Privacy-Enhancing Face Obfuscation Guided by Semantic-Aware Attribution Maps

Jingzhi Li, Hua Zhang, Siyuan Liang, Pengwen Dai, and Xiaochun Cao, *Senior Member, IEEE*

***Abstract*— Face recognition technology is increasingly being integrated into our daily life, e.g. Face ID. With the advancement of machine learning algorithms, the personal information such as age, gender, and race can be easily deduced from the recorded face images in these applications. This poses a serious privacy threat to individuals who do not want to be profiled, as face images are collected for biometric purposes. Existing methods mostly focus on adding the invisible adversarial perturbations into the images to make automatic inference infeasible. However, the application scenarios of these methods are limited due to the perturbations depending on the specific model. In this paper, we introduce a novel face privacy-enhancing framework by obfuscating the stored faces, which could maintain the data utility (face identity) while protecting the privacy of users (facial attributes). Specifically, we first develop a feature attribution module to discover the identity-related facial parts. Within this module, we introduce a pixel importance estimation model based on Shapley value to obtain a pixel-level attribution map, and then each pixel on the attribution map is aggregated into semantic facial parts, which are used to quantify the importance of different facial parts. Next, we design a privacy-enhancing module to generate the high-quality obfuscated images, which can modify the privacy semantic content and preserve the identity-related information. Using the proposed method, users can choose the single or multiple attributes to be obfuscated without affecting identity matching. Extensive experiments conducted on CelebA-HQ and VGGFace2-HQ benchmarks demonstrate the effectiveness and generalization ability of our method.**

***Index Terms*— Privacy preserving, face obfuscation, feature attribution.**

I. Introduction

PRIVACY concerns on face recognition have emerged as a global issue in recent years. Face recognition is increasingly used for biometric authentication, such as unlocking smartphones or authorizing transactions. Different from other biometrics, face images contain more information than individual's identity, e.g. age, gender, race, and health status, which represent facial characteristics with strong inherent stability and individual differences. Recent research has been able to automatically extract and infer these attributes from face images with high accuracy [1], [2]. For example, the accuracy of the state-of-the-art attribute recognition method on the Celeb-A database exceeds 90%. And a series of methods for predicting expressions, physiological signals, and sexual orientation from face images have been proposed.

Although these analyses are useful for identifying individuals, they also bring the considerable privacy risks to the general public. These risks include unauthorized facial analysis of facial recognition data to develop targeted adverting, identity theft arising from facial recognition data breaches, attackers stealing faces to create fake content, and more. Consequently, individuals want to be identified by machines for safety and access purposes, while unwilling to expose their privacy information to intelligent algorithms and human observers. Furthermore, some political regulations have been proposed to restrict biometric data to follow the data minimization principles, e.g. the General Data Protection Regulation (GDPR) [3]. Therefore, developing attribute privacy-enhancing techniques is crucial to the security of face recognition systems.

An intuitive way to protect face privacy is to remove the identity information by breaking the links between the biometric traits and the specific person [9], [10], [11], [12]. For example, we usually use the mosaic to obfuscate the original faces to protect the privacy of users, as illustrated in Fig. 1. However, these methods destroy the verification utility of face images. Recently, there are some works on exploring the privacy protection for face recognition systems [13], [14], [15], [16], [17], [18]. Though the results are impressive, existing methods suffer from two mainly limitations: 1) Weak generalization. Some methods target a specified recognition model [4], [19] or the specified attribute [4], [13] to generate the protected face, which are easily attacked by the unknown face model. 2) Poor image quality. There are several visual artifacts in the protected results [4], [14]. As shown in Fig. 1, the results from SAN model [4] have failed when attacked by the third-part attribute classifier and face recognizer, and have lower image quality. Furthermore, face editing methods [5], [6], [20] could generate the impressive attribute-modified faces. Whereas these methods do not consider the privacy issue and could not be directly used in this task. It has been demonstrated that face images generated by the GAN-based methods would introduce the domain gap [21], which

Manuscript received 20 August 2022; revised 7 April 2023 and 18 May 2023; accepted 22 May 2023. Date of publication 2 June 2023; date of current version 23 June 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100800; in part by the National Natural Science Foundation of China under Grant 62025604, Grant 62132006, and Grant 62072454; and in part by the Beijing Natural Science Foundation under Grant L212004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Walter J. Scheirer. *(Corresponding author: Hua Zhang.)*

Jingzhi Li and Siyuan Liang are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China.

Hua Zhang is with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China, and also with the School of Ocean Information Engineering, Jimei University, Xiamen 361021, China (e-mail: zhanghua@iie.ac.cn).

Pengwen Dai and Xiaochun Cao are with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China.

Digital Object Identifier 10.1109/TIFS.2023.3282384

1556-6021 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

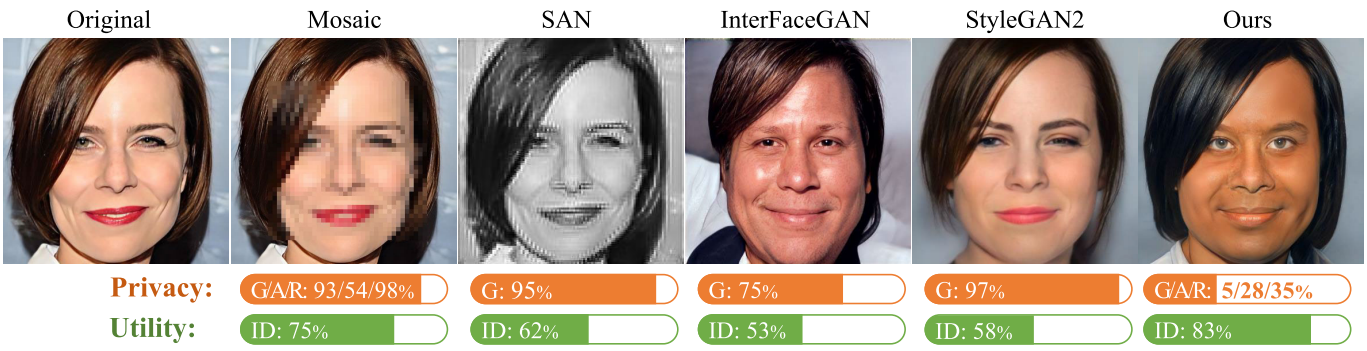


Fig. 1. Privacy-enhancing results compared with existing methods, where Mosaic (50×50) is a traditional method, SAN [4] is an attribute privacy method, InterFaceGAN [5] and StyleGAN2 [6] are face editing methods. The orange bars denote the probability of privacy attributes inferred from the protected face (G: gender, A: age, R: race). The green bars denote the identity similarity between the protected face and the original face. We employ the third-part attribute classifier FairFace [7] and face recognizer CosFace [8] to perform the attribute prediction and face verification, respectively.

influences the utility of generated face images. As shown in Fig. 1, the results of InterFaceGAN [5] and StyleGAN2 [6] have significantly reduced the recognition utility, and have a high attribute prediction accuracy (high risk for privacy leaks).

To solve these issues, we develop a face privacy-enhancing framework that could protect multi-attribute privacy while preserving the recognition utility. For the identity preserving, we need to find the invariant local visual structure of the input face to develop the invariant feature representation. To this end, we introduce the feature attribution map as a bridge to search for the critical local regions for face recognition. While for the privacy attribute protection, we design a novel StyleGAN-based face editing model, which could produce a high visual quality images and protect the attributes, as shown in Fig. 1. In the real scenario, we could deploy the face anonymizer to various applications including face payment terminal, face boarding systems, video surveillance system, etc. When the face images are uploaded to the face database, they will be modified by our anonymizer, making them unavailable for any other facial analysis except the identity matching. Furthermore, our model could also generate a cognition gap between the recorded faces and the real faces, which could not be associated by the human observer.

Specifically, we first develop a feature attribution module to discover the identity related regions for preserving the face recognition utility. To that end, we conduct a Shapley value [22] based pixel importance estimation model to obtain the attribution map. And then, each pixel on the attribution map is aggregated into semantic facial parts, which are used to quantify the importance of different parts on face recognition. Based on the semantic-aware attribution maps, we select the critical regions for face recognition named as ID-sensitive regions. After that, we build a privacy enhancement module to generate a high-quality obfuscated facial image, which could protect the specific attribute and preserving ID-related information. Different from the traditional face editing methods, we introduce three new constraints to ensure the local structure of ID-sensitive regions and their relationship to the identity. A multi-task based loss function is used to optimize the face generator to achieve the privacy protection face images. Extensive experiments conducted on CelebA-HQ [20] and VGGFace2-HQ [23] benchmarks demonstrate the effectiveness and generalization ability of our method on balancing the data utility and the multi-attribute privacy.

Our contributions could be summarized into three folds:

- We develop a general privacy-enhancing framework for face images stored in face recognition systems, which could maintain the data utility (face identity) while protecting the privacy of users (facial attributes).
- We introduce an identity attribution method to discover the ID-sensitive facial regions by estimating the Shapley values. Moreover, we quantitatively evaluate the attribution results of our method on different face recognition models, and the experimental results demonstrate that our method can discover more reliable and accurate regions.
- We propose a novel face obfuscator to modify the privacy semantic content while preserving the ID-sensitive local structure and relations. Furthermore, we also develop an effective optimization policy to constrain the latent codes for face generation.

The rest of the paper is organized as follows: Section II reviews the related works; Section III presents our privacy-enhancing model; Section IV presents the experimental results conducted on different databases; Section V provides the conclusions and limitations.

II. Related Works

A. Face Privacy

1) Face De-Identification: Focuses on preserving facial attributes e.g. gender, age, and race, while removing the identity information from the input images. Early face de-identification works mainly obfuscate face parts through masking, pixelization, blurring, and other methods. However, these occlusion methods would significantly reduce the data's utility. Furthermore, these methods have been shown to be ineffective for deep learning based face recognition [24]. The k-same algorithm-based methods are another representative work. These algorithms exploit the average face of k-closet faces to replace the given face, which make the face recognition accuracy less than 1/k [25]. Many variants [26], [27] are proposed to improve the data utility of average face. Recently, new techniques and mechanisms have been applied to enhance the face privacy. Some researchers implement

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

de-identification by adding adversarial perturbations [15], [28], [29]. Shan et al. [28] proposed a system that alter the image feature representations using perturbations imperceptible to the naked eye.

GANs [30] inspire a new vein of face de-identification techniques, which can be divided into two categories: conditional inpainting-based technique [31], [32], and face representations manipulating-based [9], [11], [33]. Sun et al. [31] generate a realistic head inpainting based on facial landmarks derived from image context. DeepPrivacy [32] employs a GAN-based head inpainting technique to generate obscured faces, which ensure thoroughly remove privacy information from the original face. Gafni et al. [9] generate the high-level representations from face images that minimize identity associations, while keeping the perceptions (pose, illumination and expression) unchanged. CIAGAN [11] leverages a vector to control the fake identity of the generated images. IdentityDP [33] combines differential privacy mechanisms with deep neural networks to achieve the adjustable privacy control. Cao et al. [34] introduce user-specific password and adjustable parameter to achieve invertible de-identification.

These methods decrease the privacy issues associated with identification, but also destroy the convenience and safety for face recognition. In contrast, we hope to maintain the identity utility in face recognition systems, and obfuscate the privacy attributes against third-party applications.

2) Identity-Preserving Face Privacy: Aims to mitigate the privacy concerns in face recognition systems. It tries to prevent other facial information to be used, except for identity recognition. Existing methods could be divided into two groups: the image-level methods and the representation-level methods.

At the image-level, Ross et al. [13] firstly introduce the concept of soft biometric privacy, and propose a face mixing approach which conceal the gender attribute while preserving the recognition utility. To improve the results, they introduce the adversarial perturbation-based method [4], [35]. Subsequently, Chhabera et al. [19] have also adopted the adversarial perturbation technique to achieve the k-facial attributes anonymization. Mirjalili et al. extend the semi-adversarial network to impart multi-attribute privacy to face images [14]. In previous works [36], [37], Li et al. propose the visual appearance obfuscation methods to protect face images that require the verification. Existing methods are mainly based on adversarial perturbations. Though these methods are shown to prevent the automatic mining of facial attributes, some protected results are highly depend on target model, and suffer from artifacts. Furthermore, the minor visual changes into the images could not ensure the privacy with respect to human observers, and there will be existing many privacy risks, for example, malicious AI face swapping without consent. In this work, we propose a high-resolution face obfuscation framework to try to address the above problems.

At the representation-level, some researchers have worked on the privacy of face descriptors [38], [39], [40]. Their goal is to get a facial representation that doesn't contain privacy-sensitive attributes. The facial representation generated by SensitiveNet [38] is eliminated the sensitive information such as gender or ethnicity via a modified triplet loss. Terhörst et al. [39], [40], [41] put forward a series of methods to suppress privacy-sensitive information on the template-level. However, these methods cannot be used to protect the images in face recognition systems, and they are highly depend on face descriptors extraction models. In this work, we propose a privacy-preserving framework to the face images rather than the representations.

B. Face Manipulation

Face manipulation involves modifying the facial attributes such as age, gender, pose, etc. Image-to-image translation algorithms are widely used for the real image editing. Earlier studies [42], [43] usually perform the image translation in two domains. And then, the algorithms for image translation between multiple domains using one model are proposed [44], [45]. To improve the attribute manipulation ability, Li et al. [46] propose a novel framework with controllable ability by disentangling the style code in latent space. Similarly, semantic face editing methods could also modify the target facial attributes. Existing works achieve impressive results by designing loss functions [47] and network architectures [20], [48]. Many works propose various optimization procedures to control the latent representation for diverse image manipulations [5], [49], [50], [51]. InterFaceGAN [5] explore the interpretion of the disentangled face representation, and achieve the controllable manipulation by leveraging the semantic encoded in the latent space. Tov et al. [50] present an encoder for image manipulation, which balances the distortion-editability tradeoff and distortion-perception within the StyleGAN latent space.

Though these methods achieve a good performance in face attribute editing, these methods do not consider the recognition utility of face images. Thus, they cannot be used for our privacy task. In this work, we perform the attribute editing with some specific constraints, whose goal is to modify the information with low relevant to face identity.

III. Proposed Method

Given an original face image I_{org}, our method aims to adaptively generate the privacy-enhancing face image I_{prv}, which have the different facial attributes comparing with the original face while preserving the identity information. As shown in Fig. 2, the proposed framework consists of two modules: ID-sensitive feature attribution and multi-attribute privacy enhancement.

Specifically, 1) ID-sensitive feature attribution: We introduce a feature attribution method based on Shapley value [22] to estimate the importance of different facial regions. The Shapley value is an important solution concept in cooperative games, which measures the contribution of each player. We first compute the Shapely value of each pixel with respect to the specific face model VGGFace [52] to obtain the feature attribution maps. After that, to achieve the semantic-level facial feature attribution, we use the face parser to divide the input image into facial parts, and then evaluate the importance of each part. 2) Multi-attribute privacy enhancement: We develop a novel face generator to obfuscate face image

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

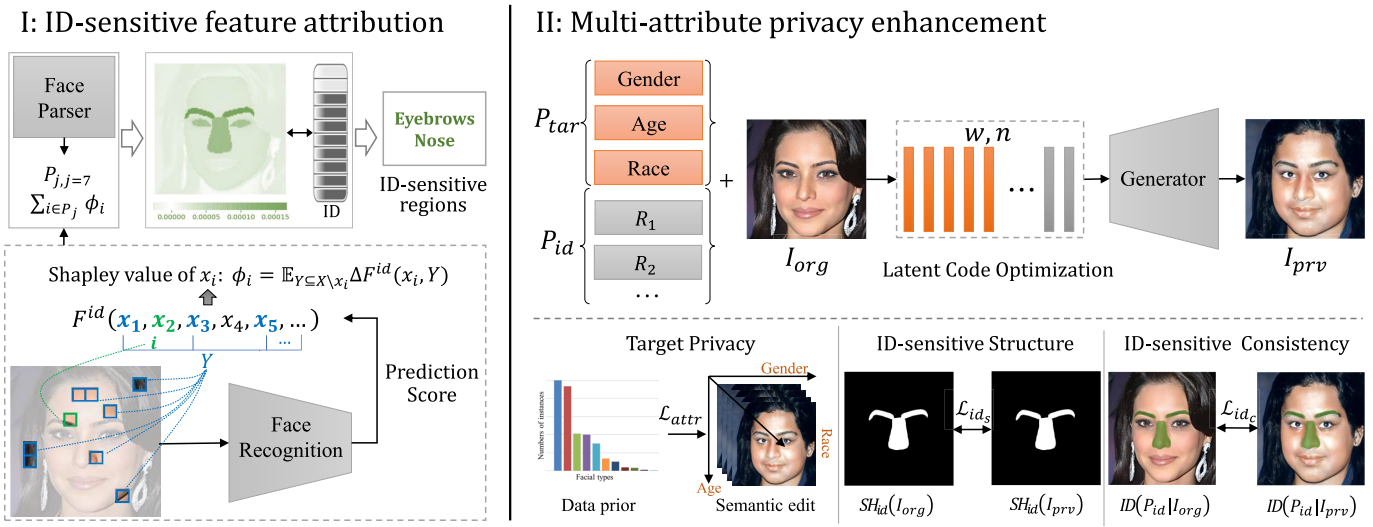


Fig. 2. The overview of the proposed multi-attribute face privacy-enhancing framework. ID-sensitive feature attribution: The goal is to discover the ID-sensitive regions. we quantify the marginal contribution of each facial regions for the face recognition model via estimating the Shapley value. Multi-attribute privacy enhancement: The goal is to generate a high-quality privacy-enhancing image, which could anonymize the specific attributes and preserving ID-related information. Based on the discovered ID-sensitive regions, we design three regularization constrains (target privacy, structure consistency, and identity feature consistency) to optimize the latent code of the stylegan-based generator.

via multi-task optimization in latent space. The optimization objective is to maintain the contribution of the discovered ID-sensitive regions to identity, while preserving the privacy.

A. ID-Sensitive Feature Attribution

In this subsection, we aim to discover the ID-sensitive facial regions for face recognition. In detail, we introduce a model-agnostic feature attribution method named Shapley value [22] to our problem, where the importance of facial features derives from their marginal contribution to the face identity. Different from the traditional feature attribution methods, the face identity attribution is more fine-grained. Traditional methods mainly locate the object region in the image, since the other regions in the image are unrelated to the object. In our problem, most areas of the face image are related to the identity item. Therefore, the results obtained by directly using the traditional methods are not unreliable. The shapely value computation can obtain more rational and accurate attribution results, because it considers the interactions between pixels.

The Shapley value is an important solution concept in cooperative games, which measures the contribution of each player. It is considered a unique unbiased method that can provide a fairly solution to allocate rewards that players earn in a cooperative manner [53]. Consider a game R with m players, where $M = 1, 2, \ldots, m$ denotes the set of all players, $S \subseteq M$ denotes a subset of M, and $R(S)$ denotes the reward earned by players in subset S. The Shapley value assigns each player a contribution value that represents the impact on the game when that player is included. Given a player i, we first obtain all subsets that not include this player, and then calculate the difference between the reward for the subset $R(S)$ and the reward for adding this player to that subset $R(S \cup \{i\})$. The marginal contribution of each player is obtained by the weighted average of all possible differences:

$$\phi(i) = \sum_{S \subseteq M, i \notin S} \frac{(|M| - |S| - 1)!|S|!}{|M|!} [R(S \cup \{i\}) - R(S)], \tag{1}$$

Shapley value has been proved to be the unique reward allocation scheme with the following desirable properties.

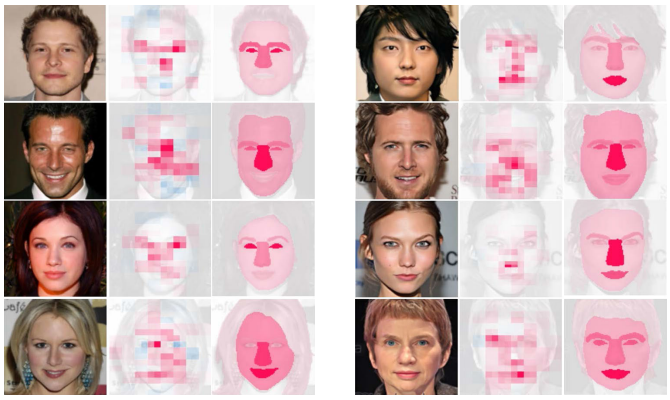
(1) Linearity: if the reward of a game $R(S)$ is linear combination of multiple games reward $R(S) = \sum_{j=1}^{k} a_j R_j(S)$, where a_j denotes the weight and R_j denotes the different games. Then the Shapley value of each player in the game R should also be a linear combination with the same weight $\phi(i) = \sum_{j=1}^{k} a_j \phi_{R_j}(i)$. (2) Nullity: For a subset $S \subseteq M$, if one player satisfies $R(S \cup \{i\}) = R(S)$, then the Shapley value is zero. (3) Symmetry: For all subsets $S \subseteq M$, if two players always make the same contribution $R(S \cup \{i\}) = R(S \cup \{j\})$, then their Shapley values are the same. (4) Efficiency: Summing up Shapley values of all players should be equal to the total reward of the game, $R(M) = \sum_{i=1}^{m} \phi(i)$.

Considering these advantages, we treat each pixel in the face image as a player and the pre-trained face recognition model as a cooperative game. The model output is regarded as the reward. Following Eq. 1, each pixel is assigned an importance value that represents the marginal contribution to the model's prediction when that pixel is included.

In details, let $X = [x_1, x_2, \ldots, x_p]$ be the set of pixels in the face image I_{org}, where p is the total number of pixels. $Y \subseteq X$ denotes the subset of X. And then, we employ the face recognition model $F^{id}(.)$ as the reward function. Therefore, the Shapley value of each pixel is computed:

$$\phi(x_i) = \sum_{Y \subseteq X, x_i \notin Y} \frac{(|X| - |Y| - 1)!|Y|!}{|X|!} [F^{id}(Y \cup \{x_i\}) - F^{id}(Y)], \tag{2}$$

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.



(a) CelebA-HQ (b) VGGFace2-HQ

Fig. 3. Visualizations of ID-sensitive feature attribution maps on CelebA-HQ and VGGFace2-HQ datasets. In each group, the left image is the original input face, the middle image is the recognition attribution map, and the right one indicates the semantic-aware feature importance attribution maps. The intensity of the color denotes the importance for recognition.

where $F^{id}(Y)$ and $F^{id}(Y \cup \{x_i\})$ represent the output probability of face identity id when the pixel x_i in Y is absent or present, respectively.

However, it would be time-consuming and infeasible to directly compute the Eq. 2 on a high-resolution images, which will require $O(2^p)$ evaluations. Sampling or clustering methods are often used when information cannot be obtained from each pixel. Here, we employ a simple pixel clustering method to achieve the superpixel regions. It uses the similarity between pixels to group, and then replaces a large number of pixels with a small number of superpixels to express image features, which greatly reduces the complexity of our Shapely value computation. In detail, we calculate the color difference between each pixel and its neighbors, and then combine them if they are within a threshold. The calculation stops when the distance value exceeds the predefined threshold, resulting in superpixels: $X^s = [x_1^s, x_2^s, \ldots, x_K^s]$, where K is the number of superpixels, and the value of K is obtained by the experiment. Next, each superpixel x_i^s is treated as one player used for estimating the Shapley value, which greatly reduces the complexity of computation.

As our goal is to discover the identity sensitive region on the given face image, we adopt the face parser [54] to segment the face image into semantic facial parts. The semantic parts consist of eyebrows, eyes, nose, mouth, skin, ears, and hair. Then, we sum all the shapley values in the facial part as the contribution of that region on predicting the identity. After that, we sort them in a descending order. Finally, we can get the corresponding facial parts through ID-preserving threshold, which used as the ID-sensitive features in the next.

The experimental results on two datasets are shown in Fig. 3. These portraits demonstrate that the ID-senstive facial parts are not consistent across different individuals. Some individuals' eyes and nose are highly activated, while the others are the mouth.

B. Multi-Attribute Privacy Enhancement

In this subsection, we present a novel face obfuscation model, which could protect the privacy attributes and preserve the identity information for the utility. The proposed face obfuscation model follows a common face semantic editing framework. In particularly, we design the novel semantic, spatial, and consistency regularization terms to achieve the high fidelity face images. Furthermore, we fuse the statistical prior into the processing of computing the latent code, which improves the performance of attribute privacy protection.

The architecture of the face obfuscation model is illustrated in Fig.4. We first project the input image I_{org} into the GAN latent space to generate the latent code w. Next, the latent code is optimized according to the target attribute semantics P_{tar} and the preserved facial regions (the ID-sensitive features) P_{id}, that is, $\psi : w \rightarrow w^*$. Finally, the privacy-enhancing image I_{prv} is generated from the optimized latent code by the generator G, that is, $I_{prv} = G(w^*)$. Specifically, there are several components used in our proposed model: 1) A stylegan-based generator G that produce the high-resolution privacy-enhancing facial images. 2) An attribute classifier C that enforces to change the specific attributes. 3) A face parser P that imposes spatial constraints to preserve the shape of ID-sensitive regions. 4) An identity feature encoder E that implements contribution consistency of ID-sensitive regions to the identity. The critical step is to optimize the latent code with these regularization constraints. We next give the detail description on each term.

1) Target Privacy Constraint: The obfuscated face image is expected to preserve users' privacy attributes to prevent the automatic information extraction. This regularization constraint is designed to anonymize the facial attributes of the original face. Therefore, we introduce the Kullback-Leibler divergence between the target attribute prediction and the smoothed ground truth probability distribution as the regularization.

$$\mathcal{L}_{attr} = \frac{1}{K} \sum_{k=1}^{k} \Delta_{KL}(C_k(P_{tar}^k | I_{prv}), \eta_k), \tag{3}$$

where Δ_{KL} is the Kullback-Leibler divergence, K is the number of target privacy attributes, P_{tar}^k denotes the target attribute (the different label to the groundtruth), $C_k(.)$ indicates a probability of the generated image belonging to the k-th target attribute, η_k denotes ground truth probability distribution of the target attribute. The value of η_k could be used to set the strength of the privacy attribute for distinct protection levels, which is also considered as the data prior.

2) ID-Sensitive Structure Constraint: The face obfuscation model aims to preserve the ID-sensitive regions and their importance in identity. However, the facial attributes we want to conceal are related with the entire image rather than a specific facial region. For example, the gender would be related with multiple facial regions e.g., mouth, eyebrow, and nose. Therefore, we can not use the traditional face manipulation method [55] to alter the specific image semantics. Thus, we propose to preserve the structure of ID-sensitive facial region to produce convincing manipulations of input images. This not only preserves the recognizability of the identity, but also increases the flexibility of image modification. We introduce to use the shape of ID-sensitive facial regions as a bridge to preserve the identity features. The ID-sensitive structure

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

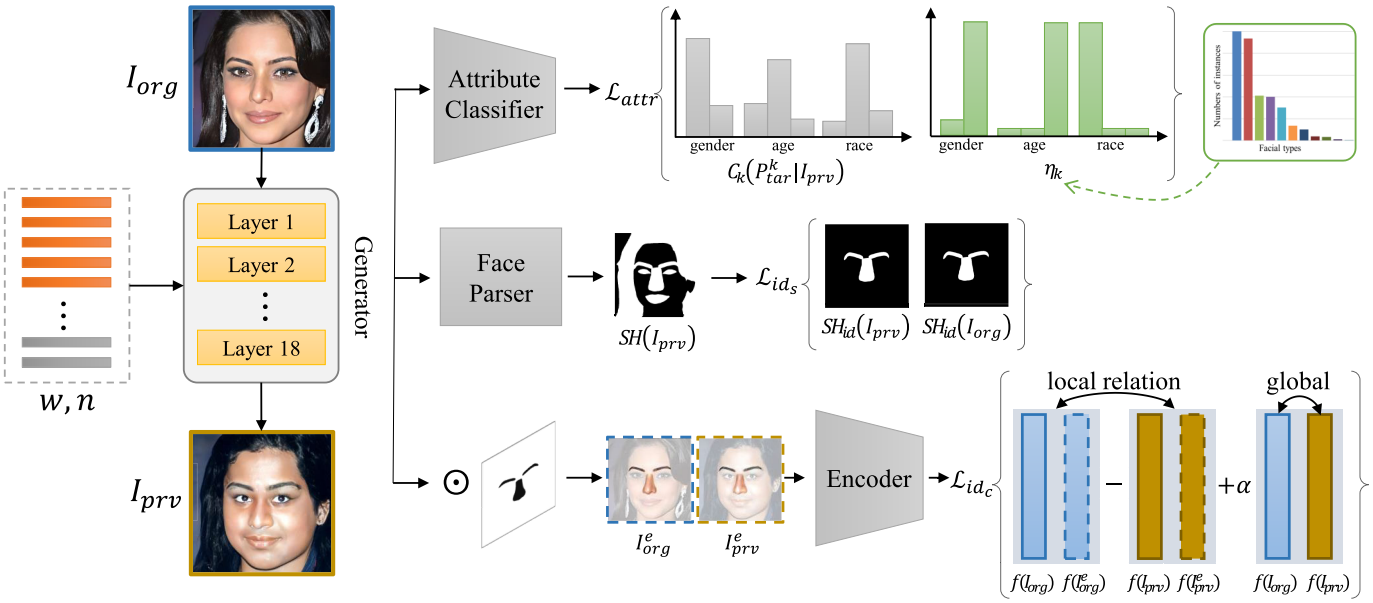


Fig. 4. The architecture of face obfuscation model. Following the traditional optimization processing of StyleGAN2 [6] to achieve the latent code w. Then, we use the target privacy constraint to maintain the probability distribution between the target-attribute prediction and the ground truth. At the same time, we use the ID-sensitive structure and consistency constraints to preserve the spatial consistency of these regions in pixel space, and the identity importance consistency of these regions in feature space.

constraint is defined as a masked Euclidean distance:

$$\mathcal{L}_{id_s} = \Delta[SH_{id}(I_{org}), SH_{id}(I_{prv}], \tag{4}$$

where Δ is the averaged pairwise Euclidean distance, SH_{id} is a probabilistic mask of the ID-sensitive spatial region produced by the face parser P, and $I_{prv} = G(w^*)$. This term ensures that the generated image I_{prv} has the similar shape with the input image I_{org} in ID-sensitive areas.

3) ID-Sensitive Consistency Constraint: To preserve the identity discriminability, we not only maintain the structure of the ID-sensitive facial regions, but also ensure their contribution to identity. The common practice is to embed an identity importance estimation module into the framework. However, this method cannot be implemented in iterations due to the time-consuming. To that end, we design a novel iterations ID-sensitive consistency constrain to maintain the contribution of the ID-sensitive regions for face recognition. This optimization objective contains two components: the local relation matching loss and the global identity matching loss. The local relation matching loss enforces that the contribution to the identity of the specific regions is consistent in the input image and the protected image. The global identity matching loss constrains the consistency of the identity representations between the input image and the protected image.

Specifically, we first erase the non ID-sensitive regions from the input image I_{org} and corresponding generated image I_{prv} to achieve the erased images I^e_{org} and I^e_{prv}, respectively.

$$\begin{aligned} I^e_{org} &= SH_{id}(I_{org}) \odot I_{org}, \\ I^e_{prv} &= SH_{id}(I_{prv}) \odot I_{prv}, \end{aligned} \tag{5}$$

where $\odot$ is the element-wise product. For the four images, we extract the feature representations via an pre-trained encoder E. Finally, the ID-sensitive consistency between input images and generated images is constrained based on the matching of their local features correlation and global features. The optimization objective is defined as:

$$\begin{aligned} \mathcal{L}_{id_c} = \langle \Delta_{cos}[f(I_{org}), f(I^e_{org})], \Delta_{cos}[f(I_{prv}), f(I^e_{prv})] \rangle \\ + \alpha \Delta_{cos}[f(I_{org}), f(I_{prv})], \end{aligned} \tag{6}$$

where $\langle ., . \rangle$ is the absolute value, Δ_{cos} is the cosine distance, $f(.)$ is the normalized image feature after applying the encoder E, α is the trade-off coefficient between the local relation matching loss and the global identity matching loss.

4) Total Objective: The objective function is defined as a linear combination of the above sub-optimization objectives:

$$\mathcal{L}_{total} = \lambda_{attr}\mathcal{L}_{attr} + \lambda_{id_s}\mathcal{L}_{id_s} + \lambda_{id_c}\mathcal{L}_{id_c}, \tag{7}$$

where λ_{attr}, λ_{id_s}, and λ_{id_c} are the weighting parameters. The optimization process is described in Section IV-C

In Fig.5, we present the face sequences generated by the face obfuscation model on two datasets. We could observe that our method can generate nature-realistic images, which have significant visual variations while maintaining the consistency in ID-sensitive regions. Due to the non-stationary nature of the facial aging pattern, the age groups do not fully capture the non-linearity in texture changes. This results in a slight changing in the appearance. However, these results can prevent the age predictor from making predictions.

IV. Experiments

In this section, we conduct a series of experiments to evaluate the effectiveness of the proposed framework. The experiments could be divided into five parts to analyze different aspects of our task. First, we validate the effectiveness of privacy protection and data utility. For multi-attribute privacy protection, we analyze the performance of protected images

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.



Fig. 5. Face obfuscation results with different combination of privacy attributes on CelebA-HQ and VGGFace2-HQ datasets. From left to right: (a) the original images, (b) the corresponding attribution feature maps via the Shapley value, and the high ID-sensitive regions obtained by thresholding, (c) face sequences generated by our method. The annotations represent the protected attributes.(G: gender, A: age, R: race).

on two attribute classifiers. For data utility, we analyze the performance of protected images on three face recognizers, as well as the realism and quality. Then, we explore the contribution of various components through ablation studies. For feature attribution module, we systematically analyze the impact of this module on identity-preserving performance. For privacy enhancement module, the impact of three main loss terms on privacy-preserving performance is analyzed. Finally, we conduct a user study to evaluate the effectiveness of this method on human observers.

A. Experimental Setup

1) Datasets: 1) CelebA-HQ [20] is a high-resolution facial image dataset. It contains 30,000 images selected from the CelebA with different demographics such as age, gender, and ethnicity. The size of each image is 1024×1024. Each image is annotated with the identity and 40 attributes. We randomly select 2,000 images for quantitative evaluations, and the remaining images are used as the training set. 2) VGGFace2-HQ [23] is another high-resolution facial image dataset, derived from the VGGFace2 [52]. It contains 1.11 million images of 8,631 identities that vary widely in pose, age, illumination, ethnicity and profession. The size of each image is also 1024×1024. We randomly select one image for each identity from the test set, a total of 5,000 images.

2) Comparison Methods: We design three groups of typical task-related baselines, which have been widely applied in this literature. The experimental settings of these baselines are presented as: The soft-biometric privacy method SAN [4] is selected, which aims to protect the gender information. The StyeGAN-based face editing methods are also employed, which aim to edit specific attributes on face images. We select three typical approaches, StyleGAN2 [6], MaskFaceGAN [55], and InterFaceGAN [5] as the baselines. Furthermore, traditional privacy-preserving methods Mosaic and DeepPrivacy [32] are also selected as baselines. Mosaic perturbs the raw image by adding an anonymization filter on

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

the detected face region. This scheme is widely used in the real life to protect user privacy. DeepPrivacy [32] achieves the privacy-preserving by generating arbitrary faces using pre-trained face generators to replace the original faces.

B. Evaluation Metrics

To validate the effectiveness of our method, we use two kinds of metrics: the privacy metrics and the utility metrics.

1) Privacy Metrics: 1) The attribute protection rate is used to measure the attribute privacy protection, which is defined as the ratio in the attribute classification accuracy of the protected images compared to that of the original images. The protection rate of multiple attributes is defined as the percentage of correct predictions for all attributes at the same time. To validate the generalization of our generated privacy-preserving model, we employ three attribute classifiers (ResNet-50 [56], FairFace [7], and UTK attribute classifier [57]). 2) The portrait protection rate is used to measure the appearance privacy protection, which is defined as the proportion of faces judged by human observers to be different from the original faces. In detail, we present the original images and the protected images to the workers similar to [58]. For each comparison, we randomly select different workers to judge if the given image pairs have the same identity.

2) Utility Metrics: 1) We employ the True Accept Rate (TAR) value at fixed False Accept Rate (FAR) value to evaluate identity utility for the face model. we compute the TAR values under different FAR values. Similar to the attribute privacy protection in real-world applications, we conduct the experiments on five face recognizers: ArcFace [59], CosFace [8], VGGFace [52], Glint360k [60], and MegaFace [61]. 2) Fréchet Inception Distance (FID) [62] is used to evaluate the performance on the realism of protected images. A lower FID score denotes that the quality of generated images is higher and closer to the real ones.

C. Implementation Details

1) ID-Sensitive Feature Attribution: utilizes two components: 1) The face parser P is based on the pretrained model [54]. The segmentation mask includes seven facial parts: eyebrows, eyes, nose, mouth, skin, ears, and hair. The face parser is finetuned on the selected datasets using weighted cross-entropy as the objective function. 2) The face recognition model is based on VGGFace2 [52]. We trained the model for 100 epochs on the selected datasets with batch size 20, and its accuracy can reach 96.8% on CelebA-HQ and 91.5% on VGGFace2-HQ.

2) Multi-Attribute Privacy Enhancement: utilizes four components: 1) The face generator G is based on the state-of-art StyleGAN2 model [6]. To improve reproducibility, we use the official version of StyleGAN2 trained on the FFHQ dataset. 2) The attribute classifier C is based on the ResNet-50. We jointly train the model on multiple attributes of the selected datasets to exploit the shared embedding space. The model is optimized with the average weighted cross entropy loss functions for the corresponding attribute groups. The training data is augmented with random horizontal flips and affine transformations to prevent over-fitting. 3) The face parser P is the same as the model used in the ID-sensitive feature attribution module. 4) The encoder E is the face recognition network used in the ID-sensitive feature attribution module after removing the fully connected layer.

The latent code optimize with several steps: First, we initialize the latent code w with the mean face latent code computed from 10,000 samples [6]. The noise components n is initialized with a standard normal distribution. Second, the latent code is optimized according to the reconstruction loss, which requires the generated face should be same with the original face. The inversion latent code could be achieved, same as styleGAN2. Third, All the loss terms are added to enforce the local structure and semantic constraints. We optimize the latent code according to Eq. 7. Similar to [63], w and n are optimized separately. We first optimize the w with $2,500$ iterations, and then w and n are jointly optimized 500 iterations. In the training process, the learning rate of the latent code w is set to 0.0004, and the learning rate of the noise component n is set to 0.01. The parameters of Adam optimizer set to $\beta 1 = 0.5$ and $\beta 2 = 0.999$. To generate comprehensive results for both privacy protection and data utility, the weighting values in Eq. 7 through hyper-parameter optimization are: $\lambda_{attr} = 0.1$, $\lambda_{id_s} = 10$, $\lambda_{id_c} = 1$. The parameter of Eq. 6 is set to $\alpha = 0.02$. The default value of Eq. 3 is set to $\eta = 0.05$ for the gender and to $\eta = 0.025$ for the other attributes. To achieve a desirable face, we set the number of iterations as $3,000$. We also want to use our proposed model on MindSpore, which is a new deep learning computing framework. These problems are left for future work.

D. Evaluation on Multi-Attribute Privacy

In this task, we consider to protect seven types of attributes: gender, race, age, and the combination of these attributes. For a given image, we can produce seven different outputs. To validate the performance on different scenarios, we conduct the attribute prediction using three attribute classifiers. The attribute classifier ResNet-50 [56] is trained on the VGGFace2 dataset [52] with attribute annotations. The other is the state-of-art attribute classifier FairFace [7] trained on their own dataset. The UTK attribute classifier [57] is pretrained on UTKFace dataset [64]. Note that the FairFace [7], and UTK attribute classifier [57] are not used during training. The performance of different classifiers on the original and protected images is measured using the attribute prediction rate. Fig. 6 shows the quantitative results of our method and baselines on ResNet-50 [56] and FairFace [7]. We can observe that the proposed model achieves the best performance on concealing facial attributes. The results of the proposed method on three attribute classifiers are shown in Table I, which demonstrate the generalization on protecting the privacy attributes. The abbreviations in the table represent the selected attributes: G: gender, A: age, and R: race.

1) Comparison With Attribute Privacy Methods: We present the protected results produced by our method and five attribute privacy methods in Fig. 8, and the quantitative results are

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

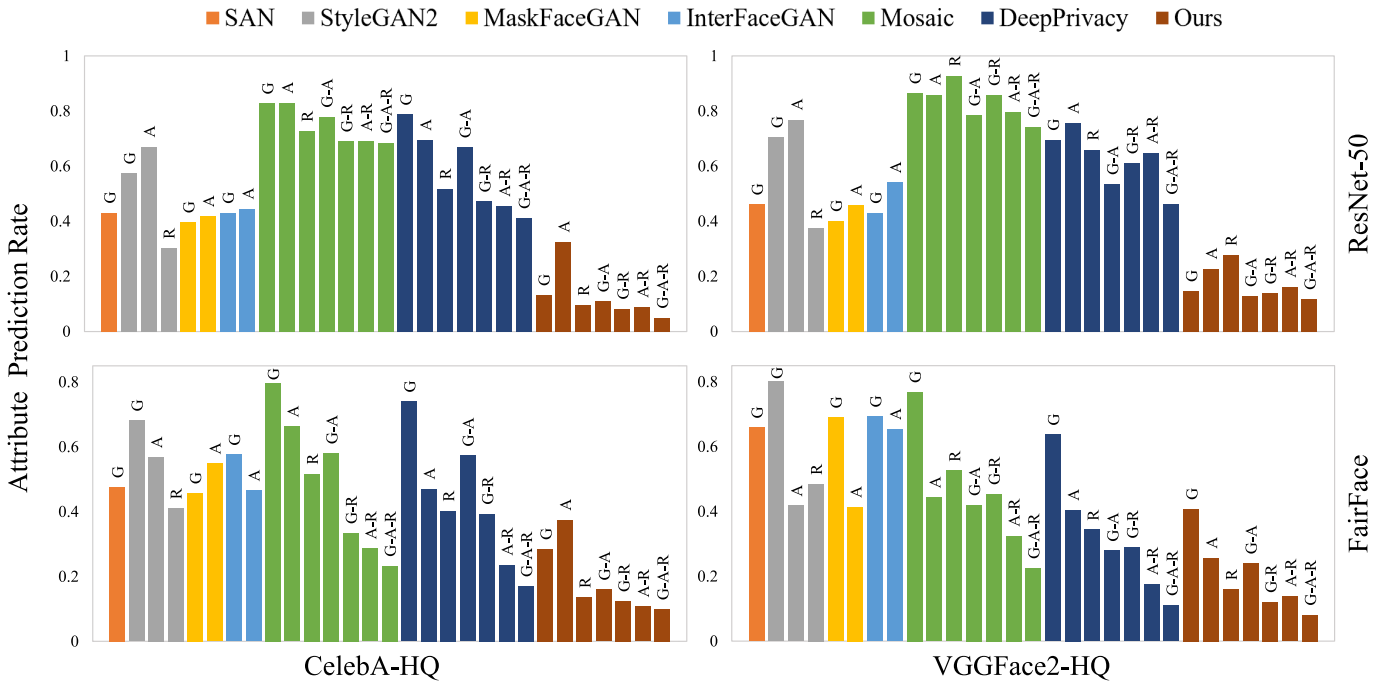


Fig. 6. Attribute privacy-enhancing performance on CelebA-HQ and VGGFace2-HQ datasets under two attribute classifiers (ResNet-50 [56] and FairFace [7]). Different outputs are marked by the selected attributes: G: gender, A: age, and R: race. The lower the value, the better the protection performance. Best viewed zoomed-in.

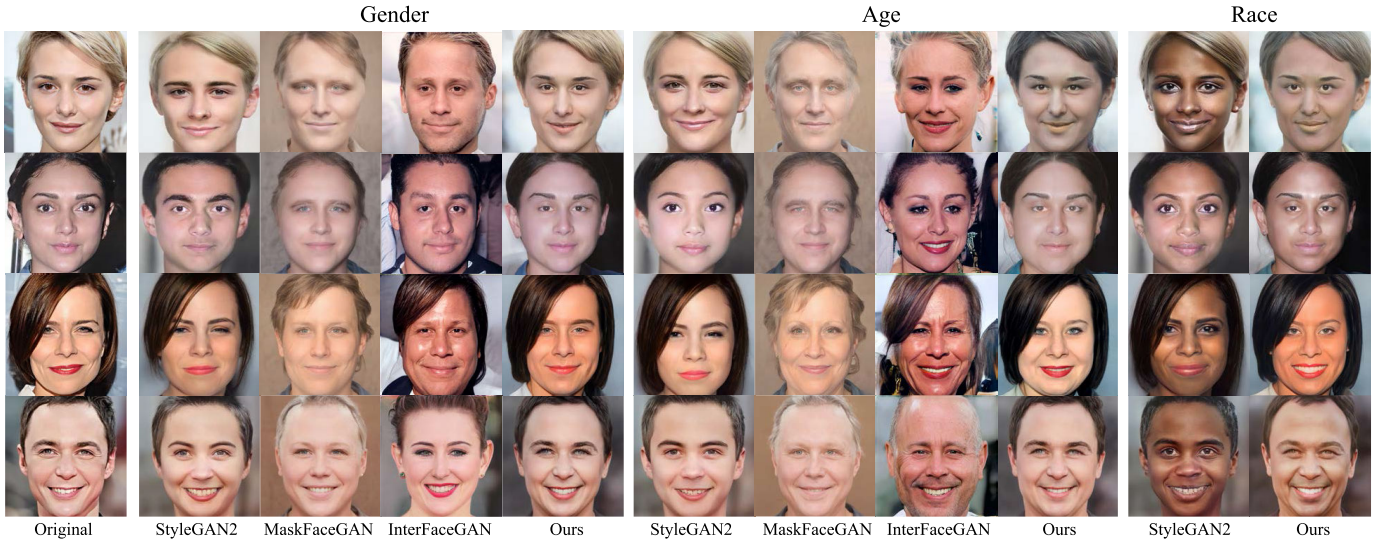


Fig. 7. Qualitative comparison with face manipulation methods. The first column shows the original face, and the rest columns present the visual results generated by StyleGAN2 [6], MaskFaceGAN [55], InterFaceGAN [5], and our method. The superscript indicates the targeted privacy attribute.

TABLE I

ATTRIBUTE PREDICTION PERFORMANCE ON TWO DATASETS UNDER THREE ATTRIBUTE CLASSIFIERS

Dataset	Attribute classifier	G	A	R	G-A	G-R	A-R	G-A-R
VGGFace2-HQ	ResNet-50	0.15	0.23	0.28	0.18	0.18	0.37	0.12
	FairFace	0.41	0.26	0.16	0.24	0.15	0.14	0.08
	UTKFace	0.23	0.27	0.20	0.23	0.29	0.19	0.13
CelebA-HQ	ResNet-50	0.13	0.49	0.10	0.19	0.08	0.09	0.05
	FairFace	0.28	0.38	0.14	0.16	0.15	0.14	0.10
	UTKFace	0.23	0.25	0.35	0.22	0.25	0.11	0.18

displayed in Fig. 6. Among them, SAN [4] is tested on VGG-face 2 dataset, and the visualization results of other methods are obtained from the original paper because the codes are not public. From the quantitative results, our method performs better in all types including the gender (G, G-A, G-R, G-A-R), compare with SAN (orange bar, gender only). Moreover, we could observe that our method achieves significant improvements in visual effects: 1) The higher quality protected results can be obtained; 2) The attribute privacy protection on visual is more effective. Which showcases a success in facial appearance anonymization and a significant improvement on confusing human observers.

2) Comparison With Face Manipulation Methods: As our privacy-enhancing task is to protect specific attributes, face manipulation methods are a feasible way. We choose the

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

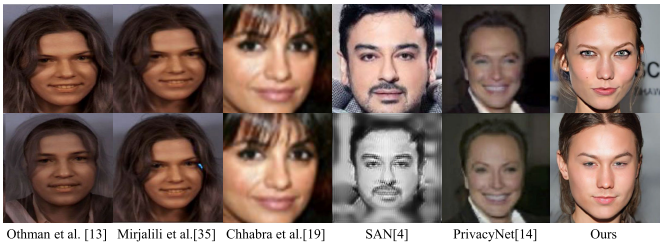


Fig. 8. Qualitative comparison with attribute privacy methods. The top row shows the original face, and the second row shows the privacy-enhancing face.

state-of-art face manipulation methods (StyleGAN2 [6], MaskFaceGAN [55], and InterFaceGAN [5]) as the baselines. Due to the nature limitation of these baselines, we adaptivly compare our model with these models. As can be seen from Fig. 6, although the attribute prediction rates of these methods are reduced, our method performs better with each combination of attributes. Our attribute protection rate is more than 20% higher than the comparison methods.

Fig. 7 compares protected results produced by our method and three face manipulation methods on several sample images from VGGFace2-HQ dataset. From the results, we can find that our method outperforms some methods in facial attribute editing, and can produce visual results of comparable quality to the comparison methods. Actually, it is challenged to maintain a high recognition performance while obfuscating facial attributes. These methods aim to edit face attributes without considering face recognition. Therefore, combined with the results in Fig. 10, these methods can generate faces that look extremely like the original one, while face recognizer still judge them incorrectly, which suggests that the information related to face identity is modified.

3) Comparison With Traditional Privacy-Preserving Methods: Two traditional methods are implemented. One is the Mosaic, where we perform pixelation operation with a size of 50×50 in the face area. The other is DeepPrivacy [32], which generates de-identification images considering the original pose and image background. We compute the attribute protection rates for all attribute combinations. From Fig. 6, we can find that our method significantly improves the privacy-preserving performance compared with these methods. Fig. 9 shows the visual results, which demonstrate that our method have a remarkable improvement in visual quality and realism.

E. Evaluation on Data Utility

To evaluate the performance on data utility for face recognition systems, we conduct two experiments: face verification, and image quality assessment. For face verification experiments, we adopt five face recognizers to match original faces with the protected faces. ArcFace [59], CosFace [8], and VGGFace [52] are trained on the VGGFace2 dataset [52]. ResNet50 is trained on Glint360k dataset [60] and MegaFace dataset [61] respectively. Note that these models are not used during training. Specifically, on CelebA-HQ dataset we randomly selected $1,000$ same identities' face pairs and $2,000$ different identities' face pairs to compute the average

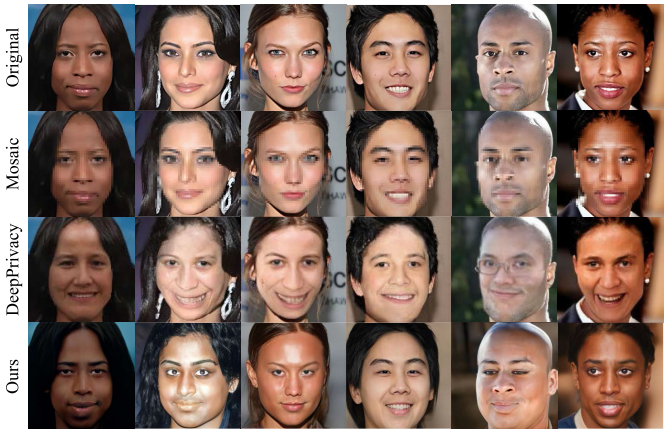


Fig. 9. Qualitative comparison with traditional de-identification methods. From top to bottom: the original faces, the faces obfuscated by Mosaic (50×50), the faces impainted by DeepPrivacy [32], our results.

TABLE II

FACE VERIFICATION PERFORMANCE ON TWO DATASETS UNDER FIVE FACE RECOGNIZERS. THE TAR IS EVALUATED AT FAR $= 0.001$

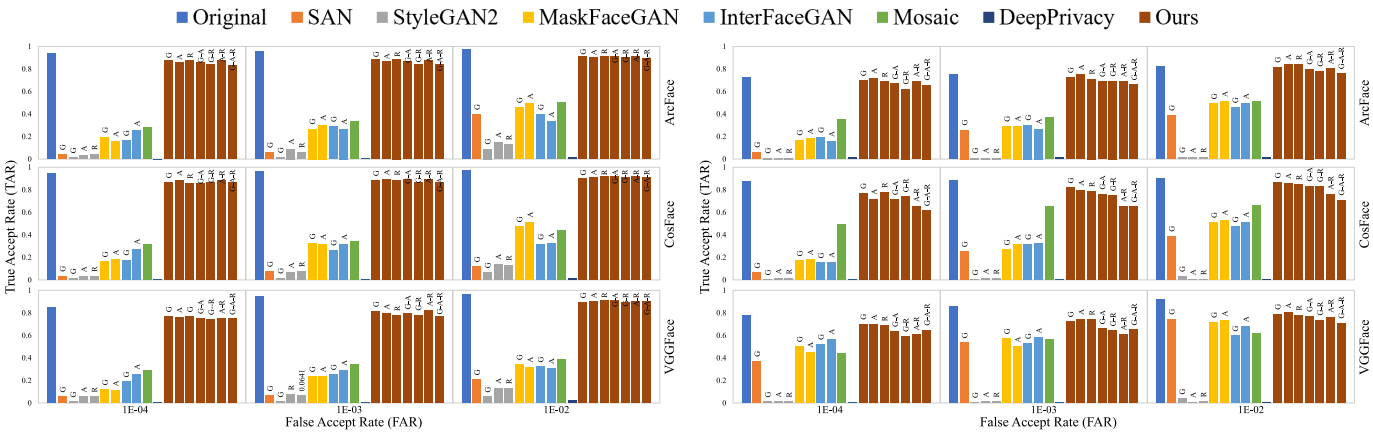
Dataset	Face recognizer	G	A	R	G-A	G-R	A-R	G-A-R
VGGFace2-HQ	ArcFace	0.72	0.75	0.71	0.69	0.69	0.69	0.66
	CosFace	0.82	0.80	0.78	0.76	0.75	0.65	0.65
	VGGFace	0.72	0.74	0.74	0.66	0.64	0.60	0.65
	Glint360K	0.75	0.78	0.75	0.78	0.76	0.79	0.77
	MegaFace	0.74	0.78	0.75	0.79	0.74	0.74	0.75
CelebA-HQ	ArcFace	0.89	0.87	0.89	0.87	0.84	0.88	0.84
	CosFace	0.88	0.89	0.89	0.89	0.87	0.89	0.87
	VGGFace	0.81	0.79	0.78	0.79	0.78	0.82	0.77
	Glint360K	0.68	0.72	0.69	0.64	0.63	0.62	0.61
	MegaFace	0.70	0.73	0.69	0.62	0.63	0.64	0.62

accuracy of face verification by the face recognizers. While on VGGFace2-HQ dataset, we randomly selected $2,000$ same identities' face pairs and $3,000$ different identities' face pairs to develop the testing list.

Similar to attribute privacy evaluation, we also compare our method with the same baselines. Our model generates 7 images for each input face according to the combination of privacy attributes. While other methods produce different attribute protection results depending on the method characteristics. Hence, we compute the True Accept Rate (TAR) values for different cases. Fig. 10 shows the quantitative results. We can observe that the TAR of our method is very close to the TAR obtained from the original images of each dataset. Compared with the baseline results, our method improves the average TAR by 30%-50%. In Table II, we report the TAR@FAR $= 0.001$ scores of the proposed method on five face recognizers. The abbreviations in the table represent the selected attributes: G: gender, A: age, and R: race. We can observe that the proposed method achieves the satisfactory performance under distinct face recognition scenarios.

We adopt FID [62] to evaluate the visual quality and realism of generated images. FID is sensitive to visual artifacts, which could indicate the realism of generated images. From the experimental results in Table III, we observe that our model achieves an expected performance comparing with related face editing model. Although face generation is not our main goal, the results illustrate that our method is able to produce high quality images.

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.



(a) CelebA-HQ (b) VGGFace2-HQ

Fig. 10. Identity-preserving performance on CelebA-HQ and VGGFace2-HQ datasets under three face recognizers (ArcFace [59], CosFace [8], and VGGFace [52]). Different outputs are marked by the selected attributes: G: gender, A: age, and R: race. Best viewed zoomed-in.

TABLE III

QUANTITATIVE COMPARISON ON IMAGE QUALITY OF SYNTHETIC IMAGES BETWEEN OUR MODEL AND COMPARISON METHODS

Method	CelebA-HQ	VGGFace2-HQ
SAN [4]	206.85	243.30
StyleGAN2 [6]	49.53	67.55
MaskFaceGAN [55]	35.50	36.54
InterFaceGAN [5]	79.79	79.76
Mosaic (50x50)	126.45	128.95
DeepPrivacy [32]	156.58	154.86
Ours	41.64	54.04

Furthermore, we present the visual results for protecting different combinations of privacy attributes on two datasets, as shown in Fig. 11. For each pair, we compute the attribute prediction accuracy and the identity similarity. The results show the generalizability of the proposed method under distinct face recognizers and attribute classifiers.

F. Ablation Study on ID-Sensitive Feature Attribution

To explicitly understand the role of the ID-sensitive feature attribution module in our proposed framework, we design some experiments to explore the impact of this module on identity-preserving performance.

First, we analyze the performance of the proposed method and the traditional feature attribution method RISE [65] on identity-preserving. We conduct the qualitative and quantitative experiments. As shown in Fig. 12, we can observe that for different individuals, the important regions obtained by RISE are always the eyes and nose, which is consistent with this method focusing on the center of the object. However, our method can achieve more discriminative results due to the consideration of relationship between pixels. As shown in Table IV, the protected images based on the RISE results receive poor performances on identity-preserving. The TAR@ FAR = 0.001 of RISE is approximately 0.4 under three face recognizers, while ours is about 0.8. This indicates that our method can successfully locate the most distinguishable facial part of the identity.

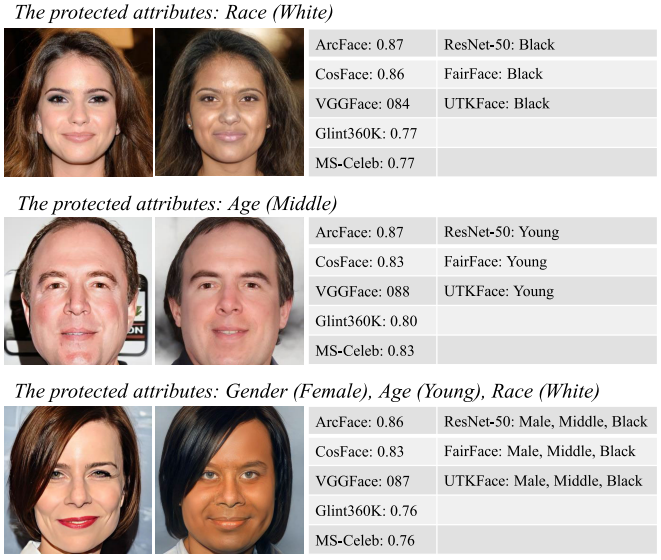


Fig. 11. Privacy-enhancing visual results with different attributes. Each pair represents the original face and the privacy-enhanced face. The annotations represent the protected attributes and their real labels. The table on the right lists the identity similarity computed by 5 face recognizers and the prediction results of 4 attribute classifiers.

TABLE IV

QUANTITATIVE COMPARISON WITH RISE ON CELEBA-HQ DATASETS UNDER FIVE FACE RECOGNIZERS

Method	RISE			Ours		
	0.0001	0.001	0.01	0.0001	0.001	0.01
VggFace	0.28	0.61	0.89	0.80	0.82	0.95
CosFace	0.37	0.40	0.78	0.91	0.92	0.96
ArcFace	0.34	0.40	0.75	0.88	0.89	0.95
Glint360K	0.37	0.40	0.78	0.81	0.89	0.95
MegaFace	0.34	0.40	0.75	0.76	0.87	0.93

Second, we perform the ID-sensitive feature attribution based on different face recognition models. As shown in Fig. 13, we could observe that the facial importance estimation results are different. However, such difference has no significant impact on achieving the ID-sensitive facial resgions. For example, the top 3 ID-sensitive facial regions obtained

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

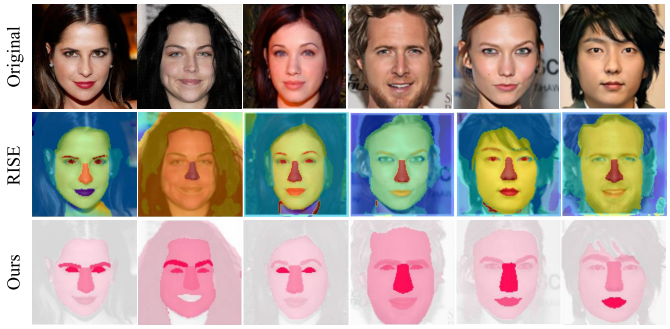


Fig. 12. Qualitative comparison with the RISE [65]. From top to bottom: the original faces, the results of RISE, our results. The intensity of the color denotes the importance for recognition.



Fig. 13. Visualizations of ID-sensitive feature attribution maps based on different face recognition models. From top to bottom: the original faces, the results based on the CosFace [8], the results based on the ArcFace [59]. The intensity of the color denotes the importance for recognition.

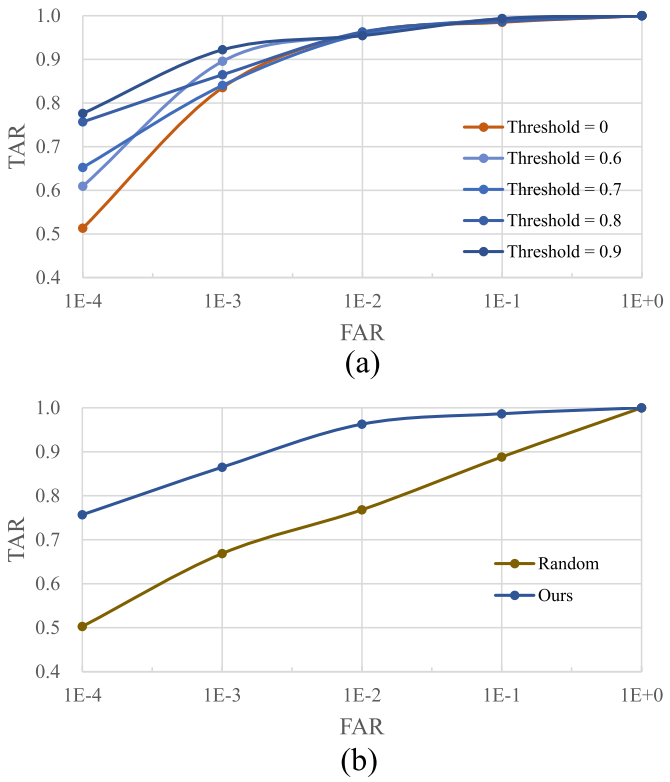


Fig. 14. Identity-preserving performance calculated by CosFace [8]: (a) The orange line denotes the results without using ID-sensitive regions, and the other lines denote the results with different ID-preserving thresholds; (b) Comparison of randomly selected regions and the threshold of 0.8.

based on the ResNet model are eyes, nose, and eyebrows, while based on the VGGFace model the top 3 ID-sensitive facial features are eyes, eyebrows, and nose. These results demonstrate the consistency of the ID-sensitive facial regions obtained on different face models. Furthermore, we validate the generalization of our method in our data utility experiments, as shown in Fig. 10.

Finally, we analysis the face verification performance of our method with different ID-preserving thresholds. We conduct this experiment on VGGFace2-HQ dataset, and employ CosFace [8] to perform the face verification. The protected images includes the following parts: 1) We directly use the generator to obtain the protected images. During the optimization process, the region shape loss term in Eq. 4 and the local relation matching loss in Eq. 6 are removed. 2) We gradually increase the threshold from 0.6 to 0.9 and generate the protected images. 3) We randomly select the face regions as the ID-sensitive facial regions P_{id} used in the generator.

Fig. 14 (a) shows the ROC curves obtained from the first two matching experiments. In detail, the threshold 0 represents the result without using ID-sensitive regions. We can find that the face verification performance decreases gradually with the threshold decreasing, which illustrates the ID-sensitive feature attribution module can significantly preserve the face identity. Fig. 14 (b) shows the results from the third experiment compared to the results with the threshold of 0.8. We can observe that our method performs much better than randomly selected regions. This demonstrate that our method can correctly discover the high ID-sensitive facial features.

Furthermore, Fig. 15 shows the number of ID-sensitive regions with different thresholds. Actually, taking different number of ID-sensitive regions will affect the balance between privacy and usability. The large number of regions encourages identity-preserving, whereas the low number of regions encourages privacy-enhancing. We can observe that the instance numbers of the first three cases are relatively balanced, when we set the threshold as 0.8. Although this numbers is also balanced at 0.9, the instance number in other cases is much larger than at 0.8. These findings indicate effective results will that yield at 0.8, which also be validated in Fig. 14 (a).

G. Ablation Study on Multi-Attribute Privacy Enhancement

To demonstrate the impact of different component of face obfuscation model, we perform an ablation study on the VGGFace2-HQ dataset. Specifically, we focus on three components of the proposed framework: the target privacy term from Eq. 3, the structure constraint term from Eq. 4, and the consistency term from Eq. 6. For a quantitative analysis, we report experimental results under different weight settings for the 3 loss terms on the identity-preserving performance, as shown in Fig. 16. In this study, we employ 5 face recognizers. Correspondingly, we report the results on attribute privacy performance, as shown in Fig. 17. From these results, we can determine the optimal combination of parameters.

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

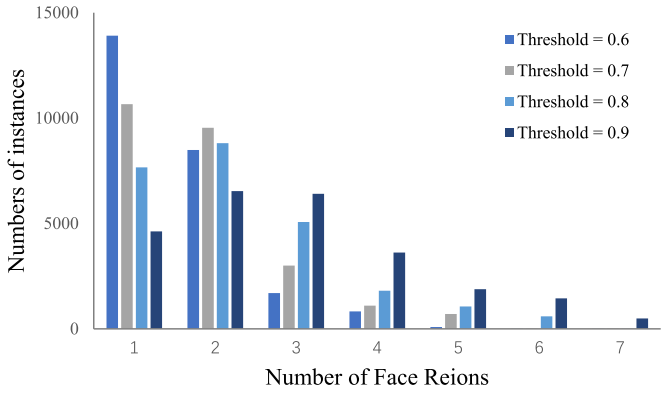


Fig. 15. Statistics on the number of ID-sensitive face regions under different thresholds.

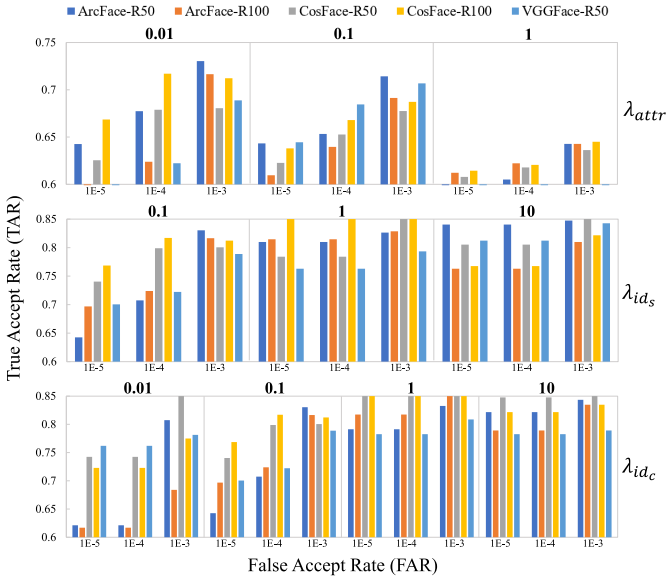


Fig. 16. Identity-preserving performance with different weighting values calculated by distinct face recognizers (ArcFace [59], CosFace [8], and VGGFace [52]).

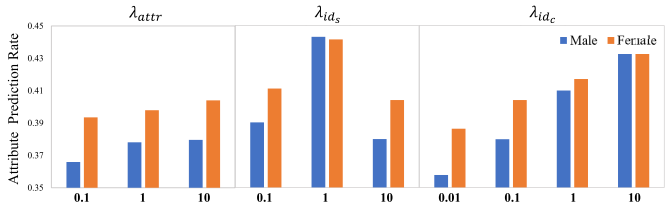


Fig. 17. Attribute privacy performance with different weighting values calculated by gender classifier (ResNet-50 [56]).

H. User Study

Besides face models, we also want to analysis the performance on appearance privacy for the human. We conduct an user study with two experiments to evaluate the effectiveness, and choose the original image and SAN [4] as the baselines. The first experiment aims to validate the protection of attribute privacy. To that end, we randomly select 150 face images from different identities, with an equal number of images those belonged to the modified attributes (gender, age, race). Users are requested to select the corresponding attribute out of our options. A total of 3750 results from 25 volunteers are collected. The average accuracy of each attribute prediction is reported in Table V. The results indicate that our method achieves the best visual obfuscation compared with the original image and the attribute privacy method SAN [4]. At the same time, the prediction rate of gender and race is lower than that of age, indicating that our method needs to improve the visual privacy of age.

TABLE V
EXPERIMENTAL RESULTS OF USER STUDY ON HUMAN-PERCEPTUAL

Method	Gender	Age	Race	Verification
Original	0.92	0.89	0.93	0.98
SAN [4]	0.87	0.85	0.87	0.93
Ours	0.52	0.66	0.49	0.36

The second experiment aims to validate the portrait privacy. Specifically, we randomly select 200 image pairs composed of 100 males and 100 females, with an equal number of pairs those belonged to the same person and those not. Then, we collected a total of 5000 results from 25 volunteers. Finally, we calculate the average accuracy of face verification to measure the ability of portrait protection, as shown in Table V. The results show that our method effectively reduces the recognition rate for the human. That is to say, when the stored images are leaked, the users' real portrait could be protected.

V. CONCLUSION

In this paper, we develop a novel privacy-enhancing face obfuscation framework for face recognition systems, which could preserving the identity discriminability while protecting multiple facial attributes. For the identity preserving, we introduce the feature attribution map as a bridge to search for the critical local regions for face recognition. For the privacy attribute protection, based on the semantic-aware attribution map, we design a novel face obfuscator to modify the privacy semantic content while preserving the ID-sensitive local structure and relations. Experiments demonstrate the effectiveness of our framework in terms of privacy protection and image utility, and also show that the semantic-aware face attribution method could find more reliable and accurate facial regions. Furthermore, our framework has a good generalization ability compared with the typical task-related baselines. In the future, we will explore the feasibility of equalizing attribute probabilities in the case of multi-class attributes. Furthermore, more privacy-sensitive information, such as health characteristics, can be considered in this privacy-enhancing framework.

REFERENCES

[1] B. Meden et al., "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4147–4183, 2021.
[2] M. Chen, X. Liao, and M. Wu, "PulseEdit: Editing physiological signals in facial videos for privacy protection," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 457–471, 2022.
[3] R. Viorescu et al., "2018 reform of Eu data protection rules," *Eur. J. Law Public Admin.*, vol. 4, no. 2, pp. 27–39, 2017.

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

[4] V. Mirjalili, S. Raschka, A. Namboodiri, and A. Ross, "Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images," in *Proc. Int. Conf. Biometrics (ICB)*, Feb. 2018, pp. 82–89.

[5] Y. Shen, C. Yang, X. Tang, and B. Zhou, "InterFaceGAN: Interpreting the disentangled face representation learned by GANs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2004–2018, Apr. 2022.

[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[7] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1547–1557.

[8] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[9] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9377–9386.

[10] T. Li and L. Lin, "AnonymousNet: Natural face de-identification with measurable privacy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 56–65.

[11] M. Maximov, I. Elezi, and L. Leal-Taixé, "CIAGAN: Conditional identity anonymization generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5446–5455.

[12] S. Yang, H. Xue, J. Ling, L. Song, and R. Xie, "Deep face swapping via cross-identity adversarial training," in *Proc. Int. Conf. Multimedia Model.*, 2021, pp. 74–86.

[13] A. Othman and A. Ross, "Privacy of facial soft biometrics: Suppressing gender but retaining identity," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 682–696.

[14] V. Mirjalili, S. Raschka, and A. Ross, "PrivacyNet: Semi-adversarial networks for multi-attribute face privacy," *IEEE Trans. Image Process.*, vol. 29, pp. 9400–9412, 2020.

[15] S. Liang et al., "A large-scale multiple-objective method for black-box attack against object detection," in *Proc. 17th Eur. Conf.* Tel Aviv, Israel: Springer, Oct. 2022, pp. 619–636.

[16] A. Liu et al., "Face anti-spoofing via adversarial cross-modality translation," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2759–2772, 2021.

[17] S. Liang, A. Liu, J. Liang, L. Li, Y. Bai, and X. Cao, "Imitated detectors: Stealing knowledge of black-box object detectors," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 4839–4847.

[18] J. Xie, C. Pun, and K. Lam, "Implicit and explicit feature purification for age-invariant facial representation learning," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 399–412, 2022.

[19] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta, "Anonymizing K facial attributes via adversarial perturbations," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 656–662.

[20] C. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5548–5557.

[21] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, pp. 1–40, Dec. 2016.

[22] L. S. Shapley, "A value for *n*-person games," in *Classics in Game Theory*, vol. 69. Princeton, NJ, USA: Princeton Univ. Press, 1997.

[23] *VGGFace2-HQ*. Accessed: Dec. 3, 2021. [Online]. Available: https://github.com/nnnnai/vggface2-hq

[24] R. Gross, L. Sweeney, F. de la Torre, and S. Baker, "Model-based face de-identification," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, 2006, p. 161.

[25] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 232–243, Feb. 2005.

[26] L. Du, M. Yi, E. Blasch, and H. Ling, "GARP-face: Balancing privacy protection and utility preservation in face de-identification," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–8.

[27] A. Jourabloo, Y. Xi, and X. Liu, "Attribute preserved face de-identification," in *Proc. Int. Conf. Biometrics (ICB)*, May 2015, pp. 278–285.

[28] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 1589–1604.

[29] S. Liang, B. Wu, Y. Fan, X. Wei, and X. Cao, "Parallel rectangle flip attack: A query-based black-box attack against object detection," 2022, *arXiv:2201.08970*.

[30] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[31] Q. Sun, L. Ma, S. Joon Oh, L. V. Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5050–5059.

[32] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, 2019, pp. 565–578.

[33] Y. Wen, B. Liu, M. Ding, R. Xie, and L. Song, "IdentityDP: Differential private identification protection for face images," *Neurocomputing*, vol. 501, pp. 197–211, Aug. 2022.

[34] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Personalized and invertible face de-identification by disentangled identity information manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3314–3322.

[35] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 564–573.

[36] J. Li et al., "Identity-preserving face anonymization via adaptively facial attributes obfuscation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3891–3899.

[37] J. Li, L. Han, H. Zhang, X. Han, J. Ge, and X. Cao, "Learning disentangled representations for identity preserving surveillance face camouflage," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9748–9755.

[38] A. Morales, J. Fierrez, R. Vera-Rodriguez, and R. Tolosana, "SensitiveNets: Learning agnostic representations with application to face images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2158–2164, Jun. 2021.

[39] P. Terhörst, D. Fährmann, N. Damer, F. Kirchbuchner, and A. Kuijper, "On soft-biometric information stored in biometric face embeddings," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 4, pp. 519–534, Oct. 2021.

[40] P. Terhörst, N. Damer, F. Kirchbuchner, and A. Kuijper, "Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations," *Appl. Intell.*, vol. 49, no. 8, pp. 3043–3060, Aug. 2019.

[41] P. Terhörst, M. Huber, N. Damer, F. Kirchbuchner, and A. Kuijper, "Unsupervised enhancement of soft-biometric privacy with negative face recognition," 2020, *arXiv:2002.09181*.

[42] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[43] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 700–708.

[44] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

[45] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.

[46] X. Li et al., "Image-to-image translation via hierarchical style disentanglement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8635–8644.

[47] Y. Wang, A. Gonzalez-Garcia, J. van de Weijer, and L. Herranz, "SDIT: Scalable and diverse cross-domain image translation," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1267–1276.

[48] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "FaceID-GAN: Learning a symmetry three-player GAN for identity-preserving face synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 821–830.

[49] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM Trans. Graph.*, vol. 42, no. 1, pp. 1–13, Feb. 2023.

[50] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for StyleGAN image manipulation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–14, Aug. 2021.

[51] Y. Nitzan et al., "MyStyle: A personalized generative prior," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–10, Dec. 2022.

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.

[52] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.
[53] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," *Comput. Operations Res.*, vol. 36, no. 5, pp. 1726–1730, May 2009.
[54] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2480–2487.
[55] M. Pernuš, V. Štruc, and S. Dobrišek, "High resolution face editing with masked GAN latent code optimization," 2021, *arXiv:2103.11135*.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[57] *UTK-Face Project*. Accessed: Mar. 29, 2023. [Online]. Available: https://github.com/arminbaz/utk-face
[58] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[59] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.
[60] X. An et al., "Partial FC: Training 10 million identities on a single machine," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1445–1449.
[61] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The MegaFace benchmark: 1 million faces for recognition at scale," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4873–4882.
[62] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 30, 2017, pp. 6629–6640.
[63] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN++: How to edit the embedded images?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8296–8305.
[64] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4352–4360.
[65] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," 2018, *arXiv:1806.07421*.



Jingzhi Li received the Ph.D. degree in cyberspace security from the University of Chinese Academy of Sciences, Beijing, China. She is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences. She was selected to join the Science and Technology Think Tank Young Talents Program of the China Association for Science and Technology. Her current research interests include image processing and face security.



Hua Zhang received the Ph.D. degree in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2015. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His current research interests include computer vision, multimedia understanding, and machine learning.



Siyuan Liang is currently pursuing the Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her current research interests include object detection and adversarial training.



Pengwen Dai received the Ph.D. degree from the Institute of Information Engineering, Chinese Academy of Sciences. He is currently an Assistant Professor with the School of Cyber Science and Technology, Sun Yat-sen University. His current research interests include pattern recognition, multimedia understanding, and AI security.



Xiaochun Cao (Senior Member, IEEE) received the B.E. and M.E. degrees in computer science from Beihang University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the University Level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University. He is currently a Professor and the Dean of the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University (SYSU). Before joining SYSU, he was a Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He has authored or coauthored over 200 journals and conference papers. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award from the International Conference on Pattern Recognition. He was on the editorial boards of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and IEEE TRANSACTIONS ON MULTIMEDIA. He is on the editorial boards of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON IMAGE PROCESSING.

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:38:25 UTC from IEEE Xplore. Restrictions apply.