StablePrivacy: Diffusion-Based Privacy Enhancement for Face Image Datasets

Andreas Leibl[1, 2]
andreas.leibl@unibw.de
Helmut Mayer[1]

[1] Universität der Bundeswehr München
Munich, Germany

[2] Zentrale Stelle für Informationstechnik
im Sicherheitsbereich (ZITiS)
Munich, Germany

Abstract

Progress in deep learning-based computer vision has been significantly accelerated by the surge in available datasets for training and testing. However, the failure to meet ethical and regulatory standards in datasets containing privacy-sensitive content such as facial images has caused public concern and even led to the withdrawal of datasets. While traditional anonymization strategies, such as pixelization, offer a seemingly straightforward solution, they lack the ability to maintain the necessary facial details crucial for applications like training face detection models. To reconcile the need for high-quality data with stringent privacy standards, we explore an innovative method for de-identification that employs Stable Diffusion using synthetically generated faces as image prompts alongside a noisy version of the original face to guide anonymization, which we term StablePrivacy. Our experiments demonstrate the capability to preserve detailed features for training high-quality face detection models while offering state-of-the-art privacy protection.

1 Introduction

One of the factors driving the recent fast-paced development of deep learning-based computer vision is the increasing availability of ever-larger datasets. As in many related fields, they have helped to advance the state of the art for tasks like face recognition or detection [15, 48, 51]. However, unlike models used for other objectives, models for these tasks require huge amounts of privacy-sensitive face images. Current practices of collecting and storing this material often violate regulations, such as the GDPR in Europe [47] and the CCPA in California [34] or at least raise ethical concerns [1]. To tackle these legal and ethical challenges, sensitive data like face images should be protected appropriately prior to publication.

There are several techniques to achieve this [29, 37, 41]. The most straightforward is to remove the sensitive information directly on the image level by pixelization, blurring, or similar methods. These standard practices have been applied to many datasets [3, 45]. However, they obfuscate the face, compromising the usefulness of the data for training or benchmarking of many deep learning-based models [16, 22, 24].

More recent approaches try to achieve face de-identification by replacing the original face

with a synthetic surrogate [17, 18, 21, 25, 28]. The advantage of this is that the utility for downstream tasks can be better preserved by retaining a face-like appearance or even more detailed attributes such as expression, gender or exact head position. It has been shown that such approaches can limit the negative effect of anonymizing training data for face detection [22] or instance segmentation [21] on performance. Therefore, they achieve a better trade-off between privacy and data utility retention [25].
However, existing approaches either offer comparatively weak protection against automatic recognition [17, 18, 21, 25] or significantly degrade image quality [28] affecting data utility for downstream applications.
To address these shortcomings, we explore the use of Stable Diffusion [39] guided by synthetically generated faces, which are not privacy-sensitive, as image prompts and a noisy version of the original image. By this means, our method can create high-quality images as it can access rough structural information of the original image without the need for completely occluding the face region before processing, which is typical for other approaches. Notably, at the same time, it does not compromise privacy as demonstrated by state-of-the-art performance for protection against recognition on the Labeled Faces in the Wild (LFW) [15] benchmark. In summary, we make the following contributions:

- To the best of our knowledge, we are the first to explore the application of a latent diffusion model guided by synthetically generated source faces as image prompts and a noisy version of the original for de-identification.
- We develop strategies such as automated source selection and dynamic anonymization depending on face size, to tailor our approach to the task of de-identifying large and diverse datasets, thereby improving privacy protection and processing speed.
- We demonstrate superior protection against automatic recognition on the LFW benchmark compared to existing approaches.
- We show that a high-quality deep learning-based face detection model can be trained on our anonymized data, outperforming models trained on data de-identified by existing approaches.

2 Related Work

Privacy Protection in Image Datasets. Obfuscation techniques, like blurring or pixelization, are probably the most widely used techniques to anonymize faces in images. They have been applied to many datasets to improve privacy and regulatory compliance [3, 45]. While they are simple to use and their computational cost is low, they significantly reduce data utility by obscuring parts of the image, making the anonymized data less useful for training and testing deep learning-based algorithms. Lee *et al.* [24] show that simply blurring faces in training data not only reduces the performance of resulting segmentation models on the anonymized classes, i.e., person, but also on those occurring concurrently with them (motorcycle, backpack, and others). The same is true for other tasks that depend on detailed depictions of the face to learn their objective such as face detection [22]. More recently, synthesis-based de-identification approaches have emerged as an alternative, promising a better privacy–data utility trade-off. They synthesize image data which is used to replace the original face image.

Synthetic Face Generation. The foundation of synthesis approaches for privacy enhancement is derived from significant advances in generative machine learning methods like Generative Adversarial Networks (GANs) [11] or diffusion models [14, 44].
Recently, diffusion models [14, 44] have been shown to beat GANs concerning image quality and diversity [7]. Nevertheless, the use of diffusion models for de-identification is underexplored. The basic idea of diffusion is to learn to reverse the gradual addition of noise to images. This is done using an approximated loss function for training the model derived from the variational lower bound of the log-likelihood. Empirically, Ho *et al.* [14] found that simplifying it as a mean squared error objective delivers good results:

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,x,\varepsilon}\left[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2\right]. \tag{1}$$

Since then, several improvements have been suggested. Nichol and Dhariwal [33] introduced classifier guidance which improves the image quality over class-conditioned diffusion models. Later, Ho and Salimans [13] proposed classifier-free guidance, simplifying guidance by removing the dependence on a separate classifier.
Building on these advances, Stable Diffusion is a state-of-the-art open-source large-scale text-to-image model [39]. Its main improvement over its predecessors, such as DALL-E [36], is the usage of a pretrained autoencoder to downsample the input to the latent space before passing it through the diffusion process. This allows the model to focus on the semantics of the data rather than barely perceptible details contained in the pixel space. It also means that training can happen on a much lower dimension making it computationally more efficient.
Our approach to improving privacy in image datasets is based on Stable Diffusion. Yet, as we will show in Sec. 3, two more recently developed modifications are necessary for de-identification. One is IP-Adapter [50], adding image prompt capabilities to Stable Diffusion. It consists of a pretrained image encoder (CLIP [35]) which is applied to the prompt image, a small projection network to convert the embedding into the required dimensionality and additional cross-attention layers to pass the image prompt into Stable Diffusion's U-Net [40] via decoupled cross-attention. Only the projection network and the added cross-attention layers need to be trained, making this a compute-efficient way to adjust Stable Diffusion to specific needs.
The other modification is SDEdit [32], allowing for guided image editing by first perturbing the image with Gaussian noise and then using the standard reverse diffusion process. As the input image is not converted to random noise but only distorted to a certain degree, the output is guided by the rough structures of the input image. The degree to which the original image is perturbed depends on the strength parameter, which can range from zero to one. The combination of Stable Diffusion, IP-Adapter and SDEdit has, to the best of our knowledge, not been explored for de-identification yet.
Synthesis-Based De-Identification. Instead, most previous synthesis-based approaches such as CIAGAN [28], DeepPrivacy [18] or Leibl *et al.* [25] rely on GANs. Both CIAGAN and DeepPrivacy first obscure the facial region of an image and then use either key points (DeepPrivacy) or detailed facial landmarks (CIAGAN) to guide the inpainting of the face. Leibl *et al.* also rely on detailed landmarks to guide the anonymization process. In contrast to all of these approaches, we do not use landmarks or key points. Instead, we employ a noisy version of the input image as guidance.
LDFA [21] uses a diffusion model similar to our approach. Yet, as it does not utilize synthetic source faces to steer the process, it can only achieve comparatively weak anonymization.
Differentiating Synthesis-Based De-Identification from Face Swapping. Though techni-

cally similar, synthesis-based de-identification and face swapping are optimized for different objectives. Face swapping is designed to create realistic images where the source person's face replaces that of the target person. These approaches typically do not include mechanisms to prevent the recovery of the target person's identity. Their unlimited access the target person's features can lead to identity leakage [25]. Moreover, face swapping typically manipulates a small number of images or videos, which makes it feasible for a user to hand-pick source and target to achieve the most realistic results. In comparison, de-identification focuses on anonymizing the targets in large-scale datasets, requiring a fully automated process and balancing image quality and data utility with privacy protection.

3 Method Overview

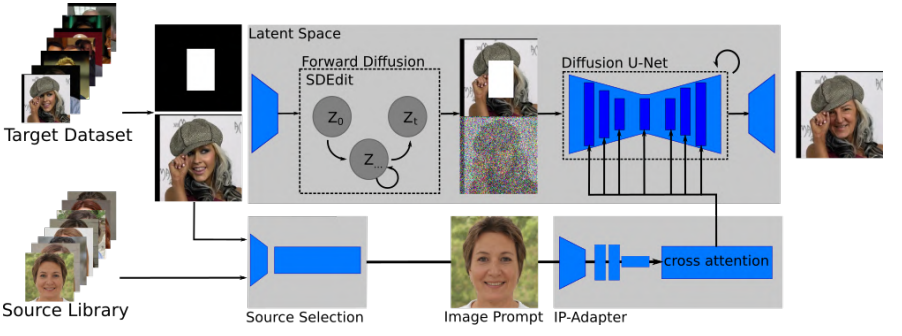

Figure 1: Visual overview of StablePrivacy. It is designed to anonymize images from a target dataset while preserving their utility. The process begins by masking the facial region of the original image. Both the original and the masked images are then converted into the latent space of Stable Diffusion. There, the previously unaltered image gets superimposed with Gaussian noise, ensuring the underlying structure remains vaguely intact with a shortened forward diffusion process (strength = 0.7) using the SDEdit [32] technique. The masked and noisy versions of the original are then passed to Stable Diffusion's U-Net for reverse diffusion. At the same time, a source image is selected from a library of synthetic face images ensuring a minimum distance to the target face in face similarity space. The source, which is not privacy-sensitive, is forwarded to IP-Adapter and used as an image prompt providing further guidance. With these inputs, the Stable Diffusion component generates the de-identified image.

In this section, we explain the details of our approach for synthesis-based face de-identification which we term StablePrivacy. It leverages Stable Diffusion for generating realistic inpaintings of the facial region of images we want to anonymize. This process is guided by a noise-modified version of the original image, employing the SDEdit technique, along with a source face image passed to Stable Diffusion using the IP-Adapter. A graphical overview can be seen in Fig. 1.

The source images serve as prompts that direct Stable Diffusion to create faces distinctly different from those in the original, effectively preventing the replication of identifiable features of the original. Prioritizing privacy protection, our approach employs sources synthetically generated by StyleGAN2 [20], ensuring that no real individuals' privacy is compromised.

For this reason, we compiled a manually curated dataset – termed the source library – containing 1,000 faces, enabling the generation of a diverse range of outputs. We discarded faces if a significant portion was covered by hair, glasses or other occlusions and balanced the gender distribution by manually choosing 500 female and 500 male images. Then, we adjusted the Euler Angles of the heads to approximately 0 degrees [30], ensuring that the image prompts contain enough relevant features to guide the transfer of the source's identity to the output.

To prevent our method from using source faces too closely resembling the original, it includes a safeguard automatically choosing a source from our library that is sufficiently different from the target. This selection process is based on maintaining a minimum distance in the face similarity space between the source and the target, as calculated using FaceNet [42]. An empirically determined threshold distance of 1.6 is applied across all experiments.

After an appropriate source is chosen, it is passed to IP-Adapter's projection network and finally into the attention layers of Stable Diffusion's U-Net using decoupled cross-attention. The degree to which this input influences the output image is controlled by the classifier-free guidance (CFG) scale.

To delineate the area for modification, our method utilizes a bounding box around the face which is given by a face detection model or dataset-provided ground truth. The exact size of the inpainting area can be adjusted by an additional margin around the bounding box.

Using the same bounding box, two versions of the original image are prepared: one with the face blacked out and another with Gaussian noise superimposed, but ensuring the underlying structure remains vaguely intact. These are used as further guidance for Stable Diffusion, to preserve data utility. We regulate the intensity of the noise through a strength parameter [32] that ranges from zero (no noise) to one (full noise), setting it at an empirically determined value of 0.7 for our main experiments.

All parameters of our approach, the CFG scale, the size of the inpainting area and the strength affect the trade-off between privacy and data utility of the anonymized output. We chose their values based on detailed experiments determining their exact influence which we present in the supplementary material (Section 2).

4 Experiments

In this section, we first introduce the evaluation metrics utilized for assessing our experiments. Subsequently, we examine the anonymization performance and image quality of StablePrivacy and give a comparative analysis with established benchmarks. Finally, we show that it is possible to train a high-quality face detector on our anonymized data.

4.1 Evaluation Metrics

Synthesis-based approaches to de-identification are commonly evaluated regarding their ability to protect images against automated recognition as well as their image quality [10, 28]. The former is often measured using face verification on LFW. For this, we anonymize one of the images of each matched pair, given by the dataset, but not the other. Mismatched pairs also stay unchanged. Then, we use a face recognition model (FaceNet [42] or ArcFace [6]) to compute the distance between the two in the embedding space. Afterward, we use the threshold distance at which the False Acceptance Rate (FAR) is 10^{-3} to compute the True Acceptance Rate (TAR), which we employ to compare the performance of different

approaches. The lower the TAR the better the protection against recognition.
Fréchet inception distance (FID) [12] is typically used to estimate the quality of generated images. For this, the Fréchet distance between the original dataset's distribution and the generated dataset's distribution of features calculated by Inception v3 [46] trained on ImageNet [5] is computed. Good (low) FID correlates with human perception of similarity. In this paper, we also use Kernel Inception distance (KID) [2], an improvement over FID which is applicable to smaller datasets.
To demonstrate the practical applicability of our approach for the de-identification of face detection datasets while preserving the data-utility, we train the Dual Shot Face Detector (DSFD) [26] on anonymized versions of WIDER FACE [48]. We evaluate the success of the training using mean Average Precision (mAP) [8] at an Intersection over Union (IoU) threshold of 0.5. As there is only one class in face detection, the mAP is equal to the AP. The value ranges from zero to one, with one indicating the perfect score.

4.2 Anonymization and Image Quality

In this subsection, we assess the anonymization efficacy of our proposed approach within the context of the LFW benchmark and compare it to other recent de-identification techniques. See Table 1.
First, we establish the baselines. The original, unaltered images set the upper limit for recognition rates, illustrating what is possible without privacy considerations. On the opposite end are images subjected to heavy pixelization, which offers the strongest privacy but at the cost of significantly reduced data utility. This trade-off is evident in the low TAR (FaceNet and ArcFace) — 0.56% and 0.33%, respectively — and high KID of 0.0417. Our method's performance was then compared to state-of-the-art techniques. It can be seen that while DeepPrivacy, DeepPrivacy2, Leibl *et al*. and LDFA all significantly reduce the probability of automatic recognition, they still leave a considerable risk. In comparison, CIAGAN gives much better protection, but cannot keep up with the image quality of other approaches. Our method, StablePrivacy, delivers reasonable KID while at the same time substantially outperforming all others (besides pixelization) on the LFW verification benchmark, for both FaceNet (0.87%) and ArcFace (1.03%).
Notably, while FaceNet generally re-identifies anonymized faces more effectively in our setting than ArcFace, the opposite holds true for Leibl *et al*. and our approach. The reason for this is that both methods use FaceNet to choose appropriate sources to ensure that the generated faces differ from the originals. This improves the performance when measured with the same recognition model by optimizing the distance between the de-identified and the original image on FaceNet's features, some of which might be specific to that embedding. However, the comparable performance on ArcFace, which is not used in the anonymization process, suggests the strategy effectively protects against different recognition models.
A visual comparison of images de-identified by the different techniques is presented in Fig. 2. It shows that our anonymizations look highly realistic and deal well with common occlusions like glasses (row one) and hats (row three) as well as difficult poses (row three). At the same time features that make faces similar in human perception like the shape of the eyes and the thickness of the lips and eyebrows [43] are clearly different. Further examples are shown in the supplementary material (Section 3).

Table 1: Comparison of anonymization performance and image quality of de-identification methods. StablePrivacy offers the best privacy protection among the synthesis-based face de-identification approaches.

De-ID Method	FaceNet (↓) [%]	ArcFace (↓) [%]	KID (↓)	FID (↓)
Original	98.60 ± 0.76	96.13 ± 1.81	N/A	N/A
Face Pixelization 16 × 16	0.56 ± 1.67	0.33 ± 0.26	0.0417 ± 0.0012	43.09
CIAGAN [23]	3.40 ± 0.65	5.83 ± 1.97	0.0105 ± 0.0007	13.30
DeepPrivacy [18]	10.90 ± 1.93	6.63 ± 2.12	0.0014 ± 0.0002	2.37
DeepPrivacy2 [7]	11.64 ± 1.97	8.60 ± 1.76	**0.0004 ± 0.0002**	**1.34**
LDFA [20]	12.92 ± 2.51	9.40 ± 2.39	0.0014 ± 0.0002	2.58
Leibl et al. [24]	9.03 ± 1.01	11.47 ± 2.25	0.0146 ± 0.0008	13.26
StablePrivacy	**0.87 ± 0.48**	**1.03 ± 0.48**	0.0017 ± 0.0003	3.36

Figure 2: Visual comparison with other anonymization methods. Our anonymizations are highly realistic and deal well with common occlusions like glasses (rows one) and hats (row three) as well as difficult poses (row three).

4.3 Training Face Detectors on Anonymized Data

While our prior experiments have illustrated our approach's capability for enhancing privacy and generating realistic images, it is important to note that synthesis-based de-identification can introduce artifacts into the modified images. As highlighted in the literature [22], these artifacts may cause face detectors trained on such datasets to overfit on them, subsequently impairing their performance when encountering real data. Therefore, in this section, we evaluate how effectively data de-identified by our method can train face detectors.

Given that face detection datasets, like WIDER FACE, typically demand anonymizing multiple faces within a single image, our method requires additional processing steps. This entails cropping each face based on the ground truth bounding boxes from the dataset (expanded by 100x100 pixels), anonymizing these faces individually, and subsequently reconstructing the images. Further details are provided in the supplementary material (Section 5). Additionally, we make another adjustment to our approach, aiming to leverage the wide variance in face sizes encountered in datasets such as WIDER FACE, where sizes range from tiny (less than 5×5 pixels) to extremely large (up to 1000×1000 pixels). Considering that smaller faces are more difficult to identify [23], we tailor the strength parameter of our approach according to the face size. Specifically, we use a strength value of 0.7 for faces exceeding 30×30 pixels, and 0.5 for smaller faces. This allows us to use more structural guidance from the original image resulting in higher image quality and to speed up computation as less steps need to be calculated during the reverse diffusion process. At the same time, the reduced

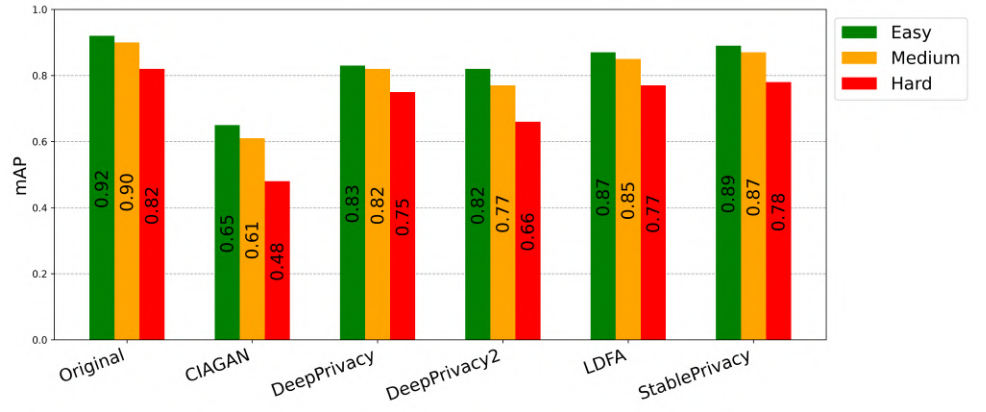

Figure 3: Comparative results for training face detectors (DSFD) on the WIDER FACE dataset: original and anonymized by different approaches. The use of StablePrivacy significantly enhances performance compared to other de-identification approaches.

privacy protection associated with lower strength values for smaller faces is offset by their limited size. In the supplementary material (Section 6), we explore alternative choices for selecting the strength value for given face sizes and measure the impact on privacy protection.

To evaluate the utility of our anonymized data, we train a face detector (DSFD) on WIDER FACE de-identified by different approaches and evaluate the performance using mAP following the procedure established by Kolmp *et al*. [22]. A comprehensive account of our training protocols is detailed in the supplementary material (Section 5). Comparing the same approaches as before, except for Leibl *et al*. which is unsuitable for this dataset, the outcomes displayed in Fig. 3 demonstrate that detectors trained on data processed by StablePrivacy perform better than those trained on other anonymized data. At the same time, StablePrivacy delivers much better privacy protection as shown in our previous experiments.

4.4 Ablation Study

In this section, we present our ablation study methodically examining the effects of removing key components of our approach: IP-Adapter, source pre-selection, and SDEdit. We assess the repercussions on image quality (KID and FID) and de-identification efficacy (FaceNet, ArcFace). The results are given in Table 2. The removal of the IP-Adapter leads to a strong decrease in de-identification performance, with the TAR increasing to 24.3 %. This suggests that without this component, StablePrivacy tends to use the guidance from the noisy version of the original face to closely reconstruct it. Removal is, therefore, an ill-suited choice, even though the image quality measured by KID improves to a value of 0.0007. Omitting source pre-selection results in a TAR of 2.2 %, signaling a significant drop in privacy protection. This highlights the source pre-selection's role in enhancing de-identification. Switching to random noise instead of utilizing SDEdit for controlling the initial face image degradation before processing with Stable Diffusion slightly enhances the TAR to 0.60 %. However, this

Table 2: Ablation Study. The effect on privacy protection and image quality when removing the IP-Adapter, the source pre-selection or the SDEdit component from our approach.

De-ID Method	FaceNet (↓) [%]	ArcFace (↓) [%]	KID (↓)	FID (↓)
StablePrivacy	0.87 ± 0.48	1.03 ± 0.48	0.0017 ± 0.0003	3.36
w/o IP-Adapter	24.30 ± 1.89	19.89 ± 2.92	0.0007 ± 0.0002	1.94
w/o source selection	2.2 ± 0.54	2.00 ± 1.06	0.0050 ± 0.0005	6.10
w/o SDEdit	0.60 ± 0.25	0.53 ± 0.27	0.0028 ± 0.0004	4.56

comes at the expense of significantly reduced image quality, with a KID of 0.0028. The ablation experiments show that each element of our approach contributes to achieving the desired privacy–data utility trade-off.

5 Limitations

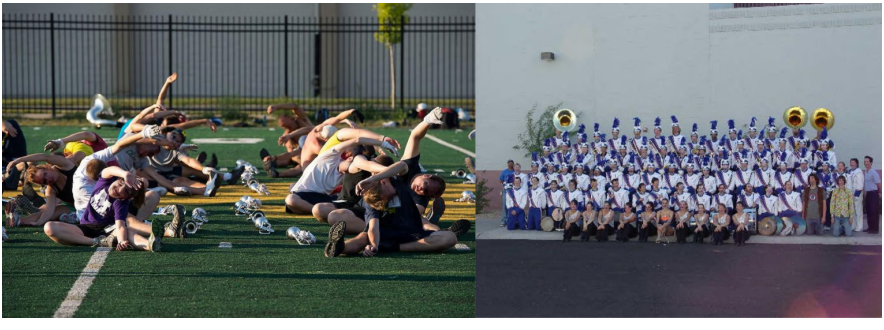

Figure 4: Limitations. Extreme poses, occlusions (left) and very small faces (right) are challenging to anonymize for StablePrivacy.

While we show in previous experiments that our method can retain the variety of features necessary for training a face detector, the WIDER FACE dataset contains several extreme poses, large occlusions and small faces that cause StablePrivacy to produce unrealistic outputs. Examples are shown in Fig. 4 and in the supplementary material (Section 4). The left image shows rotated and occluded faces that typically occur during sports or other in-the-wild situations. The output only vaguely resembles human faces. Similarly, the small faces shown in the right image are more blurry than in the original and contain visual artifacts.

Another limitation of this work is that we assumed the synthetic face images used as the source library to be free of privacy concerns. In practice, this is not necessarily the case, as generative models can accidentally memorize [9] training data or be susceptible to Membership Inference Attacks [4]. To guarantee the privacy of source library images, we plan to use generative methods based on differential privacy [19, 27] in future work.
Finally, we want to point out that while we took care to balance the source library with respect to gender, there are other sensitive biases such as skin color or age that are not considered in the current work.

6 Conclusion

This paper has introduced StablePrivacy, an innovative method leveraging Stable Diffusion to enhance privacy in face image datasets while retaining data utility critical for training deep learning-based face detection models. Our experimental results demonstrate StablePrivacy's superior performance in balancing privacy protection with data utility. Notably, our approach exhibits state-of-the-art de-identification capabilities, as evidenced by a significant reduction in True Acceptance Rate (TAR) measured on the LFW benchmark. At the same time, data anonymized with StablePrivacy is better suited for training face detector models than other anonymized data as shown by superior detection performance.

Acknowledgment

The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

References

[1] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? *IEEE Winter Conference on Applications of Computer Vision*, pages 1536–1546, 2021.

[2] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *International Conference on Learning Representations*, 2018.

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yuxin Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. *ACM SIGSAC Conference on Computer and Communications Security*, 2019. URL https://api.semanticscholar.org/CorpusID:221203089.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[8] Mark Everingham, Andrew Zisserman, Christopher K I Williams, Luc Van Gool, Moray Allan, Christopher M Bishop, Olivier Chapelle, Navneet Dalal, Thomas Deselaers, Gyuri Dorkó, et al. The 2005 pascal visual object classes challenge. *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop*, pages 117–176, 2005.

[9] Qianli Feng, Chen Guo, Fabian Benitez-Quiroz, and Aleix M Martínez. When do gans replicate? on the choice of dataset size. *IEEE/CVF International Conference on Computer Vision*, pages 6681–6690, 2021. URL https://api.semanticscholar.org/CorpusID:244398659.

[10] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 2017.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, abs/2207.12598, 2021.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[15] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[16] Hakon Hukkelas and Frank Lindseth. Does image anonymization impact computer vision training? *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 140–150, 2023.

[17] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1329–1338, 2023.

[18] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. *Advances in Visual Computing*, 2019.

[19] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. OpenReview.net, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] Marvin Klemp, Kevin Rösch, Royden Wagner, Jannik Quehl, and Martin Lauer. Ldfa: Latent diffusion face anonymization for self-driving applications. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3199–3205, 2023.

[22] Sander R Klomp, Matthew Van Rijn, Rob G J Wijnhoven, Cees G M Snoek, and Peter H N De With. Safe fakes: Evaluating face anonymizers for face detectors. *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021.

[23] Martin Knoche, Stefan Hörmann, and Gerhard Rigoll. Susceptibility to image resolution in face recognition and training strategies to enhance robustness. *Leibniz Transactions on Embedded Systems*, 8:01:1–01:20, 2022.

[24] Jun Ha Lee and Sujeong You. Balancing privacy and accuracy: Exploring the impact of data anonymization on deep learning models in computer vision. *IEEE Access*, 12: 8346–8358, 2024.

[25] Andreas Josef; Leibl, Andreas; Attenberger, Andreas; Meißner, Stefan; Altmann, and Helmut Mayer. De-identifying face image datasets while retaining facial expressions. *IEEE International Joint Conference on Biometrics*, 2023.

[26] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: Dual shot face detector. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5055–5064, 2019.

[27] Yunhui Long, Boxin Wang, Zhuolin Yang, Bhavya Kailkhura, Aston Zhang, C A Gunter, and Bo Li. G-pate: Scalable differentially private data generator via private aggregation of teacher discriminators. 2019. URL https://api.semanticscholar.org/CorpusID:245634703.

[28] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixe. Ciagan: Conditional identity anonymization generative adversarial networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5446–5455, 2020.

[29] Blaz Meden, Peter Rot, Philipp Terhorst, Naser Damer, Arjan Kuijper, Walter J. Scheirer, Arun Ross, Peter Peer, and Vitomir Struc. Privacy-enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, 16:4147–4183, 2021.

[30] Andreas Meißner, Andreas Fröhlich, and Michaela Geierhos. Keep it simple: Local search-based latent space editing. *International Joint Conference on Computational Intelligence*, pages 273–283, 2022.

[31] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. *IEEE Symposium on Security and Privacy (SP)*, pages 691–706, 2019.

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations*, 2021.

[33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *International Conference on Machine Learning*, 139:8162–8171, 2 2021.

[34] State of California Legislative Counsel. Assembly bill no. 375 – chapter 55. california consumer privacy act, 2018. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375. Accessed: 2024-03-19.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, 139:8748–8763, 2 2021.

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *International Conference on Machine Learning*, 139:8821–8831, 2 2021.

[37] Slobodan Ribaric and Nikola Pavesic. An overview of face de-identification in still images and videos. *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2015.

[38] Maria Rigaki and Sebastián García. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56:1 – 34, 2020.

[39] Robin Rombach, A Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[41] Kanthi Kiran Sarpatwar, Nalini K Ratha, Karthik Nandakumar, Karthikeyan Shanmugam, James T Rayfield, Sharath Pankanti, and Roman Vaculín. Privacy enhanced decision tree inference. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 154–159, 2020.

[42] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[43] Pawan Sinha, B Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94:1948–1962, 2006.

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *International Conference on Machine Learning*, 37:2256–2265, 2 2015.

[45] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott M Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2451, 2019.

[46] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[47] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR) A Practical Guide*, volume 1. Springer, Cham, 2017.

[48] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533, 2016.

[49] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. *ACM SIGSAC Conference on Computer and Communications Security*, 2019.

[50] Hu Ye, Jun Zhang, Siyi Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *ArXiv*, abs/2308.06721, 2023.

[51] Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Global-local gcn: Large-scale label noise cleansing for face recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7728–7737, 2020.