Validating Privacy-Preserving Face Recognition under a Minimum Assumption

Supplementary Material

This document provides detailed discussions on key definitions (Section A1), implementation details of three PPFRs in the main manuscript (Section A2), privacy inference analysis concerning both unprotected state-of-the-art face recognition systems (Section A3) and their white-box settings (Section A4). Additionally, it includes ablation studies (Section A5), quiz-based subjective evaluation (Section A6), and a collection of supplementary generated face images (Section A7).

A1. Vocabulary

Target system S_T. This is a privacy-preserving biometric recognition system that is targeted by an attacker. In this system, the recognition results are computed on the server side and sent back to clients, and the results are often represented as similarity scores (indicating the similarity between the query face and the enrolled face). Clients set their thresholds based on their specific requirements to obtain final recognition decisions locally.

Validation system S_V. We consider another face recognition system called the validation system to evaluate the degree of privacy disclosure revealed by the generated face images from the S_T. S_V could be either the privacy-protecting face recognition systems (i.e., Duetface, DCTDP, and Partialface) or unprotected face recognition systems without privacy protection (i.e., ArcFace, MagFace, and AdaFace).

'Mode' in prior space construction. The term mode typically refers to the modes of the data distribution in GANs (see [1]), which relates to the variety and diversity of samples that a GAN can generate. However, GANs face challenges in capturing all the modes or diverse patterns present in the data distribution. We adopt the concept of an image prior distribution to mitigate mode collapse, inspired by [6]. Each image prior is uniquely characterized by a parameter, denoted as θ, and is referred to as a mode to align with GAN notation.

Query efficiency. In our 1k1c assumption, we consider the attacker's behavior to mimic that of a regular user. Typically, one query attempt is made per user inputting a single-face image to access the system. Ideally, we aim to minimize these attempts. In our experiments, the optimization iterations necessary in Map^{2}V are regarded as query counts. This simulates the adversary's actions in real attack scenarios, where reconstructed images are used to query S_T. Hence, query efficiency refers to the number of query counts required to generate the final face images.

Generalizability. The ability of a privacy validation method to effectively perform on previously unseen verification models and face images of the target identity is referred to as **generalizability**. In other words, a privacy validation method with high generalizability can assess the privacy of various target systems and achieve consistently high privacy scores across different validation systems, including PSI and PSII.

A2. Implementation details of PPFRs

Table A1. Face recognition accuracy on different datasets.

Method	LFW	CFP-FP	AgeDB	CALFW
DuetFace	99.62	93.20	96.68	95.47
DCTDP	99.63	93.17	95.82	95.27
PartialFace	98.53	89.51	90.10	91.08

We conducted privacy validation on three different SOTA PPFRs, namely DuetFace [4], DCTDP [3], and PartialFace [5]. Since these methods do not have publicly available pre-trained models, we trained these models as closely as possible in this work. The specific details are as follows.

DuetFace: as a privacy-preserving face recognition system, a face image is split by channel frequency and used to co-train the client-side model and server-side model. We use the adapted ResNet50 with an improved residual unit (IR-50) as the server-side backbone and the MobileFaceNet as the client-side backbone, and they are trained with MS-celeb-1m [2] as the training set. To be fair, the latter two methods are trained on the same dataset.

DCTDP: as a face recognition system, we use ResNet50 backbone as the baseline model and ArcFace as a loss function. In order to limit face recognition service providers to only learn classification results with a certain confidence level and fail to recover the original image, a face image is first converted to the frequency domain the DC channel is removed, and then different privacy budgets are set for different elements combined with differential privacy. In order to achieve a similar recognition performance as the original paper, the initial values of the learnable budget allocation parameters are set to 2 and the learning rate is 0.001. Other Settings are consistent with the original paper.

PartialFace: a face image is transformed to frequency channels using the Discrete Cosine Transform (DCT) and then human-perceivable low-frequency components are pruned. Moreover, a randomized strategy is introduced to enhance privacy and impede easy recovery. We employ an IR-50 backbone trained on the MS-celeb-1m dataset and relevant parameters are set as in the original paper (

i.e. $(\sigma, s, r, m, n) = (10, 36, 18, 6, 6)$ and randomly select S, P).

The recognition performance on different datasets of adopted PPFRs is shown in Table A1.

A3. Privacy validation on unprotected face recognition using Map^{2}V

In addition to the validation performed on PPFR in the main text, we utilized our proposed Map^{2}V method to assess the vulnerability of the SOTA face recognition systems. The Map^{2}V employed the same parameter settings as described in the main content, conducting privacy inference on the ArcFace, MagFace, and AdaFace systems under 1k1c settings. The results in Table A2 indicate that the average PSI exceeds 90%, and the average PSII is around 80%. This strongly suggests a significant privacy leakage risk in these face recognition models. Furthermore, it further demonstrates the strong generalizability of our proposed method when facing both protected and unprotected face recognition systems.

A4. Privacy validation on PPFRs under white-box settings using Map^{2}V

We also conducted privacy validation on the target system in a white-box setting (assuming that there is a gradient to be used). The results in Table A3 demonstrate that images reconstructed from the target system consistently achieve privacy scores of over 90% for PSI and around 80% for PSII across various S_V. Taking the target system DuetFace as an example, after obtaining a face image from DuetFace on LFW, the privacy score PSI of matching this face with other PPFR systems is 94.35% and 94.64% for DCTDP and PartialFace, respectively. In the most challenging scenario for PSII, matching this face with other PPFR systems enrolled with a new face leads to PSII scores of 81.79% and 81.93% for DCTDP and PartialFace, respectively. These results showcase that the proposed Map^{2}V is capable of achieving accurate privacy inference under white-box settings.

A5. Additional ablation studies of Map^{2}V

For ablation, we compared the impact of different parameter choices on the results of privacy inference, including parameters u, top-k, and learning rate α. To improve experimental efficiency, we conducted privacy validation on DuetFace, DCTDP, and PartialFace using the LFW dataset and calculated average privacy scores PSII when attacking the same model. This is sufficient to illustrate the impact of different parameters on the proposed MAP^{2}V. The results are shown in the Fig. A2. We can observe that in our privacy validation scenario, the combination of u=16, k = 5 and α=0.1 emerges as the optimal choice, striking a balance between maintaining the highest privacy score and visual quality. These parameters may require appropriate adjustments when validating a new PPFR system.

A6. Quiz-based subjective evaluation

Quiz: This quiz is designed based on the LFW face dataset. The quiz consists of 30 multiple-choice questions, each featuring a reconstructed image and five options, including one target option, three distractors, and a 'none' option. Respond to the question: Which option below is most similar to the given image? (see Figure A3)

Subjective privacy score (SPS). In assessing the similarity between reconstructed images and target images, we calculated the subjective privacy score (SPS) for each reconstructed image. The calculation method for this score is as follows:

$$SPS = \frac{\text{Number of observers selecting target images}}{\text{Total number of observers}} \times 100\% \quad (1)$$

The overall subjective privacy score (SPS) for the reconstructed images is obtained by computing the average of all individual scores.

Subjective Recognition Rates(SRR). To assess whether the reconstructed face images, while resembling the target images, also resemble other identities, we established subjective recognition rates.

$$SRR = \frac{\text{Number of observers selecting IDi}}{\text{Total number of observers}} \times 100\% \quad (2)$$

A7. Additional demonstration of the results of Map^{2}V

In order to facilitate a comparison with the native privacy validation methods under the 2k2c assumption from PPFR works of literature, we have provided an additional visual demonstration of generated face images, as shown in the Fig.A4.

It is worth noting that the reconstructed face images under the 1k1c assumption using the proposed Map^{2}V are significantly superior to the results of auto-encoders in PPFR works. This echoes our experimental analyses in Sec. 4.4.

References

[1] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do gans learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018. 1

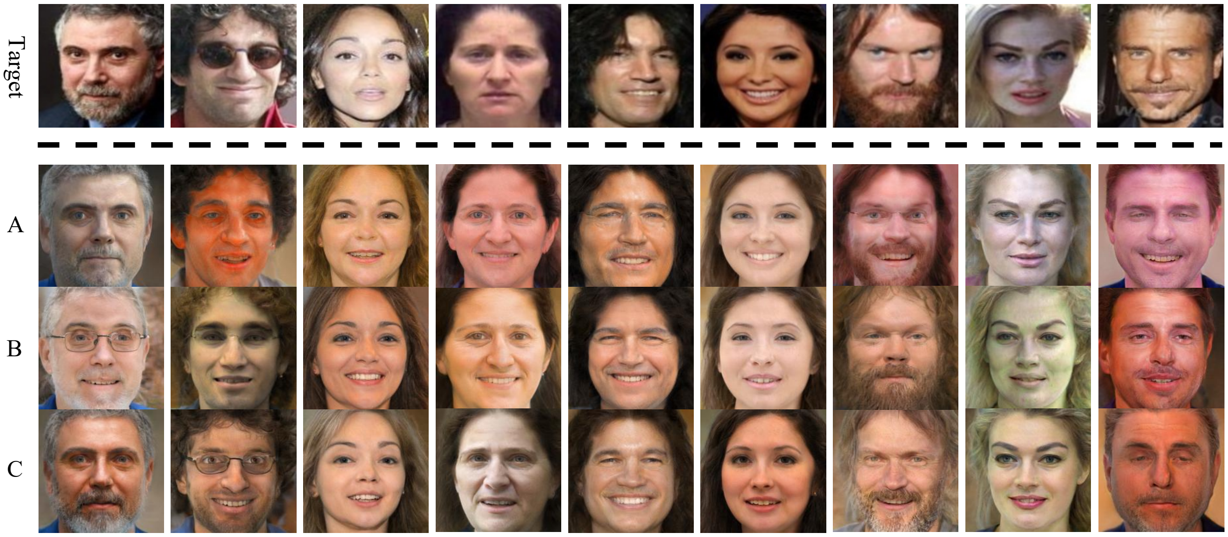
[2] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016. 1

Table A2. Privacy scores (%) against different unprotected face recognition system on LFW and CelebA datasets under the 1k1c scenario.

		Reconstructed from S_T							
Dataset	**Attack to S_V**	**ArcFace**		**MagFace**		**AdaFace**		**Average**	
		PSI	PSII	PSI	PSII	PSI	PSII	PSI	PSII
LFW	ArcFace	**97.76**	**91.14**	88.26	71.26	85.76	66.61	90.59	76.34
	MagFace	96.30	83.54	**96.80**	**87.95**	86.80	70.55	93.30	80.68
	AdaFace	96.06	77.90	89.56	75.81	**98.31**	**87.98**	94.64	80.56
CelebA	ArcFace	**96.73**	**86.78**	90.48	72.90	90.23	75.12	92.48	78.27
	MagFace	96.73	82.45	**96.23**	**82.42**	93.73	78.73	95.56	81.20
	AdaFace	93.57	75.67	88.57	70.70	**96.57**	**83.29**	92.90	76.55

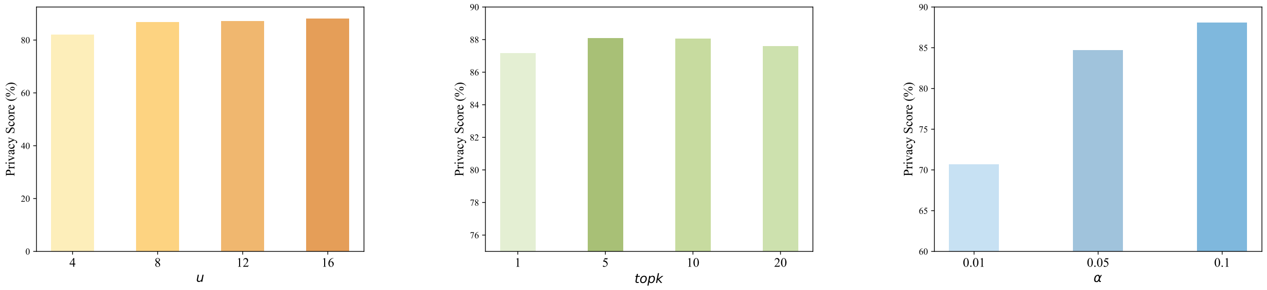
Table A3. Privacy scores (%) against different validation systems on LFW and CelebA dataset under the white-box scenario.

		Reconstructed from S_T							
Dataset	**Attack to S_V**	**DuetFace**		**DCTDP**		**PartialFace**		**Average**	
		PSI	PSII	PSI	PSII	PSI	PSII	PSI	PSII
LFW	ArcFace	96.51	81.95	93.51	76.72	91.76	69.83	93.93	76.17
	MagFace	92.80	81.69	90.05	78.77	92.30	75.78	91.72	78.75
	AdaFace	93.81	79.51	87.56	75.46	91.91	72.23	91.09	75.73
	DuetFace	**97.46**	**92.01**	94.55	81.62	95.58	79.90	95.86	84.51
	DCTDP	94.35	81.79	**96.35**	**89.59**	90.35	75.33	93.68	82.24
	PartialFace	94.64	81.93	89.39	78.17	**98.89**	**96.06**	94.31	85.39
CelebA	ArcFace	95.23	81.11	94.73	78.59	95.23	81.65	95.06	80.45
	MagFace	94.48	82.27	93.73	80.28	96.48	84.40	94.90	82.32
	AdaFace	91.57	76.80	89.82	74.58	93.57	79.04	91.65	76.81
	DuetFace	**97.94**	**88.45**	95.08	80.55	95.26	83.64	96.09	84.21
	DCTDP	95.27	81.59	**96.77**	**86.69**	92.52	79.94	94.85	82.74
	PartialFace	93.76	81.11	90.04	77.67	**97.79**	**90.69**	93.86	83.16



A: Reconstructed from DuetFace   B: Reconstructed from DCTDP   C: Reconstructed from PartialFace

Figure A1. More examples of reconstructed faces from the CelebA dataset for three SOTAs under white-box settings. The first row is the target image for the adversary's attack, A to C shows the results of Map^{2}V.



(a) (b) (c)

Figure A2. Privacy scores of different choice of parameter: u and ϵ for zeroth-order gradient estimation. When varying a single hyperparameter, other hyperparameters are fixed to the optimal value (u = 16, k = 5, α = 0.1).

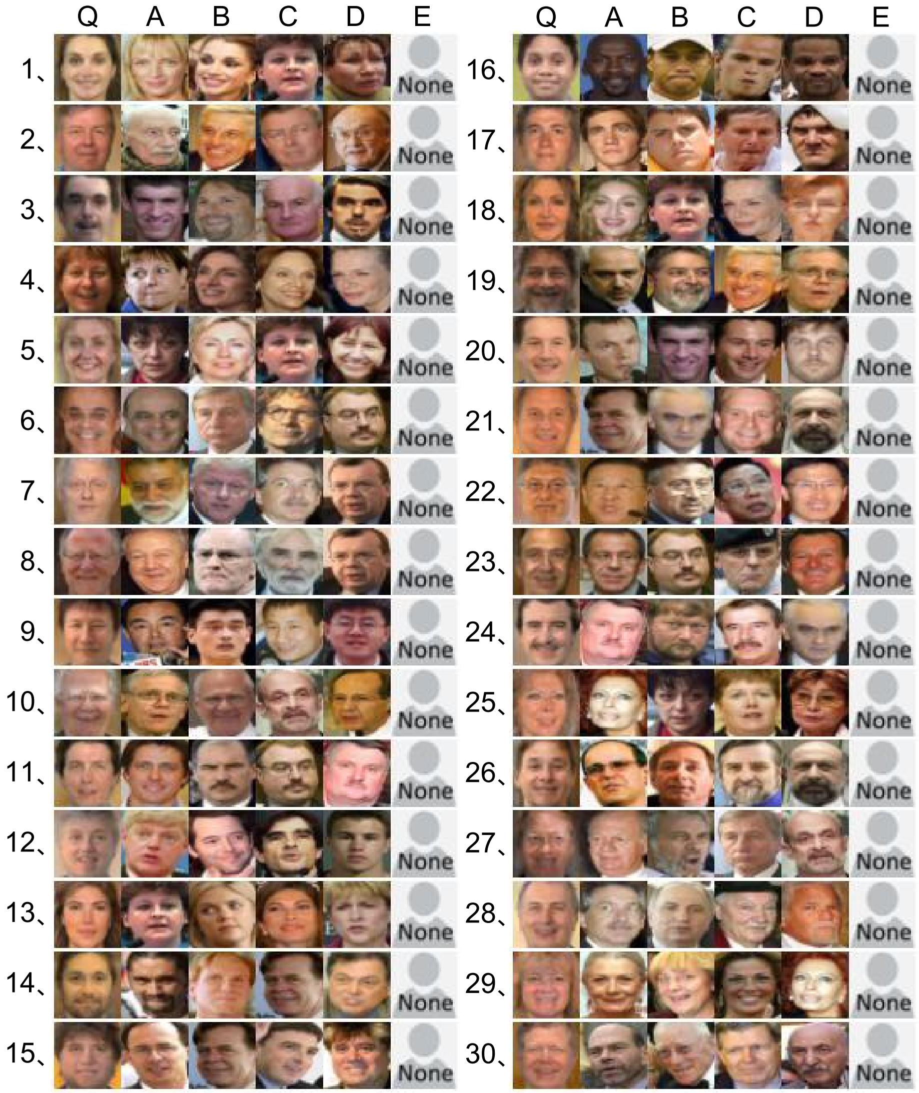


Figure A3. Quiz of Subjective evaluation. Each row represents a question, totaling 30. Q is the provided reconstructed image, and A, B, C, D are the target option and three distractor options placed randomly, while E is the 'none' option.

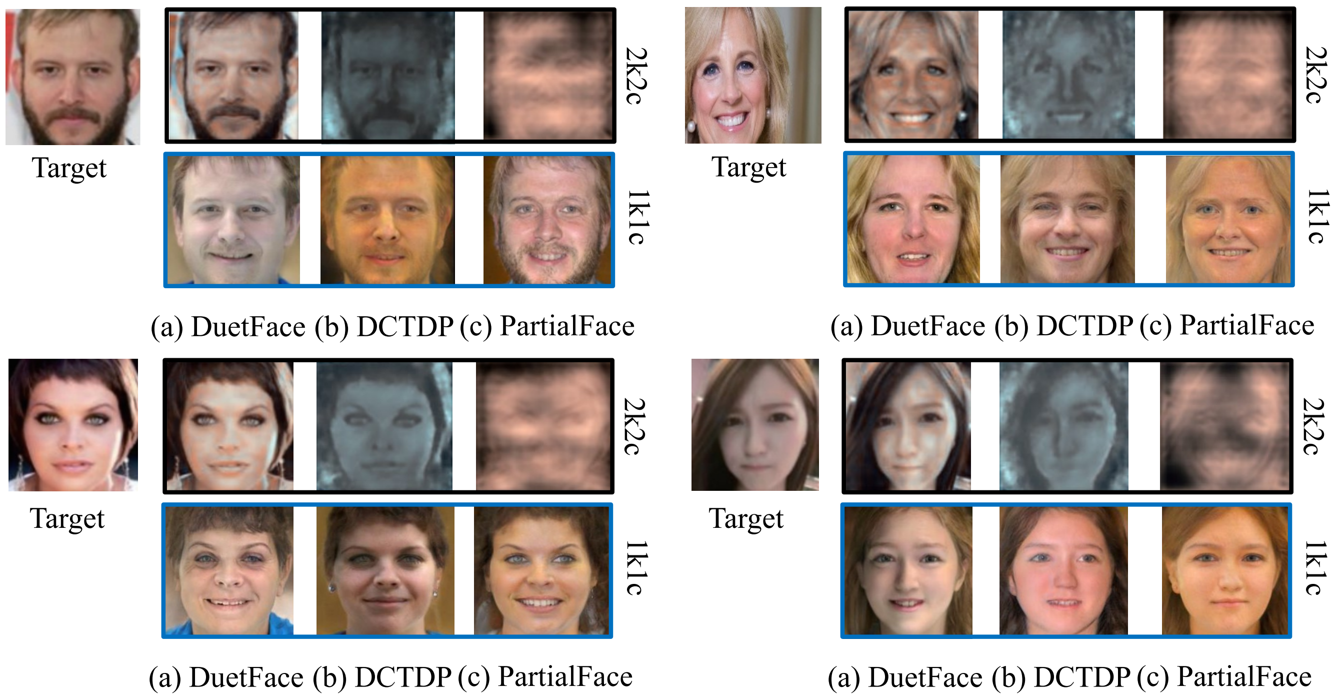

Target
2k2c
1k1c
(a) DuetFace (b) DCTDP (c) PartialFace
Target
2k2c
1k1c
(a) DuetFace (b) DCTDP (c) PartialFace
Target
2k2c
1k1c
(a) DuetFace (b) DCTDP (c) PartialFace
Target
2k2c
1k1c
(a) DuetFace (b) DCTDP (c) PartialFace

Figure A4. More exemplar face images reconstructed from SOTA PPFRs. The recovery results of proposed Map^{2}V under 1k1c settings (2nd row) outperform the rest under 2k2c settings (1st row).

[3] Jiazhen Ji, Huan Wang, Yuge Huang, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, ShengChuan Zhang, Liujuan Cao, and Rongrong Ji. Privacy-preserving face recognition with learnable privacy budgets in frequency domain. In *European Conference on Computer Vision*, pages 475–491. Springer, 2022. 1

[4] Yuxi Mi, Yuge Huang, Jiazhen Ji, Hongquan Liu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Duetface: Collaborative privacy-preserving face recognition via channel splitting in the frequency domain. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6755–6764, 2022. 1

[5] Yuxi Mi, Yuge Huang, Jiazhen Ji, Minyi Zhao, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using random frequency components. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19673–19684, 2023. 1

[6] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14, 2021. 1