# Validating Privacy-Preserving Face Recognition under a Minimum Assumption

Hui Zhang, Xingbo Dong, YenLung Lai, Ying Zhou, Xiaoyan Zhang, Xingguo Lv, Zhe Jin†, Xuejun Li
Anhui Provincial Key Laboratory of Secure Artificial Intelligence, Anhui University, China
huizhang@stu.ahu.edu.cn, jinzhe@ahu.edu.cn

## Abstract

*The widespread use of cloud-based face recognition technology raises privacy concerns, as unauthorized access to face images can expose personal information or be exploited for fraudulent purposes. In response, privacy-preserving face recognition (PPFR) schemes have emerged to hide visual information and thwart unauthorized access. However, the validation methods employed by these schemes often rely on unrealistic assumptions, leaving doubts about their true effectiveness in safeguarding facial privacy. In this paper, we introduce a new approach to privacy validation called Minimum Assumption Privacy Protection Validation (Map$^2$V). This is the first exploration of formulating a privacy validation method utilizing deep image priors and zeroth-order gradient estimation, with the potential to serve as a general framework for PPFR evaluation. Building upon Map$^2$V, we comprehensively validate the privacy-preserving capability of PPFRs through a combination of human and machine vision. The experiment results and analysis demonstrate the effectiveness and generalizability of the proposed Map$^2$V, showcasing its superiority over native privacy validation methods from PPFR works of literature. Additionally, this work exposes privacy vulnerabilities in evaluated state-of-the-art PPFR schemes, laying the foundation for the subsequent effective proposal of countermeasures. The source code is available at https://github.com/Beauty9882/MAP2V.*

## 1. Introduction

A cloud-based face recognition (FR) system usually comprises the client and server sides. The client devices capture facial images, which are then transmitted to a cloud server [38, 41]. The server employs machine learning algorithms to analyze facial features, matching them against stored data. Results are sent back to clients, enabling seamless and efficient identity verification and access control [29, 37]. As FR technology becomes increasingly prevalent, the need to safeguard the privacy of sensitive facial images has become a critical concern. This is primarily due to the fact that face is typically considered private by the client, who is hesitant to share the raw image with external entities, including the server [10, 32, 44].

Recently, various privacy-preserving face recognition (PPFR) schemes have been constructed using deep learning-based techniques to address the aforementioned challenges [6, 28, 41, 46, 47]. Despite the growth of PPFR techniques and the sound security they demonstrated, there is still a lack of a proper framework capable of comprehensively validating their model capacity to safeguard against privacy inference [14, 34]. The existing privacy validation methods [11, 22, 23] typically focus on reconstructing face images from user data stored on the server. These methods often rely on numerous assumptions and can be collectively termed as the **2k2c** framework, granting adversaries the following advantages:

1. **K**nowledge to the PPFR system, be it within a black-box or white-box settings.
2. **K**nowledge to the protected user data collectible from the server or database within a PPFR system.
3. **C**apabilities to query the server indefinitely, aiming to obtain input-output pairs for network training.
4. **C**apabilities to independently train a face reconstruction network (e.g., using autoencoders) and reverse-engineer the face from partial information obtained from the server.

The aforementioned **2k2c** framework is indeed a less realistic attacking environment for the potential adversary. To elaborate, granting the adversary the capability to collect protected user data from the server or database effectively provides the attacker access to the PPFR system with minimal (or even zero) effort. On the other hand, allowing the adversary to query the server indefinitely suggests that no PPFR system can withstand such a brute-force approach without relying on computationally bounded assumptions.

Therefore, we refined **2k2c** assumptions into a more nuanced **1k1c** framework, relaxing the adversary's capability to only require:

1. **K**nowledge of the PPFR system but under a black-box

†Corresponding author.

setting: The adversary mimics normal user behavior, limited to a few query attempts and interaction within the user interface outside the server, observing authentication results presented as similarity scores.

2. **C**apability of exploiting generic image priors: Image priors may include public face images or generic priors from a Generative Adversarial Network (GAN) [26].

Such refined **1k1c** assumptions notably avoid unnecessary overpowered computational capability for potential adversaries, creating a much more realistic environment for any adversary aiming to compromise the targeted PPFR system. A visual comparison of the attack pipeline between 2k2c and 1k1c assumptions is illustrated in Fig. 1. In particular, we emulate a potential adversary as a typical user, interacting with the system through queries and gathering exploitable information from the returned similarity scores (see Fig. 1 (c)). This exploration, characterized by *minimized assumptions* to adversary's computing power, not only provides insights into the vulnerabilities of the PPFR system but also holds immediate relevance in combating database attacks through efficient system queries.

Achieving the privacy inference under the 1k1c assumption is challenging. We take a detour by leveraging generic deep image priors [26] and zeroth-order optimization based on gradient estimation to establish privacy-preserving validation method, referred to as **Map$^2$V**. Map$^2$V formulates any adversary's privacy attack on the system as a zeroth-order optimization problem aimed at obtaining the best solution from public image priors, with a specific emphasis on query efficiency and generalizability. Specifically, our main findings in this work cover three distinct domains:

- Conceptually, we introduce a novel privacy validation framework under a minimal assumption, where the adversary operates from outside the PPFR system and aims to breach privacy barriers with minimal effort by posing as a normal user and acquiring knowledge of system outputs. This scenario represents a more realistic and credible threat model, marking a significant paradigm shift in privacy validation practices.
- Technically, we introduce an explicit construction called Map$^2$V that realizes our conceptual idea. By framing the privacy inference as a zeroth-order optimization problem, we propose a novel optimization strategy using deep image priors and gradient estimation. Leveraging rich priors enables a query-efficient reconstruction of original facial images without training, making it a potential generalizable tool for PPFR evaluation.
- Experimentally, we provide a comprehensive assessment of the latest state-of-the-art (SOTA) PPFR works using our Map$^2$V approach. Employing both human and machine vision, we assess privacy leakage on three major tasks: reconstruction visual quality analysis, visual privacy analysis, and identity privacy analysis. Our findings highlight the potential of Map$^2$V as a general privacy validation method, and the vulnerability of existing PPFR methods on popular LFW and CelebA face datasets.

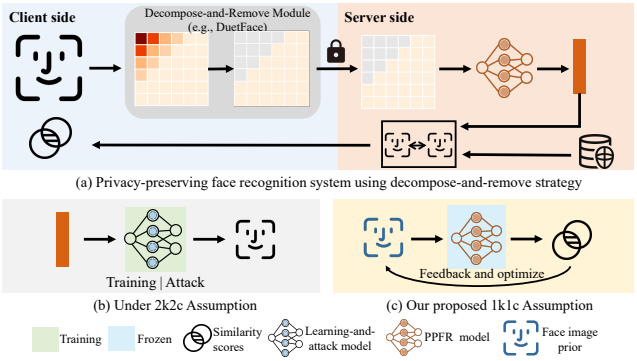



Figure 1. This paper is targeted at validating the privacy of SOTA privacy-preserving face recognition. (a) is a general decompose-and-remove-based privacy-preserving face recognition framework, and (b) is a prior learning-based privacy validation method under the 2k2c assumption. (c) is our proposed privacy validation method under the 1k1c assumption, featuring a more realistic attack environment than (b), aimed at uncovering vulnerabilities in (a).

## 2. Related Works

### 2.1. Privacy-Preserving Face Recognition

Early PPFR efforts predominantly centered on encryption-based methods, transforming faces into encrypted domains for privacy protection [4, 5, 17, 20, 30, 33]. In contrast, transformation-based methods take a different route, converting facial images into perturbed or regenerated representations to reduce distinguishability by untrusted entities [6, 31, 44, 45]. However, these approaches often involve a trade-off between privacy preservation and recognition performance [16, 18, 36, 43].

To mitigate the inherent trade-off challenge, recent studies concentrate on a novel approach: decomposing the face image and subsequently removing human-perceivable features. This process renders the remaining face images visually indistinguishable while maintaining model accuracy. We term it as "**D**ecompose-a**n**d-**R**emove" (DnR). Figure 1 (a) shows a typical example of DnR pipeline.

Numerous notable DnR approaches have been proposed for PPFR in the literature. For instance, the PPFR-FD scheme by Wang et al. [40] utilizes fast face masking through random channel shuffling, aiming to filter out human-perceptible channel information from facial images. The filtered facial data is subsequently transmitted to the server for secure storage and recognition purposes. On the other hand, Ji et al. [11] introduced the DCTDP scheme, leveraging the concept of differential privacy by introducing differential privacy perturbations into the generated face

Table 1. Comparision of privacy validation methods on SOTA PPFRs.

| Validation method | PPFR | VP | RP | IIP | Assumption | Train-free | Query count | Generalizable | Available code | Venue |
|---|---|---|---|---|---|---|---|---|---|---|
| Learning-and-attack | PPFR-FD[40] | YES | YES | YES | - | No | - | - | No | AAAI-2022 |
| | DCTDP[11] | YES | No | No | 2k2c | No | ≈ Millions | NO | Yes | ECCV-2022 |
| | DuetFace[22] | YES | No | No | 2k2c | No | ≈ Millions | NO | Yes | ACMMM-2022 |
| | PartialFace[23] | YES | No | No | 2k2c | No | ≈ Millions | NO | No | ICCV-2023 |
| **Map$^2$V** | DnRs [11, 22, 23] | **YES** | **YES** | **YES** | **1k1c** | **YES** | **Thousands** | **YES** | **YES** | - |

features. The difficulty of recovering the original face images is enhanced by allocating different privacy budgets to decomposed frequency domain features. Under a semi-honest server threat model, the DuetFace scheme was proposed by Mi et al. [22]. It incorporates the use of Block Discrete Cosine Transform along with channel splitting. This process involves identifying and preserving essential channels within the high-frequency spectrum, which is crucial for recognition performance while removing the non-essential ones. In their latest work, Mi et al. [23] have introduced another DnR strategy, namely PartialFace, involving the extraction of complementary local features from a diverse image set through mixed training. This approach equips the recognition model to gain a holistic understanding of collective data while minimizing individual information exposure risk.

### 2.2. Privacy Validation of PPFRs

While DnR conceals visual information from the server, providing a conceptually sound approach to privacy protection. It remains unclear to which extent that transmitting only visually imperceptible data to the server can effectively counter potential privacy inference by adversaries interacting directly with the system. This underscores the need for comprehensive privacy validation for DnR approaches, encompassing tasks such as reconstructing visually inaccessible query images and inferring the face's identity through intercepted or unauthorized redistributed information, as highlighted in existing literature [24, 25, 27, 39].

The generic privacy validation method aims to evaluate three primary privacy preservation requirements, which are not absolute necessities but are viewed as sufficient conditions for a secure PPFR, as follows:

- **Visual Privacy (VP)**: On the very basis, the server should not have the capability to extract meaningful information from the visual appearance of the face image.
- **Reconstruction Privacy (RP)**: The server should not be able to effectively reconstruct the original face image (or its related information) from its protected counterpart.
- **Identity Inferred Privacy (IIP)**: The server can redistribute reconstructed images or intercept transmitted messages. However, without accessing the recognition model, the information learnable from the reconstructed image should not reveal the user's identity.

Table 1 summarizes the privacy validation methods of PPFR schemes. It can be observed that SOTA PPFR systems construct autoencoder models, such as UNet, based on the 2k2c assumption for privacy validation. This involves determining compliance with the three privacy preservation requirements mentioned above. Attack models based on autoencoder models heavily rely on training data obtained through queries to the PPFR system. This approach is not aligned with the limited access nature of practical face recognition applications. Moreover, when applied to different targeted PPFR, the need for retraining limits the model's generalizability and significantly raises validation costs across a broad spectrum of distinct PPFR systems. In light of the above reasoning, constructing a query-efficient and generalizable privacy validation method with minimized assumptions is crucial for exposing the vulnerabilities of PPFR and promoting its development. This motivation serves as the core driving force in this work.

## 3. Methodology

### 3.1. Formalization

**Objective.** Let $x$ and $\hat{x}$ denote the human-imperceptible decomposed facial data derived from the target facial image and reconstructed facial image. It is more intuitive to represent the PPFR model as a function $f(x) : \mathcal{D} \to \mathbb{R}^{512}$ that transforms input data $x \in \mathcal{D}$ into a real-valued feature vector of 512 dimensions. Let $\phi : \{0, 1\} \to \mathbb{R}$ denote a function that output the resultant similarity score $s \in \mathbb{R}$ corresponding to the system decision $D \in \{0, 1\}$, i.e., the decision $D(x) = 1$ means granted access and $D(x) = 0$ means reject. In practice, attackers following the 1k1c assumption aim to reconstruct the target face image solely from the system's output $s$. This scenario can be simulated using $x$ and $\hat{x}$ with the adoption of face image priors to generate facial images, aiming to **minimize** the following objective:

$$\begin{aligned} &\max \mathbb{E}_{x\sim\mathcal{D}_{real}}[\phi(D(x))] - \mathbb{E}_{\hat{x}\sim\mathcal{D}_G}[\phi(D(\hat{x}))] \\ &= \max \mathbb{E}_{x\sim\mathcal{D}_{real},\hat{x}\sim\mathcal{D}_G}\left[\left|\phi(D(x)) - \phi(D(\hat{x}))\right|\right]. \end{aligned} \quad (1)$$

To estimate $\mathbb{E}_{x\sim\mathcal{D}_{real}}[\phi(D(x))]$ using limited number of input data, one utilizes $\frac{1}{t}\sum_t \phi(D(x_i))$. The empirical version of the real distribution is therefore denoted as $\hat{\mathcal{D}}_{real}$ (likewise for $\hat{\mathcal{D}}_G$), where each $x_i$ is assigned equal probability $\frac{1}{t}$.

Essentially, facial images in our work can be either drawn from a public face dataset or a public GANs' deep

generator $G$. We adopt $G$ in our work. Given $f(\cdot)$ is 1-Lipschitz, above objective have a solution [1] where:

$$\nabla d_W(\hat{\mathcal{D}}_{real}, \hat{\mathcal{D}}_G) \approx -\mathbb{E}_{z\sim Z}[\nabla f(G(z))]. \tag{2}$$

Here, the latent vector $z$ is supposed to follow Gaussian distribution, and the Wasserstein distance $d_W$ between $\hat{\mathcal{D}}_{real}$ and $\hat{\mathcal{D}}_G$ described as:

$$d_W(\hat{\mathcal{D}}_{real}, \hat{\mathcal{D}}_G) = \sup_f \left| \mathbb{E}_{x\sim\hat{\mathcal{D}}_{real}, \hat{x}\sim\hat{\mathcal{D}}_G}[d(f(x), f(\hat{x})] \right|, \tag{3}$$

where $d(\cdot)$ represents the Euclidean distance measurement.

**Challenges.** Nonetheless, enforcing a 1-Lipschitz constraint on the PPFR model can be challenging. This constraint imposes strict bounds on the gradient, and achieving it reliably can be complex. Another challenge lies in the inherent issue of GAN generalization, known as mode collapse, where the model tends to produce a restricted range of outputs, limiting its ability to explore the complete training data distribution and capture its diversity and richness. In face privacy inference, mode collapse is undesirable as it restricts the diversity of GAN-generated outputs, limiting the range of possible facial features and hindering the comprehensive representation of the user's face. Consequently, it may lead to reconstructed results deviating significantly from the actual privacy inference level, resulting in a lack of generalizability in validation models.

**Countermeasures.** To address the above challenges, we decided to directly minimize the Wasserstein distance between $\hat{\mathcal{D}}_{real}$ and $\hat{\mathcal{D}}_G$ as indicated by Eq. 3. When $f(\cdot)$ represents a neural network-trained PPFR model, Eq. 3 becomes equivalent to the neural network distance between $\hat{\mathcal{D}}_{real}$ and $\hat{\mathcal{D}}_G$. Suppose the GAN successfully minimizes the neural network distance between the empirical distributions, i.e., $d(\hat{\mathcal{D}}_{real}, \hat{\mathcal{D}}_G)$. In that case, we can infer that the neural network distance between the original distributions $d(\mathcal{D}_{real}, \mathcal{D}_G)$ is also small (as per Theorem 3.1 in [2]). Therefore, this approach allows us to achieve GAN generalization for our intended goal. It is important to note that the Wasserstein distance (as defined in Eq. 3) is sensitive to the support of $f(\cdot)$. To be more explicit, we define:

$$\begin{aligned} &\mathbb{E}_{x\sim\hat{\mathcal{D}}_{real}, \hat{x}\sim\hat{\mathcal{D}}_G}[d(f(x), f(\hat{x}))] \\ &= \frac{1}{mt}\sum_{i=1}^{m}\sum_{j=1}^{t} d(f(x_i), f(\hat{x_j})). \end{aligned} \tag{4}$$

This scenario accounts for situations where an attacker may attempt to reconstruct $m$ instances of users' face images using a potential $t$ number of decomposed facial data $\hat{x}$. Therefore, the support of $f(\cdot)$ should be proportional to $m$ and $t$.

Besides, it is worth noting that Eq. 4 only considered a specific mode of the empirical distribution for both $\hat{\mathcal{D}}_{real}$ and $\hat{\mathcal{D}}_G$. To achieve a richer diversity and mitigate mode collapse, it is appropriate to consider a multimodal scenario, which can be described as $\hat{\mathcal{D}}_G^{(1)}, \hat{\mathcal{D}}_G^{(2)}, \ldots, \hat{\mathcal{D}}_G^{(n)}$. Here, the superscript indicates particular mode numbers totaling $n$ modes.

Considering the above, the objective for an adversary is to identify specific modes that optimize the criteria below.

$$\arg\min_{\theta} \mathbb{E}_{x\sim\hat{\mathcal{D}}_{real}, \hat{x}\sim\hat{\mathcal{D}}_G^{(\theta)}}[d(f(x), f(\hat{x}))], \tag{5}$$

where $\theta$ correspond to a specific mode that minimize $d(f(x), f(\hat{x}))$. This formalized ultimate objective entails the implicit parameterization of the potential modes during GAN training and their respective supports for the Wasserstein distance between the compared distributions in terms of the values of $n$, $m$ and $t$. This approach is highly relevant to our face privacy validation goal, where the adversary's capabilities are limited to a specific number of queries for $f(x_i)$ and $f(\hat{x}_j)$. Therefore, we can validate the privacy protection capacity of the PPFR $f(\cdot)$ about the number of queries, and this quantification is directly tied to the values of $n$, $m$, and $t$, establishing a proportionality between them.

### 3.2. Explicit Construct: Map$^2$V

Based on the above formal definition, we introduce an explicit construction for PPFR validation: Map$^2$V, with three modules: **1)** Prior space construction ($PSC$), **2)** Rank-and-Ensemble Initialization ($REI$) and **3)** Zeroth-order gradient estimation. The first and third modules aim to improve query efficiency, while the second is intended to ensure generalizability.

**Prior Space (Modes) Construction.** While Eq. 5 guarantees an optimal minimum for a specific mode characterized by $\theta$ within $n$ queries for an adversary to request $f(x)$ with a particular $x$, it entails the need to construct empirical distributions for both $\hat{\mathcal{D}}_{real}$ and $\hat{\mathcal{D}}_G^{(\theta)}$ across a total of $n$ possible modes. Taking inspiration from Tov et al.[35], we can rely on StyleGAN[13] to accomplish this. This is because, in the context of validating PPFR, the adversary has the capability of exploiting generic public image priors under the 1k1c assumption. In [35], two fundamental principles are proposed for designing encoders that offer precise control over the proximity of inversions to regions originally covered by StyleGAN training data. We embrace this encoding strategy using the StyleGAN2 generator ($G$) to construct our prior space from publicly available facial datasets, consisting of $n$ distinct modes as initialization. In contrast to optimizing from the noise space, the prior space construction strategy takes into account the adversary's knowledge and capabilities for initial space optimization, mining the optimal set of $n$ modes to form the prior space.

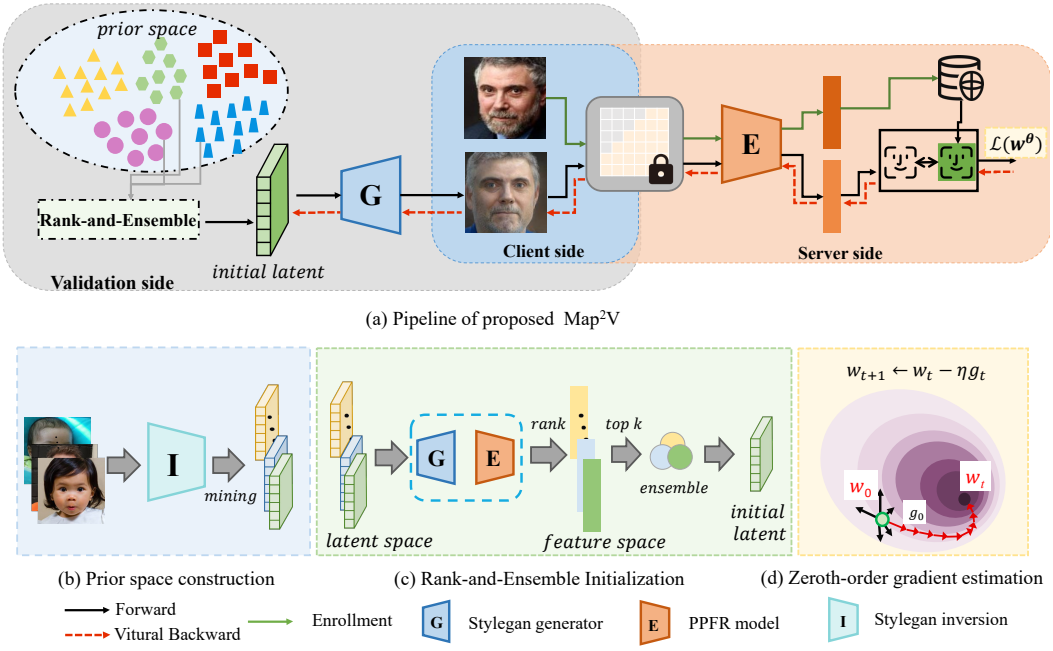



Figure 2. Overview of the proposed Map$^2$V. (a) outlines our privacy inference pipeline. (b), (c) and (d) describe the details of three modules: Prior Space Construction, Rank-and-Ensemble Initialization, and Zeroth-order gradient estimation. By collaborating these three modules, we obtain a universal privacy validation model that is efficient and generalizable.

And then we associated individual mode to single latent code $w^{(\theta)} \sim \mathcal{W}^+$, which is then fed into each layer of $G$ through AdaIN [9]. In this context, $\mathcal{W}^+$ is a subset of $\mathbb{R}^{\ell \times 512}$, where $\ell$ denotes the number of layers, e.g., $\ell = 14$. The integration of a layer-wise latent space promotes disentangled features, leading to enhanced differentiation between different modes. By doing so, it reduces the chance for $w^{(\theta)} \in \mathcal{W}^+$ veering towards neighboring points that are closer to the boundaries of unintended modes. With this association of latent codes within unique modes, we can proceed to refine Eq. 5 as follows:

$$\arg\min_{w^{(\theta)}} \mathbb{E}_{x \sim \hat{\mathcal{D}}_{real}, \hat{x} \sim \hat{\mathcal{D}}_G^{(\theta)}} \Big[ d(f(x), f(\Phi(G(w^{(\theta)})))) \Big], \quad (6)$$

where $\Phi(G(w^{(\theta)})) = \hat{x}$ simply output the decomposed data of the generated image.

**Rank-and-Ensemble Initialization.** The face reconstructed from the target system should be tested across multiple face recognition systems or PPFR systems to enhance the credibility of the privacy verification results of the target system. However, when there is a disparity between the models used in the target and validation systems, the testing results diminish. To address this issue, we propose a rank-and-ensemble initialization strategy inspired by model averaging protocol that boosts domain generalization [3, 48]. Adversary first generates face images using the latent vectors $w^{(\theta)} \sim \mathcal{W}^+$ of $n$ modes through stylegan generator $G$, which are then input into the target face recognition model $E$. $E$ calculates the similarity score between the extracted feature vectors and the enrolled features in the database and returns it to the adversary. Next, the $n$ similarity scores are sorted, and the latent vectors corresponding to the top-$k$ scores are selected. Finally, the top-$k$ latent vectors are combined through averaging and ensemble methods to create an initial latent vector, whose mode optimally aligns with our objective is determined, s.t. the distance $d(f(x), f(\Phi(G(w^{(\theta)}))))$ is minimum among all $n$ modes. The proposed rank-and-ensemble strategy enhances the privacy inference performance and strengthens the generalization capabilities of Map$^2$V, whether the target and validation systems are identical or differ.

**Zeroth-Order Gradient Estimation.** Under 1k1c assumption, we shall proceed with the optimization of Eq. 6 without accessing the gradient information of (pre-trained) $G$ and $E$ (or $(f\cdot)$). We propose a zeroth-order gradient estimation to achieve a training-free and query-efficient solution. After rank-and-ensemble initialization, an optimal initial latent vector is determined; one straightforward approach to minimize Eq. 6 is gradient decent, denoted as:

$$w_{t+1}^{(\theta)} \leftarrow w_t^{(\theta)} - \alpha \frac{\partial \mathcal{L}}{\partial w_t^{(\theta)}}. \quad (7)$$

Here, $t$ refers to the iteration count, which also corresponds to the number of samples for $\hat{x}_j$ within a particular mode $(\theta)$ (see Eq. 4), $\alpha$ represents a fixed learning rate, and

$$\mathcal{L}(w_t^{(\theta)}) \coloneqq d(f(x), f(\Phi(G(w_t^{(\theta)})))). \quad (8)$$

This approach allows for precise modulation of the latent update using a small parameter $\alpha$. However, under the 1k1c assumption, it is impossible to compute gradients $\frac{\partial \mathcal{L}}{\partial z}$ (denoted $g$ for simplicity). Motivated by zeroth-order optimization used in reinforcement learning, we attempt to approximate the gradient in this black-box setting as follows:

$$\hat{g} = \frac{d}{t} \sum_{j=1}^{t} \frac{\mathcal{L}(w^{(\theta)} + \epsilon u_j) - \mathcal{L}(w^{(\theta)})}{\epsilon} u_j \,, \quad u_j \sim \mathcal{U}(\mathcal{S}^{d-1}), \tag{9}$$

where $d$ is the dimension of the latent code, $u_j$ denotes a perturbation noise sampled randomly from a $d$-dimensional unit sphere $\mathcal{S}^{d-1}$, and $\epsilon$ stands for a small positive constant known as the smoothing parameter.

Based on the information provided earlier (Eq. 4), we can deduce that the attack complexity required for the adversary to achieve the formalized objective (Eq. 6) is *at least* in the order of $O(nmt)$. This provides a lower bound for our reference to privacy validation of the PPFR system. In other words, it implies the minimum number of queries necessary to attain a particular privacy inference performance within our 1k1c framework.

# 4. Experimental evaluation

## 4.1. Experimental setup

**Datasets**: Two widely recognized face recognition benchmarking datasets Labeled Faces in the Wild (LFW) [8] and CelebFace Attributes (CelebA) [19] are used to evaluate the potential of Map$^2$V and the vulnerability of the latest PPFR system in the literature. We sampled 200 individuals randomly for each dataset in the evaluation (i.e.,$m = 200$).

**Configuration**: We use cosine distance as the distance metric in Eq. 6, and then the Adam optimizer with a fixed learning rate of 0.1 for 1k1c black-box scenarios. We perform optimization for 400 iterations (i.e., $t = 400$). For the initialization step, we randomly sample 500 latent vectors (i.e., $n = 500$) from constructed prior spaces and use the ensemble of the top-5 latent vectors as the initial point. Meanwhile, we specify the parameters $\epsilon = 0.1$ and $u = 16$ for gradient estimation. This leads to a total of 6900 query accesses to the PPFR system.

In the identity inference analysis, the reconstructed images from a target system (denoted as $S_T$ for simplicity) are used to impersonate a user in another recognition system (denoted as $S_V$ for simplicity) for validation. $S_T$ and $S_V$ can be either an unprotected FR system (e.g., ArcFace [7], MagFace [21], AdaFace [15]) or a PPFR (e.g., DuetFace, DCTDP, PartialFace, with training configurations and more experiments detailed in the supplementary materials).

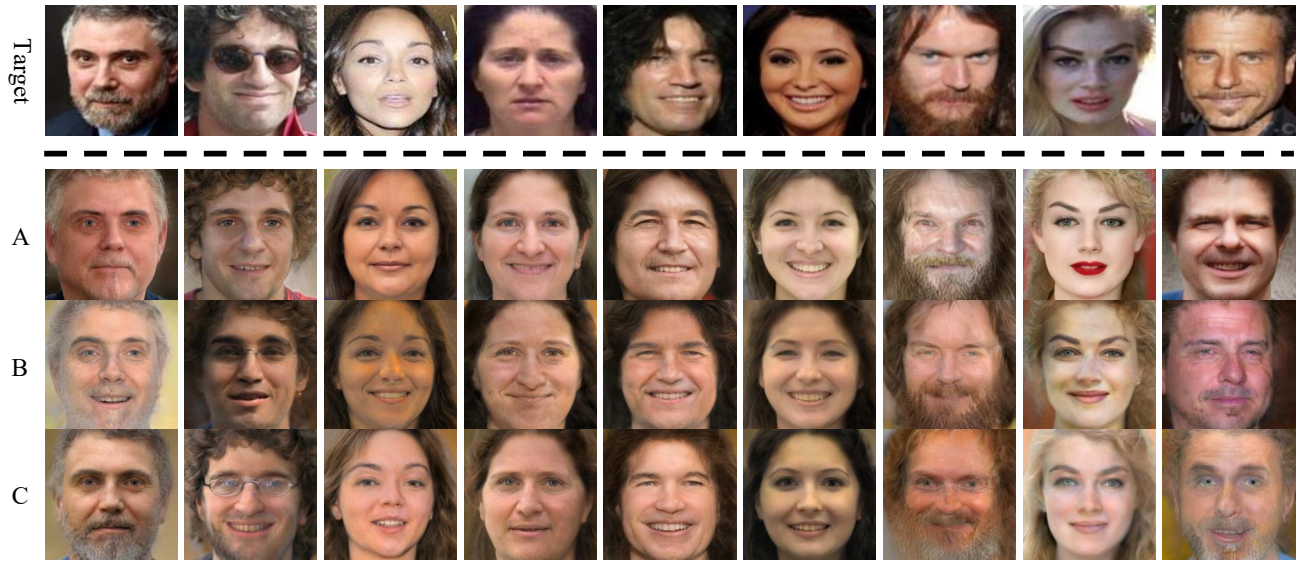
**Evaluation Metrics**: For each generated face image from $S_T$, we compare it with the original image of the same identity and other images of this user using $S_V$. Consequently, we derive two recognition accuracy metrics, termed privacy scores: PSI and PSII. Lower privacy scores indicate reduced facial privacy risks. It is worth mentioning that we set up both 1k1c black-box and white-box scenarios (see supplementary), and also consider the case where the target system and the validation system are the same and different. As the goal is to reveal the user's identity and threaten the security of the system, we also set up human observers for subjective evaluation.

## 4.2. Privacy inference analysis of Map$^2$V

**Reconstruction quality.** Examples of original images and generated face images based on randomly selected subjects in CelebA are shown in Fig. 3. We can observe that the proposed Map$^2$V can generate high-resolution face images based on the 1k1c assumption. We can also observe that personal attributes, such as gender, hairstyle, and skin color, are well preserved in our final inferred face images across existing popular PPFR systems, i.e., DuetFace, DCTDP, and PartialFace.

**Identity privacy analysis.** Table 2 presents the privacy scores for attack images within the state-of-the-art naive FR and PPFR systems in a 1k1c setting. Taking CelebA as an example: 1) when inferring an image from the same PPFR for identical images, the PSI scores are notably high, reaching 97.94%, 96.77%, and 97.79% for DuetFace, DCTDP, and PartialFace, respectively. However, comparing the reconstructed image to the same PPFR but a different image of the same user results in a slightly decreased PSII score, which remains at a concerning level for privacy risks; 2) inferring the image from one PPFR and inferring the identity from another PPFR system shows a decrease in both PSI and PSII scores but still maintains a high privacy risks. For instance, after obtaining a face image from DuetFace, the privacy score of matching this face with other PPFR systems is 92.02% and 86.29% 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 75.32% and 70.82% for DCTDP and PartialFace, respectively. These results showcase that the proposed Map$^2$V is capable of achieving privacy inference under minimal assumptions.

**Visual privacy analysis by machine vision.** We employed a face attribute classifier [12] that was trained on the FairFace dataset, encompassing attributes such as gender, race, and age. The results, showcased in Table 3, reveal that the average attribute matching rate between the reconstructed and target faces surpasses 70% in LFW and 60% in CelebA. This suggests a significant retention of specific target face attributes in our reconstructed faces, potentially disclosing private information. Meanwhile, there is a significant variation in the quality of the dataset, leading to biases in the initial guess of gender when the quality of target images is




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

Figure 3. Examples of reconstructed faces from the CelebA dataset for three SOTAs under 1k1c settings. The first row is the target image for the adversary's attack, A to C shows the results of Map$^2$V.

Table 2. Privacy scores (%) against different validation systems on LFW and CelebA dataset under 1k1c settings.

| Dataset | Attack to $S_V$ | Reconstructed from $S_T$: DuetFace PSI | DuetFace PSII | DCTDP PSI | DCTDP PSII | PartialFace PSI | PartialFace PSII | Average PSI | Average PSII |
|---|---|---|---|---|---|---|---|---|---|
| LFW | ArcFace | 90.25 | 69.26 | 91.75 | 73.46 | 89.50 | 67.46 | 90.5 | 70.06 |
| | MagFace | 82.08 | 70.21 | 86.83 | 75.78 | 90.08 | 73.67 | 86.33 | 73.22 |
| | AdaFace | 81.27 | 65.65 | 81.27 | 69.51 | 88.27 | 68.11 | 83.60 | 67.76 |
| | DuetFace | **97.46** | **82.61** | 93.64 | 76.38 | 91.60 | 72.62 | 94.23 | 77.20 |
| | DCTDP | 88.43 | 71.42 | **96.43** | **88.26** | 88.18 | 69.08 | 91.01 | 76.25 |
| | PartialFace | 82.40 | 66.84 | 85.65 | 73.95 | **98.90** | **93.42** | 88.98 | 78.07 |
| CelebA | ArcFace | 92.48 | 74.71 | 93.73 | 76.53 | 93.98 | 78.43 | 93.40 | 76.56 |
| | MagFace | 91.73 | 74.89 | 90.73 | 77.52 | 94.23 | 82.30 | 92.23 | 78.24 |
| | AdaFace | 85.82 | 68.01 | 85.32 | 71.04 | 92.32 | 75.97 | 87.82 | 71.67 |
| | DuetFace | **97.94** | **82.15** | 94.37 | 78.57 | 93.86 | 80.70 | 95.39 | 80.47 |
| | DCTDP | 92.02 | 75.32 | **96.77** | **85.66** | 90.52 | 78.09 | 93.10 | 79.69 |
| | PartialFace | 86.29 | 70.82 | 87.54 | 74.48 | **97.79** | **89.31** | 90.54 | 78.20 |

poor. This also indicates a weaker correlation between gender and identity features, consistent with face anonymization studies such as Li et al. (2023) and Shamshad et al. (2023) that highlight the ability to modify identity features while maintaining gender attributes.

**Visual privacy analysis by human vision.** If a reconstructed image can fool the system but is visibly different from the target image to human observers, it poses only a partial threat to PPFR. Therefore, we designed a quiz with 30 reconstructed images randomly selected of distinct individuals using the LFW dataset for 100 human observers to assess the subjective privacy (i.e., visual decision of being the same person) between reconstructed images and target identities. To ensure a fair evaluation, we included three real images with the same race, gender, and age group as the target identity as distractors, along with a 'none' option. Overall Subjective Privacy Score (SPS) and Subjective Recognition Rates(SRR) quantify the results of human subjective ratings, with specific details available in the supplementary materials. According to the quiz results shown in Fig. 4, SPS is 93.97%. Among the 30 IDs, SRRs for 27 IDs exceed 90%, with ID-6 and ID-7 reaching 100%. Even the lowest recognition rate attained 84%. Interference options and the 'none' option constitute only a marginal portion. In conclusion, our study showcases that faces reconstructed can also deceive human observers.

Table 3. Face attributes matching rates (%) between reconstructed faces and target faces.

| Dataset | PPFR | Race | Gender | Age | Average |
|---|---|---|---|---|---|
| LFW | DuetFace | 76.50 | 80.50 | 71.00 | **76.00** |
| | DCTDP | 82.50 | 80.50 | 65.50 | **76.17** |
| | PartialFace | 82.50 | 81.50 | 76.50 | **80.17** |
| CelebA | DuetFace | 74.50 | 49.00 | 65.00 | **62.83** |
| | DCTDP | 76.50 | 48.00 | 74.50 | **66.33** |
| | PartialFace | 76.00 | 47.50 | 71.00 | **64.83** |

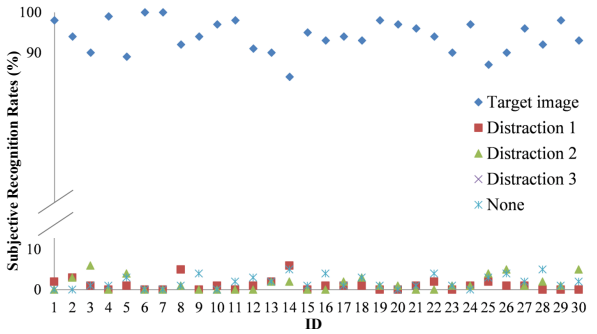
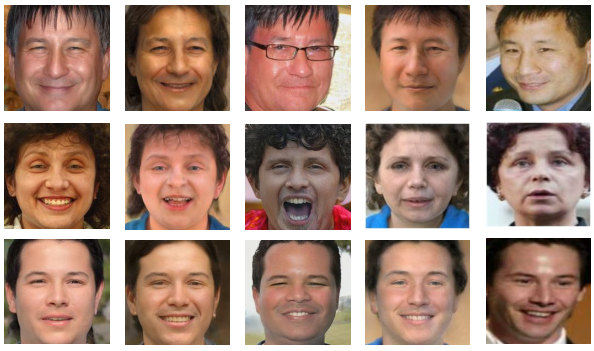



Figure 4. Based on quizzes with 100 individuals, it shows subjective recognition rates for 30 IDs. The scatter of each color represents the distribution of subjective privacy scores for recognizing the target face, three distractors, and 'none.'



(a) w/o PSC&REI (b) w PSC (c) w REI (d) w PSC&REI (e) Target

Figure 5. Visual comparisons on the influence of modules $PSC$ and $REI$. Map$^2$V with $PSC$ and $REI$ modules can achieve a higher degree of similarity to the target faces.

Table 4. Ablation study on $PSC$ and $REI$ of Map$^2$V.

| Module | | Same system | Different system |
|---|---|---|---|
| $PSC$ | $REI$ | | |
| × | × | 97.34 | 79.82 |
| ✓ | × | 97.55 | 75.72 |
| × | ✓ | 97.38 | 84.57 |
| ✓ | ✓ | **97.46** | **84.89** |

### 4.3. Ablation study

Assessing the impact of the $PSC$ and $REI$ modules on privacy validation results, we analyzed the PSI of Map$^2$V with and without each module on DuetFace, using the LFW dataset. Other parameters align with the 1k1c scenario.

**Effectiveness of $PSC$.** Efficiently aligning reconstructed faces with target faces requires reinforcing the prior space. Directly fitting a latent vector from a noise vector space to the target face assumes alignment between face image and noise distributions, often inconsistent in real-world scenarios [42]. To address this, we employ an encoding strategy, transforming public face images into latent space to construct our initial space. In Table 4, the first row displays Map$^2$V's PSI score under 1k1c assumptions without $PSC$ and $REI$, while the second row demonstrates the PSI improvement for using the same PPFR by integrating the $PSC$ module into Map$^2$V.

**Effectiveness of $REI$.** Although introducing the $PSC$ improves PSI within the same system, it reduces generalizability. The PSI decreases for different systems, as seen in the second row of Table 4. In the third row of Table 4, we can observe that the introduction of the $REI$ module increases privacy scores across different system settings, enhancing Map$^2$V's generalizability. Consequently, combining both modules yields the best results in attack scenarios, as demonstrated in the fourth row of Table 4.

### 4.4. Comparison with existing common privacy validation methods

To emphasize the advantages of proposed privacy validation method over native ones from DuetFace, DCTDP, and PartialFace, we conducted a comparison of privacy scores and query counts based on the LFW dataset. The results from Table 5 show that the target PPFR system can achieve a privacy score of over 95% with 6900 queries, indicating a significantly higher degree of face privacy leakage than what the authors had claimed. Our method exhibits significantly lower privacy inference costs. Fig. 6 visually portrays that the native privacy validation method fails to reveal the privacy leakage risks. In contrast, Map$^2$V shows higher similarity to the target image and a more natural appearance, showcasing the high privacy risks.

Table 5. Comparison of privacy scores and query counts for different validation methods.

| Target | **2k2c**[11, 22, 23] | **1k1c**($n = 500$) | **1k1c**($n = 1000$) |
|---|---|---|---|
| **DuetFace** | 96.52 (∼1 million) | 97.46 (6900) | 97.53 (7400) |
| **DCTDP** | 79.60 (∼1 million) | 96.43 (6900) | 96.35 (7400) |
| **PartialFace** | 65.35 (∼1 million) | 98.90 (6900) | 98.89 (7400) |


(a) DuetFace (b) DCTDP (c) PartialFace

Figure 6. Visual comparisons of privacy verification results: the black box for 2k2c assumption, the blue box for our results.

## 5. Conclusion

In this paper, we proposed a general and query-efficient privacy validation approach under a minimum assumption coined Map$^2$V to assess the level of privacy leakage in PPFR systems. We use rich image priors to construct a prior space as a replacement for random initialization from the noise latent space. We then use a rank-and-ensemble strategy to discover an optimal initial vector and optimize it based on the integration of gradient estimation to a common gradient decent technique. Map$^2$V represents the first attempt to validate the privacy of PPFR, achieving efficient reconstruction under the minimal assumption, can serve as a generalizable tool for PPFR assessment and holds the potential for further expansion. Our experiments demonstrated that Map$^2$V not only exposes the vulnerabilities of SOTA PPFRs but can also be employed to validate the privacy protection capacity of naive FR and PPFR systems.

**Acknowledgments.** This work was supported by the National Natural Science Foundation of China (Nos. 62376003) and Anhui Provincial Natural Science Foundation (No. 2308085MF200)

## References


[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 4

[2] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International conference on machine learning*, pages 224–232. PMLR, 2017. 4

[3] Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *Advances in Neural Information Processing Systems*, 35:8265–8277, 2022. 5

[4] Pia Bauspieß, Jascha Kolberg, Pawel Drozdowski, Christian Rathgeb, and Christoph Busch. Privacy-preserving preselection for protected biometric identification using public-key encryption with keyword search. *IEEE Transactions on Industrial Informatics*, 2022. 2

[5] Vishnu Naresh Boddeti. Secure face matching using fully homomorphic encryption. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10. IEEE, 2018. 2

[6] 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. 1, 2

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[8] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 6

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 5

[10] Anil K Jain, Debayan Deb, and Joshua J Engelsma. Biometrics: Trust, but verify. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(3):303–323, 2021. 1

[11] 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, 2, 3, 8

[12] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021. 6

[13] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 4

[14] Mahdi Khosravy, Kazuaki Nakamura, Yuki Hirose, Naoko Nitta, and Noboru Babaguchi. Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system. *IEEE Transactions on Information Forensics and Security*, 17:357–372, 2022. 1

[15] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022. 6

[16] Shu-Min Leong, Raphaël C-W Phan, Vishnu Monn Baskaran, and Chee-Pun Ooi. Privacy-preserving facial recognition based on temporal features. *Applied Soft Computing*, 96:106662, 2020. 2

[17] Anran Li, Hongyi Peng, Lan Zhang, Jiahui Huang, Qing Guo, Han Yu, and Yang Liu. Fedsdg-fs: Efficient and secure feature selection for vertical federated learning. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, pages 1–10, 2023. 2

[18] Yuancheng Li, Yimeng Wang, and Daoxing Li. Privacy-preserving lightweight face recognition. *Neurocomputing*, 363:212–222, 2019. 2

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 6

[20] Zhuo Ma, Yang Liu, Ximeng Liu, Jianfeng Ma, and Kui Ren. Lightweight privacy-preserving ensemble classification for face recognition. *IEEE Internet of Things Journal*, 6(3): 5778–5790, 2019. 2

[21] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021. 6

[22] 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, 3, 8

[23] 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, 3, 8

[24] J Andrew Onesimu, J Karthikeyan, D Samuel Joshua Viswas, and Robin D Sebastian. Security and privacy challenges of deep learning: A comprehensive survey. *Research Anthology on Privatizing and Securing Data*, pages 1258–1280, 2021. 3

[25] Dailé Osorio-Roig, Christian Rathgeb, Pawel Drozdowski, and Christoph Busch. Stable hash generation for efficient privacy-preserving face identification. *IEEE Transactions*

*on Biometrics, Behavior, and Identity Science*, 4(3):333–348, 2021. 3

[26] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7474–7489, 2021. 2

[27] Solon A Peixoto, Francisco FX Vasconcelos, Matheus T Guimarães, Aldísio G Medeiros, Paulo AL Rego, Aloísio V Lira Neto, Victor Hugo C de Albuquerque, and Pedro P Reboucas Filho. A high-efficiency energy and storage approach for iot applications of facial recognition. *Image and Vision Computing*, 96:103899, 2020. 3

[28] Yuwen Pu, Jiahao Chen, Jiayu Pan, Diqun Yan, Xuhong Zhang, Shouling Ji, et al. Facial data minimization: Shallow model as your privacy filter. *arXiv preprint arXiv:2310.15590*, 2023. 1

[29] Joseph P Robinson, Can Qin, Yann Henon, Samson Timoner, and Yun Fu. Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Transactions on Image Processing*, 2023. 1

[30] Roberto Román, Rosario Arjona, Paula López-González, and Iluminada Baturone. A quantum-resistant face template protection scheme using kyber and saber public key encryption algorithms. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2022. 2

[31] Alamgir Sardar, Saiyed Umer, Ranjeet Kumar Rout, and Chiara Pero. Face recognition system with hybrid template protection scheme for cyber–physical-social services. *Pattern Recognition Letters*, 174:17–24, 2023. 2

[32] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20595–20605, 2023. 1

[33] Xin Sun, Chengliang Tian, Changhui Hu, Weizhong Tian, Hanlin Zhang, and Jia Yu. Privacy-preserving and verifiable src-based face recognition with cloud/edge server assistance. *Computers & Security*, 118:102740, 2022. 2

[34] Muhammad Tayyab, Mohsen Marjani, NZ Jhanjhi, Ibrahim Abaker Targio Hashem, Raja Sher Afgun Usmani, and Faizan Qamar. A comprehensive review on deep learning algorithms: Security and privacy issues. *Computers & Security*, page 103297, 2023. 1

[35] 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. 4

[36] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Jiankang Deng, Xinchao Wang, Hakan Bilen, and Yang You. Facemae: Privacy-preserving face recognition via masked autoencoders. *arXiv preprint arXiv:2205.11090*, 2022. 2

[37] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1

[38] Shulan Wang, Qin Liu, Yang Xu, Hongbo Jiang, Jie Wu, Tian Wang, Tao Peng, and Guojun Wang. Protecting inference privacy with accuracy improvement in mobile-cloud deep learning. *IEEE Transactions on Mobile Computing*, 2023. 1

[39] Tao Wang, Yushu Zhang, Ruoyu Zhao, Wenying Wen, and Rushi Lan. Identifiable face privacy protection via virtual identity transformation. *IEEE Signal Processing Letters*, 2023. 3

[40] Yinggui Wang, Jian Liu, Man Luo, Le Yang, and Li Wang. Privacy-preserving face recognition in the frequency domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2558–2566, 2022. 2, 3

[41] Zhousheng Wang, Geng Yang, Hua Dai, and Yunlu Bai. Dafl: Domain adaptation-based federated learning for privacy-preserving biometric recognition. *Future Generation Computer Systems*, 150:436–450, 2024. 1

[42] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2022. 8

[43] Wencheng Yang, Song Wang, Hui Cui, Zhaohui Tang, and Yan Li. A review of homomorphic encryption for privacy-preserving biometrics. *Sensors*, 23(7):3566, 2023. 2

[44] Lin Yuan, Linguo Liu, Xiao Pu, Zhao Li, Hongbo Li, and Xinbo Gao. Pro-face: A generic framework for privacy-preserving recognizable obfuscation of face images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1661–1669, 2022. 1, 2

[45] Yushu Zhang, Tao Wang, Ruoyu Zhao, Wenying Wen, and Youwen Zhu. Rapp: Reversible privacy preservation for various face attributes. *IEEE Transactions on Information Forensics and Security*, 2023. 2

[46] 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, 2023. 1

[47] Yaoyao Zhong and Weihong Deng. Opom: Customized invisible cloak towards face privacy protection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3590–3603, 2022. 1

[48] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 5