

Privacy-preserving Image Generation Based on Self-Attention

Zhihui Wang✉^[0000-0002-8024-7753] and Zijian Li

School of Computer Science, Fudan University, Shanghai, China
zhhwang@fudan.edu.cn
Shanghai Key Laboratory of Data Science, Shanghai, China

Abstract. In recent years, differential privacy techniques have become the de facto standard for privacy protection. With rigorous mathematical proofs, it only requires to add a small amount of noise to the deterministic functions to achieve privacy protection. However, most existing data generation methods based on differentially private GANs (Generative Adversarial Networks) are very difficult to train. Especially for privacy-preserving image generation, existing methods are difficult to generate high-quality results. We rethink the noise addition in differentially private GAN and design a self-attention differentially private GAN (DP-SAGAN). Even working in the case of privacy preservation, our DP-SAGAN still scores comparable to the current mainstream models in IS and FID metrics, and also generates higher resolution images stably.

Keywords: Differential Privacy · Self-Attention · Image Generation · Generative Adversarial Network.

1 Introduction

Generative Adversarial Network (GAN)[8] is an attractive topic in the field of artificial intelligence in recent years and GAN has shown impressive performance in modeling the underlying data distribution with its variants because of its unique training approach. However, the generation method of sampling from a distribution does not guarantee that the generated data will not have private information about the training dataset. Recent studies have shown that machine learning models may leak sensitive information about the training samples. We can launch an attack against the target model and then infer the membership of the training set[22] or reconstruct the entire training dataset[7]. The same problem is naturally not avoided by GAN. The work of [11] designed a white-box attack against the publicly released discriminator of GAN and showed that this attack could even achieve 100% accuracy in some cases.

The proposed concept of differential privacy[6] provides a promising direction to address the problem of generating data with guaranteed privacy. Now, differential privacy techniques have become the de facto standard, providing strict privacy guarantees through rigorous mathematical derivations, and a number of

researchers have successively proposed various techniques to satisfy differential privacy for designing generative models.

To address these issues and to enable the application of differential privacy techniques to today’s more complex models, we have rethought the approach of perturbing gradients on models and designed a differentially private GAN, which we call DP-SAGAN. We find that the training of our DP-SAGAN is more stable in image generation compared to other differentially private GANs and can consistently obtain higher resolution images, and the images generated by DP-SAGAN perform no worse than other GAN works in downstream classifiers.

Our contributions in this paper are as follows. Firstly, we reconsidered the way of gradient perturbation and designed a method which may substantially improve the training speed of the differentially private GAN. Secondly, we proposed a differentially private GAN based on a self-attentive mechanism. Thirdly, we compared the image generation quality of current mainstream methods with that of our method, and found that our method had better effectiveness than other methods especially in high-resolution image generation. Finally, the images generated by our method also performed well in downstream classifiers.

The rest of this paper is organized as follows. In section 2, we introduce the related work. In section 3, we describe the preliminary knowledge of this paper. And we propose our approach and present its details in section 4. The experimental results are described and analyzed in section 5. Finally, we summarize our research work in section 6.

2 Related Work

In recent years, researchers have been interested in deep learning data generation tasks, and many of them have turned their attention to the application of GANs. The proposal of DP-GAN[27] opened the door to the study of differentially private GANs. Subsequently, more differentially private GANs based on DP-SGD[1] have been proposed, such as DP-CGAN[25], dp-GAN[28], DPautoGAN[24], RDP-GAN[18]. However, all these GANs are very difficult to train because they perturb the gradients of all parameters of the discriminator D .

Some recent research aimed to make a breakthrough in this problem. PATE-GAN[14] obtains differential privacy labels through the PATE mechanism and trains student models based on the datasets obtained from these labels, which can avoid the problem of too many parameters of the discriminator. G-PATE is an extension of the work of PATE-GAN, and it also trains the discriminator in a non-private way D . However, it need to address the problem that the high-dimensional gradient of the PATE framework needs to be downscaled. GS-WGAN[4] is the closest work to our approach, which perturbs the gradient of discriminator D toward generator G when training generator G , resulting in high-quality results. The experiments show that our approach can obtain the similar high-quality results as GS-WGAN, and even better than GS-WGAN for some datasets.

3 Preliminaries

In this section, we review the concept of differential privacy and self-attention.

3.1 Differential Privacy

Definition 1. (Differential privacy $((\varepsilon, \delta)\text{-DP})$ [5, 6]) A randomized mechanism \mathcal{M} with domain $\mathcal{N}^{|\mathcal{X}|}$ is (ε, δ) -differentially private if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $\mathcal{D}, \mathcal{D}' \in \mathcal{N}^{|\mathcal{X}|}$ such that $\|\mathcal{D} - \mathcal{D}'\|_1 \leq 1$:

$$\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq \exp(\varepsilon) \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{S}] + \delta \quad (1)$$

where the parameters ε and δ are non-negative real numbers. If $\delta = 0$, we say that \mathcal{M} is ε -differentially private.

3.2 Self-Attention

Most GANs are built by using convolutional layers as intermediate layers. Convolutional layers are inefficient in modeling long-distance dependence on images because they can only process local information, which is well compensated by the self-attention mechanism.

In order to use the self-attentive mechanism, it is necessary to transform the image features $x \in \mathbb{R}^{C \times N}$ of the implicit layer into two different feature spaces f and g , where $f(x) = W_f x$, $g(x) = W_g x$. Here C denotes the number of image channels, and N denotes the number of feature regions of the previous hidden layer feature. Then we can calculate the following equation:

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^N \exp(s_{ij})}, s_{ij} = f(x_i)^T g(x_j) \quad (2)$$

where $\beta_{j,i}$ denotes the degree of attention paid by the model to the i^{th} region when synthesizing the j^{th} region. The output of the attention layer is represented as $o = (o_1, o_2, \dots, o_N) \in \mathbb{R}^{C \times N}$, where

$$o_j = v \left(\sum_{i=1}^N \beta_{j,i} h(x_i) \right), h(x_i) = W_h x_i, v(x_i) = W_v x_i \quad (3)$$

Here W_g , W_f , W_h , W_v are implemented as 1×1 convolutions which are learned weight matrix.

Finally, the output of the attention layer is multiplied by a scale parameter γ and summed with the input feature map.

$$y_i = \gamma o_i + x_i \quad (4)$$

4 Our Approach

In this section, we present the problems with most of the current differentially private GANs and rethink the way differential privacy can be applied in model training in light of these problems. Then, we will introduce our proposed Differentially Private GAN based on Self-Attention (DP-SAGAN).

4.1 Rethinking Noise Addition in GAN Training

In the current mainstream research on differentially private GAN, more classical works such as DP-GAN[27], DP-WGAN[2], dp-GAN[28], and DP-CGAN[25] use a gradient perturbation method called DP-SGD[1], which is characterized by the fact that we add noise to the gradients derived from each sample in the model operation and bound these gradients to their L_2 norm, followed by the specification of the weights of the model in the range $[-C, +C]$ by using the constant C .

Although we can obtain data with differential privacy properties by using the DP-SGD method, the time we spent in model training may be increased by more than several times. Furthermore, since DP-SGD is a gradient perturbation of all parameters of the discriminator D , the whole training process may become very unstable. Therefore, we need a method to tackle these problems, so that we can add less noise and also get the same effect.

According to our observation, we find that we do not need to do gradient perturbation on the gradients of all parameters of D when training the discriminator D . We only need to ensure that the data synthesized by the generator G is differentially private according to the post-processing nature of differential privacy by perturbing the gradients of the flow from D to G when training the generator G . This allows us to add less noise and also enables the discriminator to guide the generator more accurately. Given the above observation, we design an approach of image generation with differential privacy, called DP-SAGAN. With the same privacy cost, our experiments show that DP-SAGAN can perform better than DP-GAN.

4.2 DP-SAGAN

Both the discriminator D and the generator G of our DP-SAGAN use four convolutional layers/deconvolutional layers and one self-attentive layer, where the self-attentive layer is located in the penultimate layer. Fig. 1 shows the self-attention mechanism of our DP-SAGAN. The functions f , g and h in the figure are $1 * 1$ convolutional layers, and \otimes stands for matrix multiplication. In particular, we deflate the self-attention feature map by using a learnable scalar γ at the final output of the self-attentive layer, and then add it to the input feature map. Such a structure is more similar to ResNet[12], which is more likely to learn features from the data. The gradient perturbation operation is based on differential privacy, and is performed on the self-attentive layer of the discriminator D in our proposed DP-SAGAN.

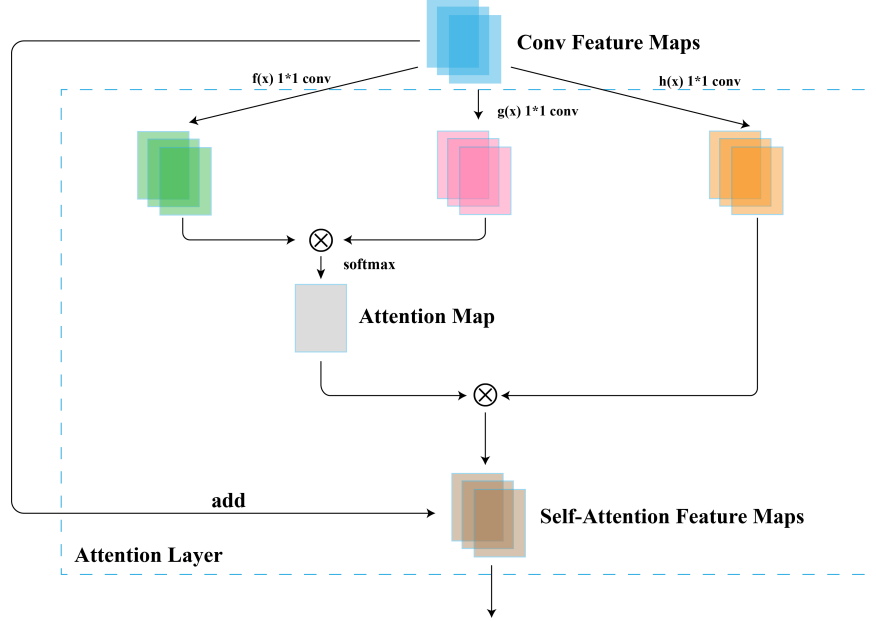


Fig. 1: The self-attention mechanism of our DP-SAGAN

After the gradient perturbation operation on the self-attentive layer, we do weight clipping on the self-attentive layer and its subsequent layers in discriminator D to restrict the parameter range to a constant C . However, whether we do gradient cropping or parameter cropping, the gradient information is significantly corrupted, so we use the Wasserstein distance[3] as the loss, and we also use the gradient penalty method[9] to ensure that the function of discriminator D always have the 1-Lipschitz property.

5 Experimental evaluation

To show the performance of DP-SAGAN, we design experiments from both quantitative and qualitative perspectives. In this section, we will first compare the images synthesized by DP-SAGAN with other mainstream models, then analyze our method qualitatively by quality metrics in GAN, and finally, show the classification performance of our method in downstream classifiers.

5.1 Experimental setting

In this subsection, we will describe the parameter settings we used in our experiments. To ensure uniformity and consistency of the experiments, our parameters of differential privacy are fixed at $(\epsilon, \delta) = (10, 10^{-5})$.

Datasets. We used MNIST[16], Fashion-MNIST[26], and CIFAR-10[15] as the datasets for our quantitative experiments.

Evaluation Metrics. The utility of the data generated by our approach are judged in two ways. Firstly, it is the quality aspect of the images generated by the model, which we evaluate by using the Inception Score[21] (IS) and the Frechet Inception Distance[13] (FID). Then, in terms of utility for downstream tasks, we test the classification accuracy of the generated images by using logistic regression, SVM (Support Vector Machine) and MLP (Multilayer Perceptron) methods.

Baselines. To highlight the prospective nature of our work, we use the four most commonly used models in the current differentially private GAN field: DP-MERF[10], G-PATE[17], GS-WGAN[4], and DP-GAN[27] as baseline models. Also, we use two non-differentially private models, DCGAN[20] and SAGAN[29], to compare the difference with differentially private GANs.

5.2 Qualitative Analysis

Table 1: Generated samples with differential privacy $(\epsilon, \delta) = (10, 10^{-5})$


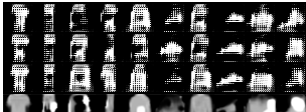
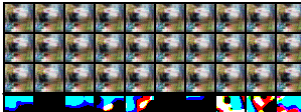
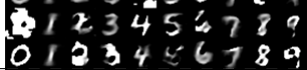

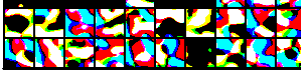

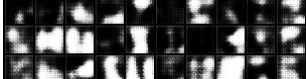
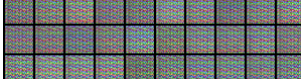
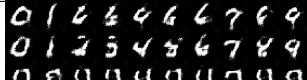



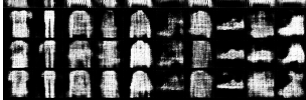
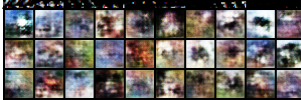
Method	MNIST	Fashion-MNIST	CIFAR-10
DP-GAN			
DP-MERF			
G-PATE			
GS-WGAN			
Ours			

Table 1 shows the samples generated by various models with the differential privacy $(\epsilon, \delta) = (10, 10^{-5})$. It is clear that DP-GAN, trained by using the traditional DP-SGD method, generates samples with a strong sense of dispersion, which is evident in the samples generated by Fashion-MNIST. DP-MERF uses a random feature kernel approach with differential privacy to train the model

rather than a form of GAN, and it has the fastest training speed of the comparison models, but it generates samples with significant flaws in some of the samples that the other models do not have. G-PATE is based on the PATE framework and uses an ensemble of teacher discriminators to train student generator, and while the approach outperforms previous work in the PATE[19] domain, it performs the worst compared to other models for the same parameters of differential privacy. GS-WGAN uses a similar approach to DP-SAGAN, but our approach is better in terms of details. For example, in our experiments on MNIST datasets, GS-WGAN prefers to generate the number 0 while we are able to retain the details better.

5.3 Quality of Generated Samples

Table 2: Quantitative results with differential privacy $(\epsilon, \delta) = (10, 10^{-5})$

Method	MNIST		Fashion-Mnist		CIFAR-10	
	Diff IS	FID	Diff IS	FID	Diff IS	FID
DCGAN	4.60	151.51	1.44	109.81	3.17	168.73
SA-GAN	0.42	108.76	1.39	113.74	3.11	128.12
DP-GAN	5.31	177.67	2.45	277.48	4.97	486.08
G-PATE	6.53	251.61	4.40	199.13	5.07	470.47
GS-WGAN	2.83	58.38	2.57	103.46	3.86	281.36
Ours	2.32	110.75	1.87	151.37	3.72	172.76

Table 2 shows our quantitative results for various models with differential privacy $(\epsilon, \delta) = (10, 10^{-5})$. The Diff IS metric indicates the difference in IS between our generated samples and the real samples. Considering the network structure, we only use the GAN-based models here. It can be seen from the IS difference that our method outperforms all other differentially private GAN methods on all datasets. Surprisingly, GS-WGAN performs even better than the native SA-GAN on FID. The work of [23] demonstrates that the effect of differential privacy is related to the structure of the native model, so we speculate that it is caused by the ResNet skeleton used by GS-WGAN being able to learn the features in the dataset better. It should be noted that in contrast to the non-differentially private SA-GAN comparison, our approach only loses some points.

5.4 Utility of Model-generated Samples in Downstream Tasks

In this subsection, we train the classifiers by using synthetic datasets and evaluate the utility of the model in downstream classifiers by using real datasets for k -fold cross-validation. For convenience, we only show the results when $k = 5$.

Table 3 shows the average accuracy of the three downstream classifiers with three different datasets. From the table, we can clearly see that our DP-SAGAN

Table 3: Average classification performance of the generative model with differential privacy $(\epsilon, \delta) = (10, 10^{-5})$ on three downstream classifiers of logistic regression, SVM, and MLP with 5-fold cross-validation

Method	MNIST	Fashion-MNIST	CIFAR-10
DP-GAN	0.33	0.30	0.11
DP-MERF	0.76	0.70	0.16
G-PATE	0.16	0.15	0.09
GS-WGAN	0.62	0.55	0.10
Ours	0.77	0.62	0.25

performs the best among all models for the MNIST dataset and the CIFAR-10 dataset. Only for the Fashion-MNIST dataset, DP-MERF is slightly better than ours, which may be due to the use of a random feature kernel function for DP-MERF. More notably, the CIFAR-10 samples generated by our DP-SAGAN have the classification performance of more than 0.2 instead of 0.1 in all three classifiers, which means that our DP-SAGAN does learn useful features from the dataset.

6 Conclusion

In this paper, we first discuss the current problems that still exist for differentially private GANs and rethink the differential privacy implementation in GANs to address these problems. Based on our thinking, we propose to achieve differential privacy by perturbing the gradient of the discriminator D toward the generator G while training the generator G and thus achieving differential privacy. Based on this method, and considering the fact that convolution cannot model images over long distances, we propose DP-SAGAN and experimentally confirm that DP-SAGAN is not only more stable in training but also can obtain better high-resolution images.

Acknowledgments. This work was supported in part by the Scientific & Technological Innovation 2030 - “New Generation AI” Key Project (No. 2021ZD0114001; No. 2021ZD0114000), and the Science and Technology Commission of Shanghai Municipality (No. 21511102200).

References

1. Abadi, M., Chu, A., Goodfellow, I.J., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016. pp. 308–318. ACM (2016)
2. Alzantot, M., Srivastava, M.: Differential Privacy Synthetic Data Generation using WGANs, 2019. URL https://github.com/nesl/nist_differential_privacy_synthetic_data_challenge

3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. CoRR **abs/1701.07875** (2017)
4. Chen, D., Orekondy, T., Fritz, M.: GS-WGAN: A gradient-sanitized approach for learning differentially private generators. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual* (2020)
5. Dwork, C.: Differential Privacy: A Survey of Results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) *Theory and Applications of Models of Computation*. pp. 1–19. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
6. Dwork, C., Roth, A.: The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* **9**(3–4), 211–407 (2014)
7. Fredrikson, M., Jha, S., Ristenpart, T.: Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. pp. 1322–1333. CCS '15, Association for Computing Machinery, New York, NY, USA (2015)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. *Commun. ACM* **63**(11), 139–144 (2020)
9. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp. 5767–5777 (2017)
10. Harder, F., Adamczewski, K., Park, M.: DP-MERF: differentially private mean embeddings with random features for practical privacy-preserving data generation. In: *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event. Proceedings of Machine Learning Research*, vol. 130, pp. 1819–1827. PMLR (2021)
11. Hayes, J., Melis, L., Danezis, G., Cristofaro, E.D.: LOGAN: evaluating privacy leakage of generative models using generative adversarial networks. CoRR **abs/1705.07663** (2017)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pp. 770–778. IEEE Computer Society (2016)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. pp. 6626–6637 (2017)
14. Jordon, J., Yoon, J., van der Schaar, M.: PATE-GAN: generating synthetic data with differential privacy guarantees. In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net* (2019)
15. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases* **1**(4) (2009)
16. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324 (1998)
17. Long, Y., Wang, B., Yang, Z., Kailkhura, B., Zhang, A., Gunter, C.A., Li, B.: G-PATE: scalable differentially private data generator via private aggregation of teacher discriminators. In: *Advances in Neural Information Processing Systems*

- 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 2965–2977 (2021)
18. Ma, C., Li, J., Ding, M., Liu, B., Wei, K., Weng, J., Poor, H.V.: RDP-GAN: A rényi-differential privacy based generative adversarial network. *IEEE Trans. Dependable Secur. Comput.* **20**(6), 4838–4852 (2023)
19. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, Ú.: Scalable private learning with PATE. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
20. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
21. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 2226–2234 (2016)
22. Shokri, R., Stronati, M., Song, C., 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. IEEE Computer Society (2017)
23. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic data - anonymisation groundhog day. In: 31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022. pp. 1451–1468. USENIX Association (2022)
24. Tantipongpipat, U.T., Waites, C., Boob, D., Siva, A.A., Cummings, R.: Differentially private synthetic mixed-type data generation for unsupervised learning. *Intell. Decis. Technol.* **15**(4), 779–807 (2021)
25. Torkzadehmahani, R., Kairouz, P., Paten, B.: DP-CGAN: differentially private synthetic data and label generation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 98–104. Computer Vision Foundation / IEEE (2019)
26. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR* **abs/1708.07747** (2017)
27. Xie, L., Lin, K., Wang, S., Wang, F., Zhou, J.: Differentially private generative adversarial network. *CoRR* **abs/1802.06739** (2018)
28. Zhang, X., Ji, S., Wang, T.: Differentially private releasing via deep generative model. *CoRR* **abs/1801.01594** (2018)
29. Zhao, H., Jia, J., Koltun, V.: Exploring Self-Attention for Image Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2020)