# Invisible-Face: Rethinking Facial Attribute Privacy in Social Media Photo Sharing

Yong Yang, Changjiang Li, *Graduate Student Member, IEEE*, Yi Jiang, Jinbao Li, *Member, IEEE*, Xuhong Zhang, Zonghui Wang, Shouling Ji, *Member, IEEE*, and Wenzhi Chen, *Member, IEEE*

***Abstract*—As social media gains popularity, users frequently share personal photos without recognizing the risks of exposing their faces to advanced facial attribute detection technologies. These technologies can extract sensitive attributes such as age, race, sexual orientation, and potential health information from facial images, raising significant privacy concerns. Despite the availability of various anonymization techniques, our research reveals that current methods inadequately protect facial attribute privacy. They often fail to balance effectiveness and utility, underscoring the pressing need for more robust solutions in today's pervasive photo-sharing culture. To remedy this gap, we introduce Invisible-Face, a tool designed to safeguard users' facial attribute privacy using advanced adversarial perturbation techniques. Invisible-Face uses local, directional, and resilient perturbation generative strategies to obfuscate multiple facial attributes effectively, thus ensuring privacy while retaining the utility of the facial images. Our comprehensive evaluation across various datasets and model architectures shows that Invisible-Face significantly outperforms existing privacy-preserving methods in terms of effectiveness while maintaining high image naturalness. Furthermore, our extensive real-world evaluations on four popular MLaaS platforms—Baidu Brain, Tencent Cloud, Aliyun, and Face$^{++}$—reveal that Invisible-Face achieves comparable privacy protection results while preserving the visual naturalness of images, outperforming existing methods. These findings boost public awareness about the importance of facial attribute privacy and urge online social platforms to improve their protection measures.**

Received 25 October 2024; revised 16 April 2025 and 28 May 2025; accepted 29 May 2025. Date of publication 13 June 2025; date of current version 24 June 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3102100; in part by NSFC under Grant U244120033, Grant U24A20336, Grant 62172243, Grant 62402425, and Grant 62402418; in part by China Postdoctoral Science Foundation under Grant 2024M762829; in part by Zhejiang Provincial Natural Science Foundation under Grant LD24F020002; in part by Zhejiang Provincial Priority-Funded Postdoctoral Research Project under Grant ZJ2024001; and in part by the National Natural Science Foundation of China under Grant 92373205 and Grant 62374146. The associate editor coordinating the review of this article and approving it for publication was Dr. Ran He. *(Corresponding author: Zonghui Wang.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the College of Biomedical Engineering and Instrument Science, Zhejiang University.

Yong Yang, Yi Jiang, Zonghui Wang, Shouling Ji, and Wenzhi Chen are with the College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang 310027, China (e-mail: yangyong2022@zju.edu.cn; jiangyi2021@zju.edu.cn; zhwang@zju.edu.cn; sji@zju.edu.cn; chenwz@zju.edu.cn).

Changjiang Li is with the School of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA (e-mail: meet.cjli@gmail.com).

Jinbao Li is with the School of Mathematics and Statistics, Qilu University of Technology, Jinan, Shandong 250353, China (e-mail: lijinb@sdas.org).

Xuhong Zhang is with the School of Software Technology, Zhejiang University, Ningbo, Zhejiang 315048, China (e-mail: zhangxuhong@zju.edu.cn).

Digital Object Identifier 10.1109/TIFS.2025.3579592

***Index Terms*—Soft Biometrics, Face Attribute, Privacy, Deep Learning.**

## I. Introduction

THE rapid advancement of *Artificial Intelligence* (AI) has led to developments in facial detection technologies driven by deep learning and neural networks. These technologies now penetrate various sectors including security [1], surveillance [2], marketing [3], and personalized user experiences [4]. However, facial detection technologies' swift expansion and widespread application present a paradox. While they offer numerous benefits, they also expose severe privacy risks [5], particularly on digital platforms like social media, where users frequently share images. These images, once uploaded, can be accessed without authorization and misused, leading to increased privacy concerns.

The need for robust privacy protection in the digital age has become increasingly apparent. Unfortunately, current privacy mechanisms primarily focus on identity anonymization through techniques like face de-identification [6], [7], [8], often overlooking the vital issue of facial attribute privacy. Some facial attribute detection technologies can infer attributes such as age, gender, race, sexual orientation, and potential genetic diseases [5], [9], [10], severely infringing on user privacy. The leakage of facial attribute information in online shared images poses a serious threat and requires high vigilance. While some of this information is visible to the naked eye, the systematic collection and accurate analysis of such data often go unnoticed. Malicious actors may exploit this data for targeted invasive advertising [11] and, more alarmingly, manipulate political ad targeting on social media using illicitly obtained facial attributes [12]. By analyzing these attributes, they can discern individuals' pReferences and demographic attributes, allowing them to tailor political ads precisely to influence public opinion and behavior. Despite the urgent need for protective measures, existing efforts to protect the privacy of facial attributes have considerable limitations. We have explored the currently available facial image privacy protection tools for social platforms and find that their primary focus is protecting the anonymity of identities [6], [7], [13]. Through rigorous experiments and theoretical analysis, we demonstrate that the privacy protection offered by these face de-identification tools in concealing facial attributes needs to be improved. Our evidence suggests that while these tools may obscure identity, they fail to protect the breadth of information conveyed through facial attributes. Moreover,

1556-6021 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. 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: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

although recent works focus on the privacy protection of specific facial attributes, many prioritize the effectiveness of protection without considering its impact on image naturalness [14], [15], [16], which degrades user experience. Additionally, most studies on facial attribute privacy protection are designed to resist only white-box attacks [15], [17], limiting their practicality in real-world scenarios like social media sharing. Our study aims to bridge this gap.

### A. Challenges

In the current research, the privacy protection of facial attributes in online shared images faces three main challenges.

1) The present facial attribute detection techniques can analyze multiple attributes concurrently. However, accomplishing comprehensive privacy protection by making minimal changes to images is difficult.
2) Ensuring adequate facial attribute privacy while retaining an image's natural appearance is challenging.
3) Maintaining the effectiveness of privacy protection against various malicious and unknown facial attribute detection models proves challenging.

### B. Our Proposal

Driven by our findings and the apparent gaps in current privacy protection strategies, we propose Invisible-Face, a novel tool designed to enhance facial attribute privacy. Invisible-Face utilizes adversarial perturbations to obfuscate facial attributes effectively. Unlike traditional perturbation generation techniques, this tool aims to generate perturbations that are nearly invisible yet robust against various facial attribute detection models. These perturbations aim to preserve the naturalness of the image while providing adequate privacy protection for facial attributes, ensuring that Invisible-Face maintains both privacy and utility.

To maintain the natural appearance of facial images, we aim to limit the perturbation distribution across the facial image and eliminate redundancy. Our analyses show that different facial regions correlate with each attribute, suggesting that targeted local perturbations can achieve adequate facial attribute privacy protection. Based on this insight, we propose a regional perturbation strategy divided into two stages: local region selection and local perturbation generation. For the former, we utilize a heuristic-based approach that employs facial feature points as candidate regions and uses simulated annealing for specific facial attribute searches. It can determine the optimal local region. In the latter stage, we incorporate a multi-attribute adversarial loss into a *Generative Adversarial Network* (GAN)-based encoder-decoder framework to generate local perturbations capable of obfuscating multiple facial attributes concurrently in a single operation.

To enhance the effectiveness of GAN-generated perturbations, we draw inspiration from gradient-based adversarial perturbation optimization strategies [18], [19]. Our approach focuses on directing the perturbation to optimize along the most effective path, facilitated by the direction of gradient optimization. Based on this insight, we propose a directional perturbation module that integrates the direction of gradient optimization into the initial perturbation generated by the GAN. This module dynamically adjusts the optimization direction of the perturbation during training, ensuring it always aligns with the direction of the optimal solution.

To ensure the local perturbation effectively counters various unknown and malicious facial attribute detection models in real-world scenarios, Invisible-Face must exhibit robust transferability. We propose an intra-attribute and inter-attribute gradient variance reduction strategy within the directional perturbation module. This technique reduces gradient variance within an attribute through gradient regularization and decreases variance between attributes by averaging the gradient across attributes. This process mitigates the overfitting of overall perturbations to the surrogate model, thereby enhancing the transferability of our method. In addition, to reduce the risk of targeted perturbation damage, we introduce a resilient perturbation module. This module simulates potential damage to perturbations during training, thereby enhancing the robustness of our approach against targeted attacks. Random seeds control the intensity of the resilient perturbation, thereby increasing the randomness of the perturbations.

### C. Evaluation

We evaluate Invisible-Face in both white-box and black-box settings. The results indicate that Invisible-Face outperforms mainstream methods. In white-box scenarios, Invisible-Face maintains better overall effectiveness than the baseline methods and enhances the naturalness of processed facial images by approximately 24% more than the baselines. Additionally, when applied to different target face detection models, the overall effectiveness of Invisible-Face decreases by only 17.2%, whereas the performance of the baseline methods diminishes by over 50%. We assess the performance using popular *Machine Learning as a Service* (MLaaS) platforms in black-box scenarios. Invisible-Face achieves comparable privacy protection results while preserving the visual naturalness of images, outperforming existing methods.

### D. Contributions

To summarize, we make the following contributions.

- We design and implement Invisible-Face, a novel facial attribute privacy protection tool. It integrates three perturbation generation strategies—local, directional, and resilient perturbations—to ensure both privacy protection and utility preservation.
- We evaluate Invisible-Face across various datasets and attack models, demonstrating its effectiveness compared to mainstream methods. Invisible-Face not only obfuscates facial attributes effectively but also maintains the natural appearance of facial images, achieving higher naturalness than these mainstream methods.
- We conduct extensive evaluations of Invisible-Face on four popular MLaaS platforms: Baidu Brain, Tencent Cloud, Aliyun, and Face$^{++}$. Our results indicate that Invisible-Face significantly outperforms existing methods.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

## II. Background

### A. Neural Network Classifier

Given an input image, represented by a feature vector $x \in \mathbb{R}^n$, a neural network classifier works to categorize it into one of $m$ possible classes by propagating it through $L$ layers. Each layer $l$ involves a linear or nonlinear transformation of the previous layer's output, represented by a weight matrix $W^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ and a bias vector $b^{(l)} \in \mathbb{R}^{d_l}$, followed by the application of a nonlinear activation function $\sigma$ to generate the output activations $h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)})$. The final output layer $L$ applies a softmax function to yield a probability distribution over the $m$ classes: $p(y = j|x) = \frac{e^{h_j^{(L)}}}{\sum_{k=1}^{m} e^{h_k^{(L)}}}$. The network aims to minimize a loss function $\mathcal{L}$, representing the disparity between the predicted probabilities and the actual labels, by iteratively updating weights and biases via backpropagation and gradient descent.

### B. Facial Attribute Detection

The rapid development of deep learning has resulted in a remarkable performance in face detection tasks. Deep learning technology can predict a person's inherent attributes and semantic information from their facial biometric data. Some research has focused on developing methods to detect facial attributes such as gender, age, and race. *Convolutional neural networks* (CNNs) were initially utilized for facial attribute detection and have demonstrated high accuracy in this task [20]. Researchers proposed model improvement techniques and new facial datasets to improve accuracy and robustness. For instance, Karkkainen and Joo [21] proposed the FairFace dataset to address facial attribute training data bias.

In addition to detecting inherent facial attributes, extracting other sensitive information from facial features is also possible. Gurovich et al. [9], [22] proposed a deep learning algorithm framework that can identify over a hundred genetic disorders from facial images. Moreover, deep neural networks can detect a user's sexual orientation from facial images [10], with an accuracy rate exceeding 80% in identifying homosexuality.

### C. Risk of Facial Attribute Leakage

As mentioned above, facial attribute detection technology can identify highly sensitive personal information, including race, sexual orientation, and genetic diseases. Such data breaches pose serious privacy infringements and increase the risk of social discrimination and exclusion. Communities harboring biases against particular races or sexual orientations may engage in deliberate exclusion or discrimination once such information is disclosed [23], [24].

While attributes such as gender are visible to the naked eye, the large-scale analysis of facial attributes typically relies on deep learning systems for the detection and extraction of attribute information. This information is frequently exploited to target users on specific social platforms or geographic regions, thereby posing substantial collective privacy threats. For instance, unethical merchants might exploit this data for biased promotional campaigns [25]. Furthermore, malicious actors can exploit illicitly obtained facial attributes from user photos to manipulate political ad targeting on social media [12]. By analyzing these facial attributes, these actors can identify individuals' preferences and demographic attributes, enabling them to tailor their political advertisements precisely to influence public opinion and behavior.

Given these concerns, the US *Federal Trade Commission* (FTC) [25] has voiced significant worries about the misuse of consumer biometric data. Likewise, European regulatory bodies such as the *European Data Protection Board* (EDPB) [26] have issued warnings about the grave risks associated with the recognition of biometric features in public spaces. Consequently, safeguarding the privacy of facial attributes on public platforms is imperative.

## III. Related Work

Recently, some works have proposed and investigated facial privacy protection techniques [6], [16], [27], [28], [29]. In particular, Meden et al. [30] provides a broad overview of current methods and summarizes the trade-offs involved. In this section, we categorize relevant prior works based on their underlying strategies.

### A. Perturbation-Based Privacy Protection

*1) Identity Anonymization:* Deep learning's vulnerability to adversarial perturbations has spurred research into using such perturbations to diminish the effectiveness of face privacy detection systems and enhance privacy protection. For example, the Fawkes system, developed by Shan et al., employed pixel-level modifications to prevent unauthorized facial recognition [6]. Similarly, Cherepanova et al. introduced LowKey, a method designed to facilitate effective face de-identification under black-box operational conditions [7]. Additionally, the TIP-IM method introduced by Yang et al. maintains the visual authenticity of images while concealing personal identities [8]. However, these endeavors primarily center on protecting identity privacy, overlooking the potential for safeguarding a wider array of facial attributes.

*2) Facial Attribute Anonymization:* There are also some recent perturbation-based facial attribute anonymization methods. Chhabra et al. advocated for the concurrent anonymization of k-facial attributes through adversarial perturbations but fell short of providing a comprehensive quantitative evaluation of the visual quality of images [27]. Mirjalili et al. proposed PrivacyNet [16], a framework based on GANs to obfuscate facial attributes. However, PrivacyNet does not address the balance between privacy effectiveness and visual realism. Furthermore, these adversarial perturbation-based methods demonstrate their effectiveness in white-box scenarios. However, their real-world applicability suffers due to the unpredictable nature of the models used by adversaries to infringe upon facial privacy. The transferability of such defensive mechanisms remains to be explored.

### B. Face Editing-Based Privacy Protection

*1) Identity Anonymization:* As image generation and editing technologies have advanced, a new spectrum of privacy

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

protection strategies based on facial editing has emerged. DeepPrivacy [14], a popular face editing methodology, anonymizes identities by adjusting facial image features while preserving the original data distribution. Similarly, Chen et al.'s *Perception Indistinguishable network* (PI-Net) [31], using StyleGAN and differential privacy, generates realistic facial images for anonymization. While these facial editing techniques offer privacy protection by modifying attributes, they also alter the original appearance of the face. Moreover, the protection efficacy for multiple facial attributes simultaneously has yet to be thoroughly examined.

*2) Facial Attribute Anonymization:* In the domain of facial attribute privacy protection, Zhang et al. proposed RAPP [17], which uses attribute confusion and adversarial networks to protect facial attributes. However, RAPP is primarily designed for binary attributes (e.g., gender or race) and relies on binary conditional inputs during training, making it incapable for multi-class attributes commonly encountered in real-world privacy scenarios such as age and race. The introduction of diffusion model-based face editing, exemplified by Huang et al.'s *Collaborative Diffusion* (CollDiff) [32], supports multi-modal face generation and editing through the combined use of text prompts and masks for modifying facial attributes such as age. Recently, Zhang et al. [33] proposed a generalized framework for visual privacy protection and instantiated it in the context of facial privacy using CollDiff. Despite the potential effectiveness of these methods, their practicality, especially in scenarios where the attacker's model is unknown and the balance between privacy protection effectiveness and utility, requires further exploration. Thus, a need exists to refine and develop better ways to balance privacy protection with utility.

## IV. Threat Model

We consider a scenario where adversaries harvest publicly shared face images on online platforms and use facial attribute detection models to extract privacy-sensitive information at scale. The goal is to analyze facial attribute distributions and perform large-scale detection without user consent. Adversaries are assumed to have access to facial images, sufficient computational resources, and public datasets to train their own facial attribute detection models, referred to as *target models*.

To mitigate this threat, defenders (e.g., platform administrators) can process images before public distribution, ensuring that target models cannot accurately infer facial attributes while maintaining the images' natural appearance and recognizability. This aligns with the needs of social platforms like Instagram and WeChat, where users prefer realistic photos. The defender may not have access to the target model's architecture, and thus uses a *surrogate model* as an approximation for training. We evaluate two scenarios: in the *white-box scenario*, the defender has full access to the target model, while in the *black-box scenario*, the target model is unknown. Although some facial attributes may remain visually perceivable to human observers, the defender's goal is to prevent automated, large-scale extraction by the target model, reducing the risk of population-level privacy inference.

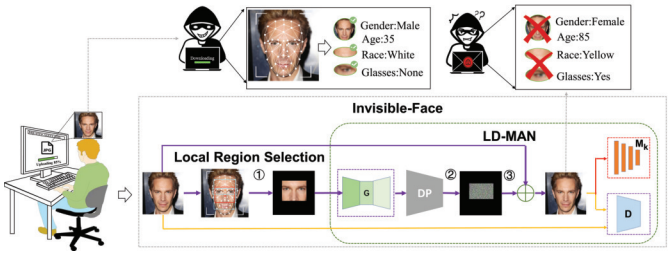



Fig. 1. Overview of invisible-face.

## V. Design of Invisible-Face

### A. Overview

We show the overview of Invisible-Face in Figure 1. Invisible-Face pipeline obfuscates facial attributes to protect user privacy through three steps: local region selection, perturbation generation, and perturbation application.

In the first step, Invisible-Face identifies specific local regions on the face that are most relevant to facial attributes. A heuristic-based approach is employed, using facial landmarks and a simulated annealing algorithm to optimize region selection. This ensures that the perturbations target critical areas for attribute obfuscation while minimizing perceptual distortion to preserve the image's natural appearance.

In the second step, a GAN is utilized to generate imperceptible perturbations. Further, two modules are designed to enhance robustness and transferability: the directional perturbation module, which optimizes perturbation direction using gradient information and a variance reduction algorithm, and the resilient perturbation module, which employs a pre-trained encoder-decoder model to simulate and resist perturbation degradation, incorporating random intensity adjustments to prevent targeted disruptions.

Finally, the perturbations are applied to the selected regions, ensuring the image retains its natural appearance while effectively deceiving facial attribute detection models.

### B. Heuristic-Based Local Region Selection

We propose a local perturbation strategy to preserve the visual quality of facial images while ensuring strong attribute privacy protection. This approach is based on the observation that different facial attributes are linked to specific regions [34], [35]. For instance, eye sockets are crucial for gender classification, as males typically have more prominent brow ridges and angular eye shapes, whereas females tend to have rounder sockets [36]. By restricting perturbations to semantically relevant areas, our method reduces unnecessary noise, preserving visual naturalness while effectively obfuscating sensitive attributes. In contrast, global perturbations often introduce redundant distortions, degrading visual quality.

While discriminative regions differ across facial attributes, our method avoids selecting separate regions for each task. Instead, it employs a unified local perturbation strategy that optimizes a single region jointly informative for all target attributes. This is achieved through simulated annealing guided by a composite objective, which balances attribute prediction consistency and perturbation sparsity. The strategy enables

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

multi-attribute obfuscation without task-specific customization, ensuring broad applicability in privacy protection.

In this section, we explain how we identify the most suitable regions for local perturbation. We initially utilize the affine matrix to align faces and their corresponding landmarks on a clean facial image $x$. Subsequently, we generate a mask image $x^M$ that exclusively encompasses specific local facial regions. The model can focus on generating perturbations within these localized regions when computing the perturbation. Formally, $x^M$ is defined as follows:

$$x^M(l_x, l_y, r_x, r_y) = \mathcal{A}(l_x, l_y, r_x, r_y) \odot x, \tag{1}$$

where $\mathcal{A}$ is a 0-1 mask function. It is employed to restrict and determine the magnitude and placement of the local region. The top left corner of the local region on the image is denoted by the coordinates $l_x$ and $l_y$, while the bottom right corner is denoted by $r_x$ and $r_y$. $\odot$ is the Hadamard product.

For $x^M$ to effectively replace $x$ for local perturbation generation, it must satisfy two requirements. Firstly, as a substitute for $x$, $x^M$ must contain the target attribute information as much as possible. In other words, the adversary facial attribute detection model should accurately predict the target's facial attributes by $x^M$. Secondly, the local region's area $S(\cdot)$ in $x^M$ should be as small as possible to create a subtle and privacy-preserving facial perturbation.

Thus, we formulate the selection of local regions as a multi-objective optimization problem:

$$\begin{aligned} &\min_{x^M} \ S(x^M(l_x, l_y, r_x, r_y)), \\ &\text{s.t. } x^M = \underset{x^M}{argmin} \ \frac{1}{K}\sum_{k=1}^{K} DE(M_k(x^M), M_k(x)), \end{aligned} \tag{2}$$

where $DE(\cdot)$ denotes a distance evaluation metric such as cross-entropy, $K$ denotes the number of facial attributes, and $M_k(\cdot)$ denotes the predicted probability of the facial attribute detection model.

To avoid dispersion and find only locally optimal solutions in the solution space of multi-objective optimization, we convert the multi-objective function into a single-objective function:

$$min \ \ SE(x^M) = \frac{\theta_m}{K}\sum_{k=1}^{K} DE(M_k(x^M), M_k(x)) + \theta_s S(x^M), \tag{3}$$

where $SE(\cdot)$ denotes the selection function of the local region, $\theta_m$ and $\theta_s$ are normalization coefficients.

To minimize $SE(\cdot)$, we propose a heuristic-based local region selection algorithm inspired by the simulated annealing algorithm. Algorithm 1 presents the overall algorithm.

Facial attribute detection primarily relies on the facial features and their contiguous regions; therefore, we base the selection of local regions on all feature points of the face. Initially, we store these candidate feature points in a list denoted as $P$, sorted by the magnitude of their abscissa. Subsequently, we randomly select feature points from regions with significant attribute disparities [37], such as the eyes, nose, and mouth, to serve as the initial central point of the local region, denoted as $p_c$. In the preliminary search for the optimal value, we select the feature point with the smallest abscissa in $P$ as the initial point, $p_l$, for the upper left corner of the local region. Utilizing the geometric distribution of rectangles, we can determine the initial point $p_r$ of the local area's lower right corner. Hence, we can readily obtain the initial output denoted as *output*. We refer to these steps as the "Initialization of Local Regions", corresponding to lines 1-6 in Algorithm 1.

We calculate the current temperature $T$ during the main loop according to the current iteration step *step*. Then, we select the next feature point $p_l$ in index order and update the bottom right feature point $p_r$ to define a new local region. We evaluate the effect of the latest local region by computing the delta $\delta$ of the objective function. We accept this new local region if $\delta < 0$ or meets the Metropolis criterion. Otherwise, we maintain the original state. As the iteration advances and the temperature progressively decreases, Algorithm 1 transitions gradually from a global search to a local search, seeking the optimal solution within the local area. We refer to these steps as "Local Region Exploration", corresponding to lines 7-20 in Algorithm 1.

**Algorithm 1** Heuristic-Based Local Region Selection Algorithm

**Require:** List of candidate face feature points $P$, the index value *idx* of the list, starting temperature $T_{max}$, final minimum temperature $T_{min}$, cooling constant $k$, current iteration *step*, the uniform distribution $U$, original facial image $x$

1: Sorted ($P$, $key = lambda\ p{:}p[0]$)
2: Select initial central points $p_c = [c_x, c_y]$
3: Select initial top-left feature point $p_l \leftarrow P[idx]$, where $idx = 0$
4: Select initial bottom-right feature point $p_r \leftarrow [2p_c[0] - p_l[0], 2p_c[1] - p_l[1]]$
5: $x^M \leftarrow \mathcal{A}(p_l[0], p_l[1], p_r[0], p_r[1]) \odot x$
6: $output \leftarrow SE(x^M)$
7: **while** $p_l[0] < p_c[0]$ and $p_l[1] < p_c[1]$ and $T_{max} \times c^{(step)} > T_{min}$ **do**
8: $\quad T \leftarrow T_{max} \times k^{(step)}$
9: $\quad p'_l \leftarrow P[idx + 1]$
10: $\quad p'_r \leftarrow [2p_c[0] - p'_l[0], 2p_c[1] - p'_l[1]]$
11: $\quad output' \leftarrow SE(x^M(p'_l[0], p'_l[1], p'_r[0], p'_r[1]))$
12: $\quad \delta \leftarrow output' - output$
13: $\quad$ **if** $\delta < 0$ or $e^{-\frac{\delta}{T}} > U[0, 1]$ **then**
14: $\quad\quad output \leftarrow output'$
15: $\quad\quad p_l \leftarrow p'_l$
16: $\quad\quad p_r \leftarrow p'_r$
17: $\quad$ **end if**
18: $\quad step \leftarrow step + 1$
19: $\quad idx \leftarrow idx + 1$
20: **end while**
21: **return:** $p_l$, $p_r$

### C. Perturbation Generation

Upon establishing the most favorable local region, the subsequent stage encompasses the computation and generation of perturbations. The objective is to successfully obfuscate an

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

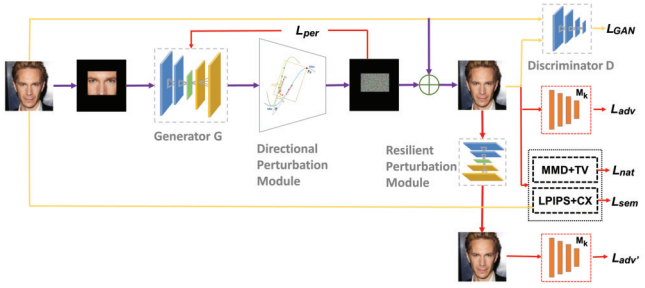
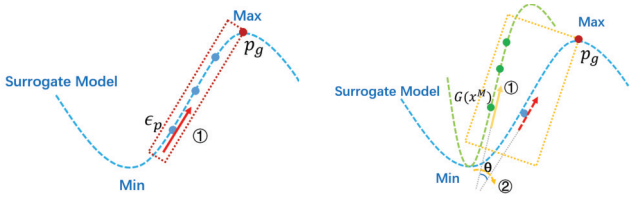



Fig. 2. The architecture of *local and directional multi-attribute adversarial network* (LD-MAN).




(a) Gradient-based adversarial perturbation. (b) GAN-based adversarial perturbation.

Fig. 3. The intuition for why adversarial perturbations for GAN-based methods are less effective than gradient-based methods. Visualized on a simplified 2D training space.

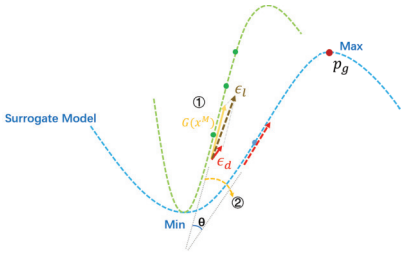



Fig. 4. The working principle of *DP*.

array of facial attributes inherent in online images while retaining their naturalistic visual representation. In mathematical terms, we can elucidate the goal of ensuring privacy protection for online images as follows:

$$\max_{x^A} \sum_{k=1}^{K} DE(M_k(x), M_k(x^A)),$$
$$\text{s.t. } x^A = argmax\ QE(x, x^A), \tag{4}$$

where $x^A$ is the adversarial facial image, $QE$ is the image quality evaluation metric such as *structural similarity index measure* (SSIM) and *peak signal to noise ratio* (PSNR). $M_k(\cdot)$ and $DE$ have the same meanings as in Equation 2.

We propose a novel GAN-based framework, termed *Local and Directional Multi-Attribute Adversarial Network* (LD-MAN), designed for multi-attribute adversarial perturbation. The LD-MAN architecture comprises a generator $G$, a discriminator $D$, a directional perturbation module $DP$, and a resilient perturbation module $RP$, as shown in Figure 2. To ensure proper regularization of the perturbation generation, we define four loss functions: multi-attribute adversarial loss, perturbation loss, naturalness loss, and semantic loss. These loss functions enable us to generate local perturbations that meet multiple attribute requirements while preserving the naturalness and semantics of the image.

*1) Generator G and Discriminators D:* The generator $G$ is responsible for generating the initial perturbation. $G$ takes the image $x^M$ as input, obtained after local region selection, which enables $G$ to focus mainly on a specific image area when generating the initial perturbation. This results in a local perturbation.

$D$ measures the difference distribution between $x^A$ and $x$. $D$ encourages $x^A$ to be more perceptually natural. To achieve this, both $G$ and $D$ use a uniform expression for their loss functions as follows [38]:

$$\mathcal{L}_{GAN} = \mathbb{E}_x \log D(x) + \mathbb{E}_x \log G(x^A). \tag{5}$$

While GAN-based approaches confer certain benefits over gradient-based methods, such as preserving naturalness in adversarial samples, their efficacy tends to fall short of that provided by gradient-based methods. We illustrate this concept in Figure 3, using a simplified 2D rendition of the training space.

As shown in Figure 3(a), the perturbation $\epsilon_p$ direction in gradient-based adversarial methods, such as FGSM, is a fixed anti-gradient direction. The size of the perturbation only needs to be continuously adjusted to get close to the global optimum $p_g$. In contrast, GAN-based adversarial methods need to consider both the direction and size of the perturbation, as shown in Figure 3(b). The perturbation $G(x^M)$ generated by GAN requires a two-step optimization process. Firstly, we must continuously optimize the size of $G(x^M)$ to reach an optimal state. Secondly, the angle $\theta$ of the $G(x^M)$ needs to be constantly adjusted so that its direction gets closer to $p_g$.

Therefore, GAN-based adversarial methods often have high complexity and slow convergence speed, which makes it difficult for the generated perturbations to reach the global optimum. The limitation mentioned above hinders the effectiveness of GAN-based adversarial methods for preserving the privacy of faces in images.

*2) Directional Perturbation Module DP:* To ensure the optimal performance of LD-MAN, we have incorporated a directional perturbation module, $DP$, into our framework. Figure 4 demonstrates the working principle of $DP$ in conjunction with our approach. The newly generated local perturbation can continuously approach the global optimal direction by adding an auxiliary direction gradient $\epsilon_d$ to $G(x^M)$. We achieve this by taking advantage of the geometric properties of in-plane tensor addition, where adding two tensors results in a tensor sum whose direction lies between the individual directions. Since $\epsilon_d$ contains gradient information, its direction is close to that of $p_g$. Therefore, incorporating $\epsilon_d$ into the optimization process can reduce the angle $\theta$, bringing the newly generated local perturbation $\epsilon_l$ closer to $p_g$. Mathematically, the expression of $\epsilon_l$ is as follows:

$$\epsilon_l = G(x^M) + \beta\epsilon_d, \tag{6}$$

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

where $\beta$ is the decay coefficient used to limit the size of $\epsilon_d$. Since $\epsilon_d$ is only used to guide the optimization direction, $\beta$ can be set to a minimal value.

Therefore, the expression of the adversarial image $x^A$ is as follows:

$$x^A = x + \epsilon_l. \tag{7}$$

The determination of $\epsilon_d$ is a critical factor for the success of *DP*. Simply employing an inverse gradient for directional perturbation can result in local perturbations that overfit the surrogate model, thus compromising perturbation transferability. Motivated by the *Stochastic Variance Reduced Gradient* (SVRG) algorithm [39], we found that reducing gradient variance can notably enhance the transferability of local perturbations.

In preserving the privacy of facial attributes, there is a need for gradients that can simultaneously resist the detection of multiple attributes. However, it is impractical to amalgamate all gradients and then apply variance reduction. This is because different attribute detection tasks have disparate decision boundaries, leading to the anisotropic distribution of gradients among various attribute detections.

**Algorithm 2** Variance Reduction Algorithm for Intra- and Inter-Attribute Gradients

**Require:** Input image $x$, the number of facial attributes $K$ and the corresponding label $y_k$, the loss function $J_k$, the number of neighbor samples $N$ with upper bound $\epsilon$, the uniform distribution $U$

Initialize the intra-attribute gradient $g_{intra}$ and inter-attribute gradient $g_{inter}$
**for** $k = 0$ to $K - 1$ **do**
  $g_{inter} \leftarrow g_{inter} + \nabla_x J(x, y_k)$
  Initialize gradient variance $v$.
  **for** $i = 0$ to $N - 1$ **do**
    $x^i = x + clip(U(0, 1), -\epsilon, \epsilon)$
    $v \leftarrow v + \frac{1}{N}(\nabla_x J(x, y_k) - \nabla_{x^i} J(x^i, y_k))$
  **end for**
  $g_{intra} \leftarrow g_{intra} + \frac{\nabla_x J(x,y_k)-v}{\|\nabla_x J(x,y_k)-v\|_1}$
**end for**
$g_{inter} \leftarrow \frac{1}{N} g_{inter}$
$output \leftarrow \text{sign}(g_{intra} + g_{inter})$
**return:** *output*

To augment the transferability of local perturbations, we introduce a variance reduction algorithm that considers both intra-attribute and inter-attribute gradients, treating each attribute as a distinct boundary. Algorithm 2 further outlines the details of this proposed algorithm.

To reduce gradient variance in intra-attributes, we calculate their gradient variance using $N$ samples in the neighborhood of the input image $x$. We perform the calculation as follows [40]:

$$v(x) = \nabla_x J(x, y) - \frac{1}{N} \sum_{i=1}^{N} \nabla_{x^i} J(x^i, y), \tag{8}$$

where $J(\cdot)$ is the loss function, and $x^i$ is a sample randomly generated by $x$ in the neighborhood of $[-\epsilon, \epsilon]$. $v(x)$ denotes the estimated variance of gradients. Considering the iterative nature of $v(x)$, we refer to it as $v$ in the subsequent discussions for clarity. After reducing the gradient variance in intra-attribute, the resulting gradient $g_{intra}$ in intra-attribute is as follows:

$$g_{intra} = \sum_{k=1}^{K} \frac{\nabla_x J(x, y_k) - v}{\|\nabla_x J(x, y_k) - v\|_1}, \tag{9}$$

where $y_k$ denotes the label corresponding to the $k$-th facial attribute. $K$ denotes the number of facial attributes. We take the gradient mean of $K$ attributes to reduce the gradient variance in inter-attributes. Thus, gradient $g_{inter}$ in inter-attributes is as follows:

$$g_{inter} = \frac{1}{N} \sum_{k=1}^{K} \nabla_x J(x, y_k). \tag{10}$$

From this, we can calculate $\epsilon_d$ generated by *DP*, as follows:

$$\epsilon_d = \text{sign}(g_{intra} + g_{inter}). \tag{11}$$

*3) Multi-Attribute Adversarial Loss:* To generate privacy-adversarial perturbation, we introduce a multi-attribute adversarial loss function $\mathcal{L}_{adv}$. The expression of $\mathcal{L}_{adv}$ is as follows:

$$\mathcal{L}_{adv} = \sum_{k=1}^{K} \mathbb{E}_{x,k}[\log P(M_k(x^A) \neq y_k \mid x^A)]. \tag{12}$$

*4) Perturbation Loss:* To limit the size of the perturbation, we also minimize the $l_2$ loss for local perturbation. The expression of the perturbation loss function $\mathcal{L}_{per}$ is as follows [38]:

$$\mathcal{L}_{per} = \mathbb{E}_x \|\epsilon_l\|_2. \tag{13}$$

*5) Naturalness Loss:* However, even with the $l_2$ norm constraint, the perturbation may still be noticeable, which can reduce the naturalness of the image [8]. To address this issue, we introduce a naturalness loss function, denoted as $\mathcal{L}_{nat}$, which aims to improve the naturalness of the image further.

$\mathcal{L}_{nat}$ is composed of two main parts. Firstly, we use the *maximum mean difference* (MMD) [41] as the loss function, denoted as $\mathcal{L}_{MMD}$, to reduce the imperceptible difference between the two data distributions. The expression for $\mathcal{L}_{MMD}$ is given by [41]:

$$\mathcal{L}_{MMD} = \frac{1}{N_m} \| \sum_{i=1}^{N_m} \phi(x_i) - \sum_{j=1}^{N_m} \phi(x_j^A) \|_{\mathcal{H}}^2, \tag{14}$$

where $\mathcal{H}$ is the *reproducing kernel Hilbert space* (RKHS) with a Gaussian kernel, and the function $\phi(\cdot)$ maps the original samples to RKHS. $N_m$ is the number of samples for comparison.

Secondly, we use the *total variation* (TV) loss for noise reduction, which helps maintain the image's smoothness. We formulate TV loss function $\mathcal{L}_{TV}$ as follows [42]:

$$\mathcal{L}_{TV} = \sum_{i,j} ((x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2)^{\frac{1}{2}}, \tag{15}$$

where $x_{i,j}$, $x_{i+1,j}$, and $x_{i,j+1}$ represent the pixels in the input image respectively.

Therefore, the final $\mathcal{L}_{nat}$ is the weighted sum of the above losses:

$$\mathcal{L}_{nat} = \lambda_{MMD} \mathcal{L}_{MMD} + \lambda_{TV} \mathcal{L}_{TV}, \tag{16}$$

where $\lambda_{MMD}$ and $\lambda_{TV}$ are hyper-parameters.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

*6) Semantic Loss:* Although adversarial attacks via local perturbations can preserve privacy, they should not affect semantic information beyond privacy preservation. Any distortion of the original semantic expression of the picture can undermine users' intention to share images online. To mitigate this issue, we propose a semantic loss function $\mathcal{L}_{sem}$.

$\mathcal{L}_{sem}$ mainly considers semantic loss from global and local perspectives. On the one hand, we reduce the overall semantic loss by adopting *learned perceptual image patch similarity* (LPIPS) as the loss function $\mathcal{L}_{LPIPS}$ [43]. On the other hand, we reduce the local semantic loss by employing the context loss function $\mathcal{L}_{CX}$. $\mathcal{L}_{CX}$ is formulated as follows [44]:

$$\mathcal{L}_{CX} = -\log \sum_{i} (CX(\Phi^i(x), \Phi^i(x^A))), \tag{17}$$

where $\Phi^i(\cdot)$ is the feature map extracted in the $l$-th layer of the perception network $\Phi$, and $CX(\cdot)$ is the context similarity function [44].

The expression of $\mathcal{L}_{sem}$ is as follows:

$$\mathcal{L}_{sem} = \lambda_{LPIPS}\mathcal{L}_{LPIPS} + \lambda_{CX}\mathcal{L}_{CX}, \tag{18}$$

where $\lambda_{LPIPS}$ and $\lambda_{CX}$ are hyper-parameters.

*7) Resilient Perturbation Module RP:* The capability to counter a wide range of unknown disturbances is paramount. We put forth *RP*, a pre-trained encoder-decoder model such as Autoencoder [45] and U-Net [46] to address this requirement. The model primarily simulates behaviors that may cause damaging perturbations. The primary purpose is to remove adversarial perturbations in the adversarial image $x^A$.

In our pursuit to yield a perturbation of resilient nature, we expect that perturbations processed by $RP(x^A)$ should continue to follow the principles of multi-attribute adversarial losses. This postulation is hinged on the notion that the model can generate perturbations robust enough to resist elimination or minimization attempts by *RP*, thereby maintaining their efficacy for adversarial purposes. To further ensure the resilience of *RP*, we set a random seed $s$ to dynamically adjust the intensity of *RP*, generating random perturbations. This strategy prevents adversaries from conducting targeted disruptions to perturbations. Mathematically, this process is as follows:

$$\mathcal{L}'_{adv} = \sum_{k=1}^{K} \mathbb{E}_{x,k,s}[\log P(M_k(RP(x^A, s)) \neq y_k \mid RP(x^A, s))]. \tag{19}$$

*8) Total Loss:* The total loss $\mathcal{L}_{total}$ is as follow:

$$\mathcal{L}_{total} = (\mathcal{L}_{adv}, \mathcal{L}_{per}, \mathcal{L}_{nat}, \mathcal{L}_{sem}, \mathcal{L}'_{adv})\Lambda^T, \tag{20}$$

where $\Lambda = (\lambda_{adv}, \lambda_{per}, \lambda_{nat}, \lambda_{sem}, \lambda'_{adv})$ represents the hyper-parameters.

## VI. Experiment Setup

### A. Datasets

To experimentally evaluate the white-box model, we utilize three datasets: UTKFace [47], CelebA [48], and CelebA-HQ [49]. The UTKFace consists of roughly $20K$ facial images, which include three types of attribute information: *gender, age*, and *race*. Given that age predictions may vary in real-world scenarios, we employ multi-category labeling with 20 years as the interval in our experiment. The CelebA comprises approximately $200K$ facial images annotated with 40 attribute categories. For our experiment, we select 10 attributes to evaluate, namely *attractive, black hair, blond hair, chubby, eyeglasses, heavy makeup, male, no beard, oval face*, and *young*. The CelebA-HQ is an upgraded version of CelebA and features around $30K$ facial images, each with 40 attribute annotations. We use this dataset to assess the naturalness of the images visually.

To evaluate black-box models and real-world MLaaS platforms, we consider a dataset of approximately $6K$ facial images to simulate realistic test samples. This dataset includes about $2K$ facial images randomly sourced from online platforms and an additional $4K$ images selected from the UTKFace and CelebA, covering multiple attributes: *gender, age, race*, and *facial expression*.

### B. Surrogate Models

For the white-box evaluation experiments, we select facial attribute detection models with varying widths and depths, including ResNet-18 [50], ResNet-34 [50], ResNet-50 [50], MobileNet [51], GoogLeNet [52], and VGG-16 [53]. It allows us to evaluate the effectiveness and transferability of InvisibleFace comprehensively.

### C. Target Models

In the white-box scenario, we use the surrogate models as the target models. In the black-box scenario, we consider four popular MLaaS platforms: Baidu Brain [54], Tencent Cloud [55], Aliyun (Alibaba Cloud) [56], and Face$^{++}$ [57].

### D. Baseline Methods

Our study evaluates four distinct methodological categories for protecting facial attribute privacy. The first group comprises perturbation-based methods such as FGSM [18] and PGD [19], which we repurpose for defense.

The second group comprises the *state-of-the-art* (SOTA) perturbation-based transfer attack methods, including MI [58], TI [59], and VMI [40], which we also adapt for defensive purposes.

The third group includes perturbation-based anonymization methods, which generate adversarial perturbations to protect identity or facial attributes. For identity anonymization, we adapt LowKey [7], Fawkes [6], and TIP-IM [8] by replacing their original loss functions with cross-entropy loss for attribute classification. These adapted versions are denoted as LowKey-*m*, Fawkes-*m*, and TIP-IM-*m*. We also compare with perturbation-based facial attribute anonymization methods, including PrivacyNet [16], which uses a GAN to obfuscate facial attributes, and Chhabra et al. [27], who propose adversarial perturbations for multi-attribute obfuscation. As their code is not publicly available, we re-implemented both methods following the descriptions provided in their original papers.

The fourth group includes popular face-editing-based anonymization methods like DeepPrivacy [14], PI-Net [31], *Collaborative Diffusion* (CollDiff) [32], and Zhang et al. [33].

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

### E. Parameter Settings

In the simulated annealing process for local region selection, we set $T_{\max} = 10$, $T_{\min} = 0.1$, and $k = 0.95$, with the temperature updated as $T = T_{\max} \cdot k^{step}$. During training, we use 120 epochs, a batch size of 64, and a learning rate of 0.001. For DP module, the decay coefficient of $\epsilon_d$ is set to $\beta = 0.03$, and the number of neighbor samples for intra-attribute gradient variance estimation is $N = 10$. For loss balancing, we set $\lambda_{TV} = 0.0001$ and $\lambda_{MMD} = 1$ in the naturalness loss, while $\lambda_{LPIPS} = 0.5$ and $\lambda_{CX} = 0.5$ in the the semantic loss. The total loss weights are set as $\Lambda = (0.4, 0.1, 0.15, 0.15, 0.2)$. In the pixel range $[0, 255]$ experiments, the maximum perturbation size is $\epsilon = 8$ under the $l_p$ norm constraint. Other baselines use default settings. These hyperparameters are initially selected based on prior works [8], [60] and then fine-tuned using grid search on a validation set to balance privacy effectiveness and image naturalness.

### F. Metrics

*1) Effectiveness:* To evaluate the performance of privacy protection, we propose a metric called *Multi-Attribute Obfuscation Success Rate* (MOSR), based on the commonly used *Attack Success Rate* (ASR) in adversarial attacks. Formally, the definition of MOSR is as follows:

$$\text{MOSR} = \frac{|\{x \in \mathcal{D} : \forall i \in \{1, \ldots, n\}, f_{a_i}(x^A) \neq y_{a_i}\}|}{|\mathcal{D}|}, \tag{21}$$

where $\mathcal{D}$ is the evaluation dataset, $x$ represents an image prior to privacy protection, and $x^A$ denotes the same image after privacy protection. In this context, $a_i$ is the identifier for the $i-th$ facial attribute, $f_{a_i}(\cdot)$ is the detection model specific to facial attribute $a_i$, and $y_{a_i}$ is the corresponding ground truth. Specifically, we consider privacy protection successful when we simultaneously obfuscate multiple target attributes. On the contrary, if there is one attribute of obfuscation failure, it will be regarded as a failure.

*2) Naturalness:* To evaluate the naturalness of the generated adversarial images, we utilize two metrics: *Structural Similarity Index Measure* (SSIM) [61], [62] and *Peak Signal to Noise Ratio* (PSNR) [63], [64]. The higher the values of these metrics, the better the quality and naturalness of the images.

*3) Privacy–Utility Trade-off:* To quantify the balance between privacy protection and image naturalness, we introduce a new metric called *H-PN Score* (Harmonic mean of Privacy and Naturalness). This metric is inspired by the F1 Score, commonly used in information retrieval, which employs the harmonic mean to reflect the trade-off between two possibly conflicting objectives. H-PN Score jointly considers privacy effectiveness, measured by MOSR, and image naturalness score, measured by a combined score of SSIM and normalized PSNR.

$$\text{H-PN Score} = \frac{2}{\frac{1}{\text{MOSR}} + \frac{1}{S_{nat}}}, \tag{22}$$

where $S_{nat}$ denotes the naturalness score computed as a weighted combination of SSIM and normalized PSNR:

$$S_{nat} = \lambda \cdot \text{SSIM} + (1 - \lambda) \cdot \text{PSNR}^*, \tag{23}$$

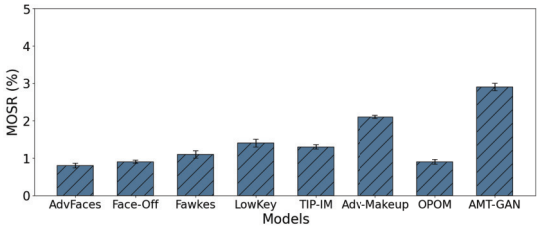



Fig. 5. Assessing the efficacy of advanced face de-identification techniques in facial attribute obfuscation.

where $\lambda \in [0, 1]$ is a balancing coefficient (defaulting to 0.5), and $\text{PSNR}^*$ is the PSNR value normalized to the range $[0, 1]$.

*4) User Study:* To validate the utility of Invisible-Face and baseline methods, we conduct a user study with 30 volunteers. These volunteers, including 10 doctoral students in computer vision, 10 average internet users, and 10 professional internet users, evaluate 30 randomly selected sets of images processed by Invisible-Face and baseline methods. Using two survey questions from Hasan et al. [65], we assess each method's utility:

- **Information Sufficiency:** How much original information does the processed image preserve compared to the original image?
- **Visual Naturalness:** How much original visual naturalness does the processed image maintain compared to the original image?

Information sufficiency assesses the perceived information retention after privacy protection, while visual naturalness measures the visual difference between the protected and original images. Participants rate their responses on a scale from 1 to 10, where higher scores indicate greater information sufficiency and visual naturalness. To ensure an unbiased evaluation, participants are not informed about the processing methods, and the order of image presentation is randomized to prevent sequence effects.

## VII. Evaluation

### A. Effectiveness in White-Box Scenarios

Before examining the effectiveness of Invisible-Face, it is essential to evaluate current privacy-preserving methods related to face de-identification, given the uncertain efficacy of these techniques in safeguarding facial attributes. We comprehensively assess 8 SOTA face de-identification methods from the past three years. The results in Figure 5 reveal a notably low success rate for these methods in anonymizing facial attributes, falling below 3%. This indicates that existing face de-identification methods are inadequate for tasks requiring robust protection of facial attribute privacy. Anonymizing identity and facial attributes are distinct challenges, and current measures do not sufficiently safeguard facial attribute privacy.

The assessment results in Table I provide a detailed analysis of different methods in white-box scenarios. The evaluation focuses on three facial attributes: gender, age, and race. The results demonstrate that almost all methods have high MOSR. However, their performance declines when faced with target models with a large number of parameters, approximately 138

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

TABLE I

MOSR OF INVISIBLE-FACE AND BASELINE METHODS FOR OBFUSCATING THREE FACIAL ATTRIBUTES ON THE UTKFACE DATASET

| Method | Target Model | | | | | |
|---|---|---|---|---|---|---|
| | ResNet-18 | ResNet-34 | ResNet-50 | MobileNet | GoogLeNet | VGG-16 |
| Invisible-Face | **100.0%** | **100.0%** | **100.0%** | 99.9% | 99.9% | 97.5% |
| FGSM | 98.4% | 98.3% | 96.2% | 84.6% | 87.3% | 71.5% |
| PGD | **100.0%** | **100.0%** | **100.0%** | 99.9% | **100.0%** | 80.5% |
| MI | **100.0%** | **100.0%** | **100.0%** | 99.9% | **100.0%** | 79.7% |
| TI | **100.0%** | 99.9% | 98.4% | 99.0% | 99.8% | 74.9% |
| VMI | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | 80.9% |
| LowKey-*m* | 93.0% | 93.0% | 93.0% | 89.6% | 87.9% | 76.8% |
| Fawkes-*m* | 83.1% | 82.9% | 83.2% | 80.7% | 81.3% | 69.3% |
| TIP-IM-*m* | 88.5% | 87.8% | 87.8% | 83.5% | 83.1% | 72.8% |
| PrivacyNet | 53.7% | 53.9% | 51.9% | 47.5% | 50.3% | 46.8% |
| Chhabra *et al.* | 97.5% | 96.6% | 96.6% | 90.7% | 90.7% | 78.1% |
| DeepPrivacy | 91.0% | 91.0% | 91.0% | 91.0% | 91.0% | 91.0% |
| PI-Net | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** |
| CollDiff | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** | **100.0%** |
| Zhang *et al.* | **89.5%** | **89.5%** | **89.6%** | **90.0%** | **89.8%** | **89.0%** |

million, such as VGG-16, resulting in a MOSR decrease of about 20%. Despite optimizing adversarial loss concerning facial attributes for LowKey-*m*, Fawkes-*m*, and TIP-IM-*m*, their MOSR performance remains suboptimal, with the mean MOSR below 90%. PrivacyNet [16], which trains its own discriminator to assess privacy protection success, does not rely on specific white-box models. As a result, its performance is slightly lower compared to other methods. In contrast, facial editing methods such as DeepPrivacy, PI-Net, CollDiff, and Zhang et al. demonstrate effectiveness and stability, with an average MOSR exceeding 90%. Invisible-Face performs consistently well across six target models, particularly under VGG-16, where its performance significantly surpasses most baseline methods.

### B. Naturalness

We evaluate the naturalness of images generated by Invisible-Face and baseline methods. The evaluation results are presented in Table II. Several key insights are observed after careful examination and analysis of the results. Methods employing perturbation-based adversarial attacks and transfer attacks still fall short in maintaining the natural appearance of images, with an average SSIM of approximately 0.71 and a mean PSNR of about 27.3, indicating perceptible perturbations [61]. Similarly, a third set of anonymization methods shows poor performance in maintaining image naturalness, as they do not consider the balance between image naturalness and privacy protection effectiveness. However, Chhabra et al. stands out, as it draws inspiration from C&W algorithm [66], which enables preservation of image naturalness. However, its MOSR performance is not impressive. Interestingly, the fourth group, which includes face-editing-based methods, shows the greatest decline in preserving image naturalness, with an SSIM index dropping to 0.5 or lower, indicating substantial alterations to the image structure.

Invisible-Face generates perturbations that have minimal impact on the overall appearance of the images. Compared to the baseline methods, Invisible-Face results in an average increase of 24.4% and 24.7% for SSIM and PSNR values, respectively, illustrating the efficacy of our proposed method.

The results we have obtained visually further highlight the disparities in various approaches. As illustrated in Figure 6, the image processed by the Invisible-Face displays the least visual disparity compared to the original. While certain explicit attributes like gender remain perceptible to the human eye, our primary objective is to impede the precise detection of user attributes by unauthorized facial attribute detection models, while preserving the visual naturalness of the user's face image.

TABLE II

SSIM AND PSNR OF IMAGES PROCESSED BY INVISIBLE-FACE AND BASELINE METHODS FOR OBFUSCATING THREE FACIAL ATTRIBUTES ON THE UTKFACE DATASET

| Method | Metric | Target Model | | | | | |
|---|---|---|---|---|---|---|---|
| | | ResNet-18 | ResNet-34 | ResNet-50 | MobileNet | GoogLeNet | VGG-16 |
| Invisible-Face | SSIM | **0.94** | **0.93** | **0.93** | **0.92** | **0.94** | **0.91** |
| | PSNR | **37.01** | **36.95** | **36.95** | **36.55** | **36.49** | **36.28** |
| FGSM | SSIM | 0.62 | 0.63 | 0.63 | 0.66 | 0.63 | 0.73 |
| | PSNR | 26.25 | 26.26 | 26.59 | 27.02 | 26.34 | 27.17 |
| PGD | SSIM | 0.73 | 0.72 | 0.74 | 0.77 | 0.78 | 0.78 |
| | PSNR | 28.55 | 28.23 | 28.72 | 29.46 | 28.27 | 28.89 |
| MI | SSIM | 0.67 | 0.66 | 0.66 | 0.68 | 0.70 | 0.77 |
| | PSNR | 27.30 | 27.26 | 27.21 | 27.94 | 27.58 | 28.01 |
| TI | SSIM | 0.73 | 0.74 | 0.75 | 0.76 | 0.76 | 0.82 |
| | PSNR | 26.57 | 26.58 | 26.79 | 26.82 | 26.80 | 27.49 |
| VMI | SSIM | 0.67 | 0.67 | 0.67 | 0.69 | 0.69 | 0.77 |
| | PSNR | 27.11 | 27.13 | 27.26 | 27.78 | 27.37 | 28.83 |
| LowKey-*m* | SSIM | 0.64 | 0.69 | 0.70 | 0.73 | 0.79 | 0.75 |
| | PSNR | 25.60 | 26.42 | 27.07 | 27.34 | 28.49 | 26.80 |
| Fawkes-*m* | SSIM | 0.77 | 0.77 | 0.76 | 0.76 | 0.79 | 0.70 |
| | PSNR | 29.03 | 28.96 | 28.99 | 29.16 | 29.63 | 28.37 |
| TIP-IM-*m* | SSIM | 0.73 | 0.74 | 0.74 | 0.78 | 0.71 | 0.69 |
| | PSNR | 26.45 | 25.98 | 26.23 | 26.87 | 25.73 | 25.66 |
| PrivacyNet | SSIM | 0.74 | 0.74 | 0.74 | 0.74 | 0.74 | 0.74 |
| | PSNR | 23.72 | 23.72 | 23.72 | 23.72 | 23.72 | 23.72 |
| Chhabra *et al.* | SSIM | 0.87 | 0.88 | 0.88 | 0.87 | 0.86 | 0.87 |
| | PSNR | 32.98 | 32.98 | 32.81 | 32.94 | 33.23 | 33.31 |
| DeepPrivacy | SSIM | 0.34 | 0.34 | 0.34 | 0.33 | 0.33 | 0.34 |
| | PSNR | 10.68 | 10.70 | 10.67 | 10.53 | 10.56 | 10.67 |
| PI-Net | SSIM | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 | 0.50 |
| | PSNR | 13.74 | 13.74 | 13.74 | 13.74 | 13.74 | 13.71 |
| CollDiff | SSIM | 0.43 | 0.43 | 0.43 | 0.43 | 0.43 | 0.43 |
| | PSNR | 12.19 | 12.19 | 12.19 | 12.19 | 12.19 | 12.19 |
| Zhang *et al.* | SSIM | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 | 0.44 |
| | PSNR | 13.28 | 13.28 | 13.28 | 13.28 | 13.28 | 13.28 |

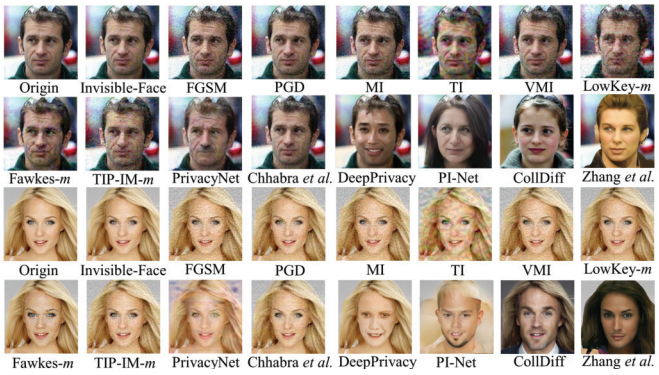



Fig. 6. Visualization of images processed by invisible-face and baseline methods.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

TABLE III

HUMAN EVALUATION RESULTS ABOUT INVISIBLE-FACE AND BASELINE METHODS. AVG DENOTES AVERAGE, AND SD DENOTES STANDARD DEVIATION

| Method | Information Sufficiency | | Visual Naturalness | |
|---|---|---|---|---|
| | Avg (↑) | SD (↓) | Avg (↑) | SD (↓) |
| Invisible-Face | **9.3** | 0.5 | **9.6** | **0.5** |
| FGSM | 8.1 | 1.3 | 6.8 | 0.9 |
| PGD | 8.5 | 1.0 | 8.2 | 0.8 |
| MI | 8.0 | 0.9 | 7.0 | 0.9 |
| TI | 6.3 | 1.8 | 4.9 | 0.7 |
| VMI | 8.1 | 0.9 | 7.3 | 0.5 |
| LowKey-*m* | 7.3 | 1.8 | 6.5 | 0.8 |
| Fawkes-*m* | 7.8 | 1.6 | 6.2 | 1.1 |
| TIP-IM-*m* | 7.1 | 1.3 | 6.7 | 0.8 |
| PrivacyNet | 8.0 | 1.5 | 6.9 | 1.3 |
| Chhabra *et al.* | 8.8 | 0.7 | 9.0 | **0.5** |
| DeepPrivacy | 1.2 | **0.4** | 2.3 | 1.9 |
| PI-Net | 1.5 | 0.5 | 3.1 | 1.9 |
| CollDiff | 1.3 | 0.5 | 2.8 | 2.4 |
| Zhang *et al.* | 1.5 | 0.5 | 3.3 | 2.1 |

Table III describes the human evaluation results for various privacy-preserving baseline methods. The evaluation reveals a pronounced preference for perturbation-based methods when considering information sufficiency. Except for Invisible-Face, high standard deviations indicate that significant perturbations may impair the user's ability to discern the original content of the image. The trend of visual naturalness evaluation results is similar to information sufficiency, with most users preferring perturbation-based methods. The reason is that they like to maintain the original appearance of the person in the image. Conversely, face-editing-based anonymization methods garner limited favor, as only a few participants perceive these as offering enhanced personalization. Most participants believe that the face-editing-based anonymization methods are unsuitable for social media photo sharing scenarios. This preference variance has resulted in notable standard deviations in assessing face-editing methods. Invisible-Face outperforms other baseline methods in the overall evaluation results, aligning with the analytical outcomes presented in Table II.

## C. Transferability

To gauge both the transferability of Invisible-Face and baseline methods, we undertake an array of experiments incorporating six different models. Using one model as a surrogate, we generate face images with perturbations and subsequently test the adversarial perturbation's transferability across the remaining five target transfer models. Figure 7 shows the average MOSR values for all tested methods when transferring from the surrogate to the target model. Among these baseline methods, LowKey-*m*, TIP-IM-*m*, and Fawkes-*m* show some promise, keeping their MOSR decrease within a 50% to 62.5% range. Conversely, other methods witness substantial performance declines post-transfer, with MOSR dropping by over 60%. In contrast, Invisible-Face only experiences an average decrease in MOSR of 17.2%. These findings suggest that Invisible-Face outperforms baseline methods regarding transferability across various surrogate models.

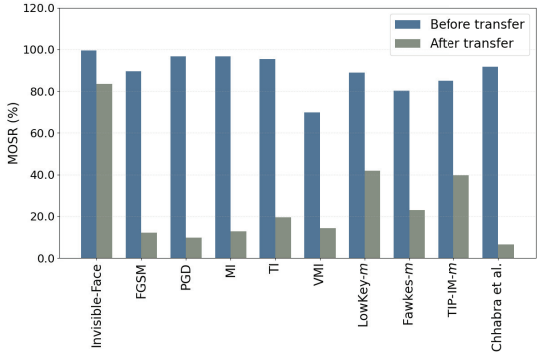



Fig. 7. Comparative analysis of mean MOSR before and after transfer. Note that methods like face-editing-based anonymization methods and PrivacyNet, which do not depend on the target model for privacy protection, are not suitable for this analysis.

TABLE IV

COMPARISON OF DIFFERENT METHODS ACROSS MLAAS PLATFORMS

| Metric | Method | Baidu Brain | Tencent Cloud | Aliyun | Face$^{++}$ |
|---|---|---|---|---|---|
| MOSR | Invisible-Face | 47.9% | 51.2% | 58.8% | 50.7% |
| | FGSM | 18.1% | 25.7% | 25.5% | 20.6% |
| | PGD | 18.7% | 26.0% | 29.1% | 20.7% |
| | MI | 18.9% | 25.9% | 28.3% | 23.9% |
| | TI | 32.8% | 30.1% | 36.5% | 32.2% |
| | VMI | 20.5% | 24.9% | 26.8% | 22.2% |
| | LowKey-*m* | 25.4% | 29.8% | 33.0% | 28.7% |
| | Fawkes-*m* | 19.7% | 23.3% | 26.5% | 19.0% |
| | TIP-IM-*m* | 22.2% | 29.3% | 31.9% | 21.9% |
| | PrivacyNet | 40.5% | 43.2% | 44.9% | 43.5% |
| | Chhabra *et al.* | 16.1% | 17.6% | 17.8% | 16.5% |
| | DeepPrivacy | 81.5% | 83.3% | 86.2% | 84.9% |
| | PI-Net | 85.2% | 88.6% | 88.8% | 88.6% |
| | CollDiff | **94.7%** | **95.8%** | **95.9%** | **95.8%** |
| | Zhang *et al.* | 86.5% | 86.8% | 87.3% | 86.0% |
| H-PN Score | Invisible-Face | **0.59** | **0.61** | **0.66** | **0.61** |
| | FGSM | 0.27 | 0.34 | 0.34 | 0.30 |
| | PGD | 0.29 | 0.36 | 0.39 | 0.31 |
| | MI | 0.28 | 0.35 | 0.37 | 0.33 |
| | TI | 0.41 | 0.39 | 0.43 | 0.41 |
| | VMI | 0.30 | 0.34 | 0.36 | 0.32 |
| | LowKey-*m* | 0.35 | 0.39 | 0.41 | 0.38 |
| | Fawkes-*m* | 0.30 | 0.34 | 0.37 | 0.29 |
| | TIP-IM-*m* | 0.32 | 0.39 | 0.41 | 0.32 |
| | PrivacyNet | 0.46 | 0.48 | 0.49 | 0.48 |
| | Chhabra *et al.* | 0.26 | 0.28 | 0.28 | 0.27 |
| | DeepPrivacy | 0.34 | 0.34 | 0.35 | 0.35 |
| | PI-Net | 0.42 | 0.42 | 0.42 | 0.42 |
| | CollDiff | 0.38 | 0.38 | 0.38 | 0.38 |
| | Zhang *et al.* | 0.41 | 0.41 | 0.41 | 0.41 |

## D. Effectiveness in Black-Box Scenarios

To assess the efficacy of Invisible-Face in real-world scenarios, our study further validates Invisible-Face and baseline methods against four popular MLaaS platforms: Baidu Brain, Tencent Cloud, Aliyun, and Face$^{++}$. Given these services' facial attribute detection capabilities, we focus our testing on three specific attributes: gender, age, and facial expression. Table IV illustrates the results. The MOSR values indicate the effectiveness of each method in preserving facial privacy. Among these methods, CollDiff outperforms all others.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

TABLE V
ABLATION STUDY OF INVISIBLE-FACE ON PERTURBATION REGION SELECTION STRATEGIES

| Metric | Method | |
|---|---|---|
| | Full Perturbations | Local Perturbations |
| MOSR | 100.0% | 100.0% |
| SSIM | 0.87 | 0.94 |
| PSNR | 33.26 | 37.01 |

TABLE VI
ABLATION STUDY OF INVISIBLE-FACE ON DIFFERENT MODULES

| Metric | Method | | | |
|---|---|---|---|---|
| | Invisible-Face | Invisible-Face w/o DP | Invisible-Face w/o RP | Invisible-Face w/o DP&RP |
| MOSR | 91.8% | 57.6% | 69.3% | 51.3% |

TABLE VII
ABLATION STUDY ON THE IMPACT OF LOSS FUNCTIONS IN INVISIBLE-FACE

| Method | MOSR | SSIM | PSNR |
|---|---|---|---|
| Invisible-Face | 100.0% | 0.94 | 37.01 |
| w/o $\mathcal{L}_{sem}$ | 97.3% | 0.88 | 35.40 |
| w/o $\mathcal{L}_{nat}$ | 100.0% | 0.81 | 29.72 |
| w/o $\mathcal{L}_{per}$ | 100.0% | 0.90 | 36.12 |

TABLE VIII
THE AVERAGE TIME COST (IN SECONDS) TO PROCESS ONE IMAGE FOR OBFUSCATING THREE FACIAL ATTRIBUTES. THE COMPUTATION IS ACCELERATED BY ONE NVIDIA RTX 3090 GPU

| Metric | Component | | |
|---|---|---|---|
| | Local Region Selection | Perturbation Generation | Total Time |
| Time Cost (s) | 1.15 | 0.3 | 1.45 |

Although Invisible-Face achieves only a moderate MOSR, it has the highest value among all perturbation-based anonymization methods. In contrast, FGSM, PGD, and Chhabra et al. show much lower MOSR values, suggesting they are less effective against black-box attacks. PrivacyNet, due to its independence from white-box environments, performs well under black-box attacks.

In terms of balancing privacy effectiveness and image naturalness, Invisible-Face achieves the highest H-PN Score across all platforms, ranging from 0.59 to 0.66. This demonstrates its strong ability to strike a balance between privacy protection and visual fidelity. In contrast, methods based on face-editing anonymization (such as DeepPrivacy, PI-Net, CollDiff, and Zhang et al.) show much lower H-PN Scores, indicating that they struggle to preserve image naturalness while ensuring effective privacy protection. This analysis further highlights the limitations of face-editing-based anonymization methods in scenarios like social media photo sharing, where both privacy and image realism are crucial. These methods fail to balance effective privacy protection with visual naturalness, making them less suitable for such contexts.

## E. Ablation Study

To deeply explore the performance of Invisible-Face, we conduct several ablation studies. We use multiple variants of Invisible-Face to obfuscate the gender, age, and race attributes of facial images in the UTKFace dataset. Finally, adversaries use ResNet-18 as an attack model to detect these facial attributes.

We first conduct an ablation study comparing full-image perturbations and local perturbations generated using our heuristic-based local region selection algorithm. As shown in Table V, both methods achieve accurate obfuscation in terms of MOSR. However, full-image perturbations noticeably reduce image quality, while our local perturbation strategy significantly improves SSIM and PSNR, indicating better visual naturalness. These results confirm that local perturbations offer a superior privacy-utility trade-off.

We then evaluate the impact of removing the DP module and the RP module individually and jointly. To evaluate the effectiveness of the RP module, we simulate potential adversarial interference by applying the DnCNN denoising model [67] to the perturbed images. As shown in Table VI, the removal of either DP or RP leads to a significant performance drop. This confirms that the DP module is critical for enhancing obfuscation effectiveness, while the RP module improves robustness against perturbation degradation.

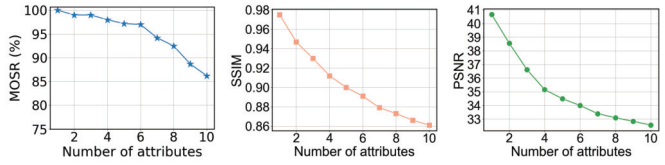
We further evaluate the contribution of the three main loss terms: semantic loss ($\mathcal{L}_{sem}$), naturalness loss ($\mathcal{L}_{nat}$), and perturbation loss ($\mathcal{L}_{per}$). As shown in Table VII, removing $\mathcal{L}_{sem}$ slightly reduces MOSR and moderately degrades SSIM and PSNR, indicating its role in preserving facial semantic structure. Removing $\mathcal{L}_{nat}$ does not impact MOSR but leads to a substantial decline in SSIM and PSNR, suggesting it is essential for maintaining perceptual naturalness. Excluding $\mathcal{L}_{per}$ has minimal effect on MOSR but results in moderate degradation in visual quality, confirming its importance in constraining the spatial distribution of perturbations.

## F. Time Cost

To evaluate the computational overhead of Invisible-Face, we conduct runtime measurements on a system equipped with a single NVIDIA RTX 3090 GPU. As shown in Table VIII, when obfuscating three facial attributes (gender, age, and race), the local region search via simulated annealing takes approximately 1.15 seconds per image, and the total average processing time per image is about 1.45 seconds. While the current runtime does not meet real-time constraints, we believe it is acceptable and practical for our target scenario—privacy-preserving image sharing on social media platforms. In such applications, the protection process typically occurs at the moment of image upload or publication, and a short processing delay (on the order of 1–2 seconds) is generally tolerable.

However, for scenarios with higher real-time requirements, such as online meetings, the current latency still presents a challenge. A promising direction is to train universal pertur-

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.




(a) MOSR (b) SSIM (c) PSNR

Fig. 8. Stability evaluation of invisible-face for obfuscating 1 to 10 facial attributes on the CelebA dataset.

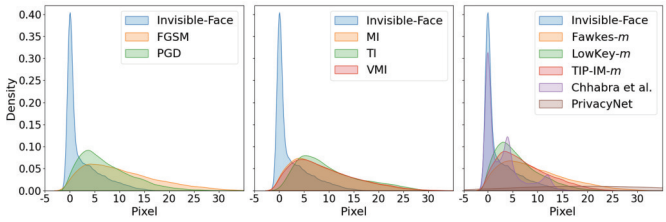



Fig. 9. Kernel density estimation for perturbation distributions. Note that perturbation analysis is only applicable to perturbation-based anonymization methods.

bations to reduce the frequency of perturbation generation, and to reuse local regions selected via simulated annealing across adjacent inputs, thereby minimizing redundant computations and improving efficiency. We plan to further explore and refine this direction in future work.

### G. Stability

As the number of facial attributes that users may wish to protect varies in photo-sharing scenarios, we assess the stability of Invisible-Face across 1 to 10 attributes from the CelebA dataset (*attractive*, *black hair*, *blond hair*, *chubby*, *eyeglasses*, *heavy makeup*, *male*, *no beard*, *oval face*, *young*). As shown in Figure 8, although MOSR gradually decreases with more attributes, it remains above 86% even when protecting 10 attributes, indicating consistent privacy performance. Meanwhile, SSIM and PSNR remain high, suggesting that the perturbations are highly localized and visually unobtrusive. These results confirm that Invisible-Face effectively balances privacy and visual naturalness with more protected attributes.

### H. Perturbation Analysis

In our investigation of the relationship between perturbations and the naturalness of images, we conduct a visualization analysis to graphically depict the dispersion of perturbed pixel values. These results are illustrated in kernel density plots, as shown in Figure 9. The distributions derived from three distinct perturbation sets show that the perturbed pixel values produced by Invisible-Face closely mirror a narrow peak of a normal distribution centered at 0. It signifies that most pixel values densely congregate around 0. This observation implies that the perturbation is fine-grained, helping to keep the original features of the image intact. It also explains why Invisible-Face imposes the slightest influence on image naturalness. In contrast, other baseline methods exhibit broader perturbation distributions, ranging from 5 to 10 or more. These wider peaks indicate coarser perturbations. Notably, Chhabra et al. shows slightly narrower perturbations, which aligns with the image naturalness observed in the evaluation results.

TABLE IX

EFFECTIVENESS CHANGE OF INVISIBLE-FACE BEFORE AND AFTER IMAGE DENOISING

| Metric | Original | DnCNN | BSRGAN | FFdnet |
|---|---|---|---|---|
| MOSR | 100.0% | 91.6% | 89.8% | 90.5% |

TABLE X

IMPACT OF OTHER ADAPTIVE ATTACKS ON THE MOSR OF INVISIBLE-FACE. II DENOTES IMAGE INPAINTING, IR DENOTES IMAGE RECONSTRUCTION, AND DST DENOTES DOMAIN-SPECIFIC TRANSFORMATION

| Metric | Original | II | IR | DST |
|---|---|---|---|---|
| MOSR | 100.0% | 69.7% | 83.9% | 80.7% |

## VIII. POSSIBLE ADAPTIVE ATTACKS

We explore potential adaptive attacks against Invisible-Face. Under our threat model, We assume that adversaries can obtain pairs of protected and unprotected images. To counteract the perturbations introduced by Invisible-Face, adversaries aim to train a denoising model that takes the protected image as input and reconstructs the unprotected version. We also assume that the adversary uses ResNet-18 as the target model.

To simulate such adaptive attacks, we employ three prevalent denoising models—DnCNN [67], BSRGAN [68], and FFdNet [69]—to train denoisers specifically against the perturbations introduced by Invisible-Face. After denoising, adversaries apply their facial attribute detection models to assess the effectiveness of the attack.

Our experimental results (see Table IX) under image denoising show that MOSR decreased only approximately 10% after denoising, compared to before the process. Invisible-Face maintains robust resistance to these attempts. We attribute the robustness of Invisible-Face against denoising to two main factors. First, the RP module in Invisible-Face introduces degradation simulation during training, which helps the perturbation generalize against common post-processing. Second, the perturbations created by Invisible-Face are granular, rendering it difficult for the denoising model to accurately learn the noise information during training. Consequently, image denoising is ineffective against Invisible-Face.

We also consider other adaptive attack strategies introduced in PrivacyProber [70], including image inpainting, image reconstruction, and domain-specific transformation (i.e., background removal), to evaluate their effectiveness in disrupting the perturbations generated by Invisible-Face. These attacks are implemented following the configurations adopted in PrivacyProber.

As shown in Table X, the MOSR of Invisible-Face decreases to 69.7% under image inpainting, suggesting that this adaptive attack has a noticeable impact on privacy protection, which aligns with observations reported in PrivacyProber. However, Invisible-Face retains a considerable degree of MOSR, indicating some degree of resilience under this challenging setting.

For image reconstruction, we follow the structure-to-signal autoencoder [70], [71], which reconstructs images from low-

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

level structural features to suppress semantic gradients and remove adversarial noise. While effective against typical gradient-based perturbations, this approach struggles with Invisible-Face, where adversarial signals are embedded in non-structural features such as texture and localized contrast. As a result, the reconstructed images lack sufficient semantic cues for accurate attribute recovery, leading to a moderately reduced MOSR of 83.9%. Domain-specific transformation, implemented via background removal, has some effect on disrupting facial attribute obfuscation. However, Invisible-Face still retains a moderate level of obfuscation performance, as its local perturbation strategy is specifically designed to concentrate adversarial perturbations within facial regions, making it less affected by background-based transformations.

## IX. Discussions

### A. Effectiveness and Utility

This paper addresses the challenge of protecting privacy in facial images shared online while maintaining the user experience. User studies show that most participants prefer minimal visual differences between the original and privacy-enhanced images, as excessive alterations can distort the intended message. In social media photo sharing, utility is closely linked to the visual naturalness of the images [5], [8], [65]. Users generally want to share facial images that remain recognizable and realistic to their friends and social contacts. For instance, when posting a group photo on platforms like Instagram, users may wish to hide sensitive attributes such as gender or age, while still ensuring the photo reflects their true self. If privacy protection distorts the image too much, making it appear synthetic or unnatural, the photo loses its communicative and social value. Therefore, maintaining visual naturalness is a crucial aspect of utility in this context.

### B. Adversarial Attribute Inversion

While the CelebA dataset focuses on binary attributes, our work uses the UTKFace dataset, which includes multi-class and continuous attributes like age (discretized into ordinal categories) and race (with five categories). This makes inversion attacks, which may apply to binary attributes like gender, less feasible. For binary attributes, we acknowledge the theoretical concern that adversaries can invert predictions if aware of the image modification. However, our method aims to reduce classifier confidence and promote misclassification, making inversion attacks more challenging, especially in black-box scenarios where adversaries lack model access.

### C. Applications in Real-World Scenarios

This work targets social media platforms, where users aim to protect facial attribute privacy while maintaining the natural appearance of shared photos. This concern has been highlighted by incidents like the Facebook-Cambridge Analytica scandal [72], where inferred biographical traits enabled unauthorized political targeting. Beyond social media, Invisible-Face also applies to other domains where privacy and image naturalness are critical, such as online identity verification (e.g., KYC systems) and retail environments using facial recognition for customer analysis. Invisible-Face mitigates these privacy risks while preserving visual authenticity.

### D. Future Work

In this paper, we introduce a tool to protect the privacy of facial attributes when users share photos online, ensuring that images maintain their natural appearance—an aspect overlooked by prior studies. However, similar to earlier face confusion systems [73], Invisible-Face may also be prone to biases, an issue we plan to explore in future work.

### E. Ethics Consideration

We conduct a user study to assess the naturalness of facial images. According to the research plan, the leading institution solely conducts the survey. We follow principles outlined in the Menlo Report [74] and the local regulations to protect the rights of human participants.

## X. Conclusion

In this work, we identify that existing face privacy protection measures are inadequate for defending against facial attribute privacy vulnerabilities. To address this issue, we propose a tool named "Invisible-Face", which protects users' facial attribute privacy while preserving the utility of shared photos online. This method applies imperceptible perturbations to a user's facial image, causing unauthorized malicious facial attribute recognition models to yield incorrect attribute information. Through extensive testing on various benchmark datasets, Invisible-Face demonstrates both efficacy and practicality. This tool promotes enhanced privacy protection in social networking applications and contributes to fostering a safer online environment.

## References

[1] C. Vijayalakshmi, S. M. Florence, D. Gokulakrishnan, G. Prakash, and A. Ponmalar, "Face detection for secure online payment with proxy detection," in *Proc. Int. Conf. Commun., Comput. Internet Things (IC3IoT)*, Mar. 2022, pp. 1–4.

[2] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Dec. 2009, pp. 1–8.

[3] C. Chai and M. Wang, "Neck entertainment fitness system based on face recognition," in *Proc. IEEE 11th Int. Conf. Computer-Aided Ind. Design Conceptual Design*, vol. 1, Nov. 2010, pp. 792–796.

[4] V. Singh, R. Shanmugam, and S. Awasthi, "Preventing fake accounts on social media using face recognition based on convolutional neural network," in *Proc. Sustain. Commun. Netw. Appl., ICSCN*, Jan. 2021, pp. 227–241.

[5] C. Liu, T. Zhu, J. Zhang, and W. Zhou, "Privacy intelligence: A survey on image privacy in online social networks," *ACM Comput. Surveys*, vol. 55, no. 8, pp. 1–35, Aug. 2023.

[6] 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. (USENIX Secur.)*, 2020, pp. 1589–1604.

[7] V. Cherepanova et al., "LowKey: Leveraging adversarial attacks to protect social media users from facial recognition," 2021, *arXiv:2101.07922*.

[8] X. Yang et al., "Towards face encryption by generating adversarial identity masks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3877–3887.

[9] Y. Gurovich et al., "Identifying facial phenotypes of genetic disorders using deep learning," *Nature Med.*, vol. 25, no. 1, pp. 60–64, Jan. 2019.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

[10] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *J. Personality Social Psychol.*, vol. 114, no. 2, pp. 246–257, Feb. 2018.

[11] J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "Online advertising: Analysis of privacy threats and protection approaches," *Comput. Commun.*, vol. 100, pp. 32–51, Mar. 2017.

[12] M. Crain and A. Nadler, "Political manipulation and Internet advertising infrastructure," *J. Inf. Policy*, vol. 9, pp. 370–410, Dec. 2019.

[13] G. Ma, K. Li, Q. Pei, and Y. Zhan, "A fine-grained face privacy protection scheme in social networks," *Netinfo Secur.*, vol. 17, no. 8, pp. 26–32, 2017.

[14] 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.

[15] H.-P. Wang, T. Orekondy, and M. Fritz, "InfoScrub: Towards attribute privacy by targeted obfuscation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3281–3289.

[16] 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.

[17] Y. Zhang, T. Wang, R. Zhao, W. Wen, and Y. Zhu, "RAPP: Reversible privacy preservation for various face attributes," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3074–3087, 2023.

[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[20] F. Saxen, P. Werner, S. Handrich, E. Othman, L. Dinges, and A. Al-Hamadi, "Face attribute detection with MobileNetV2 and NasNet-mobile," in *Proc. 11th Int. Symp. Image Signal Process. Anal. (ISPA)*, Sep. 2019, pp. 176–180.

[21] K. Karkkainen 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. 1548–1558.

[22] Y. Gurovich et al., "DeepGestalt–identifying rare genetic syndromes using deep learning," 2018, *arXiv:1801.07637*.

[23] (2022). *Global News*. [Online]. Available: https://globalnews.ca/news/8653716/first-nations-man-files-human-rights-complaint-td-bank-bc/

[24] (2017). *Harvard Chan School*. [Online]. Available: https://www.hsph.harvard.edu/news/press-releases/poll-lgbtq-americans-discrimination/

[25] (2023). *FTC*. [Online]. Available: https://www.ftc.gov/news-events/news/press-releases/2023/05/ftc-warns-about-misuses-biometric-information-harm-consumers

[26] (2021). *EDPB*. [Online]. Available: https://www.efdpo.eu/edpb-edps-call-for-ban-on-use-of-ai-for-automated-recognition-of-human-features-in-publicly-accessible-spaces-and-some-other-uses-of-ai-that-can-lead-to-unfair-discrimination/

[27] S. Chhabra, R. Singh, M. Vatsa, and G. Gupta, "Anonymizing k-Facial attributes via adversarial perturbations," 2018, *arXiv:1805.09380*.

[28] T. Wang, Y. Zhang, Z. Yang, X. Xiao, H. Zhang, and Z. Hua, "Seeing is not believing: An identity hider for human vision privacy protection," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 7, no. 2, pp. 170–181, Apr. 2024.

[29] J. Li, H. Zhang, S. Liang, P. Dai, and X. Cao, "Privacy-enhancing face obfuscation guided by semantic-aware attribution maps," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 3632–3646, 2023.

[30] B. Meden et al., "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4147–4183, 2021.

[31] J.-W. Chen, L.-J. Chen, C.-M. Yu, and C.-S. Lu, "Perceptual indistinguishability-net (PI-Net): Facial image obfuscation with manipulable semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6474–6483.

[32] Z. Huang, K. C. K. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 6080–6090.

[33] Y. Zhang, J. Ji, W. Wen, Y. Zhu, Z. Xia, and J. Weng, "Understanding visual privacy protection: A generalized framework with an instance on facial privacy," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 5046–5059, 2024.

[34] D. A. Meyer and M. W. Quong, "The bio-logic of facial geometry," *Nature*, vol. 397, no. 6721, pp. 661–662, Feb. 1999.

[35] Z. Zhuang, D. Landsittel, S. Benson, R. Roberge, and R. Shaffer, "Facial anthropometric differences among gender, ethnicity, and age groups," *Ann. Occupational Hygiene*, vol. 54, no. 4, pp. 391–402, 2010.

[36] P. R. Husmann and D. R. Samson, "In the eye of the beholder: Sex and race estimation using the human orbital Aperture," *J. Forensic Sci.*, vol. 56, no. 6, pp. 1424–1429, Nov. 2011.

[37] Y.-C. Chen, X. Shen, Z. Lin, X. Lu, I.-M. Pao, and J. Jia, "Semantic component decomposition for face attribute manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9851–9859.

[38] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," 2018, *arXiv:1801.02610*.

[39] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, Dec. 2013, pp. 315–323.

[40] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1924–1933.

[41] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, Jul. 2006.

[42] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on ArcFace face ID system," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 819–826.

[43] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[44] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 768–783.

[45] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," 2020, *arXiv:2003.05991*.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Cham, Switzerland. Springer, 2015, pp. 234–241.

[47] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 5810–5818. Accessed: Mar. 18, 2024.

[48] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738. Accessed: Mar. 18, 2024.

[49] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. Accessed: Mar. 18, 2024.

[50] 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. Accessed: Mar. 18, 2024.

[51] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[52] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[54] *Baidu Brain*. [Online]. Available: https://ai.baidu.com/tech/face/detect

[55] *Tencent Cloud*. [Online]. Available: https://cloud.tencent.com/product/facerecognition

[56] *Aliyun*. [Online]. Available: https://ai.aliyun.com/face

[57] *Face*$^{++}$. [Online]. Available: https://www.faceplusplus.com.cn/face-detection/

[58] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[59] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4312–4321.

[60] F. Zhan et al., "Unbalanced feature transport for exemplar-based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15023–15033.

[61] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.

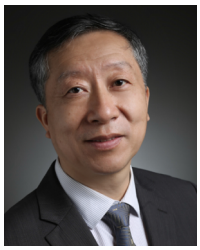

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.

[62] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2366–2369.
[63] J. W. Soh, G. Y. Park, J. Jo, and N. I. Cho, "Natural and realistic single image super-resolution with explicit natural manifold discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8114–8123.
[64] W. Zhou, Z. Wang, and Z. Chen, "Image super-resolution quality assessment: Structural fidelity versus statistical naturalness," in *Proc. 13th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2021, pp. 61–64.
[65] R. Hasan et al., "Viewer experience of obscuring scene elements in photos to enhance privacy," in *Proc. CHI Conf. Human Factors Comput. Syst.*, Apr. 2018, pp. 1–13.
[66] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.
[67] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.
[68] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4791–4800.
[69] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.
[70] P. Rot, K. Grm, P. Peer, and V. Štruc, "PrivacyProber: Assessment and detection of soft–biometric privacy–enhancing techniques," *IEEE Trans. Dependable Secure Comput.*, vol. 21, no. 4, pp. 2869–2887, Jul. 2024.
[71] J. Folz, S. Palacio, J. Hees, and A. Dengel, "Adversarial defense based on Structure-to-Signal autoencoders," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 3568–3577.
[72] (2018). *50 Million Facebook Profiles Harvested for Cambridge Analytica Major Data Breach*. [Online]. Available: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
[73] H. Rosenberg, B. Tang, K. Fawaz, and S. Jha, "Fairness properties of face recognition and obfuscation systems," in *Proc. 32nd USENIX Secur. Symp. (USENIX Secur.)*, Jan. 2021, pp. 7231–7248.
[74] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The menlo report," *IEEE Secur. Privacy*, vol. 10, no. 2, pp. 71–75, Mar. 2012.

**Yong Yang** received the master's degree from Shanghai Jiao Tong University in 2021. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University. His current research interests include AI security and privacy. He was a recipient of the IEEE S&P 2025 Distinguished Paper Award.

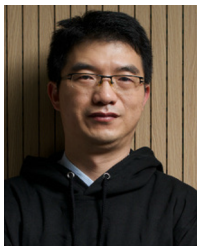
**Changjiang Li** (Graduate Student Member, IEEE) received the master's degree from the School of Computer Science, Zhejiang University, in 2020. He is currently pursuing the Ph.D. degree with the College of Computer Science, Stony Brook University. His research interests include adversarial machine learning and AI privacy.

**Yi Jiang** received the master's degree from the City University of Hong Kong. He is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University. His current research interests include AI security and privacy.



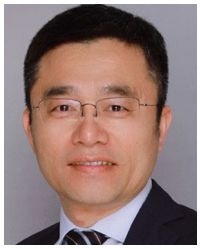
**Jinbao Li** (Member, IEEE) received the Ph.D. degree in computer science from Harbin Institute of Technology, China, in 2007. He is currently a Professor at the School of Computer Science and Technology, Qilu University of Technology, and Heilongjiang University. His research interests include databases, wireless sensor networks, mobile computing, and distributed computing. He received several awards, including the National Science and Technology Advancement Award, China, in 2005. He served on TPCs at conferences and editorial boards of journals.

**Xuhong Zhang** received the Ph.D. degree in computer engineering from the University of Central Florida in 2017. He is a ZJU 100-Young Professor with the School of Software Technology, Zhejiang University. He has authored more than 30 publications in premier journals and conferences, such as IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IEEE Symposium on Security and Privacy, USENIX Security, ACM CCS, NDSS, and VLDB. His research interests include distributed big data and AI systems, big data mining and analysis, data-driven security, and AI security.



**Zonghui Wang** received the Ph.D. degree from the College of Computer Science and Technology, Zhejiang University. He is a Senior Engineer with the College of Computer Science and Technology, Zhejiang University. His current research interests include deep learning, AI security, cloud computing, and computer architecture.

**Shouling Ji** (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology and the Ph.D. degree in computer science from Georgia State University. He is a Qiushi Distinguished Professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include AI security, software and system security, and data-driven security and privacy. He is a member of ACM and a Senior Member of CCF. He was a recipient of the 2012 Chinese Government Award for Outstanding Self-Financed Students Abroad and 10 Best/Outstanding Paper Awards, including the IEEE S&P 2025 Distinguished Paper Award and the ACM CCS 2021 Best Paper Award.



**Wenzhi Chen** (Member, IEEE) received the Ph.D. degree from the College of Computer Science and Technology, Zhejiang University. He is a Professor at the College of Computer Science and Technology and the Director of the Information Technology Center, Zhejiang University. He used to be the Vice Dean of the College of Computer Science and Technology, Zhejiang University. His current research interests include embedded systems and its application, computer architecture, computer system software, and information security. He is a member of ACM and ACM Education Council. He was a recipient of the NDSS 2024 Distinguished Paper Award.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:40:54 UTC from IEEE Xplore. Restrictions apply.