特徴量抽出器を用いた個人照合に対する敵対的サンプル生成法

柿崎　和也† 　吉田　光佑† 　高橋　翼††

† 日本電気株式会社　〒 211–8666 神奈川県川崎市中原区下沼部 1753
†† LINE 株式会社　〒 160-0022 東京都新宿区新宿 4-1-6　JR 新宿ミライナタワー 23 階
E-mail: †k-kakizaki@ay.jp.nec.com, †k-yoshida@hq.jp.nec.com, ††tsubasa.takahashi@linecorp.com

あらまし　機械学習モデルに対して出力を意図的に誤らせる敵対的サンプルによるリスクが，特徴量抽出器を用いた個人照合においても同様に存在する．リスクを適正に把握するためには，高い精度で誤照合を引き起こす強力な敵対的サンプルの生成方法の実現が課題である．本稿では，特徴量抽出器を用いた個人照合に対してより高い精度で誤照合を誘発する敵対的サンプルの生成法を提案する．評価実験を通して，提案手法が複数のデータセットに対して既存手法よりも有用性が高いことを示す．

キーワード　敵対的サンプル，深層学習

1　はじめに

近年，機械学習モデルへの入力データに人が認識できない微小な摂動を加えることで，意図的に誤った推論結果を誘導する敵対的サンプルが機械学習・AI の活用に安全上の懸念を生じさせている [1] [2] [3]．顔や声といった生体情報を用いた個人照合タスクでは，悪意を持った敵対者が特定のターゲットになりすましを行うリスクが存在する [4] [5]．

顔画像や声データを用いた個人の照合では，学習段階で全ての潜在的なユーザのデータを訓練データとして用意することが難しい．そのため，当該分野では，不特定多数の個人を識別し得る特徴量を獲得するため，収集した訓練データから特徴量抽出器を学習する [6]．そして運用段階では，あらかじめ登録されている個人データの集合（テンプレート集合と呼ぶ）を用いて，入力データと最も近い特徴量を持つ個人を照合結果とする．このような個人照合タスクの敵対的サンプルに対する脆弱性を評価するためには，特徴量抽出器を騙す敵対的サンプルについて研究することが重要である．

End-to-End の分類器を騙す従来の敵対的サンプルとは異なり，特徴量抽出器を騙す手法として，[7] や [8] が挙げられる．これらの手法は，敵対的サンプルの元になるデータ（ソースデータと呼ぶ）に摂動を加えることで，ターゲットとなる個人データ（ターゲットデータと呼ぶ）と特徴量が近い敵対的サンプルを生成することが可能である．しかしながら，なりすましを図るにあたって、これらの手法は十分とは言えない．なぜなら，テンプレート集合内全てのデータとの相対的な距離関係により照合先が決定される個人照合タスクでは，ソースデータをターゲットデータに近づけるだけでは、照合先がターゲットデータ以外の個人データになる可能性が排除できないからである．

そこで本稿では，摂動を算出する際にテンプレート集合内全体を考慮に入れた敵対的サンプルの生成手法を提案する．具体的には、テンプレート集合内の全ての個人データとの特徴量空間での距離を元に摂動の算出を行い照合先がターゲットとする個人になるようにする．図 1 は，摂動の大きさを固定した際に既存手法と提案手法がそれぞれ生成する敵対的サンプルの違いを表した例である．既存手法は，ターゲットクラスとの最小距離となる敵対的サンプルを生成するため，結果としてターゲットデータよりも近いテンプレートデータが存在し，ターゲットへのなりすましが成功しない．それに対して，提案手法は，テンプレート集合内の全ての個人データとの距離関係を考慮して，テンプレート集合内でターゲットが最も近くなる敵対的サンプルを生成するためにターゲットへのなりすましが成功する．

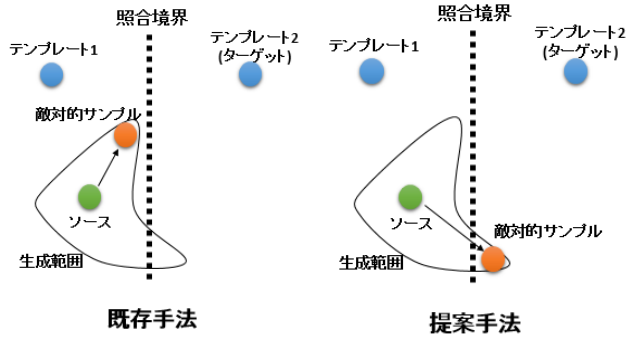


図 1: なりすまし（ターゲットクラスへ誤照合）を誘発する敵対的サンプルの探索．図内の輪は，摂動の大きさを固定したときに決定される敵対的サンプルの生成範囲を表す．既存手法 [7] [8] は生成範囲内でターゲットクラスとの最小距離点を探索するが，ターゲット以外との関係を考慮していない．提案法はテンプレート集合内でターゲットが最も近くなる点を探索する．

本稿では，複数の大規模データセットを用いた評価実験により，実際に提案手法は，既存手法と比較してターゲットに対するなりすましの成功率が高いことを示す．

本論文の貢献：本稿の貢献は以下の通りである．

(i)　特徴量抽出器とテンプレート集合を用いた個人照合において，新たな敵対的サンプル生成手法の提案する．

(ii)　実顔画像データセットを用いて既存手法との比較実験を行い、提案手法の方がターゲットに対するなりすましの成功率が高いことを実験的に示す．

本稿の以下のように構成されている．まず，2 章では，本研究に関連する既存研究を示すことで，本研究の提案手法の位置づけを明確化する．次に，3 章で個人照合タスクの詳細と脅威モデルについて説明し，4 章で提案手法の定式化を行う．5 章では，複数の顔データセットを用いて既存研究との比較実験を行い，提案手法が既存手法と比較してより高い精度でなりすましが成功する敵対的サンプルが生成できることを示す．最後に，6 章では考察と残課題について議論する．

2 関 連 研 究

本章では，関連研究について説明し，本研究の位置づけを明確化する．

分類器に対する敵対的サンプル：分類器に誤分類を引き起こす敵対的サンプルの生成に取り組んだ代表的な研究として [2] [3] [9] などが挙げられる．いずれも分類器から算出される勾配を元に，微小な摂動をソース画像に加えることで敵対的サンプルを生成する．また，誤分類を引き起こすことが可能な眼鏡やシールの生成に取り組んだ研究なども挙げられる [4] [5] [10]．これらの手法はいずれも分類器をターゲットとして提案された手法であり，本稿が対象とする特徴量抽出器に基づく個人照合に対する敵対的サンプルを生成するには，さらなる工夫が必要である．

特徴量抽出器に対する敵対的サンプル：特徴量抽出器に対する敵対的サンプル生成手法として [7] と [8] がある．[7] では，顔照合のための特徴量抽出器に対して敵対的サンプルを生成する手法を提案している．著者らは，特徴量抽出器 f が与えられたときソースとなる顔画像 x_s の特徴量と任意のターゲットの顔画像 x^t の特徴量を近づけることで，敵対的サンプルが生成されることを示している．具体的には，ソースとなる画像 x_s に対して，勾配方向 $\nabla_{x_s}\|f(x_s)-f(x^t)\|_2$ の摂動を加える．これを誤照合が引き起こされるまで繰り返し行っていくことで敵対的サンプルを生成することができる．[8] では，小さい摂動で特徴量抽出器を騙す敵対的サンプルを生成するために，以下の制約付き最適化問題を解いている．

$$\begin{aligned} &\min_{x} \quad \|f(x)-f(x^t)\|_2 \\ &s.t. \quad \|x-x_s\|_\infty < \delta. \end{aligned} \tag{1}$$

制約条件である式 (1) を加えることで，摂動が元のデータ空間で一定の値を下回ることを保証している．このとき δ が小さいほど，摂動が小さくなり人間が認識するのが難しくなる．[7] や [8] の手法を用いると，特徴量空間上でターゲットデータ x^t と距離が近い敵対的サンプルを生成することができる．しかし，これらの手法では，ターゲット t とテンプレート集合 $T=\{x^1,...,x^{|T|}\}$ に対して，ターゲット以外のテンプレート集合内のデータ $T-\{x^t\}$ との距離が考慮されていない．従って，テンプレート集合内全てのデータとの距離関係に基づく個人照合タスクに即した個人照合タスクに即した手法ではなく，なりすましが成功する精度が低くなる．

本稿は，これら先行研究を改良した手法を提案し，提案手法が既存手法よりなりすましが成功する精度が高いことを示す．

表 1: 記号表

記号	定義
x	クエリデータ (任意次元のベクトル)
x_s	ソースデータ (x と同次元のベクトル)
x_{adv}	敵対的サンプル (x と同次元のベクトル)
x^i	テンプレートデータ (x と同次元のベクトル)
x^t	ターゲットデータ (x と同次元のベクトル)
T	テンプレート集合
f	特徴量抽出器
w	代替パラメータ (x と同次元のベクトル)
l	学習率 (スカラー)
c	目的関数調節パラメータ (スカラー)
c_{init}	c の初期値 (スカラー)
c_{max}	c の最大値 (スカラー)
c_{min}	c の最小値 (スカラー)
M_1	c のバイナリサーチ回数 (スカラー)
M_2	w の更新回数 (スカラー)
ϵ	摂動の大きさ (スカラー)

3 準　　備

本章では，対象とする個人照合タスクの詳細，および提案手法の前提となる脅威モデルに関して説明する．なお，本稿で用いられる記号は表 1 の通りである．

3.1 個人照合タスク

特徴量抽出に基づく個人照合タスクでは，クエリデータ x とテンプレート集合 $T=\{x^1,...,x^{|T|}\}$ を用いて，x と最も近い特徴量をもつ $x^i \in T$ の個人 i を照合結果として返す．テンプレート集合とは事前に登録した照合先のデータ x^i の集合で，各個人に一つのデータのみ含む．特に本稿の特徴量抽出器 $f:\mathcal{X}\to\mathcal{Z}$ は，式 (2) のようにテンプレート集合内のデータ $x^i \in T$ と，入力データ x の特徴量空間における L2 距離 $\|f(x)-f(x^i)\|_2$ が最も小さくなる個人 i を照合結果とする．

$$i = \arg\min_i \|f(x)-f(x^i)\|_2. \tag{2}$$

3.2 脅威モデル

本稿では，真に特徴量抽出器のリスクを評価するために，ワーストケースを想定する．具体的には，敵対者は学習済み特徴量抽出器 f のパラメータ及びアーキテクチャに加えて，テンプレート集合 T についても知っているものと仮定する．

4 提 案 手 法

本章では，テンプレート集合内の全てのデータを考慮する新たな敵対的サンプル生成手法を提案する．本手法では，テンプレート集合内の全ての個人データとの特徴量空間での距離を元に，小さな摂動でターゲットとして照合される敵対的サンプルの生成を目指す．特に，ターゲットクラスとの距離が，それ以外のクラスよりも近くなるような点を探索することで，なりすましの成功率の向上を図る．

4.1 問題の定式化

特徴量抽出器 f，ソースデータ x_s，ターゲットデータ x^t，テンプレート集合 $T = \{x^i, ..., x^{|T|}\}$ が与えられたときに，前述の問題を解決する敵対的サンプルは，以下の制約付き最適化問題で定式化できる．

$$\min_{x_{adv}} \quad \|x_{adv} - x_s\|_2 \tag{3}$$

$$s.t. \quad h(x_{adv}, x^t, T) < 0, \tag{4}$$

ここで，

$$\begin{aligned} h(x_{adv}, x^t, T) =& \|f(x_{adv}) - f(x^t)\|_2 \\ & - \min_{x^i \in T : x^i \neq x^t} \|f(x_{adv}) - f(x^i)\|_2, \end{aligned} \tag{5}$$

である．

式 (5) は，$h(x_{adv}, x^t, T) < 0$ を満たすとき，$x^i \in T$ に対して $t = \arg\min_i \|f(x_{adv}) - f(x^i)\|_2$ が成り立ち，敵対的サンプル x_{adv} がターゲット t へのなりすましが成功することを示す．従って，制約 (4) の下で式 (3) を最小化することで，なりすましが成功する中で摂動が最小な敵対的サンプルが生成される．これにより，同じ摂動の大きさで比較した際に，既存手法に比べてなりすましの成功率が高くなると期待できる．

上記の最適化問題はラグランジュ未定乗数法を用いて，以下の目的関数を最小化する問題に帰着される．

$$J(w, x_s, x^t, T, c) = \|g(w) - x_s\|_2 + c \cdot h(w, x^t, T). \tag{6}$$

ただし，

$$\begin{aligned} h(w, x^t, T) =& \|f(g(w)) - f(x^t)\|_2 \\ & - \min_{x^i \in T : x^i \neq x^t} \|f(g(w)) - f(x^t)\|_2, \end{aligned}$$

$$g(w) = \frac{d}{2}(\tanh(w) + 1),$$

であり，c は $c > 0$ を満たす定数であり，d は x_{adv} の各要素の上限値を表す．

上記の最適化問題では，x_{adv} に代わり，新たに導入した x と同じ次元で表される代替パラメータ w に対して目的関数を最小化する．[3] と同様に，得られる画像が実現可能な定義域から外れることがないよう，任意の w で $0 \leqq g(w) \leqq d$ を満たす新たな関数 $g(w) = x_{adv}$ を用いる．本稿では，画像データを対象とするため，$d = 255$ を用いる．

また，$c > 0$ は摂動の大きさを表す第一項とテンプレート集合内のテンプレートデータとの距離を表す第二項のバランスを決めるハイパーパラメータである．

4.2 敵対的サンプル生成アルゴリズム

Algorithm 1 に敵対的サンプル生成アルゴリズムを示す．式 (6) は，定数 c により第一項と第二項のバランスが決定される．そのため，なりすましが成功する中でより摂動が小さな敵対的サンプルを生成するために，Algorithm 1 では c の値をバイナリサーチにより変更しながら，勾配法を用いることで敵対的サンプルを生成していく．Algorithm 1 は，ソースデータ x_s，

Algorithm 1 Find Adversarial Example

Require: Source:x_s, Target:x^t, Template:T, learning rate:l, Parameter:$c_{init}, c_{max}, c_{min}, M_1, M_2$
Ensure: Adversarial Example x_{adv}

1: $S = \{\}$: Set of adversarial examples
2: $c \leftarrow c_{init}$
3: **for** $i = 1$ to M_1 **do**
4: set initial value to w_0
5: **for** $j = 1$ to M_2 **do**
6: $w_j \Leftarrow w_{j-1} - l \cdot \nabla_w J(w, x_s, x^t, T, c)$
7: **if** $h(w_j, x^t, I) < 0$ **then**
8: $S \leftarrow S \cup \{x_{adv}\}$
9: **end if**
10: **end for**
11: **if** $h(w_{M_2}, x^t, I) < 0$ **then**
12: $c_{max} \leftarrow c$
13: $c \leftarrow \frac{c + c_{min}}{2}$
14: **else**
15: $c_{min} \leftarrow c$
16: $c \leftarrow \frac{c + c_{max}}{2}$
17: **end if**
18: **end for**
19: $w_{opt} \leftarrow \arg\min_{w \in S} \|g(w) - x_s\|_2$
20: **return** $g(w_{opt})$

ターゲットデータ $x^t \in T$，テンプレート集合 T，学習率 l，定数 c の初期値 c_{init}，最大値 c_{max}，最小値 c_{min}，c に対するバイナリーサーチの探索回数 M_1，勾配法による w の更新回数 M_2 がそれぞれ入力として与えられる．攻撃が成功した場合（すなわち $h(w, x^t, T) < 0$ が成立する場合）の中で，最も摂動の小さい敵対的サンプルを出力する．このアルゴリズムは，更新回数 M_1 回の外側のループと M_2 回の内側のループの二重ループで構成されている．内側のループでは，ある $c > 0$ が与えられた時に勾配法を用いて M_2 回 w の更新を行い (5〜9 行)，外側のループでは，なりすましが成功するかどうかを元にバイナリーサーチを用いて次の定数 c を決定する (11〜17 行)．ここで，バイナリサーチでは，なりすましが成功した際には，さらに大きな値を次の c として設定され (13 行)，なりすましが失敗したし際には，さらに小さな値を次の c として設定される (16 行)．

5 実　　験

本章では，実顔画像データと顔特徴量抽出器から成る個人照合に対して，提案手法と既存手法を用いて敵対的サンプルを作成し，提案手法がより高い精度でなりすましが成功する敵対的サンプルを生成できることを実験的に示す．

5.1 実 験 設 定

評価指標：本実験では，なりすまし成功率 Accuracy(ϵ) と，敵対的サンプルの摂動サイズを評価指標として用いた．

なりすまし成功率 Accuracy(ϵ) は，生成された敵対的サンプルのうち，摂動の大きさが ε 以下でなりすましが成功した敵対

表 2: VGGFace 特徴量抽出器の照合精度

照合精度	$\|T\|=3$	$\|T\|=5$	$\|T\|=10$
VGGFace データセット	0.985	0.973	0.958
Pubfig データセット	0.894	0.828	0.774

的サンプルの割合を評価する．なりすまし成功率 Accuracy(ϵ) を式 (7) で定義する．

$$\text{Accuracy}(\epsilon) = \frac{|\{x|d(x,x_s) < \epsilon, x^t = r(x,T), x \in AX\}|}{|AX|}, \tag{7}$$

ここで，

$$d(x,x_s) = \|f(x) - f(x_s)\|_2, \tag{8}$$
$$r(x,T) = \arg\min_{x^i \in T} \|f(x) - f(x^i)\|_2,$$

AX は生成された敵対的サンプルの集合，f は特徴量抽出器，x^t はターゲット画像，x_s は敵対的サンプルのソース画像，T はテンプレート集合を表す．

摂動サイズは，生成された敵対的サンプルの摂動の大きさを評価する．摂動サイズを式 (8) で定義する．

データセット: 本実験では，多様な顔画像を含む大規模なデータセットとして，VGGFace データセット [11] と Pubfig データセット [12] を用いた．VGGFace データセットは VGGFace の学習に用いられた 2,622 個人の 2,604,175 枚の顔データである．Pubfig データセットは、200 個人の 58797 枚の顔画像である．

特徴量抽出器: 本実験では，顔データに対する特徴量抽出器として学習済み VGGFace 特徴量抽出器を用いた [11]．VGGFace 特徴量抽出器は，VGGFace データセットで学習されたもので，FC7 層の出力を特徴量抽出器の出力と見なすことで，顔画像に対する特徴量抽出器として使うことができる．

VGGFace データセットと Pubfig データセットにおける，VGGFace 特徴量抽出器の照合精度を，テンプレート数 $|T| = \{3,5,10\}$ の条件で評価した．各データセットからランダムに $|T|$ 人選択し，$|T|$ 人からそれぞれランダムに選択される 10 枚のクエリ画像と評価テンプレート 1 枚を用いて照合精度を評価した．この操作を 100 回繰り返し，照合精度の平均を，各特徴量抽出器の照合精度として計測した．照合精度を表 2 に示す．いずれのテンプレート数においても，VGGFace データセットでは照合精度は 95% 以上，Pubfig データセットにおいては 77% 以上であった．

各敵対的サンプル生成手法の設定: 本実験では，提案手法, [7] の手法及び [8] の手法により敵対的サンプルを生成した．ここで，本実験における各手法の設定について説明する．

提案手法では，パラメータ $c_{init} = c^{-3}, c_{max} = 10^{10}, c_{min} = 0, M_1 = 9, M_2 = 100$ を用いて敵対的サンプルの生成を行った．

[7] の手法は，本来摂動が加えられた後ピクセルの整数値への丸め込みが行われる．しかし，丸め込みを実施しない提案手法との公正な比較を行うため，本実験では丸め込みを行わない．また，摂動を加える最大回数は 100 回とした．

[8] の手法は，摂動の L_∞ ノルムを制約として，ターゲットデータと敵対的サンプルの L_2 距離を最小化することで敵対的サンプルの生成を行う．しかし，摂動の L_2 ノルムを最小化する提案手法との公正な比較を行うため，本実験では制約式の L_∞ ノルムを L_2 ノルムに変更し敵対的サンプルの生成を行う．そして，提案手法と同様にラグランジュ未定乗数法を用い Algorithm 1 と同様に敵対的サンプルの生成を行った．この際，各種パラメータは Algorithm 1 と同様の値を用いた．

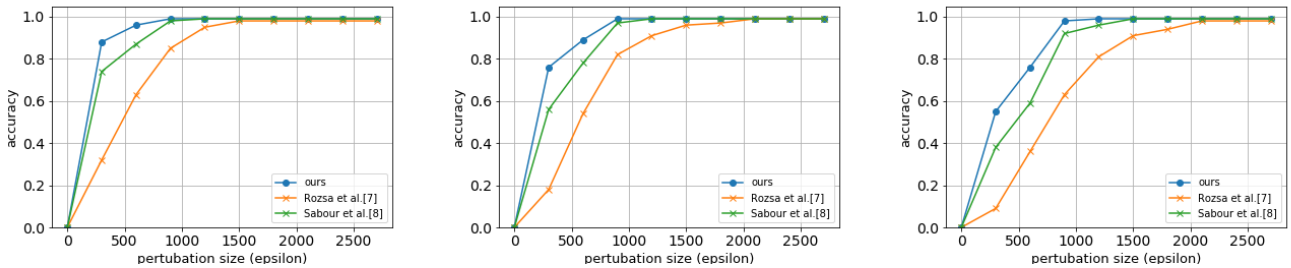
敵対的サンプル生成の手順: 本実験では，テンプレート数が異なる三つのケース ($|T| = \{3,5,10\}$) について評価を行った．それぞれのテンプレート数に対して，評価に用いるテンプレート集合 T，ソース画像 x_s 及びターゲット画像 x^t は，ランダムに 10 セット，個人の重複がないように選択した．ただし，ターゲット画像はテンプレート画像の中からランダムに選択した．また，各セットについてそれぞれ 10 個人のソース画像について敵対的サンプルを生成し，合計 10 個人 × 10 セット = 100 枚の敵対的サンプルを生成し評価した．

5.2 評価結果

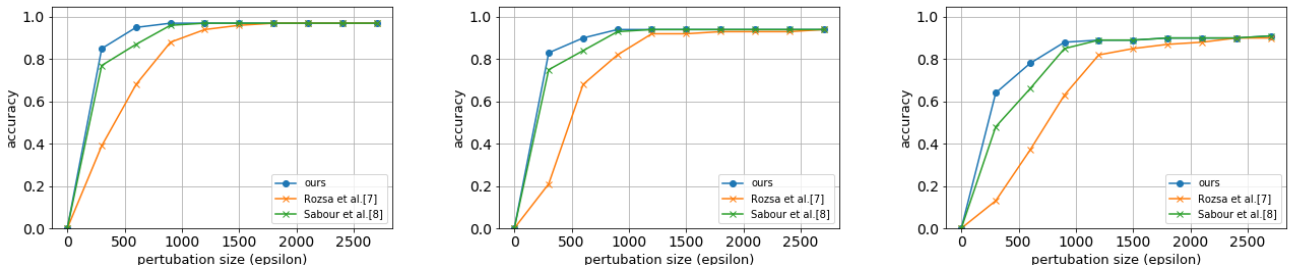
図 2 は横軸を摂動の大きさ ϵ としたとき，各手法のなりすまし成功率 Accuracy(ϵ) を示している．図 2 を見ると，いずれのデータセットおよびテンプレート数においても，全ての手法のなりすまし成功率が，ϵ が大きくなるにつれてそれぞれ同じ値に漸近していることが確認できる．また，提案手法は既存手法と比較して，より早くなりすまし成功率がある値に漸近していることが確認できる．従って，いずれの ϵ においても提案手法は既存手法と比べて同等以上のなりすまし成功率であり，小さな ϵ では提案手法の方がなりすまし成功率が高いと言える．具体的には，$\epsilon = 500$ において，VGGFace データセットでは [7] の手法と比較して最大 31%，[8] の手法と比較して最大 19% なりすまし成功率は高く，Pubfig データセットでは [7] の手法と比較して最大 36%, [8] の手法と比較して最大 16.9% なりすまし成功率は高い．よって，既存手法と比較して提案手法はより高い精度でなりすましが成功する敵対的サンプルを生成できていると言える．

また，表 3 は，同じなりすまし成功率のときの摂動の大きさの平均と標準偏差を表している．提案手法は，VGGFace データセットでは [7] の手法と比較して最大 67.5%，[8] の手法と比較して最大 42.5% 摂動平均が小さくなり，Pubfig データセットでは [7] の手法と比較して最大 64.4%, [8] の手法と比較して最大 27.5% 摂動平均が小さくなった．なりすまし成功率が同じとき，摂動が小さいほどより強力な攻撃であると言えるので，提案手法の優位性が言える．

また，図 3 に提案手法により生成された敵対的サンプル，そのソース画像，及びターゲット画像のペアの一例を示す．いずれの例でも，ターゲット画像と敵対的サンプルの間に差異を目視で確認することは難しく，小さな摂動の大きさで敵対的サンプルが生成できていることが確認できる．



$|T| = 3$(VGGFace データセット) $|T| = 5$(VGGFace データセット) $|T| = 10$(VGGFace データセット)



$|T| = 3$(Pubfig データセット) $|T| = 5$(Pubfig データセット) $|T| = 10$(Pubfig データセット)

図 2: 生成された敵対的サンプルのなりすまし成功率

表 3: 生成された敵対的サンプルにおけるなりすまし成功率と摂動の平均，標準偏差

$|T| = 3$(VGGFace データセット)

	成功率	平均	標準偏差
提案手法	0.99	**137.1**	170.0
[7]	0.99	422.1	368.4
[8]	0.99	238.8	247.5

$|T| = 5$(VGGFace データセット)

	成功率	平均	標準偏差
提案法	0.99	**233.8**	206.6
[7]	0.99	564.6	419.8
[8]	0.99	324.4	276.2

$|T| = 10$(VGGFace データセット)

	成功率	平均	標準偏差
提案法	0.99	**356.5**	266.3
[7]	0.99	772.8	523.1
[8]	0.99	479.8	328.8

$|T| = 3$(Pubfig データセット)

	成功率	平均	標準偏差
提案手法	0.97	**137.9**	160.4
[7]	0.97	371.3	378.0
[8]	0.97	190.3	240.7

$|T| = 5$(Pubfig データセット)

	成功率	平均	標準偏差
提案手法	0.93	**163.9**	163.5
[7]	0.93	460.6	393.2
[8]	0.93	215.0	214.8

$|T| = 10$(Pubfig データセット)

	成功率	平均	標準偏差
提案法	0.90	**303.2**	359.6
[7]	0.90	696.7	529.1
[8]	0.90	413.3	393.9

6 終わりに

本稿では特徴量抽出器を用いた個人照合に対する敵対的サンプル生成手法を検討した．先行研究 [7] [8] は，敵対的サンプルを生成するために，ターゲットとなる個人データと特徴量が近くなるように摂動を算出するのに対して，本稿では，テンプレート集合内の全てのデータとの特徴量空間での距離を考慮し，なりすましが成功するように摂動を算出する手法を提案した．2 つの顔画像データセットを用いた評価を通して，提案手法が既存手法よりも高い精度でなりすましが可能であることを示した．

本稿の結果を踏まえ，個人照合のために特徴量抽出器を用いる際には，敵対的サンプルへの対策を講じることを考慮すべきである．対策方法の一つとしては，敵対的サンプルを訓練データに混ぜてモデルの学習を行う Adversarial Training が挙げられる [13]．また，分類器においては，敵対的サンプルに対して所定のロバスト性を担保しつつ分類器を学習する方法 [14] [15] が近年提案されている．分類器同様，特徴量抽出器のロバスト性を担保した学習方法の実現が今後の課題として挙げられる．

文　献

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[2] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, 2017.

[4] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1528–1540, 2016.

[5] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. *arXiv preprint arXiv:1801.00349*, 2017.

[6] Wang Mei and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

[7] Andras Rozsa, Manuel Günther, and Terranee E Boult. Lots

about attacking deep features. In *International Joint Conference on Biometrics (IJCB)*, pp. 168–176, 2017.

[8] Sara Sabour, Yanshuai Cao, Fartash Faghri, and David J Fleet. Adversarial manipulation of deep representations. In *International Conference on Learning Representations (ICLR)*, 2016.

[9] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[11] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[12] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[13] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations(ICLR)*, 2018.

[14] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6542–6551, 2018.

[15] Hajime Ono, Tsubasa Takahashi, and Kazuya Kakizaki. Lightweight lipschitz margin training for certified defense against adversarial examples. *arXiv preprint arXiv:1811.08080*, 2018.

図 3: 提案手法により生成された敵対的例: 左の列はテンプレート画像の中から選択されたターゲット画像，中央の列は，敵対的サンプルのソース画像，右の列は生成された敵対的サンプルを表し，それぞれの行が，ターゲット画像，ソース画像，敵対的サンプルの一組のペアを表す．また，上 2 行は VGGFace データセットの例，下 2 行は Pubfig データセットの例を表す