SynthProv: Interpretable Framework for Profiling Identity Leakage Supplementary Material

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

This supplementary material contains additional results. We present them in the order that they are mentioned in the main paper. First Sec. A provides an overview of our notation. Sec. B provides experimental results for identity invariance in the $\mathcal{W}$ space. Next, we present the quantitative plots which are the results of our ablation studies for identity leakage detection in Sec. C. Additionally, as part of our ablation studies, we present match score distribution for the leaking reals (LRs) for all values of k, the number of parents used to create composites, in Sec. D. Lastly, we discuss the impact of the selected range of k in Sec. E.

A. Notations

Table 1 summarizes the notation that we use in our paper.

Notation	Meaning
$\mathcal{S}$	set of randomly sampled latent vectors
$\mathcal{R}$	latent vectors of training set images
l_c	composite latent vector
l_{p_j}	j^{th} parent latent vector of l_c
l_{n_i}	i^{th} non-parent latent vector of l_c
l^q	latent vector belonging to $\mathcal{Q}-$set
$l^{\mathcal{Q}}$	latent vectors of $\mathcal{Q}-$set
$G(\cdot)$	StyleGAN2 generator
$m(\cdot)$	face-matcher
I_t	face image represented by l_t given by $G(l_t)$
e_t	face-matcher embedding of I_t given by $m(I_t)$
e^q	embedding of latent vector belonging to $\mathcal{Q}-$set
$e^{\mathcal{Q}}$	embeddings of latent vectors of $\mathcal{Q}-$set
$\phi(e_1, e_2)$	match score between two embeddings
v_{ij}	vector from latent l_i to l_j given by $v_{ij} = l_j - l_i$
d^*	identity invariant direction in $\mathcal{W}$ space
$h(l_1, l_2, d^*)$	latent space identity distance between l_1 and l_2
k	number of parents of a composite latent/image
a	number of assistants of each query for provenance

Table 1. Notation used in our work.

B. Identity Invariance in Latent Space

We present experimental results of evaluating the invariance in identity features encoded by the direction d^*. Starting from a fixed latent vector l_0^s, we traverse along d^* in steps of size α. This is given by

$$l_\alpha^s = l_0^s + \alpha d^* \tag{1}$$

where α is a scalar. For a given value of α, we decode l_α^s into a face image I_α^s to qualitatively evaluate the preservation of identity features as α is increased. This is presented in Figure 1 where face images corresponding to the respective latent vectors at these points result in samples with preserved identity features w.r.t to the base images ($\alpha = 0.0$ for the leftmost column). Slight deviations occur only after very large distances in the $\mathcal{W}$ space, which are much larger than the distance values between face representations used in our work. Hence, this direction can be assumed to be globally consistent in encoding identity, and is additionally independent of any particular latent vector.

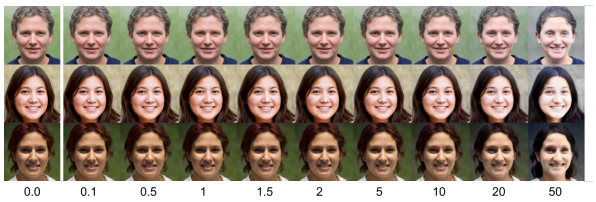

Figure 1. Results of traversing along the identity-invariant direction in the $\mathcal{W}$ space. The lowermost row indicates the sizes of the steps (α from eq. 1) taken along the direction.

C. Match Score Distributions for Identity Leakage

We present the match score distributions for higher values of k as part of our ablation studies in this section. This is done for all face-matchers and datasets, shown in Table 2, Figure 2, and Figure 3. Particularly, Table 2 shows the mean

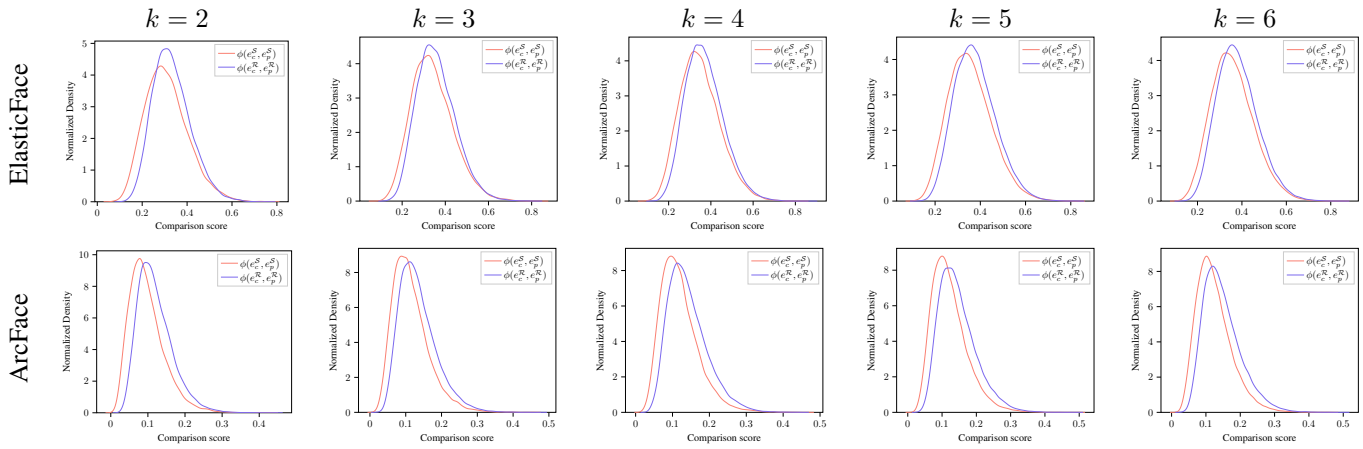

Figure 2. Match score distributions of each type of composite images with their parent images, *i.e.*, distribution of $\phi(e_c^{\mathcal{S}}, e_p^{\mathcal{S}})$ alongside $\phi(e_c^{\mathcal{R}}, e_p^{\mathcal{R}})$, for all values of k and all face-matchers, on the FFHQ dataset.

Table 2. Mean and standard deviation of comparison scores of the composites with different k parents, for all matchers and datasets.

Dataset	k values	ElasticFace		ArcFace	
		$\phi(e_c^s, e_p^s)$	$\phi(e_c^r, e_p^r)$	$\phi(e_c^s, e_p^s)$	$\phi(e_c^r, e_c^r)$
FFHQ	2	0.3093 ± 0.095	0.3318 ± 0.0842	0.0974 ± 0.0472	0.1197 ± 0.0469
	3	0.3404 ± 0.0963	0.3583 ± 0.0891	0.1138 ± 0.0501	0.1341 ± 0.0516
	4	0.3496 ± 0.0952	0.3703 ± 0.0899	0.1188 ± 0.0505	0.1415 ± 0.0529
	5	0.3551 ± 0.0959	0.3776 ± 0.0911	0.122 ± 0.0513	0.1453 ± 0.0538
	6	0.3588 ± 0.0953	0.3809 ± 0.0912	0.1239 ± 0.051	0.1475 ± 0.0543
CelebAHQ	2	0.3519 ± 0.1033	0.4547 ± 0.1049	0.1184 ± 0.0556	0.1827 ± 0.0653
	3	0.3358 ± 0.1014	0.4407 ± 0.1025	0.1098 ± 0.0528	0.173 ± 0.0626
	4	0.3264 ± 0.1017	0.431 ± 0.1011	0.1043 ± 0.052	0.1661 ± 0.0609
	5	0.3204 ± 0.1019	0.4257 ± 0.1009	0.101 ± 0.0514	0.1632 ± 0.0601
	6	0.3166 ± 0.1025	0.4209 ± 0.0995	0.0991 ± 0.0509	0.1597 ± 0.0589

and standard deviation of comparison scores for each k, dataset, and face matcher. Further, Figure 2 and Figure 3 show the distributions of $\phi(e_c^{\mathcal{S}}, e_p^{\mathcal{S}})$ with $\phi(e_c^{\mathcal{R}}, e_p^{\mathcal{R}})$ for the FFHQ and CelebAHQ datasets, respectively. The distributions of $\phi(e_c^{\mathcal{S}}, e_n^{\mathcal{S}})$ with $\phi(e_c^{\mathcal{R}}, e_n^{\mathcal{R}})$ are presented in Figure 4 and Figure 5 in the same order.

D. Distribution of Leaking Reals Match Scores

As the final part of our ablation studies, we show traceability of identity leakage for all values of k. Distributions of match scores of synthetic composite images with their respective leaking reals LRs are shown alongside the match score distributions of real composite images with their real parents. This is shown in Figure 6 and Figure 7 for all face-matching strategies and for StyleGAN2 trained on the FFHQ and CelebAHQ datasets, respectively.

E. Higher Values of k

We consider $k \in \{2, 3, 4, 5, 6\}$ as higher values of k show highly similar faces, with reduced variation of facial attributes. This is depicted in Figure 8, to show the convergence of face images towards one highly averaged face image.

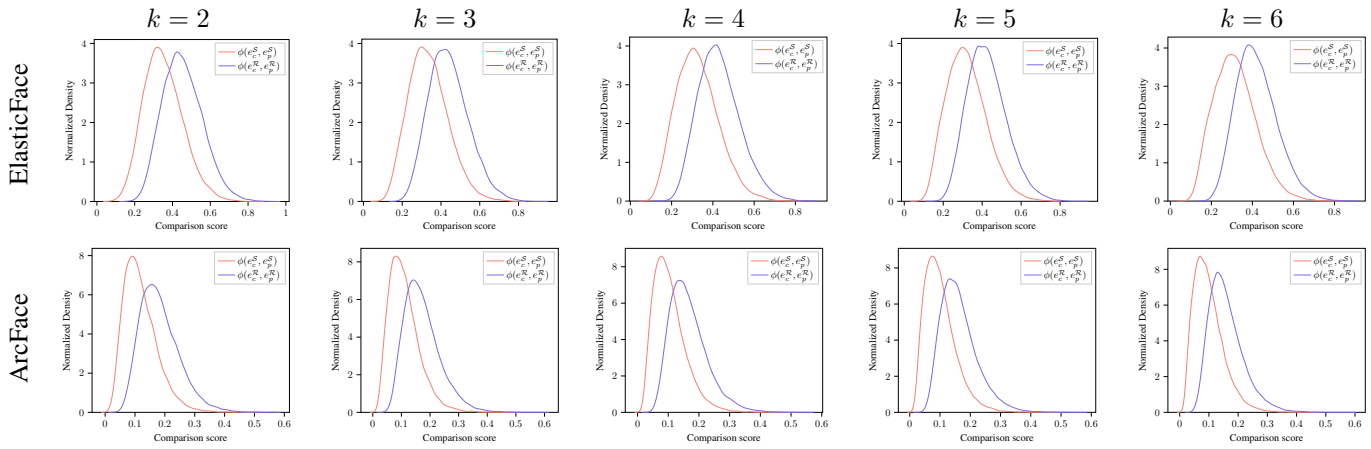

Figure 3. Match score distributions of each type of composite images with their parent images, *i.e.*, distribution of $\phi(e_c^{\mathcal{S}}, e_p^{\mathcal{S}})$ alongside $\phi(e_c^{\mathcal{R}}, e_p^{\mathcal{R}})$, for all values of k and all face-matchers, on the CelebAHQ dataset.

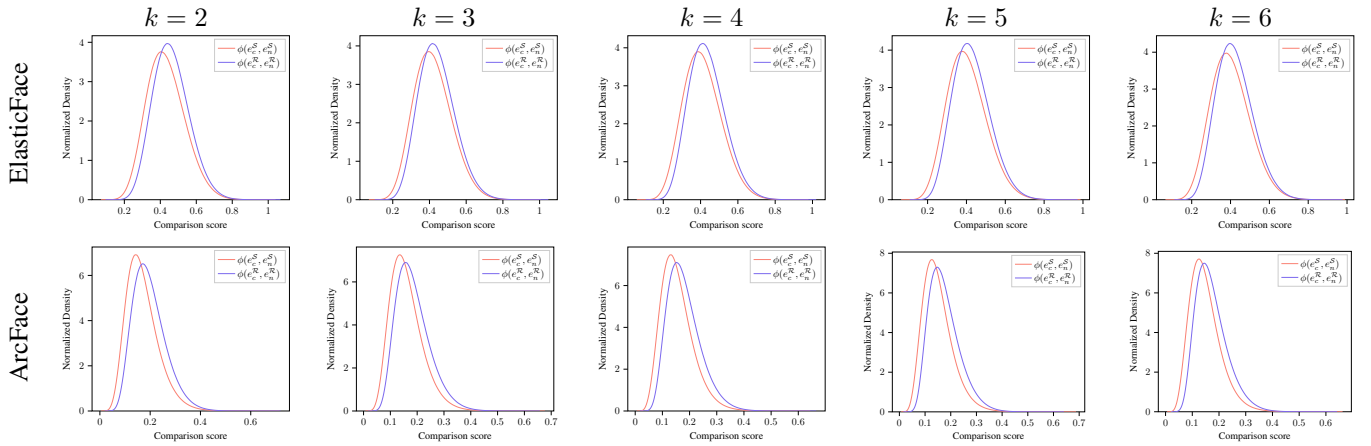

Figure 4. Match score distributions of each type of composite images with their strict non-parent images, *i.e.*, distribution of $\phi(e_c^{\mathcal{S}}, e_n^{\mathcal{S}})$ alongside $\phi(e_c^{\mathcal{R}}, e_n^{\mathcal{R}})$, for all values of k and all face-matchers, on the FFHQ dataset.

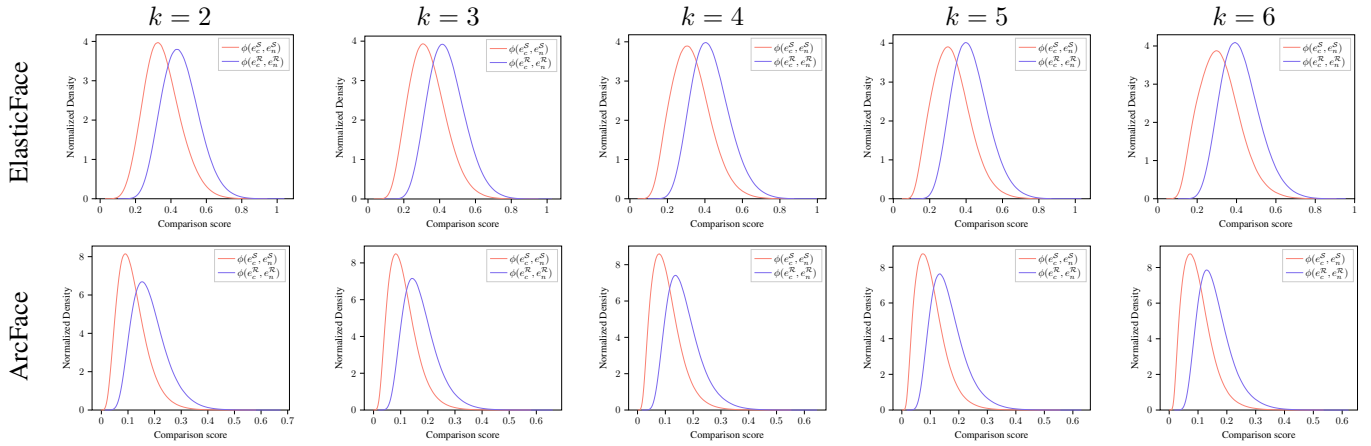

Figure 5. Match score distributions of each type of composite images with their strict non-parent images, *i.e.*, distribution of $\phi(e_c^{\mathcal{S}}, e_n^{\mathcal{S}})$ alongside $\phi(e_c^{\mathcal{R}}, e_n^{\mathcal{R}})$, for all values of k and all face-matchers, on the CelebAHQ dataset.

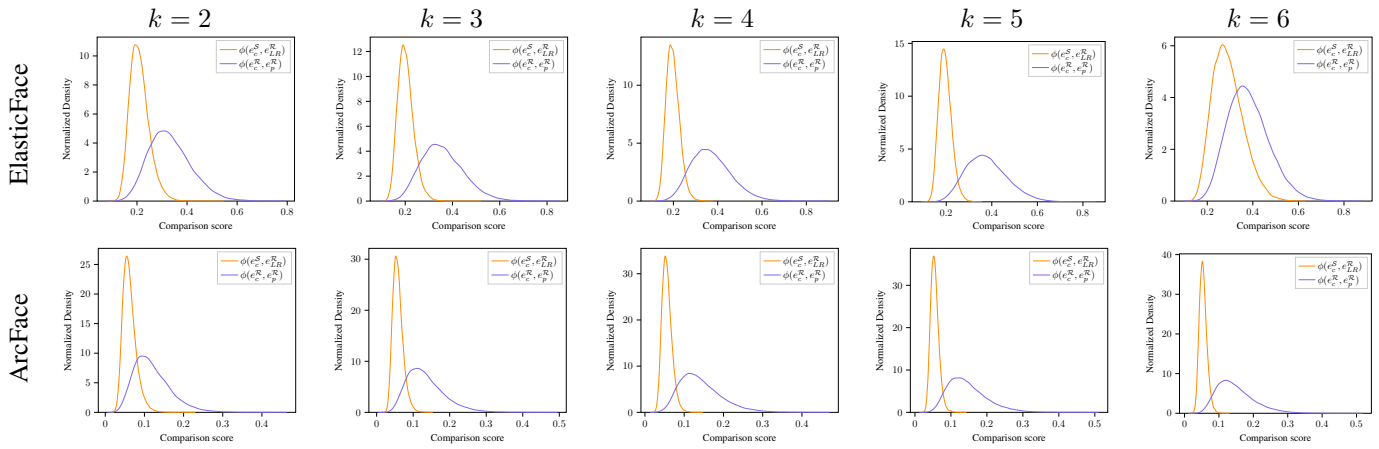

Figure 6. Match score distributions of each type of composite images with their LRs, *i.e.*, distribution of $\phi(e_c^{\mathcal{S}}, e_{LR}^{\mathcal{R}})$ alongside $\phi(e_c^{\mathcal{R}}, e_n^{\mathcal{R}})$, for all values of k and all face-matchers, on the FFHQ dataset.

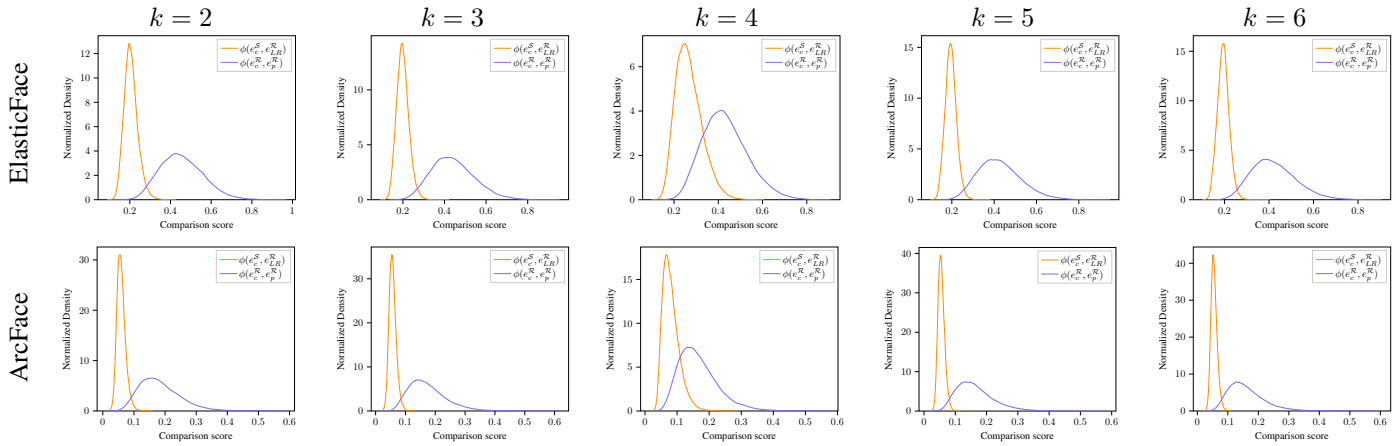

Figure 7. Match score distributions of each type of composite images with their LRs, *i.e.*, distribution of $\phi(e_c^{\mathcal{S}}, e_{LR}^{\mathcal{R}})$ alongside $\phi(e_c^{\mathcal{R}}, e_n^{\mathcal{R}})$, for all values of k and all face-matchers, on the CelebAHQ dataset.

Figure 8. By increasing the value of k we find that face images converge to one of highly-averaged facial attributes, greatly reducing the diversity of composite face images.