DP-Shield: Face Obfuscation with Differential Privacy

Muhammad Usama Saleem
UNC Charlotte
Charlotte, North Carolina
msaleem2@uncc.edu

Dominick Reilly
UNC Charlotte
Charlotte, North Carolina
dreilly1@uncc.edu

Liyue Fan
UNC Charlotte
Charlotte, North Carolina
liyue.fan@uncc.edu

ABSTRACT

An immense amount of image data is captured and shared nowadays, e.g., social media and surveillance databases. Such image data may contain sensitive information, such as faces, which can be misused if in the hands of an adversary. Widely used image privacy solutions obfuscate faces, e.g., via pixelization and blurring, before sharing with untrusted parties. However, they do not provide quantifiable privacy guarantees and are prone to inference attacks. In this demo, we present DP-Shield, an interactive framework for face image obfuscation under the rigorous notion of differential privacy. DP-Shield showcases our recently proposed obfuscation methods, namely DP-Pix and DP-SVD, and also includes two alternative methods for comparison. The audience will be able to learn about existing DP methods by interacting with them using real-world face image datasets. Furthermore, DP-Shield integrates widely used image quality measures and practical privacy risk measures (i.e., face recognition) to illustrate the efficacy of our methods.

1 INTRODUCTION

An immense amount of image data is generated from a variety of sources, such as social medial platforms and surveillance cameras. The wide release of such data would greatly benefit society, e.g., advancing computer vision research and applications. 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 social media platforms and surveillance cameras may expose human faces, which adversaries may use to track or profile an individual [7, 10]. To preserve privacy, image data must be obfuscated before sharing with untrusted parties.

Widely used face obfuscation techniques include pixelization [11] and blurring [14]. However, deep convolutional neural networks (CNNs) are successful at re-identifying faces obfuscated with those techniques [11]. Furthermore, standard obfuscation approaches do not allow privacy to be effectively bounded. In other words, they do not quantify the sensitive information that may be leaked in the obfuscated image. In light of those limitations, recent approaches [5, 6] adopted the notion of differential privacy [4] (DP) for image obfuscation. DP-Pix [5] is the first approach to extend DP to individual-level image publication, which defines neighboring images for content protection and reduces sensitivity via pixelization. DP-SVD [6] adopts metric privacy [3] and provides indistinguishability based on perceptual image features. Furthermore, DP-SVD designs a novel sampling approach in high-dimensional spaces to achieve privacy.

In this study, we present DP-Shield, an interactive framework demonstrating DP-Pix and DP-SVD for face image obfuscation. For comparison, we include two alternative methods, namely DP-Samp [15] and Snow [8]. As shown in Figure 1, DP-Shield applies differentially private image obfuscation to an input image and evaluates the obfuscated image for quality and practical privacy risk. Image quality is measured using Mean Squared Error (MSE) and Structural Similarity (SSIM). Practical privacy risk is measured via state-of-the-art face recognition techniques [13] that are trained on publicly available datasets. DP-Shield shows great promise for enhancing the privacy of image data that is shared with a wide audience.

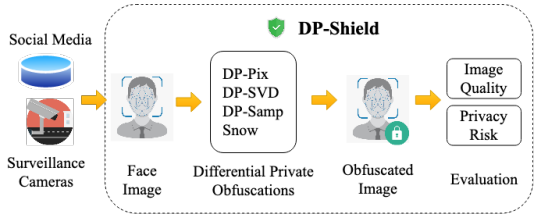


Figure 1: DP-Shield Framework

Compared with our recent work [12], DP-Shield presents the following contributions: **(1)** Users can interact with DP image obfuscation methods by varying privacy and algorithm-specific parameters. They can observe intermediate results as well as final outputs of the DP methods, which help them understand the key steps through the quantitative evaluation results. **(2)** DP-Shield focuses on face images, which are considered to be highly sensitive. To that end, this study adopts two large scale real-world face image datasets, namely VGGFace2 and CASIA-WebFace, to illustrate the feasibility of DP image obfuscation methods in real world applications. **(3)** We assess the privacy risk of sharing face images by applying state-of-the-art face recognition techniques [13] before and after applying privacy and evaluating the difference. Our approach simulates practical privacy attacks, where an adversary attempts to re-identify the obfuscated face image using public face recognition models.

The rest of the paper is organized as follows: Section 2 introduces DP-Shield and DP image obfuscation methods; Section 3 discusses the empirical evaluation conducted using real-world face datasets; Section 4 describes how the audience can interact with our demonstration; Section 5 concludes the paper and discusses directions for future work.

2 DP-SHIELD OVERVIEW

As shown in Figure 1, the core of DP-Shield is the set of differentially private image obfuscation methods. After an input image is obfuscated, DP-Shield also evaluates the image quality based on widely adopted metrics as well as practical privacy risks using state-of-the-art techniques for face recognition.

2.1 Differential Privacy Preliminaries

Differential privacy [4] is the state-of-the-art notion for quantifying privacy leakage in statistical databases. Differential privacy allows the publication of aggregate statistics about the input database via a randomized algorithm $\mathcal{M}$, so that the output of $\mathcal{M}$ remains roughly the same even if any record in the input is arbitrarily modified. Given the output of $\mathcal{M}$, an adversary will

not be able to infer much about any individual record in the input, hence protecting privacy. More formally, given any neighboring databases $\mathcal{D}$ and $\mathcal{D}'$ that differ by at most one record, a randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for any $Z \subset range(\mathcal{M})$, $\Pr[\mathcal{M}(D) \in Z] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D') \in Z] + \delta$. The parameters $\epsilon > 0$ and $\delta \in [0, 1]$ specify the degree of privacy provided by the mechanism. Smaller ϵ and δ values indicate stronger privacy protection, and vice versa.

Metric-based privacy [3] extends differential privacy to a set of secrets $\mathcal{X}$ equipped with a distance metric, i.e., $d_{\mathcal{X}}$, and guarantees a level of indistinguishably that is proportional to the distance between secrets. Adopted by DP-SVD [6], metric privacy allows the publication of individual-level data while providing provable privacy guarantees. When $d_{\mathcal{X}}$ is the Hamming distance for databases, metric privacy is equivalent to differential privacy [3].

2.2 Image Obfuscation Methods

DP-Shield presents two innovative methods which apply the principle of differential privacy to image obfuscation.

DP-Pix. Differentially private pixelization (DP-Pix) [5] was the first approach to provide differential privacy guarantees in the publication of individual-level image data. The novelty of DP-Pix is in adapting the notion of neighboring databases to the image domain through the definition of m-neighborhood.

Definition 2.1. (m-neighborhood [5]) Two images are considered neighboring if they have the same dimension and differ at most by m pixels.

Under this definition, the presence of any content represented by up to m pixels, e.g., persons or objects, is protected by differential privacy. The data owner can specify the value of m, where larger values indicate stronger privacy.

Another key idea of DP-Pix is to adopt pixelization in order to address the high sensitivity in image publication. A grid is superimposed onto the input image where each grid cell is of size $b \times b$ pixels. The average pixel value in each grid cell is reported in a differentially private manner via the classic Laplace mechanism [4]. Specifically, let I denote the input image and $P_b(I)$ denote the pixelization with parameter b. DP-Pix generates the differentially private pixelization $\tilde{P}_b(I) = P_b(I) + \text{Laplace}(\frac{255\text{m}}{\text{b}^2\epsilon})$, where $\text{Laplace}(\frac{255\text{m}}{\text{b}^2\epsilon})$ represents i.i.d. random noise drawn from a Laplace distribution with 0 mean and $\frac{255m}{b^2\epsilon}$ scale. It can be shown that DP-Pix satisfies ϵ-DP [5].

DP-SVD. DP-SVD [6] adopts the metric privacy paradigm [3] with the goal of providing indistinguishability based on perceptual image features. Unlike DP-Pix, which achieves differential privacy by directly perturbing super-pixels, the DP-SVD mechanism perturbs perceptual features derived from the input image. DP-SVD first applies singular value decomposition (SVD) to an input image and achieves metric privacy by perturbing i highest singular values. The perturbation noise is drawn using a novel sampling method, which is applicable in high-dimensional spaces. Specifically, in 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 [6]:

$$D_{\epsilon,i}(x_0)(x) = C_{\epsilon,i} e^{-\epsilon \cdot d_i(x_0,x)}, \quad C_{\epsilon,i} = \frac{1}{2}\left(\frac{\epsilon}{\sqrt{\pi}}\right)^i \frac{\left(\frac{i}{2}-1\right)!}{(i-1)!} \quad (1)$$

where d_i represents i-dimensional Euclidean distance and i is assumed even without loss of generality. The parameter i is a chosen by the user. Increasing i may result in a higher utility and better approximation of the input image but may require a large amount of noise to achieve privacy.

Other Methods. We also consider two alternative methods which provide weaker privacy guarantees. The Snow [8] method employs pixel-level noise by arbitrarily re-assigning pixel intensities to a constant value, i.e., 127 for grayscale images. The method adopts a parameter p which denotes the percentage of altered pixels. Snow achieves $(0, \delta)$-DP with $\delta = 1 - p$ and protects individual pixels in the input image, i.e., a special case with $m = 1$ in Def. 2.1. DP-Samp is our adaptation of the video sanitization method from [15] to individual images. It subsamples pixels with the most useful values for reconstructing the input image, i.e., top k pixel values. Our adaptation modifies the sampling constraints to protect up to m pixels in the input, similar to DP-Pix. More technical details about DP-Samp can be found in [12]. Although the authors of [15] provided an analysis of the ϵ-DP guarantee, a pixel candidate generation step was performed in the *public* setting, hence providing weakened privacy protection.

3 EXPERIMENTS

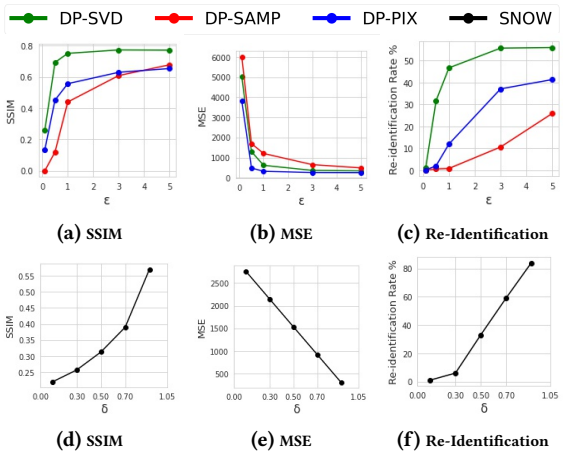
3.1 Methodology

We conduct empirical analysis on the image obfuscation methods to present the audience with a comparative evaluation. Two widely used face datasets are adopted: VGGFace2 [2] and CASIA-WebFace [16]. VGGFace2 contains 3.31 million images of 9131 subjects. CASIA-WebFace contains 494,414 face images of 10,575 subjects. We detect the facial region in each input image using MTCNN and convert it to grayscale with a standard resolution of 160×160 pixels. The default parameter values used are: $b = 4$ for DP-Pix, $i = 6$ for DP-SVD, $k = 48$ for DP-Samp, and $m = 1$ for both DP-Pix and DP-Samp.

To measure image quality, we adopt Mean Squared Error (MSE) and Structural Similarity (SSIM) as in previous works [5, 6]. Both measures are computed between the clean and obfuscated images, and the average among each dataset is reported. Lower MSE indicates higher quality; higher SSIM indicates higher quality. To measure practical privacy risks, we measure the accuracy of re-identification using state-of-the-art face recognition techniques, e.g., FaceNet with the Inception ResNet (V1) network [13]. Our rationale is that an adversary may have access to publicly available images of individuals and thus can deduce the identity of an obfuscated facial image using well-known models. Specifically, for both datasets, we randomly sample 1000 individuals for the study and 20 images per individual, partitioned between training and testing (15 : 5). An SVC classifier is trained on the FaceNet embeddings of the training partition, and we report its accuracy on the testing partition. For non-private grayscale images, the re-identification accuracy is 94.18% for VGGFace2 and 82.72% for CASIA-WebFace.

3.2 Quality vs. Privacy Results

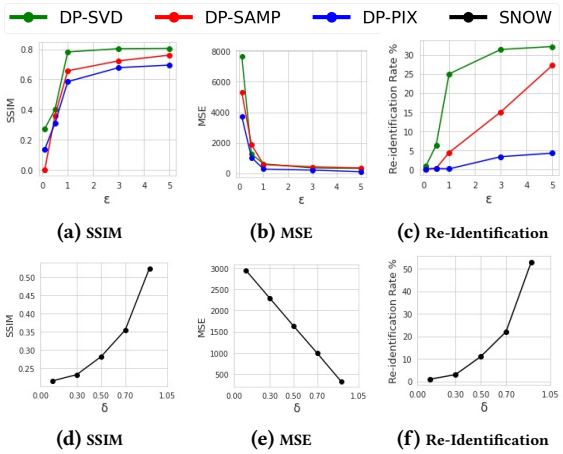
We report MSE, SSIM, and face re-identification accuracy for VGGFace2 and CASIA-WebFace in Figure 2 and Figure 3, respectively. The privacy parameters ϵ and δ indicate the level of privacy protection, and lower values indicate stronger privacy protection. While larger ϵ values have been used, e.g., in [1], our study focuses on $\epsilon \leq 5$. In classic DP [4], it is recommended to set $\delta = \frac{1}{|D|}$ to protect each record in input. However, the evaluation of Snow focuses on $\delta \geq 0.1$, as lesser values of δ give little to no



(a) SSIM (b) MSE (c) Re-Identification

(d) SSIM (e) MSE (f) Re-Identification

Figure 2: Mean Squared Error (MSE), Structural Similarity (SSIM) and Re-Identification Rate results on VGGFace2 dataset.



(a) SSIM (b) MSE (c) Re-Identification

(d) SSIM (e) MSE (f) Re-Identification

Figure 3: Mean Squared Error (MSE), Structural Similarity (SSIM) and Re-Identification Rate results on CASIA-WebFace dataset.

utility. As we increase ϵ or δ, we observe that MSE decreases and SSIM increases for all DP methods, and the re-identification rate increases as well. These results confirm that all methods exhibit a trade-off between utility and privacy.

In comparison to other methods, Snow leads to a higher re-identification risk (with higher δ values in both Figure 2 and Figure 3), indicating lower empirical privacy protection. DP-SVD yields high image quality, illustrated by the highest SSIM scores and lower MSE errors, but also leads to higher re-identification rates. Both DP-Pix and DP-Samp yield lower image quality and lower privacy risks in comparison to DP-SVD. Those results may be used by data owners to fine tune the DP parameters, in order to achieve acceptable image quality and privacy risk levels.

We observe that DP-Samp leads to the lowest image quality and re-identification rate in VGGFace2 (Figure 2). Through the analysis of Sobel gradients [9], we find that images in this dataset contain more details. As a result, the pixel subsampling approach incurs higher information loss. We believe that the algorithm-specific parameters, i.e., b, i, and k, may be tuned for each dataset to find an optimal trade-off between privacy and utility.

4 DEMONSTRATION PLANS

DP-Shield is available at https://fan-group.github.io/imageprivacy/. Three interactive use cases on real-world datasets will be demonstrated: (1) exploration of different DP methods for sanitizing images; (2) exploration of the effects of DP methods on image embeddings; (3) the quantitative and qualitative evaluation of DP methods.

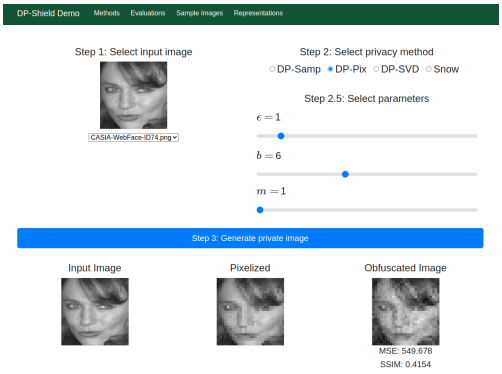


Figure 4: Exploring DP Methods

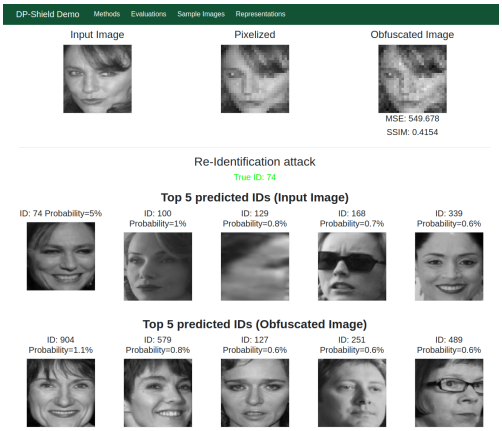


Figure 5: Re-Identification with Clean vs. Obfuscated Images

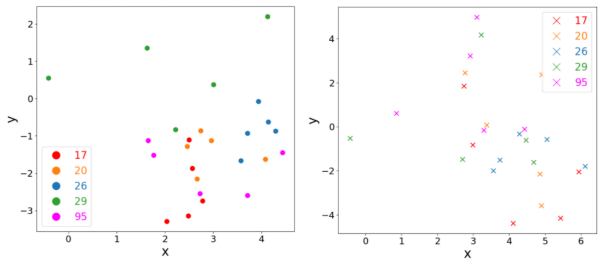
4.1 Exploring DP Methods

The "Methods" page illustrates how each DP method obfuscates images with given privacy parameters. As shown in Figure 4, the audience can select an input image from VGGFace2 or CASIA-WebFace. The audience can then select a DP method and modify its parameter(s) to see how the output image, as well as the quality measures, may vary. To help the audience better understand the DP methods, we also display the intermediate results of the methods. For example, in Figure 4 we show the pixelized image before applying Laplace noise in DP-Pix.

In addition, the audience will learn about the practical privacy protection offered by DP methods against facial recognition attacks (described in Section 3.) Specifically, as in Figure 5, we show the top 5 most likely identities predicted by the facial recognition model using the clean image vs. using the obfuscated image, along with the confidence of each predicted identity. The audience will observe that performance of facial recognition yields higher uncertainty (i.e., lower confidence) on the obfuscated image when compared to the non-private image. This illustrates that DP image obfuscation methods are effective at deterring facial recognition based attacks.

4.2 Facial Image Representations

The "Representations" page illustrates how neural representations of face images (i.e., FaceNet embeddings) vary under different DP methods and parameters. Similar to the “Methods” page, the audience can select a DP method and parameter values to



(a) Clean Images (b) Obfuscated Images

Figure 6: Representation for Clean vs. Obfuscated Images with DP-Pix with $\epsilon = 1$, $b = 6$ and $m = 1$ (best viewed in color)

observe their effects on the image representations. To visualize high-dimensional embeddings in 2D-space in a consistent manner, a Linear Discriminant Analysis (LDA) model is trained on the training partition of each dataset and applied to the clean and obfuscated images in the test partition.

Figure 6 shows the projections of the clean and obfuscated images for 5 individuals in VGGFace2, where the DP-Pix method was applied to generate the obfuscated images. We observe that privacy methods may introduce distortions in the neural representations of face images. On one hand, images of the same individual may be further apart after applying privacy methods, e.g., ID 20 in Figure 6. On the other hand, images of different individuals may not be separated after applying privacy methods, e.g., IDs 17 and 26 in Figure 6. The visualization demonstrates to the audience that DP methods can introduce uncertainty into popular facial recognition techniques and hence offer protection against practical privacy attacks.

4.3 Quantitative and Qualitative Evaluation

The audience can navigate to the "Evaluations" page to view a quantitative evaluation of each evaluated DP method on the VGGFace2 and CASIA-WebFace datasets. For an interactive and comparative evaluation, aggregated results for each dataset (using the same data as in Figure 2 and Figure 3) are shown in colored column charts.

The audience can also observe sample output images of the DP methods on various ϵ and δ values by navigating to the "Sample Images" page, as shown in Figure 7. Increasing the privacy parameter values decreases distortions or disturbances in the output of all methods, resulting in better image quality. Compared to other methods, DP-Samp yields lower image quality, even with higher ϵ values.

5 CONCLUSION

We demonstrated DP-Shield, a face image obfuscation framework that provides differential privacy guarantees. We described the differentially private image obfuscation methods as well as the evaluation methodology using large real-world datasets. DP-Shield conveys the aggregated evaluation results to the audience for a comparative evaluation among the DP methods. Furthermore, it allows the audience to interact with each method, learning to interpret the image quality measures and practical privacy risks. Future work may include open-sourcing the DP image obfuscation methods, releasing the face re-identification attacks as a standard test set, and improving the usability of DP-Shield for mobile applications.

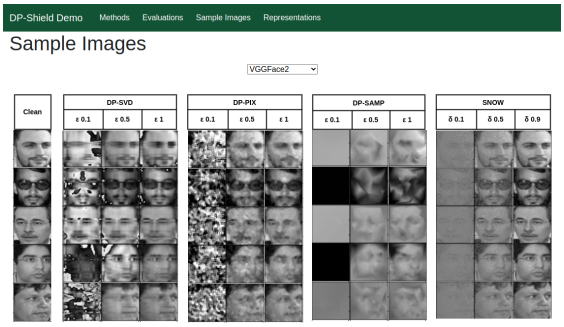


Figure 7: Qualitative Evaluation

ACKNOWLEDGMENTS

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] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. 2018. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE Computer Society, Los Alamitos, CA, USA, 67–74. https://doi.org/10.1109/FG.2018.00020

[3] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the Scope of Differential Privacy Using Metrics. In *Privacy Enhancing Technologies*, Emiliano De Cristofaro and Matthew Wright (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 82–102.

[4] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4 (Aug. 2014), 211–407. https://doi.org/10.1561/0400000042

[5] Liyue Fan. 2018. Image Pixelization with Differential Privacy. In *Data and Applications Security and Privacy XXXII*, Florian Kerschbaum and Stefano Paraboschi (Eds.). Springer International Publishing, Cham, 148–162.

[6] Liyue Fan. 2019. Practical Image Obfuscation with Provable Privacy. *2019 IEEE International Conference on Multimedia and Expo (ICME)* (2019), 784–789.

[7] Mislav Grgic, Kresimir Delac, and Sonja Grgic. 2011. SCface–surveillance cameras face database. *Multimedia tools and applications* 51, 3 (2011), 863–879.

[8] Brendan John, Ao Liu, Lirong Xia, Sanjeev Koppal, and Eakta Jain. 2020. Let It Snow: Adding Pixel Noise to Protect the User's Identity. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 43, 3 pages. https://doi.org/10.1145/3379157.3390512

[9] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. 1988. Design of an image edge detection filter using the Sobel operator. *IEEE Journal of solid-state circuits* 23, 2 (1988), 358–367.

[10] Siyuan Liu, Shuhui Wang, Feida Zhu, Jinbo Zhang, and Ramayya Krishnan. 2014. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*. 51–62.

[11] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. 2016. Defeating Image Obfuscation with Deep Learning. arXiv:cs.CR/1609.00408

[12] Dominick Reilly and Liyue Fan. 2021. A Comparative Evaluation of Differentially Private Image Obfuscation. *IEEE International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications* (12 2021).

[13] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. *CoRR* abs/1503.03832 (2015). arXiv:1503.03832 http://arxiv.org/abs/1503.03832

[14] Ryan Stevens and Ian Pudney. 2012. *Blur select faces with the updated blur faces tool.* https://youtube-eng.googleblog.com/2017/08/blur-select-faces-with-updated-blur.html

[15] Han Wang, Shangyu Xie, and Yuan Hong. 2020. VideoDP: A Flexible Platform for Video Analytics with Differential Privacy. *Proc. Priv. Enhancing Technol.* 2020, 4 (2020), 277–296. https://doi.org/10.2478/popets-2020-0073

[16] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. Learning Face Representation from Scratch. *CoRR* abs/1411.7923 (2014). arXiv:1411.7923 http://arxiv.org/abs/1411.7923