IDDiffuse: Dual-Conditional Diffusion Model for Enhanced Facial Image Anonymization

Muhammad Shaheryar[1 [0000−0003−3992−1387]], Jong Taek Lee[1 [0000−0002−6962−3148]], and Soon Ki Jung[1⋆ [0000−0003−0239−6785]]

School of Computer Science and Engineering, Kyungpook National University, Daegu, Republic of Korea
{shaheryar,jongtaeklee,skjung}@knu.ac.kr

Abstract. The increasing prevalence of computer vision applications in public spaces has raised substantial privacy concerns regarding facial image data. Traditional anonymization methods, despite their potential, often suffer from drawbacks such as limited output variety, inadequate detail, distortions in extreme poses, and inconsistent temporal patterns. This study introduces an identity diffuser based on a dual-conditional diffusion model that efficiently anonymizes facial images while preserving task-relevant features for diverse applications. Our approach ensures a clear separation from the original identity by utilizing synthetic identities and an optimized identity feature space derived from three state-of-the-art models. It maintains consistency across frames for video anonymization. Unlike existing methods, our approach eliminates the need for task-relevant feature extractors, such as those for pose and expression. Instead, it employs a dual-condition diffusion model to integrate both identity and non-identity information, offering improved anonymization without compromising data usefulness. Our technique enables seamless transitions from real to synthetic identities by incorporating a time-step-dependent ID loss, providing controllable identity anonymization. Extensive studies demonstrate that our method achieves superior de-identification rates and consistency compared to state-of-the-art techniques, preserving non-identity features with a 20% improvement in emotion recognition, handling extreme poses with enhanced image quality, output diversity, and temporal consistency. This makes it a valuable tool for privacy-preserving computer vision applications.

Keywords: Face Anonymization · Synthetic Identity · Face Privacy.

1 Introduction

The growing number of cameras has encouraged advancements in computer vision, leading to breakthroughs in autonomous vehicles and automated surveillance systems. Nevertheless, this advancement offers an increase in privacy problems as humans are under continuous surveillance. To address this, stronger

⋆ Corresponding author

data privacy legislation like the General Data Protection Regulation (GDPR) [1] and Personal Information Protection and Electronic Documents Act (PIPEDA) [6] are evolving internationally. These restrictions promote the development of privacy-preserving technology, making face anonymization essential to maintaining personal information in our increasingly surveillance-driven environment. Lamyan *et al*. [22] have recently introduced a framework that aims toward preserving privacy in surveillance systems. This framework addresses significant issues related to the technology and provides useful solutions at the same time. Applications such as emotion recognition and person detection do not require identifying individuals but still need facial features [10]. Researchers are developing anonymization approaches that disguise identities while preserving other features [27], but face quality and the risk of re-identification remain challenges. Anonymization in computer vision is crucial for privacy-preserving applications, aiming to prevent re-identification while maintaining essential features such as image quality and facial expressions. This process involves removing identifiable information that could be recognized by facial recognition systems or individuals. A key challenge in face anonymization is determining an alternative identity that preserves non-identity features like pose and expression, ensuring the utility of the anonymized images for various applications. Various methods have been proposed to address this challenge (more detail in Sec. 2, each offering unique approaches for altering identity features while preserving anonymity and non-identity features [5,12,17,27]. However, these methods often produce limited diversity in outputs, making it easier to reverse-engineer the original identities [29]. Deepfake technologies [4,24] can replace faces while preserving attributes, but do not fully guarantee privacy and utility [39]. Face swapping [8] also compromises the identity provider's confidentiality and utility. Other methods may introduce unnatural distortions, compromising data quality. This can diminish the user experience and reduce the impact of relying on anonymized visual data.

To address these challenges, we provide a novel technique named IDDiffuse, based on a dual-conditional diffusion model, which tries to improve face anonymization by mapping consistent synthetic identities. This approach promises that the anonymized version of any real individual is consistent throughout the video, keeping the integrity of non-identification aspects such as pose, emotion, hairstyle, and gaze. Our solution introduces a time-step-dependent identity loss to control the extent of the anonymization depending on different applications. Through meticulous evaluation, we demonstrate that IDDiffuse outperforms state-of-the-art techniques, achieving higher de-identification rates, preserving non-identity features with a 20% improvement in emotion recognition, and enhancing image quality by 0.04. Our work can be summarized as follows:

- We propose a dual-conditional diffusion model framework based on identity mapping, enabling consistent anonymization by mapping real identities to synthetic identities.
- Our technique contains an identity-dependent loss for controlled anonymization, allowing modification according to needs of various applications.

– We conducted extensive experiments and evaluations to thoroughly assess the performance of our method.

2 Related Work

2.1 Face anonymization

The primary goal of face anonymization is to remove identity information from an image while maintaining its usefulness. Traditional methods like blurring, pixelating, and Gaussian blur can obscure identities but often degrade visual quality, leading to ghosting effects and loss of crucial non-identity features such as gaze, pose, and expression [30, 31]. This limits the content's suitability for detailed analysis and restricts it to basic information sharing.

Generative AI has made significant advancements in face anonymization to achieve higher visual quality and better preservation of non-identity features. CIAGAN uses an identity guidance discriminator and a transposed convolutional neural network to blend selected identities with original images using facial landmarks, ensuring anonymity while retaining pose but struggling with expressions and gaze [27]. DeepPrivacy employs conditional GANs with bounding boxes and key points, effectively anonymizing faces while preserving pose, though it faces challenges with occlusions and non-traditional poses [17]. Another approach maps identities by pairing real images with their nearest synthetic counterparts using a FaRL, ViT-based image encoder, maintaining head, pose, and geometric details [2, 40]. Recent work combines differential privacy with ensemble learning to optimize identity distance while maintaining gaze and emotion recognition, but it doesn't address visual attributes like hairstyles, sunglasses, and backgrounds [5]. Cao *et al*. [7] proposed decoupling a face image into separate attribute and identity vectors, achieving anonymization by rotating the identity vector while preserving other facial characteristics. Current methods often struggle to handle diverse poses, expressions, and other non-identity features crucial for video anonymization tasks. Our approach addresses these challenges effectively by utilizing a proposed identity extractor, which efficiently maps real identities to synthetic ones.

2.2 Synthetic Data Generation

Synthetic datasets, including SynFace, SFace, and DCface, utilize advanced generative models such as DiscoFaceGAN, StyleGAN2-ADA, and Diffusion models to produce large-scale, diverse face datasets [3, 20, 33]. These methods generate synthetic identities to mitigate ethical concerns and address class imbalances present in real datasets. In our paper, we adopt DCface [20] due to its ability to generate a wide variety of styles within synthetic identities, making it an effective tool for mapping synthetic identities to real ones.

2.3 Latent Diffusion Models

Diffusion models are known for generating high-quality images and are particularly effective in tasks like text-to-image synthesis and inpainting [13, 32, 37]. These models excel due to their stable objective function, allowing for realistic generation without fine-tuning, which suggests potential for face anonymization [26]. DiffFace uses a diffusion model with an image inversion module to anonymize faces while balancing privacy and appearance, employing facial recognition data to ensure anonymization quality [12]. Latent Diffusion Models (LDMs) [35] improve computational efficiency by operating in the compact latent space of a pre-trained Variational Autoencoder (VAE). The encoder E and decoder D map images to latent variables, $z_t = E(x)$, at each timestep t, and the model estimates the added noise $\epsilon_\theta(z_t, c, t)$, where c represents conditioning inputs like text. The objective function for noise prediction is:

$$\mathcal{L}_{noise} = \mathbb{E} z_t, \epsilon(z_t), c, t \sim \mathcal{N}(0,1) \left[|\epsilon - \epsilon_\theta(z_t, c, t)|_2^2 \right]. \tag{1}$$

This loss function minimizes the squared difference between actual noise ϵ and predicted noise $\epsilon_\theta(z_t, c, t)$ over time t.

Our research explores the capabilities of conditional diffusion models for face anonymization by manipulating identity and non-identity features, aiming to preserve task-relevant information while ensuring privacy.

3 Methodology

Our proposed face anonymization approach addresses four basic requirements: ensuring complete anonymity by substituting synthetic identities, allowing user control over the extent of identity anonymization, maintaining realism for various applications, and preserving pose and temporal consistency in video sequences to facilitate tracking and action recognition. To achieve these objectives, we introduce a dual-conditional diffusion model, with the mathematical formulation and problem definition for our task detailed in the following section.

3.1 Problem Formulation

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the set of original images, and let $Y = \{y_1, y_2, \ldots, y_m\}$ be the set of original identities corresponding to the images in X. Let $\hat{X} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_p\}$ be the set of synthetic images, and let $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_q\}$ be the set of synthetic identities corresponding to the images in $\hat{X}$. The overall problem formulation can be summarized as follows: Given an original image $x \in X$, we generate an anonymized image x_{anon} through the following steps: (1) Map the original identity $y \in Y$ to a synthetic identity $\hat{y} \in \hat{Y}$ using the identity mapper: $\hat{y} = f(y)$. (2) Extract the texture information $t(x)$ from the original image x. (3) Use the dual condition diffusion model $\mathcal{D}$ to generate the anonymized image x_{anon} that retains the identity of $\hat{y}$ and the texture of x:

$$x_{\text{anon}} = \mathcal{D}(\hat{y}, t(x)). \tag{2}$$

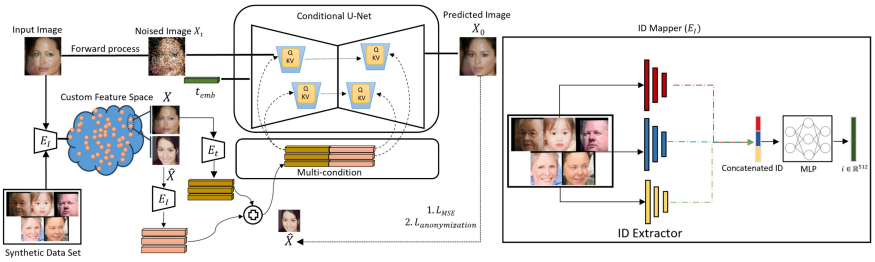


Fig. 1: Dual Conditional face anonymization framework. As a first step, E_I is trained to extend the feature space for better representation of synthetic identities and is used to map and extract identity-related features from real and synthetic images. In second step, texture information is extracted using E_t, an encoder trained with the arcface feature extractor. Both conditions are utilized to train a dual-condition diffusion model, generating realistic and texture-enriched anonymized images.

We propose a method based on dual condition diffusion model IDDiffuse, to diffuse the real identity and generate anonymized images as shown in Fig. 1. The process involves two key steps: First, we seek a mapping $f : Y \rightarrow \hat{Y}$ such that each original identity y_i is mapped to a corresponding synthetic identity $\hat{y}_j$. Given an original face image $x \in X$, our method finds a corresponding synthetic face image $\hat{x} \in \hat{X}$ using the identity mapper E_I. The goal is to map the real identity as close to synthetic identity, while preserving other non-identity features. This step is detailed in Sec. 3.2. Second, we use a texture extractor E_t to capture non-identity information from the original image x, and then train a dual condition diffusion model (DDPM) using these two extracted features. The diffusion model $\mathcal{D}$ combines the identity from the synthetic image and the texture from the real image to produce an anonymized image x_{anon}.

3.2 Identity Mapping

The first step in our anonymization process involves mapping the original identity to a synthetic one. Initially, the approach was to randomly select an identity from a synthetic dataset and condition the diffusion model to generate x_{anon}. However, this method led to inconsistent outputs and loss of non-identity features because the feature space of identity extractors encompasses both identity and non-identity-related details [5, 20, 23]. Moreover, different identity extractors are trained with various loss functions and objectives, resulting in unique biases. For instance, ArcFace [9], while finding similar identities in its feature space, tends to focus more on attributes such as hairstyle, facial expression, or lighting conditions rather than strictly the identity. This can lead to inconsistent identity mapping, where the multiple images of same real identity may be paired with different synthetic identities. Consistency is paramount in video anonymization, where changing poses and expressions of a person should map to

the same synthetic identity across frames. Relying solely on ArcFace for identity extraction led to inconsistent synthetic identities for the same person, undermining anonymization, particularly in tasks like tracking and action recognition. We discuss the limitations of using a single identity extractor in Sec. 5.

This paper explores a novel approach to address these issues and achieve optimal mapping. We propose a multi-model feature extraction technique to achieve a more robust and unbiased identity mapping process. To enhance our anonymization process, we utilize an ensemble of facial recognition models: ArcFace [9], FaRL [40], and FaceNet [36]. By combining outputs from these diverse feature extractors, we mitigate biases inherent in individual models, leading to a fair and more unbiased identity extraction process. Specifically, we extract 512-dimensional feature vectors for each image using these models and then train a Multilayer Perceptron (MLP) with triplet contrastive learning loss to achieve a richer and more consistent feature space for identity mapping. For an input image x, let $\mathbf{v}_{\text{ArcFace}}(x)$, $\mathbf{v}_{\text{FaRL}}(x)$, and $\mathbf{v}_{\text{FaceNet}}(x)$ be the 512-dimensional feature vectors extracted from the ArcFace, FaRL, and FaceNet models respectively. We concatenate these vectors to form a combined feature representation:

$$\mathbf{v}(x) = \mathbf{v}_{\text{ArcFace}}(x) \oplus \mathbf{v}_{\text{FaRL}}(x) \oplus \mathbf{v}_{\text{FaceNet}}(x), \tag{3}$$

where $\oplus$ denotes concatenation. The combined feature vector $\mathbf{v}(x)$ is then fed into a Multilayer Perceptron (MLP) with parameters θ. The MLP is trained using a triplet contrastive learning loss to refine this feature space for consistent identity mapping. The triplet contrastive learning loss $\mathcal{L}_{\text{triplet}}$, is defined as follows. Given anchor, positive, and negative samples (x_a, x_p, x_n), the loss is:

$$\mathcal{L}_{\text{triplet}} = \max\left(\|\mathbf{z}_a - \mathbf{z}_p\|_2^2 - \|\mathbf{z}_a - \mathbf{z}_n\|_2^2 + \alpha, 0\right), \tag{4}$$

where $\mathbf{z}_a = \text{MLP}_\theta(\mathbf{v}(x_a))$, $\mathbf{z}_p = \text{MLP}_\theta(\mathbf{v}(x_p))$, $\mathbf{z}_n = \text{MLP}_\theta(\mathbf{v}(x_n))$, $\|\cdot\|_2$ denotes the Euclidean distance, and α is a margin parameter. By minimizing this loss, the MLP generates a feature space that is both discriminative and consistent for identity mapping. It captures identity-specific and non-identity features, ensuring that all instances of the same identity consistently map to the same anonymized identity. This approach enhances the robustness of anonymization while maintaining key features for downstream tasks.

3.3 Patch-wise Texture Extraction

For each pair, the identity features are extracted from the synthetic image using our proposed identity feature extractor E_I in Sec. 3.2, while texture features are derived from the real image using E_t. We choose ArcFace as a backbone in E_t, for its superior ability to capture non-identity features. Texture features are extracted in a patch-wise manner as shown in Fig. 2 to focus more on texture details rather than identity, inspired by the success of DCFace [20]. Specifically, with a pre-trained ArcFace F_t. The intermediate feature representation

$$F_t(X) = I_{tex} \in \mathbb{R}^{C \times H \times W} \tag{5}$$

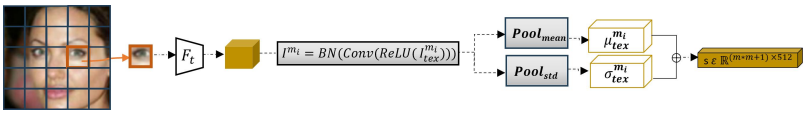


Fig. 2: Patch-wise texture extraction E_t inspired by [20]. Utilizing 5x5 patch size, the fixed feature encoder from ArcFace and the patch-wise spatial mean-variance operation effectively eliminate detailed identity information while preserving the texture of the image.

is divided into a $m \times m$ grid. For each grid element $I^{m_i}_{tex} \in \mathbb{R}^{C \times \frac{H}{m} \times \frac{W}{m}}$, a series of operations—Batch Normalization (BN), Convolution, and ReLU—are applied, followed by pooling to compute the mean and variance vectors. These vectors, both with dimensions $\mathbb{R}^C$ (where C is the number of channels), are calculated as:

$$\mu^{m_i}_{tex} = \text{mean}(I^{m_i}) and \quad \sigma^{m_i}_{tex} = \text{std}(I^{m_i}). \tag{6}$$

The texture vector is formed by concatenating texture vector from each patch. As discussed by authors in [20], a crucial property of E_t is to limit the flow of information to the next layer. The fixed feature encoder F_t and the patch-wise spatial mean-variance operation eliminate detailed identity information while preserving the image's texture.

3.4 Dual Conditional DDPM

To train the Dual Condition DDPM, we used the original DDPM objective, as described in Eq. (1), but it is insufficient for ensuring consistency between the synthetic id condition $\hat{X}$ and the predicted image X_0. To address this, we propose a time-dependent anonymization loss function to maximize the similarity between $\hat{X}$ and X_0 in the feature space of a proposed E_I. This proposed anonymization loss function is used in conjunction with the mean squared error (MSE) loss in Eq. (1) to ensure both anonymization consistency and high-quality image generation. The loss function for this training process is designed to combine identity interpolation with texture consistency:

$$\begin{aligned} L_{anonymization} = \gamma_t CS(F(X), F(X_0)) + (1-\gamma_t) CS(F(\hat{X}), F(X_0)) \\ + \lambda CS(F(X), F(X_0)), \end{aligned} \tag{7}$$

where λ is a hyper-parameter that controls the strength of style consistency across the identity interpolation. The term γ_t is a time-dependent weight that linearly changes from 0 to 1 as training progresses. X and $\hat{X}$ are the input images with different identities, so the interpolation allows the prediction to retain the same texture as the real identity Y while gradually shifting towards the synthetic identity $\hat{Y}$. The variable X_0 represents the reconstructed or predicted image whose features are compared to $\hat{X}$ for identity and to X for texture. The first term $(\gamma_t CS(F(X), F(X_0)))$ ensures that at the beginning

of the interpolation (when t is close to T), the identity is closer to real image identity Y. The second term $((1-\gamma_t)CS(F(\hat{X}), F(X_0)))$ allows the identity to transition towards synthetic identity $\hat{Y}$ as t decreases. The third term $(\lambda CS(F(X), F(X_0)))$ is a regularization term that ensures the texture stays consistent with X. By integrating these components, our methodology effectively anonymizes the real image X while preserving important texture information, ensuring both identity anonymization and texture consistency. Features extracted using $E_{\mathrm{I}}\hat{X}$ and $E_{\mathrm{t}}(X)$, can be injected into the U-Net ϵ_θ by following the convention of the DDPM-based text-conditional image generators [34]. Specifically, the cross-attention operation can be written as a modification of the attention equation [38] with an additional key K_c and value V_c from the identity (E_I) and texture (E_t) extractors are concatenated with the original K and V.

$$\text{Cross-Attn}(Q, K, V, K_c, V_c) = \text{SoftMax}\left(\frac{QW_q([K, K_c]W_k)^\top}{\sqrt{d}}\right) W_v[V, V_c]. \quad (8)$$

4 Experiments

In this section, we evaluate the performance of our anonymization framework against SOTA methods using privacy-related metrics, and pose and gaze estimation metrics. We also demonstrate the impact of video anonymization, with a focus on maintaining temporal consistency.

Datasets: We leverage the DCFace synthetic labeled dataset [20], consisting of 0.5 million images across 10,000 synthetic identities, with 50 images per identity. It was generated using an unconditional DDPM trained on FFHQ [18] for identity and CASIA-WebFace [14] for style, this dataset provides high consistency and diversity, making it ideal for synthetic identity mapping in face anonymization. Extensive experiments comparing DCFace-generated images with CASIA-WebFace show a False Match Rate (FMR) of only 0.0026% at a threshold of 0.3 [20], indicating that synthetic identities are effectively distinct from real ones, minimizing the risk of identity leakage. For training our dual-condition diffusion model, we use the CASIA-WebFace dataset [14] with 494,414 images of 10,575 real identities. To evaluate utility and anonymization performance, we employed the LFW dataset [15], containing 13,233 images and 5,749 identities.

State of the art: We compare our anonymization method with leading anonymization methods, specifically CIAGAN [27], DeepPrivacy [17], Disguise [5], Password [11], and Repaint [26].

Implementation Details: We designed an identity mapper using a basic MLP architecture consisting of three layers with sizes [1536, 2048, 1024, 512]. Features from ArcFace [9], FaRL [40], and FaceNet [36] are concatenated and input into the MLP to train for a more enriched and robust feature space. This model was trained for 50 epochs on synthetic data comprising 10,000 identities, employing triplet contrastive loss. The training process, conducted on an RTX 4090 GPU, spanned over 60 hours. Our dual-condition diffusion model is based on the architecture proposed in DCFAce [20]. We trained this model on the CASIA-WebFace

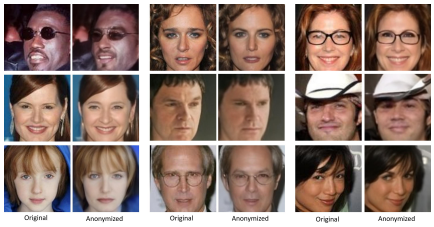


Fig. 3: Anonymization results across a range of poses and expressions, including variations with glasses and different facial orientations. Columns (a) display the original images, while columns (b) show the anonymized versions.

Table 1: Evaluation comparison of identification/validation rates and image quality on anonymized version of LFW dataset.

Methods	TPR (%) @ FPR=10^{-3} / Accuracy (%) ↓			FIQ ↑
	SphereFace	AdaFace	Average	
Original	(87.9, 96.2)	(95.4, 97.7)	(92.4, 97.0)	0.77
DeepPrivacy [17]	(2.9, 70.9)	(4.6, 68.6)	(4.9, 71.1)	0.67
DeepPrivacy2 [16]	(1.0, 61.5)	(2.2, 62.2)	(1.6, 62.1)	0.58
CIAGAN [27]	(1.0, 59.0)	(5.6, 71.0)	(2.8, 64.8)	0.65
Password [11]	(17.1, 73.5)	(51.0, 84.0)	(33.3, 78.9)	0.69
RePaint [26]	(1.1, 63.5)	(6.7, 68.5)	(2.5, 66.0)	0.64
Disguise [5]	(0.03, 50.0)	(0.0, 50.0)	(0.02, 50.0)	0.90
Ours	**(0.01, 50.0)**	**(0.00, 50.0)**	**(0.01, 50.0)**	**0.94**

dataset [14] using dual condition after generating synthetic identities through the identity mapper. The diffusion model underwent training for 20 epochs with a batch size of 128, using the AdamW optimizer at a learning rate of 0.001. This training was executed over 20 hours using four NVIDIA TITAN Xp GPUs.

Metrics: We evaluated our approach using commonly used metrics for both privacy preservation and texture preservation [5, 21, 25, 28].

Validation Rate (TPR@FPR=1e-3): Measures the success rate of anonymization at a very low false positive rate (FPR) of 1 in 1000. This indicates how well the system prevents identification while minimizing incorrectly classifying matching faces.

Verification Accuracy: Assesses the system ability to distinguish between anonymized and original faces. A random guess would achieve 50% accuracy, so achieving significantly lower accuracy indicates effective anonymization. We employed the AdaFace [19] and SphereFace [25] algorithm for distance metric learning due to its effectiveness and to ensure unbiased results. Since we never used them during training, it provided a neutral measure of anonymization strength. [1]

Facial Landmark Detection: Measured using the ℓ_2 pixel distance and Normalized Mean Error (NME) to quantify the accuracy of locating key facial points.

[1] For fair analysis and due to the unavailability of released code, we borrowed qualitative and quantitative results from [5] for comparison in our work. This ensures fair and consistent analysis, as altering the input images could impact the results.

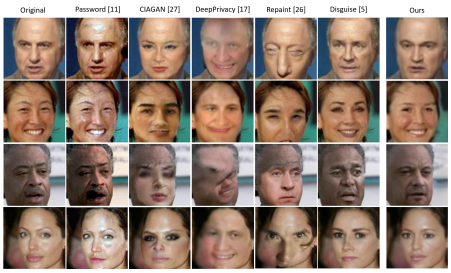


Fig. 4: Qualitative comparison between our method and current anonymization techniques, with the original faces displayed in the first column.

Gaze Estimation: Evaluated with Mean Absolute Error (MAE) to assess the precision of gaze direction prediction.
Emotion Recognition: Accuracy was used to gauge the system's ability to correctly identify emotions.
Image Quality: Preserved image quality was measured using SER-FIQ, a metric specifically designed for evaluating image anonymization impact.
Anonymization: We evaluated the methods on the LFW dataset, a standard benchmark by anonymizing the second image in each pair within the dataset and comparing the performance of various techniques. Our method demonstrates superior performance in both anonymization effectiveness shown in Fig. 3 and image quality compared to previous state-of-the-art approaches presented in Table 1. Specifically, our method achieves the lowest verification accuracy (0.01%) across both SphereFace and AdaFace, indicating a high level of anonymization, which significantly surpasses other methods. Additionally, our method maintains a high Face Image Quality (FIQ) score of 0.94, the highest among all evaluated techniques. This dual success in protecting privacy by making faces unrecognizable while preserving high image quality underscores the robustness and effectiveness of our approach. Fig. 4 visually compares the anonymized images generated by our method with those from other leading techniques. It is evident that both our method and Disguise [5] produce high-quality images, while the images from other methods appear less realistic. Importantly, both our method and Disguise [5] achieve strong anonymization by producing diverse, unrecognizable faces. However, our method surpasses Disguise in terms of FIQ score, demonstrating superior image quality. This advantage is attributed to the enhanced capabilities of our diffusion model, which ensures better image generation. Our approach not only sets a new benchmark in anonymization but also ensures that the anonymized images retain essential visual clarity. This balance between effective anonymization and maintaining image quality make our method highly valuable for applications where both privacy protection and visual integrity are crucial.
Non-Identity Attribute Preservation: Our method outperforms existing approaches in non-identity attribute preservation, excelling in both gaze estimation

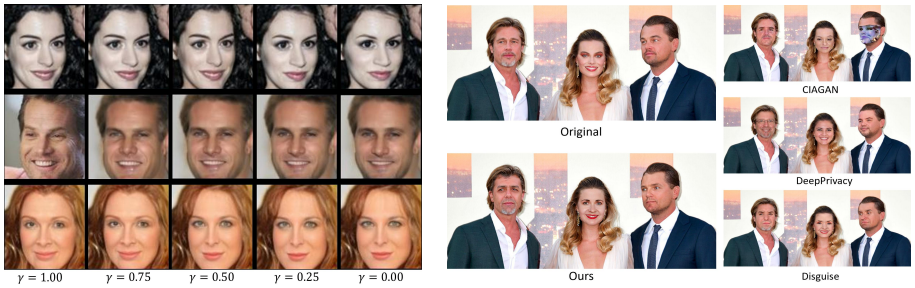
Table 2: Performance Comparison: Emotion (Accuracy % ↑), Gaze Estimation (MAE ↓), and Facial Landmarks (ℓ_2 pixel distance ↓)

Methods	Emotion	ETH-XGaze		RetinaFace			Dlib		
	DF	Pitch	Yaw	Eyes	Nose	Mouth	Eyes	Nose	Mouth
DeepPrivacy [17]	34.3	7.7	13.6	13.1	9.9	16.5	32.7	25.1	89.0
DeepPrivacy2 [16]	30.2	9.2	12.2	18.4	14.4	19.6	59.9	49.9	120.6
CIAGAN [27]	36.9	8.8	14.6	9.3	5.5	9.2	59.0	31.2	97.7
Password [11]	43.4	10.5	24.7	10.4	7.7	11.1	26.5	19.3	55.4
RePaint [26]	19.4	11.3	18.1	30.8	32.2	47.3	133.5	152.1	432.0
Disguise [5]	47.0	6.8	8.4	7.7	5.6	**8.2**	28.8	19.8	60.0
Ours	**67.2**	**5.6**	**7.5**	**4.9**	**3.7**	8.8	**15.7**	**13.6**	**55.3**

and emotion recognition, as shown in Table 2. Using the same settings as [5], and employing DeepFace for emotion classification and L2CS-Net for gaze estimation, our method achieves the highest emotion recognition accuracy of 67.2%, retaining essential facial expressions, and the lowest Mean Absolute Error (MAE) values of 3.6 for pitch and 4.9 for yaw, confirming precise gaze estimation. Furthermore, it achieves state-of-the-art ℓ_2 distances for facial landmarks using RetinaFace (5 points) and Dlib (68 points), significantly outperforming other methods in maintaining facial structure. Main key to this superior performance is our identity mapper and texture extractor, which allow for more accurate retention of critical attributes while anonymizing the identity. Competing methods like DeepPrivacy, CIAGAN, and Password fail to preserve facial features without introducing artifacts or blurring. While RePaint performs well on in-distribution faces, it struggles with out-of-distribution cases and occlusion, unlike our dual-condition model, which excels in these scenarios. This comprehensive evaluation underscores our method's superiority in balancing anonymization with utility preservation, setting a new benchmark in the field.

Controllable Identity Anonymization: In Fig. 5a, we demonstrates the process of identity anonymization using varying levels of γ. We visualize the plot of interpolation in X, demonstrating the smooth transition achieved through the interpolation process. We use the Eq. (7) with γ increasing linearly from 0 to 1. This results in a gradual transformation, smoothly altering pose and expression and changing the identity. The leftmost image in column (a) corresponds to $\gamma = 1$, and the subsequent images at (b), (c), (d) and (e) with γ values of 0.75, 0.5, 0.25, and 0 show progressively more anonymized identity. This approach allows us to control the level of anonymization based on the specific requirements of different applications. In Fig. 5b, we provide an intuitive comparison of various face anonymization techniques applied to images with multiple faces. For a fair analysis, we have borrowed the results from the Disguise [5] and performed experiments of our method on original image for better comparison. Our method demonstrates superior performance, effectively anonymizing all faces in the image while maintaining a natural and coherent appearance, as compared to previous methods.

Temporal Consistency: In Fig. 6a, we showcase the qualitative results that highlight our method's superior ability to maintain temporal consistency across video frames. Our approach excels in consistently projecting the same identity



(a) Controlling Identity Anonymization

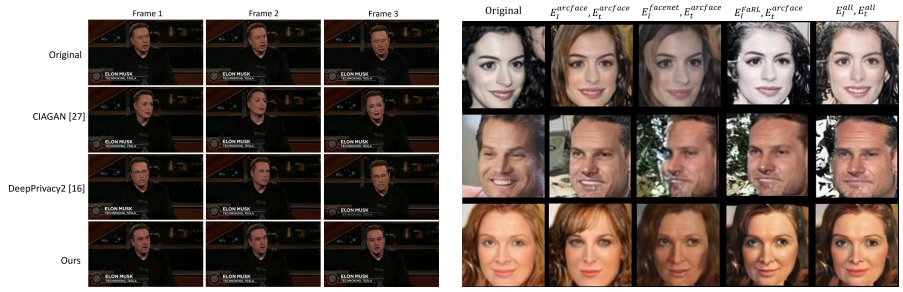
(b) Face anonymization on images with multiple faces

Fig. 5: Qualitative evaluation of our method, showcasing (a) control over identity anonymization and (b) application on images with multiple faces.

throughout sequential frames, ensuring a coherent identity representation in the video. When compared to Disguise and DeepPrivacy, our method clearly outperforms in preserving identity and other texture features. By enhancing temporal consistency and ensuring superior face anonymization, our method effectively preserves essential identity features and textures. This results in higher-quality anonymized videos, optimally balancing identity protection and visual quality, and setting a new standard for anonymization techniques in video sequences.

5 Ablation Experiments

For the ablation study, we used different configurations of E_{I} and E_{t} to determine which feature extractor performs best and why. Specifically, the notations for method configurations are as follows: $E_{\mathrm{I}}^{\mathrm{arcface}}, E_{\mathrm{t}}^{\mathrm{arcface}}$ indicate that ArcFace is used as the feature extractor for both identity and texture, while $E_{\mathrm{I}}^{\mathrm{all}}, E_{\mathrm{t}}^{\mathrm{all}}$ represent the use of all three feature extractors (ArcFace, FaceNet, and FaRL) in both the identity mapper and texture extractor. The ablation study examines these configurations across different activites as shown in Tab. 3. The configuration $E_{\mathrm{I}}^{\mathrm{all}}, E_{\mathrm{t}}^{\mathrm{arcface}}$, excels in most of the tasks. In Fig. 6, we provide visual comparisons of these configurations. The first column displays the original images, which serve as the reference for evaluating the performance of the other configurations. The second column shows results using $E_{\mathrm{I}}^{\mathrm{arcface}}, E_{\mathrm{t}}^{\mathrm{arcface}}$. Although this configuration maintains some level of identity preservation, it lacks consistency in facial expressions and pose, introducing more non-identity features and compromising the natural aspect of the faces. The third column gives the results of utilizing FaceNet as the feature extractor. While it anonymizes the identity to some extent, FaceNet suffers with non-attribute facets such as gaze and extreme poses, as seen in the second and third image of the said column. This implies that depending simply on FaceNet is insufficient to achieve efficient anonymity and feature preservation. The fourth column employs solely $E_{\mathrm{I}}^{\mathrm{FaRL}}$, which results in considerable variations from the original identity, with notable changes



(a) Video anonymization comparison on consecutive frames.

(b) Visualization of different configurations of E_I and E_t of our method.

Fig. 6: Comparison of different methods and techniques for face anonymization.

in expressions and poses. This configuration demonstrates the limitations of employing solely FaRL, which results to less effective anonymization and feature preservation, yet introducing some degree of natural appearance. Finally, the fifth column displays the results of utilizing $E_\mathrm{I}^\mathrm{all}, E_\mathrm{t}^\mathrm{all}$, which utilizes the combined strengths of ArcFace, FaceNet, and FaRL. This setting gives the best results, with images that closely match the originals in terms of pose, expression and identity. However, it results in additional texture aspects not present in the original image, showing that while this approach delivers a comprehensive representation, it is most useful for identification extraction. ArcFace, as proven in the second column, is particularly useful in capturing expressions and texture features.

Table 3: Ablation study results comparing different configurations of E_I and E_t for face anonymization.

Methods	Emotion	ETH-XGaze		RetinaFace			Dlib		
	DF	Pitch	Yaw	Eyes	Nose	Mouth	Eyes	Nose	Mouth
$E_\mathrm{I}^\mathrm{arcface}, E_\mathrm{t}^\mathrm{arcface}$	63.8	6.2	8.6	9.2	6.8	11.9	23.5	16.8	62.0
$E_\mathrm{I}^\mathrm{facenet}, E_\mathrm{t}^\mathrm{arcface}$	49.8	6.5	10.1	13.7	8.5	15.2	20.8	24.0	63.2
$E_\mathrm{I}^\mathrm{FaRL}, E_\mathrm{t}^\mathrm{arcface}$	51.8	7.0	8.5	7.0	6.0	10.0	19.7	21.4	65.0
$E_\mathrm{I}^\mathrm{all}, E_\mathrm{t}^\mathrm{all}$	47.9	6.2	8.1	5.5	4.5	12.5	27.5	22.0	65.8
$E_\mathrm{I}^\mathrm{all}, E_\mathrm{t}^\mathrm{arcface}$	**67.2**	**5.6**	**7.5**	**4.9**	**3.7**	**8.8**	**15.7**	**13.6**	**55.3**

6 Conclusion

In this research, we develop IDDiffuse, a unique dual-condition diffusion model for face anonymization. Our method leverages synthetic identities to map the original identity inside an enhanced feature space and then adopts the dual-condition diffusion model for effective face anonymization. Through extensive

Fig. 7: Our face anonymization resutls under extreme poses.

Fig. 8: MultiFace anonymization in public space setting.

experiments, IDDiffuse achieves state-of-the-art quantitative and qualitative results in facial anonymization, other feature preservation and temporal consistency. Qualitative results indicate that several approaches can easily achieve good anonymization performance. However, retaining additional features and ensuring consistent mapping to synthetic identities was difficult, which we addressed. Future work will focus on improving anonymization by studying non-identity features in identity space.

Societal Impact and Limitations

Our approach to face anonymization offers significant privacy benefits even in extreme poses displayed in Fig. 7, but it has limitations in extreme overlapping situation. In such cases, identity features can interfere with pose and expression retention, making it difficult to anonymize effectively without losing crucial non-identity information as visualized in Fig. 8. Future research should focus on improving the preservation of pose and expression in situations where identity features overlap with these attributes. Additionally, while our method reduces privacy risks, the potential misuse for deepfake generation remains a societal concern, we support community efforts to develop DeepFake detection methods and emphasize ethical use, ensuring technology is applied responsibly.

Acknowledgment

This work was supported by the IITP grant (IITP-2024-RS-2022-00156389) funded by MSIT, and the Digital Innovation Hub project (DBSD1-04) supervised by DIP, funded by MSIT and Daegu City, 2024. ※ MSIT: Ministry of Science and ICT

References

1. Legal text. general data protection regulation (gdpr). (2024, april 22). https://gdpr-info.eu/ 2
2. Barattin, S., Tzelepis, C., Patras, I., Sebe, N.: Attribute-preserving face dataset anonymization via latent code optimization pp. 8001–8010 3
3. Boutros, F., Huber, M., Siebke, P., Rieber, T., Damer, N.: Sface: Privacy-friendly and accurate face recognition using synthetic data. In: 2022 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–11. IEEE (2022) 3
4. Bu, J., Jiang, R.L., Zheng, B.: Research on deepfake technology and its application. In: International Conference on Computing, Networks and Internet of Things (CNIOT). pp. 47–51 (2023) 2
5. Cai, Z., Gao, Z., Planche, B., Zheng, M., Chen, T., Asif, M.S., Wu, Z.: Disguise without disruption: Utility-preserving face de-identification. Conference on Artificial Intelligence (AAAI) p. 918–926 (2024) 2, 3, 5, 8, 9, 10, 11
6. Canada, G.: The personal information protection and electronic documents act (2021), https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/ 2
7. Cao, J., Liu, B., Wen, Y., Xie, R., Song, L.: Personalized and invertible face de-identification by disentangled identity information manipulation pp. 3334–3342 3
8. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: International Conference on Multimedia (ACMMM). pp. 2003–2011 (2020) 2
9. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019) 5, 6, 8
10. Gafni, O., Wolf, L., Taigman, Y.: Live face de-identification in video. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9378–9387 (2019) 2
11. Gu, X., Luo, W., Ryoo, M.S., Lee, Y.J.: Password-Conditioned Anonymization and Deanonymization with Face Identity Transformers, p. 727–743 (2020) 8, 9, 11
12. He, X., Zhu, M., Chen, D., Wang, N., Gao, X.: Diff-privacy: Diffusion-based face privacy protection (2023) 2, 4
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems (NIPS), volume=33, pages=6840–6851, year=2020 4
14. Huang, G., Mattar, M., Lee, H., Learned-Miller, E.: Learning to align from scratch. Advances in Neural Information Processing Systems (NIPS), volume=25, year=2012 8, 9
15. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008) 8
16. Hukkelås, H., Lindseth, F.: Deepprivacy2: Towards realistic full-body anonymization pp. 1329–1338 (2020) 9, 11
17. Hukkelås, H., Mester, R., Lindseth, F.: DeepPrivacy: A Generative Adversarial Network for Face Anonymization, p. 565–578. Springer International Publishing (2019) 2, 3, 8, 9, 11
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4401–4410 (2019) 8

19. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18750–18759 (2022) 9
20. Kim, M., Liu, F., Jain, A., Liu, X.: Dcface: Synthetic face generation with dual condition diffusion model. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12715–12725 (2023) 3, 5, 6, 7, 8
21. Kuang, Z., Yang, X., Shen, Y., Hu, C., Yu, J.: Facial identity anonymization via intrinsic and extrinsic attention distraction. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12406–12415 (2024) 9
22. Laishram, L., Shaheryar, M., Lee, J.T., Jung, S.K.: Toward a privacy-preserving face recognition system: A survey of leakages and solutions. ACM Computing Surveys (2024) 2
23. Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., Cao, X.: Identity-preserving face anonymization via adaptively facial attributes obfuscation. In: International Conference on Multimedia (ACMMM). pp. 3891–3899 (2021) 5
24. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5074–5083 (2020) 2
25. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 212–220 (2017) 9
26. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11461–11471 (2022) 4, 8, 9, 11
27. Maximov, M., Elezi, I., Leal-Taixé, L.: Ciagan: Conditional identity anonymization generative adversarial networks. pp. 5447–5456 (2020) 2, 3, 8, 9, 11
28. Maximov, M., Elezi, I., Leal-Taixé, L.: Decoupling identity and visual quality for image and video anonymization. In: Asian Conference on Computer Vision (ACCV). pp. 3637–3653 (2022) 9
29. McPherson, R., Shokri, R., Shmatikov, V.: Defeating image obfuscation with deep learning. arXiv preprint arXiv:1609.00408 (2016) 2
30. Neustaedter, C., Greenberg, S., Boyle, M.: Blur filtration fails to preserve privacy for home-based video conferencing. ACM Transactions on Computer-Human Interaction **13**(1), 1–36 (2006) 3
31. Newton, E.M., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE transactions on Knowledge and Data Engineering **17**(2), 232–243 (2005) 3
32. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. pp. 8162–8171. PMLR (2021) 4
33. Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., Tao, D.: Synface: Face recognition with synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10880–10890 (2021) 3
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 8
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. pp. 10684–10695 (2022) 4
36. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015) 6, 8

37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 4
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 8
39. Westerlund, M.: The emergence of deepfake technology: A review. Technology innovation management review **9**(11) (2019) 2
40. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18697–18709 (2022) 3, 6, 8