# Face De-identification: State-of-the-art Methods and Comparative Studies

Jingyi Cao, Xiangyi Chen, Bo Liu, *Senior Member, IEEE,* Ming Ding, *Senior Member, IEEE,* Rong Xie, *Member, IEEE,* Li Song, *Senior Member, IEEE*, Zhu Li, *Senior Member, IEEE,* Wenjun Zhang, *Fellow, IEEE*

arXiv:2411.09863v1 [cs.CV] 15 Nov 2024

***Abstract*—The widespread use of image acquisition technologies, along with advances in facial recognition, has raised serious privacy concerns. Face de-identification usually refers to the process of concealing or replacing personal identifiers, which is regarded as an effective means to protect the privacy of facial images. A significant number of methods for face de-identification have been proposed in recent years. In this survey, we provide a comprehensive review of state-of-the-art face de-identification methods, categorized into three levels: pixel-level, representation-level, and semantic-level techniques. We systematically evaluate these methods based on two key criteria, the effectiveness of privacy protection and preservation of image utility, highlighting their advantages and limitations. Our analysis includes qualitative and quantitative comparisons of the main algorithms, demonstrating that deep learning-based approaches, particularly those using Generative Adversarial Networks (GANs) and diffusion models, have achieved significant advancements in balancing privacy and utility. Experimental results reveal that while recent methods demonstrate strong privacy protection, trade-offs remain in visual fidelity and computational complexity. This survey not only summarizes the current landscape but also identifies key challenges and future research directions in face de-identification.**

***Index Terms*—Face de-identification, Image privacy, Deep learning.**

## I. Introduction

The rapid proliferation of artificial intelligence (AI) technologies, particularly deep learning-based models, has significantly transformed the landscape of face recognition systems. These advancements have brought about a wide range of applications, such as social media platforms. Unfortunately, accompanying these flourishing advancements, privacy concerns have also grown fast. Especially, faces are regarded as one of the most privacy-sensitive biological information directly related to personal identity. The essence of face recognition is biometric authentication, whose characteristics are unique and irrevocable. Once face recognition technology is used for cross-referencing with other databases, it will further disclose the user's other sensitive information. The previous study [1] has shown how faces can be used as the link across different databases and the trails associated with their different personas, thus violating privacy.

Facial privacy issues are receiving more attention these days. Restrictive laws and regulations such as the General Data Protection Regulations (GDPR) [2] have taken effect, which stipulates that data collection, sharing, and analysis by companies without the user's knowledge is considered illegal and regular consent is required for any use of their personal data to ensure data privacy. In GDPR, privacy information was defined as "personal data that are related to an identified or identifiable natural person", so personal identity is the most important part of facial image protection. To address this privacy issue, face de-identification (aka. face anonymization and face obfuscation) techniques have emerged as a crucial means of protecting privacy by anonymizing facial features in images without compromising their utility for non-identification tasks.

Face de-identification refers to the process of concealing or altering personally identifiable facial features to prevent recognition. These techniques have a broad range of applications, from protecting individuals' identities in media interviews and video surveillance [3], and medical research [4], to ensuring privacy in surveillance footage and social platforms [5], [6]. To date, many face de-identification methods have been proposed, which aim at protecting facial sensitive information (especially identity) while preserving utility for identity-unrelated applications. In addition, several related surveys have been conducted on related topics over the past decade, which are outlined in Table I.

1) Padilla et al. [7] provided a review and classification of visual privacy protection methods, as well as an analysis of how visual privacy protection techniques are deployed in existing privacy-aware intelligent monitoring systems.
2) Ribaric et al. [8] reviewed existing face de-identification in still images and videos at the time, namely non-deep-learning image filtering methods and $k$-Same based methods.
3) Ribaric et al. [9] presented a review of de-identification methods for non-biometric identifiers, physiological, behavioral and soft-biometric identifiers in multimedia contents.
4) Liu et al. [10] targeted privacy issues in dynamic Online Social Networks from a user-centric perspective. They proposed a privacy analysis framework consisting of

Jingyi Cao, Xiangyi Chen, Rong Xie, Li Song and Wenjun Zhang are with The Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (email: cjycaojingyi@sjtu.edu.cn;chenxiangyi@sjtu.edu.cn; xierong@sjtu.edu.cn; songli@sjtu.edu.cn; zhangwenjun@sjtu.edu.cn).

Bo Liu is with Australian Artificial Intelligence Institute and School of Computer Science, University of Technology Sydney, Ultimo NSW 2007, Australia (email:bo.liu@uts.edu.au).

Ming Ding is with Data61, CSIRO, Sydney NSW, Australia (ming.ding@data61.csiro.au).

Zhu Li is with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, Kansas City, MO 64110 USA (e-mail: lizhu@umkc.edu).

three stages with different principles, observable privacy, contextual privacy and inferential privacy.

5) Cai et al. [11] systematically summarized the application of GAN in privacy and security, which discussed privacy from the perspectives of both data type and model and security from model robustness, malware detection, etc. They also provided unsolved challenges in the application scenario, model design and data utilization.
6) The relation between privacy and machine learning has been discussed in [12], which covers three categories: private machine learning, machine learning-aided privacy protection and machine learning-based privacy attack and corresponding methods.
7) The proposed taxonomy of [13] was tied to biometric recognition systems and partitioned biometric privacy-enhancing techniques into image-level, representation-level and inference-level, which also reviewed existing datasets, relevant standards and regulations.
8) Shopon et al. [14] summarized fragmented research on biometric de-identification and provided a classification mechanism based on the modalities employed and the types of biometric traits preserved after de-identification. They also discussed their applications in various domains such as cybersecurity.
9) The survey [15] started with a face recognition system and classified anti-facial recognition tools according to their targets, from data collection and model training to inference. They created a systematic framework to analyze the benefits and trade-offs of different AFR approaches and then considered the technical and societal challenges.
10) Zhang et al. [16] reviewed privacy attack methods and corresponding defense mechanisms in both visual data and visual systems under the context of deep learning.
11) Park et al. [17] focused on reviewing GAN-based facial de-identification algorithms. Particularly, their study evaluated existing methods in terms of data utility pertaining to the preservation of dermatologically interesting features such as skin color, pigmentation and texture.
12) Maity et al. [18] conducted a thorough exploration into the intersection of video analytics and privacy preservation, focusing on two core techniques: face de-identification and background blurring. They rigorously analyze the latest advancements, highlight both their efficacy and inherent limitations, so as to underscore the practical significance of these techniques through real-world applications in surveillance and the dynamic landscape of social platforms.

Despite the development of numerous techniques, the field still lacks a comprehensive survey that categorizes these methods and systematically evaluates their performance under different criteria. This paper aims to fill that gap by presenting a state-of-the-art review of face de-identification methods. We categorize existing approaches based on their operational levels (pixel, representation, and semantic) and assess their strengths and limitations in terms of privacy protection and image utility preservation.

Our key contributions are summarized as follows.

- We present a state-of-the-art survey on face de-identification, categorizing existing works by different image processing levels, i.e., pixel-level, representation-level, and semantic-level, and further discuss the characteristics of each category.
- We summarize relevant metrics from the perspectives of privacy protection and image utility, and propose a comprehensive evaluation framework for de-identification algorithms.
- We perform qualitative and quantitative experimental comparisons of several algorithms to discuss their advantages and disadvantages.
- We summarize the analysis and comparative research of face de-identification methods, and further point out open problems and possible future research directions in this field.

The remainder of this paper is organized as follows. Section II reviews preliminary knowledge on face recognition and de-identification technologies. Subsequently, we review the existing face de-identification methods in Section III and classify them into *pixel-level*, *representation-level*, and *semantic-level*, which can help readers gain a high-level understanding of current work and basic ideas. In Section IV, we show practical applications of face de-identification, including usage in specific domains and identity-agnostic computer vision tasks. Afterwards, the evaluation criteria commonly used in face de-identification algorithms are summarized in Section V. Additionally, quantitative and qualitative studies are implemented to compare representative methods in Section VI and the results provide intuitions and insights of of existing algorithms to readers. Finally, we propose some future directions in our view for this task in Section VII and conclude our work in Section VIII.

## II. Preliminary

This section introduces the preliminary knowledge in the research of face de-identification, reviews the research content of face recognition and de-identification, and introduces the principles of deep learning technology and privacy protection theory.

### A. Face Recognition and De-identification

*1) Face Recognition:* Face recognition is a biometric technology that automatically identifies individuals based on their facial features, including statistical and geometric characteristics. It is one of the most important applications of image analysis and understanding. Face recognition encompasses two main types of tasks: binary classification (face verification) and multi-classification (face retrieval). The term ”face recognition” typically refers to identity recognition and verification based on optical facial images. The face recognition process can be summarized as a series of steps: first, a computer analyzes a facial image or video; next, it detects and tracks faces within the input; then, it extracts effective facial features; and finally, it determines the identity of the face by comparing these features against known identities.

TABLE I
SUMMARY ON PRIOR SURVEY ARTICLES

| Ref | Focuses |
|---|---|
| [7] | Visual Privacy Protection Methods and Privacy-Aware Monitoring Systems |
| [8] | Traditional Face De-identification Methods |
| [9] | Privacy Protection in Multimedia Contents |
| [10] | Image Privacy in Online Social Networks |
| [11] | Generative Adversarial Networks in Privacy and Secure Applications |
| [12] | Interactions between Privacy and Machine Learning |
| [13] | Face Biometric Privacy-Enhancing Techniques |
| [14] | Biometric Systems De-Identification |
| [15] | Benefits and Trade-offs of Different Anti-Facial Recognition Technology |
| [16] | Visual Privacy Attack and Defense Methods |
| [17] | GAN-based De-Identification Methods in Dermatology Use Cases |
| [18] | Privacy preservation in video analytics |
| This Survey | Face De-identification Methods |

Face recognition research, dating back to the late 1960s, initially focused on designing feature extractors coupled with machine learning algorithms for classification. Traditional methods rely on hand-made features like edge texture description, and combine with machine learning techniques such as principal component analysis, linear discriminant analysis, and support vector machines. Early geometric-based approaches extracted contours and relationships of facial components, using these as feature vectors. While these methods offered fast recognition and low memory requirements, they often overlooked subtle local features, resulting in information loss. The latest feature point detection technology still struggles to meet accuracy requirements, highlighting the ongoing challenges in the field.

Deep learning techniques have revolutionized face recognition, significantly improving accuracy and robustness by learning representations from large datasets that account for various conditions like lighting, posture, and expressions. For example, DeepFace [19] introduced 3D model-based alignment, establishing a foundation for deep learning in this field. DeepID [20]–[22] series evolved from multi-classification to combining recognition and verification features, later incorporating advanced network architectures. FaceNet [23] employed triplet loss to map faces to Euclidean space, where distances represent image similarity. Then, SphereFace [24] and CosFace [25] introduced angular and cosine margins respectively, mapping features to hyperspherical space. Currently, ArcFace [26] is considered the state-of-the-art, using additive angular margins for tighter feature distributions and clearer decision boundaries. Although current face-recognition models achieve impressive accuracy on specific datasets, they still face significant challenges. Primary among these are the influences of illumination and posture on recognition accuracy. Furthermore, cross-racial and cross-age recognition remains important areas for ongoing research and improvement.

*2) Face De-identification:* The concept of de-identification, a key aspect of privacy protection, lacks a consistent definition in existing literature. However, Ribaric et al. [9] provided a widely referenced definition, describing de-identification in multimedia content as *"the process of concealing or removing personal identifiers, or replacing them with surrogate personal identifiers in multimedia content"*.

The main purpose of face de-identification is to conceal the identity information of a face image. For the given image $X$, the de-identification algorithm $\mathcal{F}$ aims to deceive the face recognition model $\mathcal{R}$ or human by reducing recognition accuracy, which can be expressed as,

$$\mathcal{R}(X) \neq \mathcal{R}(\mathcal{F}(X)), \tag{1}$$

where $\mathcal{R}(X)$ indicates the identity of $X$.

In the past few years, face de-identification methods have evolvied from initial traditional approaches to more sophisticated deep learning-based techniques. Early methods focused on simple perturbation operations applied to facial regions. However, deep learning has enabled more advanced approaches, significantly enhancing the quality of de-identified results. The major challenge of face de-identification task lies in balancing privacy with utility. This trade-off is inherent to the task: as privacy protection effectiveness increases, the utility of de-identified results typically decreases.

The state-of-the-art deep learning-based methods can generate more realistic face de-identification results while ensuring the effectiveness of privacy protection, and they are the focus of the discussion in Section III.

### B. Technical Background

In this section, we introduce the basic knowledge of deep learning technologies mainly used in face de-identification, especially Generative Adversarial Networks and Adversarial Perturbation.

*1) Generative Adversarial Networks (GANs):* The generative adversarial Networks are based on adversarial training proposed by Goodfellow et al [27], which is used to generate images from random noise. Tero Karras et al. [28] was the first to apply GAN to face images and generated results with celebrity characteristics and satisfactory visual perception.

**Basic Theory.** Generative Adversarial Networks are inspired by the game theory, which include two main parts: the generator $G$ and the discriminator $D$. The basic idea is to train through the confrontation between them to achieve the Nash equilibrium. The generator is trained to generate as realistic images as possible to deceive the discriminator, while the discriminator seeks to accurately distinguish the generated fake image from the real image. The corresponding optimization value function can be expressed as,

$$\begin{aligned}\min_G \max_D V(G,D) = &\mathbb{E}_{x\sim P_{\text{data}}}\left[\log D(x)\right] \\ &+ \mathbb{E}_{z\sim P_z}\left[\log\left(1 - D(G(z))\right)\right].\end{aligned} \tag{2}$$

where $P_{\text{data}}$ represents the distribution of real data sets, $P_G$ is the distribution of network-generated images, and $z$ is the random noise in latent space. Under the optimal discriminator, the process of solving the optimal generator is essentially looking for the minimum distance between the distribution of real data and generated data.

**Conditional GANs.** The classical generative adversarial network generates results by randomly sampling from a priori distribution, which lacks controllability in the generation process. Conditional GANs (cGANs) introduce the class information $y$ into the training process, and the objective function can be formulated as,

$$\begin{aligned}\min_G \max_D V(G,D) = &\mathbb{E}_{x\sim P_{\text{data}}}\left[\log D(x \mid y)\right] \\ &+ \mathbb{E}_{x\sim P_G}\left[\log\left(1 - D(G(z \mid y))\right)\right].\end{aligned} \tag{3}$$

Conditional GANs [29] are widely used in image conversion tasks such as face attribute editing [30]–[32]. Some face de-identification algorithms (Section III-D1) treat facial attributes as privacy-sensitive information and achieve image protection through attribute editing. What's more, many face de-identification frameworks are designed based on conditional GANs or the idea of introducing conditional information, like [33]–[35].

**Progressive GANs.** Progressive GAN was proposed in ProGAN [28], which starts with a small input in low resolution, then adds new layer blocks, increases the output size of the generator and the input of the discriminator until the desired image size is reached. Progressive GAN is a training methodology of increasingly learning from low to high resolution, which can allow stable training with high-quality generated results. This strategy is also popular in network design and model training in GAN-based face de-identification methods.

**StyleGAN.** Inspired by style transfer, StyleGAN [36] designed a new generator network structure, which can decouple the high-level semantic attributes of the image through unsupervised automatic learning. StyleGAN2 [37] analyzed the effectiveness of different resolution layers and designed a better network where the latent space can be more disentangled and the quality of the generated image can be higher. StyleGAN3 [38] fundamentally solved the problem of adhesion between coordinates and features and realized the invariance of image translation and rotation. Due to the powerful image generation and editing capabilities, many face swapping or face de-identification algorithms select StyleGAN as the generator [36], [39], [40].

*2) Adversarial perturbation:* Adversarial attack refers to adding adversarial noise to original samples to mislead the deep neural networks to output wrong prediction results, while humans cannot perceive the obfuscation. From different dimensions, adversarial examples can be categorized into false positive attacks and false negative attacks, white-box attacks and black-box attacks, target attacks and non-targeted attacks, and one-time attacks and iterative attacks [41]. Szegedy et al. [42] first proposed L-BFGS method to generate adversarial examples against DNNs. Another classic algorithm is FGSM (Fast Gradient Sign Method) [43] proposed by GoodFellow et al., which is a gradient-based attack method based on constant model network parameters. The process is calculating the derivative of the model and adding noise in the gradient direction. Since the amplitude of perturbation is hard to train, I-FGSM (Iterative Fast Gradient Sign Method) [44] has been proposed to add noise by iterative methods. There are other adversarial perturbation methods including RAND-FGSM [45], JSMA [46], DeepFool [47], etc.

The targets of the existing privacy protection approach based on adversarial perturbation technology can be divided into two aspects: face detection and face recognition, where the former aims to prevent the face in the image from being caught and the latter is mainly to prevent the identification.

*3) Diffusion Models:* Diffusion models have recently become one of the hottest topics in generative models. There are three sub-categories including denoising diffusion probabilistic models (DDPMs) [48], noise-conditioned score networks (NCSNs) [49], and stochastic differential equations (SDEs) [50]. Diffusion models operate in two distinct stages: (1) in the forward stage, input data undergoes a gradual transformation through the addition of Gaussian noise over multiple steps, and (2) in the reverse/backward stage, the objective is to reconstruct the original input data from the noise. The noise at each time step is estimated by a neural network, typically based on a U-Net architecture, which is a critical component of the diffusion model's learning process.

On the basis of the fundamental unconditional diffusion model, many approaches have introduced guidance into the generation process to achieve better results. For example, Guided Diffusion [51] introduces a classification network into the reverse process and calculates gradients based on the cross-entropy loss to guide the generation sampling at each step. In Semantic Guidance Diffusion [52], category guidance is further extended to include guidance based on reference images and text, allowing for more versatile editing effects. Collaborative Diffusion [53] provides a simple and effective method for the collaboration between different diffusion models, allowing each model to leverage its expertise and achieve high-quality multi-modal control in human face generation and editing. When performing face operations based on diffusion models, it's common to use face representations or identity embeddings as control conditions to guide the generation process.

### C. Privacy Theory

*1) $k$-Anonymity:* $k$-anonymity theory [54] contains two-step operations of suppression and generalization of datasets. The theory is based on the assumption of *quasi-identifiers*, that is, the data holder can identify the attributes that may appear in the external data, requiring that each record indistinguishable from $k-1$ other records. *Quasi-identifiers* refer to the attribute set that can be connected with external tables to identify individuals. For example, face data can be associated with identity information through face attributes, and *quasi identifier* can be defined as face attributes and semantic attributes can be defined as facial features or identity features.

$k$-anonymity theory ensures that the attacker cannot determine whether a record is in the data set, nor can it judge whether a specific object has sensitive attributes. If a data set satisfies $k$-anonymity, with only quasi-identifiers of one individual known, the true record can only be chosen with a probability of $1/k$, so that the attacker cannot establish an accurate corresponding relationship with a specific user. This theory provides a certain privacy guarantee but also has some limitations. When background knowledge is available to attackers, homogeneity attack [55] and background knowledge attack [56] become more effective.

*2) Differential Privacy:* Differential Privacy originally targeted at differential attacks to protect users' privacy. It aims to provide a way to maximize the accuracy when querying from statistical databases and minimize the possibility of identifying their records [57].

*Definition.* A randomized method $\mathcal{M}$ meets $(\varepsilon, \delta)$-differential privacy for any pair of neighboring datasets $D$ and $D^{'}$ and for every possible set of outcome $S$, when $\mathcal{M}$ satisfies,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \cdot \Pr\left[\mathcal{M}\left(D'\right) \in S\right] + \delta, \quad (4)$$

where $\epsilon$ refers to the privacy budget. A differential privacy mechanism is usually realized by adding random noise. According to different noise distributions, it can be further divided into Laplace mechanism, Exponential mechanism, Gaussian mechanism and so on.

Differential privacy has widely applied in real-world organizations and acted as a means of legal definitions of privacy such as FERPA [58] and GDPR [59]. Differential privacy has also been applied to machine learning, named differential private machine learning, which aims to train models with formal guarantees implemented by randomizing the training process like adding noise to the gradient. The advanced interaction between differential privacy and machine learning has been discussed in [60]–[62].

## III. Existing Methods

### A. Algorithmic Taxonomy

We propose an algorithmic taxonomy for facial de-identification methods in this survey. They are grouped according to at what level (during an image generation or classification process) and under which data representation form anonymization is addressed. We classify existing methods into three main categories: pixel-level, representation-level and semantic-level. This paper also discussed de-identification in videos and how de-identification algorithms combined with privacy metrics. The full taxonomy is shown in Fig. 1. Especially, generative models, such as Generative Adversarial Networks (GANs) and Diffusion models, are widely used in de-identification methods of deep learning at various levels for better performance, as shown in Fig. 2.

**Pixel-level.** Pixel-level methods — as its name suggests — perform facial anonymization at the atomic level, i.e. raw pixels of face images, which could be further divided into three sub-categories.

First, traditional face de-identification techniques focused on still face images, for instance, black box masking, blurring or pixelation of the facial area after detection and localization.

Second, $k$-Same family introduce $k$-anonymity to facial privacy enhancement. The original face is replaced (hence anonymized) with a synthesized surrogate face, computed through pixel-level averaging. Additional mechanisms [63]–[66] to ensure facial attribute similarity or other data and to further enforce identity change are added into developed variants of $k$-Same based methods [67].

Inspired by recent advancements in generative models, especially GANs, the third line of pixel-level de-identification methods revisits the face inpainting or face synthesis task, this time using neural networks. The general idea is to preprocess the input image (usually mask or crop out the identity-revealing facial area), condition the generator on the pre-processed image which generally contains only non-sensitive biometric information (such as background, hairstyles etc.), and leverage a generative network (in most cases conditional GANs) to reconstruct the facial area at pixel-level.

Traditional methods and pixel-level $k$-Same based algorithms are not the focus of this survey. For more detailed information on these methods, readers could refer to other surveys [8], [9], [68] in the literature.

**Representation-level.** Representation-level methods perform de-identification directly on or with the help of down-sampled latent representations of a face image in a well-chosen latent space.

Most representation-level strategies rely on generative models, such as Generative Adversarial Networks (GANs), Variational Auto-Encoders (VAEs), and Diffusion models, that are capable of generating synthetic face images with a certain degree of pre-defined characteristics and allow inference over latent features. For Generative models-based techniques, either latent vectors are manipulated directly, or indirectly with auxiliary discriminative modules and loss functions that guide the generation process towards de-identification [33], [69]–[72].

Another common group of approaches focus on how to fool the face recognition model, which first extracts face embeddings with a well-designed recognition network, and then computes adversarial perturbation to the original face in order to confound face recognition systems.

We note that the aforementioned $k$-Same based methods could be applied not only on raw image pixels but also combined with deep neural networks for better image quality where high-level image representations in latent spaces are inferred, such as [73], [74].

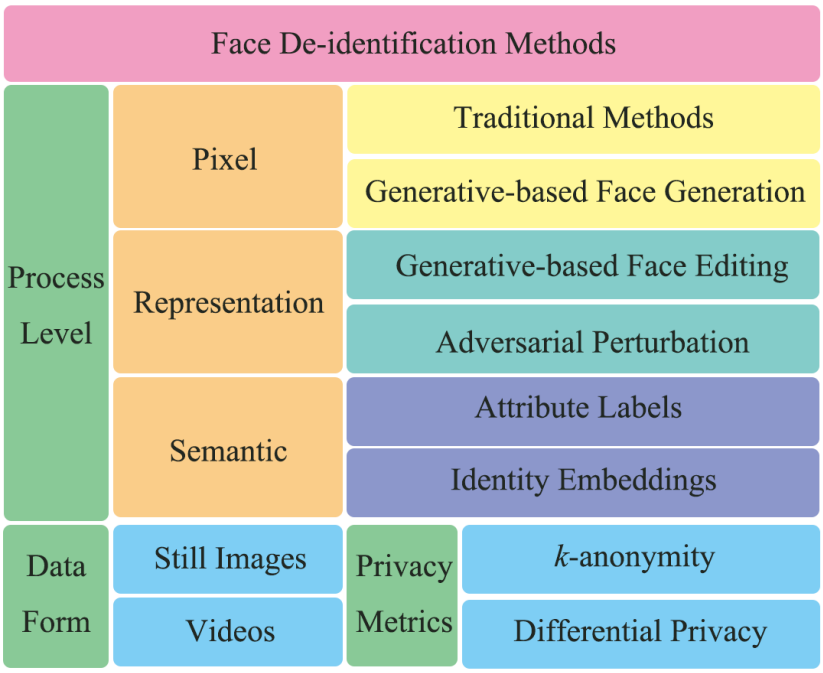


Fig. 1. Taxonomy of face de-identification methods. We group existing methods according to process level and then discuss de-identification in videos. In addition, several de-identification methods are combined with privacy theory, like $k$-anonymity and differential privacy, to provide privacy guarantees.

Alternatively, high-level features of face images could also be modified by adding a noise that follows the differential privacy scheme [75], [76]. These strategies can help to provide adjustable control on image protection and privacy guarantees.

**Semantic-level.** Semantic-level methods extract semantic features of a given face image, and achieve privacy enhancement or de-identification by performing obfuscation on certain extracted semantic features. They could be further divided into the following two sub-categories, attribute-level and identity-level.

Attribute-level methods: these methods first leverage some pre-trained networks to obtain a set of semantic attributes (gender, age, glasses, etc.) of a face image. Subsequently, a subset of the extracted features (whose obfuscation leads to the greatest amount of privacy enhancement, or to say the most privacy-revealing subset) is selected according to traditional privacy protection theories and edited towards obfuscation. These methods [77], [78] by design parameterize the de-identification process with respect to various appearance-related characteristics.

Identity-level methods: while attribute-level methods do not explore the relationship between facial attributes and identity, identity-level methods [40], [79], [80] seek to disentangle the identity embedding from an extracted latent representation and modify or obfuscate only the former. Different from attribute-level methods, identity-level methods first compute one feature map with entangled information on both facial attributes and identities, explicitly disentangle it, and alter the identity embedding but keep the remaining attribute embedding unchanged, which contributes to improving visual similarity and image utility.

Using our level-based classification framework, we now present a comprehensive comparison of the surveyed face de-identification methods in Table II. In this list, we categorize existing proposals by the following characteristics:

- **Data Type**: Data type to which the algorithm can be applied, where $I$ stands for still images and $V$ for videos.
- **Controllability**: Whether the degree of image modification, that is, the identity difference between the de-identified result and the original, can be controlled, so as to meet wider application requirements. Among them, although the value of $k$ in $k$-Same family algorithms can be adjusted, it does not correspond to the degree of identity protection directly, so here they are not considered controllable.
- **Diversity**: Whether different results can be generated in the de-identification process.
- **Reversibility**: Whether the image with original identity can be recovered if it has been de-identified.
- **Authenticity**: Whether the identity of the de-identification result corresponds to the real world.
- **Similarity**: Whether to consider preserving the similarity with the original image.
- **Additional Utility**: Whether there are extras designed to preserve specific usability such as attributes immutability, expression consistency, etc.
- **Privacy Guarantees**: Whether it is combined with traditional privacy theory to provide metric guarantees.

Especially, due to face de-identification results obtained by *Adversarial Perturbation* keep the same content with original images, where the designed noise cannot be perceived by humans. We use other features to summarize and compare them in Section III-C2, so that they aren't listed here for comparison.

TABLE II
COMPARISON OF THE SURVEYED FACE DE-IDENTIFICATION METHODS. THIS TABLE SUMMARIZES THE TECHNIQUES IN TERMS OF PROCESS LEVEL (LEVEL), DATA TYPE (DT), CONTROLLABILITY (CON.), DIVERSITY (DIV.), REVERSIBILITY (REV.), AUTHENTICITY (AUT.), SIMILARITY (SIM.), ADDITIONAL UTILITY (ADD-UTILITY) AND PRIVACY GUARANTEES (GUARANTEES). PAPERS IN EACH CATEGORY ARE SORTED BY PUBLICATION YEAR.

| Level | Methods | Year | DT | Con. | Div. | Rev. | Aut. | Sim. | Add-Utility | Guarantees |
|---|---|---|---|---|---|---|---|---|---|---|
| Pix. | Blurring | - | I | ✓ | × | × | Real | × | × | × |
| | Pixelation | - | I | ✓ | × | × | Real | × | × | × |
| | Masking | - | I | × | × | × | - | × | × | × |
| | $k$-Same [67] | 2005 | I | × | ✓ | × | Fake | ✓ | × | $k$-anonymity |
| | $k$-Same-Select [65] | 2005 | I | × | ✓ | × | Fake | ✓ | Expression, Sex | $k$-anonymity |
| | $k$-Same-M [63] | 2006 | I | × | ✓ | × | Fake | ✓ | × | $k$-anonymity |
| | $k$-Same-further [64] | 2014 | I | × | ✓ | × | Fake | ✓ | × | $k$-anonymity |
| | Meng et al. [66] | 2014 | I | ✓ | ✓ | × | Fake | ✓ | Expression | $k$-anonymity |
| | Samarzija et al. [81] | 2014 | I | ✓ | ✓ | × | Fake | × | Poses | $k$-anonymity |
| | $k$-Diff-furthest [82] | 2015 | I | ✓ | ✓ | × | Fake | × | × | $k$-anonymity |
| | Sun et al. [83] | 2018 | I | × | × | × | Fake | × | × | × |
| | Sun et al. [84] | 2018 | I | ✓ | ✓ | × | Fake | × | × | × |
| | DP-Pix [85] | 2018 | I | ✓ | ✓ | × | Real | × | × | DP |
| | Fan et al. [86] | 2019 | I | ✓ | ✓ | × | Real | × | × | DP |
| | DeepPrivacy [34] | 2019 | I/V | × | ✓ | × | Fake | × | × | × |
| | CIAGAN [35] | 2020 | I/V | × | ✓ | × | Real | × | × | × |
| | JaGAN [87] | 2021 | V | × | ✓ | × | Fake | × | × | × |
| | Kim et al. [88] | 2023 | V | ✓ | ✓ | × | Real | × | × | × |
| | RID-Twin [89] | 2024 | V | × | ✓ | × | Fake | × | Expression | × |
| Rep. | Ren et al. [90] | 2018 | V | × | × | × | Fake | × | Action | × |
| | $k$-Same-Net [73] | 2018 | I | × | ✓ | × | Fake | ✓ | × | $k$-anonymity |
| | $k$SS-GAN [74] | 2019 | I | × | ✓ | × | Fake | ✓ | × | $k$-anonymity |
| | Song et al. [91] | 2019 | I | × | × | × | Fake | ✓ | × | × |
| | PP-GAN [72] | 2019 | I | × | × | × | Fake | ✓ | × | × |
| | PPAPNet [92] | 2019 | I | × | × | × | Fake | ✓ | × | DP |
| | Gu et al. [33] | 2020 | I | × | ✓ | ✓ | Real | × | × | × |
| | DeIdGAN [69] | 2021 | I | × | ✓ | × | Fake | × | × | $k$-anonymity |
| | DeepBlur [93] | 2021 | I | ✓ | ✓ | × | Fake | × | × | × |
| | DP-Image [75] | 2021 | I | ✓ | ✓ | × | Fake | ✓ | × | DP |
| | PI-Net [76] | 2021 | I | ✓ | ✓ | × | Fake | × | Attributes | DP |
| | SF-GAN [71] | 2021 | I | × | × | × | Fake | ✓ | Attributes | × |
| | QM-VAE [70] | 2022 | I | × | × | × | Fake | ✓ | Expression | × |
| | Yang et al. [94] | 2022 | I | × | ✓ | ✓ | Real | × | × | × |
| | Zhao et al. [95] | 2022 | I | ✓ | ✓ | × | Fake | ✓ | × | × |
| | Diff-Privacy [96] | 2023 | I | ✓ | ✓ | ✓ | Fake | ✓ | × | × |
| | Savic et al. [97] | 2023 | V | × | × | × | Fake | ✓ | × | × |
| | Hanawa et al. [98] | 2023 | I | × | ✓ | × | Fake | ✓ | × | × |
| | Tian et al. [99] | 2023 | I | × | × | × | Fake | × | Race, Age, Gender | × |
| Sem. | Chi et al. [100] | 2015 | I | × | ✓ | × | Fake | ✓ | × | $k$-anonymity |
| | Meden et al. [101] | 2017 | I/V | × | ✓ | × | Fake | × | Expression | $k$-anonymity |
| | AnonymousNet [77] | 2019 | I | × | ✓ | × | Fake | × | × | $t$-closeness |
| | CleanIR [102] | 2020 | I | ✓ | ✓ | × | Fake | ✓ | × | × |
| | Wen et al. [103] | 2020 | I | ✓ | ✓ | × | Fake | ✓ | × | × |
| | CFA-Net [104] | 2021 | I | ✓ | ✓ | × | Fake | ✓ | × | × |
| | UU-Net [105] | 2021 | V | × | ✓ | ✓ | Fake | × | × | × |
| | Cao et al. [80] | 2021 | I | ✓ | ✓ | ✓ | Fake | ✓ | × | × |
| | Chuanlu et al. [106] | 2021 | I | × | ✓ | × | Fake | × | Expression | $k$-anonymity |
| | Yang et al. [107] | 2021 | I | × | × | × | Fake | ✓ | × | × |
| | FICGAN [40] | 2021 | I | × | ✓ | × | Fake/Real | ✓ | Attributes | $k$-anonymity |
| | Croft el al. [108] | 2021 | I | ✓ | ✓ | × | Fake | ✓ | × | DP |
| | Cao et al. [78] | 2022 | I | × | ✓ | × | Fake | ✓ | × | $k$-anonymity, DP |
| | IdentityDP [79] | 2022 | I | ✓ | ✓ | × | Fake | ✓ | × | DP |
| | IdentityMask [109] | 2022 | V | ✓ | ✓ | ✓ | Fake | ✓ | × | × |
| | Cao et al. [110] | 2023 | I | × | × | × | Fake | ✓ | × | × |
| | IDeudemon [111] | 2023 | I | ✓ | ✓ | × | Fake | ✓ | × | × |
| | Zhu et al. [112] | 2024 | I/V | ✓ | ✓ | × | Fake | ✓ | × | × |

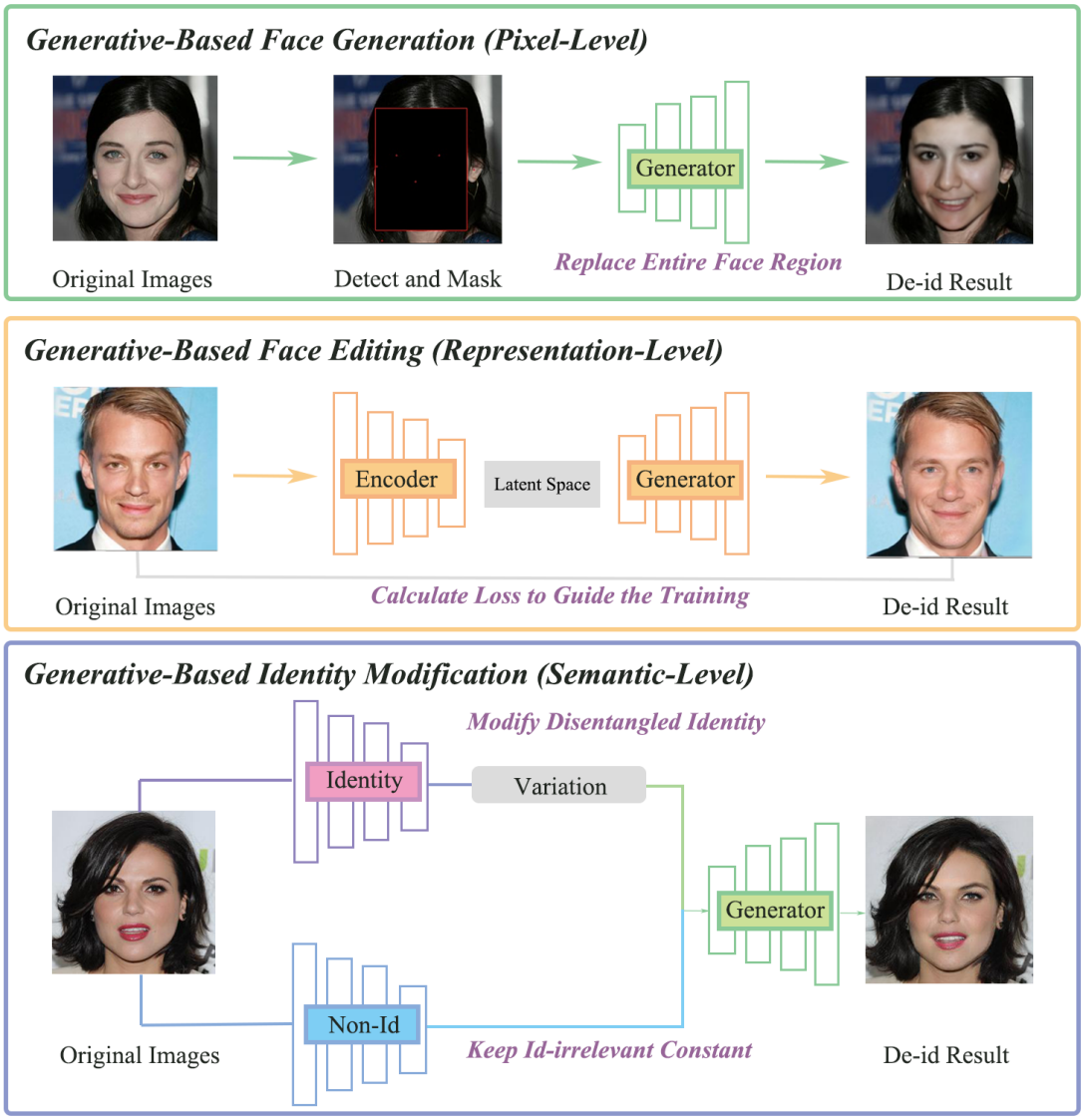


Fig. 2. Different applications of Generative Models in de-identification algorithms, where the results are in turn generated by DeepPrivacy [34], SF-GAN [71] and IdentityDP [79].

### B. Pixel-Level

*1) Traditional Methods:* **Methods:** In traditional computer vision, privacy protection mainly obfuscates face privacy-sensitive information directly. For example, blurring refers to smoothing the face area and each pixel is replaced by a weighted average of its neighborhood, pixelation reduces the resolution of the face area by dividing the image into a certain range of cells and setting all pixels in each cell as common, masking means to use a black rectangle to occlude the entire face area or key parts, and another way is adding Gaussian or other noise to the pixels.

On this basis, some methods introduce deep generation models into face blurring. DartBlur [113] uses the DNN architecture to generate blurred faces, which aims to reduce the artifacts introduced in de-identified results that are harmful to the utility of downstream tasks. The results prove a good balance between accessibility, review convenience and detection of artifact suppression.

**Discussion:** Though traditional methods have been widely used because of simplicity and ease of operation, it is worth noting that they achieve de-identification by unselectively obfuscating all biometric features of the human face in question, making it hardly reusable in standard computer vision pipelines. In addition, they have limited effectiveness in protecting privacy and are vulnerable to attacks. The study [114] shows that face recognition models based on deep learning technology, especially convolutional neural networks, can still identify people treated with these traditional techniques with high accuracy. What's worse, obfuscation-based protection methods can greatly impair image quality, where visual effects are also greatly reduced since these perturbations can be significantly perceived by humans.

*2) Generative Model-based Face Generation:* Due to the powerful face generation capability of GAN, it has been introduced to face de-identification. The core idea of achieving privacy protection through face synthesis or inpainting is that

if partial or complete face information is used in training the generator, it may bring the privacy leakage risk, so this type proposed to de-identify by reconstructing the face region at pixel-level completely.

**Methods:** Sun et al. [83] proposed an image obfuscation method by head inpainting which can be summarized as the following two stages: (1) if the input image contains the original face area, detect face landmarks with Dlib toolbox while if the original face area is blocked, generate face landmarks with the trained Landmark Generator; (2) generate the head inpainting based on the landmarks and the blackhead or blurhead image with Head Generator. Another identity obfuscation method [84] combined art parametric face synthesis with a deep generative model, where it can manipulate identity explicitly by the parametric part and the generator can help improve the realism of results. It firstly renders the reconstructed face based on 3DMM and replaces the face region. The generator then takes the rendered face and the image of the occluded face area as input, and generates the photo-realism image by adding fine details to match the surrounding image context.

The most well-known algorithm is DeepPrivacy [34], which proposed a new architecture that can automatically anonymize the face without changing the original posture and image background. Based on DeepPrivacy, JaGAN [87] was designed for face anonymization in videos, masking out faces with black patches and inpainting the missing parts with generated faces by a video generative model enforcing temporal coherence. Compared with the previous work, the innovation of CIAGAN [35] is to control the face anonymity process and guide the synthesis of de-identification results according to the given reference identity. CIAGAN implemented an identity guidance discriminator to drive the generator towards creating images whose representational features are similar to those of a landmark identity, earlier fed into the generator. Thus the anonymization procedure is naturally under full control while diversity in generated identities is simultaneously ensured by carefully selecting landmark identities. CIAGAN took the landmark as input to avoid the leakage of original facial information and the generator conditioned on the face shape can keep the posture information, which also applied the masked background to ensure there are no background changes that may interfere with the detection or tracking.

To solve the artifacts caused by the difference of face angles, ATFS [115] proposes an efficient framework for improved angle transformation and face replacement. In the framework, TransNet and SWGAN are proposed based on generative adversarial networks, where TransNet combines the predicted facial keypoints to locate the target's facial features when reconstructing, and generates multi-angle transformed images to achieve angle transformation, and SWGAN is designed to extract facial features and replace the face region of the input image with the target to realize face de-identification.

**Discussion:** De-identification by face inpainting or synthesis can achieve the highest privacy level because they do not require access to the original image and can guarantee the deletion of any privacy-sensitive information. However, they also have the disadvantage that they cannot retain the appearance similarity with the original image. A significant shortcoming is that all information contained in the face region will be removed even if the part is not related to identity. It is hoped to achieve identity protection through minor modification, so that the de-identification results can retain higher utility.

*3) $k$-Same:* **Methods:** $k$-Same family algorithms are designed based on $k$-anonymity [54] theory. Newton et al. [67] proposed to divide the face images in the dataset into $k$ classes according to facial features and calculate the average face in each class to limit the recognition success rate to $1/k$. Both $k$-Same-pixels algorithm selecting images for averaging based on Euclidean distance and $k$-Same-eigen algorithm selecting images on Principal Component Analysis coefficient space failed to obtain an average image with good visual effect.

The variants of $k$-Same family aimed at improving the utility and quality of the results. $k$-Same-Select [65] aimed at preserving facial attributes like gender or expression, which divided the dataset into mutually exclusive subsets according to the data utility function, and apply $k$-Same to different subsets independently. $k$-Same-M [63] tried to remove unpleasant artifacts because of misalignment by Active Appearance Model (AAM). Another multi-factor framework for face de-identification that unifies linear, bilinear and quadratic models was proposed in [116] for better data utility. $k$-Same-further [64] pointed out that the previous methods target minimizing the identity change through $k$-Same-closest and proposed to select the farthest image cluster to maximize the identity change, no matter how to select $k$ value, it can be perfect privacy protection. Additionally, $k$-Same-further-FET [66] combined $k$-Same-furthest for privacy protection and model-based Facial Expression Transfer for data utility. $k$-Diff-furthest [82] adopted the same iterative process of $k$-Same-furthest, where two clusters with size $k$ would be generated in each iteration and their centroids would be swapped. For the problem of poor diversity of the $k$-Same algorithm, $k$-Diff-further could generate a unique de-identified result for each face in a cluster.

Considering the faces captured in videos have various poses, a de-identification method accommodated for still images with large-angle was proposed in [81], which provided the basis for the naturalness of video de-identification. Similar to the ideas of $k$-Same-Select [65], face images were grouped in accord with poses, so that the shape could remain in the process of model fitting while the texture would be changed. It also applied $q$-far based on the Euclidean distance to ensure that the appearance distance is far enough.

**Discussion:** Although the $k$-Same family algorithms are important parts in de-identification which realize the combination of image privacy protection and traditional theories. They also have the following limitations: (1) Because $k$-anonymity theory is vulnerable to background knowledge attack, where the attacker can also reduce the protection effectiveness through the prior knowledge of sensitive information. A more detailed theoretical discussion on the susceptibilities of background knowledge was present in [108]. (2) The $k$-Same algorithms assume that the original face image set must be person specific and each object appears only once in dataset, so that the

existence of multiple images from the same object or images with similar biometrics will reduce privacy level in a specific practice. In addition, applying the $k$-Same algorithm at pixel level requires images to be strictly cropped out of the face area and aligned in order to reduce artifacts.

*4) Differential Privacy:* **Methods:** Fan et al. [85] first attempted to provide differential privacy guarantees for multimedia data and proposed $m$-neighborhood to protect sensitive information represented by $m$ pixels at most. In addition, a pixelization-based method with grid cells was designed for better privacy-utility balance, and experiments proved that differentially private pixelization could reduce privacy risk significantly. However, applying pixelation to the entire image resulted in high utility loss and the results would be unrecognizable as human faces. For better image quality, another relaxed privacy model in [86] developed an efficient mechanism in which Singular Value Decomposition (SVD) was empolyed, and generated visually similar images with the same singular matrices but different singular values. Although image quality could be improved compared with the former, the perceptual information captured by singular value decomposition is limited, and the results generated by perturbing perceptual features of $i$ highest singular values are still far from meeting the requirements of preserving image utility and photo-realism. Additionally, the interactive framework for face obfuscation under the rigorous notion of differential privacy, including DP-Pix [85] and DP-SVD [86] was designed in DP-Shield [117].

**Discussion:** Most methods of applying differential privacy at the pixel-level are to combine it with traditional de-identification algorithms (like pixelation) to provide metric privacy guarantees. Due to the rough processing form, they often failed to generate high-quality results.

### C. Representation-Level

*1) Generative Models-based Editing:* Another line of face de-identification work seeks to exploit GANs' ability to perform the image-to-image transformation. Seminal advances on GANs have provided an inspiring framework, and in order to adapt it to face de-identification scenarios, some methods attempt to modify the initial architecture of deep generative networks, in particular by attaching auxiliary modules and additional optimizing loss functions.

**Method:** SF-GAN [71] adds a unique unit to ensure that the data distribution of the synthesized face is different from anyone in the datasets and balances the trade-off between face attribute preservation and privacy protection. PP-GAN proposed by Wu et al. [72] combines the two principal objectives in face de-identification — de-identify and keep structural similarity — by adding to a basic conditional GAN backbone two additional modules, (1) a verificator that enforces the sampling range away from given samples in the identity embedding space, or to de-identify, and (2) a regulator which constrains the variation of image structure, or to retain structural similarity. Both are explicitly expressed as external losses to compensate for their contradiction in traditional GAN loss. Zhao et al. [95] designed an adjustable privacy-related loss in the training process, which indicates the identity distance between results and the origin. A pre-trained styleGAN2 was used as the generator, so that high-quality and high-resolution de-identification results can be obtained. Although different parameters can be set to generate de-identified images with different privacy levels, the parameters lie in the loss function, so the encoder needs to be retrained, where the flexibility of the model would be limited. Hanawa et al. [98] introduce a de-identification loss function and embed facial features of other persons into the face images.

Several methods transformed face images into latent space, and generated de-identified results with regard to the modified representations. DeepBlur [93] searched the latent vectors with a feature extractor, then applied a Gaussian filter on them, and finally reconstructed with the blurred representations. Various kernel sizes of Gaussian filters can be selected to obtain different results with desired privacy levels.

Noting the irreversibility and hence limited data accessibility in previous methods, Gu et al. [33] further proposed a password-based face identity transformer enabling anonymization and de-anonymization. With carefully designed optimizing loss functions, this model could (1) remove face identity information after anonymization, (2) recover the original face when given the correct password, and (3) return a wrong but photo-realistic face with a wrong password. An invertible mask for face privacy was proposed in [94]. The swapped image was generated according to the protected face and replaced face and a well-designed "mask" would be put on to obtain the masked face and calculate lost information, where the "mask" is visually indistinguishable. For authorized users, the original image could be recovered by operating a discrete wavelet transform on the masked face and auxiliary information. FaceRSA [118] is the first fully facial identity cryptography framework similar to RSA based on StyleGAN, which can control the face identity by user-given password for anonymization and deanonymization. The unrelated attributes can be kept by adding modifications only to the identity-related layers of embeddings in StyleGAN latent space. *Single loss* is used for a single pair of encryption and decryption processes, *Sequential loss* is designed for more complex requirements with multiple pairs of passwords, and *Associated loss* is applied to ensure that equivalent passwords exist during both encryption and decryption.

Based on evolutionary GANs [119], Song et al. [91] opted for another approach, which dynamically selects an optimal generator that best caters for the de-identification need. Their method differs from the previous ones in that the distance and structural similarity between the original and de-identified faces are introduced during the evaluation and selection process of generators, and thus no supplementary networks are needed.

QM-VAE (Quality Maintenance-Variational AutoEncoder) framework [70] was a multi-input generative model guided by service quality evaluation. In the first step, four sets of de-identified images without private information were generated by four typical methods (blindfold, cartoon, Laplace, and mosaic). Next, these datasets were fed into QM-VAE to calculate whether the utility was consistent with the original

data. At the same time, the calculation results were added as a part of the loss function for model training. The service quality was quantified as the difference of attributes in service scenarios, and it contributes to generating reliable faces.

Diff-Privacy [96] is the first work to apply the diffusion model to privacy protection, addressing two crucial tasks in facial privacy: anonymization and visual identity information hiding. The former aims to prevent machine recognition, while the latter requires ensuring the accuracy of machine recognition. By training the proposed Multi-Scale Image Inversion (MSI) module, a set of conditional embeddings in the SDM format is obtained from the original images. Based on these conditional embeddings, corresponding embedding scheduling strategies are designed to construct different energy functions during the denoising process.

**Discussion:** The core principle of Generative Models-based methods is to impose certain constraints by adding external guidance losses (guidance towards target identity or away from source identity, guidance to maintain certain attributes, etc.), together with their accompanying networks (namely verificators, regulators, modifiers, etc.). Generative Models-based methods generally achieve promising de-identification rates, but they sometimes struggle in applying to faces in the wild with non-frontal head poses or occlusions, and are not evidently applicable to videos [72]. Meanwhile, although they generate realistic de-identified faces, their output could still drastically alter the original faces in terms of visual perception and would, therefore, fail to live up to the possible expectation of retaining the original face appearance. Another shortcoming of Generative Models-based methods is that the quality of generated faces is highly dependent on the implied structure itself, thus general artifacts could be noticed in their outputs [33]. Taking the basic generative models as a baseline and adding appropriate constraints to improve the network, the purpose of de-identification can be effectively realized. Unfortunately, the degree of privacy protection cannot be adjusted if the model has been trained. Models with different parameters are required to meet different applications, so the flexibility may be poor.

*2) Adversarial Perturbation:* The existing face de-identification methods based on adversarial examples are mainly aimed at preventing the faces from being recognized by the deep neural network, so as to prevent private them from being used to make false images or videos.

**Methods:** The latest research on adversarial image perturbations [120] has shown that it can effectively confuse the face recognition system without introducing significant artifacts, and proposed to use of the game theory framework to achieve the opposite goals for users and recognizers to obtain the privacy guarantee of the user without the countermeasures independent of the recognizers.

A series of methods could generate strong perturbations over the entire facial images. $A^3GN$ (Attentional Adversarial Attack Generative Network) [121] introduced face recognition networks as the third player for target-attack, where a conditional variational autoencoder was applied to get latent code and attentional modules were provided to capture feature representation and dependencies. Dong et al. [122] proposed an evolutionary attack method for query-efficient adversarial attacks for the black-box scenario. The target is to model the local geometries of search directions and reduce the dimension. DFANet (Dropout Face Attacking Network) [123] applied dropout layers to the surrogate model in each iteration. Inspired by the I-FGVM [44] algorithm, P-FGVM [124] has been proposed to generate adversarial de-identified facial images that resemble the original ones in the image spatial domain, which can both protect the privacy and preserve visual facial image quality. AdvFaces [125] could focus on the set of pixel locations required by face recognition model and only add perturbation to those salient regions like eyebrows and nose. The most advanced method Fawkes [126] helps users add an invisible "cloak" to the shared image, which can successfully confuse the face recognition model. $CSP^3Adv$ [127] proposed a collaborative privacy protection method based on adversarial examples for social networks, which avoids direct image acquisition by the server and the protected images are robust to JPEG compression. Users can construct the gradient perturbation on the lightweight face recognition model, and the perturbation transfer module is trained to learn the corresponding perturbation under the complex face recognition model on the server side. Additionally, the high-frequency noises will be filtered to improve the privacy protection effect after social network compression. Ghafourian et al. [128] compare the effectiveness of two popular adversarial methods, Basic Iterative Method (BIM) and Iterative Least Likely Class (ILLC) to de-identify. They reach the preliminary conclusion that BIM works better through experiments and transferability between multiple face recognition models.

Patch-based adversarial attacks mean just adding adversarial noises to certain areas. For example, the person can be allowed to evade being recognized when wearing the generated eyeglass in [129]. AdvHat [130] was designed to create a rectangular image stuck on the hat and decrease similarity to ground-truth identity below threshold. Although these algorithms will not affect the naturalness of the generated results, these adversarial patches are perceptible to humans and mostly applied in the white-box setting.

For increased robustness and image quality, Yang et al. [131] introduced an attention module to different target classification networks to extract chief features and processed them by feature fusion, which helps to improve the transferability. Almost invisible adversarial examples were generated by perturbing the features of faces calculated by feature fusion matrix.

Some studies combined de-identification with other face editing tasks, where the adversarial noise can be hidden within the editing area to obtain better visual effects. A common example is makeup transformation. Zhu et al. [132] made the first attempt by applying two sub-networks, the makeup transfer sub-network and the adversarial attack sub-network for the white-box face recognition attack. Adv-Makeup [133] proposed to add makeup to eye regions to fool face recognition models. They applied a blending method to improve the naturalness of generated results and a fine-grained meta-learning strategy to improve the transferability of the black-box attack. AMT-GAN (Adversarial Makeup Transfer GAN) [134] proposed to generate an antagonistic face image with

TABLE III
COMPARISON OF DE-IDENTIFICATION BASED ON ADVERSARIAL PERTURBATION. PAPERS ARE SORTED BY PUBLICATION YEAR.

| Methods | Year | Model Knowledge | Attacks Type | Perturbation Region |
|---|---|---|---|---|
| Sharif et al. [129] | 2016 | White / Black-Box | Target / Non-target | Eyeglasses |
| Dong et al. [122] | 2019 | Black-Box | Target | Entire Face |
| Zhu et al. [132] | 2019 | White / Black-Box | Target / Non-target | Eye (Makeup) |
| P-FGSM [124] | 2019 | White-Box | Target | Entire Face |
| AdvFaces [125] | 2020 | Black-Box | Target / Non-target | Eyebrows, Nose |
| Fawkes [126] | 2020 | Black-Box | Non-target | Entire Face |
| DFANet [123] | 2020 | Black-Box | Target | Entire Face |
| AdvHat [130] | 2021 | White-Box | Non-target | Hat |
| $A^3$GN [121] | 2021 | White / Black-Box | Target | Entire Face |
| Adv-Makeup [133] | 2021 | Black-Box | Target | Eye Shadow |
| AMT-GAN [134] | 2022 | Black-Box | Target | Entire Face (Makeup) |
| Yang et al. [131] | 2022 | White / Black-Box | Non-target | Entire Face |
| $CSP^3$Adv [127] | 2023 | White / Black-box | Target | Entire Face |

makeup transferred from the reference image. A new regularization module and a joint training strategy were introduced to coordinate the conflict between adversarial noise and cycle consistency loss in makeup transfer, so as to achieve an ideal balance between image quality and attack success.

**Discussion:** Unlike other face de-identification algorithms, the adversarial perturbation techniques aim at disturbing the face recognition model without any modification to the image content. They cannot affect visual observation and identity recognition of humans as they retain the appearance exactly the same as the original image. Therefore, the protection performance and application scenarios of this type of face de-identification method are also different. In addition, it is a practical research that combines adversarial examples with other face modification tasks to make the adversarial noises for de-identification inconspicuous and unnoticeable.

There are still various issues with existing methods. Compared with white-box attacks, black-box attacks are more useful in real-world scenarios. Like other adversarial attack research, a major challenge is how to improve the generalization and robustness of adversarial examples. On the other hand, there exist obvious artifacts in the de-identification results and the clarity can still be improved.

*3) $k$-Same:* **Methods:** Chang et al. [135] focused on the requirement of machine learning data utility of face de-identification and proposed to adjust the data class of $k$-anonymity by feature important (FI) and margin preservation (MP), where margin preservation can not only anonymity the data but also can retain the original boundaries.

With the development of deep learning, some algorithms combine $k$-same theory with the deep generation network in order to solve the problem of poor visual perceptions. $k$-Same-Net [73] generated anonymity face by generative networks instead of image or model-parameter averaging. However, $k$-Same-Net can only generate synthetic images with trained identities, which need to be retrained for each new image. $k$-Same-Siamese-GAN [74] proposed to take advantage of the mixed precision training (MPT) technique for better privacy protection and image quality. In addition, $k$-anonymity theory has been widely applied in many face de-identification algorithms. Anonfaces [136] extracted low-dimensional features of face images and performed fixed-size clustering analysis while minimizing information loss. The trade-off between privacy and utility can be adjusted by fine-tuning the weight in identity fusion or selecting an appropriate $k$ value. Compared with $k$-Same-Net, Anonfaces could generate any de-identification result with an unknown identity.

An identity-adversarial discriminator is leveraged to guide the generator away from synthesizing the original identity in DeIdGAN [69] and pre-defined sensitive identities. DeIdGAN remove the original identity information by anonymizing both shape and style, where the de-identified results are generated with $k$-anonymized ASM (Anonymous Semantic Mask) and ASI (Anonymous Style Image).

**Discussion:** The de-identification results lack naturalness, diversity and realism, even though there is a trend of combining with deep learning to improve image quality. Each method based on $k$-Same has a common disadvantage of failing to provide unique de-identified outputs for various images [70].

*4) Differential Privacy:* **Methods:** It is a common idea to expand the traditional differential privacy protection theory when applying it to high-level features. DP-Image [75] firstly introduced a novel notion of image-aware differential privacy, which redefined traditional DP and local differential privacy (LDP) [137]. DP-Image provided provable and adjustable privacy levels to achieve image protection through adding noise to image features combined with advanced GAN techniques. The pSp framework [138] performed as feature extraction to map input image into latent space, the noise generator used to inject noise onto the feature, and StyleGAN2 [37] acted as the image reconstruction module to generate protected results with perturbed features.

Focusing on that applying DP with Laplace noise may lead to unacceptable artifacts, another method named PI-Net (Perceptual Indistinguishability-Net) [76] was proposed to study how differential privacy should be applied for face image obfuscation. Motivated by metric privacy [139], image adjacency was defined as the difference between high-level

features. PI-Net also provided a clear semantic definition of latent space which can be captured for privacy analysis. Based on the definition of perceptual distance and adjacent images, $\epsilon$-PI was proposed to provide a privacy guarantee when injecting noises to obfuscate images.

The PPAPNet [92] (Privacy-Preserving Adversarial Protector Network) proposed to add optimized noise at the latent space level, consisting of protector, critic and attacker. Protector was designed as an encoder-decoder network for obtaining latent codes, and another encoder-decoder network, attacker, tried to prevent the anonymized results from model inversion attacks [140]. Especially, a noise amplifier inside the protector was trained to generate optimal noise for latent representations of each dimension, which contributed to both image realism and privacy attack immunity.

**Discussion:** Lack of formal privacy guarantee inspired many scholars to introduce differential privacy, a provable guarantee of privacy, in face de-identification process. Due to the difference in the data form of face images, how to combine deep generative network and differential privacy theory and how to improve traditional DP to make it more suitable for face de-identification tasks are major problems worthy of attention in the future.

### D. Semantic-Level

*1) Facial Attribute Labels:* Some researchers regard facial attribute labels as privacy-sensitive information and realize image identity protection based on attributes indistinguishability. This type of algorithm is usually combined with the face attribute editing model in the implementation process. On the contrary, some algorithms aim at de-identification with attribute preservation to ensure the utility in attribute classification tasks, which will be discussed in Section IV-B.

**Method:** AnonymousNet [77] proposed a de-identification framework based on attribute editing and designed privacy-preserving attribute selection (PPAS) algorithm based on traditional privacy protection theory to make the attribute distribution of the protected face close to that in real life, which is the first to prove that face privacy is measurable and provide privacy guarantees. AnonymousNet consists of four steps: facial feature extraction, semantic-based attribute obfuscation, de-identified face generation based on StarGAN [31], and adversarial perturbation. Further considering minor changes to protected images while providing metric privacy, a differential privacy $k$-anonymity algorithm [78] was proposed to calculate the obfuscated attributes as the de-identified target attributes. In UU-Net [105], four attributes *ID, gender, ethnicity, hairstyle* labels were taken into consideration, where *ID* is always selected to change for de-identification. UU-Net can have full control over the properties of anonymized faces in poor-resolution video streams.

**Discussion:** Since facial attribute tags can be represented by binary discreteness, most of these algorithms can be combined with traditional privacy protection theories like $k$-anonymity [54], thereby enhancing interpretability. However, most algorithms do not further consider the relationship between face attributes and identity. Most methods focus on the attributes which have a great impact on the appearance rather than the attributes with higher privacy level. For example, we think external features like hair color have weak privacy to be protected but have a great impact on appearance. Identity preserving face privacy protection is a related topic, which tries to anonymize facial attributes but preserve the recognition utility. The soft biometric privacy was discussed in [141], [142] and some researches targeted at multi-attribute privacy, such as PrivacyNet [143] and SensitiveNet [144]. Li et al. [145] proposed a face anonymization method that is composed of an identity-aware region discovery module and an identity-aware face confusion module. Different from removing identity information, these methods could obfuscate visual appearance while retaining identity discriminability. There are also some studies that use attribute inference as the goal of privacy. For example, InfoScrub [146] employed a bi-directional discriminator for attribute inversion in the first step, and performed attribute obfuscation by maximizing entropy on inferences over targeted privacy attributes. They are also interesting research fields, and the impact of identity and attribute labels on face privacy is also worth studying.

*2) Identity Embeddings:* Disentanglement is based on the assumption that each component of the representation learned by neural networks is independent of each other. It aims at transforming the entangled data into another latent space that can be separated. How to disentangle identity and attributes has also received extensive attention in face swapping and face de-identification tasks. In this section, we reviewed face swapping based on disentanglement, Generative Models-based identity modification for de-identification, and combining $k$-anonymity or differential privacy to identity encryption.

Face swapping refers to transferring the identity features from the source image to the target while keeping other attributes, such as posture, expression, and lighting conditions invariant. An important difference is that the de-identified results may not correspond to a real identity but the swapped results should keep consistent with the source in recognition. In order to avoid unnecessary face modification, some Generative Models-Based algorithms have been proposed to only modify the disentangled identity representation. Since the identity can be represented as a vector, it is also conducive to the further integration of privacy protection theory.

The basic framework can be summarized as an encoder-decoder structure, where the encoder can be divided into identity encoder and attribute encoder [39], [102], [103], [147], [148], and the decoder acts as a generator. Due to the development of face recognition technology [23], [25], [26], identity features can be represented as a vector. Relatively, attribute features should include all other identity-irrelevant information, so that the network structure and the representation will be more complex. Most methods applied a pre-trained face recognition model as the identity encoder to obtain identity embedding to guide the training of the attribute encoder and generator.

**Face Swapping.** Better design of network structure or auxiliary module is a common improvement idea. FaceShifter [148] designed the adaptive embedding integration network (AEI-Net), where an adaptive attentional denormalization generator

can achieve adaptive integration of identity and attributes. Additionally, a novel heuristic error acknowledging refinement network (HEAR-Net) is applied in the second stage to address face occlusions. ZFSNet (Zero-shot Face Swapping Network) [149] proposed an additional de-identification module and an attention spatial style modulation to blend features. A classifier designed on a multi-layer perception network is employed in the de-identification module to obtain identity-related features. FaceSwapper [147] is another end-to-end framework proposed for one-shot progressive face swapping. There are two major parts, the disentangled representation module and the semantic-guided fusion module. Some studies have improved the connotation of identity. For example, HifiFace [150] proposed 3D shape-aware identity based on the idea that face shape is the feature corresponding to identity. HifiFace introduced a pre-trained 3D face reconstruction model [151] as a shape feature encoder combined with a state-of-the-art face recognition model to get the identity information including geometric structure.

Some methods directly use the pre-trained StyleGAN [36] as the generator for high-quality image generation. Nitzan et al. [39] proposed to learn a mapping network to map disentangled face representations $z$ into $W$ space. Xu et al. [152] disentangled the latent semantics of StyleGAN by utilizing the progressive nature of the generator and deriving structure and appearance from different layers.

There are also some approaches that take advantage of the diffusion model for the face-swapping task. DiffFace [153] proposed an ID Conditional DDPM, which can generate face images with the desired identity. Some facial expert models are introduced in the sampling process to preserve target attributes. DiffSwap [154] introduces a high-fidelity controllable face-swapping approach, treating it as a conditional restoration task, executed by a powerful diffusion model guided by facial identity and landmarks. Additionally, a midpoint estimation method is proposed to efficiently recover reasonable diffusion results for the swapped faces in just two steps, enabling the introduction of identity constraints to enhance the quality of face swapping.

**Generative Models-Based Identity Modification.** Many de-identification algorithms of this type include two steps, disentanglement and identity conversion. CleanIR [102] adopted the VAE architecture with skip connections to split encoded features into identity-related and attributes-related parts, and the Identity Changer module was designed to calculate a new identity for de-identification. Wen et al. [103] proposed to add adjustable privacy to the individual identity embedding, which achieves the target of generating natural de-identification results and controlling the degree of privacy protection according to application scenarios. ITNet [155] randomly assigns a unique virtual identity mask added to the disentangled identity. In addition, the identity mask is optimized in de-identification using the identities extracted from the user's $n$ images. CFA-Net (Controllable Face Anonymization Network) [104] disentangled face identity from other contents, and projected identity onto spherical space for manipulation. The anonymized results changed in two dimensions, anonymous extent and anonymous identity variety. Due to the identity distance can be expressed as cosine similarity between features calculated by face recognition model, when the angle of identity representations in spherical space is larger than a pre-set threshold, the identity distance will exceed the threshold of face recognition model, achieving the target of de-identification. A personalized and invertible face de-identification method was proposed in [80], where the encryption process in hyperspherical space is independent of a deep generative network. Various de-identification results can be generated with the user-specific password, the degree of identity variation can be controlled by privacy level parameters and the original identity can be restored with the correct password. Similarly, MRDD-FID [156] proposes a manipulable, reversible, and diverse de-identification framework that enables individuals to only modify identity-related representations. Random passwords are used for identity variation to ensure the randomness of the identity modification. MRDD-FID applied StyleGAN2 as a high-quality face generator and injected identity features by the weight modulation/demodulation layer. The degree and direction of de-identification can be precisely controlled according to the intensity and style specified by the user, without affecting the image quality.

Another systematical face de-identification method [107] proposed to de-id for humans by face swapping and de-id for face recognition model by adversarial perturbations. Similarly, a pre-trained face recognition model is used as the identity encoder and another Inception-V3 network acts as the expression and pose encoder, where the latent vector is obtained by combining the two parts. Additionally, the adversarial vector mapping network was designed to calculate the adversarial latent vector by propagating the gradient back instead of adding adversarial noise to the original images. Finally, the de-identified results could be generated with the adversarial latent vector by StyleGAN.

Some recent methods introduce more advanced 3D face generation models into face de-identification. IDeudemon [111] employs a "divide and conquer" strategy, gradually preserving identity and utility while maintaining good interpretability. In Step I, the 3D reconstruction ID codes calculated by the NeRF model are blurred to protect identity. In the second step, a face restoration GAN with adjusted losses has been introduced for visual similarity assistance. Noticed that 2D generation models are difficult to produce identity-consistent results for multiple views, Cao et al. [110] focus on identity disentanglement in the latest 3D-aware face generation models and propose a comprehensive framework that can be applied to single-view, multi-view and a group images of the same identity.

**$k$-Same.** Meden et al. [101] exploited GNNs to generate surrogate faces, which first extract feature vector by a pre-trained face network and select $k$ most similar identities in the gallery for de-identified results generation. Chi et al. [100] proposed to generate "average face" by suppressing facial identity-preserving (FIP) features, where FIP features can reduce intra-identity variances while maintaining inter-identity distinctions. At the same time, FIP features are invariant to pose and illumination, and the de-identification results are generated by replacing that of the original with the averaged FIP features of $k$ individuals. Based on the hypothesis that

face appearance contains identity factor and utility factor (expression, sex, etc.), Chuanlu et al. [106] proposed to separate the multi-dimensional representation with regard to identity and expression with Information Gain (IG) feature selection method. In the de-identification process, $k$-anonymity procedure is just applied in the identity subspace while keeping the utility subspace consistent.

FICGAN [40] was developed as an autoencoder-based conditional generative model, where the ID and non-ID latent codes disentanglement and layer-wise generator structure ensure the controllability between de-identification and attribute preservation. Two types of contrastive verification loss [157], [158] were introduced for a better identity. Non-ID latent code is inserted into the generator for structural attributes and ID latent code is processed by adaptive instance normalization (AdaIN) [159]. FICGAN provided various privacy protection modes including manifold $k$-same algorithm [54], controlling degree, attribute control, example-based control and face swapping.

**Differential Privacy.** Differential Privacy has also been introduced into face de-identification methods at the semantic level, especially identity representations. Croft et al. [108] proposed the first framework to apply differential privacy to facial identity obfuscation via generative machine learning models. The network took one or more class vectors as input and each identity vector was considered as a specific class to be obfuscated. In order to solve the problem that unknown identities cannot be generated, it is proposed to approximate an unknown identity as a weighted sum of the known identities and formulate the approximation as an optimization problem. IdentityDP [79] was the first to introduce the rigorously formulated DP theory into face anonymization and applied differential privacy mechanisms to deep neural networks for adjustable privacy control. The framework includes facial representations disentanglement, $\varepsilon$-IdentityDP perturbation and image reconstruction, where various de-identification results can be generated according to different privacy budgets. When adding differential privacy perturbation to the identity representation, those attributes remain unchanged for sharing more similarities with the original.

**Discussion:** At present, it is considered that adding appropriate disturbance to the disentangled identity embedding is an effective method to preserve the visual similarity as possible, which can obtain visually similar face photos while hiding identity information. How to design the model to learn a more accurate identity representation and how to disentangle the identity and attributes more effectively are the focus of further research.

### *E. De-identification in Videos*

Compared to image processing, the processing of video is more complex, including the consistency of adjacent frames, the temporal consistency of identity, the naturalness of background in generated images and so on.

**Method:** Enforcing spatio-temporal constraints is important for extending image-based methods to video face de-identification. For CIAGAN [35], the same framework could be used for both images and videos because its landmark-based representation of the input face ensures pose preservation and rudimentary temporal consistency. In a similar attempt and inspired by Vid2Vid GAN [160], Balaji et al. [87] utilized an initially image-based GAN to anonymize video frames, and coupled it with several mechanisms to guarantee temporal coherence whilst video processing. Notably, a conditional video discriminator, a pre-processing phase, and a "burn-in" stage, respectively, enforce consistency between consecutive frames, within each sequence and across sequences. FIVA [161] is able to maintain the same face anonymization consistently over frames with a suggested identity-tracking and guarantees a strong difference from the original face.

Gafni et al. [162] proposed a novel feed-forward encoder-decoder network to de-identify live videos. The model holds a host of novelties. First, the representation from an existing face recognition network is concatenated to enrich the latent space, thereby augmenting the auto-encoder to be able to feed-forwardly treat new identities unseen during training. Second, a new attractor-repeller perceptual loss is introduced to maintain source expression, pose and lighting conditions, while surrogating the target identity. This is achieved by employing a perceptual loss between the source and generated output on several low-to-medium abstraction layers, while distancing the high abstraction layer perceptual loss between the target and generated output.

Savic et al. [97] designed a semi-adversarial framework which processes an input video by adding unobtrusive perturbations that remove biometric privacy while preserving the rPPG signal and visual appearance. Unlike the previous works that deal with pronounced facial spatial features, they aim at preserving the underlying physiological signals that depend on subtle spatial-temporal features. This means that the applied perturbations need to be temporally consistent and not detrimental to the subtle visual cues that compose the rPPG signal. Therefore, they take interframe dependencies into account when introducing an rPPG preserving constraint, and choose a 3D-CNN as the autoencoder for spatial-temporal modeling. The features are only spatially upsampled and downsampled to not degrade the delicate temporal information.

Some other researchers tackle video anonymization by swapping faces. Among face-swapping models, the most famous one is faceswap/deepfake project [163]. It first trains two auto-encoders with a common encoder but different decoders, one on the target face dataset and the other on the source face dataset. Then face swap is achieved by exchanging the two decoders. Zhu et al. [4], showcased the utility of this method in the video de-identification task and proved that it is key point invariant.

Afterwards, Mukherjee et al. [89] proposed RID-Twin, which stands for Re-enacting Inpainting based De-identified-Twin. Specifically, they decouple human identity from motion and re-enact a de-identified twin actor, thereby transferring motion to a de-identified image. They evaluated their pipeline regarding the de-identification level, identity consistency, and expression preservation.

Furthermore, Park et al. [164] introduced a framework that combines face verification-enabled deidentification techniques

with face-swapping methods, tailored for video surveillance environments. This method maintains high face recognition performance (98.37%) across various facial recognition models while achieving effective de-identification.

Most methods in the literature require facial landmarks to perform video de-identification. However, Proença [105] proposed UU-Net, a landmarks-free and reversible face de-identification method for video surveillance data. UU-Net is composed of two sequential modules, one that receives and de-identifies the raw stream, and the other that reconstructs, disclosing the original identities. Joint optimization of these two entities intrinsically implies shared knowledge on encoded features, assuring proper reconstruction, hence reversibility. Also, he keep full control over the facial attributes that are preserved/changed between the raw and de-identified streams. This is achieved by inferring a pre-trained multi-label CNN classifier to estimate agreeing/disagreeing labels (ID, gender, etc.) between images, whose pairwise responses are later used to constrain the properties of the de-identified elements, as well as temporal consistency in the same sequence.

Zhu et al. [112] introduced an innovative framework for video face de-identification that leverages diffusion autoencoders to disentangle identity and motion features from videos. Unlike existing GAN-based methods, this model offers a robust approach that excels in both reconstruction and de-identification tasks due to its foundation in diffusion models.

While previous methods emphasize designing novel networks to process face videos frame by frame, Wen et al. [109] was the first to introduce motion flow into video de-identification tasks to avoid per-frame evaluation. Their method first infers a motion flow generator from [165] to compute the relative dense motion flow between pairs of adjacent original frames and runs the de-identification only on the first frame. The de-identified video will be obtained based on the anonymous first frame via the relatively dense motion flow.

Meanwhile, because videos produced through live streaming can cause problems such as infringement of others' portrait rights and exposure of trademarks, Kim et al. [88] attempted to de-identify real-time videos. They apply de-identification to the landmark area on the object's edge, so it can only de-identify the area that connects the facial feature points, which allows for the preservation of information other than the face. In particular, this method allows for arbitrary selection of de-identification targets.

**Discussion:** The main challenge in video de-identification lies in how to ensure seamless modification of the original stream without causing many visual artifacts such as flickering or distortion while preserving temporal consistency across frames. Different methods address it differently, smoothing by spline interpolation over neighboring frames [35] or learned blending mask [162], enforcing resemblance of temporal dynamics to real sequences and adding past frames into input tensors [87]. Also, data augmentation on input images proves to enhance semantic mapping and robustness on color, lightning [162], [163]. Besides, further exploiting data redundancy and continuity in videos [89], [97], [109] could provide an interesting line of research.

## IV. Applications

We present some representative examples of face de-identification applications, which can be divided into targeting specific application scenarios and designing for other computer vision tasks.

### A. Specific Areas

Zhu et al. [4] proposed to protect patient privacy in medical research by deepfakes. Traditional methods wiped out facial information or body keypoints entirely, which are important in medical diagnoses. Different from other algorithms, there is no need for de-identification in medical research to consider the appearance similarity preservation with the original, so face swapping technology could be applied to eliminate the ethical restrictions of medical data. Similarly, Tian et al. [99] generate a context image database with more than 300 fake identity images by StyleGAN and the target is selected by race, age and gender. Then, the target image with a fake identity will be fused with the original image to modify the identity in dental patient images.

Wearable cameras have been widely used for capturing lifevlogs, and a face de-identification system designed for wearable camera images [166] has been proposed to replace the original faces with synthetic images. The personal information contained in the face can be deleted, but the face is allowed to remain intact, which health apps can benefit. The proposed system consists of six components: face detection, face alignment, face recognition (ArcFace [26]), synthetic face generator, and face swapping (SimSwap [167]). The surrogate face is chosen by calculating an embedding centroid and the cosine similarity should be lower than the threshold to ensure it has a different identity.

FPGAN [5] could be applied for de-identification on social robot platforms to ensure the data safety of robot users. First, both the workstation and MAT [168] (an experimental platform for research on social robots are installed in the FPGAN, and the social robot periodically updated weights with face de-identification on the workstation. When the camera captures face images, it should use the networks to ensure that the user's visual privacy is not breached.

Lee et al. [6] proposed a privacy-preserving UAV system to protect identity privacy in the first-person videos taken by UAVs, which just modified face regions instead of blurring or masking the whole area. The trained face-anonymizer could be mounted on the UAV system, so that the people in the video could not be recognized as anyone in the dataset. Additionally, the perception performance required for performing UAV's essential functions, such as simultaneous localization and mapping, would not be degraded when using such anonymized videos.

"My Face My Choice" (MFMC) argues that access to real identities on social networks should be designed by "face", with people in different places on the friendship graph seeing different results for the same original image. It is the first complete system designed for using deepfakes to solve the problem of face ownership on social media platforms. When uploading images from the client, users who are connected

to the uploader will be selectively tagged, and other faces in the image will be replaced with deepfakes. Tagged friends will get their own unique results based on the friendship, especially outsiders will get the result with all fake faces.

### B. Identity-Agnostic Computer Vision Tasks

Considering that it is extremely difficult to retain the utility of all aspects at the same time, some algorithms were designed for specific identity-independent tasks. In this section, we discuss how face de-identification is designed for three broad computer vision tasks: attribute classification, expression analysis, and action detection. Similarly, de-identification can be applied to a wider range by adding corresponding modules to the model design.

*1) Attribute Classification:* In order to enhance the utility in classification tasks, some de-identification algorithms pay attention to how to achieve identity protection without affecting attributes. A natural thought is to retain deep facial attribute information by loss functions. For example, SF-GAN [71] designed the attribute similarity loss to ensure the generated image can share as much attribute similarity as possible with the original one. A task-agnostic anonymization framework [169] was proposed to preserve facial attributes while obfuscating identity by applying identity loss and feature-matching loss [170] to optimize latent codes.

The PRN (Privacy Removal Network) [171] proposed to remove identity-related information and keep facial attributes with two encoders of different architectures, an internal autoencoder and an external autoencoder. FSN (Feature Selection Network) and PEN (Privacy Evaluation Network) can also be chosen according to the application scenarios. This framework can be used to retain a selected single attribute or multiple attributes and achieve de-identification at the same time, and has higher data availability compared to the $k$-same algorithm.

MAP-DeID [172] consists of a multi-channel attribute preservation module and a reconstruction-guided face de-identification module. The attribute preservation module includes three channels, non-facial area, keypoints and facial appearance. The de-identification module implements the modification of the identity while maintaining the authenticity of facial attributes and the naturalness of the image.

Some $k$-anonymity based algorithms implement attribute retention by adding constraints to the selected $k$ images. APFD [173] proposed to estimate the best weights to combine $k$ images by modeling attributes preservation, where the $k$ images selected to share the most similar attributes with the test image to save the corresponding shape and appearance parameters. The attributes preserving face de-identification method [174] designed to remove identity information while retaining all facial attributes, which can be summarized as: (1) first randomly select $k$ faces which may contain the same or different face attributes with the test image; (2) edit attributes of $k$ selected faces by ELEGANT [175] to make them the same as those of the test image; (3) generate the averaged $k$ faces as the de-identified result. GARP-Face [176] balanced utility preservation in de-identification by determining *Gender, Age and Race* attributes and seeking corresponding representatives to preserve them.

*2) Expression Recognition:* Some other de-identification methods emphasize obfuscating identity information while maintaining and recognizing facial expressions or emotion, so that the de-identified results can still be used in corresponding classification tasks.

AnonyGAN [177] is a landmark-guided de-identification solution that maintains the facial pose of the source face. They achieve facial pose preservation by exploiting a Graph Convolution Network to learn the spatial relations between the source and condition landmarks, which are modeled as a bipartite graph. Learned geometrical reasoning is subsequently aggregated to the target's appearance representation. Chen et al. [178] proposed a network that learns an expression-discriminative representation of a face image, disentangled from its identity information. Following a GAN paradigm, the discriminator is designed as a multi-task classifier consisting of three separate auxiliary networks. In addition to distinguishing real face images from synthesized ones, the discriminator is also explicitly imposed to (1) align the expression representation of output face with that of the input face, (2) align the output faces's identity code with the target identity code, earlier fed directly into the generator. Though not specifically designed for face anonymisation, it could be effortlessly adapted to cope with de-identification tasks with minor modifications on the identity discriminator. With a slightly different goal but still under the same approach of amending the discriminator, Aggarwal et al. [179] also attached an auxiliary network serving as an emotion verificator to the discriminator, in order to guarantee maximum emotion similarity after de-identification. GANonymization [180] achieved keeping facial expressions while anonymizing faces. The pipeline consists of four steps: face extraction, face segmentation, face landmarks detection and re-synthesis, where the face generation network was based on pix2pix architecture to convert facial landmarks to a random and high-quality face image with the same landmarks.

Given an input face, Agarwal et al. [181] proposed to select from a set of StyleGAN generated artificial — thus anonymous — faces a proxy face with the "same" emotion and closest facial pose representation. Next, non-biometric attributes are extracted from the input and merged with the chosen proxy face. Intuitively, by directly starting from a proxy face with the same emotion, emotion information is sure to be preserved despite its high computational cost. Similarly, Leibl et al. [182] proposed to generate the synthetic face images at five yaw angles by StyleGAN2 [38] to form a library. Secondly, the source face with a similar pose is selected by head pose estimation [183]. Finally, the de-identified result is generated by swapping the source face using FSGANv2 [184] which can retain pose or expressions. The framework [185] proposed to classify the gender by DeepFace [19] to select a proxy image from a synthetic image dataset generated by StyleGAN [36], and the de-identified result is transformed by SimSwap [167].

*3) Action Detection:* Ren et al. [90] proposed a GAN-based video face anonymizer that performs de-identification of each person's face, while maximally maintaining the information used for action recognition. An adversarial paradigm between a face modifier aiming to minimize face identification

accuracy, and a face identifier as its discriminator enables anonymization. Action preservation, on the other hand, is attained through an action detector attached to the face modifier, trained on modified frames with ground-truth action labels.

*4) Face Detection:* Accurate identification of road users, such as pedestrians and cyclists, is important in intelligent transportation systems. Since data regulations require that the datasets used in training must meet certain privacy standards, LFDA [186] proposes a two-stage approach based on the latent diffusion models. Experimental results prove that LFDA can achieve higher face detection mAP values than other anonymization methods when the face detection model is trained on the de-identified datasets and tested on other non-anonymized datasets.

## V. Evaluation Metrics

At present, there are not yet universally acknowledged evaluation criteria for face de-identification. A mere quantitative evaluation framework for de-identification was proposed in [187], from three perspectives of de-identification, expression preservation and photo-reality. On this basis, we sort out the evaluation indicators used in existing papers and present the definition or explanation of the metrics in this section.

### A. Privacy Protection

- *Identity Distance (ID)*: Almost all face verification models determine whether two images share the same identity by comparing the distance between their identity embeddings. Therefore, we use the distance between identity-vectors $e_{id}$ extracted from the face recognition model, which can be formulated as,

$$\textit{Id-dis} = Dis(e_{id}(X), e_{id}(\mathcal{F}(X))), \tag{5}$$

  where $X$ indicates the original image, $\mathcal{F}(X)$ represents the de-identification result and the specific form of $Dis$ is determined by the face recognition model used to obtain identity embedding $e_{id}$.
- *Successful De-identification Rate (SDR)*: In face verification, if the identity distance exceeds the model's reference threshold, the two images are considered to have different identities. A successful de-identification occurs when the de-identified image's identity differs from the original. Thus, we compare the identity distance with the threshold to calculate the success rate of de-identification, formulated as:

$$\textit{SDR} = 1 - \frac{1}{N}\sum_{i=1}^{N} f_{ver}(X, \mathcal{F}(X)), \tag{6}$$

  where $f_{ver} = 0$ when *Id-dis*$> \tau$, otherwise $f_{ver} = 1$, $N$ is the number of testing.

### B. Utility Preservation

*1) General Utility:*

- *Face detectability (FD)*: The key factor for data utility is that the generated images should appear natural and realistic. This can be measured by assessing how well de-identified faces are detected by face detectors. Two approaches can be used: (1) Face detection rate (%), which is the proportion of de-identified faces detected by a face detector, or (2) face detectors like [188] that provide a confidence score for face detection. Face detectability can then be evaluated by comparing detection scores or softmax probabilities from a face recognition network for both raw and de-identified faces.
- *Landmark Distance (LD)*: Accurate localization of face regions and keypoints is essential in face modification tasks. It is important that pixel-level landmark deviations remain unaffected during the face de-identification process.
- *Peak signal-to-noise ratio (PSNR)*:

$$\textit{PSNR} = 10\log_{10}(\frac{MAX^2}{MSE}), \tag{7}$$

  where $MAX$ is the maximum pixel value of the image and MSE refers to the mean squared error between two images $I$ and $K$ with the resolution of $m \times n$, calculated by $MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$.
- *Structural similarity (SSIM)*: We aim to preserve the similarity between the original and de-identified images, rather than randomly altering the face. Specifically, we want to remove only the privacy-related characteristics, but keep the visual similarity, i.e., contours and luminous condition. The structural similarity (SSIM) is defined as follows:

$$\textit{SSIM} = l(X, D(X))^{\alpha} \cdot c(X, D(X))^{\beta} \cdot s(X, D(X)))^{\sigma}, \tag{8}$$

  where $l(X, D(X))$, $c(X, D(X))$, and $s(X, D(X))$ denote respectively brightness similarity, contrast similarity and structural similarity, and $\alpha$, $\beta$, $\gamma$ their weighting coefficients, usually set to 1. It is important to note that both PSNR and SSIM just focus on objective image quality evaluation and fail to capture many nuances of human perception [189].
- *Fréchet Inception Distance (FID)* [190]: FID measures the distance between the distribution of real and synthesized images, comparing the visual quality of generated samples to real ones. A lower FID indicates greater similarity between real and generated images. Wang et al. [160] proposed a variant of FID for video evaluation, which accesses both visual quality and temporal consistency. They modified a video recognition network by removing its last layers, using the remaining part as the "inception" network. Spatio-temporal feature maps are extracted from each video, and means and covariance matrices are computed for real and synthesized video feature vectors.
- *Learned Perceptual Image Patch Similarity (LPIPS)* [191]: The LPIPS Distance, also known as *perceptual loss*, measures the similarity between two images based on deep features. It was designed to prioritize perceptual similarity, learning the inverse mapping from generated images to ground truth. LPIPS has been shown to correlate better with human perception compared to traditional metrics like MSE, PSNR, SSIM, and FSIM. A lower

LPIPS value indicates higher similarity between images, while higher values indicate greater differences.

*2) Customized Utility:*

- *Attribute preservation*: To evaluate the proposed model's performance in attribute preservation, we use separate classifiers for each attribute to calculate the accuracy rate (%) on demographic features such as hair, gender, age, and skin tone.
- *Expression preservation*: According to the Facial Action Coding System (FACS) [192], Action Units (AUs) represent fundamental muscle movements and serve as a proxy for overall facial expressions. Bursic et al. [187] proposed extracting AUs from both original and de-identified faces and calculating the root-mean-square error (RMSE) averaged across all AUs. For video de-identification tasks, Pearson's Correlation Coefficient (PCC) is also computed to assess how well AUs correlate over time.
- *Temporal Consistency*: For video de-identification, preserving temporal consistency is crucial for data utility. The adapted FID metric can measure temporal consistency in synthesized videos, while empirical issues like flickering or identity warping indicate its absence. To quantify temporal coherence, Balaji et al. [87] proposed the *Identity Invariance Score*, which calculates the identity distance between consecutive frames, averaged over the entire video. This metric assumes that temporal coherence correlates with the invariance of altered identity. Notably, this method can be generalized to assess temporal consistency in other attributes as well.

## VI. Experimental Comparison

We select a few representative algorithms from different categories for qualitative and quantitative experimental comparison, where the selected algorithms are shown in Table IV. There is currently no unified dataset for face de-identification, and we choose the CelebA-HQ dataset [193] based on the suitability of the selected methods, which is a new dataset of $256 \times 256$ derived from CelebA [194] containing 30k images of celebrity faces, and the faces have been cropped based on facial landmarks, so that each of them has a normalized position and rotation. De-identification results generated by different methods are presented in Fig. 3.

The quantitative comparison results of the privacy and utility indicators in Section V are shown in Table V and Table VI, respectively. Since different pre-processing is required between different algorithms, the "original image" used in the evaluation is the result after pre-processing, such as image resolution adjustment, for a fairer performance comparison. In the privacy evaluation, we use FaceNet [1] and face recognition library [2] to calculate the identity distance. In the utility evaluation, we use dlib library [3] for face and landmark detection.

On the basis of the experimental results, we summarize the advantages and disadvantages of each technology category as follows.

- The traditional methods are simple to operate, but the protection performance is limited. When the added disturbance degree is small, it has little impact on image quality but poor protection performance. It is found that increasing the disturbance degree contributes higher protection effectiveness, but the image quality and utility will be damaged greatly, such as the inability to detect faces.
- Deep learning-based methods can deal with the trade-off better with higher-quality de-identification results. *Generative Models-Based Face Generation* methods replace the whole face region at pixel level, they can't consider the similarity with the original. *Generative Models-Based Face Editing* methods at the representation level achieve de-identification by designing loss functions, they often cannot adjust the protection degree after model training, which has insufficient flexibility.
- For the approaches of semantic-level, *attribute-based* methods have diminishing effectiveness in privacy because there is no clear correspondence between attributes and identity. Generally speaking, *identity-based* algorithms have the best performance in controlling the trade-off, and achieves more effective protection through minor but more targeted modification.
- Although adding adversarial noise seems like an ideal protection method, which can affect the recognition accuracy without affecting the image content. However, it is hard to apply to various face recognition models in practice. For example, although Fwakes [126] performs well for Amazon Rekognition Face Verification API while it fails to obtain successful de-identification results in our experiments. On the other hand, the designed noise cannot achieve an "invisible" quality in the current state of the art. There still exists perceptual noise texture even if being combined with other face painting tasks like AMT-GAN [134].

## VII. Future Directions

In this section, we provide a summary of the main challenges and open issues associated with face de-identification methods and then point out future directions in our view.

**Targeted Modification.** Based on the summary and experimental comparison, we think that more fine-grained processing on the part targeted at identity can achieve both more effective protection effects and improved utility. Obviously, if the identity-related parts can be separated and operated, better protection can be achieved under the same disturbance degree. How to disentangle identity and attribute is still an active area with much ongoing research work, including exploring a more explicit latent space, more comprehensive identity representation [150], the introduction of contrast loss [40] in training, etc.

**Controllable and Reversible Privacy.** Different scenarios require varying levels of de-identification, and user needs can differ even within the same context. Therefore, an ideal face de-identification method should be adjustable and controllable to suit diverse applications. One way to increase flexibility

[1] https://github.com/davidsandberg/facenet
[2] https://github.com/ageitgey/face_recognition
[3] https://github.com/davisking/dlib

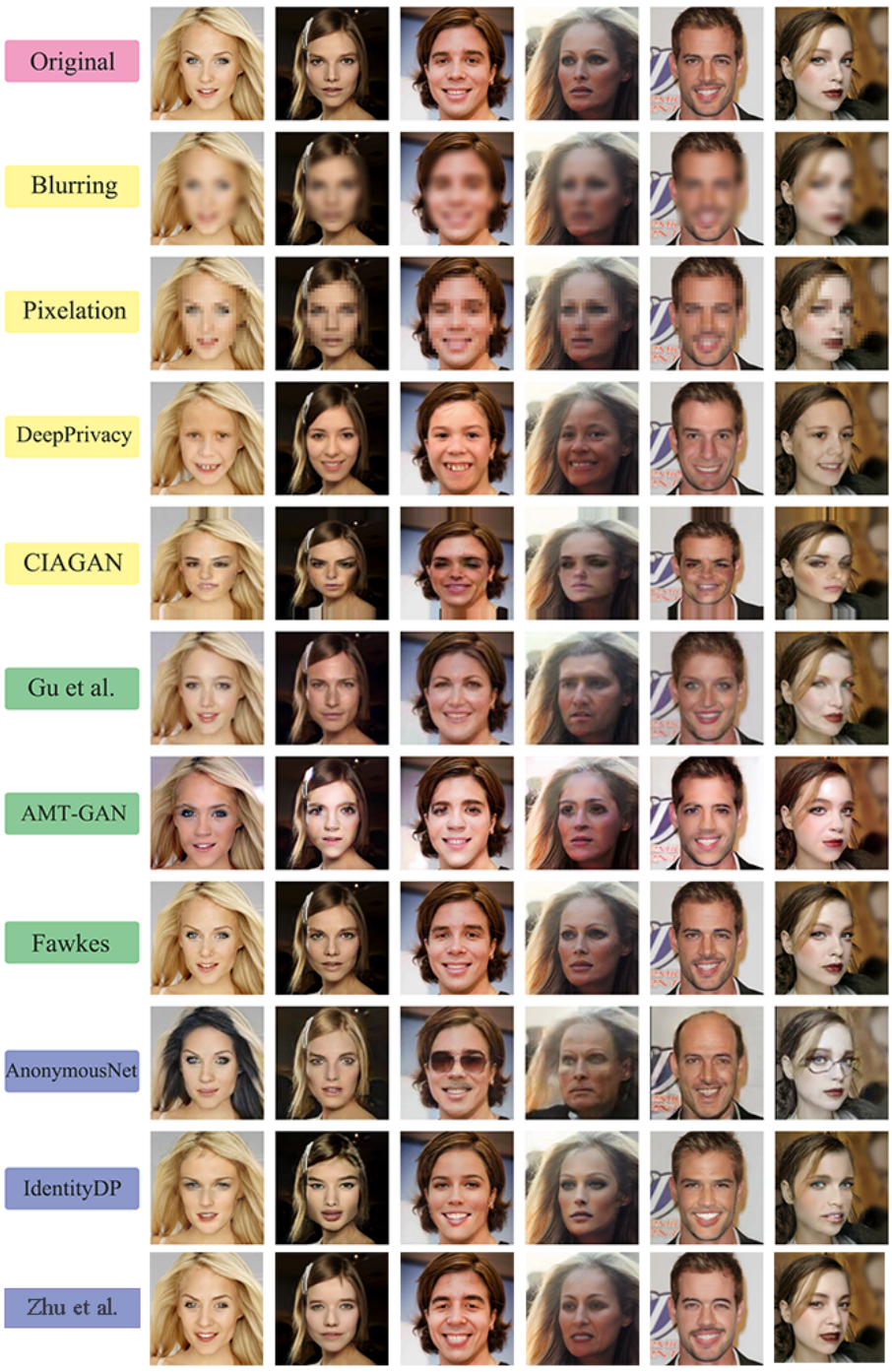


Fig. 3. Qualitative comparison results of representative face de-identification methods on CelebA-HQ dataset [193].

is to separate the protection process from network training. In certain cases, such as crime tracking, the original images may be preferable to the de-identified ones. Currently, most research focuses on protection, with few methods offering recoverable de-identification. Exploring restoration processes in the future could be highly valuable.

**Generalization Across Diverse Applications.** Some algorithms rely on the pre-trained face recognition model to obtain identity embeddings, especially *Generative Models-Based Identity Modification* algorithms and adversarial perturbation, where the former group has better generalization and the latter is often effective only for specific face recognition models. In addition, deep neural networks are always trained in a data-driven manner, so the performance will be influenced greatly by datasets. There is often no uniform standard for face images in practical applications, so it is very challenging to break through the dataset deviation and produce satisfactory results for different inputs.

**Privacy Guarantees.** Provable privacy is another important aspect of face de-identification methods. Existing privacy

TABLE IV
SELECTED METHODS FROM DIFFERENT CATEGORIES FOR COMPARISON.

| Process Level | Methods | Category | Dataset | Resolution |
|---|---|---|---|---|
| Pix-Level | Blurring / Pixelation | III-B1 | - | - |
| | DeepPrivacy [34] | III-B2 | WIDER-Face | 128 × 128 |
| | CIAGAN [35] | III-B2 | CelebA | 128 × 128 |
| Rep-Level | Gu et al. [33] | III-C1 | CASIA / LFW / FFHQ | 128 × 128 |
| | AMT-GAN [134] | III-C2 | Makeup Transfer / CelebA-HQ / LAND | 256 × 256 |
| | Fawkes [126] | III-C2 | WebFace / VGGFace2 | 224 × 224 |
| Sem-Level | AnonymousNet [77] | III-D1 | CelebA | 128 × 128 |
| | IdentityDP [79] | III-D2 | CelebA-HQ | 256 × 256 |
| | Zhu et al. [112] | III-D2 | VoxCeleb | 256 × 256 |

TABLE V
COMPARISON OF THE SELECTED REPRESENTATIVE ALGORITHMS ON CELEBA-HQ [193] UNDER PRIVACY PROTECTION EFFECTIVENESS METRICS, WHERE THE RED ONE REPRESENTS THE BEST AND THE BLUE ONE INDICATES THE SECOND.

| Process-Level | Method | FaceNet (VGGFace2) | | FaceNet (CASIA) | | Face Recognition | |
|---|---|---|---|---|---|---|---|
| | | ID↑ | SDR↑ | ID↑ | SDR↑ | ID↑ | SDR↑ |
| Pix-Level | Blurring ($r$=20) | 1.082 | 0.785 | 1.051 | 0.710 | 0.510 | 0.371 |
| | Pixelation (4×4) | 0.938 | 0.348 | 0.940 | 0.362 | 0.519 | 0.171 |
| | DeepPrivacy [34] | 1.206 | 0.940 | 1.172 | 0.897 | 0.749 | 0.948 |
| | CIAGAN [35] | 0.591 | 0.412 | 0.562 | 0.352 | 0.353 | 0.452 |
| Rep-Level | Gu et al. [33] | 1.261 | 0.936 | 1.238 | 0.935 | 0.838 | 0.978 |
| | AMT-GAN [134] | 0.939 | 0.322 | 0.915 | 0.286 | 0.556 | 0.405 |
| | Fawkes (low-level) [126] | 0.468 | 0.0 | 0.495 | 0.0 | 0.318 | 0.0 |
| Sem-Level | AnonymousNet [77] | 0.724 | 0.452 | 0.806 | 0.395 | 0.489 | 0.534 |
| | IdentityDP ($\varepsilon$=0.57) [79] | 1.225 | 0.912 | 1.184 | 0.906 | 0.753 | 0.965 |
| | Zhu et al. (T=100) [112] | 1.214 | 0.903 | 1.152 | 0.873 | 0.756 | 0.949 |

metrics like $k$–anonymity is based on strong assumptions and fail to handle real-world problems. It is worth considering how to better combine privacy theory with deep learning technology to enhance de-identification algorithms' measurement and interpretability. Additionally, due to the particularity of image data, some improvement based on traditional privacy mechanisms is required to be adapted to the specific tasks.

**Evaluation Standards.** Although we summarize the evaluation metrics commonly used in current de-identification algorithms from both privacy and utility in Section V, they still cannot comprehensively quantify the algorithm performance. Most of the utility metrics are based on the similarity between the de-identified results and the original image. To the best of our knowledge, there is no generally accepted evaluation system or indicator for de-identification. Proposing new evaluation criteria for the characteristics of de-identification algorithms is a driving force for future development in this field.

**Systematic Design for Scalability and Real-time Performance.** There is now a growing need for scalable and efficient de-identification methods that can handle high-resolution images and video streams in real-time. Future work should prioritize the development of lightweight algorithms that maintain robust privacy protection without sacrificing computational efficiency. Techniques such as model compression, pruning, and edge computing could be leveraged to enhance scalability and deployment across resource-constrained environments.

**Broader Multi-modal Privacy Protection Systems.** The increasing sophistication of AI systems means that facial data is often integrated with other biometric identifiers, such as voice, gait, or even behavioral patterns. Future research should investigate how face de-identification techniques can be integrated with multi-modal privacy protection frameworks that secure not only visual data but also other sensitive information. For instance, combining face de-identification with voice anonymization or gesture obfuscation could create a more comprehensive and effective privacy solution, particularly in scenarios where multiple forms of biometric data are simultaneously captured.

TABLE VI
COMPARISON OF THE SELECTED REPRESENTATIVE ALGORITHMS ON CELEBA-HQ [193] UNDER UTILITY PRESERVATION METRICS, WHERE THE RED ONE REPRESENTS THE BEST AND THE BLUE ONE INDICATES THE SECOND.

| Process-Level | Method | FD↑ | LD↓ | PSNR↑ | SSIM↑ | FID↓ | LPIPS↓ | MAE↓ |
|---|---|---|---|---|---|---|---|---|
| Pix-Level | Blurring ($r$=20) | 0.869 | 1.816 | 25.402 | 0.792 | 74.920 | 0.246 | 0.623 |
| | Pixelation (4×4) | 0.971 | 1.398 | 25.927 | 0.777 | 42.747 | 0.276 | 0.657 |
| | DeepPrivacy [34] | 1.0 | 2.225 | 22.084 | 0.823 | 23.412 | 0.383 | 0.582 |
| | CIAGAN [35] | 0.979 | 4.355 | 14.342 | 0.404 | 30.703 | 0.314 | 1.059 |
| Rep-Level | Gu et al. [33] | 0.998 | 3.608 | 23.248 | 0.744 | 42.465 | 0.259 | 0.699 |
| | AMT-GAN [134] | 0.995 | 2.305 | 20.826 | 0.773 | 35.390 | 0.105 | 0.816 |
| | Fawkes (low-level) [126] | 0.990 | 0.426 | 43.960 | 0.994 | 1.092 | 0.005 | 0.269 |
| Sem-Level | AnonymousNet [77] | 0.867 | 3.415 | 20.302 | 0.798 | 55.047 | 0.656 | 0.745 |
| | IdentityDP ($\varepsilon$=0.57) [79] | 0.997 | 1.856 | 24.013 | 0.861 | 23.217 | 0.231 | 0.416 |
| | Zhu et al. (T=100) [112] | 0.992 | 2.185 | 25.142 | 0.878 | 30.495 | 0.311 | 0.402 |

## VIII. CONCLUSION

This survey provided a comprehensive survey of face de-identification for protecting sensitive facial information, especially identity. Based on the proposed taxonomy of image-processing level, we have categorized existing methods in *pixel-level*, *representation-level* and *semantic-level*. With the increasing importance of privacy preservation in the era of AI-driven facial recognition systems, de-identification methods have become essential for protecting individuals' sensitive data while maintaining the usability of images for non-identification tasks. We reviewed state-of-the-art methods and proposed an evaluation framework that measures the effectiveness of these methods in terms of privacy protection and image utility.

In addition, we present practical applications that include specific domains and identity-independent computer vision tasks. Furthermore, we conduct qualitative and quantitative experimental comparisons for each category of algorithms to better demonstrate their performance. Our experimental comparison of representative methods reveals that deep learning-based approaches, especially those leveraging Generative Adversarial Networks (GANs) and diffusion models, have significantly improved the trade-off between privacy protection and utility preservation. However, challenges remain in achieving high fidelity while ensuring robust privacy, particularly under adversarial conditions or complex image distortions. For example, while semantic-level methods offer more flexibility by manipulating high-level features, they often struggle with consistency across varied scenarios. In contrast, pixel-level methods, though more straightforward, frequently degrade image quality, limiting their practical applications.

Looking ahead, several research challenges need to be addressed. First, more robust methods are needed to provide fine-grained, controllable, and guaranteed privacy with image utility across diverse contexts. Second, the field would benefit from a deeper exploration of system designs for more scalable and real-time applications. Third, the integration of privacy-preserving techniques with multi-modal privacy protection systems could create a more comprehensive and effective privacy solution.

In conclusion, while significant progress has been made, face de-identification remains a vibrant area of research with many open questions. This survey offers a foundation for further exploration and highlights both the opportunities and the obstacles in designing next-generation de-identification systems that are robust, efficient, and secure.

## REFERENCES


[1] A. Acquisti, R. Gross, and F. D. Stutzman, "Face recognition and privacy in the age of augmented reality," *Journal of Privacy and Confidentiality*, vol. 6, no. 2, p. 1, 2014.

[2] K. A. Houser and W. G. Voss, "Gdpr: The end of google and facebook or a new paradigm in data privacy," *Rich. JL & Tech.*, vol. 25, p. 1, 2018.

[3] A. Senior, "Privacy protection in a video surveillance system," in *Protecting Privacy in Video Surveillance*. Springer, 2009, pp. 35–47.

[4] B. Zhu, H. Fang, Y. Sui, and L. Li, "Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 414–420.

[5] J. Lin, Y. Li, and G. Yang, "Fpgan: Face de-identification method with generative adversarial networks for social robots," *Neural Networks*, vol. 133, pp. 132–147, 2021.

[6] H. Lee, M. U. Kim, Y. Kim, H. Lyu, and H. J. Yang, "Development of a privacy-preserving uav system with deep learning-based face anonymization," *IEEE Access*, vol. 9, pp. 132 652–132 662, 2021.

[7] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "Visual privacy protection methods: A survey," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4177–4195, 2015.

[8] S. Ribaric and N. Pavesic, "An overview of face de-identification in still images and videos," in *2015 11th IEEE International conference and workshops on automatic face and gesture recognition (FG)*, vol. 4. IEEE, 2015, pp. 1–6.

[9] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey," *Signal Processing: Image Communication*, vol. 47, pp. 131–151, 2016.

[10] C. Liu, T. Zhu, J. Zhang, and W. Zhou, "Privacy intelligence: A survey on image privacy in online social networks," *ACM Computing Surveys (CSUR)*, 2020.

[11] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey toward private and secure applications," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–38, 2021.

[12] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1–36, 2021.

[13] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc, "Privacy–enhancing face biometrics: A comprehensive survey," *IEEE Transactions on Information Forensics and Security*, 2021.

[14] M. Shopon, S. N. Tumpa, Y. Bhatia, K. P. Kumar, and M. L. Gavrilova, "Biometric systems de-identification: Current advancements and future directions," *Journal of Cybersecurity and Privacy*, vol. 1, no. 3, pp. 470–495, 2021.

[15] E. Wenger, S. Shan, H. Zheng, and B. Y. Zhao, "Sok: Anti-facial recognition technology," *arXiv preprint arXiv:2112.04558*, 2021.

[16] G. Zhang, B. Liu, T. Zhu, A. Zhou, and W. Zhou, "Visual privacy attacks and defenses in deep learning: a survey," *Artificial Intelligence Review*, pp. 1–55, 2022.

[17] C. Park, H. K. Jeong, R. Henao, M. Kheterpal *et al.*, "Current landscape of generative adversarial networks for facial deidentification in dermatology: Systematic review and evaluation," *JMIR Dermatology*, vol. 5, no. 2, p. e35497, 2022.

[18] A. Maity, R. More, G. Kambli, and S. Ambadekar, "Preserving privacy in video analytics: A comprehensive review of face de-identification and background blurring techniques," *Authorea Preprints*, 2023.

[19] Y. Taigman, Y. Ming, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2014.

[20] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[21] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 1988–1996.

[22] Y. Sun, D. Liang, X. Wang, and X. Tang, "Deepid3: Face recognition with very deep neural networks," *Computer Science*, 2015.

[23] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[24] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.

[25] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[26] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[30] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE transactions on image processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[31] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[32] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3673–3682.

[33] X. Gu, W. Luo, M. S. Ryoo, and Y. J. Lee, "Password-conditioned anonymization and deanonymization with face identity transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 727–743.

[34] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *International Symposium on Visual Computing*. Springer, 2019, pp. 565–578.

[35] M. Maximov, I. Elezi, and L. Leal-Taixé, "Ciagan: Conditional identity anonymization generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5447–5456.

[36] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[37] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[38] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.

[39] Y. Nitzan, A. Bermano, Y. Li, and D. Cohen-Or, "Face identity disentanglement via latent space mapping," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–14, 2020.

[40] Y. Jeong, J. Choi, S. Kim, Y. Ro, T.-H. Oh, D. Kim, H. Ha, and S. Yoon, "Ficgan: Facial identity controllable gan for de-identification," *arXiv preprint arXiv:2110.00740*, 2021.

[41] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 9, pp. 2805–2824, 2019.

[42] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

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

[44] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[45] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[46] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[47] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[48] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[49] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[50] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[51] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[52] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, "More control for free! image synthesis with semantic diffusion guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 289–299.

[53] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6080–6090.

[54] L. Sweeney, "K-anonymity: A model for protection privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[55] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd international conference on data engineering*. IEEE, 2006, pp. 106–115.

[56] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, no. 1, pp. 3–es, 2007.

[57] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[58] K. Nissim, A. Bembenek, A. Wood, M. Bun, M. Gaboardi, U. Gasser, D. R. O'Brien, T. Steinke, and S. Vadhan, "Bridging the gap between computer science and legal approaches to privacy," *Harv. JL & Tech.*, vol. 31, p. 687, 2017.

[59] R. Cummings and D. Desai, "The role of differential privacy in gdpr compliance," in *FAT'18: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2018.

[60] Z. Ji, Z. C. Lipton, and C. Elkan, "Differential privacy and machine learning: a survey and review," *arXiv preprint arXiv:1412.7584*, 2014.

[61] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[62] A. D. Sarwate and K. Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal processing magazine*, vol. 30, no. 5, pp. 86–94, 2013.

[63] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-based face de-identification," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, 2006, pp. 161–161.

[64] L. Meng and Z. Sun, "Face de-identification with perfect privacy protection," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2014, pp. 1234–1239.

[65] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *PET'05 Proceedings of the 5th international conference on Privacy Enhancing Technologies*, 2005, pp. 227–242.

[66] L. Meng, Z. Sun, A. Ariyaeeinia, and K. L. Bennett, "Retaining expressions on de-identified faces," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2014, pp. 1252–1257.

[67] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, 2005.

[68] R. Gross, L. Sweeney, J. Cohn, F. d. l. Torre, and S. Baker, "Face de-identification," in *Protecting privacy in video surveillance*. Springer, 2009, pp. 129–146.

[69] Z. Kuang, H. Liu, J. Yu, A. Tian, L. Wang, J. Fan, and N. Babaguchi, "Effective de-identification generative adversarial network for face anonymization," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3182–3191.

[70] Y. Qiu, Z. Niu, B. Song, T. Ma, A. Al-Dhelaan, and M. Al-Dhelaan, "A novel generative model for face privacy protection in video surveillance with utility maintenance," *Applied Sciences*, vol. 12, no. 14, p. 6962, 2022.

[71] Y. Li, Q. Lu, Q. Tao, X. Zhao, and Y. Yu, "Sf-gan: Face de-identification method without losing facial attribute information," *IEEE Signal Processing Letters*, vol. 28, pp. 1345–1349, 2021.

[72] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-protective-gan for privacy preserving face de-identification," *Journal of Computer Science and Technology*, vol. 34, no. 1, pp. 47–60, 2019.

[73] B. Meden, Ž. Emeršič, V. Štruc, and P. Peer, "k-same-net: k-anonymity with generative deep neural networks for face deidentification," *Entropy*, vol. 20, no. 1, p. 60, 2018.

[74] Y.-L. Pan, M.-J. Haung, K.-T. Ding, J.-L. Wu, and J.-S. Jang, "K-same-siamese-gan: K-same algorithm with generative adversarial network for facial image de-identification with hyperparameter tuning and mixed precision training," in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2019, pp. 1–8.

[75] B. Liu, M. Ding, H. Xue, T. Zhu, D. Ye, L. Song, and W. Zhou, "Dp-image: differential privacy for image data in feature space," *arXiv preprint arXiv:2103.07073*, 2021.

[76] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6478–6487.

[77] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[78] J. Cao, B. Liu, Y. Wen, Y. Zhu, R. Xie, L. Song, L. Li, and Y. Yin, "Hiding among your neighbors: Face image privacy protection with differential private k-anonymity," in *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2022, pp. 1–6.

[79] Y. Wen, B. Liu, M. Ding, R. Xie, and L. Song, "Identitydp: Differential private identification protection for face images," *Neurocomputing*, 2022.

[80] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Personalized and invertible face de-identification by disentangled identity information manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3334–3342.

[81] B. Samarzija and S. Ribaric, "An approach to the de-identification of faces in different poses," in *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2014, pp. 1246–1251.

[82] Z. Sun, L. Meng, and A. Ariyaeeinia, "Distinguishable de-identified faces," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 4. IEEE, 2015, pp. 1–6.

[83] Q. Sun, L. Ma, S. Joon Oh, L. V. Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5050–5059.

[84] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele, "A hybrid model for identity obfuscation by face replacement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 553–569.

[85] L. Fan, "Image pixelization with differential privacy," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2018, pp. 148–162.

[86] ——, "Practical image obfuscation with provable privacy," in *2019 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2019, pp. 784–789.

[87] T. Balaji, P. Blies, G. Göri, R. Mitsch, M. Wasserer, and T. Schön, "Temporally coherent video anonymization through gan inpainting," *arXiv preprint arXiv:2106.02328*, 2021.

[88] S.-O. Kim, D.-W. Jeong, and S.-Y. Lee, "Method of facial de-identification using machine learning in real-time video," in *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. Springer, 2023, pp. 201–208.

[89] A. Mukherjee, M. N. Choudhury, and D. B. Jayagopi, "Rid-twin: An end-to-end pipeline for automatic face de-identification in videos," *arXiv preprint arXiv:2403.10058*, 2024.

[90] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 620–636.

[91] J. Song, Y. Jin, Y. Li, and C. Lang, "Learning structural similarity with evolutionary-gan: A new face de-identification method," in *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*. IEEE, 2019, pp. 1–6.

[92] T. Kim and J. Yang, "Latent-space-level image anonymization with adversarial protector networks," *IEEE Access*, vol. 7, pp. 84 992–84 999, 2019.

[93] T. Li and M. S. Choi, "Deepblur: A simple and effective method for natural image obfuscation," *arXiv preprint arXiv:2104.02655*, vol. 1, 2021.

[94] Y. Yang, Y. Huang, M. Shi, K. Chen, and W. Zhang, "Invertible mask network for face privacy preservation," *Inf. Sci.*, vol. 629, pp. 566–579, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256756216

[95] Y. Zhao, B. Liu, T. Zhu, M. Ding, and W. Zhou, "Private-encoder: Enforcing privacy in latent space for human face images," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 3, p. e6548, 2022.

[96] X. He, M. Zhu, D. Chen, N. Wang, and X. Gao, "Diff-privacy: Diffusion-based face privacy protection," *arXiv preprint arXiv:2309.05330*, 2023.

[97] M. Savic and G. Zhao, "De-identification of facial videos while preserving remote physiological utility," in *34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023*, 2023.

[98] G. Hanawa, K. Ito, and T. Aoki, "Face image de-identification based on feature embedding for privacy protection," in *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2023, pp. 1–6.

[99] M. Tian, W. W. Lu, K. W. C. Foong, and E. Loh, "Generative adversarial networks for dental patient identity protection in orthodontic educational imaging," 2023.

[100] H. Chi and Y. H. Hu, "Face de-identification using facial identity preserving features," in *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2015, pp. 586–590.

[101] B. Meden, R. C. Mallı, S. Fabijan, H. K. Ekenel, V. Štruc, and P. Peer, "Face deidentification with generative deep neural networks," *IET Signal Processing*, vol. 11, no. 9, pp. 1046–1054, 2017.

[102] D. Cho, J. H. Lee, and I. H. Suh, "Cleanir: Controllable attribute-preserving natural identity remover," *Applied Sciences*, vol. 10, no. 3, p. 1120, 2020.

[103] Y. Wen, B. Liu, R. Xie, Y. Zhu, J. Cao, and L. Song, "A hybrid model for natural face de-identiation with adjustable privacy," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 269–272.

[104] T. Ma, D. Li, W. Wang, and J. Dong, "Cfa-net: Controllable face anonymization network with identity representation manipulation," *arXiv preprint arXiv:2105.11137*, 2021.

[105] H. Proença, "The uu-net: Reversible face de-identification for visual surveillance video footage," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 496–509, 2021.

[106] L. Chuanlu, W. Yicheng, C. Hehua, and W. Shuliang, "Utility preserved facial image de-identification using appearance subspace decomposition," *Chinese Journal of Electronics*, vol. 30, no. 3, pp. 413–418, 2021.

[107] S. Yang, W. Wang, Y. Cheng, and J. Dong, "A systematical solution for face de-identification," in *Chinese Conference on Biometric Recognition*. Springer, 2021, pp. 20–30.

[108] W. L. Croft, J.-R. Sack, and W. Shi, "Obfuscation of images via differential privacy: From facial images to general images," *Peer-to-Peer Networking and Applications*, vol. 14, no. 3, pp. 1705–1733, 2021.

[109] Y. Wen, B. Liu, J. Cao, R. Xie, L. Song, and Z. Li, "Identitymask: Deep motion flow guided reversible face video de-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[110] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Achieving privacy-preserving multi-view consistency with advanced 3d-aware face de-identification," in *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, ser. MMAsia '23. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3595916.3626407

[111] Y. Wen, B. Liu, J. Cao, R. Xie, and L. Song, "Divide and conquer: a two-step method for high quality face de-identification with model explainability," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5148–5157.

[112] Y. Zhu, J. Cao, B. Liu, T. Chen, R. Xie, and L. Song, "Identity-consistent video de-identification via diffusion autoencoders," *2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271574301

[113] B. Jiang, B. Bai, H. Lin, Y. Wang, Y. Guo, and L. Fang, "Dartblur: Privacy preservation with detection artifact suppression," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 479–16 488.

[114] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," *arXiv preprint arXiv:1609.00408*, 2016.

[115] C.-S. Tsai, H.-C. Wu, W.-T. Chen, and J. J.-C. Ying, "Atfs: A deep learning framework for angle transformation and face swapping of face de-identification," *Multim. Tools Appl.*, vol. 83, pp. 36 797–36 822, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259909053

[116] R. Gross, L. Sweeney, F. De La Torre, and S. Baker, "Semi-supervised learning of multi-factor models for face de-identification," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[117] M. U. Saleem, D. Reilly, and L. Fan, "Dp-shield: Face obfuscation with differential privacy." in *EDBT*, 2022, pp. 2–578.

[118] Z. Zhang, T. Wei, W. Zhou, H. Zhao, W. Zhang, and N. Yu, "Facersa: Rsa-aware facial identity cryptography framework," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, pp. 7423–7431, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28573

[119] C. Wang, C. Xu, X. Yao, and D. Tao, "Evolutionary generative adversarial networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 6, pp. 921–934, 2019.

[120] S. Joon Oh, M. Fritz, and B. Schiele, "Adversarial image perturbation for privacy protection–a game theory perspective," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1482–1491.

[121] L. Yang, Q. Song, and Y. Wu, "Attacks on state-of-the-art face recognition using attentional adversarial attack generative network," *Multimedia tools and applications*, vol. 80, no. 1, pp. 855–875, 2021.

[122] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7714–7722.

[123] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1452–1466, 2020.

[124] E. Chatzikyriakidis, C. Papaioannidis, and I. Pitas, "Adversarial face de-identification," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 684–688.

[125] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[126] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *29th USENIX security symposium (USENIX Security 20)*, 2020, pp. 1589–1604.

[127] Z. Pan, J. Sun, X. Li, X. Zhang, and H. Bai, "Collaborative face privacy protection method based on adversarial examples in social networks," in *Advanced Intelligent Computing Technology and Applications*, D.-S. Huang, P. Premaratne, B. Jin, B. Qu, K.-H. Jo, and A. Hussain, Eds. Singapore: Springer Nature Singapore, 2023, pp. 499–510.

[128] M. Ghafourian, J. Fierrez, L. F. Gomez, R. Vera-Rodriguez, A. Morales, Z. Rezgui, and R. Veldhuis, "Toward face biometric de-identification using adversarial examples," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023, pp. 723–728.

[129] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1528–1540. [Online]. Available: https://doi.org/10.1145/2976749.2978392

[130] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 819–826.

[131] J. Yang, W. Zhang, J. Liu, J. Wu, and J. Yang, "Generating de-identification facial images based on the attention models and adversarial examples," *Alexandria Engineering Journal*, vol. 61, no. 11, pp. 8417–8429, 2022.

[132] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang, "Generating adversarial examples by makeup attacks on face recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2516–2520.

[133] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," *arXiv preprint arXiv:2105.03162*, 2021.

[134] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 014–15 023.

[135] R.-I. Chang, C.-Y. Lee, P.-W. Chen, and C.-H. Wang, "Machine learning of k-anonymity data by using feature importance and margin preservation," in *2023 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, 2023, pp. 91–96.

[136] M.-H. Le, M. S. N. Khan, G. Tsaloli, N. Carlsson, and S. Buchegger, "Anonfaces: Anonymizing faces adjusted to constraints on efficacy and security," in *Proceedings of the 19th Workshop on Privacy in the Electronic Society*, 2020, pp. 87–100.

[137] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*. IEEE, 2013, pp. 429–438.

[138] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2287–2296.

[139] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi, "Broadening the scope of differential privacy using metrics," in *International Symposium on Privacy Enhancing Technologies Symposium*. Springer, 2013, pp. 82–102.

[140] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[141] A. Othman and A. Ross, "Privacy of facial soft biometrics: Suppressing gender but retaining identity," in *European Conference on Computer Vision*. Springer, 2014, pp. 682–696.

[142] V. Mirjalili and A. Ross, "Soft biometric privacy: Retaining biometric utility of face images while perturbing gender," in *2017 IEEE International joint conference on biometrics (IJCB)*. IEEE, 2017, pp. 564–573.

[143] V. Mirjalili, S. Raschka, and A. Ross, "Privacynet: semi-adversarial networks for multi-attribute face privacy," *IEEE Transactions on Image Processing*, vol. 29, pp. 9400–9412, 2020.

[144] R. Vera-Rodriguez and R. Tolosana, "Sensitivenets: Learning agnostic representations with application to face images," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 43, no. 6, 2021.

[145] J. Li, L. Han, R. Chen, H. Zhang, B. Han, L. Wang, and X. Cao, "Identity-preserving face anonymization via adaptively facial attributes obfuscation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3891–3899.

[146] H.-P. Wang, T. Orekondy, and M. Fritz, "Infoscrub: Towards attribute privacy by targeted obfuscation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3281–3289.

[147] Q. Li, W. Wang, C. Xu, and Z. Sun, "Learning disentangled representation for one-shot progressive face swapping," *arXiv preprint arXiv:2203.12985*, 2022.

[148] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[149] H. Li, Y. Li, J. Liu, Z. Hong, T. Hu, and Y. Ren, "Zero-shot face swapping with de-identification adversarial learning," in *International Conference on Parallel and Distributed Computing: Applications and Technologies*. Springer, 2022, pp. 101–112.

[150] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, "Hififace: 3d shape and semantic prior guided high fidelity face swapping," *arXiv preprint arXiv:2106.09965*, 2021.

[151] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[152] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7642–7651.

[153] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, and K. Lee, "Diffface: Diffusion-based face swapping with facial guidance," *arXiv preprint arXiv:2212.13344*, 2022.

[154] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8568–8577.

[155] T. Wang, Y. Zhang, R. Zhao, W. Wen, and R. Lan, "Identifiable face privacy protection via virtual identity transformation," *IEEE Signal Processing Letters*, vol. 30, pp. 773–777, 2023.

[156] D. Xiao, J. Xia, M. Li, and M. Zhang, "Manipulable, reversible and diversified de-identification via face identity disentanglement," *Multimedia Tools and Applications*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267976121

[157] M. Kang and J. Park, "Contragan: Contrastive learning for conditional image generation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 357–21 369, 2020.

[158] H. Zhang, Z. Zhang, A. Odena, and H. Lee, "Consistency regularization for generative adversarial networks," in *International Conference on Learning Representations*, 2019.

[159] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[160] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," *arXiv preprint arXiv:1808.06601*, 2018.

[161] F. Rosberg, E. E. Aksoy, C. Englund, and F. Alonso-Fernandez, "Fiva: Facial image and video anonymization and anonymization defense," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 362–371.

[162] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9378–9387.

[163] "Faceswap. github project," 2017, https://github.com/deepfakes/faceswap.

[164] S. Park, H. Na, and D. Choi, "Verifiable facial de-identification in video surveillance," *IEEE Access*, 2024.

[165] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[166] B. Puangthamawathanakun, C. Arpnikanondt, W. Krathu, G. Healy, and C. Gurrin, "Towards face de-identification for wearable cameras," *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266598411

[167] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2003–2011. [Online]. Available: https://doi.org/10.1145/3394171.3413630

[168] J. Yang, S. Li, Z. Wang, and G. Yang, "Real-time tiny part defect detection system in manufacturing using deep learning," *IEEE Access*, vol. 7, pp. 89 278–89 291, 2019.

[169] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, "Attribute-preserving face dataset anonymization via latent code optimization," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8001–8010, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257631514

[170] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 676–18 688, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244908314

[171] J. Liu, Z. Zhao, P. Li, G. Min, and H. Li, "Enhanced embedded autoencoders: An attribute-preserving face de-identification framework," *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9438–9452, 2023.

[172] Y. Cao, Y. Zhang, J. Wu, and Y. Fang, "Multi-channel attribute preservation for face de-identification," *Multimedia Tools and Applications*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270011234

[173] A. Jourabloo, X. Yin, and X. Liu, "Attribute preserved face de-identification," in *2015 International Conference on Biometrics (ICB)*, 2015, pp. 278–285.

[174] B. Yan, M. Pei, and Z. Nie, "Attributes preserving face de-identification," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 1217–1221.

[175] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging latent encodings with gan for transferring multiple face attributes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 168–184.

[176] L. Du, M. Yi, E. Blasch, and H. Ling, "Garp-face: Balancing privacy protection and utility preservation in face de-identification," in *IEEE international joint conference on biometrics*. IEEE, 2014, pp. 1–8.

[177] N. Dall'Asen, Y. Wang, H. Tang, L. Zanella, and E. Ricci, "Graph-based generative face anonymisation with pose preservation," in *International Conference on Image Analysis and Processing*. Springer, 2022, pp. 503–515.

[178] J. Chen, J. Konrad, and P. Ishwar, "Vgan-based image representation learning for privacy-preserving facial expression recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1570–1579.

[179] A. Aggarwal, R. Rathore, P. Chattopadhyay, and L. Wang, "Epd-net: A gan-based architecture for face de-identification from images," in *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*. IEEE, 2020, pp. 1–7.

[180] F. Hellmann, S. Mertes, M. Benouis, A. Hustinx, T.-C. Hsieh, C. Conati, P. Krawitz, and E. André, "Ganonymization: A gan-based face anonymization framework for preserving emotional expressions," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258461439

[181] A. Agarwal, P. Chattopadhyay, and L. Wang, "Privacy preservation through facial de-identification with simultaneous emotion preservation," *Signal, Image and Video Processing*, vol. 15, no. 5, pp. 951–958, 2021.

[182] A. Leibl, A. Meißner, S. Altmann, A. Attenberger, and H. Mayer, "De-identifying face image datasets while retaining facial expressions," in *2023 IEEE International Joint Conference on Biometrics (IJCB)*, 2023, pp. 1–10.

[183] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2155–215 509, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:4880325

[184] Y. Nirkin, Y. Keller, and T. Hassner, "Fsganv2: Improved subject agnostic face swapping and reenactment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 560–575, 2023.

[185] M. Shopon and M. L. Gavrilova, "I got your emotion: Emotion preserving face de-identification using injection-based generative adversarial networks," in *Advances in Visual Computing: 18th International Symposium, ISVC 2023, Lake Tahoe, NV, USA, October 16–18, 2023, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 417–430. [Online]. Available: https://doi.org/10.1007/978-3-031-47969-4_33

[186] M. Klemp, K. Rösch, R. Wagner, J. Quehl, and M. Lauer, "Ldfa: Latent diffusion face anonymization for self-driving applications," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3199–3205, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257019613

[187] S. Bursic, A. D'Amelio, M. Granato, G. Grossi, and R. Lanzarotti, "A quantitative evaluation framework of video de-identification methods," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 6089–6095.

[188] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[189] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[190] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[191] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[192] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[193] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[194] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.