Supplementary Material for
What can Discriminator do? Towards Box-free Ownership Verification of Generative Adversarial Networks

Ziheng Huang[1†], Boheng Li[1†], Yan Cai[1], Run Wang[1*], Shangwei Guo[2], Liming Fang[3], Jing Chen[1], Lina Wang[1]

[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China

[2] College of Computer Science, Chongqing University, China

[3] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China

[†] Equal contribution [*] Corresponding author. E-mail: wangrun@whu.edu.cn

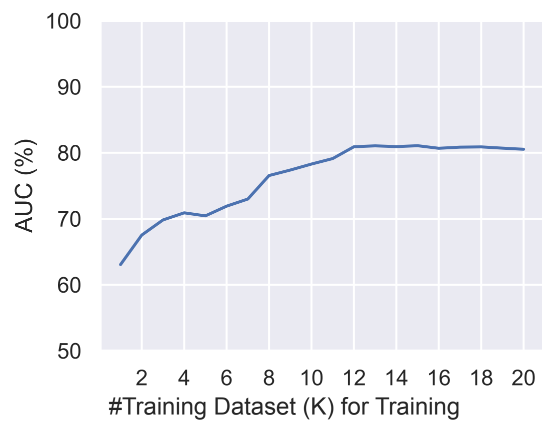
1. Overview

In this supplementary material, we present a deep understanding of our proposed method and more experimental results and analysis.

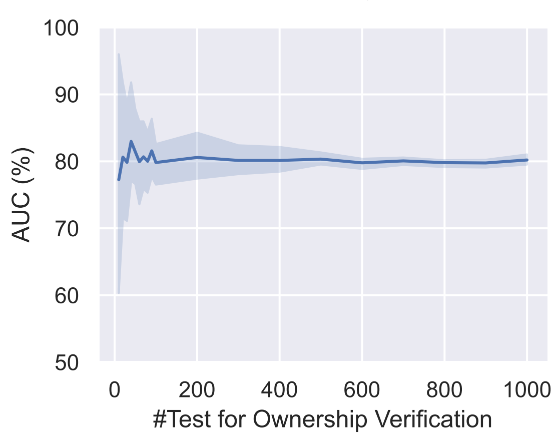
- We present the ablation study to explore the impact of training set and test set in determining the ownership verification performance.
- We analyze the impacts of using Pearson coefficient loss and Mean Squared Error (MSE) loss respectively on GAN training.
- We show the comprehensive experimental results of effectiveness evaluation with AUC and AP.
- We provide further discussion on the scalability of our method and demonstrate why it can be applied to more complex scenarios.
- We analyze the similarities and differences between our method and anomaly detection from both theoretical and experimental perspectives.

2. Ablation Study

In this section, we explore the impact of training set and test set in determining the performance of our one-class classifier. Experimental results in Figure 1 (a) illustrate that the performance tends to be stable when the number of training samples lies in a range between 12,000 and 15,000. In exploring the number of input samples for determining the ownership, we feed different batches of samples to investigate the relationship between test samples and AUC score. Figure 1 (b) demonstrated that our method could determine the ownership of synthesized images effectively and steadily when the number of input samples is over 500.

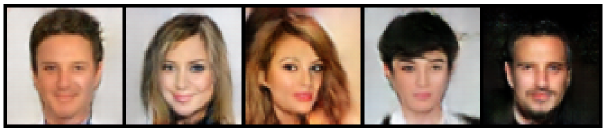


(a) Numbers of training set.

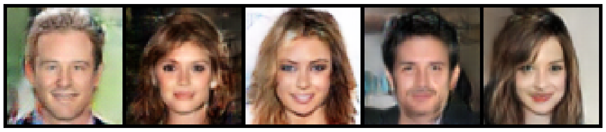


(b) Numbers of test set.

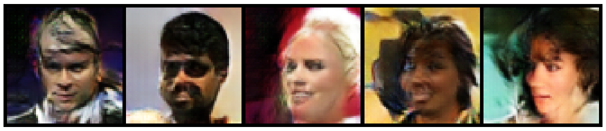
Figure 1: Ablation studies on the number of training set and test set.



(a) Original



(b) w/ Pearson Correlation Loss



(c) w/ MSE Loss

Figure 2: Visualization of synthesized images generated under different loss function conditions.

	no additional loss	Pearson Correlation loss	MSE loss
DCGAN	34.03	35.78	**71.54**
SNDCGAN [9]	31.15	30.46	**68.33**
ResDCGAN [4]	32.90	32.14	**66.29**

Table 1: FID score (↓ means better) of the generated images under three conditions. The GANs are trained on CelebA dataset.

3. Pearson Correlation Loss vs. MSE Loss

In this section, we respectively evaluate the effects of using the Pearson correlation coefficient and the MSE loss as the additional loss term on the training of the GAN, compared to not using any additional loss.

Tab. 1 presents the FID score [6] of the generated images of three DCGAN [10] models trained on the CelebA dataset [8] under three different conditions. The experimental results indicate that incorporating the Pearson correlation negatively impact on the performance of the GAN. The quality of the generated images does not decrease when the Pearson correlation coefficient is incorporated. In contrast, leveraging the Mean Squared Error (MSE) loss has a detrimental effect on the quality of the generated images. Fig. 2 are the samples of the generated images under the three different conditions.

4. Effectiveness Evaluation with AP

Due to the limited space in the main manuscript, we present the effectiveness evaluation measured by both AUC and AP in this section. Tab. 2 and Tab. 3 present the effectiveness of our proposed method in determining the ownership of GANs measured by both AUC and AP for the entire image synthesis. Tab. 4 shows the results for the image-to-image translation. Experimental results demonstrated the effectiveness of our methods and reach the same conclusion presented in the main manuscript.

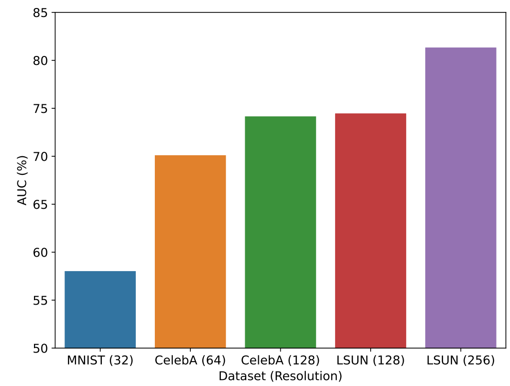


Figure 3: The AUC trend on complexity of tasks GAN. The content inside the parentheses refers to the resolution of the corresponding training set.

5. Scalability

In this section, we provide further discussion and analysis on scalability. We conduct an experiment to analyze the trend of our method's performance when the task becomes more complex. Specifically, we gradually increase the resolution of the training and generated images. For generating higher-resolution images, we use more advanced model architectures.

The results in Fig. 3 show that the AUC score has a gradual upward trend, which indicates that our method has high confidence in spotting different GANs when applied to GANs that cope with more challenging tasks. A possible explanation lies in that the synthesized images on complex tasks carry more unique artifacts, providing a clearer signal for ownership verification. Experimental results show the potential of our method in tackling challenging datasets with SOTA GAN architectures in real scenarios.

6. Comparison and Discussion with Anomaly Detection.

In this section, we compare our method with anomaly detection and discuss their similarities and differences.

Anomaly detection aims to identify rare or abnormal events, patterns or behaviors in data [1]. The fundamental objective of anomaly detection is to distinguish between normal data and anomalous ones. Anomalies are typically rare data instances, in contrast to normal instances that comprise a majority of the data population. It is difficult to collect a large amount of labeled abnormal instances.

Due to imbalanced data classes, our method and anomaly detection techniques utilize only a single class of data (i.e., one-class classification). The concept of learning a hypersphere has also been applied in many other data description methods and anomaly detection methods. However, our

Models	DCGAN		SNDCGAN		ResDCGAN	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
DCGAN	51.18±0.22	50.84±0.86	70.79±0.56	68.80±0.40	69.97±0.84	71.26±0.59
SNDCGAN	72.21±0.93	72.89±0.97	51.18±0.22	50.73±0.39	71.13±0.96	69.28±0.48
ResDCGAN	71.81±0.35	70.24±0.22	72.30±0.49	74.08±0.76	50.90±0.25	53.10±0.58

(a) Model architectures

Train. Set	Subset 1		Subset 2		Subset 3	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
Subset 1	51.18±0.22	50.92±0.81	75.63±0.18	74.64±0.49	74.80±0.88	71.79±0.32
Subset 2	75.66±1.37	77.89±1.09	50.89±0.36	51.94±0.41	78.81±0.80	76.54±0.96
Subset 3	73.40±1.20	70.95±0.49	75.54±1.64	77.03±0.78	52.02±0.61	50.38±0.79

(b) Training sets

Seeds	Seed 1111		Seed 2222		Seed 3407	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
Seed 1111	51.18±0.22	50.04±0.58	61.24±0.56	59.55±0.39	60.17±0.30	62.59±0.63
Seed 2222	60.21±0.32	62.53±0.50	50.79±0.34	52.12±0.37	62.31±1.28	60.47±0.56
Seed 3407	59.93±0.19	62.29±0.61	60.15±0.41	60.72±0.69	50.87±0.23	52.83±0.40

(c) Random seeds.

Table 2: Effectiveness evaluation measured by AUC and AP on LSUN dataset for entire image synthesis.

Models	DCGAN		SNDCGAN		ResDCGAN	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
DCGAN	50.42±0.43	51.66±0.85	70.33±1.26	72.45±0.50	73.18±0.47	70.51±0.61
SNDCGAN	73.20±1.24	69.85±0.75	50.13±0.17	51.28±0.31	75.61±1.68	72.31±1.51
ResDCGAN	74.72±0.89	74.23±0.44	72.95±1.10	70.84±0.94	50.56±0.40	52.41±0.51

(a) Model architectures.

Training Set	Subset 1		Subset 2		Subset 3	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
Subset 1	50.42±0.43	52.16±0.80	78.41±1.79	75.65±1.34	76.35±0.83	73.57±0.77
Subset 2	76.77±1.52	75.52±1.66	51.06±0.33	50.89±0.65	75.86±0.74	76.77±0.34
Subset 3	78.63±1.48	76.58±0.99	77.32±0.49	78.10±0.73	51.44±0.27	50.95±0.56

(b) Training sets.

Seeds	Seed 1111		Seed 2222		Seed 3407	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
Seed 1111	50.42±0.43	52.38±0.87	64.94±0.97	65.31±1.56	63.14±0.48	61.10±0.69
Seed 2222	61.73±0.61	61.17±0.30	50.76±0.32	52.30±0.41	60.43±0.37	62.07±1.45
Seed 3407	60.04±0.26	58.99±0.28	59.97±0.50	62.12±0.64	50.66±0.51	52.43±1.55

(c) Random seeds.

Table 3: Effectiveness evaluation measured by AUC and AP on CelebA dataset for entire image synthesis.

method has an essentially different objective compared to existing anomaly detection methods. While anomaly detection is expected to accurately distinguish between normal and abnormal data points, our method focuses on detecting differences between data distributions. It can determine if there is a significant difference between the overall performance of a batch of images, without the strict constraints of identifying each individual data point as normal or abnormal. Moreover, the existing anomaly detection methods, analogous to the GAN attribution methods discussed in the main paper, suffer the limitations including i) all techniques require training a powerful external one-class classifier, which is both time and resource-consuming; ii) similar to attribution classifiers, the adversary can easily reproduce the one-class discriminator and perform ambiguity attacks and iii) the complexity of GAN learned distributions make anomaly detection methods a bad performance. Actually, some anomaly detection methods can be directly applied to training, and therefore we select several representative methods [13, 12, 11, 3] to compare with our method. We train two DCGAN models on

Models	StarGAN		AttGAN		STGAN	
	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑	AUC (%) ↑	AP (%) ↑
StarGAN [2]	50.86±1.63	52.03±0.79	74.58±1.20	72.97±0.47	77.48±1.74	78.90±0.58
AttGAN [5]	74.23±0.75	72.06±0.90	50.92±0.96	51.46±0.41	73.84±0.67	71.53±0.93
STGAN [7]	74.42±0.81	72.73±0.64	76.59±1.42	74.78±0.93	51.47±0.66	52.84±0.74

Table 4: Effectiveness evaluation measured by AUC and AP on CelebA dataset for image-to-image translation.

CelebA dataset, with only the initial seeds different. Then, we follow the original paper to train the anomaly detection models with the source model's generated images and use the models to differentiate given images. Experimental results in Tab. 5 show that our method outperforms state-of-the-art anomaly detection methods with a notable margin.

However, one must note that, although some methods seemingly have an acceptable performance (e.g., $\sim 58\%$), the ambiguity attack is whatsoever a sword of Damocles (refer to Section 4.4 in the main paper). After an attacker steals the generative model, he can train an anomaly detection model that performs similarly to the owner's one. With both the attacker and owner having the model as the credential, ownership is in doubt. In contrast, our method can resist ambiguity attack because we utilize the discriminator which is unique during each training and irreproducible. Consequently, even if the attacker knows our method's training pipeline, he can not obtain a model that is equivalent to the owner's. Thus, this fundamental limitation makes the anomaly detection methods infeasible to the task of ownership verification, which also essentially differs our method from the anomaly detection.

Training strategy	AUC (same) ↓	AUC (different) ↑
DCAE [13]	51.83	53.98
AnoGAN [12]	54.49	58.35
Deep SVDD [11]	51.91	55.21
DROCC [3]	50.64	54.84
Ours	**50.42**	**64.94**

Table 5: Performance in ownership verification among different anomaly detection methods. We separately train two DCGAN models with only initial seeds different. The column *same* indicates the verification of the images generated by the paired G, while the column *different* denotes the verification for different GAN models.

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 2

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 4

[3] Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *International conference on machine learning*, pages 3711–3721. PMLR, 2020. 3, 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478, 2019. 4

[6] 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*, 30, 2017. 2

[7] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3673–3682, 2019. 4

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2

[9] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2

[10] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[11] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 3, 4

[12] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone,*

NC, USA, June 25-30, 2017, Proceedings, pages 146–157. Springer, 2017. 3, 4

[13] Philipp Seeböck, Sebastian Waldstein, Sophie Klimscha, Bianca S Gerendas, René Donner, Thomas Schlegl, Ursula Schmidt-Erfurth, and Georg Langs. Identifying and categorizing anomalies in retinal imaging data. *arXiv preprint arXiv:1612.00686*, 2016. 3, 4