Vol.2019-CVIM-217 No.12
2019/5/30

敵対的生成ネットワークを用いた角膜表面反射画像からのシーン識別

枝本 祐典[1,a)]　中澤 篤志[1]　西田 豊明[1]

概要：本論文では，目の表面反射（角膜表面反射）画像からシーンを高精度に自動的に識別する手法を提案する．角膜表面反射画像にはシーンの映像が反射していることが知られているが，低い解像度や虹彩テクスチャ，まつげやその陰などが含まれており，シーン識別タスクに用いるには困難であることが知られている．そこで本研究では，深層学習の手法，具体的には 1 ）Generative Adversarial Network（GAN）を用いて角膜表面反射画像を高精度化しシーン識別に用いる手法，2 ）GAN の Discriminator の出力ベクトルによりシーン識別を行う手法，を実装する．また，全方位ディスプレイにシーン画像を投影し，その中で被験者の角膜表面反射を撮影することでデータセットを作成し，評価を行った．

Scene Identification from Corneal Surface Reflection Images Using Generative Adversarial Networks

Yusuke Edamoto[1,a)]　Atsushi Nakazawa[1]　Toyoaki Nishida[1]

1. 序論

近年，画像を使った研究は様々な応用がなされており我々の生活に根ざしており，また，カメラから得られたシーン画像から場所を特定する Visual Odometry 技術は古くから研究がなされている [1]．しかし，記念写真や自己撮影画像などは，背景シーンが映っていない状況も多く，背景シーンを用いたシーン識別が困難である．これに対し，人物画像中の角膜表面反射画像からシーン識別をすることを行うことは，犯罪捜査への活用が期待されるともに，新たなプライバシー問題への対策検討につながる重要なタスクである．

角膜表面反射画像はシーン識別に用いるために，撮影された画像から目領域を切り出し，目の幾何モデルを考慮して展開することで，通常のカメラ画像と同様の透視投影による画像にすることができる．しかし，この画像を用いてそのままシーン識別すると，角膜表面反射画像に特有の以下のようなノイズにより識別精度が上がらない [2]．

- 虹彩のパターンが角膜表面反射したシーン画像と混ざり合う．
- まつげが角膜に直接覆いかぶさる．
- 帽子やまぶたが影になりシーン画像を遮って角膜まで到達できなくなる．

この理由から，角膜表面反射画像からシーン識別精度を向上にはこれらのノイズを取り除く必要がある．

本研究では，近年画像変換の技術として盛んに研究が行われている敵対的生成ネットワーク（GAN）を用いて角膜表面反射画像特有のノイズを取り除き，その画像を用いてシーン認識を高精度で行う手法を提案する．具体的には，以下の 2 つの手法を提案する．1 つ目は，GAN により角膜表面反射画像をシーン画像のドメインに変換した画像から抽出した Vector of Locally Aggregated Descriptors（VLAD）特徴量を用いてシーン識別を行うという手法である．2 つ目は，GAN の Discriminator にベクトルを出力させ，そのベクトルを特徴量として用いてシーン識別を行うという手法である．以上の 2 つの提案手法によりシーン識別精度を向上させることを目指した．

[1] 京都大学，〒 606-8501 京都府京都市左京区吉田本町
[a)] edamoto@ii.ist.i.kyoto-u.ac.jp

2. 関連研究

2.1 画像によるシーン識別に関する研究

画像によるシーン識別は Visual Odometry とも呼ばれ多くの研究が存在するが，基本的には，識別対象とする画像の特徴量と予め得ておいた画像特徴量を比較し，最も一致するものからシーンを同定する手法が多い．これらに属する研究として，Torii ら [3] は Vector of Locally Aggregated Descriptors（VLAD）[4] という特徴量を用いてシーン識別する手法を提案した．VLAD は任意の局所特徴量を集約することができ，位置ずれに強く次元数を大幅に圧縮することのできる特徴量である．さらに主成分分析などによる次元圧縮も可能で，多数のデータの貯蓄ができるため，検索性能が高いことが知られている．

2.2 敵対的生成ネットワークに関する研究

Generative Adversarial Networks（GAN）[5] は, 生成器 (generator) と識別器 (discriminator) から構成される画像生成モデルである．Generator は，潜在変数から生成した画像を Discriminator が本物の画像と判定するように学習し，Discriminator は，本物の画像か生成画像かを正しく判定するように学習することで，Generator がより本物らしく画像を生成することを可能にしている．

Conditional GAN（cGAN）[6] は，GAN を条件付きモデルに拡張したものである．Generator は潜在変数とラベルを入力として画像を生成し，Discriminator は画像とラベルに基づき本物の画像か生成画像かを判定することで，Generator がラベルに基づく画像を生成することを可能にしている．

pix2pix[7] は，cGAN を拡張することで，2 つのドメイン間の変換を学習することを可能にしたモデルである．ドメイン X から Y への学習するために，ドメイン X の各画像に 1 対 1 対応するドメイン Y の画像を用意して，cGAN における入力ラベルをドメイン X の画像に，本物の画像をドメイン Y の画像とすることで，generator がドメイン X から Y に画像を変換することを可能にしている．

CycleGAN[8] は，GAN を拡張することで，1 対 1 対応する訓練データを用いずに 2 つのドメイン間の相互変換を学習することを可能にしたモデルである．GAN を 2 セット用意し，それぞれの generator の潜在変数を他方の本物の画像とすることで，2 つの generator がそれぞれドメイン X から Y，Y から X への画像変換を学習する．2 つのドメイン間に共通する構造を保ったまま変換するために，再構築誤差 (cycle consistency loss) を generator の損失関数に追加している．

Auxiliary Classifier GAN（ACGAN）[9] は，GAN の discriminator がクラス識別も行うように拡張したモデルである．generator は cGAN のように潜在変数とラベル (クラス) を入力として画像を生成し，discriminator は，本物の画像か生成画像かの判定と画像に対するクラス出力をするように拡張し，generator と discriminator が協力して，出力クラスを生成画像なら生成時に入力したクラス，本物の画像ならその画像の属するクラスとなるように学習することで，generator の生成画像をより高精度なものにしている．

3. 提案手法・アルゴリズム

3.1 提案手法の概要

本研究では，2 種類の手法を用いてシーン識別を行った．

1 つ目は，GAN による画像のドメイン変換と VLAD によるシーン識別である．これは，VLAD のみを用いる従来手法を拡張したものである．図 1 のように処理を行う．まず，角膜表面反射画像を平面状に展開した画像を GAN の Generator を用いてシーン画像のドメインに変換する．次に，その生成画像から VLAD 特徴量を抽出する．最後に，抽出した VLAD 特徴量を用いて，予め抽出しておいたデータセットの全てのシーン画像の VLAD 特徴量との類似度を計算し，類似度が高い順に画像を出力する．

2 つ目は，GAN の Discriminator によるシーン識別である．図 2 のように処理を行う．まず，角膜表面反射画像を平面状に展開した画像を GAN の Generator を用いてシーン画像のドメインに変換する．次に，その生成画像を GAN の Discriminator に入力してベクトルを出力させる．また，データセットの全てのシーン画像も Discriminator に入力して出力ベクトルを得る．最後に出力されたベクトルを特徴量として用いて，全てのシーン画像との類似度を計算し，類似度が高い順に画像を出力する．

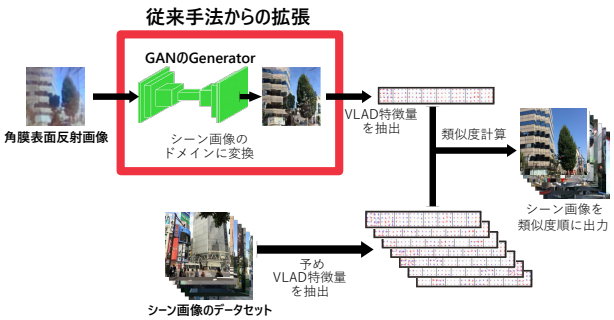

図 1 画像のドメイン変換と VLAD によるシーン識別の手法

3.2 画像のドメイン変換と VLAD によるシーン識別

シーン識別に用いる VLAD 特徴量は，任意の局所画像特徴量を利用して変換することができるが，本研究では拡大縮小・回転・照明変化に強い Dense SIFT 特徴量を利用する．さらに計算された VLAD 特徴量を主成分分析により 4096 次元に次元圧縮し正規化する．

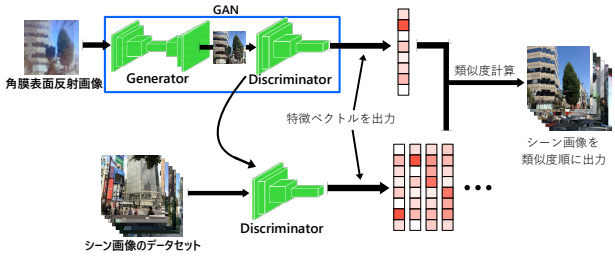

図 2　GAN の Discriminator によるシーン識別の手法

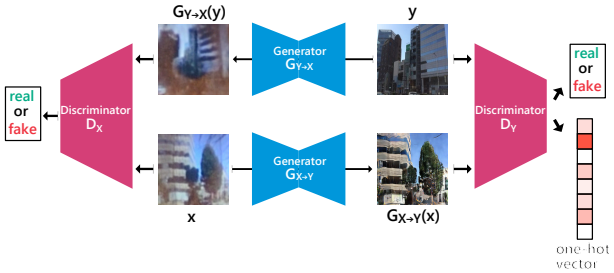

図 3　CycleACGAN のネットワーク構造

正規化されている特徴量ベクトルを用いるので，ドット積によって得られる cos 類似度を画像間の類似度として用いる．

画像のドメイン変換には以下の小節で述べる GAN を用いる．

3.2.1　pix2pix

Zhu らの公開しているコード [*1] を基に実装した．

ネットワークは Generator と Discriminator からなる．Generator は U-net[10] をベースとして構成した．Discriminator は PatchGAN[7][11] をベースとして構成した．

3.2.2　CycleGAN

Zhu らの公開しているコード[*1] を基に実装した．

ネットワークは 2 つの Generator と 2 つの Discriminator からなる．2 つの Generator, 2 つの Discriminator はそれぞれ同じネットワーク構造である．Generator は ResNet[12] をベースとして構成した．Discriminator は pix2pix と同じものを用いている．

3.2.3　CycleACGAN

ACGAN のように discriminator がクラスを出力し generator と discriminator が協力してクラス識別を学習する仕組みを CycleGAN に加えた．具体的には，CycleGAN のシーン画像の真偽を判定する Discriminator の出力にクラスを表す one-hot vector を加え，角膜表面反射画像からシーン画像のドメインに変換する generator とシーン画像の真偽を判定する Discriminator が協力してクラス識別を学習するようにしたものである．

小節 3.2.2 で述べた CycleGAN を拡張することで実装した．図 3 のように，CycleGAN を基に構成して，シーン画像を判定する Discriminator に変更を施している．

角膜表面反射画像のドメインをドメイン X，シーン画像のドメインをドメイン Y として扱う．Generator $G_{X\mapsto Y}$,$G_{Y\mapsto X}$，Discriminator D_X,D_Y はそれぞれ以下の損失関数 $L_{G_{X\mapsto Y}}$，$L_{G_{Y\mapsto X}}$，L_{D_X}，L_{D_Y} を最小化するように学習させる．

$$L_{G_{X\mapsto Y}} = L_{\mathrm{GAN}}(G_{X\mapsto Y}, D_Y) + L_{\mathrm{cyc}}(G_{X\mapsto Y}, G_{Y\mapsto X}) + L_{\mathrm{idt}}(G_{X\mapsto Y}) + L_{\mathrm{class}}(G_{X\mapsto Y}, D_Y) \quad (1)$$

$$L_{G_{Y\mapsto X}} = L_{\mathrm{GAN}}(G_{Y\mapsto X}, D_X) + L_{\mathrm{cyc}}(G_{X\mapsto Y}, G_{Y\to X}) + L_{\mathrm{idt}}(G_{Y\mapsto X}) \quad (2)$$

$$L_{D_X} = -L_{\mathrm{GAN}}(G_{Y\mapsto X}, D_X) \quad (3)$$

$$L_{D_Y} = -L_{\mathrm{GAN}}(G_{X\mapsto Y}, D_Y) + L_{\mathrm{class}}(G_{X\mapsto Y}, D_Y) \quad (4)$$

$$L_{\mathrm{class}}(G_{X\mapsto Y}, D_Y) = E_{x\sim p_{data}(x)}[L_{\mathrm{CE}}(V_{\mathrm{onehot}}(x), V_{D_Y}(G_{X\mapsto Y}(x)))] + E_{y\sim p_{data}(y)}[L_{\mathrm{CE}}(V_{\mathrm{onehot}}(y), V_{D_Y}(y))] \quad (5)$$

$$L_{\mathrm{CE}}(t,p) = -\sum_{i=0} t_i \log p_i \quad (6)$$

$V_{D_Y}(y)$ は画像 y を Discriminator D_Y に入力して得られるベクトルを表す．

$V_{\mathrm{onehot}}(x)$ は画像 x のクラスを表す one-hot vector を表す．

3.2.4　CycleACVLADGAN

CycleACGAN のシーン画像の Discriminator の出力ベクトルを，VLAD 特徴量を主成分分析により次元圧縮したものと同じ次元数である 4096 次元のベクトルに変更したものである．

小節 3.2.3 で述べた CycleACGAN を改良することで構成した．図 4 のように，CycleACGAN を基に構成して，シーン画像の真偽を判定する Discriminator に変更を施している．

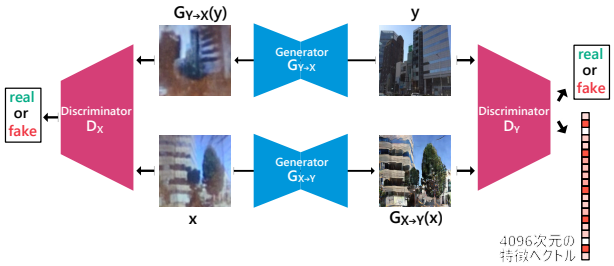

図 4　CycleACVLADGAN のネットワーク構造

角膜表面反射画像のドメインをドメイン X，シーン画像のドメインをドメイン Y として扱う．Generator

[*1] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

$G_{X\mapsto Y}$,$G_{Y\mapsto X}$，Discriminator D_X,D_Y はそれぞれ以下の損失関数 $L_{G_{X\mapsto Y}}$，$L_{G_{Y\mapsto X}}$，L_{D_X}，L_{D_Y} を最小化するように学習させる．

$$
\begin{aligned}
L_{G_{X\mapsto Y}} &= L_{\mathrm{GAN}}(G_{X\mapsto Y}, D_Y) + L_{\mathrm{cyc}}(G_{X\mapsto Y}, G_{Y\mapsto X}) \\
&\quad + L_{\mathrm{idt}}(G_{X\mapsto Y}) + L_{\mathrm{simi}}(G_{X\mapsto Y}, D_Y) \\
L_{G_{Y\mapsto X}} &= L_{\mathrm{GAN}}(G_{Y\to X}, D_X) + L_{\mathrm{cyc}}(G_{X\mapsto Y}, G_{Y\mapsto X}) \\
&\quad + L_{\mathrm{idt}}(G_{Y\mapsto X}) \\
L_{D_X} &= -L_{\mathrm{GAN}}(G_{Y\mapsto X}, D_X) \\
L_{D_Y} &= -L_{\mathrm{GAN}}(G_{X\mapsto Y}, D_Y) + L_{\mathrm{simi}}(G_{X\mapsto Y}, D_Y) \\
L_{\mathrm{simi}}&(G_{X\mapsto Y}, D_Y) \\
&= E_{x\sim p_{data}(x)}[1 - V_{D_Y}(G_{X\mapsto Y}(x)) \cdot V_{\mathrm{VLAD}}(y_{X\mapsto Y}(x))] \\
&\quad + E_{y\sim p_{data}(y)}[1 - V_{D_Y}(y) \cdot V_{\mathrm{VLAD}}(y)]
\end{aligned}
$$

$V_{\mathrm{VLAD}}(y)$ は画像 y の VLAD 特徴量を次元圧縮した 4096 次元のベクトルを表す．$y_{X\mapsto Y}(x)$ は角膜表面反射画像 x のクラスに対応するシーン画像 y を表す．$V_{D_Y}(y)$ は画像 y を Discriminator D_Y に入力して得られる 4096 次元のベクトルを表す．

3.3 GAN の Discriminator によるシーン識別

GAN の Discriminator の出力ベクトルを用いて類似度を求める．出力ベクトルのドット積をとることにより，cos 類似度を求め，これを画像間の類似度として用いる．

利用する GAN は，小節 3.2.3，3.2.4 で述べた CycleACGAN，CycleACVLADGAN である．CycleACGAN の出力ベクトルは正規化されていないので，ドット積を求める前に正規化を行っている．

4. データセット

4.1 シーン画像データセット

4.1.1 24/7 Tokyo dataset

Torii ら [13] によって作成された 1125 枚の画像からなるデータセットである．撮影には Apple 社の iPhone 5S と Sony 社の Xperia のスマートフォンを用いている．図 5 にその一例を示す．

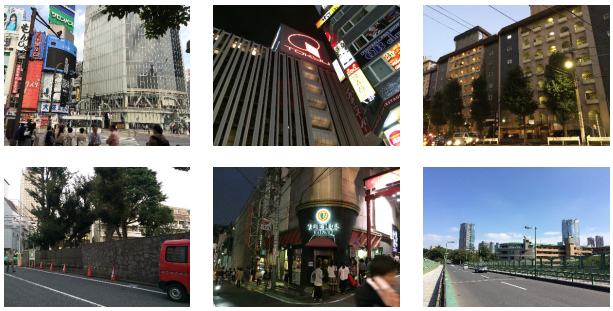

図 5　24/7 Tokyo dataset のシーン画像の一例

4.1.2 京都大学周辺のシーン画像データセット

京都大学周辺の 25 シーンにおいて 1 枚ずつ撮影された 25 枚の既存のデータセットである．画像は RICOH 社の RICOH THETA を用いて撮影されたものである．図 6 にその一例を示す．

図 6　京都大学周辺で撮影したシーン画像の一例

4.2 角膜表面反射画像データセット

4.2.1 24/7 Tokyo dataset を用いて撮影した角膜表面反射画像データセット

図 7 のように八面ディスプレイ内にカメラとチンレスト（顎乗せ台）を配置した環境で撮影を行った．被撮影者の頭部をチンレストによって固定し，ディスプレイに表示したパターン画像の中心点に視線を向けてもらうことで目の位置を固定する．この状態で 2 秒間隔の画像の自動切り替えとインターバル撮影によって自動で撮影を行った。画像の表示には八面ディスプレイの前方 3 面を用いた．表示する画像は 5 点のパターン画像とシーン画像であり，これを交互に表示する．シーン画像には 24/7 Tokyo dataset のうち無作為に選んだ 100 枚を用いた．パターン画像は次のシーン画像の位置合わせに用いる．撮影は RAW 撮影で行った．

上記の自動撮影によって得た画像セット（シーン画像を表示した時の撮影画像とその直前のパターン画像表示した時の撮影画像のセット）に，図 8 のように以下の処理を施した．

(1) RAW 現像により明るさを調整と目の周辺の切り抜きを行う．
(2) シーン画像を表示した時の撮影画像をパターン画像表示した時の撮影画像のパターンの周囲 4 点の座標をもとに切り抜きと平面展開を行う．
(3) シーン画像と合わせるため，左右反転させる．

以上の処理により得た，1053 枚の角膜表面反射画像のデータセットである．図 9 にその一例を示す．

4.2.2 京都大学周辺シーンを用いて撮影した角膜表面反射画像データセット

京都大学周辺の 25 シーンを用いて、小節 4.2.1 で述べた手法により得た，99 枚の角膜表面反射画像のデータセットである．図 10 にその一例を示す．

図 7　撮影の様子

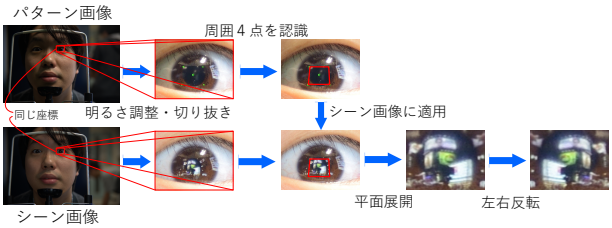

図 8　撮影画像から角膜表面反射画像を得る処理

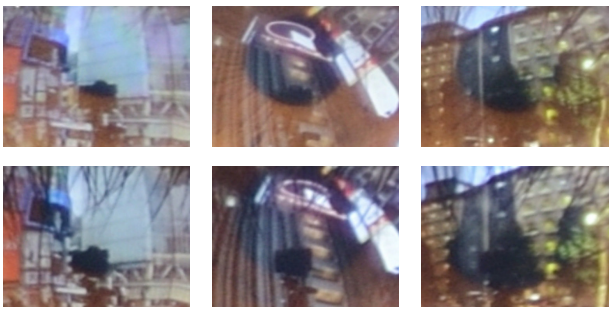

図 9　24/7 Tokyo dataset を用いて撮影した角膜表面反射画像の一例

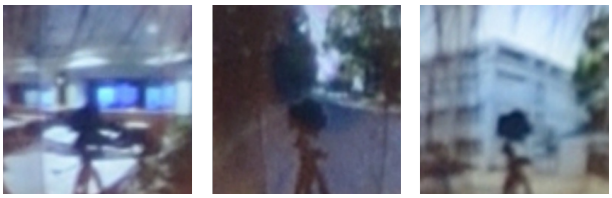

図 10　京都大学周辺シーンを用いて撮影した角膜表面反射画像の一例

4.2.3　京都大学周辺シーンの角膜表面反射画像データセット

京都大学周辺の 25 シーンで撮影された 62 枚の既存のデータセットである．画像はアイカメラを用いて撮影されたものである．図 11 にその一例を示す．

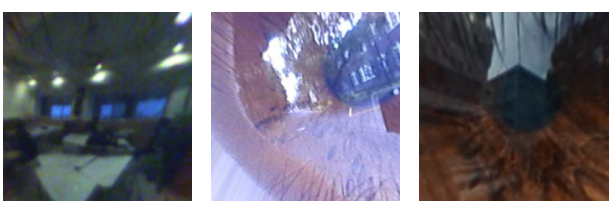

図 11　京都大学周辺で撮影した角膜表面反射画像の一例

5.　実験

本研究において提案した，1）GAN により角膜表面反射画像からシーン画像のドメインに変換したものから抽出した VLAD 特徴量を利用したシーン識別を行う手法，2）GAN に入力した角膜表面反射画像に対する GAN の Discriminator の出力ベクトルによりシーン識別を行う手法について，従来手法である角膜表面反射画像から直接抽出した VLAD 特徴量を利用してシーン識別を行う手法と比較，精度検証を行った．

5.1　実験設定

学習･評価に用いる画像は，RGB 画像でサイズは 256×256 に resize している．GAN の generator の出力画像は，RGB 画像でサイズは 256 × 256 である．シーン識別で用いた VLAD 特徴量は，Torii ら [3] による，24/7 Tokyo dataset から抽出した約 2500 万の Dense SIFT 特徴量もとに学習を行った VLAD 特徴空間と主成分分析を用いて得た，4096 次元の VLAD 特徴量である．

CycleACGAN のクラス数・出力ベクトルの次元は 100 とする．クラスは，以下で述べる学習データである撮影に用いた 100 シーンをそれぞれ表す．

学習データ

24/7 Tokyo ディスプレイデータセット（シーン画像：24/7 Tokyo dataset の 1125 枚，角膜表面反射画像：24/7 Tokyo dataset を用いて撮影した 1053 枚）

評価データ

評価に用いるデータセットとして以下の 3 つを用いた．

(1) 24/7 Tokyo ディスプレイデータセット

(2) 京都大学ディスプレイデータセット（シーン画像：京都大学周辺の 25 枚＋ 24/7 Tokyo dataset の 1125 枚の 1150 枚，角膜表面反射画像：京都大学周辺シーンを用いて撮影した 99 枚）

(3) 京都大学データセット（シーン画像：京都大学周辺の 25 枚＋ 24/7 Tokyo dataset の 1125 枚の 1150 枚，角膜表面反射画像：京都大学周辺で撮影された 62 枚）．

5.2　評価方法

各角膜表面反射画像に対して類似度の高い順にシーン画像を並べ，上位 k 件の中に正解画像が含まれている割合（これを $Acc(k)$ とする）により評価を行う．

5.3　結果と考察

5.3.1　24/7 Tokyo ディスプレイデータセット

24/7 Tokyo ディスプレイデータセットに対する各手法の $Acc(k)$ を表 1 にまとめる．また，図 12 に GAN の生成画像を示す．

	従来手法	画像のドメイン変換と VLAD				discriminator の出力	
k 件	eye	pix2pix	CGAN	CAGAN	CAVGAN	CAGAN	CAVGAN
1	0.8481	0.8386	0.7246	0.8196	0.8262	0.0190	**0.9934**
2	0.9136	0.9003	0.7949	0.8803	0.8822	0.0199	**0.9991**
3	0.9288	0.9117	0.8167	0.9050	0.8993	0.0199	**0.9991**
4	0.9345	0.9231	0.8310	0.9202	0.9050	0.0199	**1.0000**
5	0.9383	0.9316	0.8433	0.9259	0.9126	0.0199	**1.0000**

表 1 24/7 Tokyo ディスプレイデータセットに対する $Acc(k)$

表中の CGAN，CAGAN，CAVGAN はそれぞれ，CycleGAN，CycleACGAN，CycleACVLADGAN を表す．以下の図・表でも用いる．

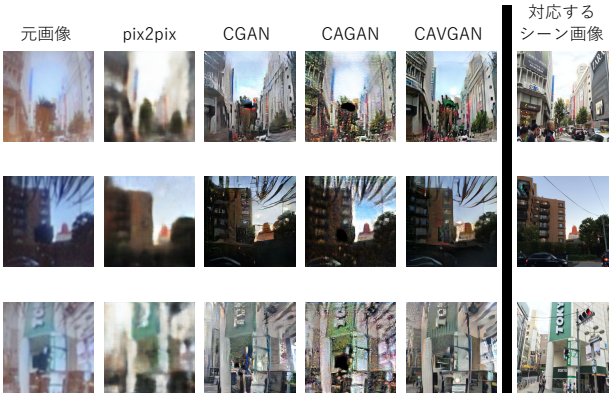

図 12 24/7 Tokyo ディスプレイデータセットに対する GAN の生成画像

VLAD 特徴量を用いる手法は，どの手法も高い精度となっている．これは，24/7 Tokyo dataset を VLAD の学習に用いたことが理由だと考えられる．pix2pix，CycleACGAN，CycleACVLADGAN が，CycleGAN と比較して識別精度が高くなっている．これは 24/7 Tokyo ディスプレイデータセットは GAN の学習に用いたものであり，これら 3 つの GAN は教師データとして対応する画像またはクラスを与えていることが理由だと考えられる．

discriminator の出力を用いる手法において，CycleACGAN による識別精度が非常に低くなっている．これは学習に用いるシーンのクラスを表す one-hot vector を特徴量として用いると，特徴量ベクトルは学習データに非常に強く依存し，汎化性がほぼなくなっていると考えられる．CycleACVLADGAN による識別精度がとても高くなっている．これも教師付きデータを用いていることが理由だと考えられる．

pix2pix による生成画像は，まつげや虹彩のパターンをほぼ取り除けているがぼやけており，これは角膜反射画像とシーン画像の位置合わせを完全には行えていないので，学習時に L1loss の影響が大きすぎたことが原因であると考えられる．CycleGAN，CycleACGAN，CycleACVLADGAN による生成画像は，色彩情報がシーンに近づいているが，強いまつげの影響は取り除くことができず，この影響で不自然な着色がなされているものがある．

VLAD 特徴量を用いる手法において，CycleGAN などによる変換で画像の見た目の品質向上が行われたにもかかわらず，元の角膜表面反射画を用いた手法による識別精度が最も高い．この原因として，24/7 Tokyo dataset を VLAD の学習に用いているので，虹彩の影響による多少の色味の違いは識別にはあまり影響せず，GAN による全体的な色彩情報向上よりも少しの不自然な着色が識別精度に影響したということが考えられる．

5.3.2 京都大学ディスプレイデータセット

京都大学ディスプレイデータセットに対する各手法の $Acc(k)$ を表 2 にまとめる．また，図 13 に GAN の生成画像を示す．

	従来手法	画像のドメイン変換と VLAD				discriminator の出力	
k 件	eye	pix2pix	CGAN	CAGAN	CAVGAN	CAGAN	CAVGAN
1	0.8889	0.6566	0.8182	0.8586	**0.9495**	0.0101	0.7778
2	0.9293	0.7273	0.8586	0.9394	**0.9495**	0.0202	0.8788
3	0.9495	0.7475	0.8687	**0.9596**	0.9495	0.0202	0.8990
4	0.9495	0.7576	0.8687	**0.9697**	0.9495	0.0202	0.9293
5	0.9596	0.7677	0.8788	**0.9697**	0.9596	0.0404	0.9293

表 2 京都大学ディスプレイデータセットに対する $Acc(k)$

図 13 京都大学ディスプレイデータセットに対する GAN の生成画像

$Acc(1)$ の値を見ると CycleACVLADGAN と VLAD 特徴量を用いる手法が最も識別精度が高かった．24/7 Tokyo ディスプレイデータセットの場合と比べて識別精度が少し向上している理由として，京都大学周辺のシーン画像データセットはそれぞれのシーンの差が大きく判別しやすいことが考えられる．また，この評価データは VLAD の学習に用いた 24/7 Tokyo dataset と異なるので，GAN による色彩情報向上による影響が出ていると考えられる．pix2pix の識別精度が低いのは，学習データの影響を受け不自然な着色がなされていることが原因だと考えられる．また，CycleGAN は学習データの影響を受け少し不自然な着色がなされ識別精度が低いが，CycleACGAN，CycleACVLADGAN は学習データの影響をあまり受けずに着色しており識別精度が向上したと考えられる．

discriminator の出力を用いる手法において，CycleACVLADGAN による識別精度が下がっているが，これは CycleACVLADGAN の discriminator の出力する特徴ベクトルは学習データの影響を受けていると考えられる．

5.3.3 京都大学データセット

京都大学データセットに対する各手法の $Acc(k)$ を表 3 にまとめる．また，図 14 に GAN の生成画像を示す．

	従来手法	画像のドメイン変換と VLAD				discriminator の出力	
k 件	eye	pix2pix	CGAN	CAGAN	CAVGAN	CAGAN	CAVGAN
1	**0.2903**	0.0161	0.2581	0.2097	0.2581	0.0000	0.0968
2	**0.3548**	0.0161	0.2742	0.2581	0.2742	0.0000	0.1290
3	**0.3548**	0.0161	0.2903	0.2742	0.2903	0.0000	0.1452
4	**0.3548**	0.0323	0.3226	0.2903	0.3387	0.0000	0.1613
5	0.3548	0.0323	0.3226	0.3226	**0.3710**	0.0000	0.1613

表 3 京都大学データセットに対する $Acc(k)$

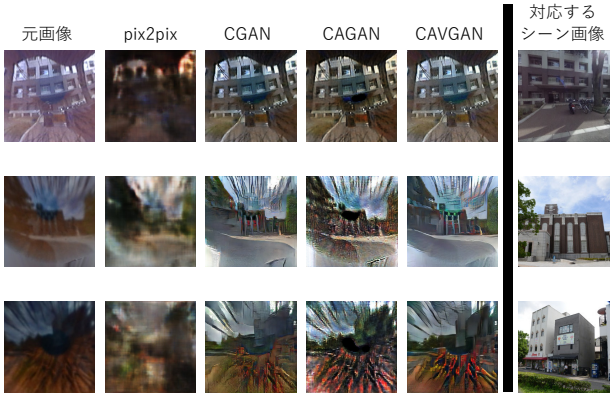

図 14 京都大学データセットに対する GAN の生成画像

従来手法が最も識別精度が高かった．VLAD 特徴量を用いる手法において，pix2pix による識別精度が非常に低いことから，pix2pix は学習データの影響を多大に受けていると考えられる．CycleGAN，CycleACGAN，CycleACVLADGAN による識別精度は pix2pix と比較すると高いので，これら 3 つの GAN の画像変換は pix2pix よりは汎化性能が高いと考えられる．

pix2pix による生成画像は，まつげや虹彩のパターンを多少取り除けているが，ぼやけていおり，不自然な着色がなされている．この不自然な着色は学習データの影響によるものと考えられる．CycleGAN，CycleACGAN，CycleACVLADGAN による生成画像は，色彩情報がシーンに多少近づいているが，まつげや虹彩のパターンによる影響はほとんど取り除けず、これらの影響で不自然な着色がなされている．

GAN の画像の生成結果が良くない原因として，GAN の学習に用いた 24/7 Tokyo ディスプレイデータセットの角膜表面反射画像は室内で撮影したが，京都大学データセットは室外で撮影されたもので，赤外線の影響で虹彩のパターンが強く出ていることや，まつ毛が多く含まれていることが考えられる．

6. 結論と今後の展望

本研究では，角膜表面反射画像からのシーン識別精度向上を目的として，GAN を用いて角膜表面反射画像をシーン画像のドメインに変換したものから抽出した VLAD 特徴量を利用するシーン識別を行う手法，及び，GAN に入力した角膜表面反射画像に対する GAN の Discriminator の出力ベクトルを特徴量としてシーン識別を行う手法を提案した．実験では従来手法である角膜表面反射画像から直接抽出した VLAD 特徴量を用いてシーン識別と，提案手法の精度の比較を行った．その結果，京都大学ディスプレイデータセットに対しては，CycleACVLADGAN による画像変換により従来手法より良い結果が得られた．しかし，京都大学データセットにおいては，画像の見た目の品質向上が行われたにもかかわらず，GAN を用いた手法は直接角膜表面反射を使用する方法よりも高い精度は得られなかった．その理由として，以下の問題点が考えられる．

(1) pix2pix は完全には位置合わせができていないデータを用いて学習した場合，かなりぼやけた画像を生成する．
(2) CycleGAN などによる画像変換では，まつ毛が取り除けず，その影響で不自然な着色を施すことがある．
(3) 角膜表面反射画像に虹彩パターンがあまりでていない学習データを用いて GAN を学習させたため，強い虹彩パターンをほとんど取り除くことができなかった．

これらの問題点の解決法として以下のことが考えられる．

問題点 1 の解決法として，位置合わせをより正確に行った学習データを用意することが考えられる．これには，2 つの方法が考えられる．1 つ目は，本稿で述べた八面ディスプレイによる撮影においてインターバル間隔を短くするという方法である．しかし，シャッタースピードを考慮すると，カメラのインターバル間隔とディスプレイの切り替えのタイミングの制御がかなり困難になり，撮影に失敗するリスクが増える．2 つ目は，位置合わせの手法で既存のデータセットを厳密に位置合わせする方法である．どちらの手法を用いる場合でも，pix2pix は本研究における学習データ数ではその影響を強く受けるので学習データをさらに多く集める必要がある．

問題点 2 の解決法として，2 つの方法が考えられる．1 つ目は，まつ毛の影響を受けている学習データを増やすという方法である．しかし，本研究における生成画像を見ると，CycleGAN は着色はできるが，形状を完全に変更するほどの着色はできず，この解決法では，まつ毛の影響を本研究による生成画像よりは取り除くことができるが，完全には取り除けないと考えられる．2 つ目は，まつ毛の位置を学習データに加え，この情報を基に変換を学習できるネット

ワークを用いるという方法である．しかし，人の手で行う場合，まつ毛位置のアノテーションには高いコストがかかるという問題がある．また，まつ毛の位置の取得にネットワークを利用する場合，どの程度の精度でまつ毛の位置を取得できれば，GAN の学習に利用できるかを検証する必要がある．

問題点 3 の解決法として，太陽光の赤外線により虹彩パターンがはっきりと出ている撮影データを，学習データに追加することが考えられる. しかし，その場合，室外での撮影となり，本稿で述べた八面ディスプレイによる撮影のように大量のデータを低コストで集めることが困難であると考えられる．

謝辞　本研究は科研費 17H01779, 26249029, 15H02738, および, JST CREST, JPMJCR17A5 の支援を受けている．

参考文献

[1] Nistér, D., Naroditsky, O. and Bergen, J.: Visual odometry, *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, Vol. 1, Ieee, pp. I–I (2004).

[2] 江川佳輝，小川太士，中澤篤志ほか：深層学習を用いた自己撮影画像の撮影場所検索，研究報告コンピュータビジョンとイメージメディア (CVIM)， Vol. 2018, No. 56, pp. 1–5 (2018).

[3] Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T. and Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307 (2016).

[4] Jégou, H., Douze, M., Schmid, C. and Pérez, P.: Aggregating local descriptors into a compact image representation, *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, IEEE, pp. 3304–3311 (2010).

[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Advances in neural information processing systems*, pp. 2672–2680 (2014).

[6] Mirza, M. and Osindero, S.: Conditional generative adversarial nets, *arXiv preprint arXiv:1411.1784* (2014).

[7] Isola, P., Zhu, J.-Y., Zhou, T. and Efros, A. A.: Image-to-image translation with conditional adversarial networks, *arXiv preprint* (2017).

[8] Zhu, J.-Y., Park, T., Isola, P. and Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, *arXiv preprint* (2017).

[9] Odena, A., Olah, C. and Shlens, J.: Conditional image synthesis with auxiliary classifier gans, *arXiv preprint arXiv:1610.09585* (2016).

[10] Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical image computing and computer-assisted intervention*, Springer, pp. 234–241 (2015).

[11] Li, C. and Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks, *European Conference on Computer Vision*, Springer, pp. 702–716 (2016).

[12] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016).

[13] Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M. and Pajdla, T.: 24/7 place recognition by view synthesis, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1808–1817 (2015).