F-003

Pix2pix を用いた運転中の一人称視点から三人称視点の多段階的生成
Multi-step generation of first-person to third-person

戸村 尋稀† 平石 広典†
Hiroki Tomura Hironori Hiraishi

1. はじめに

本研究では，普段の運転特性をより理解するために，ドライブレコーダで撮影した映像を利用する．ドライブレコーダの映像は運転席から前方を捉えた一人称視点の映像である．しかし，自分の運転を正確に理解するためには，自分の車を後ろから捉えた第三人称視点の映像の方が，車の動きをより正確に理解することが可能である．

そのため，本研究では，第一人称視点から第三人称視点映像の生成を試み，敵対的生成ネットワーク(GAN)の 1 種である pix2pix を使用した[1,2]．これは生成器と識別器と呼ばれる 2 つのネットワークをお互いに競わせて学習を行い，画像を生成するものである．

事前研究では，データの数に焦点を当て，トレーニングデータの直線画像とカーブ画像の比率を比較した学習とトレーニングデータを無作為に収集し学習した実験を行った[3]．事前研究の結果では，直線画像を約 5000 枚，カーブ画像を 1000 枚で学習した時の結果が最も良い精度であった．しかし，評価データを使用した評価実験では，どの実験モデルでも生成精度に大きな違いはなかった．

本研究では，画像のカラー情報を無くして学習を行い，3 人称視点のグレースケール画像を生成し，生成した画像を入力画像として学習を行い，カラー画像の生成を行い，事前研究でも行ったカラー画像からカラー画像の学習・生成を行い比較した．

2. システム概要

学習データの作成には 3D シミュレータを使用した．実際に使用した画像を図 1 に示す．ドライビングシミュレータから入力画像として第一人称視点の画像（図 1 左）とリアル画像(教師画像)として第三人称視点の画像（図 1 右）をペアとしたデータセットの作成をした．

Input Image Real Image

図 1 入力画像と教師画像

† Ashikaga University, Faculty of Engineering

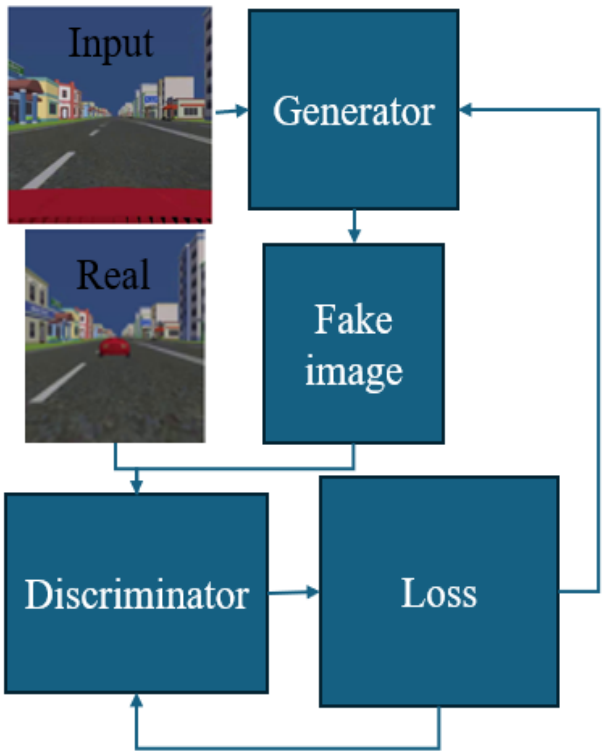

図 2 GAN のネットワーク構造

本研究で，使用する敵対的生成ネットワークの構造を図 2 に示す．敵対的生成ネットワークでは，Generator によって画像が生成され生成された画像が，Discriminator でリアル画像との比較が行われて，生成画像の真偽が判断される．その時の Loss が Generator と Discriminator に帰され，Generator には Loss が最小化するように学習され，Discriminator では，Loss が最大化するように学習される．

3. pix2pix の多段階的生成比較

本研究では，pix2pix を多段階的に使用しての学習とカラー画像同士で pix2pix の学習を行い比較する．図 3 に多段階的生成のフローチャート図を示す．

Pix2pix を使用した多段階的生成では，初めにグレースケール化をした入力画像と教師画像で学習を行い，グレースケールの 3 人称視点を生成する．次に，生成した 3 人称視点のグレースケール画像を入力画像として学習を行う．最後に，評価データを使用して，グレースケール 3 人称視点を生成し，生成画像からさらにカラー画像を生成し評価を行う．カラー画像同士の学習では，入力画像と教師画像をカラー画像で読み込み学習を行う．

事前研究ではトレーニングデータを増やして実験を行っていたが，今回の多段階的生成の比較実験では，事前研究の結果を踏まえて直線画像 500 枚，カーブ画像 100 枚として，トレーニングデータを 600 枚用意した．バッチサイズ

を 16，エポック数を 1000 回として学習を行い，学習で最も良かった Loss の時のモデルを使用して評価を行った．また，画像サイズを 256 ピクセルから 512 ピクセルに変更し，解像度を良くして実験を行った．

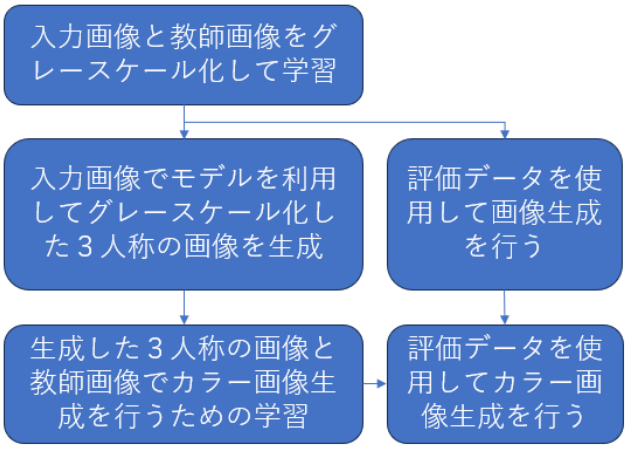

図 3　pix2pix の多段階的生成フローチャート図

結果として，学習結果は多段階生成の方が上手く学習できた．それぞれの学習時の平均絶対誤差を表 1 に示す．表 1 に示した平均絶対誤差(MAE)は，0 に近づくほど学習がうまくできたことを意味するため，カラー画像同士の学習よりも多段階生成の学習が上手くできていることが分かる．

表 1　学習時の平均絶対誤差

	カラー画像同士	多段階生成
グレー画像の学習		3.18
グレー画像からカラー画像の学習		3.18
カラー画像同士の学習	11.32	

3.1 評価実験

評価実験では，評価用データとして 1 分 40 秒の動画を使用する．画像枚数にすると 3038 枚である．しかし，3038 枚には，同じ画像が多くあるので，評価用データを 30 枚に厳選し，30 枚の画像を使用して生成した画像の評価を行う．

表 2 では，評価実験での平均絶対誤差を示す．表 2 を見ると，学習時の結果とは異なりカラー画像同士モデル生成した画像の方が，良い生成精度になったが多段階生成との差はほとんどなかった．また，学習時の生成精度と比べると多段階生成は，明らかに生成精度が悪く過学習が起きていると考えられる．

表 2　評価実験の平均絶対誤差

	カラー画像同士	多段階生成
グレー画像の生成		9.54
グレー画像からカラー画像生成		13.27
カラー画像同士の生成	13.21	

図 4 に評価データを使用して生成した画像を示す．生成画像を比べるとカラー画像同士の生成は，全体的にノイズは少ないが車や建物の生成がうまくいっておらず，多段階的生成では，全体的にノイズが多い結果となった．

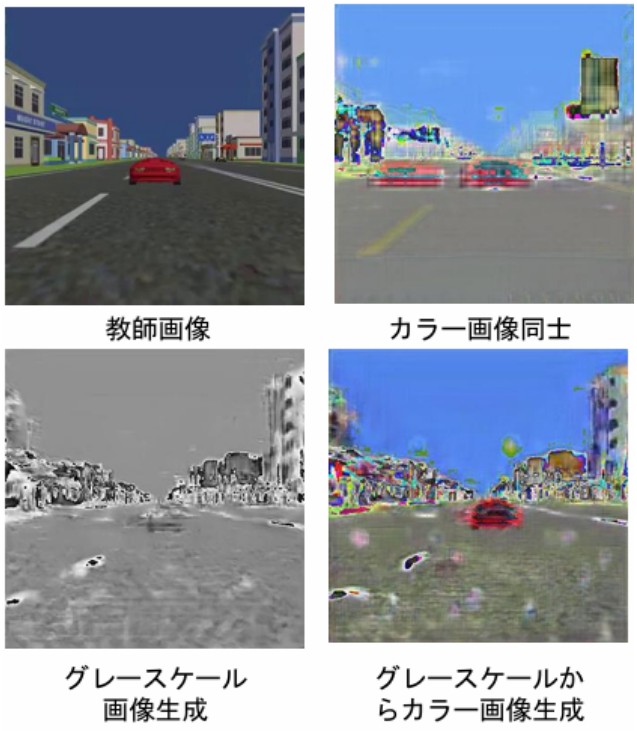

図 4　評価結果

4. おわりに

本研究では，運転評価を目的として，ドライブレコーダのような第一人称視点の運転映像から第三人称視点映像の生成を試みた．そのため，ＧＡＮのモデルの一つである pix2pix を利用した．ドライブシミュレータを利用して，1 人称映像と 3 人称映像の作成を行い，その映像を利用して評価を実施した．

本研究の実験では pix2pix を多段階的に使用しての学習とカラー画像同士で pix2pix の学習を行い比較した．学習時の結果は多段階生成の方が良い結果になったが，評価データを使用した結果では，カラー画像同士の評価が良い結果となった．また，多段階生成の学習時の平均絶対誤差と評価結果を比べると多段階生成では過学習が起きていると考えることができる．

参考文献

[1] P. Isola, J. -Y. Zhu, T. Zhou and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 5967-5976

[2] L. A. Gatys, A. S. Ecker and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2414-2423

[3] Hiroki Tomura, Hironori Hiraishi, "First-Person View to Third-Person View Generation Using Pix2pix in Driving," The Thirtieth International Symposium on Artificial Life and Robotics 2025 (AROB 30th 2025),2025,1, pp.59-62