Hiding Among Your Neighbors: Face Image Privacy Protection with Differential Private k-anonymity

1st Jingyi Cao
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
cjycaojingyi@sjtu.edu.cn

2nd Bo Liu
School of Computer Science
University of Technology Sydney
Sydney, Australia
bo.liu@uts.edu.au

3rd Yunqian Wen
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
wenyunqian@sjtu.edu.cn

4th Yunhui Zhu
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
zhuyunhui@sjtu.edu.cn

5th Rong Xie
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
xierong@sjtu.edu.cn

6th Li Song
Institute of Image Communication and Network Engineering
Shanghai Jiao Tong University
Shanghai, China
song_li@sjtu.edu.cn

7th Lin Li
Migu Cultural Technology Co., Ltd.
Shanghai, China
lilin@migu.cn

8th Yaoyao Yin
China Mobile Communications Co., Ltd.
Shanghai, China
yinyaoyao@chinamobile.com

***Abstract*—The development of modern social media allows millions of private photos to be uploaded and shared, which provides a wide range of image acquisition but extremely threatens personal image privacy. Face de-identification is treated as an important privacy protection tool in multimedia data processing by modifying image identity information. Although there exist many traditional methods widely used to hide sensitive private information, they all fail to balance the trade-off between privacy and utility in qualitative and quantitative manners and cannot generate de-identified results with satisfactory visual perception. In this paper, we propose a novel face image privacy protection method with differential private k-anonymity, which can not only generate de-identified results with good image quality but also control the balance between privacy protection and image utility according to different application scenarios. The framework consists of the following three steps: facial attributes prediction, privacy-preserving attributes obfuscation, and naturally realistic de-identificated image generation. Our extensive experiments demonstrate the stability and effectiveness of the proposed model.**

***Index Terms*—Image Privacy, Face De-identification**

I. Introduction

The advances in internet as well as the popularity of smartphones have made it possible for lots of personal photos to be shared on social media every day. While technology brings increasing convenience to our lives, it also poses a certain threat to image privacy. As a consequence, it is crucial for us to learn a method which can protect the sensitive private information of face images before uploading and sharing them with an unknown third party.

There exist many traditional methods to enhance facial privacy in computer vision, where the most widely used methods including blurring, pixelation and masking all hope to obfuscate sensitive information directly. However, it has been proved that these techniques are vulnerable to be defeated and typically preserve neither privacy nor the image utility [1]. Some deep learning models can still identify faces in images encrypted with these techniques with high accuracy. Furthermore, the images protected by these manners result in unsatisfying perception since humans can easily capture the interference. With the development of deep learning, new mechanisms are proposed and applied to enhance image privacy. More state-of-the-art methods [2]–[4] have been proposed to improve the quality and realism of de-identified results using the generative adversarial networks (GAN).

The major challenge of face de-identification is the trade-off between privacy and image utility. The ideal de-identification method should be able to control the balance to adapt to extensive applications. Most GAN-based methods fail to quantify this matter until Li *et al.* [5] proposed that facial privacy is measurable and provided a privacy preservation way with an attribute selection method based on privacy metrics such as k-anonymity [6], l-diversity [7], and t-closeness. However, AnonymousNet modified facial attributes of protected image close to its real world distribution without considering the control of image disturbance degree. We hope to add minor changes for better utility preservation with the condition of privacy protection.

(a) Original (b) Blurring (c) Noise (d) Pixelated (e)DeepPrivacy (f) CIAGAN (g)Anonymous (h) Ours

Fig. 1: Qualitative comparison of various traditional and state-of-the-arts de-identification methods, where (a) input image, (b)-(d) are traditional methods including (b) blurred image, (c) adding Gaussian noise to the pixel, (d) pixelated image, and (e)-(g) are GAN-based methods including (e) DeepPrivacy [2], (f) CIAGAN [3], (g) AnonymousNet [5], and (h) ours.

In this paper, we propose a face image privacy protection method with differential private k-anonymity, including the following two key features: (1) it first finds the average face attributes of the k nearest neighbors of the given image, then edits it towards the direction of the average face, which can hide the identity while ensuring the modification is small. (2) Differential Privacy (DP) is introduced to add randomness and provides further protection on top of the former, because the first step is a deterministic process and limited in protection effectiveness. As the de-identification results shown in Fig.1, our approach can generate the naturally realistic faces and keep similarity with the original images.

II. Related Work

A. Privacy-Preserving Machine Learning

The focus of privacy-preserving machine learning is how to prevent leaking sensitive information in both model and dataset. Differential private machine learning has been widely used in perturbation, which aims to train models with formal guarantees implemented by randomizing the training process such as adding noise to the gradient. Visual privacy attacks and defenses in deep learning has been analyzed in [8]. A model-agnostic approach named "Private Aggregation of Teacher Ensembles" (PATE) [9] introduces a model aggregation strategy and injects randomness in the aggregation process. Another more data-efficient algorithm named Private kNN [10] is the first practical differentially private deep learning solution for large-scale computer vision that can achieve comparable or better consequence than PATE while reducing privacy loss. Inspired by privacy-preserving strategies, we design the obfuscation algorithm in attributes aggression process and apply it to the face de-identification task.

B. Face De-identification Methods

The traditional obfuscation-based methods simply used blurring, masking or pixelation to the face region, which always result in limited utility because of facial information loss. The k-Same family algorithms are based on k-anonymity [6], which can guarantee that each de-identification image indiscriminately relates to at least k faces in the gallery. Owing to the popularity of deep generative models, more novel GAN-based methods have been proposed. DeepPrivacy [2] proposed to replace the whole face region with a fully anonymized image to realize complete protection of sensitive information. CIAGAN [3] introduced a conditional GAN to remove identification characteristics of images and videos while retaining pose features. Additionally, some recent methods [11]–[13] focus on the disentangled identity information in latent space. However, most deep learning methods lack privacy guarantees and cannot meet the adaption of various privacy metrics. AnonymousNet [5] firstly proposed that facial privacy is measurable and designed the privacy-preserving attribute selection (PPAS) algorithm to de-identify images by editing facial attributes. Unfortunately, AnonymousNet only proceed from identity protection and edit the attributes close to the real-world distribution. In our approach, we hope that it can maintain more similarity with the original image and de-identify with a small modification.

III. Preliminaries

A. Face De-identification

The major purpose of de-identification task is to protect identity information. For a given facial image X, the de-identification function $\mathcal{F}$ intends to deceive the face recognition model I and decrease recognition accuracy, which can be formulated as,

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

where $I(X)$ represents the identity information of X. Considering image utility for both users and computer vision tasks, we hope that the de-identified results can retain as much similarity as the original and keep the necessary facial information to allow face detectors can apply. Additionally, better image quality and more satisfactory visual perception are also preferred.

B. Differential Privacy

Differential privacy is a rigorous mathematical definition of privacy and probability is used to take over randomness, which is a strong guarantee since it is based on the statistical property of the mechanism without the requirement of auxiliary information [14].

Definition 1 (ε-differential privacy). Let ε be a positive real number (privacy parameter) and the randomized algorithm $\mathcal{A}: Y \rightarrow \Theta$ is said to provide ε-differential privacy if for all neighboring datasets $D, D' \in Y$ that differ on at most a single element, and all random subsets $S \subset \Theta$ satisfy:

$$\Pr\left[\mathcal{A}\left(D\right) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{A}\left(D'\right) \in S\right]. \tag{2}$$

There are three commonly used mechanisms in differential privacy according to data types: *Laplace*, *Gaussian* and *exponential mechanism*. The overall idea of the exponential mechanism is that when receiving a query, it returns a certain probability value calculated by the scoring function q instead of a deterministic result, thereby achieving differential privacy.

Definition 2 (Exponential Mechanism). Let $q(D, r)$ be a function of dataset D which selects and outputs an element $r \in R$, then an exponential mechanism $\mathcal{M}$ is ε-differential privacy if

$$\mathcal{M}(D) = \left\{ \text{ return } r \text{ with probability} \propto \exp\left(\frac{\varepsilon q(D,r)}{2\Delta q}\right)\right\}, \tag{3}$$

where Δq represents the sensitivity of function q.

IV. Our Approach

We will describe our three-step approach in detail in this section. First of all, we employ the facial attribute classifier to predict original attributes. Then we calculate the obfuscation attributes with differential private k-anonymity algorithm. Finally, we employ the face attribute editing network to generate de-identification results.

Step 1: Attributes Prediction. Firstly, we train a facial attribute extraction network to predict labels of query X, which has two major functions in subsequent operations. On the one hand, when calculating the obfuscation attributes, it will be used as feature extractor to get deep features. On the other hand, when generating the de-identification, we take the different attributes as input, so we need the original prediction for reference.

For the c-label classification problem, we adopt MultiLabelSoftMarginLoss as loss function, which creates a criterion that optimizes a multi-label one-versus-all loss based on max-entropy. For each sample in the minibatch:

$$\begin{aligned}\mathcal{L}(u, v) = -\frac{1}{c}\sum_{i}^{c} v[i]\log\left((1+\exp(-u[i]))^{-1}\right) \\ + (1-v[i])\log\left(\frac{\exp(-u[i])}{1+\exp(-u[i])}\right),\end{aligned} \tag{4}$$

where $v[i] \in \{0, 1\}$. Prediction label u and ground truth v are with the same shape of (n, c), where n is the batchsize while c represents the number of classes. At the end of this step, we can get the original attributes $\mathbb{P}$ of the given image.

Step 2: Obfuscation We design differential private k-anonymity algorithm to acquire the obfuscation attributes, which can be summarized as the following two parts and we will further describe their respective functions in Section V-C.

(a) k-anonymity Average Attributes. For the given no-label query X, we sample a random subset D_γ with the Poisson sampling of probability γ. Both X and D_γ will be mapped into the feature space by a pre-trained feature extractor φ. Then we select k nearest neighbors according to the feature Euclidean distance between $x = \varphi(X)$ and $f = \{f_i = \varphi(d_i) \mid \forall d_i \in D_\gamma\}$. Notice that for a binary classification task, the global sensitivity is 2, while for a problem with c-labels, the global sensitivity will be extended to $2c$, which will make the following noisy-adding mechanisms inefficient. In order to limit the range of global sensitivity, we apply τ-approximation [10] limitation which means each neighbor can only vote for τ attributes at most.

Definition 3 (τ-approximation). Considering the binary multi-label task, the vote of neighbor j upon query X can be expressed as a c-way vector, we apply

$$\hat{v}_{j,i} = v_{j,i} \cdot \min\left(\frac{\tau}{\mid v_j(X) \mid}, 1\right), i \in [1, c], \tag{5}$$

where $|v_j(x)|$ is the $\mathcal{L}_1$ norm of original neighbor j's voting results and $\hat{v}_j$ is the neighbor j's voting results with τ-approximation. The global sensitivity of a randomized algorithm $\mathcal{M}_\tau$ can be reduced to 2τ with this setting.

(b) Differential Privacy. After obtaining the k-anonymity average attributes and the corresponding votes $\mathbb{V} = \{v_1, v_2, \ldots, v_c\}$. To introduce more randomness for privacy protection, we further apply *exponential differential privacy* to voting process as privacy metrics. We divide all privacy-sensitive attributes $\mathbb{A}$ into *independent attributes* $\mathbb{B}$ and *conflict attributes* $\mathbb{C}$, which satisfy $\mathbb{A} = \mathbb{B} \cup \mathbb{C}$ and $\mathbb{B} \cap \mathbb{C} = \varnothing$.

- Independent Attributes $\mathbb{B}$. There is no correlation between independent attribute a_i and other attributes in $\mathbb{B}$, that is, we can individually determine whether to choose it. Therefore, we count the voting results of *with* this attribute v_i as the value of score function q and select the obfuscation attributes based on the probability calculated by

$$p = \frac{\exp\left(\frac{\varepsilon v_i}{2\Delta q}\right)}{\exp\left(\frac{\varepsilon v_i}{2\Delta q}\right) + \exp\left(\frac{\varepsilon(k-v_i)}{2\Delta q}\right)}. \tag{6}$$

- Conflict Attributes $\mathbb{C}$. Considering the exclusivity between attributes, we further divide conflict attributes set $\mathbb{C}$ into groups as $\mathbb{C} = \{\mathbb{G}_1, \mathbb{G}_2, \ldots, \mathbb{G}_m\}$, where is no mutual influence between different groups. Generally speaking, two or more attributes in the same group $\mathbb{G}_i$ will not be selected simultaneously. We respectively count the votes of each attribute as the score function q value

and the probability of selecting attribute a_{i_n} in $\mathbb{G}_i = \{a_{i_1}, a_{i_2}, \ldots, a_{i_m}\}$ by

$$p_{i_n} = \frac{\exp\left(\frac{\varepsilon v_{i_n}}{2\Delta q}\right)}{\sum_{j=1}^{m} \exp\left(\frac{\varepsilon v_{i_j}}{2\Delta q}\right)}. \tag{7}$$

Step 3: Image Generation We adopt a generative adversarial network (GAN) to generate de-identification images according to the obfuscation attributes. For better generation and feature accuracy, we customize the facial attribute editing model based on STGAN [15] which improves manipulation ability by presenting selective transfer units incorporated with encoder-decoder. Different from StarGAN [16] and AttGAN [17], which both take target attributes as input, STGAN only focuses on the changed attributes $attr_{\text{diff}}$ that represents the difference between predicted original facial attributes $\mathbb{P}$ and the obfuscation attributes $\mathbb{O}$ in our approach.

The loss function includes adversarial loss $\mathcal{L}_{adv}$, reconstruction loss $\mathcal{L}_{rec}$ and attribute manipulation loss $\mathcal{L}_{attr}$. The adversarial loss [18] is applied for constraining the generated results to be indistinguishable from real images. We follow Wasserstein GAN (WGAN) and WGAN-GP [19] to define the adversarial loss as,

$$\begin{aligned}\max_{D_{adv}} \mathcal{L}_{D_{adv}} =& \mathbb{E}_X D_{adv}(X) - \mathbb{E}_{\hat{Y}} D_{adv}(\hat{Y}) + \\ & \lambda \mathbb{E}_{\hat{X}} \left[\left(\left\| \nabla_{\hat{X}} D_{adv}(\hat{X}) \right\|_2 - 1 \right)^2 \right],\end{aligned} \tag{8}$$

$$\max_{G} \mathcal{L}_{G_{adv}} = \mathbb{E}_{X, attr_{\text{diff}}} D_{adv}\left(G\left(X, attr_{\text{diff}}\right)\right), \tag{9}$$

where $\hat{X}$ is uniformly sampled between a pair of original and generated images and $\hat{Y} = G\left(X, attr_{\text{diff}}\right)$.

The reconstruction loss is defined as,

$$\mathcal{L}_{rec} = \|X - G(X, 0)\|_1, \tag{10}$$

where the $\mathcal{L}_1$ distance is adopted for ensuring the quality and clarity of the reconstructed images and $G(X, 0)$ is the reconstructed images sharing the same attributes with original.

To improve the accuracy of attributes editing, we introduce the attribute manipulation loss $\mathcal{L}_{attr}$. The attribute classifier D_{attr} shares the common convolution layers with D_{adv} and the attribute manipulation loss is designed as,

$$\begin{aligned}\mathcal{L}_{D_{attr}} = -\sum_{i=1}^{c} \Big[& attr_p^{(i)} \log D_{attr}^{(i)}(X) + \\ & \left(1 - attr_p^{(i)}\right) \log\left(1 - D_{attr}^{(i)}(X)\right),\end{aligned} \tag{11}$$

$$\begin{aligned}\mathcal{L}_{G_{attr}} = -\sum_{i=1}^{c} \Big[& attr_o^{(i)} \log D_{attr}^{(i)}(\hat{Y}) + \\ & \left(1 - attr_o^{(i)}\right) \log\left(1 - D_{attr}^{(i)}(\hat{Y})\right),\end{aligned} \tag{12}$$

where $attr_p^{(i)}$ means the i-th value of prediction attributes $\mathbb{P}$, $attr_o^{(i)}$ indicates the i-th value of obfuscation attributes $\mathbb{O}$ and $D_{attr}^{(i)}(X)$ represents the i-th value of attribute classification results of X by the attribute classifier D_{attr}.

Taking the above losses into account, the overall loss function of discriminator D can be formulated as,

$$\mathcal{L}_D = -\mathcal{L}_{D_{adv}} + \lambda_1 \mathcal{L}_{D_{attr}}, \tag{13}$$

and that for the generator G is

$$\mathcal{L}_G = -\mathcal{L}_{G_{adv}} + \lambda_2 \mathcal{L}_{G_{attr}} + \lambda_3 \mathcal{L}_{rec}, \tag{14}$$

where λ_1, λ_2, and λ_3 are the model tradeoff parameters.

V. Experiments

A. Dataset

We use Large-scale CelebFaces Attributes (CelebA) Dataset [20] which contains 202,599 aligned facial images and 10,177 identities with 40 *with or without* attributes labels of boolean values. In experiments, we use about half of the dataset, of which 75,160 images for training and 26,216 images for test.

B. Implementation Details

Attributes Prediction. We train the facial attributes classification network on CelebA dataset using the Resnet-50 structure. We conduct the batch size of 128, set a base learning rate of 4×10^{-4} reducing by a polynomial decay with a gamma of 0.1 and the weight decay is 5×10^{-4}.

Attributes Obfuscation. When performing de-identification for the given image, we firstly downsample the training set in proportion to γ to get a random subset D_γ and then extract deep features from the fully connected layers of the facial attributes classification network. In our experiments, we consider 13 attributes to protect, including *Bald, Bangs, Black Hair, Blond Hair, Brown Hair, Bushy Eyebrows, Eyeglasses, Male, Mouth Slightly Open, Mustache, No Beard, Pale Skin* and *Young*, due to that they are more distinctive in appearance. Among the attributes considered, we define two sets of conflicting attributes: $\mathbb{G}_1$={*Black Hair, Blond Hair, Brown Hair*} and $\mathbb{G}_2$={*Mustache, No Beard*}, while the other are all defined as independent attributes.

Image Generation Network. We utilize the facial attributes editing to generate de-identified images after obtaining the obfuscation attributes. We train on CelebA dataset for the considered attributes following the settings in [15] where the tradeoff parameters in Equations (13) and (14) are set to $\lambda_1 = 1$, $\lambda_2 = 10$ and $\lambda_3 = 100$.

C. Performance Analysis

Fig.2 illustrates some de-identification results in pairs, where the left presents the original image and the right is the de-identified result generated by our approach. The pure apply of k-anonymity fails to protect sensitive information from the homogeneity attack and is vulnerable to the attacks based on background knowledge [21]. Moreover, the protection effectiveness is limited especially when the value of k is large. Therefore, we employ differential privacy to provide more randomness in obfuscation process of more reliable privacy guarantees. The influence of two main parameters k and ε on the attribute obfuscation is shown in Fig.3, where the accuracy displayed on the y-axis is represented between obfuscation

Fig. 2: Some de-identification results generated by our approach. In each pair, the left is original and the right is de-identified.

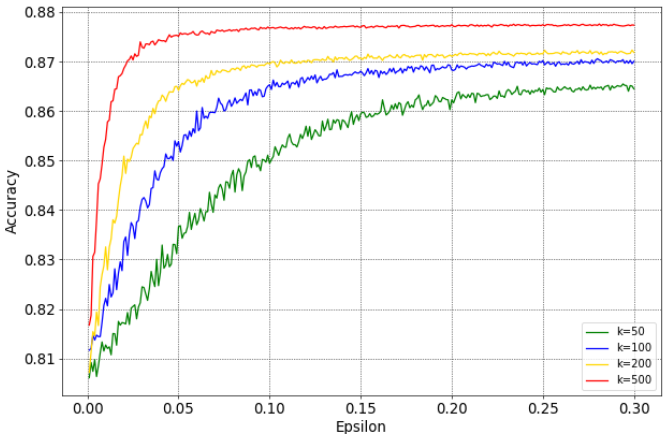

Fig. 3: The influence of different k and ε values on the obfuscation degree, where the y-axis represents *attributes accuracy* between obfuscation attributes and the prediction.

attributes $\mathbb{O}$ and prediction attributes $\mathbb{P}$. We only perturb the considered attributes while the other attributes without privacy protection keep the same as the predicted. When we set $\varepsilon = 0.0$, it means randomly selecting either *with* or *without* independent attributes and choosing one of the conflict attributes in same group, both with the same probability. As ε increases, the extent of disturbance decreases, the accuracy will increase. The attributes accuracy will be greater as k increases with the same ε and the impact of k values has been magnified after the introduction of differential privacy, because function q mainly depends on the voting results. Particularly, due to the design of conflicting attributes mechanism and the prediction deviation of k-anonymity, it will eventually stabilize instead of reaching 100%.

D. Quantitative Evaluation

We use the following metrics to evaluate our approach comparing with existing de-identification methods from both identity protection effectiveness and image utility.

1) **Identity protection effectiveness**: Most of face verification models judge whether two images have the same identity by comparing identity embedding distance. We use the Face Recognition to calculate the **identity distance (*Id-dis*)**, which is based on the deep learning model of dlib and the model tested with Labeled Faces in the Wild datasets can achieve the accuracy of 99.38%.
2) **Image utility**: **(a) Image quality**: We use peak signal-to-noise ratio(PSNR) and structure similarity(SSIM) to measure image distortion at the pixel level. Fréchet Inception Distance [22] is used to measure image distance in feature space. When applying system distortion, lower FID indicates higher image quality. Learned perceptual image patch similarity [23] distance is applied to measure visual similarity which is closer to human perception than traditional metrics. **(b) Utility for computer vision tasks**: We evaluate whether the de-identification results are still usable for identity-independent computer vision tasks by performing face detection with *opencv*. We define the proportion of faces can still be detected in the protected images as **Face Detectability (*Face-det*)**.

The comparison results are presented in Table I. Since the strict threshold for judging whether two images have the same identity is 0.5 in the face recognition model, we choose the values of k and ε to make *Id-dis* basically meet the threshold. We select two sets of parameters with a smaller obfuscation and a larger in traditional methods including blurring, noise and pixelation, and it can be concluded that when adding a small disturbance, there is little impact on image quality but almost no effects on identity protection. Increasing the degree of disturbance contributes higher protection effectiveness, but the image quality and utility will be damaged greatly. Compared with traditional methods, the GAN-based methods can balance the tradeoff better. Additionally, compared with the de-identification methods based on entire face synthesis like DeepPrivacy, our algorithm takes the reduction of modification degree into consideration, so that de-identified results can maintain higher perception similarity (lower LPIPS) with the original. CIAGAN is the identity-swapping-based anonymization methods so that the de-identified face still corresponds to a real identity information, which may cause identity leakage in dataset. Due to the requirement of face landmark and masked background in CIAGAN, it is not convenient in practical applications. Our approach can adjust privacy-protection level by controlling the parameters k and ε, so as to meet the application in different scenarios.

TABLE I: Comparison with other methods under different metrics.

	Id-dis↑	PSNR↑	SSIM↑	FID↓	LPIPS↓	Face-det↑
Blurring(r=5)	0.2573	24.931	0.8005	66.866	0.0654	0.8600
Blurring(r=20)	0.4203	22.666	0.7419	91.623	0.0755	0.6917
Noise(σ=10)	0.2565	21.917	0.7739	32.126	0.0534	0.8136
Noise(σ=30)	0.2911	17.968	0.6281	83.169	0.1265	0.2832
Pixelation(4×4)	0.3251	**25.221**	**0.8278**	**26.073**	0.0326	0.9302
Pixelation(8×8)	0.6908	22.686	0.7010	83.666	0.0915	0.0211
DeepPrivacy [2]	**0.7232**	20.046	0.7605	27.569	0.0868	0.9606
CIAGAN [3]	0.5740	19.014	0.5349	36.719	0.0782	0.9455
AnonymousNet [5]	0.4891	19.102	0.7380	55.047	0.0965	0.8224
Ours(k=100,ε=0.05)	0.5608	19.069	0.7726	52.888	0.0411	**0.9728**
Ours(k=100,ε=0.10)	0.5269	20.308	0.7588	52.214	0.0345	0.9614
Ours(k=200,ε=0.05)	0.4795	21.029	0.8024	38.315	**0.0323**	0.9502

VI. Conclusion

In this paper, we focus on the problem of image privacy and face de-identification. In order to confuse the identity information with minor modifications, we propose a face image privacy protection method to provide metric privacy based on attributes indistinguishability. Our approach consists of three steps: attributes prediction, privacy-protection attributes obfuscation and de-identification image generation. We design the differential private k-anonymity algorithm which combines exponential differential privacy mechanism to introduce additional randomness to the average attributes of k-nearest neighbors in random subset. The method we propose can achieve pleasant visual perception and balance the tradeoff between privacy and utility by adjustable parameters. Experiments demonstrate that our method is effective in identity protection and utility preservation.

Acknowledgment

This work was supported by MoE-China Mobile Research Fund Project (MCM20180702), the 111 Project (B07022 and Sheitc No.150633) and the Shanghai Key Laboratory of Digital Media Processing and Transmissions.

References

[1] R. Gross, E. Airoldi, B. Malin, and L. Sweeney, "Integrating utility into face de-identification," in *International Workshop on Privacy Enhancing Technologies*. Springer, 2005, pp. 227–242.

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

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

[4] Y. Liu, J. Peng, J. James, and Y. Wu, "Ppgan: Privacy-preserving generative adversarial network," in *2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2019, pp. 985–989.

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

[6] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, p. 557–570, 2002. [Online]. Available: https://doi.org/10.1142/S0218488502001648

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

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

[9] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *ICLR*, 2017.

[10] Y. Zhu, X. Yu, M. Chandraker, and Y.-X. Wang, "Private-knn: Practical differential privacy for computer vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 851–11 859.

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

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

[13] Y. Zhao, B. Liu, T. Zhu, M. Ding, and W. Zhou, "Private-encoder: Enforcing privacy in latent space for human face images," *Concurrency and Computation: Practice and Experience*, p. e6548, 2022.

[14] C. Dwork, "Differential privacy: A survey of results," ser. TAMC'08. Berlin, Heidelberg: Springer-Verlag, 2008, p. 1–19.

[15] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3668–3677.

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

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

[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, 2014, p. 2672–2680.

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017.

[20] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[21] W. L. Croft, J.-R. Sack, and W. Shi, "Obfuscation of images via differential privacy: From facial images to general images," *Peer-to-Peer Networking and Applications*, pp. 1–29, 2021.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6629–6640.

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