Vol.2020-CG-180 No.6
Vol.2020-DCC-26 No.6
Vol.2020-CVIM-223 No.6
2020/11/5

CG学習・実画像推論のための画像スタイル変換とランドマーク検出の同時学習

大友 賢太郎[1,a)]　浮田 宗伯[1,b)]

概要：近年，交通事故削減の対策の一つとして自動運転技術が注目されているが，自動運転レベル 3 以下では，ドライバへ運転の主導権が委譲される場合がある．よって，交通事故の削減と自動運転時の安全な権限移譲の為に，ドライバの覚醒度合いをシステムが認識する必要がある．ドライバの覚醒度合いの検出指標には，視線など眼球運動が利用されており，本研究では眼球運動の認識精度を向上させる為に，目のランドマークの座標位置を検出するネットワークを利用する．但し，学習データとして実画像を利用する場合，アノテーション済の画像が大量に必要となるが，アノテーションは非常にコストのかかる作業である．また生成が簡単な CG 画像を学習データとして用いた場合，実画像を入力しても高精度に検出できない事が考えられる．そこで，GAN を利用し，大量の CG 画像と少数の実画像を学習データとして，CG 画像を実画像に近づくよう変換する生成器の学習を行う．また，生成画像と CG 画像のアノテーション情報を，ランドマーク座標検出ネットワークの学習データとして利用し，ランドマーク座標検出の精度の変化を検証する．本手法では，生成器と識別器で構成される画像変換ネットワークへランドマーク座標検出器を追加し，アノテーション情報と生成画像のランドマーク座標を推定した結果との間で算出された誤差を生成器へ逆伝播する事によって，変換された画像が CG 画像のアノテーション情報を保持する為の制約を強め，変換精度を向上させると共に，end-to-end で学習出来るネットワークを作成した．

キーワード：GAN，ランドマーク座標検出ネットワーク，スタイル変換，同時学習

1. はじめに

近年，自動車の安全装置の普及により, 交通事故件数は年々減少傾向にあるものの, 交通死亡事故発生件数は, 他の法令違反による死亡事故件数と比較して，漫然運転や脇見運転など安全運転義務違反によって引き起こされる事故が，高い割合で占められている．こうした需要から，交通事故削減の対策の一つとして自動運転技術が注目されているが，自動運転レベル 3 以下では，システムからドライバへ運転の主導権が移譲される場合がある為，交通事故の削減と自動運転時の安全な権限移譲の為には，ドライバの覚醒度合いをシステムが認識する必要があり，ドライバの状態を認識する為の装置として，ドライバの顔をカメラで撮影するドライバーモニタリングシステム (DMS) が開発されている．しかし，顔向きや瞬目などは取得できるが，視線の移動など眼球運動を高精度に取得する事は難しいといった問題がある．そこで，眼球運動を高精度に取得する為に，視線の検出精度が高くなることが求められている．

視線検出は数多くの研究がされており [1], [2], [3]，多くの従来法は，虹彩や瞼などの領域のランドマーク座標の検出精度に依存しており，検出精度を高める為に，深層学習モデルを活用している．しかし，ランドマーク座標検出ネットワークの学習には，学習データとして，大量の実際に撮影された画像と画像に対応したアノテーションデータが必要となるが，撮影画像にランドマーク座標の位置などをアノテーションする事は，非常にコストのかかる作業である．そこで，Seonwook ら [4] は，学習データの生成コストを下げる為に，容易かつ短時間で大量にデータを生成可能なアノテーション済み CG 画像を学習データとして利用し，畳み込みニューラルネットワーク（CNN）を学習することによって，実画像のランドマーク座標を推定した．但し，実画像と CG 画像間には，画像内の目の大きさや照明環境など，ドメイン差が存在し，CG 画像で学習したネットワークで実画像のランドマーク座標を推定しても，精度が上がらないといった問題がある．また，この手法では一般的な RGB 画像を入力としている為，CG 画像と実画像のドメイン差は小さいが，図 1 に示すように，提案手法で

[1] 豊田工業大学
Toyota Technological Institute
[a)] sd19409@toyota-ti.ac.jp
[b)] ukita@toyota-ti.ac.jp

Vol.2020-CG-180 No.6
Vol.2020-DCC-26 No.6
Vol.2020-CVIM-223 No.6
2020/11/5

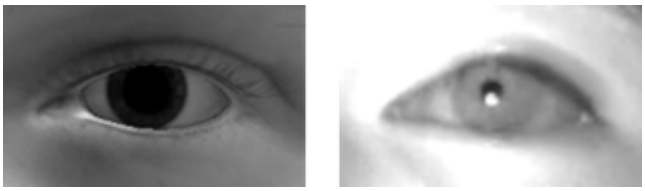

図 1 左の CG 画像と比較して，右の実画像は画像全体の輝度値が高く，ドメイン差が大きい為，ランドマーク座標の検出精度が低下する．

用いた実画像はドメイン差の大きい画像を用いている．このようなドメイン差の大きい画像を入力した場合，検出精度が更に低下してしまう事が考えられる．

実画像を CG 画像に近づけるための手法として，Ashish ら [5] は，CG 画像と少数の実画像を学習データとして敵対的生成ネットワーク（Generative Adversarial Network: GAN）[6] を学習することによって，CG 画像を実画像に近づくように変換する SimGAN を提案した．しかし，前述したように，CG 画像と本論文で使用している実画像のドメイン差は大きい為，CG 画像のアノテーションデータと変換した画像が一致せず，ランドマーク座標検出ネットワークの学習データとして使用する事は難しい．

そこで本研究では，画像生成ネットワークである SimGAN の生成精度を高め，生成画像を用いてランドマーク座標検出ネットワークを学習し，ランドマーク座標検出の精度の変化を検証する．本論文の新規性は以下である．

(1) 画像生成ネットワークにランドマーク座標検出ネットワークを追加し，画像生成ネットワークとランドマーク座標検出ネットワークを end-to-end で学習できるネットワークを作成する．
(2) 生成画像のヒートマップとアノテーションデータから作成した CG 画像のヒートマップ間で算出された誤差を生成器に逆伝播することによって，生成精度を向上させる．

2. 関連研究

本章ではまず，GAN について説明をし，GAN を用いた CG 画像と実画像の変換手法について述べる．次に目のランドマーク座標検出ネットワークに関する手法について述べる．

2.1 敵対的生成ネットワーク (GAN)

GAN の構造を図 1 に示す．GAN は，生成器（Generator）と識別器（Discriminator）の 2 つのネットワークで構成された生成ネットワークである．生成器は，データの特徴を示す潜在変数と呼ばれる値や，ノイズを入力とし，新しいデータを生成するモデルである．識別器は，生成器が生成したデータを生成されたデータだと識別できるように学習をしていく．[6] は，この様なモデルを学習する時の損失関

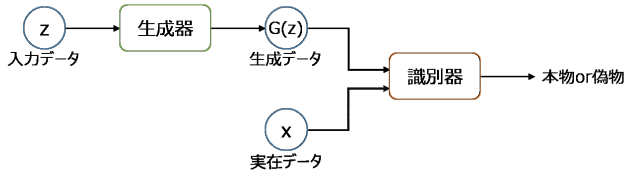

図 2 GAN の概念図．生成器の生成したデータを偽物，実在データを本物と識別するように識別器を学習する．生成器は識別器が本物だと識別するようなデータを生成するように学習する．

数を次の様に定義している．

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[logD(x)] + E_{z\sim p_z(z)}[log(1-D(G(z)))]$$

生成器に潜在変数又はノイズ z が入力され，$G(z)$ が出力される．識別器には，実在データ x と，生成器の出力データ $G(z)$ がそれぞれ入力され，識別器に実在データが入力された時のスコア $D(x)(0 \leq D(x) \leq 1)$ と，生成器の出力データが入力された際のスコア $D(G(z))(0 \leq D(G(z)) \leq 1)$ が出力される．識別器の学習時には，$D(x)$ が大きくなり，$D(G(z))$ が小さくなるようにする．すなわち，識別器が実在データは本物，生成データは偽物だと識別できるように学習をしていく．一方，生成器の学習時には，$D(G(z))$ が大きくなるようにする．つまり，生成器は生成したデータを，識別器が本物と識別するように学習をしていく．この様に，生成器と識別器を交互に学習していくことで，本物に近いデータを生成することができるようになる．GAN は主に画像の生成に用いられ，それ以外に画像の変換 [7] や超解像 [8] などに用いられる．

2.2 GAN を用いた画像変換手法

2.2.1 CG 画像の変換

GAN を用いた CG 画像変換手法は，Ashish ら [5] が SimGAN を提案した．大量の CG 画像と少数の実画像を学習データとして利用し，生成器には CG 画像が入力され，実画像に近づくように変換した画像を生成する．また，識別器には生成された画像と実画像が入力され，入力された画像が生成画像か実画像かを識別する．
識別器の損失関数を以下に示す．

$$L_D(\phi) = -\sum_i log(D_\phi(\tilde{\mathbf{x}}_i)) - \sum_j log(1-D_\phi(\mathbf{y}_j)) \quad (1)$$

$\tilde{\mathbf{x}}_i$ は生成された画像，$\mathbf{y}_j$ は実画像であり，識別器は生成された画像は偽物，実画像は本物と識別できるように学習をしていく．識別器の学習を行う際に，ある特徴量に大きく依存してしまった場合，生成に失敗してしまう可能性がある．その為，生成画像が実画像に近い場合，両画像の各局所領域は区別がつかないと考え，SimGAN では識別器への入力画像をパッチ分割し，分割された領域ごとに識別

Vol.2020-CG-180 No.6
Vol.2020-DCC-26 No.6
Vol.2020-CVIM-223 No.6
2020/11/5

を行っている．

次に生成器の損失関数を以下に示す．

$$L_G(\theta) = -\sum_i log(1 - D_\phi(R_\theta(x_i))) + \lambda \|\psi(R_\theta(\mathbf{x}_i)) - \psi(\mathbf{x}_i)\|_1 \quad (2)$$

$\mathbf{x}_i$ は CG 画像，λ は第二項に対する重み，ψ は特徴空間への写像である．第一項は，識別器が本物の画像と識別してしまうように学習を行うための損失である．第二項は自己正則化項であり，目の画像の変換を行う際に実画像に変換するだけではなく，CG 画像と生成画像の間で，瞼や虹彩の位置などを保持できる様，CG 画像と生成画像の間で各ピクセルごとの L1 損失を算出する．これらから，生成器は CG 画像を実画像に近づけた画像を生成する．本手法の利点として，容易に生成可能な CG 画像とアノテーションデータを学習に用いる事によって，学習データの生成コストを下げる事が出来る．しかし，CG 画像と実画像の間には，瞼や虹彩の位置に関してドメイン差が存在する為，自己正則化項で制約を与えても，CG 画像のアノテーションデータと生成画像が完全に一致しないといった問題がある．

2.2.2 実画像の変換

近年，Conditional GAN[9] を利用した，生成器に実画像と視線方向を入力する事で，入力した視線方向に変更された実画像を生成するネットワーク [10] が提案されている．Zhe ら [10] の提案した手法は，実画像と変更したい視線方向を入力し，視線方向が変更された画像を生成する．識別器は，実画像か生成された画像かを識別するのみでなく視線角を検出し，生成器に入力した視線方向との差分から，視線角の損失を算出する．また，生成画像と入力した実画像へ戻すような視線方向を再度生成器に入力する事によって, 逆変換した画像と入力画像の間で再構成誤差を算出する．この手法は，実画像の視線方向を任意の方向へ変更した画像が生成出来るので，CG 画像を変換する手法と比較して，ドメイン差分による影響を小さくする事が出来る利点がある．但し，学習する為には，実画像及び実画像のアノテーションデータが必要となる為，学習データの生成コストが高くなるといった問題がある．

2.3 ランドマーク座標検出ネットワーク

従来の視線検出手法の多くは，虹彩や瞼などの目の領域のランドマーク座標の検出精度に依存しており，視線の検出精度を高めるには，ランドマーク座標の検出精度を高める必要がある．Seonwook ら [4] は，ランドマーク座標の検出精度を高める為に，局所的特徴と大域的特徴の両方を同時に認識可能な構造を持つ Hourglass Network[11] を利用しており，入力された目の画像から，瞼と虹彩，虹彩中心のヒートマップを推定し，推定されたヒートマップからランドマーク座標を検出する手法である．この手法では，学習データの生成コストを下げる為に，学習画像に CG 画像を用いているが，ドメイン差の大きい実画像を入力した際に，検出精度が下がってしまう問題がある．これらの事から，本研究では学習データの生成コストの少ない SimGAN に対して，ドメイン差による影響を減少させることによって，生成精度を向上させ，ランドマーク座標検出ネットワークに適した学習データを生成し，ランドマーク座標の検出精度を向上させる事を目的とする．

提案手法との比較を表 1 に示す．

表 1 提案手法と 2.2 節の手法との比較.

	CG 画像の変換	実画像の変換	提案手法
アノテーションコスト	小	大	小
ドメイン差の影響	大	小	小

3. 提案手法

本研究では，SimGAN にランドマーク座標検出ネットワークを追加することで，画像生成ネットワークとランドマーク座標検出ネットワークの同時学習を行い，その上で画像生成精度を向上させる手法を提案する．提案する同時学習ネットワークの概略図を図 3 に示す．

3.1 同時学習ネットワーク

Li ら [12] の提案した TripleGAN は，生成器と識別器で構成される GAN に分類器を追加し，分類器で生成データのラベルを予測する事によって，生成データの意味的情報を考慮したデータ生成を可能にした．本手法では，ランドマーク座標検出ネットワークを追加し，生成器と識別器で構成された画像生成ネットワークとの同時学習を行うことで，生成器の画像生成精度を向上させると同時に，ランドマーク座標推定ネットワークの推定精度を向上させる．

3.1.1 画像生成ネットワーク

学習に使用する CG 画像は，UnityEyes[13] と呼ばれる目領域周辺の CG 画像と，瞼や虹彩などのアノテーションデータを同時に生成可能なツールを用いて作成した．生成されたアノテーションデータの虹彩と瞼の座標位置にガウシアンフィルタを適用することにより，ヒートマップ画像を作成し，生成器の入力データとして CG 画像とヒートマップ画像を利用する．

瞼や虹彩の位置を保持するための制約として，Ashish ら [5] の手法では，CG 画像と生成画像の間で各ピクセルごとの L1 損失を算出していた．しかし，図 1 に示す様に，今回使用している実画像は全体的に輝度値が高い画像となっており，CG 画像と比較して，虹彩や強膜，皮膚の部分で輝度値に大きな差がない．その為，図 4 の様に，L1 損失を

Vol.2020-CG-180 No.6
Vol.2020-DCC-26 No.6
Vol.2020-CVIM-223 No.6
2020/11/5

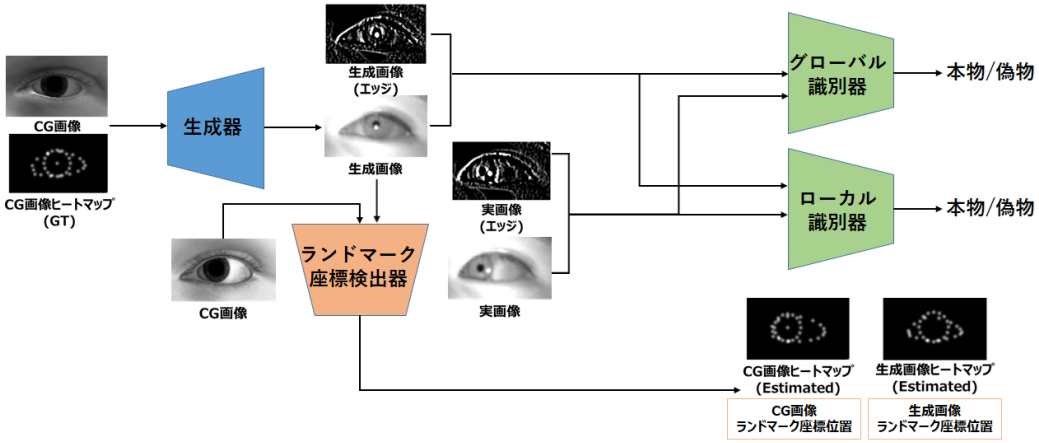

図 3 提案する同時学習ネットワークの概略図.(1)CG 画像とヒートマップの画像を生成器に入力し，実画像感を付与した画像を生成する．(2) 生成画像とエッジ画像，実画像とエッジ画像をグローバル識別器とローカル識別器に入力し，入力された画像が実画像なのか生成画像なのかを識別する．(3) 得られた生成画像と CG 画像をランドマーク検出器に入力して，ヒートマップ画像とランドマーク座標を推定する．

算出しても，視線方向の一致している (a) の L1 損失と比較して，一致していない (b) の L1 損失の方が小さくなってしまい，瞼や虹彩の位置を保持することが出来ない．そこで，本手法では、はじめにランドマーク座標検出ネットワークに CG 画像と生成画像を入力し，それぞれの推定されたヒートマップ画像を出力させる．次に，各ヒートマップ間で損失を算出し，生成器の学習に利用することによって，CG 画像と生成画像のランドマーク座標の位置を合わせるように学習を進めていく．

識別器は画像の大域的領域から識別するグローバル識別器と局所的領域から識別するローカル識別器の 2 種類を利用する．また，本手法では，従来手法 [6] と比較して，勾配消失問題やモード崩壊が起こりにくい WGAN[14] の手法で算出される損失を利用して学習を行った．実画像のクロップには OpenFace[15] と呼ばれる顔全体のランドマーク検出器を用いた．近赤外カメラで撮影された動画を OpenFace に入力し，取得した目のランドマーク座標位置から目の部分をクロップした画像を，実画像として学習に用いる．

本手法では，識別器の学習に、通常の画像だけではなく，エッジ画像も利用している．図 5 からわかるように，実画像と比較して CG 画像のエッジは濃く出ており，両エッジ画像の差分は大きい為，識別精度を向上させる事が可能となる．識別器は CG 画像と実画像，また両方のエッジ画像で事前学習されたモデルを利用する．

3.1.2 ランドマーク座標検出ネットワーク

ランドマーク座標検出ネットワークには，CG 画像と生成画像を入力し，瞼の領域及び虹彩の領域で各 16 点，虹彩中心の全 33 点を表すヒートマップを推定し，ヒートマップの最発火座標から検出したランドマーク座標位置を出力する．CG 画像と生成画像を入力データに利用し，学習を進めていくことによって，実画像に対しても瞼や虹彩のランドマーク座標を正確に検出する事が可能となる．ランドマーク座標検出器の構造には，[4] 同様，局所的特徴と大域的特徴の両方を同時に認識する事の出来る Hourglass Network[11] を利用する．Hourglass Network の構造を 6 に示す．局所的特徴とは，虹彩や瞳孔の中心など，画像中の細かな情報を捉える特徴であり，画像内に小さく映っている虹彩や瞳孔の中心の位置を検出する事が可能となる．

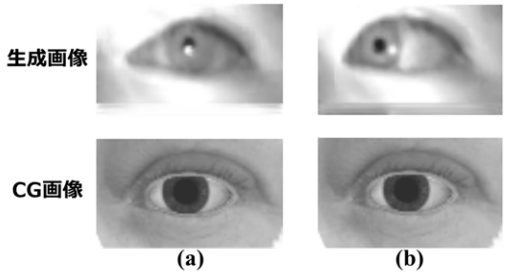

図 4 視線方向の一致している (a) の L1 損失は 0.2337，一致していない (b) の L1 損失は 0.2167 と一致していない場合の方が低い値となっている

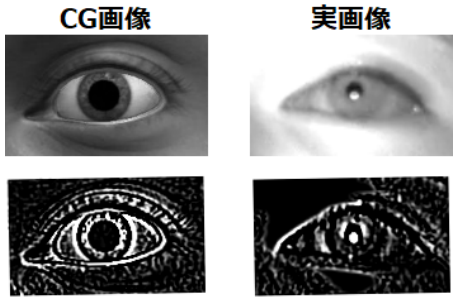

図 5 実画像のエッジ画像と比較して，CG 画像のエッジ画像は瞼や虹彩のエッジが濃く出ており，エッジ画像における差分は大きい

大域的特徴とは，広範囲な画像情報を捉える特徴であり，画像内に大きく写っている瞼や虹彩の位置関係を認識することが可能となる．局所的特徴と大域的特徴を考慮することによって，視線方向が変化している際に，虹彩の一部が隠れてしまう場合があるが，隠れている箇所を含めて虹彩を円として認識する事が出来，瞼や虹彩，虹彩中心などの位置関係を判断する事が可能となる．Hourglass Network は複数連結することで検出性能が向上する為，本手法では，3 個の Hourglass Network を連結して利用する．

ランドマーク座標検出ネットワークの学習は，最適化手法に Adam を用い，学習率は 0.0004 とした．損失は，CG 画像の推定されたヒートマップ画像と GT のヒートマップ画像間，推定されたランドマーク座標と GT のランドマーク座標間で平均二乗誤差 (MSE) を算出，また生成画像でも同様の損失を算出したものを利用する．また，ランドマーク座標検出ネットワークは CG 画像で事前学習されたモデルを利用する．

次に，画像生成ネットワークとランドマーク座標検出ネットワークの同時学習手順について説明する．

(1) CG 画像を生成器に入力し，生成された画像とエッジ画像を両識別器に入力し，識別器の損失を算出する．また，ランドマーク座標検出ネットワークに生成された画像に生成画像を入力し，ヒートマップ画像間の損失を算出する．これらの損失を利用して生成器を学習する．

(2) 同 Iteration で，ランドマーク座標検出ネットワークに CG 画像と生成画像を入力し，ヒートマップ画像間及びランドマーク座標間の損失を利用して，学習する．

(3) 偶数 Iteration のみ，識別器に生成画像と実画像及びそれぞれのエッジ画像を入力し，WGAN[14] の手法で算出される損失を利用して，識別器を学習する．
以上の手順を繰り返し，同時学習ネットワークの学習を行う．

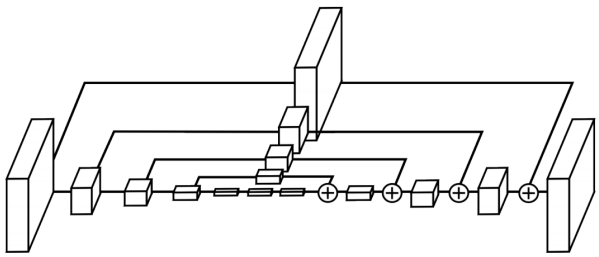

図 6 ランドマーク座標検出ネットワークの構造として用いる Hourglass Network[11] の模式図.

3.2 実験条件

初めに，今回の実験で使用する学習用データと検証用データについて説明を行う．

学習用の CG 画像のデータには，UnityEyes[13] で作成し，正面を向いた画像を 1 万枚，視線方向を上下 15 度に変更した画像，左右 30 度に変更した画像を各 1 万枚の計 5 万枚を作成した．画像と同時に生成されるアノテーションデータには，瞼や虹彩などのランドマーク位置が含まれており，虹彩のランドマーク位置から重心座標を算出し，重心座標を中心に 160×96 のサイズにクロップ，その後グレースケールに変換した画像を利用する．

学習用の実画像は，近赤外カメラで撮影された動画から CG 画像と同サイズにクロップされた画像を計 6,491 枚利用した．また，検証用データセットには，正面、上下左右各 5 枚ずつの計 25 枚の実画像データセットの虹彩に vatic[16] と呼ばれるアノテーションツールを用いて，手動でアノテーションを行い，作成されたバウンディングボックスの重心位置を虹彩中心として算出した．

画像生成ネットワークについて，生成器の学習は，最適化手法に Adam[17] を用い，学習率は 0.0002 とし，識別器の学習は，最適化手法に Adam を用い，学習率は 0.0001 とした．

提案手法の検証の為，以下の比較実験を行った．

(1)CG 画像で Seonwook らの手法 [4] を学習する．

(2)Li らの手法 [12] 同様，分類器を追加して学習を行い，生成された画像を用いて Seonwook らの手法を学習する．

(2) について，CG 画像及び実画像に正面，上下左右の 5 クラスでラベル付けを行い，分類器の学習は，損失関数として Binary Cross Entropy(BCE), 最適化手法には Adam を用い，学習率は 0.0002 とした．また，ランドマーク位置の保持の為に，SimGAN で使用されている L1 損失と分類器で算出された損失を利用した．提案手法及び (1)(2) の手法について，検証用画像をランドマーク座標検出ネットワークに入力し，虹彩中心のランドマーク座標を推定する．各視線方向に対して，推定した虹彩中心のピクセル誤差を指標として性能評価を行う．

3.3 実験結果

提案手法及び比較実験の検出結果の例を図 9 に示す．また，検証用データの 25 枚の画像に対して，虹彩中心のピクセル誤差の平均により，評価した結果を表 2 と 3 に示す．提案手法は，手法 (1) と比較して精度は上がったが，手法 (2) と比較すると，x 座標は一部の視線方向のみ精度が上がり，y 座標は精度が下がってしまうという結果となった．(2) の手法について，図 8 より視線方向が左向きの場合に虹彩の位置，下向きの場合に瞼の形状が，画像と推定結果で一致していないことがわかる．

生成画像でランドマーク座標が一致していない原因として，

Vol.2020-CG-180 No.6
Vol.2020-DCC-26 No.6
Vol.2020-CVIM-223 No.6
2020/11/5

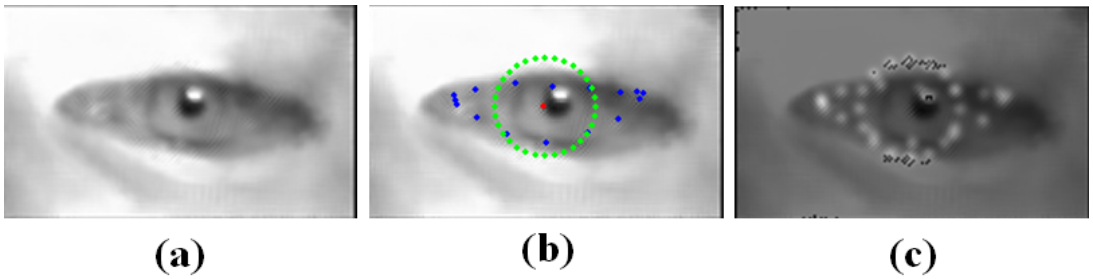

図 7 (a) 生成画像．(b) 生成画像に CG のアノテーションデータを重畳表示した画像．(c) 生成画像に推定されたヒートマップ画像を重畳表示した画像．
(a) の生成画像内に線が発生しており，(b) に重畳表示している CG 画像のアノテーションデータと一致している事がわかる．また (c) について，推定されたヒートマップも線に対して検出している事がわかる．

表 2 各視線方向における虹彩中心推定結果の x 座標ピクセル誤差

	正面 (x[pix])	上 (x[pix])	下 (x[pix])	左 (x[pix])	右 (x[pix])
提案手法	4.476	3.925	11.991	7.740	20.449
手法 (1)	14.041	18.376	28.949	19.993	4.994
手法 (2)	7.976	4.966	9.549	17.011	0.574

表 3 各視線方向における虹彩中心推定結果の y 座標ピクセル誤差

	正面 (y[pix])	上 (y[pix])	下 (y[pix])	左 (y[pix])	右 (y[pix])
提案手法	2.014	2.117	4.003	2.992	3.999
手法 (1)	1.987	8.054	2.995	5.065	1.275
手法 (2)	0.935	1.989	1.492	1.594	1.061

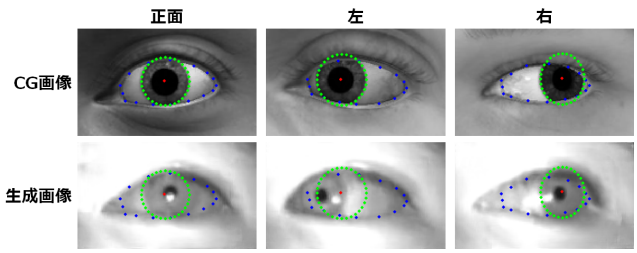

図 8 CG 画像と生成画像にランドマーク座標を重畳表示した例．
生成画像は CG 画像の瞼や虹彩の位置を保持する事が出来ていない為，ランドマーク座標が一致していない．

分類器は 5 クラスの視線方向で画像を分類している為，生成器は視線方向が一致した画像を生成する事が可能である．しかし，瞼や虹彩の位置の保持については，分類器で制約する事は出来ず，L1 損失のみで制約しており，図 4 に示した様に，L1 損失では保持することが出来ない事がわかっている．その為，ランドマークの位置が CG 画像と一致していない画像が生成され，その画像でランドマーク座標推定ネットワークを学習した事によって、推定結果に誤差が生じたと考えられる．次に，提案手法のピクセル誤差が大きくなってしまった要因として，図 7(a) に示す様に，CG 画像のランドマーク座標位置に合わせて生成画像上に線が発生しており，この様な生成に失敗した画像でランドマーク座標検出ネットワークを学習してしまっている事が挙げられる．生成が失敗している原因として，生成器は虹彩などの位置を保持する制約に，CG 画像と生成画像のヒートマップ間で算出した損失を利用している．しかし，損失を下げる為に生成画像のランドマーク座標位置を修正するのではなく，生成器は図 8(b) のランドマーク座標位置に線を描いた画像を生成してしまっている．図 8(c) を見るとわかるように，発生している線に対してヒートマップを推定している為，線が発生していない実画像を入力した際にランドマーク座標の推定精度が下がってしまう事が考えられる．
本来，画像生成ネットワークはヒートマップ間の損失によって，生成画像のランドマーク座標位置を CG 画像に合わせつつ，実画像感を付与した画像を生成するように学習し，またランドマーク座標検出ネットワークは，瞼や虹彩を正しく検出できる出来る様、学習をする必要性がある．

4. まとめと今後の課題

本研究では，画像生成ネットワークにランドマーク座標検出器を追加し，画像生成ネットワークとランドマーク座標検出ネットワークを同時学習できるネットワークを提案した．また実験から，ヒートマップ間で損失を算出し，それを用いて生成器を学習することによって，画像の生成精度が向上し，生成画像で学習したランドマーク座標検出ネットワークの虹彩中心の検出誤差は，CG 画像で学習したネットワークよりも小さくなることがわかった．一方で，一部の視線方向については，生成画像に線が発生してしまい，その画像で学習を行ってしまった為，検出誤差が大きくなってしまった．生成画像に線が発生してしまう問題を解決する方法として，ローカル識別器のネットワークサイズを変更し，識別器の受容野の大きさを変更すること

が挙げられる．現状，実画像に存在しないはずの線が生成画像上に発生しており，発生している線が非常に細い為、ローカル識別器では検出する事が出来ていない．その為に，ローカル識別器のネットワークサイズを現状よりも大きくし，受容野を小さくすることによって，線を検出することが可能となり，線の有無も含めて識別を行い，その結果から生成器が学習する事によって，線が発生しない画像を生成する事が出来る可能性がある．

次に，瞼や虹彩の位置が CG 画像と生成画像で一致しない問題を解決する方法として，Jun ら [7] によって提案された Cycle GAN の手法を利用することが挙げられる．Cycle GAN では生成画像を入力画像に再構成するネットワークを追加し，入力画像と再構成された画像で損失を算出する事によって，入力画像を実画像に近付ける為に変換すべき領域についても学習する事が可能である．この様な再構成ネットワークを本手法に追加することで，CG 画像の皮膚や虹彩の部分など変換すべき領域のみを変換しつつ，ランドマーク座標の位置を保持することが出来る可能性がある．以上の様に，ローカル識別器のネットワークサイズの変更や再構成ネットワークを導入することによって，画像生成ネットワークの問題点を解決し，ランドマーク座標検出ネットワークを学習する事で検出精度の更なる向上を目指す．

参考文献

[1] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017.

[2] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018.

[3] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. *IEEE Transactions on Image Processing*, 29:5259–5272, 2020.

[4] Seonwook Park, Xucong Zhang, Andreas Bulling, and Otmar Hilliges. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10, 2018.

[5] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[7] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[8] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[9] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[10] Zhe He, Adrian Spurr, Xucong Zhang, and Otmar Hilliges. Photo-realistic monocular gaze redirection using generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6932–6941, 2019.

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[12] LI Chongxuan, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Advances in neural information processing systems*, pages 4088–4098, 2017.

[13] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 131–138, 2016.

[14] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[15] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[16] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International journal of computer vision*, 101(1):184–204, 2013.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vol.2020-CG-180 No.6
Vol.2020-DCC-26 No.6
Vol.2020-CVIM-223 No.6
2020/11/5

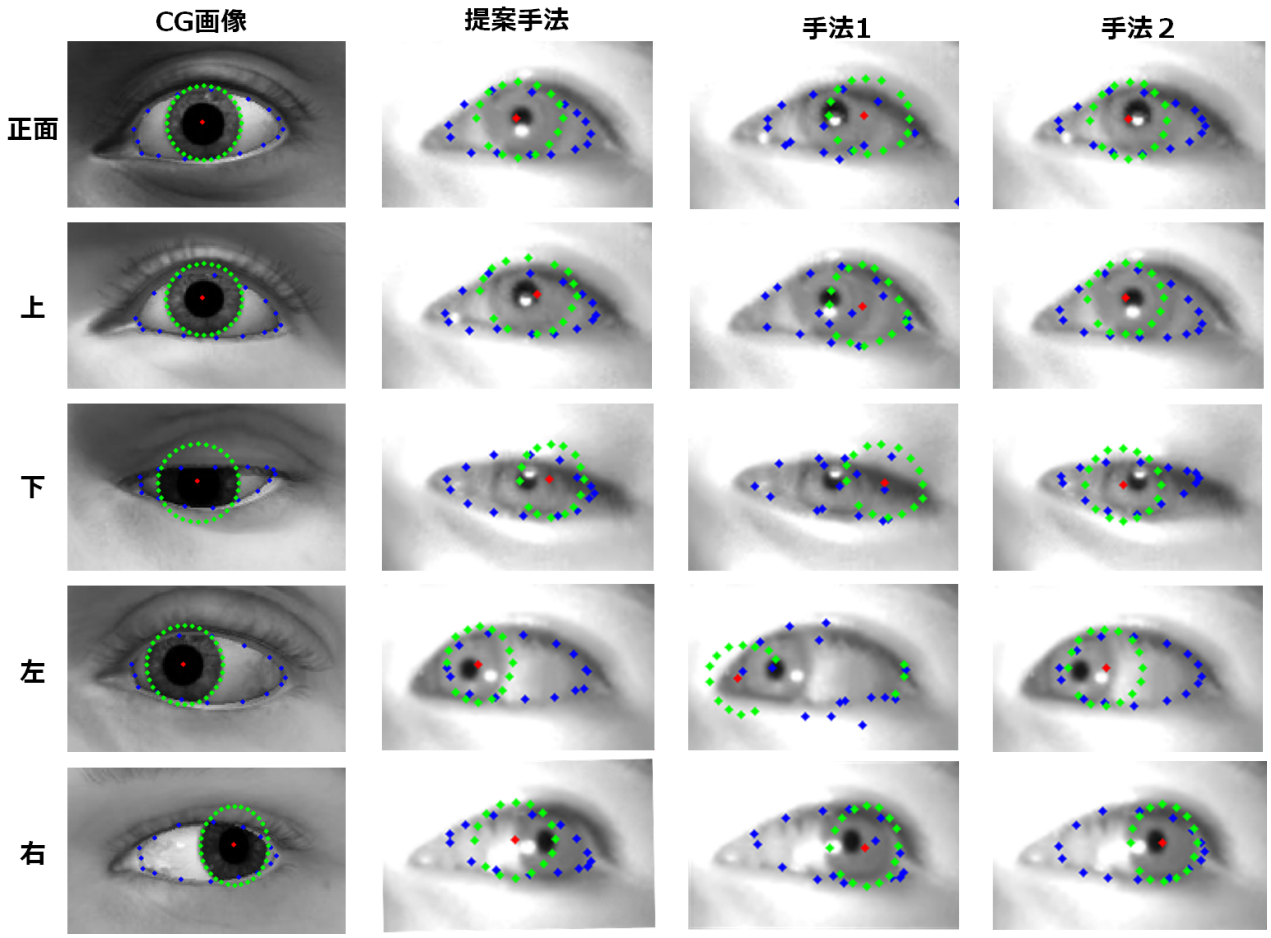

図 9 提案手法及び比較実験のランドマーク座標検出結果例.
青が瞼，緑が虹彩，赤が虹彩中心のランドマーク座標を表している．提案手法は手法 (1) と比較して，虹彩中心の位置が正しく推定されていることがわかる．また，手法 (2) と比較して，正面や左向きの画像については正しく推定されているが，下や右については精度よく推定する事が出来ていない．