Haha! Caught you: Evaluating StyleID, a tool for anonymizing facial images using Record Linkage Attacks

Sargam Gupta and Vicenç Torra, Umeå University

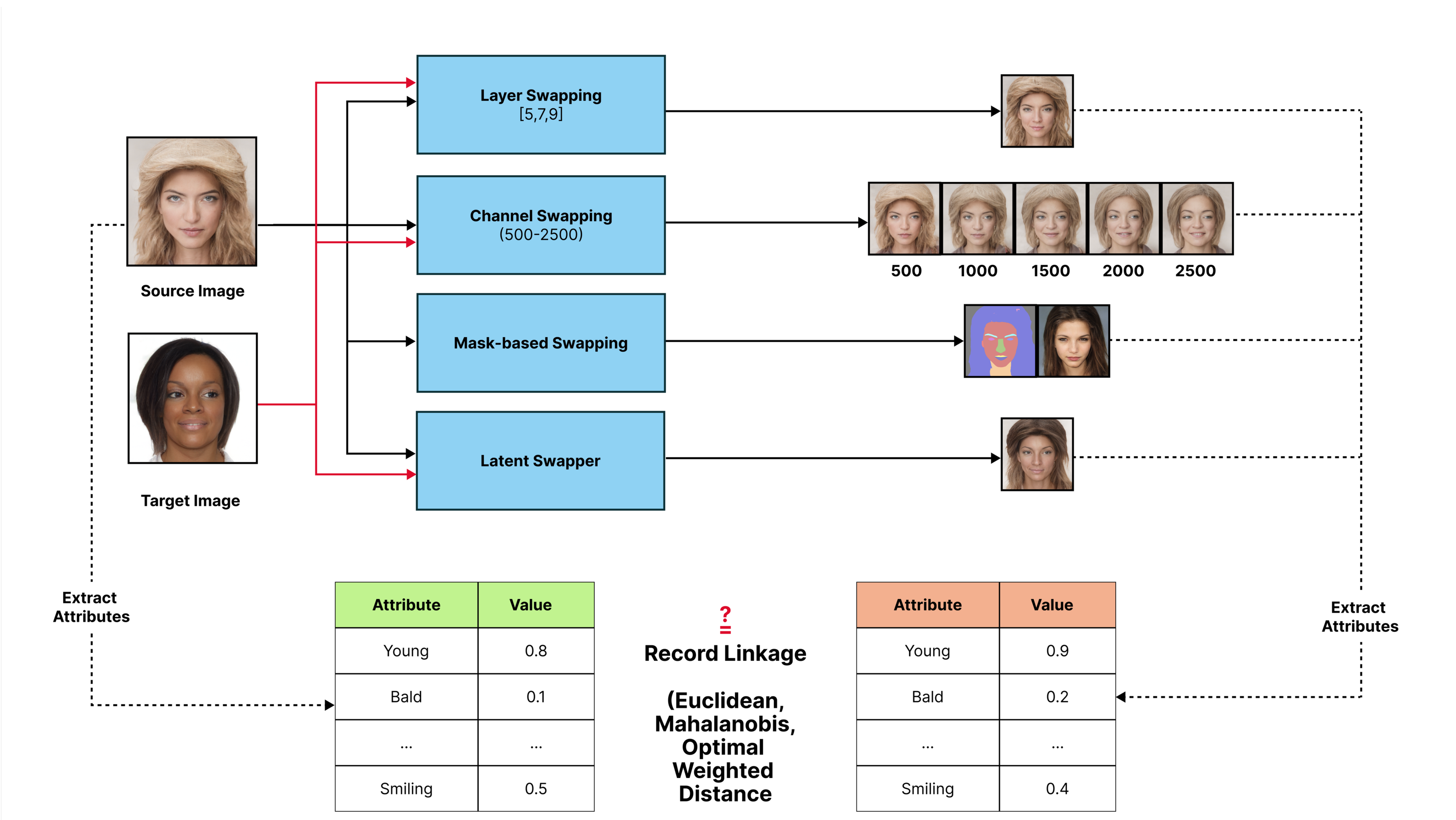
Department of Computing Science

UMEÅ UNIVERSITET

Abstract

The uploading and sharing of facial images of individuals is on the rise in this growing age of social media. But, this exchange of images may lead to some serious privacy threats. To preserve the identity of the individuals, several face de-identification tools have been proposed in the past. In our work, we evaluate the privacy-preserving nature of the different disentanglement methods proposed by Minh-Ha et al. [1] and compare them to find out which one is the best for anonymizing facial images. We have used **record linkage attacks** under various settings for this evaluation. Our experiments were able to link more than **50% of the anonymized records** to the original image in some cases which exceeds an acceptable limit for privacy.

Our Evaluation Framework



- We have used three variants of the record linkage attacks for the evaluation of the StyleID framework.
- **Euclidean Distance:** $d(a,b)^2 = \sum_{i=1}^{n} \left(\frac{V_i^X(a) - V_i^Y(b)}{\sigma(V_i^X - V_i^Y)} \right)^2$
- **Mahalanobis Distance:** $d(a,b)^2 = (a-b)' \left[\mathrm{Var}(V^X) + \mathrm{Var}(V^Y) - 2\mathrm{Cov}(V^X, V^Y) \right]^{-1} (a-b)$
- **Parametric Distance-based weighted mean:** $dWM(d(V_1(a_i), V_1(b_i)), ..., d(V_n(a_i), V_n(b_i))) < dWM((d(V_1(a_i), V_1(b_j)), ..., d(V_n(a_i), V_n(b_j)))$

$$\text{Min: } \sum_{i=1}^{N} K_i$$

$$\text{Subject to: } \sum_{i=1}^{N} \sum_{j=1}^{N} dWM_i^2$$

$$((d(V_1(a_i), V_1(b_j)), ..., d(V_n(a_i), V_n(b_j))) -$$

$$(d(V_1(a_i), V_1(b_i)), ..., d(V_n(a_i), V_n(b_i)))$$

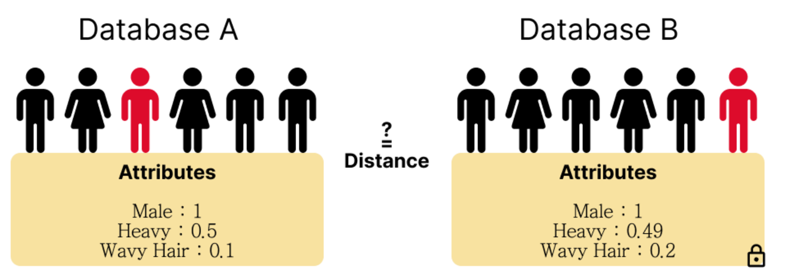
$$+CK_i > 0$$

$$K_i \in 0, 1$$

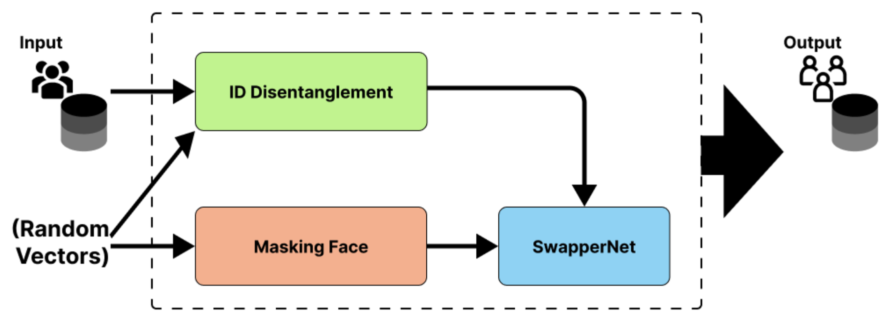
$$\sum_{i=1}^{n} p_i = 1$$

$$p_i \geq 0$$

- For every correctly linked record i, the aggregation of the values $d(V_k(a_i), V_k(b_i))$, say, with a weighted mean, for all k is smaller than the aggregation of the values $d(V_k(a_i), V_k(b_j))$ where $i \neq j$.
- Each block has a set of all distances between one record from the original data file and all the records from the protected data file. K_i will be the decision variable associated to a_i in the objective function.
- The value of K_i can be 0 or 1. **K_i = 0 if the constraints are accomplished for a_i or K_i = 1 if not. We want to minimize the number of K_i equal to 1.**

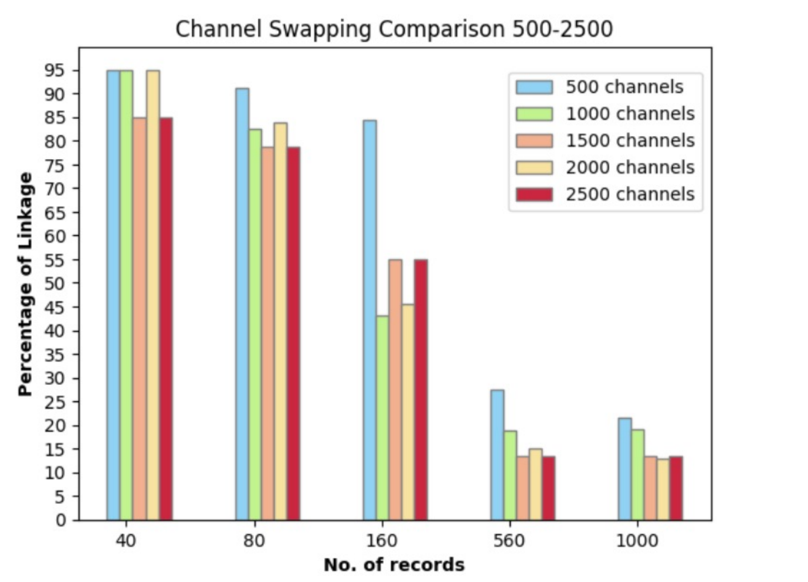


Record Linkage Attacks

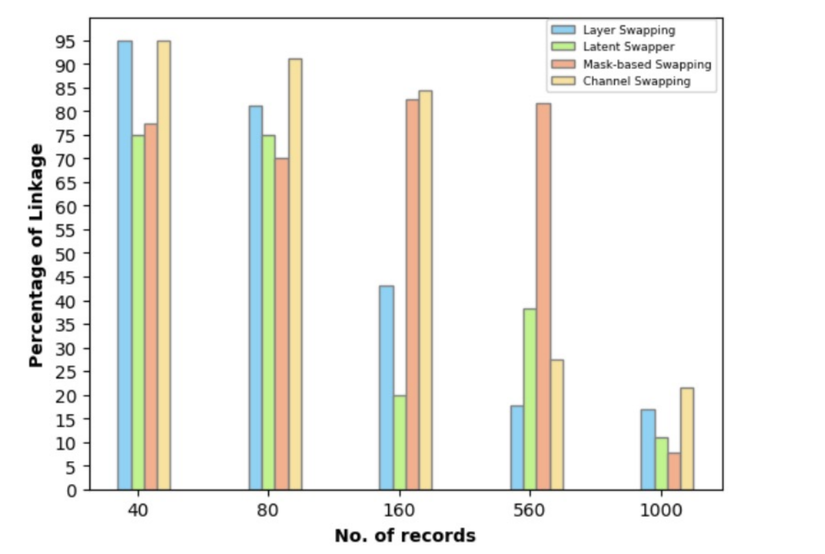


Simplified StyleID Framework

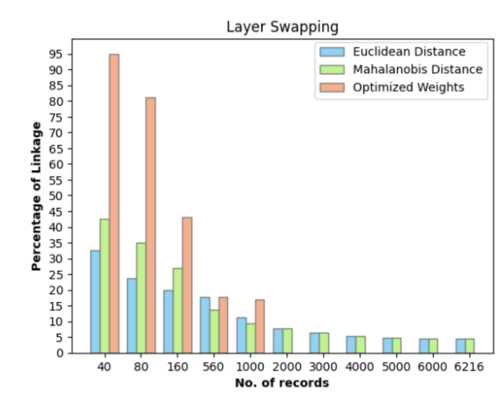
Results

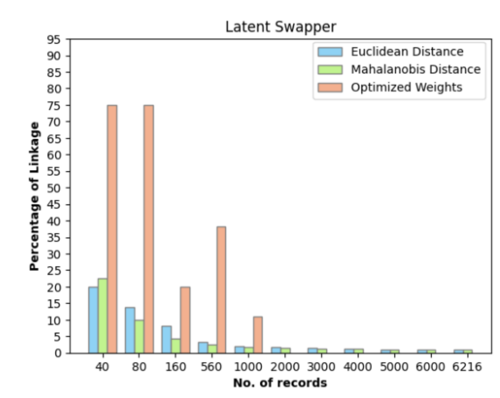


(a) Comparison between different numbers of top channels using the optimized weighted distance-based record linkage



(b) Comparison of different disentanglement methods using the optimized weighted distance-based record linkage





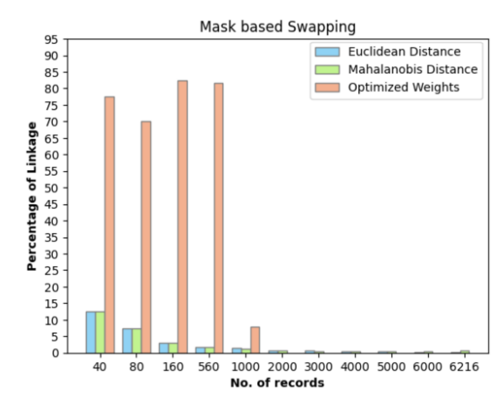


Table 2: Optimized Record Linkage for Mask-based Swapping

No of records	Value of k	%linkage	Gap
40	9	77.5%	55.54%
80	24	70%	50%
160	94	82.5%	99.96%
560	102	81.70%	98.29%
1000	922	7.8%	99.56%

Table 5: Optimized Record Linkage for Layer Swapping

No of records	Value of k	%linkage	Gap
40	2	95%	50%
80	15	65%	86.63%
160	91	43.12%	97.80%
560	460	17.8%	100.0%
1000	830	17%	100.0%

Table 3: Optimized Record Linkage for Latent Swapper

No of records	Value of k	%linkage	Gap
40	10	75%	56.6%
80	20	75%	75.71%
160	128	20%	96.7%
560	346	38.21%	98.98%
1000	891	10.9%	99.89%

Table 4: Optimized Record Linkage for Channel Swapping

No. of records	Level	Value of k	%linkage	Gap
40	500	2	95%	0%
	1000	2	95%	0%
	1500	5	87.5%	0%
	2000	2	95%	0%
	2500	6	85%	0%
80	500	7	91.25%	45.59%
	1000	14	82.5%	78.57%
	1500	15	81.5%	35.88%
	2000	13	83.75%	32.30%
	2500	17	78.75%	8.5%
160	500	25	84.38%	92%
	1000	91	43.13%	97.80%
	1500	83	48.13%	99.69%
	2000	87	45.63%	100.0%
	2500	72	55.00%	98.57%
560	500	406	27.50%	100.0%
	1000	455	18.75%	100.0%
	1500	475	15.10%	100.0%
	2000	475	15.10%	100.0%
	2500	484	13.50%	100.0%
1000	500	786	21.40%	100.0%
	1000	810	19.00%	100.0%
	1500	848	15.20%	100.0%
	2000	870	13.00%	100.0%
	2500	865	13.50%	100.0%

Conclusion

We have presented a disclosure risk assessment using distance- based record linkage attacks to evaluate StyleID, a feature-preserving anonymization framework for facial images. We have done a comparison of the different disentanglement techniques on the basis of how well can they link to the original image after anonymization. We have shown that for some of the disentanglement techniques the identity disclosure risk can be quite high, and unless the number of images in the database is large, the produced images can still be sensitive. The results in the paper show that among the available techniques, segmentation mask-based swapping seems to be a good approach for preserving privacy. We plan to work with stronger attacks on the framework to validate more anonymization techniques thoroughly.

References

1. Minh-Ha Le and Niklas Carlsson. 2023. StyleID: Identity Disentanglement for Anonymizing Faces. Proceedings on Privacy Enhancing Technologies 1 (2023), 264–278.
2. Daniel Abril, Guillermo Navarro-Arribas, and Vicenç Torra. 2012. Improving record linkage with supervised learning for disclosure risk assessment. Information Fusion 13, 4 (2012), 274–284.
3. Vicenç Torra. 2022. Guide to Data Privacy. Springer Cham. https://doi.org/10.1007/978-3-031-12837-0.
4. Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 4401–4410.

Contact Information

Sargam Gupta

sgupta@cs.umu.se

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM