

A Comparative Evaluation of Differentially Private Image Obfuscation

Dominick Reilly*, Liyue Fan
Department of Computer Science
UNC Charlotte
{dreilly1, liyue.fan}@uncc.edu

Abstract—Image data may contain sensitive information, such as face and iris, which can be misused if in the hands of an adversary. As image data is continuously being collected and shared, it is imperative to ensure the privacy of image data. Widely used image obfuscation methods apply blurring or pixelization to those sensitive regions. However, they are prone to inference attacks, and do not provide quantifiable privacy guarantees. Recently, several obfuscation approaches have been proposed to satisfy the rigorous notion of differential privacy. The goal of this work is to provide a comparative evaluation of those previously proposed approaches in the context of obfuscating face and iris images. We synthesize existing differentially private obfuscation methods and analyze their privacy guarantees. Furthermore, we conduct an extensive empirical evaluation regarding practical utility and privacy protection, with real-world face and iris image datasets. We find that DP-SVD outperforms other methods on several privacy and utility measures. Moreover, we provide an in-depth discussion of our results and point to several considerations when applying those differentially private image obfuscation methods.

Keywords—Image Obfuscation, Comparative Evaluation, Differential Privacy

I. INTRODUCTION

An immense amount of image data is captured from a variety of sensors, and the wide-scale release of such data would be of great benefit to society. For example, image data has proven to be an invaluable asset for researchers, allowing for advancements in intelligent traffic monitoring [14] and early screening of mental illnesses [19]. However, as image data may contain sensitive information, the privacy of individuals captured in the data may be put at risk. For instance, images from traffic cameras may expose a wide array of information, such as faces and license plate numbers, which may be used by adversaries track an individual. As another example, eye-tracking images captured by virtual reality headsets may expose a user’s iris, allowing an adversary to obtain or misuse biometric information. In light of those privacy concerns, image data must be protected by privacy enhancing techniques before sharing with untrusted parties.

Standard image obfuscation approaches consist of methods such as pixelization [17] and blurring [22]. More recently, complex approaches have been proposed to in-paint sensitive image regions [20], [24]. However, those approaches have shown to be susceptible to re-identification attacks, which may

utilize convolutional neural network models [11], [17] and context cues outside the obfuscated regions [24]. Furthermore, the aforementioned obfuscation approaches do not allow the privacy to be effectively bounded. In other words, they do not quantify the sensitive information that may be leaked in the obfuscated image.

In order to address these challenges, several image obfuscation techniques have been proposed recently [7], [8], [13], [25] under the principles of differential privacy [5]. Differential privacy is the state-of-the-art paradigm for quantifying privacy leakage in statistical databases. A differentially private mechanism provides a guarantee that the outputs of any two neighboring databases differing by one entry will be indistinguishable to an adversary.

In this study, we conduct a comparative evaluation of the current differentially private image obfuscation methods. Specifically, our study considers four methods, namely DP-Pix [7] and DP-Samp (an adaptation of [25]) which provide ϵ -differential privacy, DP-SVD [8] which satisfies metric privacy [3], and Snow [13] which satisfies (ϵ, δ) -differential privacy. Our specific contributions are: **(1)** We provide an in-depth analysis of existing DP algorithms for image obfuscation with consistent notation. We made our implementation publicly available at [1]. **(2)** We conduct a systematic evaluation among existing DP methods using real eye and face datasets. We adopt widely used privacy and utility measures, such as ϵ and δ for differential privacy and MSE/SSIM measures for image quality. **(3)** We further evaluate existing methods regarding their usefulness for specific applications, e.g., in eye-tracking tasks, and regarding their practical privacy protection, e.g., in re-identification attacks. **(4)** Last but not least, we thoroughly discuss the observations, analyze the trade-off between privacy and utility, and provide suggestions to domain applications.

The rest of the paper is organized as follows: we describe fundamental concepts around differential privacy in Section II. In Section III, we describe and analyze each DP image obfuscation method. In Section IV, we present our evaluation methodology and the complete set of empirical evaluation results. In Section V, we interpret the observed results and provide recommendations for applications. Section VI concludes the paper with brief discussions on future work.

*Work done while an undergraduate student.

II. PRELIMINARIES

Differential Privacy. Differential privacy [5] is the state-of-the-art notion for quantifying privacy leakage in statistical databases containing sensitive data. Given two neighboring databases, $\mathcal{D}$ and $\mathcal{D}'$, that differ by at most one record, a randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ) -differential privacy [6] if for any $Z \in \text{range}(\mathcal{M})$,

$$\Pr[\mathcal{M}(\mathcal{D}) = Z] \leq e^\epsilon \cdot \Pr[\mathcal{M}(\mathcal{D}') = Z] + \delta. \quad (1)$$

The ϵ and δ parameters specify the degree of privacy provided by the mechanism, also known as the privacy budget. Here, $\epsilon > 0$ bounds the difference between output probabilities of two neighboring databases $\mathcal{D}, \mathcal{D}'$. In addition, $\delta \in [0, 1]$ accounts for the probability of *bad events* that might lead to a privacy breach. ϵ -DP, often called *pure DP*, can be achieved by setting $\delta = 0$. Typically, smaller ϵ and δ values indicate stronger privacy protection, and vice versa.

An advantage of DP is its resistance to post-processing [6], i.e., any computation performed on the output of a DP mechanism would not incur additional privacy cost. Other benefits of DP include the lightweight computation and ease of control over the information leakage with the help of ϵ, δ parameters. Naturally, there exists a trade-off between preserving privacy and maintaining data utility.

Metric-based Privacy. Metric-based privacy [3], i.e., $\epsilon \cdot d_{\mathcal{X}}$ -privacy, extends differential privacy to a set of secrets $\mathcal{X}$ that are equipped with a distance metric, i.e., $d_{\mathcal{X}}$, and guarantees a level of indistinguishability that is proportional to the distance between secrets. Specifically, a mechanism $K : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Z})$ satisfies $\epsilon \cdot d_{\mathcal{X}}$ -privacy, iff $\forall x, x' \in \mathcal{X}$:

$$K(x)(Z) \leq e^{\epsilon \cdot d_{\mathcal{X}}(x, x')} K(x')(Z) \quad \forall Z \in \mathcal{F}_{\mathcal{Z}} \quad (2)$$

where $\mathcal{Z}$ is the output space of K , $\mathcal{F}_{\mathcal{Z}}$ is a σ -algebra over $\mathcal{Z}$, and $\mathcal{P}(\mathcal{Z})$ is the set of probability measures over $\mathcal{Z}$.

To extend the principle of differential privacy to image data, several image obfuscation methods adopt the classic DP definition, such as DP-Pix [7], DP-Samp [25], and Snow [13], to protect pixel values in the input image. The DP-SVD [8] method adopts metric-based privacy, to protect significant singular values of the input image.

III. OBFUSCATION METHODS

A. DP-Pix

Differentially private pixelization [7] (DP-Pix) is the first method that provides differential privacy guarantees for publishing individual images. Given the large number of pixels contained in a typical image, directly perturbing each pixel in the source image would lead to low utility. To balance privacy and utility, DP-Pix adopts pixelization and the Laplace mechanism to satisfy differential privacy.

Algorithm 1: DP-Pix

Input : Input image $\mathcal{I}$, Privacy budget ϵ ,
Block size b , Number of pixels m
Output: Obfuscated image satisfying ϵ -DP

```

1 blocks  $\leftarrow$  partition  $\mathcal{I}$  into blocks of size  $b \times b$ 
2 foreach block in blocks do
3   average  $\leftarrow$  average pixel intensity of block
4   noise  $\leftarrow$  noise drawn from  $\text{Laplace}(0, \frac{255m}{b^2\epsilon})$ 
5   assign intensity of pixels in block to average + noise
6 end
7  $\hat{\mathcal{I}} \leftarrow$  output image constructed from blocks
```

1) *m-neighborhood*: The work [7] proposes a *m-neighborhood* notion to define neighboring images in the context of differential privacy. Two images, I_1 and I_2 , are considered neighboring if they differ by at most m pixels. By varying the value of m , the data owner can control the privacy protection offered by DP-Pix: higher m values indicate indistinguishability in a larger neighborhood, thus stronger privacy protection.

2) *Private Pixelization*: DP-Pix leverages pixelization to reduce the amount of noise required for differential privacy. Pixelization, *a.k.a.* mosaicing, decomposes an image into blocks by superimposing a grid on the source image, where each grid cell (i.e., super-pixel) contains $b \times b$ pixels. The value of each super-pixel is determined by averaging all pixels contained in the grid cell. To achieve ϵ -DP, a perturbation noise is sampled from a Laplacian distribution with mean 0 and scale $\frac{255m}{b^2\epsilon}$ and added to each grid cell. Algorithm 1 depicts the steps taken by DP-Pix.

B. DP-Samp

A recent work [25] proposed a pixel-sampling method to protect visual elements (e.g., persons and objects) in videos. In this study, we design a new method named DP-Samp, which adapts [25] to protecting up to m pixels in a source image and is presented in Algorithm 2. DP-Samp consists of four steps, namely pixel clustering, budget allocation, pixel sampling, and interpolation.

1) *Pixel Clustering*: The goal of clustering is to identify pixel intensities that are useful for reconstructing the image. An intuitive approach is to select the most frequent intensities in the image; however, this approach may not capture the structures of images containing large regions with slightly varying intensities. The work of [25] adopted multi-scale analysis [27] to partition each visual element in a video in k cells; as such methods do not apply to a single image, we propose to generate k pixel clusters with K-means¹. The most frequent intensity in each cluster, $\Psi_{1:k}$, will be candidates for pixel sampling. Note that the multi-scale analysis step in [25] is conducted in the *public* setting, as does the clustering step in our adaptation. The integration with differentially private clustering methods, such as [23], is possible but out of the scope of this study. Figure 2 presents the clustering

¹Note that other clustering methods (e.g., hierarchical clustering) may also apply.

Algorithm 2: DP-Samp

Input : Input image $\mathcal{I}$, Privacy budget ϵ ,
Number of clusters k , Number of pixels m
Output: Obfuscated image satisfying ϵ -DP

- 1 perform pixel clustering to generate k clusters
- 2 calculate most frequent intensity in each cluster
($\Psi_1, \dots, \Psi_k$)
- 3 // Budget allocation
- 4 **foreach** $\Psi_i, i \in [1, k]$ **do**
- 5 | compute the privacy budget $\epsilon(\Psi_i)$ with Eq. 3
- 6 **end**
- 7 // Pixel sampling
- 8 **foreach** $\Psi_i, i \in [1, k]$ **do**
- 9 | compute maximum x_i with Eq. 4
- 10 | randomly select x_i pixels from $\mathcal{I}$ with intensity Ψ_i to
 preserve in output image $\hat{\mathcal{I}}$
- 11 **end**
- 12 // Interpolation
- 13 linear interpolate non-sampled pixels in $\hat{\mathcal{I}}$



(a) Input Image (b) Sampled Pixels (c) Interpolated Image
 Fig. 1: Sampled and interpolated images generated by DP-Samp with $\epsilon = 3$, $k = 15$, $m = 1$

results on eye and face images at varying values of k . Pixels chosen as sampling candidates are highlighted in green. From the images, we see that a higher number of clusters may capture important structures better in both datasets, potentially allowing an more effective reconstruction of the image.

2) *Budget Allocation*: The privacy budget ϵ is split amongst all selected intensities in Ψ . DP-Samp allocates higher privacy budgets to intensities occurring more frequently. Let $\text{Freq}(\Psi_i)$ be the number of pixels in the source image with intensity Ψ_i , then the privacy budget for Ψ_i can be computed as:

$$\epsilon(\Psi_i) = \frac{\epsilon \cdot \text{Freq}(\Psi_i)}{\sum_{j=1}^k \text{Freq}(\Psi_j)} \quad (3)$$

3) *Pixel Sampling*: From each intensity Ψ_i , we randomly sample x_i pixels from the source image, which preserve their location and intensity. The value of x_i is solved by:

$$\max x_i, \text{ s.t. } \binom{c_i}{x_i} / \binom{c_i - m}{x_i} \leq e^{\epsilon(\Psi_i)}, \quad (4)$$

where c_i is the count of pixels with intensity Ψ_i in the input, m is the number of pixels that are allowed to differ between neighboring images. Selecting x_i according to Equation 4 satisfies ϵ -DP, and the analysis is similar to that of [25].

4) *Interpolation*: DP-Samp performs linear interpolation on the sampled pixels to estimate the values of those non-sampled pixels. Utilizing the post-processing property of DP [6], the interpolation does not inflict additional privacy loss in the output image $\hat{\mathcal{I}}$. The sampled pixels and the final interpolated image are shown in Figure 1, we see that pixel interpolation

Algorithm 3: DP-SVD

Input : Input image $\mathcal{I}$, Privacy budget ϵ ,
Number of eigenvalues i
Output: Obfuscated image satisfying $\epsilon \cdot d_i$ -privacy

- 1 decompose $\mathcal{I}$ using SVD: U, Σ, V^T
- 2 $\hat{\Sigma} \leftarrow i$ largest singular values of Σ . ($\hat{\Sigma}_1, \dots, \hat{\Sigma}_i$)
- 3 $\mathcal{N} \leftarrow$ noise vector sampled according to Eq. 5
- 4 **for** $j \in [1, i]$ **do**
- 5 | $\hat{\Sigma}_j \leftarrow \hat{\Sigma}_j + \mathcal{N}_j$
- 6 **end**
- 7 pad $\hat{\Sigma}$ with 0s for the discarded singular values
- 8 $\hat{\mathcal{I}} = U \cdot \hat{\Sigma} \cdot V$

produces an accurate estimation of the input, despite sparsely sampled pixels.

C. DP-SVD

The work of [8] (DP-SVD) adopted to singular value decomposition to decompose a source image into constituent feature matrices that capture perceptual and geometric features in the image. Furthermore, the work developed a novel sampling mechanism in high-dimensional space which achieves metric-based privacy [3]. Algorithm 3 provides an overview of the method.

1) *Singular Value Decomposition*: The singular value decomposition (SVD) is in the form $I = U\Sigma V^T$, where I is the source image, U and V consist of singular vector matrices that capture geometric features, and Σ consists of the singular value matrix that capture the magnitudes of features in U and V . It can be shown that SVD can capture perceptual information in input images, which can be used to robustly hash visually similar images [15], such as after compression, rotation, and cropping. The singular values are considered as sensitive information in [8].

2) *Private Sampling*: The intuition of DP-SVD is to provide indistinguishability guarantees to visually *similar* images, similar to the geo-indistinguishability [2] for location data. To that end, metric-based privacy, i.e., $\epsilon \cdot d_{\mathcal{X}}$ -privacy, was adopted in [8] for protecting image data. DP-SVD achieves metric privacy by perturbing the first i singular values using a novel sampling method. Other singular values are discarded, and the output image is produced with the privacy-enhanced singular values.

Specifically, DP-SVD perturbs the first i singular values according to a specific probability distribution. In an i -dimensional space, let x_0 denote the input vector, i.e., containing the real singular values. A mechanism that samples the output vector x according to following probability distributions satisfies $\epsilon \cdot d_i$ -privacy [8]:

$$D_{\epsilon, i}(x_0)(x) = C_{\epsilon, i} e^{-\epsilon \cdot d_i(x_0, x)} \quad (5)$$

where d_i represents Euclidean distance in the i -dimensional space and

$$C_{\epsilon, i} = \frac{1}{2} \left(\frac{\epsilon}{\sqrt{\pi}} \right)^i \frac{\left(\frac{i}{2} - 1 \right)!}{(i - 1)!} \quad (6)$$

Algorithm 4: Snow**Input** : Input image $\mathcal{I}$, Privacy budget δ **Output**: Obfuscated image satisfying $(0, \delta)$ -DP

```

1  $p \leftarrow (1 - \delta)$ 
2  $S \leftarrow$  random subset of  $p \cdot \mathcal{I}_{width} \cdot \mathcal{I}_{height}$  pixels in  $\mathcal{I}$ 
3  $\hat{\mathcal{I}} \leftarrow \mathcal{I}$ 
4 foreach pixel in  $S$  do
5   | set intensity of pixel to 127 in  $\hat{\mathcal{I}}$ 
6 end

```

Dataset	DP-Pix (b)	DP-Samp (k)	DP-SVD (i)
CASIA	6	28	6
AT&T	4	48	4

Table I: Default algorithm parameters

where i is assumed to be even without loss of generality. Details of sampling according to Equation 5 are described in [8]. Increasing i may lead to a better approximation of the input image, while inflicting a higher perturbation to achieve privacy.

D. Snow

The intuition of Snow [13] is the introduction of noise to an image via randomly re-assigning pixel intensities to a constant value, i.e., 127 for grayscale images. The method is outlined in Algorithm 4. The parameter p controls the proportion of pixels that will be re-assigned and is related to the privacy parameter δ . It is shown in [13] that the method achieves $(0, \delta)$ -differential privacy with $\delta = 1 - p$.

IV. EXPERIMENTS

A. Methodology

We implement the above four differentially private methods in Python 3 and empirically evaluate their performance on utility and privacy measures. Our experiments were conducted on a Linux machine with a 2.20 GHz processor and 12 GB RAM.

1) *Datasets*: We analyze the performance of the methods on two widely used datasets: CASIA-IrisV2 [10] (CASIA) and AT&T Database of Faces [21] (AT&T). CASIA is a collection of iris images containing 2400 images from 60 subjects, and the images have a resolution of 640×480 pixels. AT&T contains 400 face images from 40 subjects, and each image has a resolution of 92×112 pixels.

2) *Default Parameter Values*: The privacy parameters, i.e., ϵ and δ , indicate the level of privacy protection. For ϵ -DP methods (DP-Pix, DP-Samp, and DP-SVD), our evaluation focuses on the range of $\epsilon \in [0.1, 0.5, 1, 5, 10]$; and for Snow, we focus on the range of $\delta \in [0.1, 0.33, 0.4, 0.5, 0.6, 0.7, 0.8]$, as smaller δ values offer little to no usefulness. Other algorithm-specific parameters also help balance privacy and utility, such as b of DP-Pix. The default parameter values used are listed in Table I. Note that the default parameter values are adjusted to the image resolution of each dataset. Furthermore, a smaller number of clusters (k) is sufficient for DP-Samp on CASIA. Figure 2 visualizes the sampling candidate pixels while varying the number of clusters, k . We see that less data

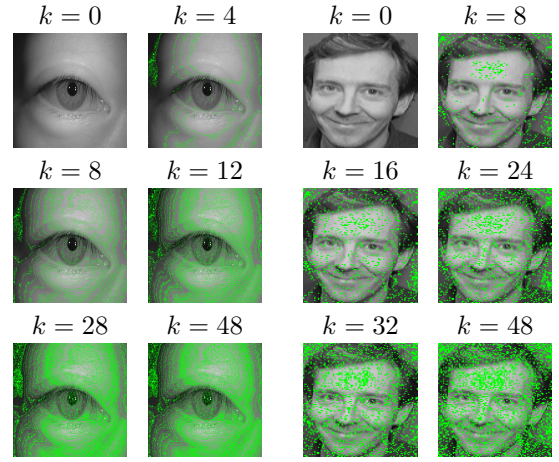


Fig. 2: Pixels in green as sampling candidates for DP-Samp at varying values of k : CASIA (two columns on left) and AT&T (two columns on right)

is required to capture the data structure in CASIA compared to for AT&T, e.g. when $k = 8$.

3) *Generic Utility Measures*: Mean Squared Error (MSE) and Structural Similarity (SSIM) [26] are adopted to quantify the usefulness of the image obfuscation methods. Both measures are computed between the source and the obfuscated images and the average value in each dataset is reported. Intuitively, MSE measures the pixel-wise difference between two images; SSIM captures the differences in the perceived quality (e.g. structure, lighting, contrast) of images.

4) *Task Based Utility*: We adopt *pupil confidence* and *gaze error* as task based utility measures, to support eye-tracking applications [18], [29]. Specifically, for each image in the CASIA dataset, we utilize the DeepVOG [28] framework to compute a pupil confidence score and to estimate the gaze in both the x (yaw) and y (pitch) directions. We report the percentage of images with a confidence score ≥ 0.8 , similar to [13]. Furthermore, we compute the gaze error (in $^\circ$) for gaze estimates obtained from the source image and the obfuscated image and report the average across the CASIA dataset.

5) *Privacy Risk Measures*: In addition to the differential privacy guarantees, we evaluate the practical privacy protection offered by the existing methods. Since each image can be obfuscated locally, this study focuses on the privacy risks associated with sharing obfuscated images.

Correct Recognition Rate (CRR). For iris images, an important privacy risk is that an obfuscated image of a target individual may be misused by an adversary for authentication [4]. For example, the adversary may aim to unlock the target's online account or device, which stores the target's iris baseline (e.g., template), with the obfuscated iris image. To evaluate such risks, we adopt widely used iris segmentation and recognition solutions [9], [16] to extract a binary iris signature for CASIA images. Specifically, we partition the dataset such that 2 randomly selected images for each individual are set aside as *baselines*, and the rest of images for the individual will be used for recognition as in [12]. To authenticate an image, Hamming distances are computed between its iris signature

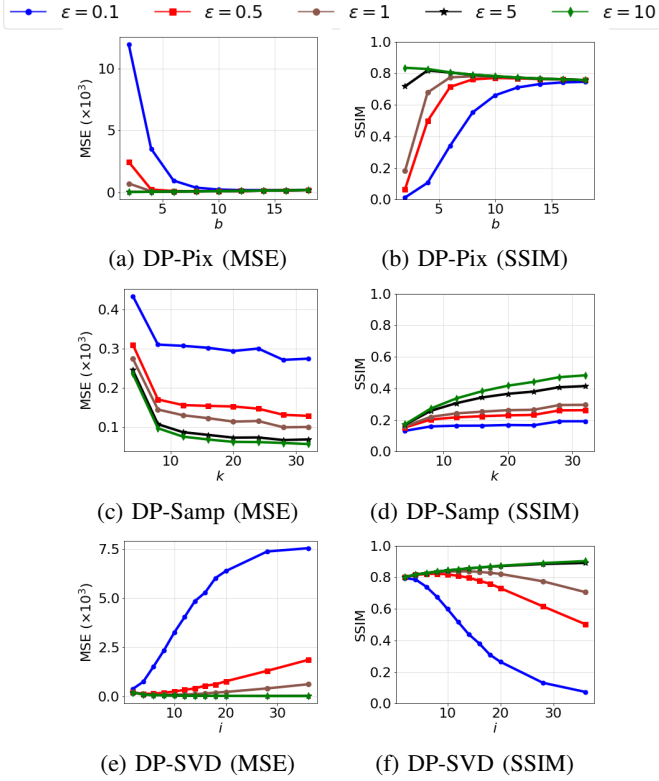


Fig. 3: Mean Squared Error (MSE) and Structural Similarity (SSIM) results of varying algorithm parameters on CASIA dataset

and two baseline signatures of the individual. If the lower distance is below a pre-defined threshold, the authentication is considered successful. In our evaluation, 0.35 is used as the threshold for a successful match, in order to achieve a low false positive rate on real CASIA images. For obfuscated images, we compute the correct recognition rate (CRR) as the percentage of obfuscated images successfully matched with their corresponding baselines. Higher values of CRR indicate higher privacy risks.

Face Re-Identification. For face images, a widely adopted privacy measure is the risk of re-identification based on convolutional neural networks (CNN) [7], [17]. In this setting, an adversary has access to some clear face images of all individuals (e.g., from social media); the adversary can apply any obfuscation method to those images and train a CNN model to predict the identity of an obfuscated image. When a new obfuscated image is available (e.g., a pixelized face in a news article), the adversary applies the trained CNN model to infer the identity of the individual. Similar to [7], [17], we partition the AT&T dataset by randomly selecting 8 images for each individual as training and the remaining 2 images for each individual as testing. A CNN model is trained for each obfuscation method and each parameter value. The accuracy on the testing set is reported in our results, with higher values indicating higher privacy risks.

B. Varying Parameters

In this section, we vary the parameters of the privacy methods and study their effects on utility. Furthermore, we

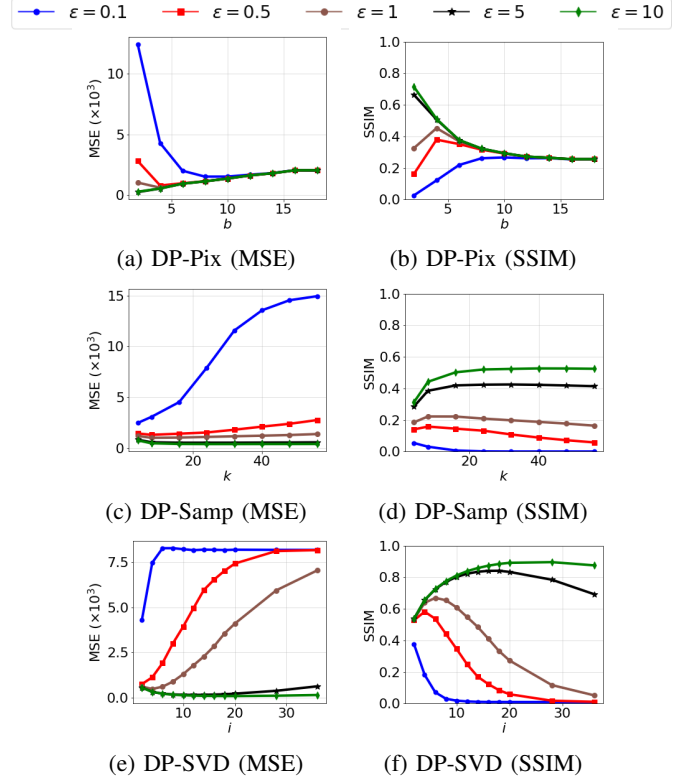


Fig. 4: Mean Squared Error (MSE) and Structural Similarity (SSIM) results of varying algorithm parameters on AT&T dataset

study those effects under different privacy levels, i.e., by varying ϵ and δ values. The results for CASIA and AT&T are reported in Figure 3 and Figure 4, respectively.

1) *Varying b in DP-Pix:* First we evaluate the effect of b in the DP-Pix method. Recall that b specifies the block width in pixels used for image pixelization. In Figure 3a and Figure 3b, we observe that for lower ϵ values, MSE first decreases as b increases and begins to increase when b is larger. The “elbow” point is different for each ϵ value. For instance, in Figure 3b, the elbow point is $b = 4$ for $\epsilon = 0.5$ and 1, and $b = 10$ for $\epsilon = 0.1$. The reason is that increasing b incurs a higher loss of information, i.e., via pixelization, but it helps reduce the magnitude of the Laplace perturbation error introduced by differential privacy, i.e., with scale $\frac{255m}{b^2\epsilon}$. The observed elbow point indicates the b value that minimizes the combined information loss and perturbation error. When $\epsilon > 1$, we do not observe elbow points. It is because when the perturbation error is small, it is always beneficial to adopt a small b value to minimize the information loss. The results of SSIM (Figure 3b and Figure 4b) are consistent with the MSE results. Note that SSIM is a similarity measure, hence the higher the better.

2) *Varying k in DP-Samp:* For this method, we evaluate the number of pixel intensities selected for sampling. On both datasets, we observe an initial decrease in MSE for smaller values of k (e.g. $k \leq 10$ in Figure 3c and Figure 4c.) As k grows, we observe in Figure 3c a decrease in MSE on CASIA for low ϵ values ($\epsilon \leq 1$), and a plateau for other ϵ values. On AT&T, we observe the increase of k leads to an increase in MSE for smaller ϵ values, e.g., $\epsilon = 0.1$ in Figure 4c. The

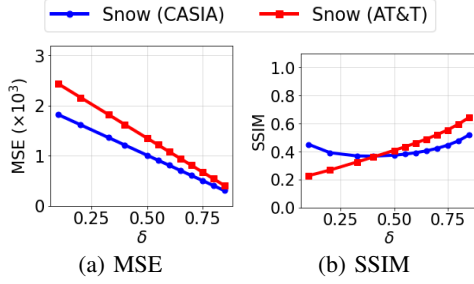


Fig. 5: Varying δ parameter of Snow on CASIA and AT&T datasets

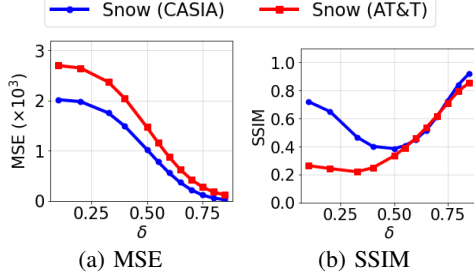


Fig. 6: Varying δ parameter of Snow with median blur on CASIA and AT&T datasets

reason is when sampling, the privacy budget is allocated to each intensity according to Equation 3. Given a sufficiently small ϵ , the allocated budgets for k intensities may be too small, leading to a higher MSE in the obfuscated image. For larger ϵ values, we observe an elbow point in MSE, e.g., $k = 8$ when $\epsilon \geq 0.5$ for AT&T, which indicates a trade-off between more pixel intensities and a smaller privacy budget for each intensity, as k increases. Note that CASIA dataset has a higher resolution and therefore the trade-off is not obvious. The SSIM results for both datasets in Figure 3d and Figure 4d show that DP-Samp does not preserve the structural information in the obfuscated image: the SSIM measure is constantly lower than 0.6, despite increasing the privacy budget ϵ . Moreover, the results show that the structure of AT&T images is more sensitive to changes in k than CASIA images.

3) *Varying i in DP-SVD*: We investigate the effects of the number of eigenvalues (i) preserved in DP-SVD. Similarly to the other methods, increasing i may lead to different effects at different privacy levels. In Figure 3e and Figure 4e, increasing i leads to higher MSE errors in both CASIA and AT&T with small ϵ values (e.g., $\epsilon = 0.1$). For larger ϵ values, e.g., $\epsilon \geq 5$, the MSE first decreases and then increases. The reason is a higher number of eigenvalues allows more information of the original image to be preserved, but would inflict larger perturbation errors by private sampling in higher dimensional spaces. As a result, an elbow point, i.e., lowest total error, is observed when ϵ is sufficiently large; and such ϵ values also depend on the input image, e.g., its resolution and structural complexity. The SSIM results in Figure 4f show the trade-off more clearly. Both Figure 3f and Figure 4f show that higher SSIM scores can be achieved if relaxing privacy: DP-SVD outperforms DP-Pix and DP-Samp given the same ϵ value.

4) *Varying δ in Snow*: Recall that Snow employs a single parameter for pixel sampling, i.e., δ , which also indicates the

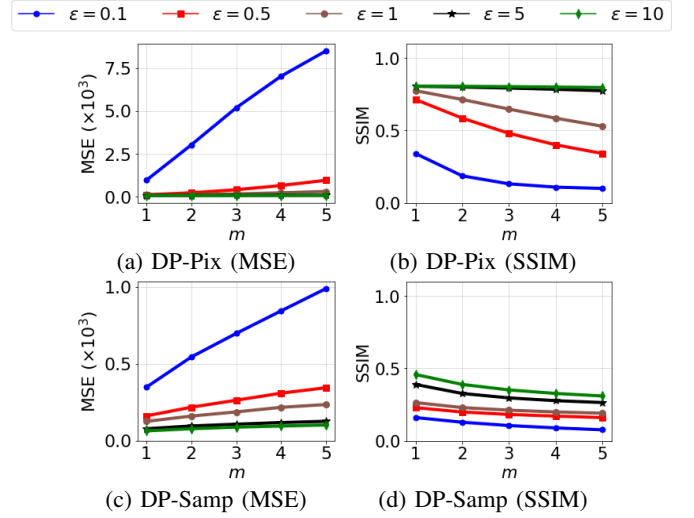


Fig. 7: Results of varying m on CASIA dataset

probability of breaching ϵ -DP. In DP studies, δ is usually set to a small value to ensure adequate privacy protection [6]. For instance, we may set $\delta = \frac{1}{n}$ where n is the number of pixels in the input image, in order to protect each pixel. However, such an δ value will lead to graying the majority of the image (see Algorithm 4), hence no practical usefulness. In Figure 5, we vary δ between 0.1 and 0.85 to study the utility empirically. For both CASIA and AT&T, increasing δ leads to lower MSE errors (Figure 5a), as fewer pixels are grayed out. In Figure 5b, we observe the SSIM first decreases and then increases for the CASIA dataset. We believe that due to the simple structure of CASIA images, SSIM does not capture the difference between the original image and the obfuscated image when most pixels are grayed out (i.e., smaller δ). We also evaluated an extension of Snow (Figure 6), by applying median blur with a 3×3 kernel to the output image. As can be seen in Figure 6b, the median blur improves the image quality when $\delta > 0.5$, i.e., when the majority of the pixels are preserved. For the CASIA dataset, applying median blur in low δ settings (e.g., $\delta \leq 0.2$) also leads to higher SSIM scores, due to the limitation of SSIM when the majority of the pixels are gray, as discussed previously. We see that face images from AT&T, which have more complex structures, are affected by this lack of robustness to a lesser extent.

C. Varying m for neighboring images

In differential privacy, the definition of neighboring databases is an important factor. Similarly, in the context of image obfuscation, the parameter m specifies the number of pixels that may differ between two neighboring images. Larger values of m provide stronger privacy protection, i.e., stronger indistinguishability guarantees, which may require larger perturbation errors. In this evaluation, we adapt DP-Pix and DP-Samp to different m values, while DP-SVD and Snow are not applicable. The utility results are reported for CASIA and AT&T datasets in Figures 7 and 8, respectively. As can be seen, increasing m incurs larger MSE errors and lower SSIM scores for DP-Pix and DP-Samp in both datasets.

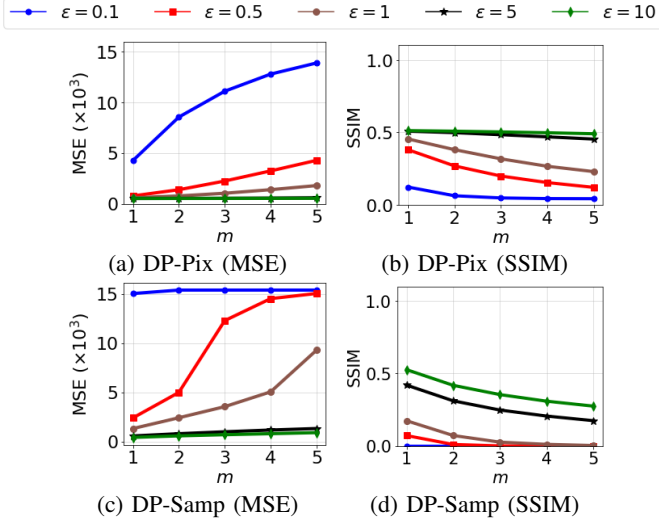


Fig. 8: Results of varying m on AT&T dataset

Table II: Task Based Utility

ϵ	Pupil Confidence			Gaze Error ($^\circ$)		
	DP-Pix	DP-Samp	DP-SVD	DP-Pix	DP-Samp	DP-SVD
0.01	47%	0%	30%	5.48	-	56.98
0.05	46%	2%	46%	4.30	10.41	44.39
0.1	47%	6%	47%	4.38	10.92	21.53
0.5	43%	40%	79%	5.79	5.66	2.58
1	45%	56%	77%	5.09	5.09	2.19
5	45%	77%	77%	4.98	3.02	2.08
10	46%	83%	77%	4.50	2.72	2.07

For CASIA datasets, DP-Samp inflicts lower MSE errors than DP-Pix, enjoying the benefits of pixel sampling in higher resolution images; but in AT&T dataset, DP-Samp does not have that advantage. The results of SSIM show that DP-Pix outperforms DP-Samp when increasing m and provides high quality consistently in high ϵ settings, thanks to preserving the high-level image structure with pixelization.

D. Practical Utility and Privacy Measures

Next, we discuss task based utility measures and practical privacy risks of the methods at varying privacy levels. Results are reported in Tables II, IV, III, and V. Note that in gaze error results, a dash (-) is used to indicate that the gaze error could not be determined due to a 0% pupil detection rate.

1) *Task Based Utility*: In Table II, we observe that when the privacy protection is stronger ($\epsilon \leq 0.1$), DP-Pix provides higher pupil confidence scores and lower gaze errors, compared to DP-Samp and DP-SVD; however, those utility measures do not improve when increasing ϵ , due to the information loss incurred by pixelization. Increasing ϵ for DP-Samp and DP-SVD leads to higher pupil confidence and lower gaze errors. We observe that DP-SVD quickly achieves high utility at a low privacy cost, e.g., $\epsilon = 0.5$. In Table III, it can be seen that increasing δ values in Snow improves pupil confidence and gaze error, with or without median blur. Note that with median blur, Snow achieves better utility in both measures at low privacy settings, i.e., $\delta \geq 0.7$, as the median blur removes noise effectively when sufficient pixels are sampled from the input image.

Table III: Task Based Utility - Snow

δ	Pupil Confidence		Gaze Error ($^\circ$)	
	Snow	Snow-Med	Snow	Snow-Med
0.1	0%	0%	-	-
0.33	30%	0%	3.20	-
0.4	64%	2%	2.17	6.58
0.5	80%	50%	1.76	3.06
0.6	87%	82%	1.42	1.86
0.7	92%	94%	0.95	0.88
0.8	94%	99%	0.58	0.28

Table IV: Privacy Risk Measures

ϵ	CRR - CASIA			Re-ID - AT&T		
	DP-Pix	DP-Samp	DP-SVD	DP-Pix	DP-Samp	DP-SVD
0.01	0%	0%	0%	3%	0%	1%
0.05	0%	0%	0%	9%	3%	3%
0.1	0%	0%	0%	10%	11%	58%
0.5	0%	0%	0%	68%	35%	63%
1	0%	0%	0%	68%	51%	56%
5	0%	4%	0%	83%	77%	59%
10	0%	7%	0%	81%	83%	60%

Table V: Privacy Risk Measures - Snow

δ	CRR - CASIA		Re-ID - AT&T	
	Snow	Snow-Med	Snow	Snow-Med
0.1	0%	0%	4%	4%
0.33	0%	0%	13%	5%
0.4	0%	0%	23%	4%
0.5	0%	0%	75%	9%
0.6	1%	0%	86%	54%
0.7	5%	25%	85%	81%
0.8	10%	77%	91%	91%

2) *Privacy Risks*: We observe in Table IV that both DP-Pix, DP-Samp, and DP-SVD are shown to be resistant to iris authentication based attacks at all privacy levels, i.e., 0% CRR, while DP-Samp allows a small percentage of matches at high ϵ settings. It shows that the privacy perturbation inflicted by DP-Pix and DP-SVD successfully prevents the obfuscated image to be matched to existing templates. As DP-Samp outputs pixels from the real image, some obfuscated images may be matched at high ϵ settings. As seen in Table V, Snow leads to at most 10% CRR; however, applying median blur leads to up to 77% CRR at high δ settings. Since median blur can be applied by any adversary on the output image, we conclude that Snow does not provide strong defense against iris authentication attacks.

The face re-identification attack shows whether deep learning models can *adapt* to the evaluated obfuscation methods. It can be seen in Table IV that DP-Pix and DP-Samp lead to up to 83% re-identification rates at high ϵ settings, while DP-SVD inflicts lower risks despite the increase in ϵ . In Table V, we observe that Snow and Snow-Med lead to high re-identification risks, up to 91%. Even in lower δ settings (e.g., $\delta \leq 0.5$), Snow inflicts up to 75% re-identification risks due to the disclosure of real pixels.

3) *CPU Time*: We measure the runtime for each obfuscation method to sanitize a single image and the results are reported in Figure 9. In general, higher runtimes can be observed on the CASIA dataset for every method, due to a higher resolution. Across all methods, the privacy level (ϵ or δ) does not significantly affect the runtime performance of

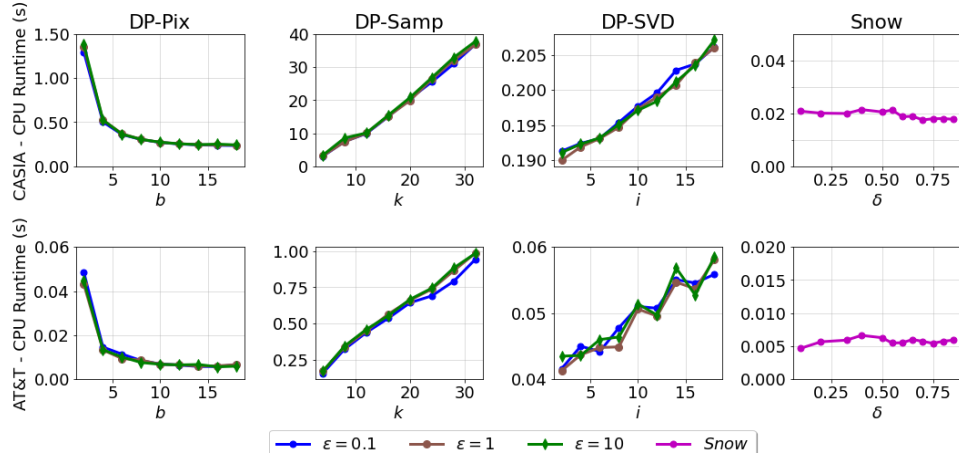


Fig. 9: Runtimes of DP-Pix, DP-Samp, DP-SVD, and Snow for sanitizing one image

the obfuscation. DP-Samp inflicts the highest runtimes among all methods, due to the computational costs inflicted by pixel clustering and sampling; a higher number of clusters (k) leads to higher runtimes. For DP-Pix, increasing the block width b reduces the runtime as perturbation is performed on a smaller number of blocks. For DP-SVD, increasing the number of eigenvalues i increases the runtime as the method conducts private sampling in a higher dimensional space.

E. Qualitative Evaluation

For qualitative evaluation, we present sample output images produced by the obfuscation methods with different parameter values and privacy levels in Figure 10. For CASIA dataset, DP-Samp introduces higher distortions to the eyes with $\epsilon = 0.5$, compared to other methods. DP-SVD does not capture detailed features, e.g., eyebrows, while DP-Pix and Snow introduce “salt and pepper” noise. Increasing ϵ to 1 reduces the distortions and perturbations in DP-Samp and DP-SVD, and improves the output quality of DP-Pix greatly. Applying median blur to Snow may aggravate the gray noise at low δ settings and help remove such noise at high δ settings.

For AT&T dataset, DP-Samp produces low quality output images among all methods. DP-SVD introduces distortions due to matrix singular value decomposition and private sampling, but distortions incurred by sampling are alleviated effectively by increasing ϵ to 1. DP-Pix outputs show the effects of pixelization and privacy perturbation, and the perturbation effect may be reduced by adopting a higher b value and/or a higher ϵ value. Due to a lower resolution, the outputs of Snow and Snow-Med are affected by the gray noise much more than for CASIA.

V. DISCUSSION

A. Interpreting the results

We have made several observations which may provide insights for adopting existing obfuscation methods and developing new image obfuscation methods.

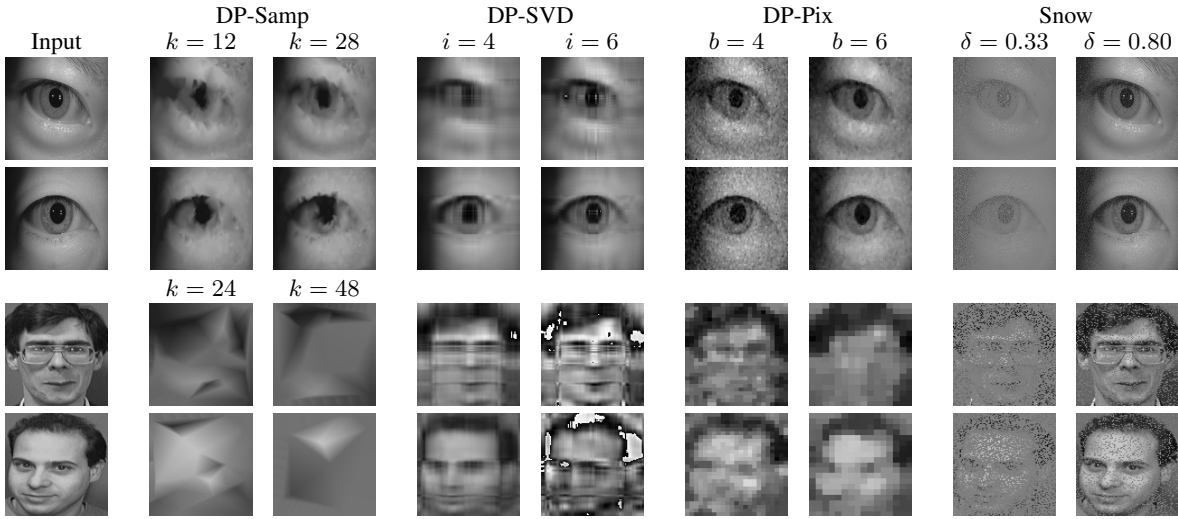
Firstly, although lower pixel-level errors, i.e., MSE, often lead to higher SSIM scores, it does not always hold for images with simpler structures, e.g., for CASIA eye images. Our

results in Figure 3c showed that DP-Samp inflicts MSE errors which are equal to or smaller than those of DP-Pix (Figure 3a) and DP-SVD (Figure 3e), while its SSIM scores are lower in most instances. For Snow, the MSE errors on CASIA monotonically decrease when increasing δ (see Figures 5a and 6a), while the SSIM scores (see Figures 5b and 6b) at $\delta = 0.1$ are higher than those of other δ values. The sample output images in Figure 10 show that the measures alone may not be sufficient to reflect the quality for images.

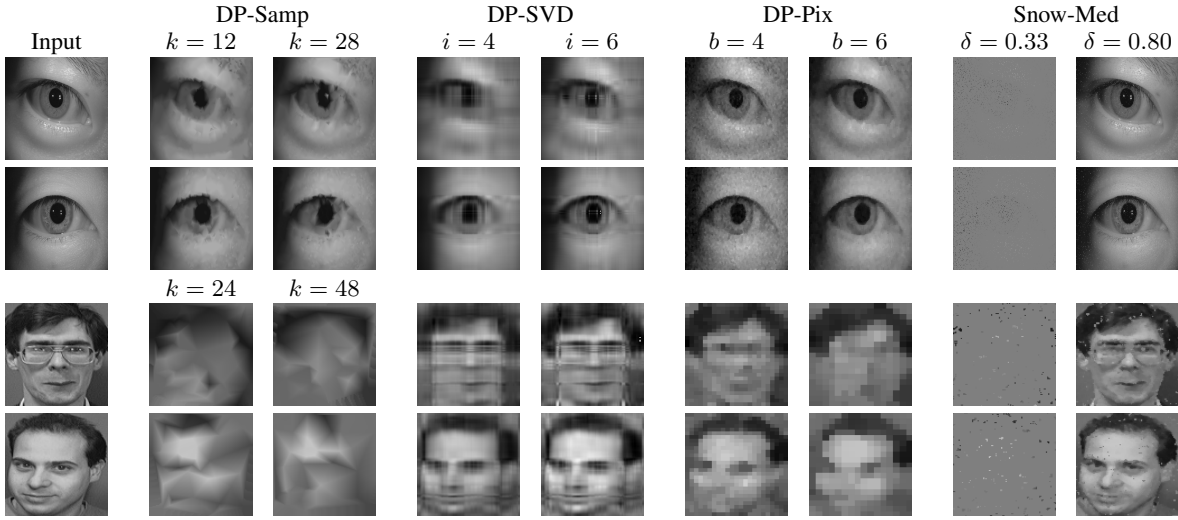
Secondly, we observe that low generic utility (MSE and SSIM) does not always lead to low task-based utility. For example, DP-Pix in low ϵ settings (e.g., $\epsilon \leq 0.1$) leads to high MSE errors and low SSIM scores, shown in Figures 3a and 3b; however, it provides higher pupil confidence (46%) and lower gaze errors (5.48°) than other methods when $\epsilon \leq 0.1$, as in Table II. It shows that pixelization-based obfuscation can provide usefulness for image applications, even at low ϵ settings.

Thirdly, we observe that the obfuscation methods exhibit distinct trade-off behaviors between privacy and task-based utility. As seen in Tables II and III, DP-Pix provides stable pupil confidence and gaze errors when increasing ϵ ; DP-Samp and DP-SVD show rapid improvements in those measures between $\epsilon = 0.1$ and $\epsilon = 0.5$; Snow gradually improves as δ increases. We believe that the design of the method plays an important role: DP-Pix employs pixelization, which inflicts a loss of detailed information independent of the privacy perturbation. DP-Samp and DP-SVD utilize the global structure/features in the input image, which may not be accurately captured at low ϵ settings. Snow outputs each pixel independently, which results in the gradual utility improvement by increasing the sampling probability.

Last but not least, we observe that although theoretical privacy guarantees often correlate with practical privacy protection, post-processing may change the level of practical privacy protection. In Table IV, we see that the practical privacy risks increase when ϵ increases for DP-Pix, DP-Samp, and DP-SVD. In Table V, we see that applying median blur after Snow significantly increases the CRR, e.g., from 10% to 77% for $\delta = 0.8$, although median blur does not weaken the



(a) $\epsilon = 0.5$ for DP-Pix, DP-Samp, and DP-SVD. Snow is not affected by ϵ .



(b) $\epsilon = 1$ for DP-Pix, DP-Samp, and DP-SVD. Snow-Med is not affected by ϵ .

Fig. 10: Sample output images produced by DP-Samp, DP-SVD, and DP-Pix at $\epsilon = 0.5$ and $\epsilon = 1$, as well as by Snow and Snow-Med (i.e., with median blur) at $\delta = 0.33$ and $\delta = 0.80$

differential privacy guarantees.

B. How to choose

It is important to recognize that applications may have different priorities and requirements for image obfuscation. There is no “one size fits all” solution. For applications requiring provable privacy guarantees, DP-Pix provides ϵ -DP guarantees and can be adapted to protecting more than one pixels in the input image. Both DP-SVD and Snow provide relaxed DP guarantees, i.e., metric privacy and (ϵ, δ) -DP, respectively. DP-Samp provides ϵ -DP for individual pixels in some steps but not in the pixel clustering step (similar to [25]). For applications requiring strong practical privacy protection, DP-Pix with $\epsilon \leq 0.1$ is a great option, thanks to low privacy risks in Table IV and much better utility measures compared to other methods in Table II. For applications wishing for a balance between privacy and utility, DP-SVD or DP-Samp with $\epsilon = 0.5$ may be considered: as shown in Table II, DP-SVD provides 79% pupil confidence and 2.58°

gaze error, while DP-Samp provides 40% pupil confidence and 5.66° gaze error; but DP-SVD leads to a higher Re-ID risk than DP-Samp in Table IV, i.e., 63% vs 35%. With even weaker privacy guarantees, applications may consider Snow with $\delta = 0.5$, which achieves 80% pupil confidence, 1.76° gaze error, and 75% Re-ID risk (see Tables III and V). Additional considerations such as computational resources and runtime requirements should also be taken to account.

C. Extension to Multi-Channel Images.

We evaluated current DP image obfuscation methods with grayscale image data, as in [7], [8], [13]. Considering image data with multiple channels, such as RGB (red-green-blue) and HSV (hue-saturation-value) images, each channel may not be independent of the other channels. A straight-forward extension of current image obfuscation is to apply the composition theorems [6] by splitting the privacy budget across multiple channels, as mentioned in [7]. An alternative idea is to treat each pixel as a vector, as in [25].

VI. CONCLUSION

We performed a comprehensive evaluation of four image obfuscation methods, namely DP-Pix, DP-Samp, DP-SVD, and Snow, which provide provable privacy guarantees based on differential privacy. We adopted real eye and face image datasets in the empirical evaluation and included both generic and task based utility measures as well as privacy risk measures against attacks. We found that DP-Pix achieves the best task-based utility in strong privacy settings (i.e., $\epsilon \leq 0.1$) with low empirical privacy risks, while providing pure ϵ -DP. For moderate to low privacy settings (i.e., $\epsilon \geq 0.5$), DP-SVD and DP-Samp provide a trade-off between privacy and utility, while DP-SVD achieves lower gaze errors and lower privacy risks even in low privacy settings (e.g., $\epsilon = 10$). Snow may achieve higher task-based utility at the cost of high practical privacy risks, e.g., 77% CRR for eye images.

The following directions may be considered for future work. 1) The development of new image obfuscation methods for specific domain applications: Snow and DP-SVD adopt different approaches to serve the target domains (i.e., eye-tracking applications and face images, respectively). Future work should take into account of the characteristics of the data and applications, in order to achieve high usefulness. 2) The development of new privacy risk measures: this study focused on *identity* based privacy attacks due to its sensitivity, while the disclosure of other types of information may also be considered, such as emotion, activity, etc. 3) The development of new obfuscation methods to produce more natural looking images: current methods may lead to visually unappealing artifacts. Future research may leverage latest machine learning techniques to generate natural looking images, while providing rigorous privacy guarantees.

ACKNOWLEDGMENT

This work has been supported in part by NSF CNS-1949217, CNS-1951430, CNS-2027114, and UNC Charlotte. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

- [1] <https://fan-group.github.io/imageprivacy/>.
- [2] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, CCS '13*, page 901–914, New York, NY, USA, 2013. Association for Computing Machinery.
- [3] K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, and C. Palamidessi. Broadening the scope of differential privacy using metrics. In E. De Cristofaro and M. Wright, editors, *Privacy Enhancing Technologies*, pages 82–102, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- [4] J. Daugman. How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier, 2009.
- [5] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- [6] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, Aug. 2014.
- [7] L. Fan. Image pixelization with differential privacy. In F. Kerschbaum and S. Paraboschi, editors, *Data and Applications Security and Privacy XXXII*, pages 148–162, Cham, 2018. Springer International Publishing.
- [8] L. Fan. Practical image obfuscation with provable privacy. *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 784–789, 2019.
- [9] A. Gangwar, A. Joshi, A. Singh, F. Alonso-Fernandez, and J. Bigun. Irisseg: A fast and robust iris segmentation framework for non-ideal iris images. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016.
- [10] Z. He, T. Tan, Z. Sun, and X. Qiu. Toward accurate and fast iris segmentation for iris biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1670–1684, 2009.
- [11] S. Hill, Z. Zhou, L. Saul, and H. Shacham. On the (in)effectiveness of mosaicing and blurring as tools for document redaction. *Proceedings on Privacy Enhancing Technologies*, 2016, 02 2016.
- [12] B. John, S. Jörg, S. Koppal, and E. Jain. The security-utility trade-off for iris authentication and eye animation for social virtual avatars. *IEEE Transactions on Visualization and Computer Graphics*, 26:1880–1890, 2020.
- [13] B. John, A. Liu, L. Xia, S. Koppal, and E. Jain. Let it snow: Adding pixel noise to protect the user’s identity. In *ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Adjunct*, New York, NY, USA, 2020. Association for Computing Machinery.
- [14] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi. Traffic monitoring and accident detection at intersections. *IEEE Transactions on Intelligent Transportation Systems*, 1(2):108–118, 2000.
- [15] S. S. Kozat, R. Venkatesan, and M. K. Mihcak. Robust perceptual image hashing via matrix invariants. In *Image Processing, 2004. ICIP '04. 2004 International Conference on*, volume 5, pages 3443–3446 Vol. 5, Oct 2004.
- [16] L. Masek. *Recognition of Human Iris Patterns for Biometric Identification*. PhD thesis, The University of Western Australia, 2003.
- [17] R. McPherson, R. Shokri, and V. Shmatikov. Defeating image obfuscation with deep learning. *CoRR*, abs/1609.00408, 2016.
- [18] S. Park, X. Zhang, A. Bulling, and O. Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. *ETRA '18*, New York, NY, USA, 2018. Association for Computing Machinery.
- [19] A. Reece and C. Danforth. Erratum to: Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6, 12 2017.
- [20] Z. Ren, Y. J. Lee, and M. S. Ryoo. Learning to anonymize faces for privacy preserving action detection. *CoRR*, abs/1803.11556, 2018.
- [21] F. Samaria and A. Harter. Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pages 138–142, 1994.
- [22] R. Stevens and I. Pudney. Blur select faces with the updated blur faces tool. <https://youtube-eng.googleblog.com/2017/08/blur-select-faces-with-updated-blur.html>, August 2017. [Online; posted 21-August-2017].
- [23] D. Su, J. Cao, N. Li, E. Bertino, M. Lyu, and H. Jin. Differentially private k-means clustering and a hybrid approach to private optimization. *ACM Transactions on Privacy and Security (TOPS)*, 20(4):1–33, 2017.
- [24] Q. Sun, L. Ma, S. J. Oh, L. Van Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. pages 5050–5059, 06 2018.
- [25] H. Wang, S. Xie, and Y. Hong. Videodp: A flexible platform for video analytics with differential privacy. *Proc. Priv. Enhancing Technol.*, 2020(4):277–296, 2020.
- [26] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [27] Y. Yang, J. Liu, and M. Shah. Video scene understanding using multi-scale analysis. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1669–1676. IEEE, 2009.
- [28] Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Ophey, V. L. Flanagan, P. zu Eulenburg, and S.-A. Ahmadi. Deepvov: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of neuroscience methods*, 2019.
- [29] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015.