Vol.2020-MPS-127 No.4
2020/3/2

GANs による歩行者の行動予測

何 巴特[1]　馮 軒昂[1]　北 栄輔[1]

概要：画像から歩行者の歩行行動予測技術は多くの場面で役立つ．自動車に装備すれば，歩行者との交通事故を回避するために利用できる．公衆の監視カメラと組み合わせて利用すれば，不審者を検出されることができる．本研究では，生成型敵対的ネットワーク（GANs）を用いて，画像から歩行者の行動を予測するアルゴリズムを生成する．入力データとして歩行者の連続する歩行画像をとり，出力データとして入力データに続く未来の予測画像をとる．本研究では，歩行者行動の予測に適用するため，GANs アルゴリズムを改良している．GANs のネットワークを評価するために，ピーク信号対雑音比（PSNR）を採用する．つづけて，主観的な評価方法として CNN 分類器を訓練して使用する．また，歩行者画像を修得するために Microsoft KINECT（以下，KINECT）を用いる．提案した GANs を LSTM と CNN を組み合わせた従来手法と比較したところ，良い結果が出ることが分かった。

キーワード：敵対的生成ネットワーク(GANs)，画像処理，行動予測

Pedestrian Behavior Prediction by using Generative Adversarial Networks

BATE HE[†1]　XUANANG FENG[†1]
EISUKE KITA[†1]

Abstract: Pedestrian behavior prediction is a very interesting technique in many scenarios. If this system is equipped at the car, the traffic accident against a pedestrian can be avoided. Suspicious people can be detected by analyzing pedestrian's data which came through the surveillance cameras. In this research, we focus on the prediction of pedestrian behavior by using improved Generative Adversarial Networks (GANs). The successive images of pedestrians are taken as the explanatory variables, and the future images of the pedestrians are as the objective variable. In this study, we transform the structure of the improved GANs algorithm for the purpose of making the better prediction of pedestrian behavior. Since the GANs' training is unstable and it is hard to find an objective way to evaluate the network, we used Peak signal-to-noise ratio (PSNR) as the objective evaluation method and we trained a CNN classifier as the subjective evaluation method. The experimental data that was self-made in this research was collected by Kinect.

Keywords: Generative Adversarial Networks (GANs), Image Processing, Behavior Prediction

1. はじめに

画像から歩行者の将来の歩行行動を予測することができれば，様々な場面で応用することができる．一つの例として，自動車での応用が考えられる．道路交通の場面においては，自動車ドライバーは運転操作をすると同時に，短時間で路肩の歩行者が自動車の前を横切るかどうかを判断する必要がある．もし，歩行者の歩行行動予測技術が自動車に設置・利用されていれば，自動車は歩行者に対する潜在的な危険を予測して，自動的に回避できる．地下鉄駅等の公共エリアには様々な場所に監視カメラが設置されている．これらのカメラ画像に歩行者の将来の歩行行動予測技術を適用できれば，犯罪や危険行為を事前防止することができるようになる．

歩行者の行動予測に関する研究は，様々な研究者によって行われている[4][5][7][13]．従来の研究では，歩行者の動線を分析し，それを通して行動を予測する方法が主流である．このとき，数理モデルはマルコフチェーンモデルによって定義されている[4]．これに対して，本研究では，連続する歩行者の画像データから，次の画像データを推定する方法について研究する．しかし，画像から行動分析を行うためには，多数の画像データが必要となる．そこで，画像データを保管するために生成型敵対的ネットワーク (Generative Adversarial Networks, GANs) を用いる．本研究では， GANs を一連の連続画像から未来の歩行者の画像を生成するために用いる．

実験に用いる画像データセットは３種類ある．第１は，被験者に参加をお願いして MicroSoft Kinect を用いて撮影したデータセットである．第２と第３は，比較実験のため公開されている二つのデータセットである．

本論文の構成は以下のようになっている．第２章では，提案アルゴリズムについて述べる．第３章では，解析結果を示し，第４章はまとめである．

2. 提案手法

本研究では，オリジナルの GANs アルゴリズムを，本研

1 名古屋大学
Nagoya University

究の目的にあわせて修正して用いる．修正点をまとめると以下のようになる．

2.1 入力変数と出力変数

オリジナルの GANs の生成器への入力はランダムノイズであるのに対して，提案した GANs の入力は連続的な画像データであること．

また，オリジナルの GANs の出力は「偽」画像であるのに対して，提案した GANs の出力は予測画像であること．つまり，提案した GANs では，4 つの連続画像データから一つの未来画像を出力する．

2.2 損失関数

オリジナル GANs と提案した GANs では損失関数が異なっている．

オリジナルの GANs の損失関数は次式で与えられる．

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

ここで， $\boldsymbol{x}$ は実データ，$\boldsymbol{z}$ は生成器に入力されるノイズである．G(x)は生成器が生成したデータ，D(z)は判別器が判断した”真”のデータである確率，Pdata は”真”のデータセットを示す．

これに対して，提案した GANs の損失関数は次式で与えられる． ***(X, Y)***をデータセットからの連続画像データとする． たとえば，***X*** は 4 つの連続画像データ，***Y*** は ***X*** につづいて現れる５つ目の画像データである．提案した GANs では，入力の***(X, Y)***をクラス 1 に，入力の***(X, G(X))***をクラス 0 に分類するように判別器をトレーニングする．したがって，提案する GANs の判別器のトレーニングに関する損失関数は次式で与えられる．

$$L^D(X,Y) = L_{bce}(D(X,Y),1) + L_{bce}(D(X,G(X)),0) \quad (2)$$

ここで，W_D は判別器のネットワーク構造の重みを示す．判別器の損失関数（式(2)）を減少させるように W_D を更新する。判別器は***(X, Y)*** と ***(X, G(X))*** の特徴を学習して、それぞれ 0 と 1 に分類できるようにネットワーク構造の重み W_D を更新する．

生成器は，判別器が***(X, G(X))***をクラス 1 に分類しように，判別器を訓練する．したがって，生成器のトレーニングに関する損失関数は次式で与えられる．

$$L^G(X,Y) = L_{bce}(D(X,G(X)),1) \quad (3)$$

ここで，L_{bce} は交差エントロピー損失関数であり，次のように定義される：

$$L_{bce}(Y,Y') = -\sum_i Y_i' \log(Y_i) + (1 - Y_i') \log(1 - Y_i) \quad (4)$$

ここで，W_G は生成器のネットワーク構造の重みを示す．生成器の損失関数（式(3)）を減少させるために W_G を更新する。生成器は***(X, Y)*** と似ていた画像***(X, G(X))*** を生成する目標で W_G を更新する、つまり画像の類似度は高いほど判別器が生成した画像***(X, G(X))*** を 1 に分類する確率が高いということになる．

判別器の訓練目的は，判別器の損失関数L^Dをできるだけ小さくすることである．これは，生成器が生成した画像をはっきり認識できるように訓練することを意味する．また，生成器の訓練目的は，生成器の損失関数L^Gをできるだけ小さくすることである．これは，生成した画像が判別器を騙すように訓練することである．

本研究で提案する GANs アルゴリズムを表 1 に示す．M 個のデータセットを用意する．判別器と生成器のパラメータl^Gとl^D，重み係数λを入力する．繰り返し係数Nを定義する．予測した出力画像を出力する．判別器と生成器を学習する．また，実験のフローチャートを図 4 に示す．

TABLE 1. IMPROVED GANS ALGORITHM

Algorithm 1 Training adversarial networks for next frame generation
Input:
(1)M data samples:$(X,Y) = (X^{(1)},Y^{(1)}),\cdots,(X^{(M)},Y^{(M)})$;
(2)Generator and Discriminator's learning rates: l_G and l_D;
(3)The weights λ;
(4)The number of iteration frequency N;
Output: The future images that predicted.
for number of training iterations N **do**
  Update the discriminator D:
  Get M data samples:$(X,Y) = (X^{(1)},Y^{(1)}),\cdots,(X^{(M)},Y^{(M)})$;
  $W_D = W_D - l_D \sum_{i=1}^{M} \frac{\partial L^D(X^{(i)},Y^{(i)})}{\partial W_D}$
  Update the Generator G:
  Get M new data samples:$(X,Y) = (X^{(1)},Y^{(1)}),\cdots,(X^{(M)},Y^{(M)})$;
  $W_G = W_G - l_G \sum_{i=1}^{M} \frac{\partial L^G(X^{(i)},Y^{(i)})}{\partial W_G}$
endfor

本研究では，画像内に歩行者が 1 人だけ歩行する場合を考えており，背景には情報が含まれていない．そこで，提案した GANs の判別器では全結合層を畳み込み層に切り替える．

全結合層では，最後の畳み込み層に接続されている完全結合層のパラメーターサイズが大きすぎるため，トレーニングとテストの計算量が増加し，計算速度が低下する．また，モデルではドロップアウトなどの方法を使用するが，ドロップアウトレートなどのハイパーパラメーターをあらかじめ設定する必要がある．

画像データは，縦・横・チャンネル方向の３次元の情報を有している．全結合層では，このデータセットが有する空間情報がなくなってしまう．これには，ピクセルや RBG の各チャンネルの間の空間的な近似性がある．距離の離れたピクセル同士はあまり関わりがなかった。３次元の形状

の中にはくみ取るべき本質的なパターンが潜んでいるばあいがあるが，全結合層では，これらの情報が削除されてしまう．これに対して，畳み込み層は形状を維持する。

改良した GANs モデルの有効性を自作のデータセットによって確認する．実験では四つの連続画像と二つの出力画像を設定して実験をする．

x 軸は訓練するエポック数を表す．また，***y*** 軸は，生成器のロス，判別器のロス，PSNR 値を表す． 生成器と判別器の損失関数の値は小さいほど良く，モデルの信頼性が高くなる． PSNR の値は大きいほど良い．改善された GANs アルゴリズムのパフォーマンスは良好であると結論付けることができる．

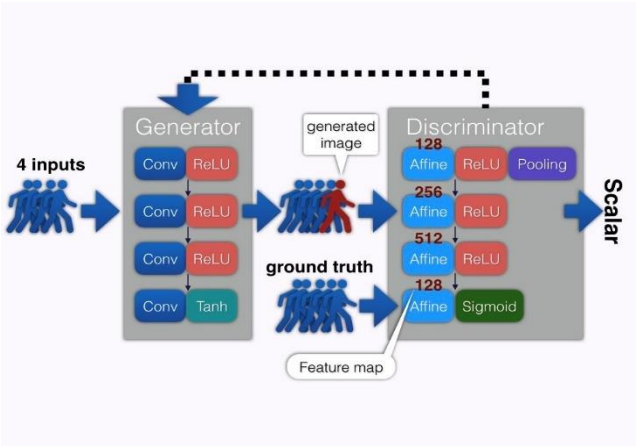

図 5　改良する前の GANs モデル

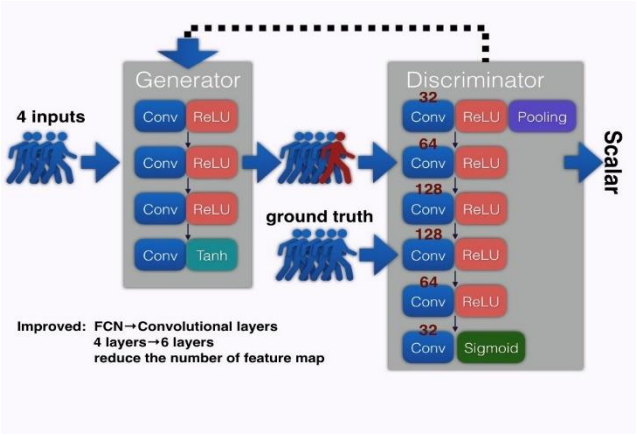

図 6　改良した後の GANs モデル

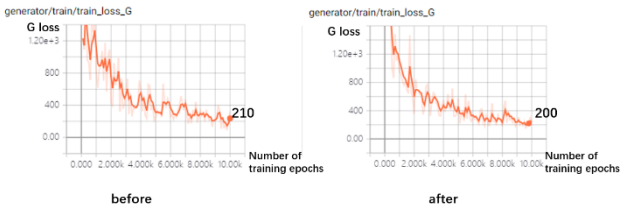

図 7

3. 実験と結果

3.1 実験データセット

本研究は三つのデータセットを実験に用いる．

第 1 のデータセットは，研究室で実験のために作成したデータセットである．これは Kinect の RGB カメラを使用して，歩行者の歩行状態を撮影して作成したデータセットである．5 人の被験者から撮影した．Kinect センサーを 2 メートルの高さに設定して，各人が 8 方向に歩行している様子をビデオ撮影する．各ビデオは 25 fps 含んでいる．これらの画像に，背景を除く前処理を施す．画像の背景を削除する理由は次の通りである．モデルのトレーニング中に背景は静止したままであり，フレーム間の移動はないため，背景を削除することで，計算メモリを節約し，計算時間を短縮することができる．

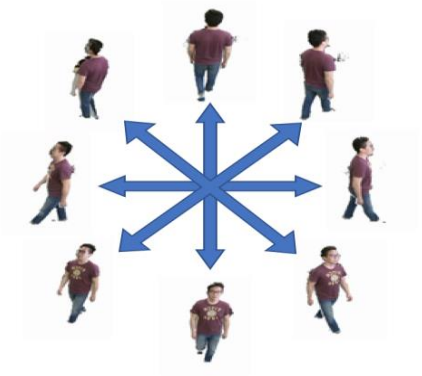

図 8　データセット 1

図 9 データセット 1 の加工

図 10　Actions as Space-Time Shapes

第 2 のデータセットは，公開されているデータセット「Actions as Space-Time Shapes」である（図 9）.「Actions as Space-Time Shapes」は 50fps のビデオ 90 個を含んでいる．解像度は 180×144，“run,” “walk,” “skip,” “jumping-jack” (or shortly “jack”), “jump-forward-on-two-legs” (or “jump”), “jump-in-place-on-two-legs” (or “pjump”), “gallopsideways” (or “side”), “wave-two-hands” (or “wave2”), “waveone- hand” (or “wave1”), or “bend.”という 10 種類の行動を行う 9 人のビデオデータからなっている．

第 3 のデータセットは，公開されているデータセット「Recognition of human actions」である（図 10）.「Recognition of human actions」は 25fps のビデオ 600 個を含んでいる．解像度は 180×144，walking, jogging, running, boxing, hand waving and hand clapping という 6 種類の行動を行う 25 人のビデオデータからなっている． 1 人が一つの動作を行う一つの姿勢に対して次の 4 つのシナリオがある．つまり，outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3 and indoors s4 である．

3.2 実験目的

本研究は，実験プログラムを python3.6 によって作成する．用いる python パッケージには，tensorflow，numpy，skimage，scipy がある．また，計算には GTX 1060G と GTX 730 GPU の 2 種類の GP-GPU を用いる．

本研究では， 2 つの実験を行う．第 1 の実験では，提案した GANs モデルの性能を LSTM+CNN のモデルと比較する．第 2 では，ハイパーパラメーターの影響を検討する．

3.3 評価基準

結果を比較検討するために，2 つの評価方法を用いる．第 1 は，ピーク信号対雑音比（PSNR）である． PSNR は，信号の最大可能電力と，その表現の忠実度に影響を与える破損ノイズの電力との比率を表す工学用語． PSNR は，非可逆圧縮コーデックの再構築の品質を測定するために最も一般的に使用される．この場合，信号は元データであり，ノイズは圧縮によって導入されたエラーである．PSNR は単純に二乗平均誤差（MSE）で定義される． 2 つの m * n の解像度の画像 I および K，それらの平均二乗誤差は次のように定義される．

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{5}$$

これを用いて PSNR は以下の数式で定義される．

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{6}$$

今までの未来の画像を生成する研究の多くでは主観的な視覚的評価(PSNR など)を使って生成した画像の品質を評価する．しかし、PSNR は画像のシャープネスを評価する指標である．そこで，二番目の評価方法を提案する。本研究で使用したもう一つの評価方法は生成された画像品質を客観的にテストできる CNN 分類器であり、歩行者がどの方向に進むかを判断することができる。図 11 で示すように、8000 個の画像データを使用して、CNN 分類器モデルを 8 方向にモデルをトレーニングする。それから図 12 で示すように、改良した GANs で生成した画像をすでに訓練した CNN 分類器に入力して、8 方向の一つに分類する、これで分類器に分類する確率で GANs モデルを評価する。

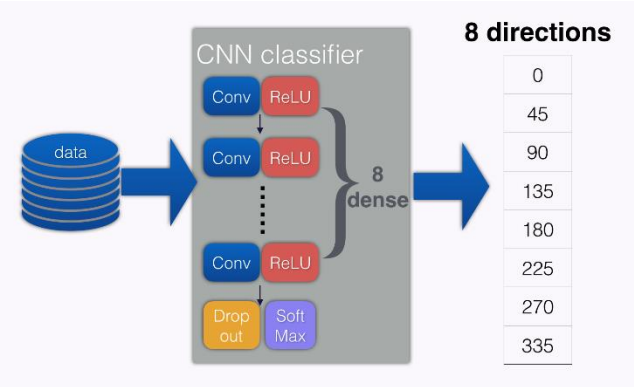

図 11 CNN 分類器

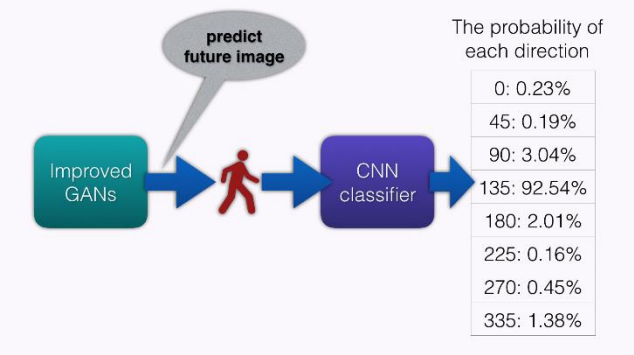

図 12 CNN 分類器で生成した画像を評価する

3.4 実験結果

第 1 に，提案した GANs モデルを LSTM＋CNN のモデルと比較する．このとき，入力データとするフレームの数を 4，出力データとするフレームの数を 2 として実験を行う．

得られた結果を表 2 に示す.

Table 2　PSNR 値

データセット	LSTM＋CNN	提案手法
自作	25.3	27.5
Actions as Space-Time Shapes	27.0	28.1
Recognition of human actions	29.0	31.6

この表から見ると，改善した GANs モデルは LSTM＋CNN のモデルより PSNR の値が優れているということがわかる.

二つ目の実験結果について，ディープラーニングでは，勾配降下法と呼ばれる簡単な 1 次収束アルゴリズムを使用する. 提案する GANs には 2 つのネットワークがあるので，2 つの学習率を設定する必要がある．ネットワークをトレーニングするために生成器を修正し，異なる学習率で判別器の損失関数を比する．そして，トレーニングネットワークの判別器を修正し，生成器の学習率を比較する．本実験は，4 フレームの入力と 2 フレームの出力で行う．結果を図 13，14，15 に示す．生成器の学習率が小さすぎると，生成器と判別器の損失関数の両方が悪化することが分かる．さまざまな学習率で実験を行った結果，生成器の学習率= 0.00004 および判別器の学習率= 0.02 とすることで，生成器の損失と判別器の損失関数を最小化できることが分かった.

入力フレームと出力フレームの数が 2 つ大きい場合，生成されたフレームが押しつぶされ，判別器と生成器の損失関数が収束できなくなっている．つまり，実験で最も適切な入力フレームの数と出力フレームの数があることがわかる．実験より，入力画像数と出力画像数の組み合わせが 2-2，4-2，4-4，6-2，6-4，6-6，8-2 ，8-4，8-6，8-8，10-2，10-4，10-6，10-8 のとき最も良い制度を示していることがわかる． PSNR の値と CNN 分類器の精度を比較することにより，最適な実験値を見つけることができている（表 3）.

Table 3　Right Pair

Input and output number	PSNR	The accuracy of CNN classifier
2-2	25.2	96.6%
4-2	27.5	94.8%
4-4	26.9	95.4%
6-2	28.2	96.8%
6-4	27.4	93.0%
6-6	25.5	95.9%
8-2	**29.1**	**98.3%**
8-4	27.4	**98.0%**
8-6	26.0	96.8%
8-8	24.7	95.1%
10-2	26.3	**98.5%**
10-4	25.0	97.7%
10-6	24.9	95.5%
10-8	24.5	93.0%

提案する GANs が生成した画像が多いほど，生成された画像からより多くの将来の情報を取得できることを意味するため, PSNR と生成された画像の CNN 分類器の精度を比較する形式を除き，より多くの画像が生成されることが望ましいという考えを加えて，(8 入力と 2 出力）というペアが一番よいパフォーマンスを発揮できる.

入力画像の例，生成した画像の例,「真」の未来の画像の例を図 13，14，15 に示す．図 13 には 8 枚の連続入力画像の例を示す. input_0.png から input_7.png までは歩行者の歩いる状態の連続画像，それから図 15 は input_7.png の続きに出てくる連続の二つ画像を示す. 図 14 は改良した GANs モデルで生成した画像を示す．図 14 と図 15 を比較したところ，顔の部分ははっきり生成しなかったが，歩行者の歩行方向の判別が可能であることがわかった.

図 13 8 枚の連続入力画像の例

図 14 2 枚の ground truth(入力画像の時系列で次の画像)

図 15　生成した画像

4.　結論

本研究では，改善された GANs モデルが歩行者の行動を予測するために訓練された． GANs は，ディープラーニング領域の教師なし学習で，モデルを評価するために，PSNR と CNN 分類器の 2 つの方法が提案されていた．改良された GANs アルゴリズムは，完全に接続されたレイヤーから畳み込みレイヤーに改善され，計算がより効率的になり，損失関数が減少した. 研究のデータセットは独創性があり，CNN を訓練するためにデータセットの量が増加した. また，二つの公開のデータセットを使って新しい手法である LSTM＋CNN のモデルとの比較実験で GANs アルゴリズムの結果がよいことが分かった．それから GANs モデルを訓練するのにいろいろ工夫をして GANs の訓練が良くなることが分かった.

本研究はいろいろ改善すべきところがあって，今後の研究でアルゴリズムと実験の設計をパーフェクトになるまで

改良したいと思います.

参考文献

[1] M. Mathieu, C. Couprie, Y. LeCun. Deep multi-scale video prediction beyond mean square error. *In International Conference on Learning Representations (ICLR).* 2016.

[2] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, Tao Mei. To Create What You Tell: Generating Videos from Captions. *In Proceedings of the 2017 ACM on Multimedia Conference*, pages 1789–1798. ACM, 2017.

[3] C. Vondrick, H. Pirsiavash, A. Torralba. To Generating Videos withScene Dynamics. *In Neural Information Processing Systems (NIPS)*. 2016.

[4] Y. F. Chen, M. Liu, S.- Y. Liu, J. Miller, J. P. How. Predictive modeling of pedestrian motion patterns with bayesian nonparametrics. *In AIAA Guidance Navigation and Control Conference*, page 1861, 2016.

[5] N. Schneider, D. M. Gavrila, J. Weickert, M. Hein, B. Schiele. Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study. *In Pattern Recog-nition ser. Lecture Notes in Computer Science*, Springer Berlin Heidelberg, vol.8142, pages 174-183, 2013.

[6] Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing,Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., Bengio,Yoshua. Generative adversarial nets. *In Neural Information ProcessingSystems (NIPS)*, 2014.

[7] Sarah Bonnin, Thomas H. Weisswange, Franz Kummert, Jens Schmuedderich. General behavior prediction by a combination of scenariospecific models. *In IEEE Transactions on Intelligent Trans-portation Systems*. pp. 1–11, 2014.

[8] Radford, L. Metz, S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. InarXiv preprint arXiv:1511.06434.2015.

[9] Krizhevsky, I. Sutskever, G. E. Hinton. Imagenet classifica-tion with deep convolutional neural networks. *In Neural Information Processing Systems (NIPS)*, 2012.

[10] J. Long, E. Shelhamer, T. Darrell. Fully convolutional networks for semantic segmentation. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[11] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville. Improved training of wasserstein gans. InarXiv preprintarXiv:1704.00028,2017.

[12] Z. Zhang. Microsoft kinect sensor and its effect. IEEE MultiMedia, 2012.

[13] Bera, Aniket, Dinesh Manocha. PedLearn: Realtime Pedestrian Tracking, *Behavior Learning, and Navigation for Autonomous Vehicles*.

[14] Shmelkov, Konstantin, Cordelia Schmid, Karteek Alahari. Howgood is my GAN?. *European Conference on Computer Vision*, 2018.

[15] Huynh-Thu, Quan, Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics letters* 44.13(2008): 800-8012008.

[16] Ranzato, Marc'Aurelio, Szlam, Arthur, Bruna, Joan, Mathieu, Michael, Collobert, Ronan, Chopra, Sumit. Video (language) modeling: a baseline for generative models of natural videos. CoRRabs/1412.6604, 2014.

[17] Revaud, Jerome, Weinzaepfel, Philippe, Harchaoui, Zaid, Schmid, Cordelia. EpicFlow: Edge Preserving Interpolation of Correspondences for Optical Flow. *In Computer Vision and Pattern Recognition*, 2015.

[18] Ziebart, B. D., Ratliff, N., Gallagher, G., Mertz, C., Peterson, K., Bagnell, J. A., Srinivasa, S. (2009, October). Planning-based prediction for pedestrians. *In IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009. IROS 2009., pp. 3931-3936. IEEE. 2009.

[19] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, Un-supervised pixel-level domain adaptation with generative adversarial networks. *In CVPR.* 2017.

[20] Molano-Mazon, M., Onken, A., Piasini, E., Panzeri, S. Synthesizingrealistic neural population activity patterns using generative adversarial networks. *In ICLR*, 2018.

[21] Dosovitskiy, Alexey, Jost Tobias Springenberg, Thomas Brox. SLearning to generate chairs with convolutional neural networks. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* pp.1538-1546. 2015.