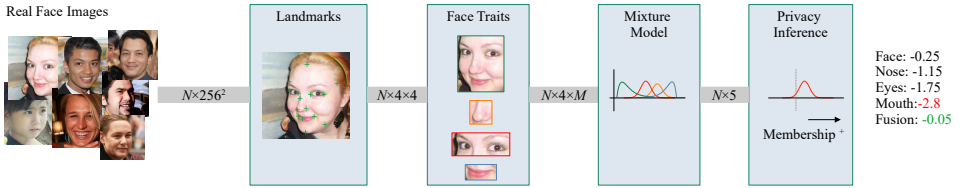


Figure 1: Our strategy is to detect the facial landmarks (Section 3.1) we crop the face, eyes, nose and mouth and train a feature extraction model for each trait (Section 3.2). We model the extracted features using a probabilistic approach to determine privacy score (Section 3.3)

vector from an image, which then serves as input latent code for a generator that reconstructs the image. The method produces significantly better visual results than common strategies (e.g. blur, mosaics, pixel manipulation). The main challenge of **DP** methods is to achieve high image quality compared to non-**DP** counterparts.

Another family of approaches to protect the training set leverages federated learning, where the model learns without directly having access to the data [66]. However, even within this framework, generated data might violate the privacy of samples in the training set [20] and care must be taken to address the issue of malicious federated contributors. Recent approaches provide techniques to mitigate individual malicious contributors attempting to recover the original data by ensuring that aggregations are protected [29]. An alternative set of approaches attempts to protect identities prior to training by using de-identification techniques [9], which make it difficult to generate synthetic samples highly similar to the training set [23]. Although these efforts show good progress [9], it is worth noting that the synthetic data can have a significant number of facial artifacts [22, 88]. These models rely on noise and distortion for de-identification and are not designed to measure the similarity of synthetic samples with the training set.

As methods are created to generate data in a privacy preserving way, other works attempt to challenge the integrity of the training dataset, primarily using re-identification techniques [83]. These methods involve creating models to identify individuals across multiple samples within the training set [17]. They require extensive data and numerous samples from a target identity to verify membership [89]. Essentially, they address whether a query person matches across different samples, which are assumed to be part of the training set [20]. While useful for identifying individuals, these methods do not effectively determine the extent to which face samples belong to the training set [85]. Additionally, it is generally understood that only real samples, not synthetic ones, are used in re-identification tasks. Varkarakis *et al.* [82] investigated whether synthetic faces are unique compared to a real reference set. They found that parameter tuning greatly affects the similarity measure for synthetic faces. Their study also examined identity uniqueness within the reference set. Our work shares a similar motivation to [82], but differs in key aspects. We focus on the probability distribution of facial traits rather than overall face similarity. Additionally, we do not consider identity in our analysis, instead concentrating solely on facial characteristics.

3 Proposed Method

Our proposed strategy, as illustrated in Figure 1, begins with the detection and demarcation of facial traits of interest using an off-the-shelf landmark detector, carefully preserving facial geometry. We then train feature extractors for both individual traits taken separately and all the traits taken together. These extracted features are subsequently compared against the

training dataset using a distance metric. To transform these computed distances into probabilities, we apply a **Bayesian Mixture Model (BMM)**, generating a **Probability Distribution Function (PDF)**. The mixture model’s posterior score serves as the key determining factor of whether a query trait belongs to the training dataset. A low score indicates a private sample, suggesting that no close match exists in the training dataset. This approach produces consistently low scores for data not present in the training set, regardless of whether the query sample is real or synthetic. The rationale being that there is a significant shift between the real and synthetic **PDFs** with respect to the chosen distance metric. Consequently, the gap between these **PDFs** is used to establish the privacy score.

3.1 Landmarks

To identify the location of each facial trait, we use the **PFLD (Practical Facial Landmark Detector)** [14] (see Figure 1). To preserve facial geometry, we establish a set of proportional metrics based on face and eye detection, which guide the cropping of each facial trait. We crop the face using the maximum and minimum coordinates of detected landmarks.

For the eyes, we extend the cropped region 50% horizontally and 25% vertically from the inter-eye distance. The nose region is defined by extending 90% above and 25% below the nose, using the inter-eye distance as a reference. The mouth region extends 30% vertically and 25% horizontally from the lip corners. These proportional cropping methods ensure consistent trait extraction across different face sizes and orientations. We apply these cropping techniques to generate a dataset of facial traits. Figure 2 illustrates examples of these extracted facial traits.

3.2 Feature Extraction

Our feature extraction process for facial traits consists of two stages. In the first stage, we train separate **Vision Transformer (ViT)**-32 models [15] for each facial trait using the labels (distinct identities) from our training set. In the second stage, we use these trained models to convert all images into feature vectors. We then calculate distances between these vectors, which serve as inputs for training a probabilistic model.

It is worth noting that some architectures can be used directly without labels and produce very descriptive features, e.g., autoencoders [16]. However, following a similar approach to people re-identification, using identity directly can generate more descriptive features [5, 7]. For example, such features may be more informative than reconstruction-based features, as the target function can consider face variations, e.g., size, shape, color etc., from the labeling process. We’re also influenced by recent work on general-purpose face recognition pipelines that generate highly descriptive features [8, 9]. These methods involve training a feature extractor and optimizing its output for a specific task. For instance, Carlini *et al.* [9] use existing feature models (e.g. ResNet and ViT) to find the most similar identity generated from text prompts. Similarly, Deng *et al.* [8] use a specialized loss function to distinguish identities. Formally, our feature extraction model solves the following problem:

$$\operatorname{argmax}_c f(x, y_c), \quad (1)$$

where f is the model mapping the trait x to the label y_c . The RGB image $x \in \mathbb{R}^{n_x \times n_y \times 3}$ has spatial dimensions $n_x \times n_y$. The label $y_c \in [0, 1^c]$ is a one-hot encoded vector, where c is the number of distinct identities (classes).

We train the model from scratch, consistently cropping and scaling the facial trait inputs. Once the model converges, following [15], we use the weights for feature extraction:

$$z = \prod_{l=1}^L \sigma(\mathbf{W}^{l,r} x^r) \quad (2)$$

here, L represents the number of ViTs in the ensemble, with one ViT dedicated to each trait.. x^r is the image trait patch r , $\mathbf{W}^{l,r}$ the image encoder from the l -th ViT in the ensemble and σ the GeLU activation function. This process produces an M -dimensional feature vector, $z \in \mathbb{R}^M$.

To determine similarity between image samples, we construct a symmetric distance matrix $D \in \mathbb{R}^{N \times N}$, where N is the number of samples, using the full set of features $z \in \mathbb{R}^{N \times M}$. We set the diagonal with large values to avoid using distances between identical samples. The embeddings' distance, using the p -norm, is calculated as:

$$D_{i,j} = \begin{cases} 10000 & \text{if } i = j \\ (|z_i - z_j|)^{1/p} & \text{if } i \neq j \end{cases} \quad (3a)$$

To prepare the input for our probabilistic model, we introduce a design matrix \hat{X} . This design matrix represents, for each sample, the average of the top- k minimum distances to any other sample in the dataset, excluding itself. We compute this for each facial trait t (face, eyes, nose, and mouth) separately. For each trait t , we construct a design matrix \hat{X}_t as follows:

$$\hat{X}_t = \frac{1}{k} \sum_{j=1}^k D_{i,j}^t \quad (3b)$$

where \hat{X}_t represents the input feature for the probabilistic model, D^t the distance matrix for trait t and k is the number of smallest distances used to calculate the average.

3.3 Probabilistic Model

Our pipeline uses two distinct modules to model facial traits: one for individual traits and another for their combination. The probabilistic model in each module is trained using the embedding distances described earlier, first for each trait individually and then for their fusion. We utilize a BMM for this training process, leveraging Bayesian inference on the features extracted from facial traits. In this framework, we consider the observation matrix $\hat{X} = \{\hat{x}^1, \hat{x}^2, \dots, \hat{x}^N\}$ as a collection of N independent and identically distributed (i.i.d) samples drawn from an observable distribution. Each \hat{x}^i represents a sample vector corresponding to a specific facial trait. We define our mixture model as follows:

$$p(\hat{X}|\theta) = \sum_{m=1}^M \pi_m \mathcal{N}(\hat{X}|\mu_m, \Sigma_m) \quad (4)$$

where $p(\hat{X}|\theta)$ is the probability of observing \hat{X} given the model parameters θ ; M is the total number of components in the mixture model, π_m is the mixing coefficient (or weight) for the m -th component and $\mathcal{N}(\hat{X}|\mu_m, \Sigma_m)$ represents a multivariate Gaussian distribution with mean μ_m and covariance matrix Σ_m . The model parameters $\theta = \{\pi, \mu, \Sigma\}$ consist of the set of mixing coefficients $\pi = \pi_1, \pi_2, \dots, \pi_M$, the set of mean vectors $\mu = \mu_1, \mu_2, \dots, \mu_M$ and the

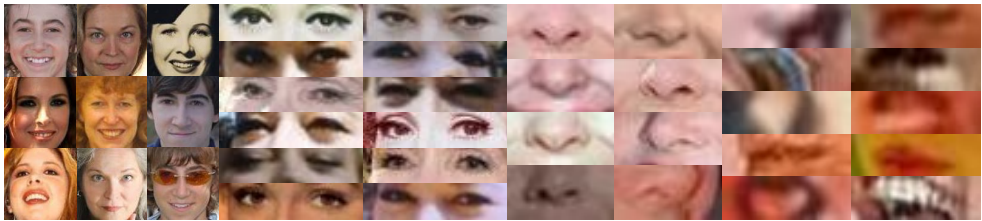


Figure 2: CELEBA samples with extracted facial traits from left to right: Face, Eyes, Nose, Mouth.

set of covariance matrices $\Sigma = \Sigma_1, \Sigma_2, \dots, \Sigma_M$. We then select the number of components q , that maximizes the posterior probability:

$$q = \underset{q}{\operatorname{argmax}} p(\hat{X} | \pi_q, \mu_q, \Sigma_q) \quad (5)$$

Each component represents a multivariate Gaussian distribution used in our mixture model. For each facial trait, we found that using just one or two components typically produces satisfactory results. We use the model’s output scores to infer if a sample was in the training set. Higher scores indicate a greater likelihood that the sample was part of the training data. This aligns with our method’s goal of maximizing the posterior probability to best represent the real data distribution. This is shown below:

$$p(\hat{x}_t | \pi, \mu, \sigma) > \gamma. \quad (6)$$

To determine if a query sample is private for trait t , we compare its privacy score \hat{x}_t to a threshold γ . If $\hat{x}_t > \gamma$, we consider the sample private. To automatically determine γ we run a linear greedy search over the validation set. We create a range of 50 evenly spaced values for γ , spanning from the minimum to the maximum posterior in the set. Starting from the lower bound, we gradually increase γ until we reach the maximum value. It’s worth noting that both the minimum and the maximum posteriors give an accuracy of approximately 0.5. Once we have the optimal value, it’s then used to evaluate the test set.

4 Experiments

Our experiments utilize 2 face image datasets: [Flick Faces High Quality \(FFHQ\)](https://github.com/NVlabs/ffhq-dataset)¹ and [Large-scale CelebFaces Attributes \(CELEBA\)](https://github.com/tkarras/progressive_growing_of_gans)² [13, 25]. These datasets contain 70K and 30K real face images respectively, with multiple samples per identity. Figure 2 displays samples from CELEBA, showcasing both the original images and their corresponding cropped facial traits. For generating synthetic faces, we use StyleGAN2 [24] and SGAN-XL [30] models, each trained from scratch on the 2 datasets. To explore the latent space of these synthesizers more thoroughly, we use a latent code optimization technique [27]. This method allows us to augment the synthetic dataset by generating several variations for each synthetic face by applying control inputs to specific regions of interest, namely the eyes, the nose and the mouth. Figure 3 illustrates this process, showing two subjects and several samples produced through this process of controlled variation generation. By exploring the latent space of the generated images more comprehensively, we enhance the robustness of our evaluation.

¹<https://github.com/NVlabs/ffhq-dataset>

²https://github.com/tkarras/progressive_growing_of_gans



Figure 3: DragGAN [14] introduces variations in synthetic faces, focusing on the eyes, nose and mouth. This helps to more thoroughly assess if the latent space of the synthesizer has captured traits from the reference set.

4.1 Best Backbone Feature Selection

We’ve conducted ablation experiments to identify the best feature extraction model, as detailed in section 3. To this end, we use the **CELEBA** dataset and evaluated the models on 3 criteria: the **mAp** score across all three individual traits and the whole face, the number of parameters in the model and the processing times for both training and validation.

As Table 1 shows, the **ViT-32b** variant achieves competitive **mAp** scores while utilizing significantly fewer parameters and requiring less processing time. This makes it an optimal choice for our backbone model. Our experiments also revealed that the **ViT-32b** model converges around 15 epochs, at which point the loss stabilizes. This makes it faster to converge and more stable than the other feature extractors.

4.2 Training Separate Traits Models

We train separate models for each trait, maintaining the original face geometry. Instead of using the original **ViT**’s input dimensions, we crop samples (see section 3.1), resize them and train from scratch. This maintains trait-specific details while standardizing inputs.

Figure 2 shows a few samples used for training, each focusing on a single facial trait. We use 80% of the data for training and reserve 20% for testing. To ensure data quality, we exclude identities with fewer than 9 samples. This prevents scenarios where we have samples for training but not for testing, which would limit the model’s learning.

Model	RunTime (seconds) (↓)	Parameters (↓)	mAp (↑)
RS-101 [14]	2498	61M	0.2348
RS-34 [14]	899	26M	0.1936
DN-101 [14]	9975	16M	0.2105
VGG-19 [14]	6890	177M	0.1642
ViT-16b [14]	2379	92M	0.2153
ViT-16L	6152	312M	0.2195
ViT-14	30980	642M	0.2612
ViT-32b (chosen)	941	94M	0.2454
ViT-32L	3897	315M	0.1978

Table 1: Backbone feature extraction model performance, in terms of **mAp** scores across all four individual face traits.

The main purpose of our model is to learn a feature representation of facial traits that allows us to determine whether a query sample contains elements from the training dataset, either in whole or in part. We train a separate **ViT** model for each trait, with each model learning to distinguish approximately 9,500 identities. The model is trained for 25 epochs using a cross-entropy loss function and stochastic gradient descent optimizer. We employ sample augmentation techniques and implement learning rate plateau detection across each

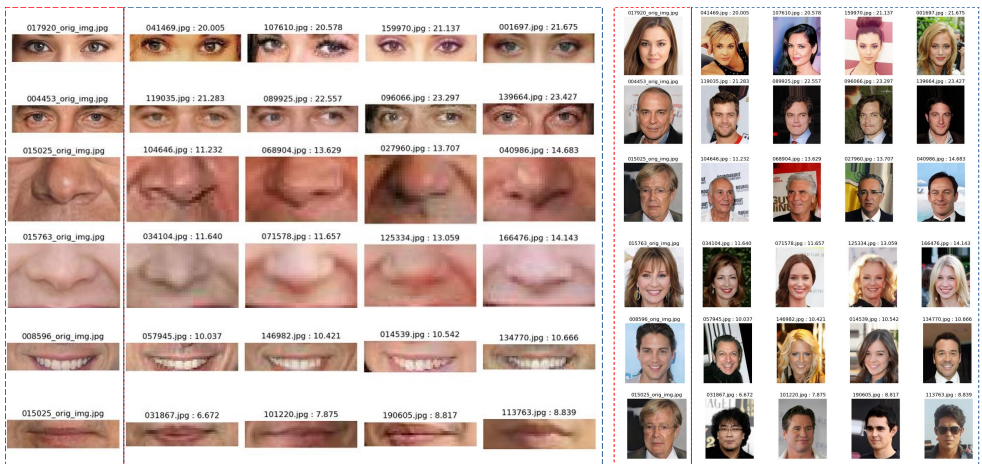


Figure 4: Red/Blue dotted lines signify Synthetic / Real images. Top-k matches for synthetic faces across the eyes, nose and mouth traits. Panel on the left shows the matched individual traits whereas panel on the right shows their corresponding whole faces. To select the top matches, we use ℓ_2 -distances in embedding space.

of the seven iterations. To expedite training, we parallelized the process across four GTX-4090 GPUs.

Table 2 presents the best performance in terms of **mean Average precision (mAp)** for each of our trait models. For context, random classification would yield an accuracy of approximately $\frac{1}{9500} \approx 10^{-4}$. Our results show that identities are most easily classified using the whole face. The individual traits, in descriptive order are: the eyes, the mouth and the nose. This aligns with previous research [L8] that identified the eyes as the most descriptive facial feature, followed by the mouth. For a given synthetic face, we can now use the embeddings (Eq. 2) from our trained models to retrieve top-k matches against the training set.

Figure 4 visualizes these matches for the eyes, nose and mouth models. It compares synthetic features against the top-k matched features from the training set using ℓ_2 -distances in embedding space. The matches for the different features are run on the same set of people.

Trait	RunTime (seconds)	Parameters	Split	Batch	mAp (\uparrow)
Face	975	94M	0.8	512	0.3941
Eyes	834	94M	0.8	4096	0.2278
Mouth	789	94M	0.8	2048	0.1674
Nose	845	94M	0.8	8192	0.1307

Table 2: Facial traits model performances on identity classification. Using the whole face gives the best performance followed by the models that use the eyes, the mouth and the nose separately.

For the eyes, we note that the synthetic ones closely resemble their real counterparts. For the nose, we find that the best matched noses do not come from the same subject as the best matched eyes. This suggests that our individual facial trait analysis can detect similarities in specific traits independently of the whole face. For the mouth, it is important to highlight that the mouth pose, whether open or closed, has a significant impact on the matches. As Figure 4 shows, smiling mouths are matched with other smiling mouths. The mouth exhibits more variation in pose compared to noses and eyes, which may explain the mouth model’s lower performance in identity classification. We hypothesize that the mouth requires more samples with variations for effective modeling compared to the mostly static nose.

















Synthetic									
Eyes	-36.19	-31.94	-23.41	-27.58	-22.14	-35.23	-28.70	-17.94	
Nose	-18.15	-19.66	-15.29	-15.25	-19.96	-23.09	-16.71	-19.49	
Mouth	-23.54	-16.25	-21.26	-15.07	-27.29	-32.62	-17.46	-25.30	
Face	-15.50	-10.94	-15.11	-14.13	-17.97	-19.94	-21.19	-19.96	
Fusion	-31.00	-21.88	-30.22	-28.26	-35.94	-39.87	-42.37	-39.92	
Real									
Eyes	-1.17	-3.10	-4.98	-1.27	-1.21	-3.95	-2.50	-11.74	
Nose	-2.41	-3.54	-4.78	-1.82	-1.69	-5.21	-3.69	-12.14	
Mouth	-2.22	-4.46	-5.21	-1.96	-1.34	-4.19	-3.22	-13.24	
Face	-0.69	-2.66	-3.87	-1.06	-0.39	-3.88	-2.47	-11.09	
Fusion	-4.24	-5.15	-7.22	-3.45	-3.16	-7.24	-4.96	-26.41	

Figure 5: Posterior scores for synthetic (First row) and real (Second row) images from the individual and fused models. Larger value implies higher likelihood of the trait being from the training set.

4.3 Probabilistic Models

Our final set of experiments focuses on inference models that produces probabilities by using both individual traits separately, as well as their fusion. We train separate probabilistic models for each trait to generate individual scores, as well as a fusion model that combines all traits. Figure 5 shows the posteriors produced by these models.

Table 3 presents the results for our models when tested against synthesizers trained on CELEBA and FFHQ separately. The findings indicate that using the whole face is a reliable method for determining whether a face belongs to the training set. However, the best scores are achieved if we consider the other facial traits in conjunction with the whole face. Since the posterior is the product of composed probabilities, we can infer that incorporating more traits will consistently enhance our privacy inference model’s performance. Fig 5 shows the scores produced by the probability models for a few samples.





Trait	CELEBA		FFHQ	
	SGAN-XL 	StyleGAN2 	SGAN-XL 	StyleGAN2 
Face	<u>0.9317</u>	<u>0.9417</u>	<u>0.8109</u>	0.8636
Eyes	0.8422	0.9085	0.7374	0.7698
Mouth	0.8017	0.8201	0.6022	0.6574
Nose	0.7741	0.8674	0.7274	0.7485
Fusion	0.9585	0.9534	0.8210	<u>0.8403</u>

Table 3: Comparison of mAp scores for the probabilistic models, calculated on images from SGAN-XL and StyleGAN2 models trained on either the CELEBA or FFHQ datasets. Fusion models generally perform the best, followed by face-only models. Highest scores are in bold, second-highest underlined.

The fusion model produces lower scores for synthetic samples when compared with the other models and higher scores for real samples. This is an important outcome for accountability. While the fusion model provides the best overall performance, the individual trait models allow us to pinpoint which specific facial feature caused a query sample image to be flagged for low privacy – a capability the fusion model lacks. The results also show that the method has very competitive accuracy in detecting synthetic facial traits. We also compared our probabilistic model against two prominent approaches: Carlini *et al.* [2] and ArcFace [6]. For Carlini *et al.*’s method, we used the CLIP [23] architecture, specifically it’s image-only variant. Following the authors’ recommendations, we trained ResNet-50 and ViT-32 models from scratch for each trait. This experimental configuration ensures a fair comparison against

our backbone. We use the cosine similarity defined in [10] to evaluate the feature extractor. It’s worth noting that all methods used the same greedy approach to determine the optimal threshold. For the ArcFace [11] comparison, we utilized the weight pipeline defined in [24], which uses CNN architectures with 32 and 64 filters, generating a 512-dimensional feature embedding. We observed no notable differences in accuracy between these versions, though they demanded higher computational resources. We adapted the radial loss function from ArcFace [11], using the angular similarity distance metric to evaluate the feature extractors. For both the compared methods [10, 11], we used the embedded space distance to determine the privacy score at the trait level. Since our focus is not on identity determination, we simply threshold the ℓ_2 -distances in feature space, as suggested by Carlini *et al.* [10], for both the methods. It’s important to note that neither CLIP nor ArcFace have fusion embedding capacities. Therefore, our comparison is limited to individual facial traits and doesn’t involve the fused model.

Method	Trait			
	Eyes	Nose	Mouth	Face
CLIP/ViT-32 [10]	0.8112	0.7422	<u>0.8141</u>	0.9104
CLIP/RS-50 [10]	0.7969	0.7548	0.7922	0.8869
CNN/ArcFace/32 [11]	0.8317	0.7756	0.8121	0.9256
CNN/ArcFace/64 [11]	<u>0.8396</u>	0.7612	0.8265	<u>0.9304</u>
ViT-32/BMM (ours)	0.8422	<u>0.7741</u>	0.8017	0.9317

Table 4: Comparison of mAp compared models calculated on images generated from SGAN-XL models trained on CELEBA. Bold marks the highest score and underline marks the second highest.

Table 4 shows that our method achieves the best performance for face and eyes traits. Our approach has the unique ability to fuse information and analyze all facial traits simultaneously, while still achieving outstanding performance on individual traits. The other methods can only analyze traits in isolation.

5 Conclusions

Our paper introduces a framework that assigns a score to synthetic faces, indicating their similarity to samples from the dataset used to train the synthesizer. This assessment applies to both whole faces and individual facial traits. Our concern stems from the potential for privacy violations when a synthesizer replicates not just entire faces, but also distinctive traits like eyes, nose and mouth. We’ve developed models for these individual traits as well as a fusion model that combines analysis of the whole face with isolated individual traits. Our work shows that the fusion model excels in detection accuracy and when used in conjunction with the individual traits models, enhances accountability by allowing us to identify which specific traits contribute to the overall score and to what degree. Our results show that the privacy score’s accuracy varies across facial traits, with the eyes showing the highest precision. This is in line with previous work in the field. We believe that our framework can be a valuable tool in evaluating synthetic datasets. It offers a quantitative measure of privacy concerning replicated traits, which can complement other privacy techniques such as DP. Moreover, by comparing scores across multiple synthetic datasets produced by the same synthesizer, we can gauge the level of privacy that can be expected from that particular synthesizer relative to the training set.

Acknowledgements

This work was supported by the Bill & Melinda Gates Foundation [INV-001309].

This work was supported in part by JADE: Joint Academic Data science Endeavour - 2 under the EPSRC Grant EP/T022205/1, & The Alan Turing Institute under EPSRC grant EP/N510129/1.

References

- [1] Fadi Boutros, Vitomir Struc, Julian Fierrez, and Naser Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, 135:104688, 2023. ISSN 0262-8856. doi: <https://doi.org/10.1016/j.imavis.2023.104688>. URL <https://www.sciencedirect.com/science/article/pii/S0262885623000628>.
- [2] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium, SEC '23, USA, 2023*. USENIX Association. ISBN 978-1-939133-37-3.
- [3] Mahawaga Arachchige Pathum Chamikara, Peter Bertok, Ibrahim Khalil, Dongxi Liu, and Seyit Camtepe. Privacy preserving face recognition utilizing differential privacy. *Computers & Security*, 97:101951, 2020.
- [4] Cunjian Chen. Pytorch face landmark: A fast and accurate facial landmark detector, 2021. URL https://github.com/cunjian/pytorch_face_landmark.
- [5] Qingrong Chen, Chong Xiang, Minhui Xue, Bo Li, Nikita Borisov, Dali Kaarfar, and Haojin Zhu. Differentially private data generative models. *arXiv preprint arXiv:1812.02274*, 2018.
- [6] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. doi: 10.1109/TPAMI.2021.3087709.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [8] Cynthia Dwork. Differential Privacy. In *Automata, Languages and Programming*, pages 1–12. Springer, Berlin, Germany, 2006. ISBN 978-3-540-35908-1. doi: 10.1007/11787006_1.
- [9] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9378–9387, 2019.
- [10] Tobias Hann. Why removing identical matches in synthetic data risks privacy: The swiss cheese problem. <https://mostly.ai/blog/why-removing-identical-matches-in-synthetic-data-risks-privacy-the-2024>.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

- [12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [13] 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.
- [14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.
- [15] Jiseob Kim, Kyuhong Shim, Junhan Kim, and Byonghyo Shim. Vision transformer-based feature extraction for generalized zero-shot learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [16] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [17] Bahram Lavi, Ihsan Ullah, Mehdi Fatan, and Anderson Rocha. Survey on reliable deep learning-based person re-identification models: Are we there yet? *arXiv preprint arXiv:2005.00355*, 2020.
- [18] Nova Hadi Lestriandoko, Raymond Veldhuis, and Luuk Spreeuwiers. The contribution of different face parts to deep face recognition. *Frontiers in Computer Science*, 4, August 2022. ISSN 2624-9898. doi: 10.3389/fcomp.2022.958629.
- [19] Roberto Leyva, Gregory Epiphaniou, Carsten Maple, and Victor Sanchez. Unsupervised face synthesis based on human traits. In *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2023. doi: 10.1109/IWBF57495.2023.10157232.
- [20] Roberto Leyva, Victor Sanchez, Gregory Epiphaniou, and Carsten Maple. Data-agnostic face image synthesis detection using bayesian cnns. *Pattern Recognition Letters*, 183:64–70, 2024. ISSN 0167-8655. doi: <https://doi.org/10.1016/j.patrec.2024.04.008>. URL <https://www.sciencedirect.com/science/article/pii/S0167865524001090>.
- [21] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 5–16, 2021.
- [22] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [23] Yuezun Li and Siwei Lyu. De-identification without losing faces. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, pages 83–88, 2019.
- [24] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.
- [25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.
- [26] Carsten Maple. Security and privacy in the internet of things. *Journal of cyber policy*, 2(2): 155–184, 2017.

- [27] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701597. doi: 10.1145/3588432.3591500. URL <https://doi.org/10.1145/3588432.3591500>.
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [29] Amrita Roy Chowdhury, Chuan Guo, Somesh Jha, and Laurens van der Maaten. Eiffel: Ensuring integrity for federated learning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2535–2549, 2022.
- [30] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.
- [31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 09 2014.
- [32] Viktor Varkarakis, Shabab Bazrafkan, Gabriel Costache, and Peter Corcoran. Validating seed data samples for synthetic identities – methodology and uniqueness metrics. *IEEE Access*, 8: 152532–152550, 2020. doi: 10.1109/ACCESS.2020.3016097.
- [33] Wenyu Wei, Wenzhong Yang, Enguang Zuo, Yunyun Qian, and Lihua Wang. Person re-identification based on deep learning—an overview. *Journal of Visual Communication and Image Representation*, 82:103418, 2022.
- [34] Hanyu Xue, Bo Liu, Ming Ding, Tianqing Zhu, Dayong Ye, Li Song, and Wanlei Zhou. Dp-image: Differential privacy for image data in feature space. *arXiv preprint arXiv:2103.07073*, 2021.
- [35] Seongyeop Yang, Byeongkeun Kang, and Yeejin Lee. Sampling agnostic feature representation for long-term person re-identification. *IEEE Transactions on Image Processing*, 31:6412–6423, 2022.
- [36] Xuefei Yin, Yanming Zhu, and Jiankun Hu. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)*, 54(6):1–36, 2021.
- [37] Asmat Zahra, Nazia Perwaiz, Muhammad Shahzad, and Muhammad Moazam Fraz. Person re-identification: A retrospective on domain specific open challenges and future trends. *Pattern Recognition*, 142:109669, 2023. ISSN 0031-3203. doi: <https://doi.org/10.1016/j.patcog.2023.109669>. URL <https://www.sciencedirect.com/science/article/pii/S0031320323003709>.
- [38] Liming Zhai, Qing Guo, Xiaofei Xie, Lei Ma, Yi Estelle Wang, and Yang Liu. A3gan: Attribute-aware anonymization networks for face de-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5303–5313, 2022.
- [39] Guoqing Zhang, Jie Liu, Yuhao Chen, Yuhui Zheng, and Hongwei Zhang. Multi-biometric unified network for cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 32:4555–4566, 2023.

- [40] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model (technical report). *arXiv preprint arXiv:1801.01594*, 2018.
- [41] Shan Zhao, Lefeng Zhang, and Ping Xiong. Priface: a privacy-preserving face recognition framework under untrusted server. *Journal of Ambient Intelligence and Humanized Computing*, 14(3):2967–2979, Mar 2023. ISSN 1868-5145. doi: 10.1007/s12652-023-04543-7. URL <https://doi.org/10.1007/s12652-023-04543-7>.