Generating Private Data Surrogates for Vision Related Tasks

Ryan Webster, Julien Rabin, Loïc Simon, Frédéric Jurie

Generating Private Data Surrogates for Vision Related Tasks

Ryan Webster [1] **Julien Rabin** [1] **Loïc Simon** [1] **Frédéric Jurie** [1]

Abstract

With the widespread application of deep networks in industry, membership inference attacks, *i.e.* the ability to discern training data from a model, become more and more problematic for data privacy. Recent work suggests that generative networks may be robust against membership attacks. In this work, we build on this observation, offering a general-purpose solution to the membership privacy problem. As the primary contribution, we demonstrate how to construct surrogate datasets, using images from GAN generators, labelled with a classifier trained on the private dataset. Next, we show this surrogate data can further be used for a variety of downstream tasks (here classification and regression), while being resistant to membership attacks. We study a variety of different GANs proposed in the literature, concluding that higher quality GANs result in better surrogate data with respect to the task at hand.

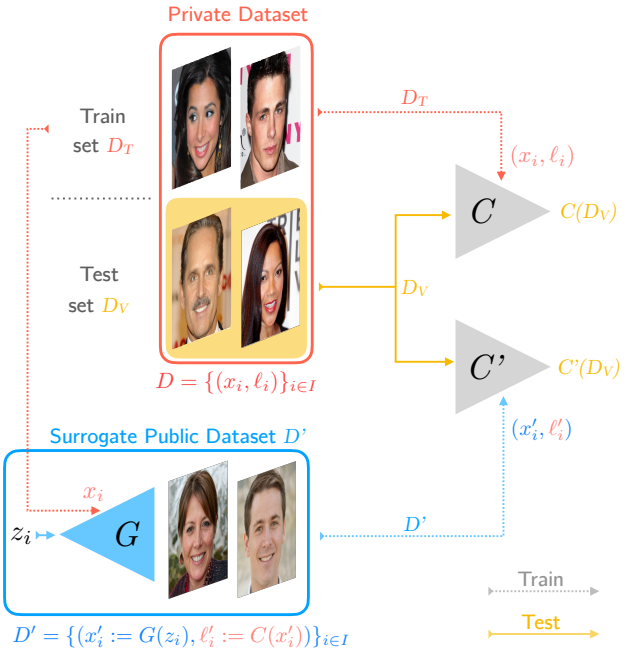


Figure 1. Overview of the proposed framework for creating private data surrogates and its application to train a private task-driven network. In a nutshell, the data surrogate D' is simply obtained by combining image samples from a generator network G (*e.g.* using GAN) and associating them with plausible labels obtained from a classifier C trained on the private train dataset D_T. From this public dataset, it is possible to train a privacy preserving classifier C' displaying similar performance and accuracy (in practice by comparing $C'(D_V)$ and $C(D_V)$ on a separate validation set D_V). We further demonstrate empirically that the obtained public dataset D' (and by composition the network C') is robust to membership attack that is described in Algorithm 1.

1. Context and Motivation

The fantastic recent advances in deep learning are strongly and inextricably related to the existence of public datasets. Such datasets not only allow researchers to learn from and experiments with the data but also to measure progress and challenge themselves with other researchers in competitions. If Pascal VOC (Everingham et al., 2010) was among the first, many followed such as ImageNet (Deng et al., 2009) for object classification or recently CelebA-HQ (Karras et al., 2018a) as a benchmark for generative models. The contemporary machine learning research landscape would undoubtedly be very different without such datasets.

However, several important application areas do not fully benefit from the progress of machine learning because of the lack of massive public datasets. Indeed, building such datasets is difficult due to privacy issues that will inevitably arise.

Privacy issues not only prevent the release of public datasets: the distribution of already trained deep networks itself poses severe risks of information leakage. Modern convolutions networks have parameters in millions, and, as shown by Zhang *et al.* (Zhang et al., 2016) and Yeom *et al.* (Yeom et al., 2018) classification networks can memorize images, posing direct threats to privacy. Such models can leak private training data to an attacker via their bias towards train-

[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC. Correspondence to: Ryan Webster <Ryan.Webster@ensicaen.fr>.

SPML'19, ICML Workshop on Security and Privacy of Machine Learning, Long Beach, California, 14th June 2019.

ing samples.

One common attack against machine learning models is the *membership attack*, which discerns data points that were used for training. Neural networks performing classification on images are known to be vulnerable to such attacks. For instance, in (Shokri et al., 2017), a membership attack was successfully performed against an MNIST model, even if the network parameters were not available to the attacker. In essence, these attacks exploit neural networks tendency to overfit, and can use simple cues such as output logit entropy.

In response to such membership attacks, various heuristic and theoretical attack defenses have also been proposed. There is the mathematical framework of *differential privacy*, which is often used to provide privacy guarantees in machine learning frameworks. For example, classifiers with privacy guarantees can be achieved with the teacher ensemble mechanism (Papernot et al., 2018) (PATE), or through knowledge transfer (Papernot et al., 2016). Empirical defenses have also been proposed for membership attacks in particular. For example, (Nasr et al., 2018a) proposes adding an adversary during training which simulates an attacker and therefore minimizes a utility privacy trade-off.

1.1. Privacy of Generative Models

While generative adversarial networks (GANs) (Goodfellow et al., 2014) are a relatively new advent, they certainly have changed the landscape of machine learning, for example, achieving state of the art in image synthesis (Karras et al., 2018b; Miyato et al., 2018; Iizuka et al., 2017) and a plethora of other generative tasks. Recent work has been devoted to defining and measuring overfitting in generative models, for example, in (Webster et al., 2019; Im et al., 2018; Gulrajani et al., 2019). (Webster et al., 2019) proposes directly recovering training and test images with the generator and comparing recovery error, with the hypothesis that training images should exhibit smaller error.

Finally, some recent works approach the question of deep learning with privacy via direct sanitation of data for release. Mirjalili *et al.*(Mirjalili et al., 2018) used convolutional autoencoders to remove other information (*e.g.* , genre) than the one related to identity from training face images. Sokolic *et al.*(Sokolic et al., 2017) proposed a data sanitization mechanism during which users' data is modified to prevent specific attacks before these data are actually used for training. The same goal is sought by Bertran *et al.* (Bertran et al., 2018), Wu *et al.*(Wu et al., 2018) or Rezaei *et al.*(Rezaei et al., 2018), by defining a collaborative sanitization function retaining valuable information for the tasks while eliminating private information. Finally, PATE-GAN (Yoon et al., 2019) proposes a framework to generate data using a GAN like framework and adopting privacy via the PATE mechanism.

In this work, we also seek to directly release data immune to membership attacks by presenting two contributions: First, a simple but efficient surrogate technique is proposed to generate a synthetic dataset that can be released in public, which ideally achieves the same level of performance as the original dataset while ensuring its protection to membership attack. The methodology for generating, labeling and evaluating this data is documented in Fig. 1.

Second, two recent membership attacks against generative models are executed against GAN generators and shown to be ineffective. Finally, while this study is preliminary, we evaluate two face datasets, CelebA-HQ and UTK-Face, across a variety of state of the art GAN generators as well as various tasks to demonstrate consistency of our observations.

The rest of the paper is organized as follows: Section 2 exposes the construction of the surrogate dataset as well as its evaluation. Section 3 details the proposed membership attack protocol to assess the efficiency of our surrogate dataset. Section 4 shows experimental results, followed by a discussion in Section 5.

2. Surrogate Data Creation and Evaluation

Typically, methods offering privacy in machine learning do so with a cost and consider privacy bounds alongside the final utility of the model (Nasr et al., 2018b; Yoon et al., 2019; Papernot et al., 2018). For example PATE-GAN (Yoon et al., 2019) demonstrates the utility of generated data through unsupervised tasks under various privacy guarentees.

We propose evaluating utility on supervised tasks by first labeling unconditional samples and finally taking the standard validation accuracy on real images. Figure 1 details the entire pipeline:

- train a classifier C and a generator G from the private training dataset $\mathcal{D}_{\mathcal{T}}$;
- build the publicly released dataset $\mathcal{D}'$ from G and C, combining randomly generated samples $x' = G(z)$ where $z \sim \mathbb{P}_Z$ with predicted labels $\ell = C(x')$ from the private classifier;
- train a classifier C' from the surrogate dataset $\mathcal{D}'$;
- validate and compare C' and C on a validation set $\mathcal{D}_{\mathcal{V}}$.

Experiments detailed in Section 4 demonstrate for various GAN approaches that the surrogate classification network C' can achieve similar performance than the one trained directly on private data. We now present two membership attacks used to assess privacy of surrogate data.

3. Assessing privacy of Generative Models by Membership Attacks

In this work, we consider two membership attack models against generative networks. Both approaches utilize overfitting of the generative network G, with respect to some attack function A. Following LOGAN (Hayes et al., 2019), we discern the training set by simply sorting the values of A, taking the lowest values to be the training set. For instance in LOGAN, A is taken to be the trained discriminator D, in which case the training set is taken as those samples which the discriminator most confidently predicts to be real. Algorithm 1 details precisely how this attack is performed.

Algorithm 1 Membership attack

Input: Training set $\mathcal{D}_\mathcal{T}$, validation set $\mathcal{D}_\mathcal{V}$

1: Set the attack score function A, either from the recovery loss function in Eq. (1) or the discriminator D.

2: Let $x_i \in \mathcal{D}_\mathcal{T} \cup \mathcal{D}_\mathcal{V}$, such that

$$\begin{cases} x_i \in \mathcal{D}_\mathcal{T} & \text{if } i \leq N \\ x_i \in \mathcal{D}_\mathcal{V} & \text{if } N+1 \leq i \leq 2N \end{cases}$$

3: Sorted indices: $I \leftarrow \arg\operatorname{sort}\{A(x_i)\}_{1 \leq i \leq 2N}$

Output:

4: Estimated set of training images: $\mathcal{T} \leftarrow \{x_{I(i)}\}_{1 \leq i \leq N}$

5: Membership attack accuracy: $Acc \leftarrow |I \cap \{i : 1 \leq i \leq N\}|/N$

3.1. Recovery Attacks

In (Webster et al., 2019), training and validation images were recovered using optimization. Generative networks were said to overfit if the statistics of training and validation recovery errors were different in some measure, for example the difference of medians. In (Liu et al., 2018), recovery errors were used similarly to perform membership attacks, where the optimization was performed over an input network to the latent space, rather than input codes.

Inspired by these approaches, we define the following latent recovery loss function

$$f_G(x_i) := \|\phi(G(E(x_i))) - \phi(x_i)\|_2^2 \qquad (1)$$

where $x_i \in \mathcal{D}_\mathcal{T} \cup \mathcal{D}_\mathcal{V}$ is in either the training or validation sets, ϕ are image features (such as convolution layers of VGG-19), G is the GAN generator trained on $\mathcal{D}_\mathcal{T}$ and E is the attack encoder. Indeed, contrarily to the aforementioned methods that rely on latent recovery optimization for every single image from a validation set, we resort here to a feed-forward recovery method that consists in using an encoder from the image domain that is exclusively trained on generated samples $G(z)$ by solving $\min_E \mathbb{E}_{z \sim \mathbb{P}_Z} f_G(E(G(z)))$.

We consider using perceptual features for ϕ, as they are generally considered to be well suited for synthesis tasks (Johnson et al., 2016), but also consider using other feature networks, such as VGG-Face (Parkhi et al.) or even $\phi = Id$ for an image domain loss. Furthermore, notice that we only train E on generated samples $G(z)$, so that E is invariant to the training and validation set split of G.

3.2. Discriminative Attacks

Following the work of (Hayes et al., 2019), *Discriminative* membership attacks are performed against GANs by using the discriminator D to sift between test and train images. Recall that D is trained along with the generator G on the training set $\mathcal{D}_T$. For convenience, we assume here that D is trained to score 0 for real images $x_i \in \mathcal{D}_T$ and 1 for generated images $x'_i = G(z_i)$.

4. Results

We train the following GANs and will use the abbreviations in parentheses: Progressive GANs presented in (Karras et al., 2018a) (PGGAN) using official code, the zero-centered gradient penalty Resnet in (Mescheder et al., 2018) with the official code (MESCH) and finally deep convolutional GAN (Radford et al., 2015) (DCGAN) and least squares GAN (Mao et al., 2017) (LSGAN) with our own implementation.

4.1. Attribute Recognition on CelebA-HQ

The CelebA-HQ dataset (Karras et al., 2018a) is a typical GAN benchmark dataset, which includes 40 attributes l_i, from which we considered the following ones: *gender*, *smiling*, *young*, *glasses*, *blond*.

Table 1 (and Table 3 in appendix) shows that a classifier C' trained with a surrogate dataset D' (generated with the proposed approach described in Section 2) performs as well as a classifier C directly trained on the private dataset D. As demonstrated by the FID scores that evaluate the quality of the generator (lower is better), the drop in performance strongly correlates with the quality of the synthesized images used for training.

4.2. Safety to Membership Attacks

We assess safety of every GAN models trained versus the membership attacks described in Section 3. Table 2 (and Table 4 in the appendix) shows that membership attacks described in Algorithm 1 are largely unsuccessful when the size of the training set $|\mathcal{D}_\mathcal{T}|$ is large enough. For evaluation of membership attack accuracy, we have used $N = 2000$ images from $\mathcal{D}_\mathcal{T}$ and $\mathcal{D}_\mathcal{V}$.

The discriminator attack of LOGAN is overall most success-

			Gender	Smiling	Average (5 attributes)	Change in Performance	FID
VGG-Face Features	C	Real Data	94.50	85.20	90.64	-	-
	C'	DCGAN	91.90	82.10	86.50	4.14	67.07
		MESCH	92.60	81.45	88.90	1.74	26.31
		LSGAN	92.10	80.80	88.35	2.29	42.01
		PGGAN	93.10	83.05	89.35	**1.29**	**19.17**

Table 1. Performance of various surrogate datasets on the CelebA-HQ (Karras et al., 2018a) binary attribute recognition task. Top row represents a classifier C trained on the original dataset $\mathcal{D}_T$, subsequent rows represent classifiers C' trained with GAN images that are labelled with C (see Section 2 for details). Accuracy represents percent correct on a validation set $\mathcal{D}_V$.

	L_2 Recovery	VGG-Face Recovery	VGG-19 Recovery	Discriminator D
DCGAN	54.1	54.5	51.6	57.1
MESCH	53.9	50.8	52.5	50.1
LSGAN ($\lvert\mathcal{D}_\mathcal{T}\rvert = 26k$)	54.8	54.1	54.0	62.9
LSGAN ($\lvert\mathcal{D}_\mathcal{T}\rvert = 5k$)	58.1	56.2	57.8	99.4
PGGAN	52.0	50.3	52.1	N/A

Table 2. Membership attack accuracies (in %) for various GAN methods trained on the CelebA-HQ dataset and various attack methods (see Algorithm 1). When not specified otherwise, the size of the training dataset is $\lvert\mathcal{D}_\mathcal{T}\rvert = 26k$ and for the validation set $\lvert\mathcal{D}_\mathcal{V}\rvert = 2k$. GAN methods are reported in the first column. The next three columns use latent recovery attack with loss function f_G (see Eq. 1), with ϕ taken to be the identity, VGG-Face or VGG-19 features respectively. The final column reports the discriminative attack accuracy with the discriminator D from the GAN training (the discriminator of PGGAN requires feeding a whole batch which prevented us to implement this attack). As a baseline, the same discriminative attack is done on LSGAN with a smaller training dataset (5k) demonstrating that in such setting the discriminator network is capable of memorizing almost perfectly the entire training dataset.

ful, for example achieving nearly 63% accuracy on LSGAN even with $26k$ training images, although interestingly the regularized discriminator in MESCH will yield unsuccessful discriminator attacks even for small datasets, while appearing to be slightly vulnerable to recovery attacks. Note that this observation is not in contradiction to the results in LOGAN (Hayes et al., 2019), which showed successful discriminator attacks across the board, as that study primarily studied small training sets and focused on the DCGAN training technique. To show our observation is consistent with LOGAN, we also include LSGAN with a small training set of $|\mathcal{D}_\mathcal{T}| = 5k$, which yields a near perfect discriminator attack. Finally, we note that a discriminator attack is a fairly unrealistic scenario, as the discriminator parameters are typically never used from an application standpoint. Furthermore, an attacker has to have moderate knowledge of the dataset if he wants to retrain a discriminator (as in (Im et al., 2018)), which is somewhat counter intuitive to the attack in the first place. On the other hand, optimization of encoder E using Eq. (1) can be done merely by sampling G, as is done in this document, but requires the parameters of G. Finally, we note that we did not in fact use the labels given to the surrogate data D' to perform our membership attacks. We do not believe this would significantly affect the membership attack accuracy, as this information is implicitly available to the attacker, for example when using a semantic network like VGG-Face for a recovery attack.

5. Discussion and Conclusion

In this work, we presented a technique for the public release of data using GANs and verified empirically the data appear retains its utility while gaining privacy. More precisely, we have demonstrated that GANs surrogates are effective for age regression and face attribute classification. To verify privacy, two different inference attack mechanisms have been investigated. The first one is based on image recovery and the second one on the discriminator optimized during GAN training, which has been reported to overfit the training dataset. A major advantage to the presented method is that it can work off-the-shelf with any GAN generator. On the other hand, complex training procedures such as PATE-GAN only exacerbate already unstable GAN training and may result in low quality data samples. While we demonstrated data surrogates greatly reduce vulnerability to membership attacks, more insight should be shed into the mechanism behind this. Hopefully, this would allow for mathematical guarantees instead of purely empirical ones.

Future investigations will include the question of adapting this strategy to conditional GANs for situations where training an additional classifier could be avoided. Additionally, GANs need large datasets for training which raises the question of the extension of the proposed framework for small private datasets. The extension to multiple private datasets is also not straightforward, but would, however, provide a useful tool for distributed learning.

Acknowledgments This work was supported by fundings from Rgion Normandie under grant RIN NormanDeep. This study has also been carried out with financial support from the French State, managed by the French National Research Agency (ANR-17-CE39-0006).

Appendix: Additional experiments

Tables 3 and 4 report additional results on the UTK-face dataset (Zhang & Qi, 2017) that are consistent with the experiments on CelabA-HQ. Particularly, networks with higher quality samples (as measured by the FID score) are performing better on the task. This dataset is composed of 20k images of faces ranging from 0 to 116 years old. The task networks are here trained on $|\mathcal{D}_\mathcal{T}| = 10k$ images to perform age regression instead of binary attribute classification. The test set is again composed of $|\mathcal{D}_\mathcal{V}| = 2k$ images.

The only difference with the previous experiments on CelebA-HQ (in Table 1 is that the performance of the task networks C and C' for age regression is measured using median absolute difference between estimated labels $C(x_i)$ and the ground truth $\ell(x_i)$, which writes:

$$\text{MAD}(C) = \text{median}\{|C(x_i) - \ell(x_i)|, x_i \in \mathcal{D}_\mathcal{V}\}. \quad (2)$$

Finally, note that in Table 4 the discriminative attack on LSGAN is slightly more successful, likely due to the fact of the much smaller training set size of $|\mathcal{D}_\mathcal{T}| = 10k$. The discriminative attack seems far less consistent, as it performs no better than random guessing on the DCGAN network with the same training set size. Future work should investigate discriminative attacks in more detail, across a wider range of datasets, GAN techniques and dataset sizes.

In the same line and accordingly with recent results from (Webster et al., 2019), Figure 2 illustrates the ability of a generative model (here LSGAN) to overfit a dataset when trained with an insufficient number of examples.

Last, Figure 3 illustrates the interest of latent recovery attack with the proposed perceptual encoder E (see Section 3.1) on PGGAN and CelebA-HQ. As reported in (Webster et al., 2019), when trained on a sufficient number of images, recovered images with GANs are of similar quality for test and train images.

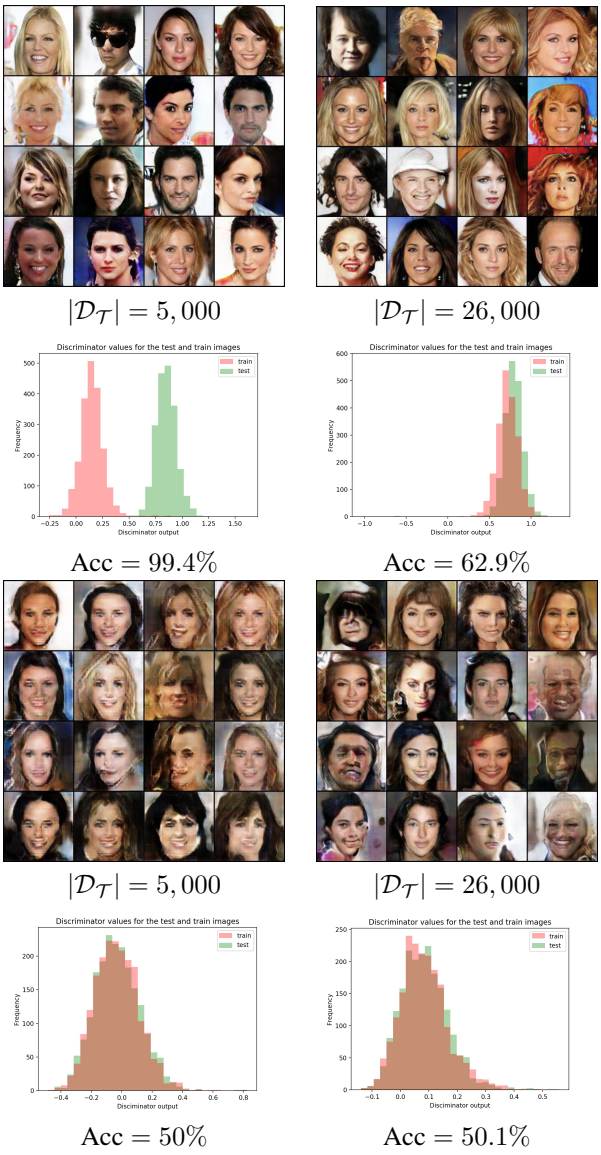


Figure 2. Histogram of attack scores based on the Discriminator D for $N = 2000$ images from the training set $\mathcal{D}_\mathcal{T}$ (in red) and the test set $\mathcal{D}_\mathcal{V}$ (in green) for LSGAN (first two rows) and MESCH (next two rows) trained on CelebA-HQ, respectively with $|\mathcal{D}_\mathcal{T}| = 5,000$ images (left column) and $26,000$ images (right column). While the quality of images does not improve a lot with a larger number of training images, the robustness to discriminative attack increases dramatically for LSGAN (average membership inference attack accuracy are given in the last row).

References

Bertran, M., Martinez, N., Papadaki, A., Qiu, Q., Rodrigues, M., and Sapiro, G. Learning to Collaborate for User-Controlled Privacy. *arXiv:1805.07410 [cs, stat]*, May 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On*, pp. 248–255. IEEE, 2009.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88 (2):303–338, 2010.

			Age (MAD error, in years)	Change in Performance (in years)	FID
VGG-Face Features	C	Real Data	5.22	-	-
	C'	DCGAN	12.03	6.81	89.68
		LSGAN	5.56	0.34	31.05
		PGGAN	5.12	**-0.10**	**30.65**

*Table 3. Performance of various surrogate datasets on the age regression task of **UTK-Face** (Zhang & Qi, 2017). Top row represents a classifier trained on the original dataset, subsequent rows represent classifiers trained with GAN images (see Section 2). MAD is the Median Absolute Difference (see Eq. 2) on the predicted versus ground-truth age for the validation set $\mathcal{D}_V$ (lower is better). FID scores are reported in the last column (lower is better) to assess the quality of generated images.*

	L_2 Recovery	VGG-Face Recovery	VGG-19 Recovery	Discriminator
DCGAN	52.3	53.5	52.1	50.9
LSGAN	53.4	53.9	53.6	75.8
PGGAN	54.7	56.8	54.1	-

*Table 4. Membership attack accuracies (in %) for various GAN methods trained on the **UTK-Face** dataset and various attack methods (see Algorithm 1). When not specified otherwise, the size of the training dataset is $|\mathcal{D}_\mathcal{T}| = 26k$ and for the validation set $|\mathcal{D}_\mathcal{V}| = 2k$. GAN methods are reported in the first column. The three next columns use latent recovery attack with loss function f_G (see Eq. 1), with ϕ taken to be the identity, VGG-Face or VGG-19 features respectively. The final column reports the discriminative attack accuracy with the discriminator D from the GAN training.*

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Gulrajani, I., Raffel, C., and Metz, L. Towards GAN benchmarks which require generalization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkxKH2AcFm.

Hayes, J., Melis, L., Danezis, G., and Cristofaro, E. D. LOGAN: Membership Inference Attacks Against Generative Models. *PoPETs*, 2019(1):133–152, 2019. doi: 10.2478/popets-2019-0008.

Iizuka, S., Simo-Serra, E., and Ishikawa, H. Globally and Locally Consistent Image Completion. *ACM Transactions on Graphics (Proc. of SIGGRAPH 2017)*, 36(4): 107:1–107:14, 2017.

Im, D. J., Ma, A. H., Taylor, G. W., and Branson, K. Quantitatively evaluating GANs with divergences proposed for training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJQHjzZ0-.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pp. 694–711. Springer, 2016.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *Sixth International Conference on Learning Representations (ICLR)*, 2018a.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018b.

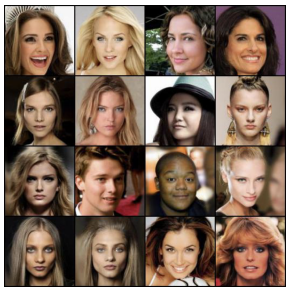
Liu, K. S., Li, B., and Gao, J. Performing Co-Membership Attacks Against Deep Generative Models. *arXiv:1805.09898 [cs, stat]*, May 2018.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2794–2802, 2017.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, pp. 3478–3487, 2018.

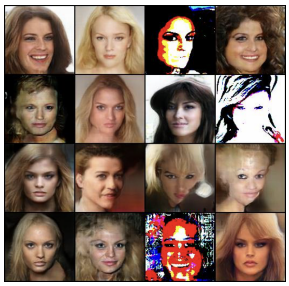
Mirjalili, V., Raschka, S., Namboodiri, A. M., and Ross, A. Semi-adversarial Networks: Convolutional Autoencoders for Imparting Privacy to Face Images. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pp. 82–89, 2018. doi: 10.1109/ICB2018.2018.00023.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.



Real images $x_i \in \mathcal{D}_\mathcal{T} \cup \mathcal{D}_\mathcal{V}$

Recovered images $G(E(x_i))$ from PGGAN with Encoder



Recovered images $G(z_i)$ from LSGAN with latent recovery optimization (Webster et al., 2019)

Recovered images $G(E(x_i))$ from LSGAN with Encoder

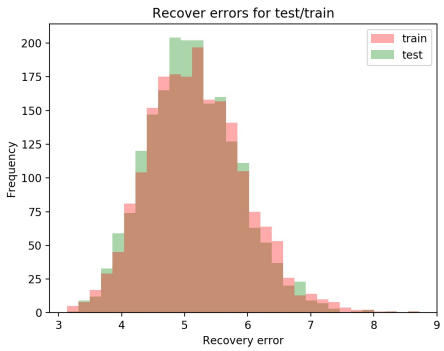


Figure 3. Illustration of latent recovery attack for PGGAN and LSGAN. Real images x_i from the test and train dataset are analyzed with the encoder E then synthesized with the generator G (right). The encoder E, trained on generated images, improves latent recovery in comparison with explicit optimization (Webster et al., 2019). The discrepancy between recovery error (see Eq. 1 is used to perform membership inference attack, which is here unsuccessfully (52% as reported in Table 2).

Nasr, M., Shokri, R., and Houmansadr, A. Machine Learning with Membership Privacy using Adversarial Regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, pp. 634–646, 2018a. doi: 10.1145/3243734.3243855.

Nasr, M., Shokri, R., and Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646. ACM, 2018b.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, U. Scalable private learning with PATE. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZB1XbRZ.

Parkhi, O. M., Vedaldi, A., Zisserman, A., et al. Deep face recognition.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rezaei, A., Xiao, C., Gao, J., and Li, B. Protecting Sensitive Attributes via Generative Adversarial Networks. *arXiv:1812.10193 [cs, stat]*, December 2018.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Sokolic, J., Qiu, Q., Rodrigues, M. R. D., and Sapiro, G. Learning to Succeed while Teaching to Fail: Privacy in Closed Machine Learning Systems. *arXiv:1705.08197 [cs, stat]*, May 2017.

Webster, R., Rabin, J., Simon, L., and Jurie, F. Detecting Overfitting of Deep Generative Networks via Latent Recovery. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Wu, Z., Wang, Z., Wang, Z., and Jin, H. Towards Privacy-Preserving Visual Recognition via Adversarial Training: A Pilot Study. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, pp. 627–645, 2018. doi: 10.1007/978-3-030-01270-0_37.

Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*, pp. 268–282, 2018. doi: 10.1109/CSF.2018.00027.

Yoon, J., Jordon, J., and van der Schaar, M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530 [cs]*, November 2016.

Zhang, Zhifei, S. Y. and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.