IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 30, 2021, accepted December 13, 2021, date of publication December 21, 2021,
date of current version December 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3137278

Evaluating Differentially Private Generative Adversarial Networks Over Membership Inference Attack

CHEOLHEE PARK[1], YOUNGSOO KIM[1], JONG-GEUN PARK[1], DOWON HONG[2], AND CHANGHO SEO[2]

[1]Electronics and Telecommunications Research Institute (ETRI), Yuseong-gu, Daejeon 34129, South Korea
[2]Department of Mathematics, Kongju National University, Gongju 32588, South Korea

Corresponding author: Dowon Hong (dwhong@kongju.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant by the Korean Government through Ministry of Science and ICT (MSIT) (Development of 5G Edge Security Technology for Ensuring 5G+ Service Stability and Availability) under Grant 2020-0-00952, and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT under Grant 2019R1A2C1003146.

ABSTRACT As communication technology advances with 5G, the amount of data accumulated online is explosively increasing. From these data, valuable results are being created through data analysis technologies. Among them, artificial intelligence (AI) has shown remarkable performances in various fields and is emerging as an innovative technology. In particular, machine learning and deep learning models are evolving rapidly and are being widely deployed in practical applications. Meanwhile, behind the widespread use of these models, privacy concerns have been continuously raised. In addition, as substantial privacy invasion attacks against machine learning and deep learning models have been proposed, the importance of research on privacy-preserving AI is being emphasized. Accordingly, in the field of differential privacy, which has become a de facto standard for preserving privacy, various mechanisms have been proposed to preserve the privacy of AI models. However, it is unclear how to calibrate appropriate privacy parameters, taking into account the trade-off between a model's utility and data privacy. Moreover, there is a lack of research that analyzes the relationship between the degree of differential privacy guarantee and privacy invasion attacks. In this paper, we investigate the resistance of differentially private AI models to substantial privacy invasion attacks according to the degree of privacy guarantee, and analyze how privacy parameters should be set to prevent the attacks while preserving the utility of the models. Specifically, we focus on generative adversarial networks (GAN), which is one of the most sophisticated AI models, and on the membership inference attack, which is the most fundamental privacy invasion attack. In the experimental evaluation, by quantifying the effectiveness of the attack based on the degree of privacy guarantee, we show that differential privacy can simultaneously preserve data privacy and the utility of models with moderate privacy budgets.

INDEX TERMS Differential privacy, artificial intelligence, deep learning, generative adversarial networks, privacy-preserving deep learning, membership inference attack.

I. INTRODUCTION

With the development of 5G communication technology that diversifies the access environment and materializes distributed networks, various types and vast amounts of data are being accumulated online. From these data, valuable results are being created through data analysis technologies. In particular, machine learning and deep learning technologies have been widely used and have shown remarkable performances in various areas such as classification, language representation, recommendations, synthetic data generation, etc. (e.g., [1]–[3]). Moreover, with the introduction of machine learning as a service (MLaaS), which is a range of machine learning functionality offered by cloud service providers, the use of artificial intelligence (AI) models is becoming more active. Typically, these models are generated

The associate editor coordinating the review of this manuscript and approving it for publication was S. K. Hafizul Islam.

by learning massive amounts of raw data, and this can lead to revealing sensitive individual information.

Indeed, along with the widespread deployment of artificial intelligence models, concerns about privacy violations have been raised. In addition, as substantial privacy invasion attacks on AI models have been proposed recently [4]–[10], the importance of research on privacy-preserving AI has been emphasized. Accordingly, various approaches have been introduced to preserve the privacy of AI models. Among them, differential privacy [11], [12], which has become a de facto privacy standard, provides a rigorous privacy guarantee, and various mechanisms that satisfy the properties of differential privacy have been proposed for designing privacy-preserving AI.

Generally, differentially private mechanisms return noisy outputs that obscure statistical differences between adjacent databases, and the magnitude of the noise that will be added to the actual output for a specific query is highly dependent on the privacy parameter ϵ, called the privacy budget. In other words, the lower the privacy budget, the larger the noise, and vice versa. Obviously, from the perspective of utility as well as privacy, the choice of ϵ is one of the most important factors, and ϵ should be calibrated with in-depth consideration of the trade-off between privacy and utility. However, the criterion for how to set an appropriate privacy budget has not been clearly established in practice, and then differentially private AI models have often set the privacy budget ϵ as a tendency to ensure acceptable utility. As a result, the utility of the models may be able to be guaranteed, but privacy may not be preserved at all. This ambiguity is a well-known problem in the field of differential privacy, and we aim to address this problem by analyzing the relationship between differential privacy and substantial privacy invasion attacks for AI models.

In this paper, we evaluate the resistance of differentially private AI models to substantial privacy invasion attacks by varying the privacy budget ϵ, and analyze how privacy parameters should be set to prevent the attacks while preserving the utility of the models. Furthermore, we study the efficacy of privacy invasion attacks under the relaxed notions of differential privacy (i.e., concentrated differential privacy [13], zero concentrated differential privacy [14], and Rényi differential privacy [15]), and analyze how much of a privacy breach occurs by relaxing the definition of differential privacy on AI models. In particular, we focus on the generative model, especially generative adversarial networks (GAN), which is one of the most sophisticated models for generating synthetic datasets and has attracted great interest recently. In the case of attack scenario, we focus on the membership inference attack, which is the most fundamental privacy invasion attack.

Currently, several results have been reported for evaluating differentially private AI models under substantial privacy invasion attacks [16]–[20]. However, these results have mainly focused on neural network-based models or regression models, and there are no results for models with the objective of generating synthetic dataset. In this respect, we aim to contribute to the evaluation of differentially private mechanisms for generative models by analyzing the relationship between the degree of differential privacy guarantee and the privacy invasion attack.

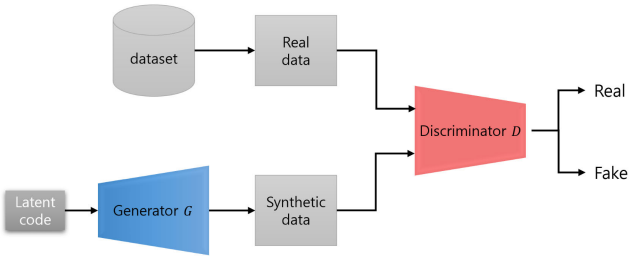


FIGURE 1. Architecture of generative adversarial networks.

The rest of this paper is organized as follows. First, there is a background review in section 2. Then, we present differentially private mechanisms and the membership inference attack for GAN models in section 3. In section 4, we describe our evaluation framework, and demonstrate experiment and evaluation results in detail. Finally, we discuss related studies in section 5, and then conclude our work in section 6.

II. BACKGROUND

In this section, we briefly illustrate generative adversarial networks (GAN), and review the definition of differential privacy and its relaxations. Then, we demonstrate mechanisms that can make GAN models differentially private.

A. GENERATIVE ADVERSARIAL NETWORKS

Generative models are designed for learning the probability distribution of a given training data, and have the purpose of generating synthetic data close to the real data. Among the various generative models, great interest has been focused on generative adversarial networks (GAN) [21], and numerous studies have been conducted to advance their performance and functionality. As shown in Figure 1, the basic architecture of GAN consists of two neural network-based components: a generator G and discriminator D. The generator G takes noise $\mathbf{z}$ (latent code) as an input and generates synthetic data $\mathbf{x}'$ with the objective of generating data that approximates the real data $\mathbf{x}$ while the discriminator D takes a dataset consisting of the synthetic data and real data with the objective of discriminating the difference between real (training data $\mathbf{x}$) and fake (synthetic data $\mathbf{x}'$). Therefore, these two components always play a game to beat each other, and are trained alternately.

More formally, let $p_{\mathbf{z}}$ be the probability distribution of the latent code and p_{data} be the probability distribution of the real data. Then the objective function $V(D, G)$ of the GAN model that consists of G and D is a minimax game, and can

be formulated as follows.

$$\min_G \max_D \mathbb{E}_{\mathbf{x}\sim p_{data}}[\log(D_{\theta_D}(\mathbf{x}))] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1 - D_{\theta_D}(G_{\theta_G}(\mathbf{z})))],$$

where θ_D and θ_G denote the parameters of the discriminator and generator, respectively. Therefore, the discriminator is trained to return a high score in a given training samples (real data), and the generator is trained to produce synthetic data that can maximize the discriminator's output. After sufficient training, if a Nash equilibrium is achieved, both the discriminator and generator settle at a point where there is no further improvement.

Since the basic concept of GAN was introduced, numerous variants have been proposed with the aim of evolving the original model by adjusting the objective function or by modifying the architecture (e.g., [22]–[27]). Among these variants, we target several significant models that have shown noticeable improvement: 1) deep convolutional GAN (DCGAN) [23], a model that combines the basic GAN architecture with a convolutional neural network. 2) Wasserstein GAN (WGAN) [25], a model that improves training stability by using the Wasserstein distance (instead of Jensen–Shannon divergence in the original GAN model) as an approximation metric between probability distributions. 3) boundary equilibrium GAN (BEGAN) [26], a model that can approximate the convergence of the training process by combining with the concept of Autoencoder.

B. PRIVACY INVASION ATTACKS ON AI MODELS

Along with the advancement of AI technology, privacy concerns have been raised simultaneously, and various attacks that can substantially invade privacy on machine learning and deep learning models have been proposed. Ateniese *et al.* [28] showed that it is possible to infer the general statistical information about a training dataset by exploiting the internal parameters of specific models (such as Support Vector Machines and Hidden Markov Models). For a collaborative recommender system, Calandrino *et al.* [29] reported that, by capturing changes between outputs, an attacker can infer specific inputs that triggered those changes. As an attack that can directly invade the privacy of training data, Fredrickson *et al.* [4], [5] proposed the model inversion attack, where an attacker can reconstruct parts of information in the training dataset by exploiting confidence vectors returned along with predictions from a target model. Tramèr *et al.* [7] introduced the model extraction attack that can extract parameters of the target model, and showed that sensitive information of the training dataset can be exposed. As a fundamental privacy invasion attack, Shokri *et al.* [6] proposed the membership inference attack, where an attacker can infer whether or not specific input data were included in the training dataset. The underlying intuition of the attack is that a machine learning model (trained model) will behave differently between the data that the model has learned (i.e., training data) and unseen data, and these differences become more severe when the model is overfitted to the training data. In order to construct a (membership inference) attack model, an adversary has to build multiple shadow models that mimic the target model, and these shadow models can be built by learning the confidence vectors obtained from the target model (by querying the target model with arbitrary inputs). Then, the attack model can be constructed by learning the results (confidence vectors) output from the shadow models for both data with and without membership. Note that, in the case of the shadow model, an attacker can know exactly whether a given data was included in the training dataset of the shadow model.

Since the concept of membership inference attack against general machine learning models was introduced, various studies have been conducted to analyze and advance the attack [8]–[10], [30]. In particular, attack methods have been proposed that focus on generative models with a different aspect from previous studies that targeted general machine learning models. Since the outputs of a generative model are synthetic data rather than predictions, it is necessary to consider a different approach from the previous methods. By capturing these points, Hayes *et al.* [8] proposed an attack method with the goal of membership inference against GAN models. Subsequently, inspired by Hayes *et al.*'s study, several attack methods have been proposed with different assumptions and attack scenarios [9], [10] in terms of distance metric. We focus on these attacks as our goal is to analyze the relationship between substantial privacy invasion attacks and privacy-preserving techniques over GAN models (a detailed analysis of each membership inference attack against GAN is covered in section 3).

C. DIFFERENTIAL PRIVACY

Differential privacy [11], [12] has become a de facto privacy standard and ensures strong privacy preservation. Intuitively, a mechanism that satisfies differential privacy returns similar outputs on adjacent datasets for a given query. It means that differentially private mechanisms provide plausible deniability against adversaries.

If two datasets $d, d' \in D$ differ in one entry, we say that the datasets are adjacent (or neighboring). Then, the definition of differential privacy is as follows.

Definition 1 ((ϵ, δ)-Differential Privacy ((ϵ, δ)-DP) [11], [12]): A randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy, for all output $S \subseteq Range(\mathcal{M})$ for any two adjacent datasets d, d', if we have:

$$Pr[\mathcal{M}(d) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(d') \in S] + \delta.$$

Obviously, a smaller value of privacy cost (privacy budget) ϵ leads to a better privacy guarantee, and the value of δ is generally set to be smaller than the inverse of any polynomial in the size of database. When the additive term δ is zero, it is called ϵ-differential privacy (pure differential privacy).

In general, differential privacy can be achieved by adding noise to the actual output for a query function, and the

magnitude of noise is estimated depending on the *sensitivity* of the query function.

Definition 2 (l_2-Sensitivity): For any two adjacent datasets d, d', the sensitivity of query function f is defined as follows:

$$\Delta f = \max_{d,d'} \|f(d) - f(d')\|_2.$$

There are several basic mechanisms that ensure differential privacy, including Laplace mechanism and Gaussian mechanism, and we focus on the Gaussian mechanism, which has been widely leveraged to achieve differential privacy for AI models.

Definition 3 (Gaussian Mechanism): For any dataset d and query function $f(\cdot)$, the Gaussian mechanism $\mathcal{M}_G$ is defined as follows:

$$\mathcal{M}_G(d) = f(d) + \mathcal{N}(0, \sigma^2),$$

where $\mathcal{N}(0, \sigma^2)$ is the Gaussian distribution with mean 0 and standard deviation σ.

Note that a single execution of the Gaussian mechanism satisfies (ϵ, δ)-DP if $\sigma > 2\ln 1.25/\delta \cdot \Delta f/\epsilon$ and $\epsilon < 1$. Basically, in terms of running multiple times, differential privacy provides the *composition* property. If two mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$ satisfy ϵ_1 and ϵ_2-DP respectively, then a family of the mechanisms $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2)$ satisfies $(\epsilon_1 + \epsilon_2)$-DP.

Since the definition of differential privacy was introduced, several notions that can relax the original definition of differential privacy have been proposed to analyze a tighter bound in terms of cumulative *privacy loss* over multiple executions by considering the fact that the privacy loss random variable is tightly concentrated around its expectation. There are three commonly used relaxed definitions of differential privacy: concentrated differential privacy [13], zero concentrated differential privacy [14], Rényi differential privacy [15].[1]

In a subsequent study of pure and (ϵ, δ)-differential privacy, Dwork *et al.* [13] introduced concentrated differential privacy (CDP) by focusing on the case where the privacy loss follows a sub-Gaussian distribution. The intuition embedded in the notion of CDP is that the privacy loss is strictly centered around its expectation and the tail is managed by the variance of the sub-Gaussian distribution.

Definition 4 (Concentrated Differential Privacy (CDP) [13]): A randomized algorithm $\mathcal{M}$ is (μ, τ)-concentrated differentially private if, for all pairs of adjacent databases d, d', we have:

$$D_{subG}(\mathcal{M}(d)\|\mathcal{M}(d')) \preceq (\mu, \tau),$$

where D_{subG} denotes the sub-Gaussian divergence.

The definition means that the expected privacy loss is bounded by μ and the distribution of the centered privacy loss (by abstracting μ) is sub-Gaussian with standard deviation τ. In terms of relevance to the previous notion, the authors showed that if a mechanism $\mathcal{M}$ satisfies ϵ-DP algorithm, then $\mathcal{M}$ ensures $(\epsilon \cdot (e^\epsilon - 1)/2, \epsilon))$-CDP (but the converse does not hold). In addition, they showed that the Gaussian mechanism defined above satisfies $(\tau^2/2, \tau)$-CDP with $\tau = \Delta f/\sigma$.

In a subsequent study on CDP, Bun *et al.* [14] proposed the notion of zero-concentrated differential privacy (zCDP). By reformulating the concept of CDP through the Rényi divergence, they analyzed a tighter bound on the cumulative privacy loss over multiple computations.

Definition 5 (Zero-Concentrated Differential Privacy (zCDP) [14]): A randomized mechanism $\mathcal{M}$ is (ξ, ρ)-zero-concentrated differentially private if, for all adjacent databases d, d' and all $\alpha \in (1, \infty)$, we have:

$$D_\alpha(\mathcal{M}(d)\|\mathcal{M}(d')) \leq \xi + \rho\alpha,$$

where $D_\alpha(\mathcal{M}(d)\|\mathcal{M}(d'))$ denotes the α-th moment's Rényi divergence between the distribution $\mathcal{M}(d)$ and the distribution $\mathcal{M}(d')$.[2]

zCDP can be directly related to previous definitions of differential privacy through the Rényi divergence. In [14], the authors showed that if a mechanism $\mathcal{M}$ satisfies ϵ-DP, then $\mathcal{M}$ is $(\frac{1}{2}\epsilon^2)$-zCDP, and moreover proved that if a mechanism $\mathcal{M}$ ensures ρ-zCDP, then $\mathcal{M}$ satisfies $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP. Furthermore, they showed that the definition of CDP and zCDP can be interpreted mutually. In the case of the Gaussian mechanism, they proved that the mechanism (in Definition 3) satisfies $(\Delta f^2/2\sigma^2)$-zCDP.

Based on the Rényi divergence, the notion of Rényi differential privacy was introduced as a natural relaxation of differential privacy [15], where the definition of differential privacy is relaxed by bounding the Rényi divergence of the privacy loss random variable for any individual moment.

Definition 6 (Rényi Differential Privacy (RDP) [15]): A randomized mechanism $\mathcal{M}$ is said to have ϵ-Rényi differential privacy of order α (or (α, ϵ)-RDP for short), if for any adjacent databases d, d' it holds that:

$$D_\alpha(\mathcal{M}(d)\|\mathcal{M}(d')) \leq \epsilon.$$

Different from the other definitions, RDP bounds the Rényi divergence of privacy loss random variable only for a single moment at a time, as shown in the definition, which allows the analysis of a tighter bound on the cumulative privacy loss. In [15], the authors showed that if a mechanism $\mathcal{M}$ satisfies (α, ϵ)-RDP, it also satisfies $(\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP for any $0 < \delta < 1$. Furthermore, they showed that the Gaussian mechanism $\mathcal{M}_G$ satisfies $(\alpha, \frac{\alpha \Delta f^2}{2\sigma^2})$-RDP.

III. MEMBERSHIP INFERENCE ATTACK AGAINST DIFFERENTIALLY PRIVATE GENERATIVE MODEL

In this section, we describe each major part in detail. First, we analyze membership inference attacks and explain how to achieve differential privacy for the GAN model. Then we

[1] Note that (ϵ, δ)-differential privacy is the most well-known relaxed notion of pure-differential privacy, where it allows the failure of differential privacy with probability δ (typically, the value of δ is taken to be cryptographically small)

[2] When $\xi = 0$, the definition is characterized as ρ-zCDP.

present our evaluation framework for analyzing the relationship between the privacy invasion attacks and differentially private GAN models.

A. MEMBERSHIP INFERENCE ATTACK AGAINST GENERATIVE MODEL

As mentioned above, previous membership inference attacks (e.g., [6], [30]) targeted general machine learning models and aimed to infer whether specific data were included in the training dataset by exploiting the confidence vectors returned from the target model. Unlike general machine learning models consisting only of discriminative components, GAN consists of two distinct components (discriminator and generator) with opposite objectives, and it is regarded that the final result of training GAN is the generative part. Therefore, since the outputs of GAN for arbitrary inputs (latent codes) are synthetic data rather than confidence values, it makes no sense to apply previous membership inference attacks against GAN model.

By capturing these differences, Hayes *et al.* [8] showed that it is feasible to implement membership inference attacks against generative models. The intuition involved in this approach is that since the discriminator tends to output (even slightly) higher probability in training data than in others (synthetic data and testing data), the discriminator can be exploited as a distinguisher for membership inference attacks. Therefore, as long as an attacker builds a shadow discriminator that mimics the discriminator of the target model, he/she can execute membership inference attacks without building an additional attack model. Note that from the attacker's point of view, it is not allowed to access the discriminator of the target model,[3] and we assume that the white-box attack where an attacker is allowed to access the discriminator is the ideal scenario.

To build a shadow discriminator, an attacker first collects enough data (synthetic data) via queries to the target generative model, and then trains a new GAN model using the data obtained. At this point, the discriminator of the GAN model constructed by the attacker becomes an attack engine (shadow discriminator) that drives membership inference. After building a shadow discriminator, the attacker can conduct membership inference on data that the attacker holds and wants to know about membership (not the data obtained by querying the target model) by feeding the data to the shadow discriminator. The final inferences for specific data are made by arranging the data according to the results of the shadow discriminator. That is, the attacker can determine that data was included in the training dataset of the target generative model if the data is ranked at the upper position (relatively high probability) in the results of the shadow discriminator. In [8], the authors showed that the ideal white-box attacker who can access the discriminator of the target model can perfectly infer membership (100% attack accuracy). In the case of a black-box scenario, they showed that a general attacker can achieve up to 63% attack accuracy, and an informed attacker who has auxiliary knowledge of the training dataset can improve the attack success rate (these outcomes are the experimental results on a facial dataset, and we deal with the same dataset as our experimental data in section 5).

[3]In general, it is considered that GAN model is a generative model, and the discriminator that was operated when training GAN is not publicly exposed.

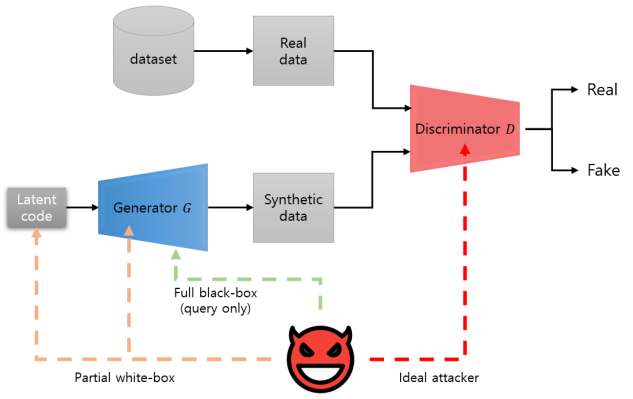


FIGURE 2. Attacker's abilities in our adversarial scenarios.

As an extended approach of the attack against the generative model, Hilprecht *et al.* [9] proposed a Monte Carlo integration-based membership inference attack in terms of distance metric between training and generated data. The intuition of this approach is that the overfitted generator tends to output synthetic data close to the training data that the target model has learned. In [9], the authors considered black-box scenarios and showed that an attacker can conduct membership inference on specific data in the Monte Carlo method by comparing the distance with synthetic data obtained from the target model.

From the perspective of distance metric, Chen *et al.* [10] proposed a more sophisticated method of attack by subdividing the attacker's knowledge about the generator of the target GAN model. They assumed black-box and white-box scenarios for the generator, and considered that the adversary can access the latent code that is the input of the model. With these assumptions, they realized a more sophisticated distance-based membership inference attack by reconstructing the synthetic data as close as possible to the target data that the attacker wants to know about membership. Note that in the case of the white-box scenario for the generator, an attacker can obtain data closer to the target data through an optimization process on the input space of the generator (e.g., gradient descent).

From the perspective of analyzing the relationship between privacy and utility for the GAN model, we deploy these membership inference attacks as privacy violation scenarios. In particular, we consider two specific attack scenarios (see Figure 2): ideal white-box (i.e., accessible to the discriminator) and realistic white-box (i.e., white-box for the generator with latent code) scenarios. For the ideal white-box scenario, the discriminator of the target model will be operated as the attack engine. In the case of the realistic white-box scenario,

TABLE 1. Comparison of differentially private GAN models in the gradient perturbation approach.

	Perturbation approach	Target model	Composition	Target data
Xie et al. [40]	Gradient perturbation (approximated sensitivity of gradient)	WGAN	Moments accountant	Image data Healthcare data
Sirivastava et al. [41]	Gradient perturbation (gradient clipping)	WGAN	Moments accountant	NIST synthetic data challenge dataset
Zhang et al. [43] Xu et al. [44]	Gradient perturbation (gradient clipping)	WGAN WGAN-GP	Moments accountant	Image data
Frigerio et al. [45]	Gradient perturbation (gradient clipping)	WGAN WGAN-GP (LSTM)	Moments accountant	Continuous and discrete data Time-series data
Torkzadehmahani et al. [46]	Gradient perturbation (gradient clipping)	CGAN	Rényi DP	Image data
Beaulieu-Jones et al. [47]	Gradient perturbation (gradient clipping)	AC-GAN	Moments accountant	Healthcare data

we assume that an adversary can access the parameters of the generator as well as latent code.

B. DIFFERENTIAL PRIVACY FOR GENERATIVE ADVERSARIAL NETWORKS

Since the notion of differential privacy emerged, extensive studies have been conducted in various fields requiring privacy preservation. Obviously, in the field of artificial intelligence, research on differentially private mechanisms has been actively conducted to preserve the privacy of AI models. Initially, the results were mainly focused on convex optimization problems and general machine learning models such as ERM [31], [32], decision tree [33], [34], regression [35], [36] etc., and progressed toward satisfying differential privacy for complex models with non-convex optimization problems such as deep learning and autoencoder (e.g., [37]–[39]).

There are three main approaches to achieve differential privacy for complex AI models: *output perturbation*, *objective perturbation*, and *gradient perturbation*. Among these approaches, most mechanisms that satisfy differential privacy for the GAN model have leveraged the *gradient perturbation* approach because of its flexibility and adaptability. Note that *output perturbation* can introduce a huge amount of noise in the parameters of the final model as differential privacy pursues the worst-case scenario, and *objective perturbation* is not generally applicable as it depends on the architecture of the model. The gradient perturbation method satisfies differential privacy by adding noise in the learning process of the model that performs gradient-based optimization. That is, at each iteration of the training, noise is added to the gradients calculated by referring to the dataset that contains sensitive information so that differential privacy can be held. The main issue with this approach is how to calculate tighter bounds on privacy loss in terms of *composition*. In this respect, Abadi *et al.*[37] recently proposed an efficient differentially private learning algorithm. In [37], the authors presented the moments accountant mechanism that can efficiently bound the cumulative privacy loss of the algorithm, and showed that differential privacy can be achieved with a modest privacy budget while preserving the utility of model. Most of the differentially private GAN algorithms leverage the moments accountant to calculate the cumulative privacy loss. Table 1 presents the characteristics for each differentially private GAN algorithm.

Xie *et al.* [40] proposed a differentially private GAN algorithm at first. They focused on the Wasserstein GAN model, and showed that differential privacy can be achieved via the same gradient-based training process as deep learning models. In principle, since GAN models consist of a discriminator and generator, unlike deep learning models that consist only of discriminative ones, it should be considered that differential privacy has to be applied to two distinct sub-models. However, Xie *et al.* [40] showed that if differential privacy is involved in the training of the discriminative model that directly references the training dataset, the generative model also satisfies differential privacy naturally due to the post-processing immunity of differential privacy. In the case of composition, they leveraged the moments

Algorithm 1 Differentially Private GAN Training

Input: data samples $\{x_1, \ldots, x_n\}$, loss function of discriminator $\mathcal{L}_D$, loss function of generator $\mathcal{L}_G$, batch size m, learning rates η_D and η_G, gradient norm bound c, noise scale σ.

Initialize w_0 and θ_0
for $t = 1, \ldots, T$ **do**
  Random sampling
    sample a batch $B_t = \{x^{(i)}\} \sim P_{\mathbf{x}}$ of data point with probability $q = \frac{m}{n}$
    sample a batch $\{z^{(i)}\}_{i=1,\ldots,m} \sim P_{\mathbf{z}}$ of latent code
  Compute gradient for discriminator
    $\mathbf{g}_{w_t,real}(x^{(i)}) \leftarrow \nabla_{w_t}\mathcal{L}_D(w_t, x^{(i)})$ for each $x^{(i)} \in B_t$
    $\mathbf{g}_{w_t,fake}(z^{(i)}) \leftarrow \nabla_{w_t}\mathcal{L}_D(w_t, G(z^{(i)}))$ for each $i \in [1, m]$
  Clip gradient
    $\bar{\mathbf{g}}_{w_t,real}(x^{(i)}) \leftarrow \mathbf{g}_{w_t,real}(x^{(i)}) / \max\left(1, \frac{||\mathbf{g}_{w_t,real}(x^{(i)})||_2}{c}\right)$ for each $x^{(i)} \in B_t$
    $\bar{\mathbf{g}}_{w_t,fake}(z^{(i)}) \leftarrow \mathbf{g}_{w_t,fake}(z^{(i)}) / \max\left(1, \frac{||\mathbf{g}_{w_t,fake}(z^{(i)})||_2}{c}\right)$ for each $i \in [1, m]$
  Add noise
    $\tilde{\mathbf{g}}_{w_t} \leftarrow \frac{1}{m}(\Sigma_i \bar{\mathbf{g}}_{w_t,real}(x^{(i)}) + \mathcal{N}(0, \sigma^2 c^2 \mathbf{I})) - \frac{1}{m}\Sigma_i \bar{\mathbf{g}}_{w_t,fake}(z^{(i)})$
  Optimize discriminator
    $w_{t+1} \leftarrow \mathrm{OPT}(w_t, \tilde{\mathbf{g}}_{w_t}, \eta_D)$
  Train generator G
    sample a batch $\{z^{(i)}\}_{i=1,\ldots,m} \sim P_{\mathbf{z}}$ of latent code
    $\mathbf{g}_{\theta_t}(z^{(i)}) \leftarrow \nabla_{\theta_t}\mathcal{L}_G(\theta_t, z^{(i)})$ for $i \in [1, m]$
    $\theta_{t+1} \leftarrow \mathrm{OPT}(\theta_t, \mathbf{g}_{\theta_t}, \eta_G)$
end for

Output: model parameters of discriminator w_T and generator θ_T, and compute the overall privacy cost (ϵ, δ)

accountant theorem. At around the same time, Srivastava and Alzantot [41] proposed a differentially private WGAN algorithm for the purpose of privacy-preserving synthetic data generation. The difference from the previous algorithm is that they simply bounded the sensitivity of gradients with the gradient clipping method. Although Xie *et al.* [40] approximated the upper bound of gradients, it can cause greater noise levels than the simple clipping approach as the approximation depends on the size of the model. Subsequently, several results have been reported that extend the previous algorithms to the improved WGAN model (WGAN-GP [42]) [43]–[45]. As another result on the differentially private GAN model, Torkzadehmahani *et al.* [46] proposed a differentially private conditional GAN algorithm that can generate both differentially private synthetic data and corresponding labels by utilizing the characteristics of the conditional GAN model. In particular, by applying Rényi differential privacy, they showed that the algorithm can improve the quality of synthetic data in the same privacy budget compared to algorithms involving the basic notion of differential privacy. For the purpose of privacy-preserving data sharing for clinical data, Beaulieu-Jones *et al.* [47] applied differential privacy to the auxiliary classifier GAN (AC-GAN [48]). As with the previous algorithms, they leveraged the differentially private SGD algorithm and calculated the overall privacy loss with the moments accountant.

By extending the capacity of these algorithms, we evaluate the impact of differential privacy regarding the privacy budget and its relaxed definitions over the substantial privacy violation scenario.[4] Algorithm 1 presents a systematic algorithm for achieving differential privacy in the training of GAN models. First, the algorithm samples a mini-batch with the sampling probability $q = \frac{m}{n}$ from the (training) dataset, and a mini-batch of size m from the latent space randomly. Then the gradients of the loss function $\mathcal{L}_D$ for the discriminator are computed with respect to the current parameters w_t of discriminator in both mini-batches, and the computed gradients are clipped by l_2-clipping with the clipping parameter c. At this point, Gaussian noise is added to the summed gradient to ensure differential privacy, and the magnitude of the noise is derived by considering only the gradient associated with the real data. After the gradient is averaged and aggregated, the parameter w_t of the discriminator is updated with the calculated gradient $\tilde{\mathbf{g}}_{w_t}$ and learning rate η_D in a gradient-based optimization method, such as SGD, Adam, or RMSProp. Note that the optimization process may include adjustments to model parameters, such

[4]We do not cover strategic approaches [49]–[51] (for example, clipping decay to reduce the magnitude of noise) because we focus on analyzing the relationship between differential privacy and the privacy invasion attack according to the privacy budget and the relaxed definitions of differential privacy

TABLE 2. Comparison of relaxed definitions of differential privacy. The DP interpretation of Concentrated DP is derived indirectly via zCDP [14].

	(μ, τ)-Concentrated DP	ρ-Zero Concentrated DP	(α, ϵ)-Rényi DP
Privacy cost in Gaussian mechanism	$\left(\frac{1}{2\sigma^2}, \frac{1}{\sigma}\right)$-CDP	$\frac{1}{2\sigma^2}$-zCDP	$\left(\alpha, \frac{\alpha}{2\sigma^2}\right)$-RDP
DP interpretation	$\left(\mu + \tau\sqrt{2\log(1/\delta)}, \delta\right)$-DP	$\left(\rho + 2\sqrt{\rho\log(1/\delta)}, \delta\right)$-DP	$\left(\epsilon + \frac{\log(1/\delta)}{\alpha-1}, \delta\right)$-DP
Gaussian noise scale	$\frac{\sqrt{2\log(1/\delta)}+\sqrt{2\log(1/\delta)+2\epsilon}}{2\epsilon}$	$\frac{\sqrt{2\log(1/\delta)}+\sqrt{2\log(1/\delta)+2\epsilon}}{2\epsilon}$	$\sqrt{\frac{\alpha}{2\left(\epsilon-\frac{\log(1/\delta)}{\alpha-1}\right)}}$

as weight clipping to ensure *Lipschitz continuity* in WGAN. This discriminator training procedure (lines 3 $\sim$ 15 in the algorithm) can be iterated in several steps internally, as in the case of WGAN. Obviously, in this case, privacy costs arise in every sub-iteration. After completing the training of the discriminator in an iteration step t, the algorithm trains the generator G, and this procedure is the same as regular generator training in the non-private scenario. Since the procedure for training G (lines 16 $\sim$ 19 in the algorithm) is a post-processing of a differentially private discriminator and does not access the training dataset, there is no need to force this procedure to ensure differential privacy. When the algorithm is finished, it outputs the final model parameters w_T and θ_T, and computes the overall privacy cost spent. We consider differential privacy and its relaxations, and Table 2 compares the noise scale in a single execution of the Gaussian mechanism according to the definitions of differential privacy. Although Algorithm 1 computes the overall privacy cost as an output process, it can be calculated during the running of the algorithm (right after optimizing the discriminator each time). In this case, a predefined entire privacy budget can be specified as termination criteria (i.e., as a threshold for spent privacy costs).

To evaluate differentially private GAN models, we first generate models via Algorithm 1, and then analyze their resistance over the membership inference attack scenarios according to the degree of the privacy guarantee and the definitions of differential privacy. Note that, the algorithm presents the process of training GAN models to satisfy differential privacy, and makes no assumptions about specific attack scenarios.

IV. EVALUATION

In this section, we conduct experiments to quantify how much privacy is leaked from differentially private GAN models. As mentioned in the previous section, we measure privacy leakage via membership inference attack in ideal and realistic adversarial scenarios.

A. EXPERIMENTAL SETUP

We first train target GAN models using Algorithm 1 with different relaxed notions of differential privacy, and compare them in terms of privacy leakage. The notions that we consider are (ϵ, δ)-DP, zero-concentrated DP (zCDP), and Rényi DP (RDP). Since concentrated DP (CDP) has the same composition property and noise scale as zCDP, as shown in Table 2, we do not include CDP in the experiments. For zCDP, we convert privacy budgets to (ϵ, δ)-DP, and use them as termination thresholds. In the case of RDP, we leveraged the RDP accountant [52], [53].

As described above, relaxing the definition of differential privacy results in a smaller noise scale for a given privacy budget. Alternatively, in terms of composition, the relaxed notions enable more differentially private operations for a given privacy budget and fixed noise scale. In this respect, we considered the latter case and set the noise scale $\sigma = 2$. With respect to the *sensitivity* in terms of differential privacy, we set the gradient clipping parameter $c = 2$. Note that, since the parameters c and σ are directly involved in the standard deviation of the Gaussian distribution in the Gaussian mechanism, large c and σ can cause large noise, and in this case, the training of GAN models may not proceed at all even with a large number of iterations.

1) TARGET MODEL

We experiment and evaluate three GAN models: 1) deep convolutional GAN (DCGAN), a model that combines the basic GAN architecture with a convolutional neural network, 2) Wasserstein GAN (WGAN), a model that improves training stability by using the Wasserstein distance as an approximation metric between probability distributions, 3) boundary equilibrium GAN (BEGAN), a model that can approximate the convergence of the training process by combining it with the concept of Autoencoder.

For DCGAN and WGAN, we built both models with the same architecture. In particular, we constructed the discriminator as three convolutional layers and a fully connected layer sequentially. Note that, since the differentially private learning algorithm computes the gradient for each single data point, we did not include the batch normalization process due to compatibility concerns. For the generator, we consisted of a fully connected layer and three de-convolutional layers (upsampling-convolutional layer) sequentially. In the case of BEGAN, we constructed the encoder as three convolutional

layers and a fully connected layer, and the decoder as a fully connected layer and three de-convolutional layers (the discriminator is constructed by encoder-decoder and the architecture of the generator is the same as that of the decoder). Unlike the other models, WGAN internally iterates discriminator training before proceeding with generator training, and we set the number of internal iterations to 5.

2) DATASET

We use two datasets to evaluate differentially private GAN models: the MNIST handwritten dataset containing 60k training samples and 10k testing samples of size 28×28 in grayscale, and the Labeled Faces in the Wild (LFW) dataset [54] containing 13,233 images of faces. In the case of LFW datset, we aligned each data to a size of 62×47 and converted it to grayscale. For both datasets, we randomly sample 10% of the data points as the training dataset. To measure the privacy leakage, we also prepare a test dataset (for attack scenario) with the same manner and size as the training dataset. From the perspective of the attack scenario, the data points in the training dataset are *members*, and the others are *non-members*. Therefore, the attack success rate of the baseline attacker can be 50%.

B. MODEL ACCURACY

Before evaluating the resistance of differentially private GAN models to the membership inference attack, we investigate the accuracy of the models according to the notions of differential privacy as well as privacy budgets in terms of the quality of generated synthetic data. Figure 3 presents generated samples from trained differentially private GAN models. Obviously, it can be seen that the quality of the generated data improves as the privacy budget increases. Likewise, in the same privacy budget, the quality of the generated data is improved as the notion of differential privacy is relaxed. To show the results from a broader perspective, we further present the outputs of differentially private GAN models trained on the CelebA dataset [55] (under the same conditions as the experiments in the MNIST and LFW datasets), which is an RGB three-color (celebrity) face dataset. Compared to other models, (differentially private) BEGAN seems to outperform even with relatively small privacy budgets.

In addition to these visual comparisons, we conduct a classification task on the generated data to evaluate the models' accuracy numerically. Although there are several metrics that can evaluate the quality of synthetic data generated from GAN models, such as inception score and SSIM (structural similarity), we adopted the classification accuracy to intuitively represent the quality of synthetic data. Note that, this approach has been applied in previous studies on differentially private GANs (e.g., [46], [49]). The process of this experiment is as follows: we first build a classification model that acts as an evaluator for model accuracy using the original training dataset, and generate synthetic datasets from the trained differentially private GAN models with the same proportion and size as the original test dataset. Then we present the classification accuracy of the evaluator on these synthetic datasets as the accuracy of differentially private GAN models. In order to label the generated data, we trained GAN models by including the class attribute as with the conditional GAN architecture [22]. In the case of the classifier, we build a general two-layer neural network model. Figure 4 shows the experimental results on the MNIST dataset. As shown in the results, we found that not only was the visual quality improved, but also the classification accuracy of the generated data, and it can be interpreted that the differentially private GAN models generate clearer data as the privacy budget is increased and the definition of differential privacy is relaxed.[5] In the case of the non-private scenario, the accuracy was measured as 0.931 and 0.952 for DCGAN and BEGAN, respectively. In the case of the differentially private scenario, we confirmed that the RDP models converge very closely to the non-private scenario at $\epsilon \geq 10$ compared to the other notions. When $\epsilon = 10$, the accuracy loss in the experiments on DCGAN was measured as 12%, 8%, and 5% in DP, zCDP, and RDP, respectively. Similarly, in the experiments on BEGAN, the accuracy loss was measured as 10%, 7%, and 4%.

C. IDEAL WHITE-BOX ATTACK SCENARIO

As described above, we assume that the attacker in the ideal white-box scenario has access to the discriminator of the trained GAN model, and exploits the discriminator as the attack engine (i.e., the attacker model) for membership inference. Figure 5 and 6 show the privacy leakage due to the membership inference attack on GAN models in the ideal white-box attack. In the case of the inference, we sorted the outputs of the discriminator for the suspect data and summarized the top ranked data, and we set 1/2 as the minimum value of the attack accuracy since the attack success rate of the baseline attacker is equal to the probability of flipping a coin. This means that there is no privacy leakage when the attack accuracy is 1/2.

Figure 5 shows the experimental results on the MNIST dataset. In the non-private scenario, the attacker achieved 93%, 93%, and 57% in DCGAN, WGAN, and BEGAN, respectively. As shown in the figure, we found that differential privacy can significantly reduce privacy leakages even with relatively large privacy budgets. In the case of DCGAN and WGAN, it was measured that the higher the privacy budget and the more relaxed the definition of differential privacy, the more vulnerable to attack (see Figure 5 (a) and (b)). As expected, (ϵ, δ)-DP showed the strongest resistance compared to the other notions, and attack accuracy was measured close to the baseline attacker. In the case of zCDP and RDP, the attack resistance was measured to be strong at $\epsilon < 10$, but the attack success rate increased at $\epsilon \geq 10$. In the case of BEGAN (Figure 5 (c)), different results from the experiments with the other models were measured.

[5] We do not present the results on WGAN because they are very similar to those in the results on DCGAN.

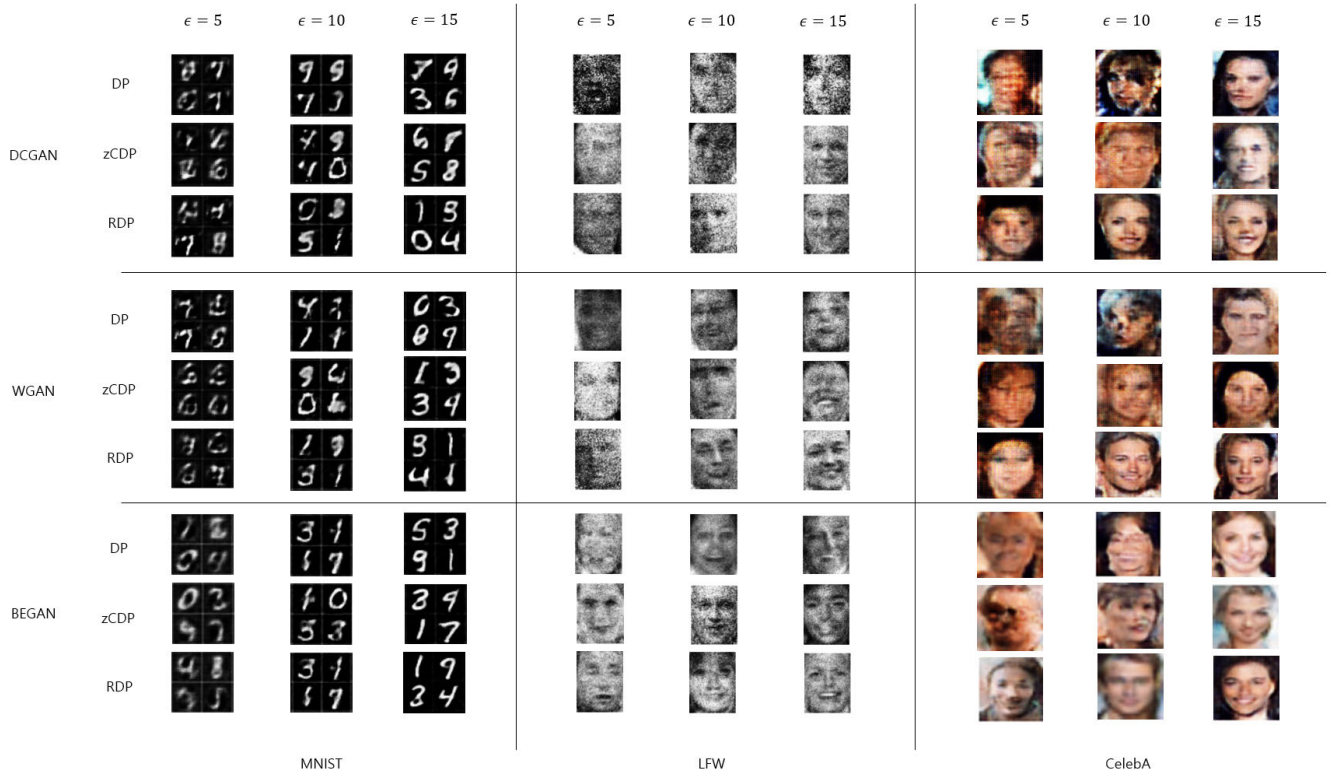
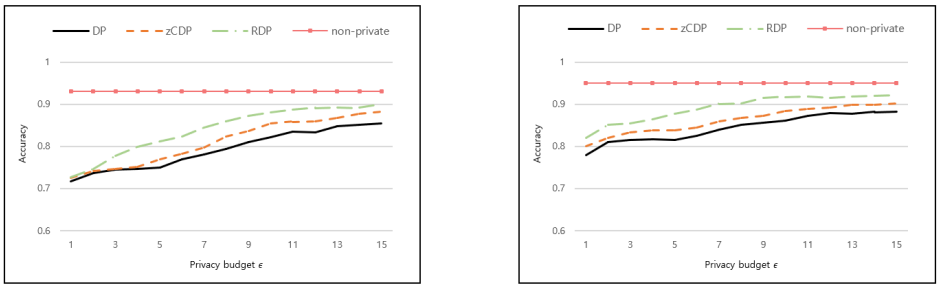


FIGURE 3. Generated samples from trained differentially private GAN models.



(a) Experimental results on DCGAN

(b) Experimental results on BEGAN

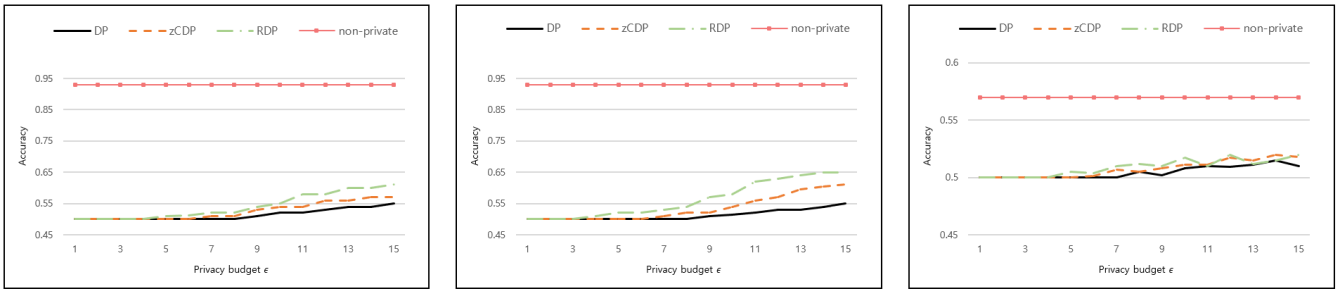
FIGURE 4. Classification accuracy on synthetic datasets generated from differentially private GAN models.

As the privacy budget grows, the attack probability seems to increase slightly. However, the attack success rate was measured with a very low probability even with large privacy budgets (the maximum attack success rate was measured to be 0.53 when $10 < \epsilon < 15$), and it was observed to have a strong resistance to the membership inference attack compared to other models.

Figure 6 shows the experimental results on the LFW dataset. In the non-private scenario, the attacker achieved 99%, 99%, and 62% in DCGAN, WGAN, and BEGAN, respectively. Compared with the experiments on the MNIST dataset, the attacker in the non-private scenario achieved higher attack success rates. Overall, the results showed a pattern similar to that of the MNIST experiments. In the case of DCGAN and WGAN, it was measured that the attacker's advantage slightly increased compared to the experiment with the MNIST dataset. However, in the case of BEGAN, the attack success rate was measured with a very low probability as in the previous experiments (the maximum attack success rate was measured to be 0.53).

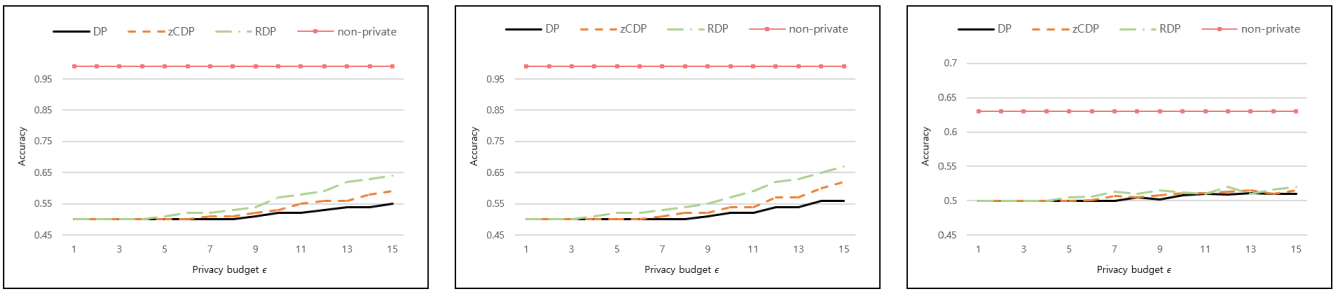
D. REALISTIC WHITE-BOX ATTACK SCENARIO

As described in the previous section, we assume that the attacker in the realistic white-box scenario has access to the generative model of the trained GAN model and latent code, and exploits them as the reconstruction engine of data for membership inference. Therefore, the attacker model in the realistic white-box attack scenario can be regarded as



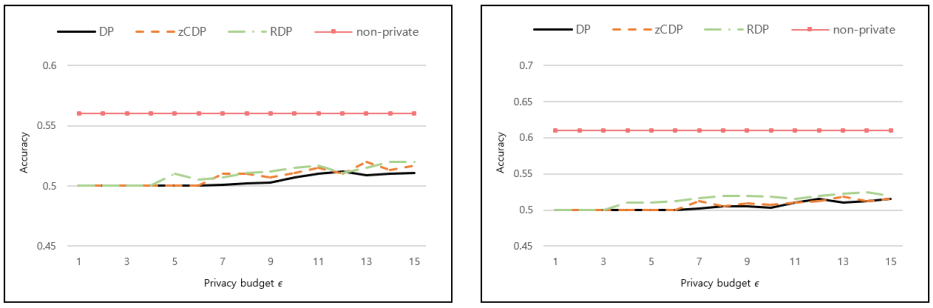
(a) Experimental results on DCGAN (b) Experimental results on WGAN (c) Experimental results on BEGAN

FIGURE 5. Ideal white-box attack performance on MNIST dataset.



(a) Experimental results on DCGAN (b) Experimental results on WGAN (c) Experimental results on BEGAN

FIGURE 6. Ideal white-box attack performance on LFW dataset.

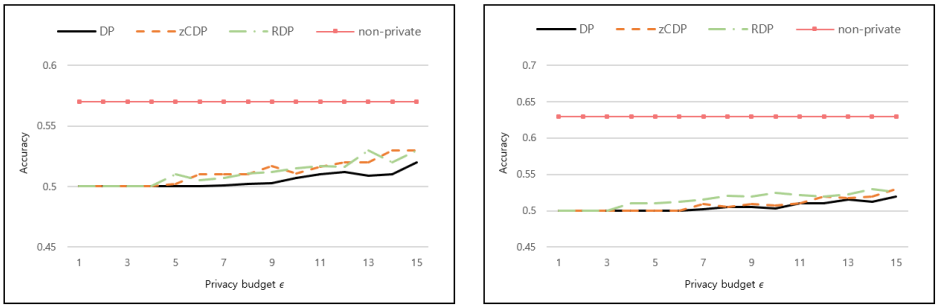


(a) Experimental results on DCGAN (b) Experimental results on WGAN

FIGURE 7. Realistic white-box attack performance on MNIST dataset.

a reconstruction process through optimization (optimization process on latent codes to generate data as close as possible to the suspicious data in terms of the distance). Figure 7 and 8 show the privacy leakage due to the membership inference attack on GAN models in the realistic white-box attack. As in the previous experiment, we sorted and summarized the distance results between synthetic and suspect data, and we set 1/2 as the minimum value of the attack accuracy and assume that there is no privacy leakage when the attack accuracy is 1/2. In addition, we excluded the experiment with BEGAN from the realistic white-box attack experiment since it showed strong resistance to the attack even in the ideal white-box scenario.

Figure 7 shows the experimental results on the MNIST dataset. In the non-private scenario, the attacker achieved 56% and 61% in DCGAN and WGAN, respectively. As shown in the figure, we found that differential privacy can reduce privacy leakages, and attack accuracy was measured very close to the baseline attacker. In both models, differential privacy showed strong resistance to the attack even with large privacy budgets and relaxed definitions, unlike the experimental results in the ideal white-box scenario. Overall, the attack success rate was measured to be less than 0.53 in both experiments. Figure 8 shows the experimental results on the LFW dataset. In the non-private scenario, the attacker achieved 57% and 63% in DCGAN and WGAN, respectively.



(a) Experimental results on DCGAN (b) Experimental results on WGAN

FIGURE 8. Realistic white-box attack performance on LFW dataset.

As expected, the measured experimental results were very similar to the experiment with the MNIST dataset.

E. DISCUSSION

From the perspective of model utility, the tighter bounds on the cumulative privacy loss by the relaxed definition of differential privacy improve the quality of synthetic data for a given privacy budget. However, in the ideal white-box attack scenario, we found that differentially private models with relaxed definitions are more vulnerable to the membership inference attack because they reduce the noise magnitude or allow more training iterations on a given dataset. Therefore, we can conclude that relaxing the definition of differential privacy comes with additional privacy risks. Nevertheless, we confirmed that differential privacy can significantly mitigate the privacy leakage compared to the non-private scenario even with relatively large privacy budgets. In particular, this advantage is more evident in the realistic white-box attack scenario. Furthermore, by experimenting with various GAN models, we found that privacy leakage is dependent on the model architecture, and applying differential privacy can amplify resistance to the membership inference attack.

In our experiment, we trained and built GAN models on subsets, which are datasets sampled with a probability of 10% from the original training datasets. In other words, GAN models can easily overfit as they are trained on very small datasets, and this can make the models very vulnerable to membership inference attacks. In [8]–[10], it has been reported that the sampling probability significantly affects the accuracy of membership inference attack, and the smaller the sampling set (i.e., the more severe the overfitting), the more vulnerable to attack. Considering these points, since differential privacy should consider the worst-case scenario, we focused on the sampling probability of 10%, which was the most reasonably vulnerable case in non-private scenarios.[6]

V. RELATED WORK

Recently, various studies have been conducted to analyzed the relationship between differential privacy and privacy invasion attacks on machine learning and deep learning models. Rahman *et al.* [16] investigated the relationship between differential privacy and the membership inference attack, focusing on neural network-based models. In particular, they analyzed the trade-off between utility and privacy by varying the privacy budget. Focusing on the model inversion attack for regression models, Wang *et al.* [17] proposed a differentially private regression model. In [17], the authors leveraged the functional mechanism to ensure differential privacy, and showed that the proposed differentially private regression model can provide resistance to the model inversion attack while preserving utility. Zhang *et al.* [18] considered an obfuscation method that injects noise into the input dataset before training the machine learning model, and showed that the data reconstructed by the model inversion attack (from the model with the obfuscation applied) is more blurred compared to the non-private scenario. Park *et al.* [20] studied the relationship between differential privacy and the model inversion attack. In particular, they focused on face recognition systems based on neural network-based models, and analyzed the trade-off between utility and privacy according to the degree of privacy guarantee in the model inversion attack scenario. Jayaraman and Evans [19] investigated the relationship beetween definitions of differential privacy and privacy invasion attacks. By focusing on the membership inference and attribute inference attack [6], [30], they analyzed the resistance of differential privacy to the attacks for logistic regression and neural network models. In contrast to previous studies that targeted neural network and regression models, we focus on generative adversarial networks, which are the most sophisticated generative models, and analyze the relationship between differential privacy and membership inference attack on GAN models.

VI. CONCLUSION

In this paper, we investigated the resistance of differentially private GAN models to the membership inference attack according to the degree of privacy guarantee. In the experimental evaluation, by quantifying the effectiveness of the attack based on the degree of privacy guarantee, we showed that differential privacy can reduce the attack success rates of membership inference while preserving the quality of

[6]In experiments with relatively large sampling sizes (roughly 30% or more), we found that differentially private GAN models have strong resistance to membership inference attacks even at relatively large epsilon values.

synthetic data. However, by investigating several notions of differential privacy, we found that relaxing the definition of differential privacy comes with additional privacy risks. Nevertheless, we confirmed that differential privacy can significantly mitigate privacy leakage compared to the non-private scenario. As a future study, it would be interesting to investigate the privacy leakage on the differentially private algorithms with strategic approaches (e.g., clipping decay).

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[2] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2773–2781.
[3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[4] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. 23rd USENIX Secur. Symp.*, 2014, pp. 17–32.
[5] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.
[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, May 2017, pp. 3–18.
[7] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 601–618.
[8] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Membership inference attacks against generative models," in *Proc. Priv. Enhancing Technol.*, 2019, pp. 133–152.
[9] B. Hilprecht, M. Härterich, and D. Bernau, "Monte Carlo and reconstruction membership inference attacks against generative models," in *Proc. Priv. Enhancing Technol.*, 2019, pp. 232–249.
[10] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 343–362.
[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.*, 2006, pp. 265–284.
[12] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2013.
[13] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," 2016, *arXiv:1603.01887*.
[14] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Proc. Theory Cryptogr. Conf.*, 2016, pp. 635–658.
[15] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 30th Comput. Secur. Found. Symp. (CSF)*, Aug. 2017, pp. 263–275.
[16] M. A. Rahman, T. Rahman, R. Laganière, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model," *Trans. Data Privacy*, vol. 11, no. 1, pp. 61–79, Apr. 2018.
[17] Y. Wang, C. Si, and X. Wu, "Regression model fitting under differential privacy and model inversion attack," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, Jun. 2015, pp. 1003–1009.
[18] T. Zhang, Z. He, and R. B. Lee, "Privacy-preserving machine learning through data obfuscation," 2018, *arXiv:1807.01860*.
[19] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *Proc. 28th USENIX Secur. Symp.*, 2019, pp. 1895–1912.
[20] C. Park, D. Hong, and C. Seo, "An attack-based evaluation method for differentially private learning against model inversion attack," *IEEE Access*, vol. 7, pp. 124988–124999, 2019.
[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–4.
[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[23] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.
[24] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2180–2188.
[25] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[26] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," 2017, *arXiv:1703.10717*.
[27] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–26.
[28] G. Ateniese, G. Felici, L. V. Mancini, A. Spognardi, A. Villani, and D. Vitali, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," 2013, *arXiv:1306.4447*.
[29] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, "'You might also like': Privacy risks of collaborative filtering," in *Proc. IEEE Symp. Secur. Privacy*, May 2011, pp. 231–246.
[30] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE 31st Comput. Secur. Found. Symp. (CSF)*, Jul. 2018, pp. 268–282.
[31] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Mach. Learn. Res.*, vol. 12, no. 3, pp. 1069–1109, 2011.
[32] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proc. IEEE 55th Symp. Found. Comput. Sci.*, Oct. 2014, pp. 464–473.
[33] A. Friedman and A. Schuster, "Data mining with differential privacy," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 493–502.
[34] B. Xin, W. Yang, S. Wang, and L. Huang, "Differentially private greedy decision forest," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 1–10.
[35] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 289–296.
[36] J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, "Functional mechanism: Regression analysis under differential privacy," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1364–1375, 2012.
[37] M. Abadi, A. Chu, I. Goodfellow, and H. B. McMahan, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 308–318.
[38] N. Phan, Y. Wang, X. Wu, and D. Dou, "Differential privacy preservation for deep auto-encoders: An application of human behavior prediction," in *Proc. 30th AAAI Conf. Artif. Intell.*, Feb. 2016, pp. 1309–1316.
[39] N. H. Phan, X. Wu, and D. Dou, "Preserving differential privacy in convolutional deep belief networks," *Mach. Learn.*, vol. 106, nos. 9–10, pp. 1681–1704, 2017.
[40] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018, *arXiv:1802.06739*.
[41] M. Alzantot and M. Srivastava. (2019). *Differential Privacy Synthetic Data Generation Using WGANs*. [Online]. Available: https://github.com/nesl/nist_differential_privacy_synthetic_data_challenge
[42] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.
[43] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model (technical report)," 2018, *arXiv:1801.01594*.
[44] C. Xu, J. Ren, D. Zhang, Y. Zhang, and Z. Qin, "GANobfuscator: Mitigating information leakage under GAN via differential privacy," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2358–2371, Sep. 2019.
[45] L. Frigerio, A. S. de Oliveira, L. Gomez, and P. Duverger, "Differentially private generative adversarial networks for time series, continuous, and discrete open data," in *Proc. Int. Conf. Syst. Secur. Privacy Protection*, 2019, pp. 151–164.
[46] R. Torkzadehmahani, P. Kairouz, and B. Paten, "DP-CGAN: Differentially private synthetic data and label generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1–7.

[47] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulat., Cardiovascular Quality Outcomes*, vol. 12, no. 7, Jul. 2019, Art. no. e005122.
[48] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2642–2651.
[49] J. Jordon, J. Yoon, and M. Van Der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–21.
[50] D. Chen, T. Orekondy, and M. Fritz, "GS-WGAN: A gradient-sanitized approach for learning differentially private generators," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–18.
[51] C. Ma, J. Li, M. Ding, B. Liu, K. Wei, J. Weng, and H. Vincent Poor, "RDP-GAN: A Rényi-differential privacy based generative adversarial network," 2020, *arXiv:2007.02056*.
[52] I. Mironov, K. Talwar, and L. Zhang, "Rényi differential privacy of the sampled Gaussian mechanism," 2019, *arXiv:1908.10530*.
[53] G. Andrew, S. Chien, and N. Papernot. *TensorFlow Privacy*. Accessed: Jun. 17, 2021. [Online]. Available: https://github.com/tensorflow/privacy
[54] L. J. Karam and T. Zhu, "Quality labeled faces in the wild (QLFW): A database for studying face recognition in real-world environments," *Proc. SPIE Hum. Vis. Electron. Imag.*, vol. 9394, Mar. 2015, Art. no. 93940B.
[55] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

CHEOLHEE PARK received the B.S. degree from the Department of Applied Mathematics, Kongju National University, in 2014, and the M.S. and Ph.D. degrees from the Department of Mathematics, Kongju National University, in 2017 and 2021, respectively. He joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea, in 2021, where he is currently working as a Postdoctoral Researcher. His research interests include data privacy, differential privacy, machine learning, deep learning, AI security, and network security.

YOUNGSOO KIM received the B.S. degree from the Department of Information Engineering, Sungkyunkwan University, Republic of Korea, in 1998, and the M.S. and Ph.D. degrees from the Department of Computer Engineering, Sungkyunkwan University, in 2000 and 2009, respectively. He joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea, in 2000, where he is currently working as a Principal Researcher. From 2012 to 2015, he was an Adjunct Professor at Chungnam National University, Daejeon. His research interests include 5G security, network security, digital forensics, cryptography, and AI security.

JONG-GEUN PARK received the B.S. and M.S. degrees from the Department of Industrial Engineering, Sungkyunkwan University, Republic of Korea, in 1997 and 1999, respectively, and the Ph.D. degree from the Department of Computer Engineering, Chungnam National University, Republic of Korea, in 2013. From 1999 to 2001, he was a Researcher at ADD, Daejeon, Republic of Korea. Then, he joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, in 2001, where he is currently working as a Principal Researcher. His research interests include mobile network security, SDN/NFV, cloud security, and AI security.

DOWON HONG received the B.S., M.S., and Ph.D. degrees in mathematics from Korea University, Seoul, South Korea, in 1994, 1996, and 2000, respectively. He has been a Principal Member of Engineering Staff of the Electronics and Telecommunications Research Institute (ETRI), South Korea, from 2000 to 2012. He joined the Department of Applied Mathematics, Kongju National University, South Korea, in 2012, and has been a Full Professor, since 2015. His research interests include cryptography, data privacy, and differential privacy.

CHANGHO SEO received the B.S., M.S., and Ph.D. degrees in mathematics from Korea University, Seoul, South Korea, in 1990, 1992, and 1996, respectively. He is currently a Full Professor with the Department of Applied Mathematics, Kongju National University, South Korea. His research interests include cryptography, information security, data privacy, and system security.