Enhancing User-Centric Privacy Protection: An Interactive Framework through Diffusion Models and Machine Unlearning

Huaxi Huang[1*], Xin Yuan[1], Qiyu Liao[1], Dadong Wang[1], Tongliang Liu[2]
[1]Data61, CSIRO, Australia
[2]The University of Sydney, Sydney NSW 2006, Australia

Abstract

In the realm of multimedia data analysis, the extensive use of image datasets has escalated concerns over privacy protection within such data. Current research predominantly focuses on privacy protection either in data sharing or upon the release of trained machine learning models. Our study pioneers a comprehensive privacy protection framework that safeguards image data privacy concurrently during data sharing and model publication. We propose an interactive image privacy protection framework that utilizes generative machine learning models to modify image information at the attribute level and employs machine unlearning algorithms for the privacy preservation of model parameters. This user-interactive framework allows for adjustments in privacy protection intensity based on user feedback on generated images, striking a balance between maximal privacy safeguarding and maintaining model performance. Within this framework, we instantiate two modules: a differential privacy diffusion model for protecting attribute information in images and a feature unlearning algorithm for efficient updates of the trained model on the revised image dataset. Our approach demonstrated superiority over existing methods on facial datasets across various attribute classifications.

Introduction

Images inherently contain a wealth of private information, such as gender, race, and age in facial datasets to license plate numbers and vehicle types in car photographs. The extensive use of such images across social networks, government databases, and industrial applications has significantly raised privacy risks, leading to profound public concerns (Times 2018; Liu et al. 2021a). This growing reliance on image data brings privacy issues to the forefront of global discussions and legislation. Countries around the world have responded by implementing stringent privacy laws, with the European General Data Protection Regulation (GDPR) (European Parliament and Council of the European Union) and the Australian Privacy Act 1988 (No 1988) being prime examples. These laws emphasize the safeguarding of "personal data," which includes any information linked to a specific or identifiable individual. Given this definition, images fall under the category of personal data due to their frequent inclusion of personal sensitive attributes such as faces, textual content, and license plates. This situation highlights the urgent need for effective methods to protect image data privacy.

Contemporary multimedia research has focused on dataset publication and model design, with a key challenge being privacy protection. Traditional methods like pixel-level image obfuscation (Fan 2018, 2019) are ineffective against advanced deep learning techniques, while recent approaches using generative adversarial networks (GANs) (Liu et al. 2021b; Wen et al. 2022; Li and Clifton) suffer from instability and poor image quality. Moreover, despite progress in preventing unauthorized access to sensitive model information, a unified solution addressing privacy in both dataset sharing and model deployment is lacking. This challenge is particularly relevant for companies providing facial image applications, such as facial recognition and emotion detection services, which require comprehensive privacy solutions to manage user requests for the removal of private data, requiring adjustments to both datasets and models.

Recognizing these theoretical and practical challenges, this paper proposes an integrated framework that bridges the gap by effectively managing privacy concerns in image data sharing and secure machine learning model release. Developed from published datasets, these models are designed to prioritize privacy protection and can efficiently unlearn specific private information upon user requests. This approach ensures a balance between privacy safeguarding and the utility and performance of the data and models. Fig. 1 provides a schematic overview of this framework, illustrating the interactive, user-centric process from customer input to model refinement, underscoring our strategy for dynamic privacy management in both data and models.

Specifically, our framework addresses the privacy preservation challenge in image sharing by introducing differential privacy (DP). To overcome the instability in GAN training and issues with image quality, we have drawn inspiration from diffusion models (Ho, Jain, and Abbeel 2020; Preechakul et al. 2022) and developed a novel approach to differential image privacy. This method employs diffusion models for both the extraction of intermediate features and the generation of new images. The fundamental advan-

*Project lead. Huaxi Huang, Xin Yuan and Qiyu Liao are Co-first authors. Corresponding Author: Dadong Wang. This work was done when Huaxi was a CREC Research Fellow at Data61.

tage of diffusion models lies in their process, which primarily involves adding noise to images and optimizing the subsequent denoising steps, circumventing the adversarial training intrinsic to GAN models and leading to more stable model training. Existing research (Ho, Jain, and Abbeel 2020; Preechakul et al. 2022) indicates that diffusion models can achieve superior image quality compared to GANs. We design the Diffusion Differential Privacy model (Diffusion-DP). By utilizing a diffusion autoencoder model (Preechakul et al. 2022), we train an auxiliary classifier to disentangle intermediate features, allowing precise control over target face images by injecting DP noises. This enhances the privacy protection of the shared images while maintaining their utility.

To address the challenge of updating models in response to dataset changes, especially in erasing privacy information, our framework incorporates an advanced machine unlearning (AMU) module. This module is designed for scenarios requiring the removal of specific private information from trained models due to dataset updates. Unlike traditional methods of retraining models from scratch, characterized by inefficiency and high costs, our method utilizes cutting-edge machine unlearning techniques (Bourtoule et al. 2021; Warnecke et al. 2023). We have adapted the machine feature unlearning algorithm (Warnecke et al. 2023) specifically for our task, allowing for rapid fine-tuning of models with minimal data input. This enables the models to efficiently unlearn specific information, significantly reducing the resources and time required while ensuring high performance. Integrated with the Diffusion-DP module and the AMU module, our framework offers an efficient solution for adapting models to dataset changes, rigorously safeguarding user privacy, and maintaining the utility and effectiveness of the models.

In summary, this paper makes the following contributions:

- We propose a user-centric interactive image privacy protection framework designed to safeguard user privacy concurrently during image data sharing and model release phases. This framework is interactive, enabling user-driven adjustments for personalized privacy considerations.
- We instantiate two specific modules within our broader framework to separately address the processing of private information in data sharing and model updates. The Diffusion-DP module focuses on editing and generating image data with an emphasis on privacy preservation, while the AMU module enables efficient and rapid model adjustments in response to dataset updates, ensuring privacy compliance.
- Comprehensive experiments are conducted to evaluate the effectiveness of our proposed framework. The results demonstrate that our approach not only achieves significant improvements in privacy protection but also maintains high utility and performance.

The rest of this paper is organized as follows. Section introduces related works. Section presents the proposed framework. In Section 6, we evaluate the proposed method on widely-used facial datasets. The conclusion is discussed in Section 6.

Related Work

Privacy Protection Methods for Image Data

Traditional privacy-preserving methods like pixel-level image obfuscation are often ineffective against advanced deep learning techniques (Fan 2018, 2019), as they can significantly degrade the utility of the image and fail to protect against sophisticated attacks (McPherson, Shokri, and Shmatikov 2016). Recent research has proposed using deep GANs for image privacy protection, where DP noise is infused into intermediate features extracted from the encoder, which are then used by the GAN's image generator to produce new images (Liu et al. 2021b; Wen et al. 2022). While this approach offers a more robust solution, it faces challenges such as GAN training instability, susceptibility to mode collapse, and the need for further improvement in the quality of generated images (Lucic et al. 2018).

This challenge extends to contemporary multimedia research, which focuses on both dataset publication (Liu et al. 2015; Lee et al. 2020; Deng et al. 2009; Karras, Laine, and Aila 2019) and specific model design (He et al. 2016; Ren et al. 2015; Ronneberger, Fischer, and Brox 2015). Privacy protection remains a critical concern, particularly in the publication and sharing of datasets. Although substantial progress has been made in developing privacy-preserving techniques for both datasets and models, a unified approach that addresses privacy concerns in both domains is still lacking. This gap is particularly evident in practical scenarios faced by companies specializing in facial image applications (FaceApp.com 2020; Corporation 2023), where there is a growing demand for solutions that safeguard privacy while maintaining the functionality of AI services based on facial data, like facial recognition (Deng et al. 2019; Schroff, Kalenichenko, and Philbin 2015), age estimation (Pan et al. 2018; Gao et al. 2018; Li et al. 2022), gender identification (Eidinger, Enbar, and Hassner 2014; Kuprashevich and Tolstykh 2023), and emotion detection (Kollias and Zafeiriou 2019; Zhang et al. 2023; Savchenko 2023). When users request the retraction of their private data to prevent privacy violations, these companies must carefully adjust both datasets and corresponding AI models, underscoring the complexity and importance of a comprehensive privacy-preserving strategy.

Machine Unlearning

In the field of machine learning, both exact and approximate machine unlearning methods have been explored to address the challenge of completely removing the influence of certain data segments from trained models. Exact machine unlearning aims to fully eliminate data influence, often necessitating some degree of model retraining. Early work by Cao *et al.* (Cao and Yang 2015) introduced a heuristic approach to convert machine learning algorithms into a summation format, facilitating the removal of data lineage. Subsequent methods, such as Sharded, Isolated, Sliced, and Aggregated training (SISA) by Bourtoule *et al.* (Bourtoule et al. 2021),

proposed a sharded, isolated, sliced, and aggregated training approach to enhance retraining efficiency. DeltaGrad (Wu, Dobriban, and Davidson 2020), another exact method, accelerates retraining by counteracting the data designated for deletion, though it is limited to specific algorithms and cannot manage mini-batch sizes.

On the other hand, approximate machine unlearning focuses on minimizing the differences between models before and after data removal, aiming to preserve model performance. Notable methods include the certified-removal approach by Guo *et al.* (Guo et al. 2020), which uses a Newton step to erase the influence of data points in L_2-regularized linear models, and a scrubbing method for deep neural networks proposed by Golatkar *et al.* (Golatkar, Achille, and Soatto 2020). These approaches often incorporate differential privacy mechanisms to obscure residual information. Additionally, specialized methods, such as those by Ginart *et al.* (Ginart et al. 2019) for K-means, highlight the development of model-specific unlearning techniques. While approximate methods offer privacy safeguards and performance preservation, they face challenges in verifying implementation and aligning with legal requirements like the "right to be forgotten."

User-centric Interactive Image Privacy Protection Framework

In this paper, we propose a user-centric interactive image privacy protection framework that converts a potentially risky model into a safe one by unlearning specific attribute features, as depicted in Fig. 1. The framework begins with a training dataset labeled as "risky" due to its inclusion of sensitive information. This dataset is used to train a risky model. The process involves identifying a sensitive subset within the training data, guided by a resource specification step, which isolates sensitive attributes or individuals. Concurrently, a transferring dataset, deemed "safe", undergoes distribution matching to ensure alignment with the sensitive subset. This allows for the safe transfer of attribute features, producing an "attribute transferred sensitive subset" that mitigates privacy risks. The core of this framework is the advanced machine unlearning (AMU) module, which employs a refined machine feature unlearning algorithm to efficiently remove sensitive attributes from the model without the need for costly retraining.

To ensure privacy preservation, the framework integrates a Diffusion-DP module, providing DP guarantees by training on the transformed sensitive subset. The result is a "safe model" that can perform accurate inferences while safeguarding sensitive information. This approach not only mitigates the risks associated with sensitive data exposure but also maintains high model performance and utility, making it an efficient solution for adapting to dataset changes while rigorously protecting user privacy.

In what follows, we will provide detailed insights into each module of the framework.

Differential Privacy

For an (ϵ, δ)-DP mechanism, $\epsilon > 0$ denotes the distinguishable bound of all outputs on adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ in a dataset[1], and δ denotes the probability that the ratio of the probabilities of two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ cannot be bounded by $\exp(\epsilon)$ after adding a privacy-preserving mechanism (McSherry and Talwar 2007). The specific definition of DP is provided as follows.

Definition 1 ((ϵ, δ)-DP (Abadi et al. 2016)) *A randomized mechanism $\mathcal{M}: \mathcal{X} \to \mathcal{R}$ with domain $\mathcal{X}$ and range $\mathcal{R}$ satisfies (ϵ, δ)-DP, if*

$$\Pr\left[\mathcal{M}(\mathcal{D}) \in \mathcal{S}\right] \leq e^{\epsilon} \Pr\left[\mathcal{M}(\mathcal{D}') \in \mathcal{S}\right] + \delta, \tag{1}$$

for all measurable sets $\mathcal{S} \subseteq \mathcal{R}$ and for any two adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}$.

Definition 2 ((α, ϵ)-RDP (Mironov 2017)) *A randomized mechanism $\mathcal{M}: \mathcal{X} \to \mathcal{R}$ with domain $\mathcal{X}$ and range $\mathcal{R}$ satisfies (α, ϵ)-RDP, if*

$$D_{\alpha}(\mathcal{M}(\mathcal{D}) \| \mathcal{M}(\mathcal{D}')) \leq \epsilon, \tag{2}$$

for all measurable sets $\mathcal{S} \subseteq \mathcal{R}$ and for any two adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}$.

Lemma 1 (RDP to $(\epsilon, \delta) - DP$ (Mironov 2017)) *If a mechanism $\mathcal{M}: \mathcal{X} \to \mathcal{R}$ satisfy (α, δ)-RDP, it also satisfies $\left(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right)$-DP for any $0 < \delta < 1$. Moreover, $\mathcal{M}$ satisfies pure ϵ-DP.*

We analyze the sensitivity and privacy performance of the proposed time-varying DP noise perturbation mechanism. We use the ℓ_2-norm sensitivity, as given by

$$\Delta s = \max_{\mathcal{D}, \mathcal{D}'} \left\| s(\mathcal{D}) - s(\mathcal{D}') \right\|, \tag{3}$$

where $s(\cdot)$ is a general function in $\mathcal{D}$.

Diffusion Differential Privacy Model

To implement DP for images, we consider a diffusion mechanism that generates images while incorporating noise to ensure privacy.

Definition 3 (Weighted Diffusion Mechanism) *Assume $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^d$, let $\mathcal{M}(\mathbf{X})$ be the mechanism that is a random function taking X as the input, and returning $\mathcal{M}(\mathbf{X})$ at the $(t+1)$-the iteration:*

$$X_{t+1}(i) = s_{t+1}(X_t(i)) + w_c^t(i) \cdot n_{t+1}(i), \tag{4}$$

$$X_{t+1} = Decoder(Encoder(X_t) + w_c^t(i) \cdot n_{t+1}(i)), \tag{5}$$

where $s_{t+1}(\cdot)$ is the contractive map, $\mathbf{w}_c^t = [w_c^t(1), \cdots, w_c^t(i), \cdots, w_c^t(d)]$ is the weighted vector obtained from a N-class classifier, and $\mathbf{n}_{t+1} \sim OU(\theta, \rho)$ at the $(t+1)$-th iteration, i.e., $\mathbf{n}_t = \mathcal{N}(e^{-\theta t}x, \frac{\rho^2}{\theta}(1 - e^{-2\theta t})\mathbb{I}_d)$. Let $\mathcal{M}(\mathbf{X}')$ be obtained from $\mathbf{X}'$ under the same mapping. For $\alpha \geq 1$, $\mathcal{M}(\mathbf{X})$ satisfies

$$D_{\alpha}(\mathcal{M}(\mathbf{X}) \| \mathcal{M}(\mathbf{X}')) \leq \frac{\alpha \|\mathbf{X} - \mathbf{X}'\|^2}{2T\left(\frac{w_c^t(i)}{\sum_{i=1}^{N} w_c^t(i)} \cdot \frac{\rho^2}{\theta}(e^{2\theta t} - 1)\right)}. \tag{6}$$

[1]Two datasets, $\mathcal{D}$ and $\mathcal{D}'$, are adjacent if $\mathcal{D}'$ can be built by adding or removing a single training example from $\mathcal{D}$.

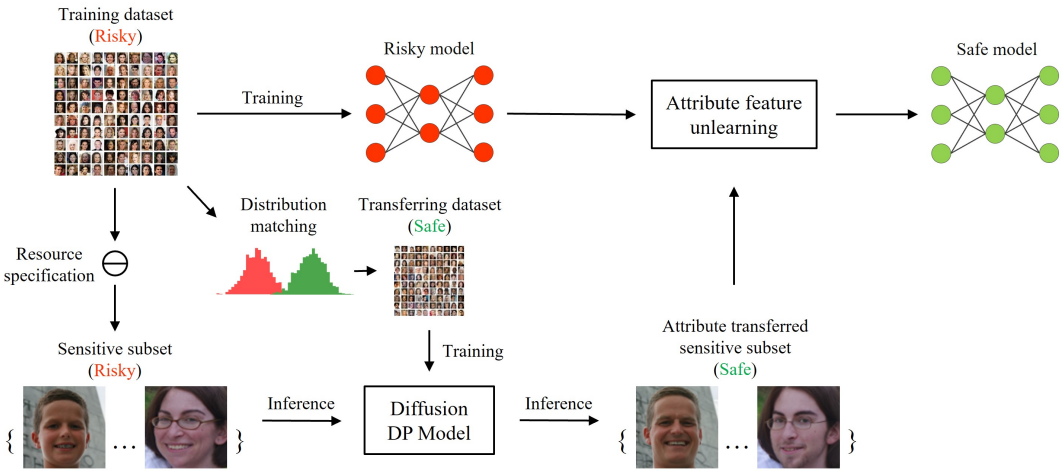

Figure 1: A user-centric interactive image privacy protection framework, which protects sensitive information in image data for machine learning. It transforms a risky training dataset and model into a safe model using attribute feature unlearning and DP techniques, ensuring privacy by modifying sensitive attributes and matching data distributions while balancing privacy protection with model performance through user feedback and adjustments.

Theorem 1 (Diffusive DP-Image) *Suppose that an input image* $\mathbf{X}_0$*, and its corresponding output image* $\mathbf{X}_T$*.* $s(\cdot)$ *maps the image into its latent space (or, feature space), and* $s'(\cdot)$ *is a 1-Lipschitz. Let* $s(\cdot)$ *have global* L_2*-sensitivity* Δ_s *and* $\mathbf{P} = (P_t)_{t\geq 0}$ *be the Ornstein-Uhlenbeck (OU) process with parameters* θ *and* ρ*. For any* $\alpha > 1$ *and* $t > 0$ *the OU mechanism* $\mathcal{M}_t^s(\mathbf{X}_t) = P_t(s(\mathbf{X}_t))$ *satisfies*

$$\left(\alpha, \frac{\alpha\theta\Delta_s^2}{2T\left(\frac{w_c^t(i)}{\sum_{i=1}^N w_c^t(i)}\rho^2(e^{2\theta t-1})\right)}\right)\text{-RDP.}$$

Proof 1 *According to* **Definition 3***, we have* $\mathcal{M}(s(\mathbf{X}))$ *for* $\alpha \geq 1$ *satisfies*

$$D_\alpha\left(\mathcal{M}(s(\mathbf{X}))\|\mathcal{M}(s(\mathbf{X}'))\right) \leq \frac{\alpha\|s(\mathbf{X}) - s(\mathbf{X}')\|^2}{2T\left(\frac{w_c^t(i)}{\sum_{i=1}^N w_c^t(i)} \cdot \frac{\rho^2}{\theta}(e^{2\theta t} - 1)\right)} \tag{7a}$$

$$\leq \frac{\alpha \max_{\{\mathbf{X},\mathbf{X}'\}}\|s(\mathbf{X}) - s(\mathbf{X}')\|_2^2}{2T\left(\frac{w_c^t(i)}{\sum_{i=1}^N w_c^t(i)} \cdot \frac{\rho^2}{\theta}(e^{2\theta t} - 1)\right)} \tag{7b}$$

$$= \frac{\alpha\Delta_s^2}{2T\left(\frac{w_c^t(i)}{\sum_{i=1}^N w_c^t(i)} \cdot \frac{\rho^2}{\theta}(e^{2\theta t} - 1)\right)}, \tag{7c}$$

where (7c) *is obtained based on* (3).

Based on the definition of (α, ϵ)*-RDP, we have* $D_\alpha(\mathcal{M}(\mathbf{X})\|\mathcal{M}(\mathbf{X}')) \leq \epsilon$*. Therefore, the OU mechanism* $\mathcal{M}_t^s(\mathbf{X}_t) = P_t(s(\mathbf{X}_t))$ *satisfies*

$$\left(\alpha, \frac{\alpha\theta\Delta_s^2}{2T\left(\frac{w_c^t(i)}{\sum_{i=1}^N w_c^t(i)}\rho^2(e^{2\theta t-1})\right)}\right)\text{-RDP.}$$

Algorithm 1: Diffusion-DP algorithm

Parameters: Noise coefficient σ, and number of iterations T.
Input: The original image X_t.
Output: Output Privacy-preserving image X_T.
while $0 \leq t < T$ **do**
  $\mathbf{z}_t = s(\mathbf{X}_t)$
  $\mathbf{z}_{t+1} = s'(\mathbf{z}_t) + \mathbf{w}_c^t \cdot \mathbf{n}_t,\ \mathbf{n}_t \sim \mathrm{OU}(\theta, \rho)$
  $\mathbf{X}_{t+1} = g(\mathbf{z}_{t+1})$
end

Advanced Machine Unlearning

This section presents our methodology for unlearning specific attributes from input images using first-order and second-order optimization techniques, aiming to remove the influence of specific attributes from the model while preserving the overall performance of other features.

First-Order Update To unlearn data, we aim to find an update $\Delta(\mathbf{X}, \tilde{\mathbf{X}})$ for our model w^*, where $\mathbf{X}$ and $\tilde{\mathbf{X}}$ are the original data and its perturbed version. If the loss $\mathcal{L}$ is differentiable, we compute the *first-order update* as:

$$\Delta(\mathbf{X}, \tilde{\mathbf{X}}) = -\eta\left(\sum_{\tilde{x}\in\tilde{\mathbf{X}}} \nabla_w \mathcal{L}(\tilde{x}, w^*) - \sum_{x\in\mathbf{X}} \nabla_w \mathcal{L}(x, w^*)\right), \tag{8}$$

where η is the unlearning rate. This update shifts the model parameters to minimize the loss on $\tilde{x}$ while removing the information in x. It differs from gradient descent by moving the model based on the gradient *difference* between original and perturbed data. The gradients can be computed in $\mathcal{O}(p)$, with p being the number of parameters in the learning model (Warnecke et al. 2023).

Second-Order Update The unlearning rate η can be eliminated if $\mathcal{L}$ is twice differentiable and strictly convex. The influence of a single data point is approximated by:

$$\left.\frac{\partial w^*_{\phi,x\to\tilde{x}}}{\partial \phi}\right|_{\phi=0} = -H^{-1}_{w^*}\left(\nabla_w \mathcal{L}(\tilde{x}, w^*) - \nabla_w \mathcal{L}(x, w^*)\right), \tag{9}$$

where $H^{-1}_{w^*}$ is the inverse Hessian of the loss at w^*. This leads to a linear approximation, as given by

$$w^*_{x\to\tilde{x}} \approx w^* - H^{-1}_{w^*}\left(\nabla_w \mathcal{L}(\tilde{x}, w^*) - \nabla_w \mathcal{L}(x, w^*)\right). \tag{10}$$

Extending this to multiple data points gives the *second-order update*:

$$\Delta(\mathbf{X}, \tilde{\mathbf{X}}) = -H^{-1}_{w^*}\left(\sum_{\tilde{x}\in\tilde{X}} \nabla_w \mathcal{L}(\tilde{x}, w^*) - \sum_{x\in X} \nabla_w \mathcal{L}(x, w^*)\right). \tag{11}$$

This update does not require parameter calibration, as it derives from the inverse Hessian of the loss function. The second-order update is preferred for unlearning in models with strongly convex and twice differentiable loss functions, and can be easily calculated with common machine learning frameworks.

Experiment

We verify the effectiveness of our proposed method on the two manually corrupted datasets: CelebA-HQ dataset (Lee et al. 2020) and CelebA dataset (Liu et al. 2015). CelebA-HQ contains 30k high-resolution celebrity images with 40 attributes. We use this CelebA-HQ dataset to train our diffusion model. CelebA has 202,599 images with 40 same attributes as CelebA-HQ. We use this dataset to test our proposed method.

Comparison Methods

We compare our method against several SOTA methods. The image-manipulated methods involve processing the input images (224x224 pixels) using different techniques: (i) Random central removal of 100x100 pixels, which involves randomly removing a central portion of the image; (ii) Whole image mosaic, which applies a mosaic filter to the entire image, blurring detailed features; and (iii) our proposed diffusion-based method, which leverages diffusion processes to selectively alter specific attributes in the images. For fine-tuning methods, we implement three unlearning approaches: (i) Retraining, which involves training the model from scratch on a modified dataset; (ii) First-order unlearning, which adjusts the model parameters using gradient descent based on the attribute-specific loss; and (iii) Second-order unlearning, our proposed method, which employs both gradient and curvature information to more effectively remove specific attributes from the model.

Evaluation Metrics

In our experiments, conducted on the CelebA dataset, we employed several evaluation metrics to assess the performance of attribute modification and image realism. Specifically, we focused on two primary metrics: Attribute Classification Accuracy (ACC) and Structural Similarity Index (SSIM).

For the assessment of attribute modification, we utilized the ACC. In this approach, we trained a linear binary classifier on the training set for each of the 40 attributes. This allowed us to evaluate the accuracy with which each attribute was correctly modified or maintained. The effectiveness of the attribute modifications across the entire image was further quantified by calculating the average ACC across all 40 attributes. This average ACC provides a comprehensive measure of the system's performance in handling multiple attributes simultaneously, ensuring that modifications are not only accurate but also consistent across the dataset.

In addition to ACC, we also assessed the realism of the generated images using the SSIM. Following the methodologies presented in (Hinojosa, Niebles, and Arguello 2021; Liu et al. 2021b), SSIM was chosen due to its ability to measure the perceptual quality of images. SSIM evaluates the similarity between two images based on luminance, contrast, and structural information, producing a similarity score that ranges from -1 to 1. Higher SSIM values indicate a greater degree of similarity between the images, reflecting the extent to which the modified images retain their visual fidelity and align with human visual perception. Conversely, lower SSIM values suggest a decline in image quality or realism.

Together, these metrics provide a robust framework for evaluating both the technical accuracy of attribute modifications and the visual quality of the resulting images. The combination of ACC and SSIM ensures that our approach is not only effective in modifying attributes but also capable of generating realistic and visually pleasing images, which is critical for applications that require high levels of visual authenticity.

Experimental Results

Table 1 provides an extensive comparison of various image manipulation methods and their impact on attribute classification performance under different levels of data change (i.e., 1%, 3%, and 5%). The methods compared include 'Mosaic', 'Random', and 'Ours', with each method's performance evaluated across multiple attributes such as `Young`, `Arched`, `Bald`, `Bangs`, and `Blurry`.

Table 1 presents the results for attributes categorized under different percentages of data changes. Notably, in the case of 1% data change, our method achieves the highest average accuracy of 89.82% across 40 attributes, outperforming both the Mosaic' (82.38%) and 'Random' (84.32%) methods. Our method also shows strong performance in specific attributes, such as `Attract` and `Bald`, indicating its effectiveness in handling minimal data changes. As the data change increases to 3% and 5%, the 'Ours' method consistently demonstrates superior performance, maintaining high accuracy in most attributes. For example, under 3% data change, our method reaches an accuracy of 89.07%, compared to 88.96% and 89.28% for 'Mosaic' and 'Random', respectively. This trend continues under the 5% data change scenario, where our method achieves an accuracy of

Table 1: Comparison of different image manipulation methods and their impact on attribute classification performance.

Change	Methods	Average	Young	Attract	Bald	Bangs	Blurry
	Clean	89.75	85.81	79.58	98.59	94.78	95.73
1%	**Mosaic**	89.06(-0.69)	82.38(-3.43)	74.1(-5.48)	98.53(-0.06)	94.6(-0.18)	95.79(+0.06)
	Random	89.31(-0.44)	**84.32(-1.49)**	76.04(-3.54)	98.52(-0.07)	94.55(-0.23)	**95.9(+0.17)**
	Ours	**89.82(+0.07)**	84.13(-1.68)	**77.8(-1.78)**	**98.59(0.00)**	**94.72(-0.06)**	95.78(+0.05)
3%	**Mosaic**	88.96(-0.79)	82.94(-2.87)	74.84(-4.74)	98.57(-0.02)	**94.56**	**95.67(-0.06)**
	Random	**89.28(-0.47)**	**84.16(-1.65)**	75.56(-4.02)	98.58(-0.01)	**94.56(-0.22)**	95.6(-0.13)
	Ours	89.07(-0.68)	83.53(-2.28)	**77.6(-1.98)**	**98.59(0.00)**	**94.56(-0.22)**	95.66(-0.07)
5%	**Mosaic**	88.96(-0.79)	79.94(-5.87)	78.84(-0.74)	**98.57(-0.02)**	**94.56(-0.22)**	**95.67(-0.06)**
	Random	88.91(-0.84)	82.87(-2.94)	71.07(-8.51)	**98.57(-0.02)**	81.97(-12.81)	95.42(-0.31)
	Ours	**89.05(-0.7)**	**82.95(-2.86)**	**79.63(+0.05)**	98.4(-0.19)	94.09(-0.69)	95.42(-0.31)

89.05%, again outperforming Mosaic (88.96%) and Random (88.91%).

Overall, the results indicate that our method is particularly robust and adaptable, offering high accuracy even as the data change percentage increases. This consistent performance across different attributes and levels of data manipulation highlights the method's potential for robust attribute classification in dynamic datasets. Such reliability is crucial for real-world applications where data may vary significantly, and accurate attribute classification is essential. Our proposed method's superior results in most categories, especially under higher data changes, highlight its effectiveness and utility in improving attribute classification performance.

Table 2 provides a detailed comparison of various unlearning methods and their impact on attribute classification performance under different levels of data change (3%, 5%, and 10%). The methods evaluated include 'Retraining', 'First_order', and 'Second_order', with their performance assessed across multiple attributes such as `Young`, `Arched`, `Bald`, `Bangs`, and `Blurry`. The differences from the 'clean' baseline are highlighted, with all selected attributes having an initial accuracy of less than 90%.

In the clean data scenario, the 'Retraining' method method achieves the highest average accuracy of 89.75% across 40 attributes, followed closely by 'First_order' (89.03%) and 'Second_order' (89.01%). As the data change increases to 3%, 5%, and 10%, the 'First_order' and 'Second_order' methods exhibit competitive performance, often surpassing 'Retraining' in specific attributes. For instance, under a 3% data change, 'Second_order' attains notable accuracies in attributes like `Bald` (98.29%) and `Bangs` (94.00%), with values close to those of the clean scenario, demonstrating robustness against minor data alterations. As the data change reaches 5% and 10%, the 'First_order' and 'Second_order' methods continue to demonstrate robust performance. Specifically, at a 10% data change, 'Second_order' maintains an average accuracy of 88.99%, excelling in challenging attributes like `Bald` (98.04%) and `Bangs` (91.15%). The 'First_order' method, although slightly lower in accuracy, still achieves a respectable average accuracy of 88.64%, showing reasonable adaptability in high data change scenarios. The method's ability to maintain high accuracy under significant data changes underscores its robustness and effectiveness.

Overall, these results demonstrate the efficacy of the 'First_order' and 'Second_order' unlearning methods in maintaining high attribute classification performance across varying levels of data change. The 'Second_order' method, in particular, shows superior adaptability and accuracy in more challenging scenarios, suggesting its potential as a reliable unlearning technique for robust attribute classification in dynamic and diverse datasets. These insights are valuable for selecting appropriate unlearning methods to enhance attribute classification performance in practical applications.

Table 3 compares the performance of various image generation models on the CelebA-HQ dataset, measured by the SSIM metric. Diff-AE leads with a near-perfect SSIM of 0.991, indicating exceptional retention of image details. Our method, with an SSIM of 0.951, ranks second, showcasing its strong capability to preserve image realism, highlighting its competitiveness and effectiveness in generating high-quality images after attribute modifications.

Table 4 summarizes the training times for different unlearning methods. The 'Clean' method takes 13,322 seconds, while the 'Retraining' method, which likely involves comprehensive retraining, required the most time at 15,185 seconds. In contrast, the 'First_order' and 'Second_order' methods drastically reduce the training times to 174.41 seconds and 506.99 seconds, respectively. Notably, the 'Second_order' method, while not as fast as 'First_order', still offers a substantial reduction in time compared to 'Clean' and 'Retraining'. This additional computational time for the 'Second_order' method suggests it may involve a more detailed unlearning process, potentially leading to better accuracy, as indicated in Table 2. This efficiency, coupled with the possibility of improved accuracy, makes the 'Second_order' method particularly valuable for applications where balancing computational resources and maintaining high performance is critical. The sharp decrease in training times with both 'First_order' and 'Second_order' methods demonstrates their potential utility in scenarios necessitating quick updates or modifications, while the slight increase in computational time for the 'Second_order' method could indicate a trade-off that favors accuracy without significantly compromising efficiency.

Table 2: Comparison of different unlearning methods and their impact on attribute classification performance.

Change	Methods	Average	Young	Attract	Bald	Bangs	Blurry
Clean	**Retraining**	89.75	85.81	79.58	98.59	94.78	95.73
	First_order	**89.03(-0.72)**	78.07(-7.74)	67.67(-11.91)	98.31(-0.28)	91.13(-3.65)	95.03(-0.70)
	Second_order	89.01(-0.74)	**79.65(-6.16)**	**72.65(-6.93)**	**98.49(-0.10)**	**91.44(-3.34)**	**95.07(-0.66)**
3%	**Retraining**	89.07(-0.68)	83.35(-2.46)	76.98(-2.60)	98.46(-0.13)	94.47(-0.31)	95.60(-0.13)
	First_order	89.00(-0.75)	82.80(-3.01)	60.62(-18.96)	98.07(-0.52)	92.53(-2.25)	95.14(-0.59)
	Second_order	**89.01(-0.74)**	**83.26(-2.55)**	**68.01(-11.57)**	**98.29(-0.30)**	**94.00(-0.78)**	**95.16(-0.57)**
5%	**Retraining**	89.01(-0.74)	83.91(-1.90)	76.96(-2.62)	98.37(-0.22)	93.94(-0.84)	95.62(-0.11)
	First_order	88.66(-1.09)	81.18(-4.63)	67.93(-11.65)	98.01(-0.58)	92.01(-2.77)	95.1(-0.63)
	Second_order	**89.02(-0.73)**	**81.27(-4.54)**	**68.18(-11.40)**	**98.23(-0.36)**	**93.00(-1.78)**	**95.15(-0.58)**
10%	**Retraining**	89.05(-0.70)	82.95(-2.86)	73.98(-5.60)	98.40(-0.19)	94.09(-0.69)	95.42(-0.31)
	First_order	88.64(-1.11)	79.10(-6.71)	61.10(-18.48)	98.03(-0.56)	**91.19(-3.59)**	95.05(-0.68)
	Second_order	**88.99(-0.76)**	**79.54(-6.27)**	**61.49(-18.09)**	**98.04(-0.55)**	91.15(-3.63)	**95.13(-0.60)**

Table 3: Image generation quality on CelebA-HQ (Lee et al. 2020).

Model	SSIM ↑
StyleGAN2 ($\mathcal{W}$) (Karras et al. 2020)	0.677
StyleGAN2 ($\mathcal{W}$+) (Karras et al. 2020)	0.827
VQ-GAN (Esser et al. 2021)	0.782
DDIM (T=100, 128^2) (Song, Meng, and Ermon 2020)	0.917
Diff-AE (T=100, 128^2) (Preechakul et al. 2022)	**0.991**
Ours (T=100, 128^2)	0.951

Table 4: Training times for different methods.

Methods	Training Time
Clean	13322s
Retraining	15185s
First_order	174.41s
Second_order	506.99s

Conclusion

This paper introduces a privacy protection framework for image data, designed to address challenges in both data sharing and model publication. By integrating a diffusion model for attribute modification with machine unlearning algorithms to secure model parameters, our approach successfully balances the dual objectives of privacy protection and model performance. The framework's user-interactive design enables customized privacy settings, ensuring that high-quality image generation and precise attribute modifications are maintained. Our results demonstrate that this framework not only preserves image realism but also enhances accuracy across facial datasets, representing a significant advancement in privacy-preserving machine learning.

References

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

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 141–159. IEEE.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, 463–480. IEEE.

Corporation, M. 2023. Microsoft Azure Face API. Web page. https://azure.microsoft.com/en-us/services/cognitive-services/face/.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Eidinger, E.; Enbar, R.; and Hassner, T. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security* 9(12): 2170–2179.

Esser, P.; Rombach, R.; Blattmann, A.; and Ommer, B. 2021. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in neural information processing systems* 34: 3518–3532.

European Parliament; and Council of the European Union. ???? Regulation (EU) 2016/679 of the European Parliament and of the Council. URL https://data.europa.eu/eli/reg/2016/679/oj.

FaceApp.com. 2020. FaceAPP. URL https://www.faceapp.com/. Last accessed: Nov. 16, 2020.

Fan, L. 2018. Image Pixelization with Differential Privacy. In *32th IFIP Annual Conference on DBSec*, 148–162. Springer International Publishing.

Fan, L. 2019. Practical image obfuscation with provable privacy. In *ICME*, 784–789. IEEE.

Gao, B.-B.; Zhou, H.-Y.; Wu, J.; and Geng, X. 2018. Age Estimation Using Expectation of Label Distribution Learning. In *IJCAI*, 712–718.

Ginart, A.; Guan, M.; Valiant, G.; and Zou, J. Y. 2019. Making AI forget you: Data deletion in machine learning. *Advances in neural information processing systems* 32.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9304–9312.

Guo, C.; Goldstein, T.; Hannun, A.; and Van Der Maaten, L. 2020. Certified data removal from machine learning models. In *Proceed-*

ings of the 37th International Conference on Machine Learning, 3832–3842.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hinojosa, C.; Niebles, J. C.; and Arguello, H. 2021. Learning privacy-preserving optics for human pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2573–2582.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS* 33: 6840–6851.

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

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

Kollias, D.; and Zafeiriou, S. 2019. Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 297. BMVA Press.

Kuprashevich, M.; and Tolstykh, I. 2023. MiVOLO: Multi-input Transformer for Age and Gender Estimation .

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Q.; Wang, J.; Yao, Z.; Li, Y.; Yang, P.; Yan, J.; Wang, C.; and Pu, S. 2022. Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20513–20522.

Li, T.; and Clifton, C. ???? Poster: Differentially Private Imaging. In *The 42nd IEEE Symposium on Security and Privacy (S&P), 2021*.

Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; and Lin, Z. 2021a. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)* 54(2): 1–36.

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

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lucic, M.; Kurach, K.; Michalski, M.; Gelly, S.; and Bousquet, O. 2018. Are gans created equal? a large-scale study. *NeurIPS* 31.

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

McSherry, F.; and Talwar, K. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103. IEEE.

Mironov, I. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, 263–275. IEEE.

No, C. 1988. Privacy Act 1988 .

Pan, H.; Han, H.; Shan, S.; and Chen, X. 2018. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5285–5294.

Preechakul, K.; Chatthee, N.; Wizadwongsa, S.; and Suwajanakorn, S. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *CVPR*, 10619–10629.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Savchenko, A. 2023. Facial Expression Recognition with Adaptive Frame Rate based on Multiple Testing Correction. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, 30119–30129. PMLR. URL https://proceedings.mlr.press/v202/savchenko23a.html.

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

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Times, F. 2018. Facebook privacy breach. *Retrieved January* 6: 2022.

Warnecke, A.; Pirch, L.; Wressnegger, C.; and Rieck, K. 2023. Machine Unlearning of Features and Labels. In *30th Annual Network and Distributed System Security Symposium, NDSS 2023*. The Internet Society.

Wen, Y.; Liu, B.; Ding, M.; Xie, R.; and Song, L. 2022. Identitydp: Differential private identification protection for face images. *Neurocomputing* 501: 197–211.

Wu, Y.; Dobriban, E.; and Davidson, S. 2020. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, 10355–10366. PMLR.

Zhang, Z.; An, L.; Cui, Z.; Xu, A.; Dong, T.; Jiang, Y.; Shi, J.; Liu, X.; Sun, X.; and Wang, M. 2023. ABAW5 Challenge: A Facial Affect Recognition Approach Utilizing Transformer Encoder and Audiovisual Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5724–5733.