# Generative Model with User Similarity in Black-Box Model Inversion

**SCIS 2022** 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

# ブラックボックス型モデル反転攻撃における ユーザ類似性を考慮した生成モデルの検討

# Generative Model with User Similarity in Black-Box Model Inversion

井田 天星 * 竹内 廉 * ヴォ ゴック コイ グエン * 西垣 正勝 *
Tensei Ida Ren Takeuchi Vo Ngoc Khoi Nguyen Masakatsu Nishigaki
大木 哲史 *
Tetsushi Ohki

**あらまし** 個人の識別タスクに利用される機械学習モデルはユーザの顔画像や音声といったセンシティブなデータを用いて訓練される．このことから，訓練済みモデルはパラメータという形でユーザのプライバシーに関わる機密情報を間接的に所有し，プライバシー面での懸念が存在する．このような訓練済みモデルに対する攻撃として，訓練に用いられたユーザの機密情報を推測するモデル反転（Model Inversion, MI）攻撃が存在する．本研究では，攻撃対象モデルがブラックボックス型の深層ニューラルネットワーク（DNN）で構築された顔画像識別器であるという前提の上で，公開データを用いて推定データの探索空間を構築する MI 攻撃の事前学習プロセスに着目する．探索空間の構築方法が攻撃精度に与える影響を調べるために，Zhang らが提案した公開データと敵対的生成ネットワーク（GAN）を用いる攻撃手法における事前学習プロセスに攻撃対象ユーザとの類似性の概念を導入することで，攻撃対象ユーザに特化した探索空間を構築可能な手法を提案する．実験では復元した顔画像によるなりすまし精度を示すとともに，提案手法の有効性について考察する．

**キーワード** AI セキュリティ，モデル反転攻撃

## 1 はじめに

機械学習 (ML) システムの普及に伴い，顔識別や音声認識など個人の識別タスクにも ML アルゴリズムが利用されるようになった．個人の識別タスクに利用される ML モデルはユーザの顔画像や音声といったセンシティブなデータを用いて訓練される [1,2]．このことから，訓練済みモデルはパラメータとしてユーザのプライバシーに関わる機密情報を間接的に所有し，プライバシー面での懸念が存在する．このような ML モデルに対する攻撃がどのような制約で実行可能であり，どの程度有効なのか知ることは，堅牢な ML システムの設計や開発，インシデント発生時の影響調査などにおいて重要である．

訓練済みモデルに対する既知の攻撃として，訓練データに用いられたユーザの機密情報を推測するモデル反転（Model Inversion，MI）攻撃が存在する [3]．MI 攻撃が成功した場合，ユーザのプライバシーが侵害されてしまう．さらに，顔画像や音声が復元された場合には，個人認証に悪用されてしまう可能性もある．MI 攻撃は，攻撃対象モデルが深層ニューラルネットワーク（DNN）のように複雑である場合や，モデル及び攻撃対象ユーザの有する知識に関して制約がある場合，訓練データの推測が難しくなることが経験的に知られている．MI 攻撃の制約に関しては，ホワイトボックス型の攻撃とブラックボックス型の攻撃が検討されている．ホワイトボックス型の場合，攻撃対象モデル内部の詳細な情報を利用して高いなりすまし精度の攻撃が可能となる [4]．ブラックボックス型の場合，攻撃対象モデルの入出力のみ利用可能で高いなりすまし精度の攻撃は困難となる [5]．高いなりすまし精度での攻撃を可能とするブラックボックス型の MI 攻撃が存在する場合，大きな脅威となる．DNN に対して高いなりすまし精度を示す既存手法として，インターネットから誰でも入手可能な顔画像の公開データを事前情報として利用し，顔画像生成モデルとして敵対的生成ネットワーク（GAN）[6] を用いる Zhang らの手法が存在する [4]．この手法は，公開データのみを用いて DNN に対する MI 攻撃が実行できるため，現実的な脅威となるが，攻撃対象モデルの特徴抽出層にアクセスが必

* 静岡大学, 静岡県浜松市中区城北 3 丁目 5-1, Shizuoka University, 3-5-1 Jo-hoku, Naka-ku, Hamamatsu City, Shizuoka, Japan

要なホワイトボックス型の攻撃である．また，GAN を用いる手法では，事前学習によって攻撃対象データが取り得る特徴空間（以下，探索空間）を学習し，これを効率的に探索することで攻撃を実現する．ここで，探索時には攻撃対象ユーザとの類似性を考慮しているが，GAN の事前学習による探索空間の構築時には類似性を考慮していない．MI 攻撃は，攻撃対象ユーザとの類似度を重要視するため，類似性が考慮されていない事前学習プロセスについて検討の余地がある．事前学習プロセスにおいて類似性を考慮することで，攻撃対象ユーザに特化した探索空間を構築することが可能となり，効率的に攻撃を実行できる可能性がある．実際の ML システムは，特定クライアントからの時間あたりのクエリ試行回数が制限されている場合があるため，ブラックボックス型の攻撃を行う際には，少ないクエリ数で高いなりすまし精度を実現可能であることが望ましい．これを実現するためのアイデアとして，なりすまし精度の高いベクトルが多い探索空間を構築した上で推定データの探索を行うことが考えられる．

そこで本研究では，攻撃対象モデルが DNN で構築された顔画像識別器であるという前提で，人間の顔画像の特徴を保ったデータを生成するために公開データを用いて推定データの探索空間を構築する，ブラックボックス型 MI 攻撃の事前学習プロセスに着目する．探索空間の構築方法がなりすまし精度に与える影響を調べるために，Zhang らの攻撃手法における事前学習プロセスに攻撃対象ユーザとの類似性の概念を導入することで，攻撃対象ユーザに特化した探索空間を構築可能な手法を提案する．類似性の概念を導入して構築される探索空間は，類似性を考慮せずに構築した探索空間に比べ，攻撃対象ユーザとの類似度が高いサンプルが多くなる．これにより，広い高次元の探索空間からランダムにサンプルを取り出した際の，類似度の期待値が既存手法より高くなることが期待される．実験では復元した顔画像によるなりすまし精度を示すとともに，提案手法の有効性について考察する．また，なりすまし精度をもとに探索空間を可視化し，既存手法と提案手法の探索空間の差を視覚的に確認する．

本研究の貢献は，以下の通りである．

- MI 攻撃の生成モデルにおける探索空間の構築時に攻撃対象ユーザとの類似度の概念を導入する．
- 探索空間の構築方法がなりすまし精度に与える影響を調査し，提案手法によって構築した探索空間が，既存手法よりも高い頻度でなりすまし精度の高い潜在ベクトルを生成できることを示した．

## 2 関連研究

### 2.1 モデル反転攻撃

MI 攻撃は，Fredrikson らによって提案された手法であり，攻撃対象モデルが出力する確信度が攻撃対象ユーザに対して高くなるような入力を探索することで達成される [3]．MI 攻撃を行う際には，探索対象の推定データを生成するために生成モデルを利用する場合 [4,5] と生成モデルを利用しない場合が存在する．生成モデルは，ランダムなノイズを推定データに変換する役割を持ち，生成モデルを予め訓練することで，推定データの探索空間を構築した上での探索が可能となる．Fredrikson らの手法では，入力の探索に生成モデルは利用せず，最急降下法 (GD) のみを用いる [3]．この手法には 2 つの問題点がある．1 つ目は，ニューラルネットワーク (NN) による画像認識のような非凸型の問題に対して GD を利用しているため，局所解に陥りやすい点である．2 つ目は，入力が顔画像であるという性質を考慮していないため，顔とはかけ離れたデータが推測される点である．これらの問題から，層が浅い NN に対してのみ有効であり，DNN に対しての攻撃は困難であった．

### 2.2 ホワイトボックス型の MI 攻撃

生成モデルを利用するホワイトボックス型の MI 攻撃として，Zhang らの手法が存在する [4]．Zhang らの手法は，生成モデルとして GAN を利用し，公開データから人間の顔の分布を事前学習した後，その分布を用いて攻撃を行う．事前学習を行うことで生成される画像が人間の顔の特徴を持つようになり，DNN で構築された顔識別システムに対しても高精度な MI 攻撃が可能となった．本手法は GAN の探索時には攻撃対象ユーザの情報を用いるが，GAN の学習は公開データのみを用いて行うことを前提としている．GAN の学習時においても攻撃対象ユーザの情報を用いることでよりなりすましに適した顔画像生成を行うことができる可能性がある．

### 2.3 ブラックボックス型の MI 攻撃

生成モデルを利用するブラックボックス型の MI 攻撃として，Aïvodji らの手法が存在する [5]．Aïvodji らの手法は，攻撃対象モデルを模倣するための代理モデルと攻撃対象モデルへの入力を生成する生成モデルを同時に訓練した後，訓練済みの生成モデルを用いて攻撃を行う．Aïvodji らの手法における制約は，生成モデルの出力の次元と値の範囲が攻撃対象モデルと代理モデルの入力に一致していることの 1 点のみである．Aïvodji らの手法は，厳しい制約下での攻撃が可能であり，ホワイトボックス型の他の手法と比べて攻撃精度は低い．また，生成モデルの訓練において，攻撃対象モデルの入力データの

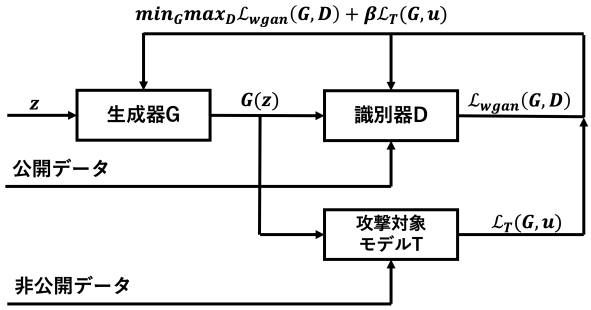



図 1: 提案手法における事前学習の概要図

特徴を考慮していないため，Fredrikson らの手法と同様に本来の分布とはかけ離れたデータが生成される可能性がある．

## 3 提案手法

### 3.1 概要

本研究では，Zhang らの手法 [4] における GAN の事前学習アルゴリズムに対し，生成器の出力画像と攻撃対象ユーザ間の類似度を導入する．これにより，攻撃対象ユーザとの類似度が高いサンプルがより多く存在する探索空間を構築することを狙いとする．図 1 に事前学習の概要を示す．GAN は，生成器 G と識別器 D から構成される．Zhang らの手法と異なる点は，GAN の損失関数からホワイトボックス要因となっていた項 $\mathcal{L}_{div}(G)$ を取り除き，生成器 $G$ の出力画像と攻撃対象ユーザ間の類似度を考慮した項 $\mathcal{L}_T(G,u)$ を新たに追加する点である．攻撃対象モデル $T$ は，DNN で実装され，攻撃者に対して非公開のデータを用いて訓練された顔画像識別器である．$\mathcal{L}_T(G,u)$ は，$T$ に対して $G$ の出力画像を入力した場合の攻撃対象ユーザ $u$ に対する損失を計算する損失関数であり，生成器 $G$ に潜在変数 $z$ を与えた際の出力画像 $G(z)$ と $T$ の最終出力となる確率ベクトル $T(G(z))$ を利用して計算する．

### 3.2 攻撃手順

攻撃手順は，大きく 2 つのステップに分けられる．ステップ 1 は GAN の訓練，ステップ 2 は攻撃対象ユーザの画像推定である．

ステップ 1 では，公開顔画像データを用いて GAN を訓練する．GAN の損失関数には，WGAN [7] で提案された Wasserstein 距離を用いた損失関数 $\mathcal{L}_{wgan}(G,D)$ を用いる．GAN の訓練における目的関数 $\mathcal{L}(G,D)$ を，式 (1) に示す．$\mathcal{L}_{wgan}(G,D)$ は，GAN が公開データの分布を学習するための項であり，$\mathcal{L}_T(G,u)$ は，攻撃対象ユーザ $u$ に類似した分布を学習するための項である．また，$\beta$ は，事前学習において，攻撃対象ユーザとの類似度をどの程度考慮するかを決定するための係数である．

$$\min_G \max_D \mathcal{L}(G,D) = \mathcal{L}_{wgan}(G,D) + \beta\mathcal{L}_T(G,u) \quad (1)$$

式 (1) における $\mathcal{L}_{wgan}(G,D)$ は，WGAN の損失関数であり，式 (2) で表される．ここで，$x$ は GAN に学習させる真のデータ分布に従う変数である．

$$\mathcal{L}_{wgan}(G,D) = E_x\left[D(x)\right] - E_z\left[D(G(z))\right] \quad (2)$$

また，$H(X,Y)$ を $X$ と $Y$ の二値交差エントロピー誤差，$Y_u$ をユーザ $u$ の識別子に対する One-hot 表現とすると，式 (1) における $\mathcal{L}_T(G,u)$ は式 (3) の通りである．

$$\mathcal{L}_T(G,u) = H\left[T(G(z)), Y_u\right] \quad (3)$$

ステップ 2 では，ステップ 1 で訓練した GAN を利用し，攻撃対象ユーザの顔画像に類似していると推測される画像を探索する．具体的には，式 (4) の目的関数を満たすように潜在変数 $z$ を探索し，得られた潜在変数 $\hat{z}$ に対応する画像を求める．

$$\hat{z} = \underset{z}{\mathrm{argmin}} - D(G(z)) - \lambda_u \log\left[T_u(G(z))\right] \quad (4)$$

ここで，$\lambda_u$ は，式 (1) の $\beta$ 同様，探索時において，攻撃対象ユーザとの類似度をどの程度考慮するかを決定するための係数である．$\hat{z}$ を求めるにあたって，確率的勾配降下法 (SGD) を用いて $z$ を最適化する．$T_u(G(z))$ は，$G(z)$ を $T$ に入力した場合の $u$ に対する確信度である．

## 4 実験・検証

### 4.1 実験概要

実験では，提案手法がなりすまし精度の高いベクトルを生成する探索空間を構築可能であるかを確認する．既存手法と提案手法それぞれで GAN の訓練を行うことで探索空間を構築し，それぞれの探索空間からランダムにサンプルしたベクトルによるなりすまし精度を比較する．さらに，提案手法において，類似度を考慮するほど推定データ探索時に早期段階でなりすまし精度が高くなるのかを確認するために，事前学習時の $\beta$ と探索時のエポック数の複数の組み合わせで攻撃を行い，なりすまし精度との関係を調べる．本実験におけるなりすまし精度は，生成器 $G$ による生成画像 $G(z)$ が攻撃対象モデル $T$ において攻撃対象ユーザ $u$ であると識別されるほど大きな値を取る指標であり，$T_u(G(z))$ の値を利用する．

### 4.2 実験手順

実験手順は，大きく分けて以下の 3 つの手順からなる．

1. 攻撃対象モデルの訓練
2. 攻撃用 GAN の訓練

3. 攻撃対象ユーザの画像推定

手順 1 では，攻撃対象モデルが非公開データの顔画像を識別できるように訓練する．

手順 2 では，手順 1 で訓練した攻撃対象モデルと公開データを用いて，攻撃用 GAN が公開データの顔画像分布に従った画像を生成可能となるように訓練する．手順 3 では，損失関数の $\beta$ の値を変化させ，$\mathcal{L}_T(G,u)$ の項がなりすまし精度に与える影響を調査する．

手順 3 では，手順 1 で訓練した攻撃対象モデルと手順 2 で訓練した攻撃用 GAN を用いて，攻撃対象ユーザに類似していると推測される画像を探索する．この際，手順 2 における $\beta$ と手順 3 における探索完了までのエポック数がなりすまし精度にどのように影響するか調べる．

### 4.3 実験に用いるデータセット

本実験では，実験用データセットとして，Large-scale CelebFaces Attributes (CelebA) Dataset [8] を用いる．このデータセットには，10,177 人の顔画像データ（以下，データとする）が存在し，合計 202,599 枚のデータが含まれている．本実験では，CelebA を非公開データと公開データに分割して利用する．ただし，公開データに非公開データと同一人物のデータが入っている場合，はじめから攻撃者が攻撃対象ユーザの顔画像を入手している前提となるため，分割する際は同一人物のデータが含まれないように注意する．

データセットを分割する前に，予め CelebA の全ての画像に対して MTCNN [9] を適用し，160×160 ピクセルの顔領域だけを抽出したものから構成される新たなデータセットを用意する．なお MTCNN によって顔領域が検出されなかった画像 297 枚をデータセットから取り除いた．

非公開データは，データセットから 1,000 人を選択し，その 1,000 人について 30 枚ずつ画像を選択したもので構成される．非公開データを 8:1:1 の割合で訓練データ，検証データ，テストデータとなるように，人物ラベルに関する層化分割を行い，攻撃対象モデルの訓練に利用する．

公開データは，データセットから非公開データを構成する際に選択された人物を除いた残りのデータから構成される．各画像は，モデルの入力サイズに合わせて拡大縮小を行い，攻撃用 GAN の訓練に利用する．

### 4.4 実験に用いるモデル

本実験では，攻撃対象モデル $T$，攻撃用 GAN の生成器 $G$ 及び識別器 $D$ の 3 つのモデルを訓練し，利用する．

攻撃対象モデル $T$ は，VGGFace2 [10] で訓練済みの FaceNet(Inception Resnet V1) [11, 12] に対して非公開データを用いて転移学習を行い，1,000 クラスの顔画像識別タスクを行えるように訓練したものを利用する．

表 1: 生成器のネットワーク構成

| Type | Kernel | Stride | Padding | Out channels |
|---|---|---|---|---|
| deconv | 4 | 1 | 0 | 512 |
| deconv | 4 | 2 | 1 | 256 |
| deconv | 4 | 2 | 1 | 128 |
| deconv | 4 | 2 | 1 | 64 |
| deconv | 4 | 2 | 1 | 3 |

表 2: 識別器のネットワーク構成

| Type | Kernel | Stride | Padding | Out channels |
|---|---|---|---|---|
| conv | 4 | 2 | 1 | 64 |
| conv | 4 | 2 | 1 | 128 |
| conv | 4 | 2 | 1 | 256 |
| conv | 4 | 2 | 1 | 512 |
| conv | 4 | 1 | 0 | 1 |

攻撃用 GAN の生成器 $G$ 及び識別器 $D$ は，DCGAN [13] のネットワーク構成と WGAN [7] の損失関数を組み合わせたモデルを利用する．生成器 $G$，識別器 $D$ のネットワーク構成の詳細は，表 1，表 2 の通りである．生成器 $G$ は，100 次元の潜在ベクトル $z$ を入力とし，64×64 ピクセルの RGB 画像を出力する．識別器 $D$ は，64×64 ピクセルの RGB 画像を入力とし，入力画像が公開顔画像データセットの分布に従う際に高い確率値を出力する．

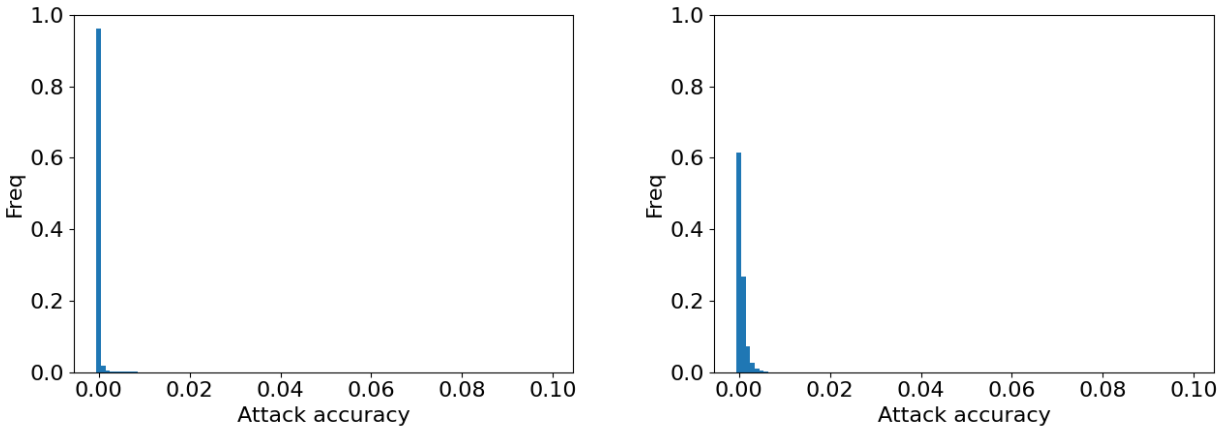
### 4.5 実験条件

実験は，AMD EPYC 7262 CPU，RTX3090 GPU，128GB RAM を搭載した，Ubuntu20.04.1(kernel 4.18) のマシン上で行った．また，プログラムの実装や実行，GPU の利用にあたって，NVIDIA driver 455.45.01，CUDA 11.1，libcudnn 8.0.5，nccl 2.8.3，Python 3.8，PyTorch 1.8.0a0+1606899 を利用した．

攻撃対象モデルの訓練では，オプティマイザに SGD を利用する．SGD のパラメータは，学習率 0.01，バッチサイズ 64，モメンタム 0.9，重み減衰 $10^{-4}$ である．訓練は 10 エポック行う．

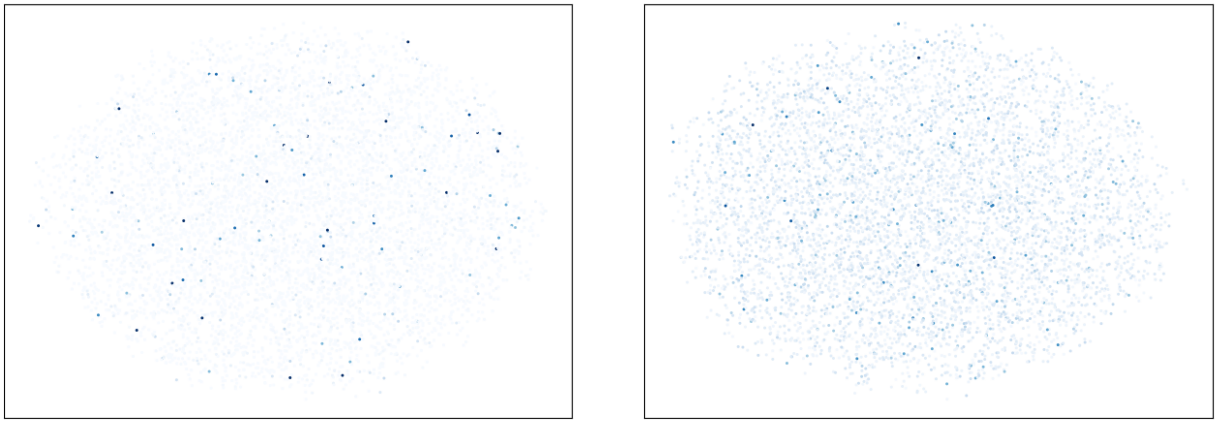
攻撃用 GAN の訓練では，オプティマイザに RMSProp を利用する．RMSProp のパラメータは，学習率 $5\times10^{-5}$，バッチサイズ 64 である．損失関数の $\beta$ の値は，{0.25, 0.5, 0.75, 1.0} の 4 種類を利用する．訓練は 1000 エポック行う．

攻撃対象ユーザの画像推定では，潜在ベクトルの探索に SGD を利用する．SGD のパラメータは，学習率 0.02，バッチサイズ 64，モメンタム 0.9 である．目的関数の $\lambda_u$ の値は，10 とする．潜在ベクトルの初期集団は，標準正規分布からランダムにサンプルした 512 個の 100 次元ベクトルである．探索は，{1, 10, 100} の 3 種類のエポック数で行う．




(a) 既存手法　　(b) 提案手法

図 2: 探索空間におけるなりすまし精度 [0.0，0.1] の出現頻度



(a) 既存手法　　(b) 提案手法

図 3: 次元圧縮によって可視化した探索空間（なりすまし精度 [0.0，0.01]）

### 4.6　実験結果

訓練データを用いて攻撃対象モデル $T$ の訓練を行った．結果として Accuracy が 95.6%の顔画像識別モデルが得られた．攻撃用 GAN の訓練及び攻撃対象ユーザの画像推定では，この訓練済みモデルを利用する．

#### 4.6.1　探索空間におけるなりすまし精度の比較

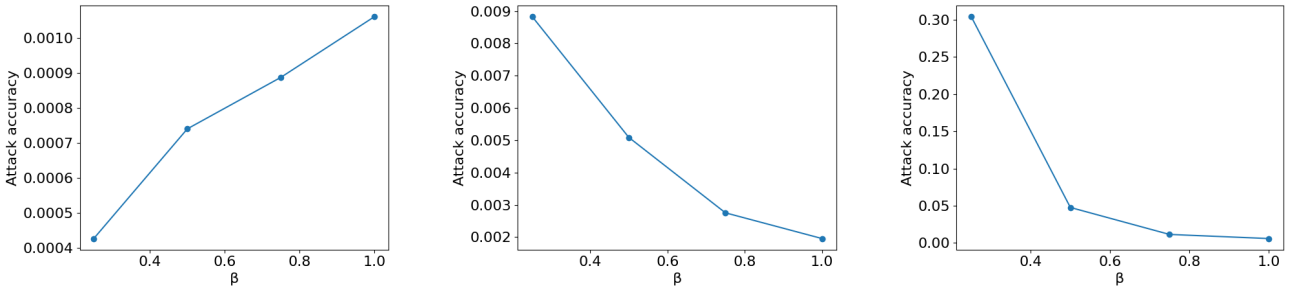
既存手法と提案手法それぞれで，攻撃用の GAN を訓練した．それぞれの探索空間の攻撃対象ユーザに対するなりすまし精度に関する調査結果を図 2(a)，図 2(b)，図 3(a)，図 3(b) に示す．ここでは，探索空間内の潜在変数 $z$ を生成器 $G$ に入力して得られる出力画像 $G(z)$ を攻撃対象モデル $T$ に入力した際の攻撃対象ユーザ $u$ に対するなりすまし精度との関係を調べた．

図 2(a)，図 2(b) は，既存手法，提案手法の探索空間におけるなりすまし精度 [0.0, 0.1] の区間の出現頻度をヒストグラムで表現したものである．既存手法と提案手法の結果を比較すると，提案手法の方が既存手法よりも探索空間内になりすまし精度が高い潜在ベクトルの出現頻度が高いことがわかる．

図 3(a)，図 3(b) は，既存手法と提案手法の探索空間からランダムに取り出した 6,400 個の 100 次元ベクトルを UMAP [14] を用いて次元削減を行うことで 2 次元に写像し，なりすまし精度 [0.00, 0.01] の区間を濃淡として表現したものである．既存手法と提案手法の結果を比較すると，提案手法は既存手法よりも全体的に濃淡が濃いことに加え，濃淡の濃い部分と薄い部分の差が小さいことが視覚的にも見てとれる．

#### 4.6.2　提案手法の $\mathcal{L}_T(G, u)$ が攻撃精度に与える影響

提案手法で GAN の訓練を行った後，1，10，100 の各エポック数で攻撃対象ユーザの画像推定を行い，$\beta$ となりすまし精度の関係をグラフにしたものを図 4(a)，図 4(b)，図 4(c) に示す．図にプロットされているのは，な




(a) 1 エポック　(b) 10 エポック　(c) 100 エポック

図 4: 提案手法の $\beta$ となりすまし精度の関係

りすまし精度の中央値である．図 4(a) より，エポック数が 1 の場合には，$\beta$ が大きいほどなりすまし精度が高くなることがわかる．また，図 4(b)，図 4(c) より，エポック数が 10 と 100 の場合には，エポック数が 1 の時とは異なり，$\beta$ が大きいほどなりすまし精度が低くなることがわかる．

## 5　議論

実験結果の図 2，図 3 より，提案手法により構築した探索空間が，既存手法よりも高い頻度で，なりすまし精度の高い潜在ベクトルを生成可能であることが分かった．これにより，ユーザ類似性を考慮した事前学習でなりすまし精度の高いサンプルが多い探索空間を作りたいという提案手法の意図通りの効果が確認できた．

図 4 より，提案手法における推定データの探索では，エポック数が小さい時は，$\beta$ が大きいほどなりすまし精度が高くなるが，エポック数が大きい時は，$\beta$ が小さいほどなりすまし精度が高くなった．これは，$\beta$，エポック数，なりすまし精度の関係は，エポック数に影響を受けずに $\beta$ が大きいほどなりすまし精度が高くなるといった予想に反する結果となった．図 2 で示したヒストグラムにおいては，提案手法は既存手法よりなりすまし精度の高い探索空間になっていると考えられる．これらの事実から，本実験においては推定データの探索が適切に行われていないことが推定できる．

探索が適切に行われない原因として，$\beta$ を増大することによって，探索空間内に攻撃対象ユーザに類似したベクトルが多くなり，なりすまし精度が極大値となる潜在ベクトルが大量に存在するようになった可能性が考えられる．極大値の数が増えたことにより，SGD による潜在ベクトルの更新ステップごとに極大値に陥る確率が高くなったのではないかと考える．探索が適切に行われない原因について更なる検討が必要である．

### 5.1　制限事項

本実験では，攻撃用 GAN と攻撃対象モデルの訓練に同じデータセット (CelebA) を分割したものを利用したが，実際に提案手法の攻撃を行う際には，攻撃用 GAN と攻撃対象モデルの訓練に用いる公開データと非公開データは異なるデータセットから構築されると考えられる．また，本実験では，攻撃対象モデルにおけるなりすまし精度のみを評価したが，ユーザのプライバシー漏洩の観点では，MI 攻撃によって推定された画像を人間が見て攻撃対象ユーザに類似していると判断されることもまた重要である．さらに，提案手法に対する防御手法については未検討である．防御手法に関して，差分プライバシーでの対策が難しいことは既存研究 [4] によって指摘されており，より有効な対策が何かを含めた検討が必要である．

## 6　結論

本研究では，生成モデルを利用するブラックボックス型 MI 攻撃において，生成モデルの訓練時にユーザ類似性を考慮することで，攻撃対象ユーザに特化した探索空間を構築可能な手法の提案及び評価を行った．評価実験では，Zhang らの手法をもとに考えた提案手法と既存手法のそれぞれで探索空間を構築し，提案手法によって構築した探索空間が，既存手法よりも高い頻度でなりすまし精度の高い潜在ベクトルを生成できることを示した．また，構築した探索空間を用いて攻撃対象モデルに攻撃を行った際のなりすまし精度を示した．

提案手法を用いて生成モデルの訓練を行うことで，ブラックボックス条件下で効率的な MI 攻撃を行える可能性があると考える．ただし，提案手法を用いて構築した探索空間の適切な探索方法は未検討であり，既存手法 [4] と同じ SGD を用いた探索を行うだけでは，高いなりすまし精度の攻撃は難しいことが実験により示された．

今後の課題は，提案手法を用いて構築した探索空間の適切な探索手法の検討及び防御手法の検討である．適切

な探索手法が見つかった場合，提案手法と組み合わせて高いなりすまし精度の MI 攻撃が可能となることが予想されるため，MI 攻撃からユーザのプライバシーを守るためにも防御手法を同時に検討する必要がある．

## 謝辞

本研究の一部は JSPS 科研費 18K11294 および 21H03442 の助成を受けて行われた．

## 参考文献

[1] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 41.1–41.12. BMVA Press, September 2015.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

[3] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 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, 2015.

[4] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 253–261, 2020.

[5] Ulrich Aïvodji, Sébastien Gambs, and Timon Ther. Gamin: An adversarial approach to black-box model inversion. *arXiv preprint arXiv:1909.11835*, 2019.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Vol. 27, pp. 2672—-2680. Curran Associates, Inc., 2014.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

[8] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

[9] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, Vol. 23, No. 10, pp. 1499–1503, 2016.

[10] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[12] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[13] Alec Radford, Luke Metz, and Soumith Chintala. 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.

[14] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.