2025 3rd International Conference on Device Intelligence, Computing and Communication Technologies (DICCT) | 979-8-3315-4335-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/DICCT64131.2025.10986474

Faces in the Fog: A Deep Dive into Face De-identification Techniques and Their Comparative Analysis

Shaonti Goswami
Department of CSE
Delhi Technological University
New Delhi, India

Sudipty Konica Paul
Department of CSE
Delhi Technological University
New Delhi, India

Sanyam Jawlia
Department of CSE
Delhi Technological University
New Delhi, India

Anurag Goel
Department of CSE
Delhi Technological University
New Delhi, India

***Abstract*—Face de-identification have drawn a lot of attention due to growing privacy concerns and the growing requirement for anonymization in visual data .This review suggests a systematic exploration of current face de-identification methods, emphasizing three main deep learning-based methods: diffusion models, latent Variable based transformations, and GAN-based models. The in-depth analysis of this literature provides a comparison of how well each strategies protect privacy while retaining data utility by focusing on the advantages and disadvantages of them. Our analysis delves deeper into the benefits and limitations of these models, focusing on important issues such as adversarial robustness, systematic biases, and trade-offs between realistic image generation and privacy preservation. This comprehensive overview synthesizes data from several studies, identifies important gaps in existing approaches and provides guidance to researchers and industry practitioners directing to more reliable and ethical de-identification frameworks.**

***Index Terms*—Face de-identification, Generative deep learning ,Privacy bias and fairness, Identity protection.**

I. Introduction

As artificial intelligence was rapidly advancing and following the extensive deployment of facial recognition systems, the concern regarding protection of personal identity became significantly larger. Governments, industries, and researchers have increasingly reached the awareness of dangers presented by an automated face, which can be used without the permission of others for unauthorized surveillance, theft of identity, or even the generation of deepfakes. Facial images are comprised of sensitive personal information, particularly in the realms of social media, surveillance video, healthcare records, and biometric authentication, which can result in severe violations of privacy if mishandled. The countermeasure against such risks is to develop face de-identification, an area of research that aims to develop algorithms that obscure personal identity effectively but maintain usability and realism of facial images.

From simple pixelation and blurring to more advanced deep learning-based methods which can, at best, produce realistic but anonymized face images, face de-identification techniques have come a long way [1]. The privacy preservation aspect has become more significant than ever due to rigorous data protection laws, such as the GDPR, and growing awareness of digital rights. Images have not been as effective with k-anonymity or differential privacy-like traditional anonymization techniques used with structured data as they would also degrade image utility while preserving anonymity [2], [3]. In an effort to combat this problem, deep learning models, particularly those based on GANs or latent space transformation techniques, were developed, thus showing some level of effectiveness with creating realistic and de-identified faces. However, as de-identification models become more advanced, new challenges arise, such as adversarial reversibility, fairness issues, and the trade-off between privacy and visual quality [2], [4], [5].

A. Motivation

Face de-identification can generally be defined as the process of modifying or replacing identifiable face features without compromising the natural appearance of the image. Various methods have been proposed, from traditional image processing to cutting-edge deep learning. In this review, we discuss face de-identification methods based on a systematic categorization of strategies and their feasibility in maintaining privacy. To extensively research development within this area, we consider the following key research questions: Face de-identification can generally be defined as the process of modifying or replacing identifiable face features without compromising the natural appearance of the image. Various methods have been proposed, from traditional image processing to cutting-edge deep learning. In this review, we discuss face de-identification methods based on a systematic categorization of strategies and their feasibility in maintaining privacy. To extensively research development within this area, we consider few key research questions shown in Table I.

B. Contributions

This review gives a systematic survey of face de-identification methods through a categorization into three major advanced methods: GAN-based methods, latent space manipulation methods, and diffusion methods for de-identification. These methods are described with a comparative categorization based on their underlying mechanism, algorithm type utilized, and their implications on privacy preservation. The review also emphasizes major contributions of various authors, such as the models utilized, their privacy preservation

979-8-3315-4335-8/25/$31.00 ©2025 IEEE

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:41:47 UTC from IEEE Xplore. Restrictions apply.

TABLE I
RESEARCH QUESTIONS & MOTIVATIONS

RQ#	Research Question	Motivation
RQ1	What is the current state-of-the-art in face de-identification?	To enumerate the latest de-identification methods in various contexts.
RQ2	How can GAN-based de-identification models be optimized to prevent full identity exposure while maintaining facial expressions?	To investigate optimizations in GAN-based approaches for best privacy as well as visual realism.
RQ3	How can latent space manipulation techniques be optimized to prevent reconstruction of sensitive features via latent space interpolation?	To investigate methodologies that are resistant to adversary attacks based on latent space loopholes. .
RQ4	What new mechanisms of privacy protection can be integrated into diffusion-based face de-identification methods for enhanced reversibility attack robustness?	To investigate the diffusion model application towards highly secure and natural-looking de-identified faces.

solutions, the utilized evaluation metrics, and the problem of privacy preservation.

The paper is structured as follows: The next section gives a comprehensive literature survey of the GAN-based, latent space-based, and diffusion-based de-identification methods. The pros, drawbacks, and privacy issues of each approach are discussed in comparative analysis that follow. Finally, the conclusion highlights the main obstacles to confront de-identification and suggests methods for further study to improve privacy protection, equity, and real-time application.

II. DETAILED LITERATURE REVIEW

State-of-the-art efforts in facial image anonymization and de-identification have explored an array of generative models, including diffusion models, latent space approaches, and GANs, to reach a tradeoff between data utility and privacy assurance. Diffusion models have been favored to generate high-quality anonymized outputs, with dual-conditional models delivering greater anonymization control [6]. In order to ensure privacy, latent space techniques have been investigated to encode and modify face features while maintaining a concise description for safe data transfer [7]–[10]. With modifications centered on maximizing stability and minimizing artifacts, GANs have been in the vanguard of the synthesis of realistic yet anonymous face images . These strategies are complemented by research efforts in computational cost and scalability for practical application to real-world practice [11]. In aggregate, these studies testify to the universality of multiple generative models in addressing the problem of privacy, leaving stable solutions to the protection of facial data across many applications [12]–[14] as shown in Table II.

A. *GAN Based Model for Face De-identification*

GAN-based models stand as essential tools for face de-identification because their generation of realistic facial images maintains both privacy standards. The several GAN based face de-identification models are summarized in Table III.

GAN Variants and Techniques Conditional GANs (cGANs), StyleGAN and Landmark-Based GANs (L-GAN) that serve different purposes in de-identification tasks. The generation of de-identified faces with cGANs operates based on specific attributes like age or gender while StyleGAN uses latent space manipulation to achieve high-quality output generation . The disruption of identity recognition happens through methods which include latent space manipulation together with face swapping and adversarial perturbation techniques while maintaining non-identity features such as facial expressions and poses [2]. The integration of GANs with autoencoders or diffusion models has become popular for enhancing both de-identification quality and utility preservation. De-identification approaches found between privacy safeguards and perceptually natural face appearances make them applicable for security, emotional analysis and medical information exchange applications.

Performance Evaluation and Output Effectiveness Assessment of the performance of GAN-based face de-identification techniques involves a trade-off between privacy protection, visual quality, and utility to downstream applications [15]. A number of key performance metrics are used.

Privacy Metrics: Privacy effectiveness is generally measured by comparison of face recognition similarity scores both prior to and following de-identification. As the similarity score gets lowers it indicates more effective privacy protection [1]. Furthermore , when advanced face recognition models (e.g., FaceNet, ArcFace) are used , adversarial evaluation are seen in order to estimate re-identification hazards .

Image Quality Metrics: To maintain visually realistic de-identified images, GAN-based models are evaluated using metrics such as Fréchet Inception Distance (FID), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR). High SSIM and low FID scores show that the generated face images are highly human-like while successfully de-identifying the person.

Utility Preservation: There are some methods which emphasize preserving usability for detecting emotion, gaze estimation, and video analysis. For instance, Facemotionpreserve performs facial de-identification in a way that medical diagnostic information is not biased, and so it beomes suitable for healthcare and biometric use cases. Additionally, feature-space adversarial perturbation-based de-identification is examined in terms of impact on machine learning models employed for human behavior analysis [12].

User Perception and Human Factors: A critical consideration of real-world deployments is whether perceptually natural de-identified images are perceived by human beings. Human perception is evaluated in experiments by Mean Opinion Score (MOS) surveys, where the subjects rate the realism of de-identified images . Additionally, in driver monitoring systems, GAN-based face de-identification is evaluated in nat-

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:41:47 UTC from IEEE Xplore. Restrictions apply.

TABLE II
FACE EDITING TECHNIQUES: FROM TRADITIONAL TO ADVANCED DL-BASED METHODS

Approaches	Subcategory	Techniques
Traditional Methods	Pixelation	Minimizing image resolution to anonymize facial details.
	Blurring	Gaussian or motion blurring to conceal facial features
	Masking	Overlaying patterns or shapes to conceal facial features..
	K-Same Algorithm	Substituting faces with average faces from a dataset to anonymize.
	Eigenface-Based Anonymization	Utilizing eigenfaces to anonymize faces in images.
Advanced DL-Based Methods	GAN-Based Models	StyleGAN: Style-based face generation for realistic face synthesis.
		StarGAN: Multi-domain face modification using a single generator.
		AttGAN: Attribute-based face editing (e.g., age, gender, hair color).
	Latent Variable Models	Latent Space Interpolation: Smooth face transitions.
		Latent Space Projection: Images projected to latent space for editing.
		Latent Space Manipulation: Individual attributes edited in latent space.
	Diffusion-Based Models	DDPM (Denoising Diffusion Probabilistic Models): Iterative denoising.
		Latent Diffusion Models (LDM): Integrates diffusion with latent space.
		Stable Diffusion: High-quality face development and editing.

uralistic driving scenarios, hence not interfering with safety-critical use cases [1].

B. Latent Variable Model for Face de-identification

Latent space-based face de-identification employs learned feature representations in generative models to anonymize facial identities without loss of significant attributes such as pose, expression, and facial structure. The overview is shown in Table IV.

Latent Variants and Techniques: Latent space mapping is one of the widely employed techniques, which manipulates identity-sensitive features without deforming the natural look of the face. This technique ensures the withholding of significant facial attributes at the same time preventing identity leakage [8]. Another strong technique is layer swapping, which determines that a limited number of layers in the latent representation are swapped to generate anonymization while retaining key facial details. By specifically modifying deep layers containing identity, this technique generates realistic but anonymized faces [3].

Apart from these transformation-based techniques, latent code optimization enhances the latent representation to perform optimal de-identification by minimizing identity similarity. This technique adjust the latent vector to reduce re-identification risks while maintaining facial attributes such as age and gender [2]. Similarly, identity-aware variational autoencoders (IA-VAE) implements a organized disentanglement of identity-related and identity-irrelevant features. By learning a disentangled enactment, this method supports efficient face-swapping while retaining natural appearance [16]. Another important technique is secure dual networks which employs reversible transformations in the latent space, while supports supervised retrieval of identity under authorized conditions [17]. These techniques altogether provide a trade-off between privacy maintainance and the retention of significant facial attributes.

Performance Evaluation and Output Effectiveness The evaluation metrics for Latent Variable based Face de-identification is estimated by :

Privacy Metrics: Face Verification Rate (FVR) and Cosine Similarity Score (CSS) evaluates privacy protection. When there is lower CSS value it reveals better anonymization, with HFORD offering exceptional privacy when CSS is below 0.3 [16], [18].

Image Realism Measures: PSNR, SSIM, and FID estimates image realism. SSIM above 0.85 yields natural faces, while below 0.6 causes distortion. FID less than 25 offers high realism, while over 40 implies excessive feature loss [17], [19].

Identity Preservation and Usefulness: L-GAN eliminates identity details but preserves expressions for applications like emotion detection. APLO preserves age and gender attributes with a success rate of over 90%, striking a balance between usability and privacy [2], [9].

Bias and Fairness Considerations: Equalized Odds Difference (EOD) and Demographic Parity (DP) facilitate minimization of bias. Models like People Taking Photos That Faces Never Share learn transformations in latent space for fair de-identification over groups [11].

C. Diffusion Based De-identification:

Diffusion-based face de-identification methods apply generative diffusion models to iteratively optimize images for privacy maintenance while preserving crucial attributes.

Diffusion Variants and Techniques: The front-running variants are score-based generative models, denoising diffusion probabilistic models (DDPMs), and conditional latent diffusion models. IDDiffuse proposes a dual-conditional diffusion model with identity priors to control anonymization and attribute preservation [6]. Diff-Privacy leverages a noise-guided diffusion process that continuously deforms and reconstructs images to achieve irreversible masking of identities [21].

Latent diffusion models (LDMs) were found to be a prospective attribute-preserving de-identification option. LDMs act upon data within a latent compressed domain to optimize computation cost at realism preservation. Piano et al. demonstrate an LDM-based anonymization criterion with assured identity-controlled removal by protecting crucial facial attributes [10]. Campos et al. apply LDMs to obscure medical images with domain applicability demonstrations [7]. Other techniques, like Realistic Anonymization using Stable Diffusion (RAD), operates pre-trained stable diffusion models to introduce visually realistic anonymized images [22].

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:41:47 UTC from IEEE Xplore. Restrictions apply.

TABLE III
COMPREHENSIVE OVERVIEW OF MODELS AND VARIANTS

Model Name	Variant Name	Authors	Technique Used	Mathematical Formulation	Performance Evaluation Metric	Strengths	Limitations
GAN-Based Models	CycleGAN	[13]	Optical face anonymization	$L_{GAN}(G,D) = \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(G(z)))]$	High-quality image translation with minimal artifacts in case of high PSNR and SSIM value	Vigorously handle adversarial attacks	Facial datasets are limited and so requires advanced hardware structures.
	Conditional GAN	[1]	Two-step face swapping framework	$L_{cGAN}(G,D) = \mathbb{E}[\log D(x,y)] + \mathbb{E}[\log(1 - D(G(z,y),y)]$	High SSIM and FID ensure realistic face swaps with preserved details.	Robust to pose and lighting variations	Computationally expensive, limited to face datasets
	Wasserstein GAN	[5]	Anomaly detection in IoT networks	$L_{WGAN}(G,D) = \mathbb{E}[D(x)] - \mathbb{E}[D(G(z))]$	High F1-Score ensures effective detection of rare anomalies.	Scalable and effective in detecting anomalies	High false-positive rate in noisy environments
	AttnGAN	[1], [14]	Text-to-image synthesis	$L_{AttnGAN}(G,D) = L_{GAN} + \lambda L_{KL}$	High R-precision and FID ensure accurate alignment of text and images.	Aligns well with text descriptions	Struggles with complex textual inputs
	Temporal GAN	[5], [15]	Video frame interpolation	$L_{Temporal}(G,D) = \mathbb{E}[\log D(x_t, x_{t+1})] + \mathbb{E}[\log(1 - D(G(z)))]$	High VMAF and PSNR ensure smooth frame interpolation with minimal artifacts.	Handles motion blur effectively	Struggles with fast-moving objects
	3D-GAN	[14]	3D object reconstruction from 2D images	$L_{3D-GAN}(G,D) = \mathbb{E}[\log D(x)] + \mathbb{E}[\log(1 - D(G(z)))]$	Low Chamfer Distance and FID ensure accurate 3D reconstructions.	Robust to noisy inputs	Limited to small-scale objects
	Progressive GAN	[15], [12]	High-resolution medical image synthesis	$L_{Progressive}(G,D) = \sum_{i=1}^{n} L_{GAN}(G_i, D_i)$	High Dice Coefficient and FID ensure high-quality medical image synthesis.	Enhances dataset diversity for medical AI models	Limited to specific medical imaging modalities
	SRGAN	[20]	Super-resolution for satellite images	$L_{SRGAN}(G,D) = L_{GAN} + \lambda L_{perceptual}$	High PSNR and SSIM ensure enhanced image resolution for remote sensing.	Enhances image resolution for remote sensing	Requires high computational resources
	StyleGAN	[12], [8]	High-resolution image synthesis	$L_{StyleGAN}(G,D) = L_{GAN} + \lambda L_{style}$	High IS and FID ensure high-quality image synthesis with fine details.	Captures fine details in images	Requires large computational resources
	AnoGAN	[20], [9]	Anomaly detection in industrial systems	$L_{AnoGAN}(G,D) = L_{GAN} + \lambda L_{anomaly}$	High Precision and Recall ensure effective detection of rare anomalies.	Effective in detecting rare anomalies	High false-positive rate in noisy environments

Denoising-based de-identification models play a role too in the process of eliminating identity. Zhu et al. propose diffusion autoencoders that advance video anonymization by promoting consistency in the frame while preserving details of motion and expression [23]. Threats to security are present nonetheless since Jiang et al. present the Denoising Diffusion Cryptanalytic Model (DDCM), an alternative that could be used for extracting anonymized images, leading to compromised privacy.

Performance Evaluation and Output Effectiveness: The metrics for diffusion based de-identification determines how well a model can anonymize faces while preserving essential visual attributes.

Privacy Metrics: Cosine Similarity Score (CSS) and Face Verification Rate (FVR) determines how well de-identification works. By reducing CSS below 0.3, IDDiffuse and Diff-Privacy perform strong anonymization and minimize the risk of re-identification [6], [21]. However, by re-generating de-identified pictures, adversarial recovery models such as DDCM reveal possible vulnerabilities and challenge robustness.

Image Quality Metrics: Visual fidelity is performed using the Fréchet Inception Distance (FID) and the Structural Similarity Index (SSIM). Natural face composition are preserved by Stable Diffusion-based models such as RAD, which maintain high SSIM (>0.85) [22]. Significant face distortions take place when SSIM falls below 0.6. When FID values is greater than 40 , it generates observable changes whereas values below 25 implies realistic results [10].

Identity Retention and Attribute Preservation: Diffusion autoencoders afflict a trade-off between anonymization and usefulness by retaining some characteristics, such as age and expression. While suppressing identity-relevant characteristics, LDM-based methods assure an attribute preservation rate of over 90%. [7], [10]. By avoiding frame-wise inconsistencies, video anonymization algorithms also preserve temporal coherence [23].

III. CHALLENGES AND COMPARISON

Different approaches are used by face de-identification models to influence a trade-off between computing productivity, realism, and privacy. Each of the three methods—GAN-based, latent variable-based, and diffusion-based—has strengths and weakness with regard to robustness, picture quality, and identity removal. GAN based models are prompt and realistic yet prone to adversarial attacks. Latent variables prefer precise control over privacy protection but they often face difficulties in complete de-identification. Advanced privacy protection is provided by diffusion-based models, but algorithmic efficiency is sacrificed in the process. The particular application will determine which method is best, whether privacy, visual quality, or real-time processing are the top priorities. Table V depicts the advantages and limitations of various face de-identification approaches.

A. *Advantages:*

1) Advantages of GAN Based Model:: **GAN-based** approaches are preferable at producing incredibly precise de-identified faces at the same time maintaining crucial non-identifiable characteristics. By learning disentangled facial features, they allow attribute-preserving de-identification, which effectively protects privacy without considerably sacrificing visual quality. Though StyleGAN-based methods guarantee utility preservation, approaches such as RBGAN and L-GAN balance anonymity and authenticity [24]. Addition to this, real-time processing is supported by GAN-based solutions, which makes them perfect for streaming settings and video analytics [9], [15].

2) Advantages of Latent Variable Based Model:: Variational autoencoders (VAEs) and latent diffusion models under **Latent space manipulation** methods provide precise benchmark over identity abolition while maintaining significant facial attributes . In spite of totally altering the original picture the methods, such as embedding-based de-identification

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:41:47 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPREHENSIVE OVERVIEW OF MODELS AND VARIANTS

Model Name	Variant Name	Authors	Technique Used	Mathematical Formulation	Performance Evaluation Metric	Strengths	Limitations
Latent Variable-Based Models	VAE-GAN	[8], [9]	Image generation with latent variable modeling	$L_{VAE-GAN} = L_{GAN} + \lambda L_{KL}$	High FID and IS ensure high-quality image generation with latent space control.	Combines generative and latent space modeling	Struggles with complex datasets
	β-VAE	[9], [3]	Disentangled representation learning	$L_{-VAE} = \mathbb{E}[D(x)] - \beta\mathbb{E}[D(G(z))]$	High disentanglement score ensures effective separation of latent factors.	Effective for unsupervised learning	Limited to small-scale datasets
	InfoGAN	[3]	Unsupervised disentangled representation learning	$L_{InfoGAN} = L_{GAN} + \lambda L_{MI}$	High mutual information ensures effective disentanglement of latent variables.	Captures interpretable latent factors	Computationally expensive
Diffusion-Based Models	DDPM	[17]	Image generation via iterative denoising	$L_{DDPM} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\|\epsilon - \epsilon_\theta(x_t, t)\|\|^2\right]$	High PSNR and SSIM ensure high-quality image generation.	Stable training and high-quality outputs	Computationally expensive
	Score-Based Generative Model	[17], [4]	Image generation via score matching	$L_{Score} = \mathbb{E}_{x\sim p(x)}\left[\|\|\nabla_x \log p(x) - s_\theta(x)\|\|^2\right]$	High FID ensures realistic image generation.	Effective for high-resolution image synthesis	Requires large computational resources

TABLE V
MODEL TYPES, ADVANTAGES, AND LIMITATIONS

Model Type	Advantages	Limitations
GAN-Based Models	a. High-quality anonymization takes place. (e.g., L-GAN, GAN-based de-identification) [9] b. The facial features are maintained at the time of anonymization (e.g., StyleGAN, L-GAN) [9] c. Ethical preservation of identity-consistency (e.g., RBGAN, L-GAN) [1] d. Capable for reversible anonymization (e.g., RBGAN, Diffuse)	a. Training and inference need high computational cost.(e.g., StyleGAN) [9] b. Issues related to model interpretability. (e.g., L-GAN) [9] c. Problems related to non-human faces and so limited to certain facial features.(e.g., RBGAN) [1] d. Issues prevails related to relaistic image generation when facing obstruction. (e.g., GAN-based models)
Latent Variable-Based Models	a. High-quality image generation even with occlusion and performs robust face de-identification (e.g., HFORD, Identity-aware VAEs) [18] b. Better maintenance of privacy preservation. (e.g., RBGAN, Face anonymization via latent space mapping) c. Reversible facial features are allowed in some methods (e.g., HFORD) [16] d. Facial attributes are protected during anonymization (e.g., Latent Space Manipulation) [8]	a. Generalization might not be well across all datasets (e.g., latent space manipulation) [8] b. Rigorous computation is seen. (e.g., VAE-based models) [16] c. In order to train effectively large datasets are required. [8]
Diffusion-Based Models	a. Well organized for privacy-preserving face de-identification. (e.g., DDCM, IDDiffuse) b. Denoising and perturbation maintains improved identity preservation. (e.g., Diff-privacy) [21] c. Re-identification might not take place . (e.g., Diffusion-based Face Privacy Protection) [6]	a. Resource consumption is potentially high.(e.g., large diffusion models) b. Natural images geberation might be less compared to GAN-based models (e.g., Diffusion-based methods) [21] c. Training phase required large labeled datasets. [10]

and latent space optimization enable structural changes that protect privacy . In sensitive applications such as medical imaging, secure dual-network and attribute-preserving latent space models enhance usability and reversibility [2], [17]. Furthermore, in comparison to direct pixel-space alterations, latent space approaches provide more advanced generalization across a variety of datasets and applications [8].

3) Advantages of Diffusion Based Model:: **Diffusion models** provide anonymized faces with exceptional persistent quality and privacy guarantees by utilizing probabilistic changes. In terms of preserving identity reconstruction attempts while retaining natural look, they perform far better than conventional GANs. Identity-consistent anonymization and enriched security in privacy-sensitive areas are offered by techniques like IDDiffuse and Diff-Privacy [6], [21]. Latent diffusion models are very favorable at anonymizing medical and forensic datasets because they further improve attribute preservation [7], [10]. Diffusion models also provide controlled de-identification at diverse noise levels, assuring flexibility in a range of real-world situations.

B. Limitations:

Although **GAN-based face de-identification** techniques have their benefits, they are prone to confronting attacks, identity leakage, and mode collapse [15], [24]. They also struggle to create a balance between privacy and functionality, which subsequently results in severe modifications that produce unrealistic images. [9] **Latent variable-based** approaches might result in undesirable biases since they conflict ensuring strong identity removal in spite of being functional for preserving facial features [2], [17]. Because of advanced training and learning processes, this model need a lot of computing resources [8]. Even though **diffusion-based methods** are more secure and of higher quality, they are computationally costly and have extensive inference times [6], [21]. Also, their clarity and control over de-identification levels are bounded which are making them incompatible for real-time applications [7], [10]. Limitations of these various techniques are shown in Table V.

IV. FUTURE RESEARCH DIRECTIONS:

Context-aware anonymization is a likely-looking approach in which the model dynamically adjusts the intensity of de-identification in response to risk estimation. Reinforcement learning or uncertainty-aware models can be used for this that measure the necessary degree of anonymization while maintaining image functionality . Since many current methods focus on combining sensitive face data for training, robust models may be trained without collecting user data by incor-

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:41:47 UTC from IEEE Xplore. Restrictions apply.

porating federated learning (FL) with self-supervised learning (SSL). FL ensures that training takes place locally on user devices, and SSL can improve feature extraction without labeled data, reducing the possibility of privacy risks. Face de-identification often rely only on visual data, but real-world privacy exposition also come from textual and biometric derivations. Another future direction can be developing multi-modal anonymization frameworks that anonymize facial images, textual descriptions, and biometric markers (e.g., gait patterns, voice data) together. As a result, privacy protection will be ensure in applications like social media, surveillance, and medical imaging.

V. Conclusion

This study examines the evolution in face de-identification methods, evaluating their efficacy, drawbacks, and implementations for privacy protection. This review's aim is to provide readers a comprehensive overview of the development of de-identification techniques, from conventional obfuscation techniques like blurring and pixelation to more complex deep learning approaches like GANs and diffusion models . By looking at different methods, this survey shows how these techniques may handle issues like identity exposure, fairness, and adverse weaknesses while maintaining a balance between privacy and data storage [15]. The paper also critically estimates how self-supervised learning and federated learning might improve privacy without disclosing sensitive data. With the use of this study, the review hopes to clarify on the advantages and disadvantages of current methods, offering a structured comparison that helps researchers and and practitioners in deciding on appropriate de-identification methods for real-world applications.

References

[1] S. Thapa and A. Sarkar, "Gan-based deidentification of drivers' face videos: An assessment of human factors implications in nds data," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–6, 2023.

[2] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, "Attribute-preserving face dataset anonymization via latent code optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8001–8010, 2023.

[3] E. BenSaid, M. Neji, and A. M. Alimi, "Natural face anonymization via latent space layers swapping," in *2023 IEEE Symposium on Computers and Communications (ISCC)*, pp. 562–567, IEEE, 2023.

[4] G. Hanawa, K. Ito, and T. Aoki, "Face image de-identification based on feature embedding," *EURASIP Journal on Image and Video Processing*, vol. 2024, no. 1, p. 25, 2024.

[5] R. More, A. Maity, G. Kambli, and S. Ambadekar, "Privacy-preserving video analytics through gan-based face de-identification," in *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*, pp. 1–6, 2024.

[6] M. Shaheryar, J. T. Lee, and S. K. Jung, "Iddiffuse: Dual-conditional diffusion model for enhanced facial image anonymization," in *Proceedings of the Asian Conference on Computer Vision*, pp. 4017–4033, 2024.

[7] F. Campos, L. Petrychenko, L. F. Teixeira, and W. Silva, "Latent diffusion models for privacy-preserving medical case-based explanations," *submitted to EXPLIMED, ECAI*, p. 3, 2024.

[8] E. BenSaid, M. Neji, and A. M. Alimi, "Faceanonym: Face anonymization model via latent space mapping," in *2023 International Conference on Cyberworlds (CW)*, pp. 480–483, IEEE, 2023.

[9] S.-s. Jang, C.-j. Kim, S.-y. Hwang, M.-j. Lee, and Y.-g. Ha, "L-gan: landmark-based generative adversarial network for efficient face de-identification," *The Journal of Supercomputing*, vol. 79, no. 7, pp. 7132–7159, 2023.

[10] L. Piano, P. Basci, F. Lamberti, and L. Morra, "Latent diffusion models for attribute-preserving image anonymization," *arXiv preprint arXiv:2403.14790*, 2024.

[11] J. Zhu, L. Gu, X. Wu, Z. Li, T. Harada, and Y. Zhu, "People taking photos that faces never share: Privacy protection and fairness enhancement from camera to user," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 14646–14654, 2023.

[12] H. Xue, B. Liu, X. Yuan, M. Ding, and T. Zhu, "Face image de-identification by feature space adversarial perturbation," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 5, p. e7554, 2023.

[13] J. Lopez, C. Hinojosa, H. Arguello, and B. Ghanem, "Privacy-preserving optics for enhancing protection in face de-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12120–12129, June 2024.

[14] J. Song, J. Kim, and J. Nang, "Face de-identification using convolutional neural network (cnn) models for visual-copy detection," *Applied Sciences*, vol. 14, no. 5, p. 1771, 2024.

[15] Y. Zhang, Y. Fang, Y. Cao, and J. Wu, "Rbgan: Realistic-generation and balanced-utility gan for face de-identification," *Image and Vision Computing*, vol. 141, p. 104868, 2024.

[16] Z. Li, Z. Zhang, S. He, Q. Meng, S. Zhang, B. Zhong, and R. Ji, "Identity-aware variational autoencoder for face swapping," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 5466–5479, 2024.

[17] Y.-L. Pan, J.-C. Chen, and J.-L. Wu, "Secure dual network for reversible facial image anonymization through the latent space manipulation.," *Electronics (2079-9292)*, vol. 13, no. 22, 2024.

[18] D. Chen, M. Zhu, N. Wang, and X. Gao, "Hford: High-fidelity and occlusion-robust de-identification for face privacy protection," *arXiv preprint arXiv:2311.08786*, 2023.

[19] Y. Cao, Y. Zhang, J. Wu, and Y. Fang, "Multi-channel attribute preservation for face de-identification," *Multimedia Tools and Applications*, pp. 1–25, 2024.

[20] Y. Wen, B. Liu, L. Song, J. Cao, and R. Xie, "High quality face de-identification with model explainability," in *Face De-identification: Safeguarding Identities in the Digital Era*, pp. 127–146, Springer, 2024.

[21] X. He, M. Zhu, D. Chen, N. Wang, and X. Gao, "Diff-privacy: Diffusion-based face privacy protection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[22] S. Malm, V. Rönnbäck, A. Håkansson, M.-h. Le, K. Wojtulewicz, and N. Carlsson, "Rad: Realistic anonymization of images using stable diffusion," in *Proceedings of the 23rd Workshop on Privacy in the Electronic Society*, pp. 193–211, 2023.

[23] Y. Zhu, J. Cao, B. Liu, T. Chen, R. Xie, and L. Song, "Identity-consistent video de-identification via diffusion autoencoders," in *2024 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, IEEE, 2024.

[24] W. Khan, L. Topham, U. Khayam, S. Ortega-Martorell, P. Heather, D. Ansell, D. Al-Jumeily, and A. Hussain, "Person de-identification: A comprehensive review of methods, datasets, applications, and ethical aspects along-with new dimensions," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

Authorized licensed use limited to: University Duisburg Essen. Downloaded on May 26,2025 at 11:41:47 UTC from IEEE Xplore. Restrictions apply.