一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2017-87(2017-10)

姿勢を表現する多様体に基づくGANsを用いた物体姿勢推定の検討

川西　康友†　出口　大輔††　井手　一郎†　村瀬　洋†

† 名古屋大学 大学院 情報学研究科　〒464-8601 愛知県名古屋市千種区不老町
†† 名古屋大学 情報連携統括本部 情報戦略室　〒464-8601 愛知県名古屋市千種区不老町
E-mail: †kawanishi@i.nagoya-u.ac.jp

あらまし　Generative Adversarial Nets（GANs）[1] は，ある事前分布から画像などのデータを生成可能なネットワークであり，乱数から様々なデータを生成可能である．本研究では，物体の姿勢変化は潜在的には多様体で表現可能であることに着目し，多様体上に定義した確率分布から画像を生成する GANs の枠組みと，それを用いて学習画像を補間しつつ姿勢推定器を学習する手法を提案する．実験では，全周を撮影した複数の物体に対し，提案手法により学習データに含まれない姿勢が補間できることを確認した．また，姿勢推定を同時に学習し，学習データに含まれない姿勢に対する姿勢推定の可能性を評価し，学習画像の枚数が少ない場合でも姿勢推定精度が大幅には低下しないことを確認した．

キーワード　多様体，敵対的生成ネットワーク，パラメトリック固有空間法

A Study on GANs based on Pose Manifold for Rigid Object Pose Estimation

Yasutomo KAWANISHI†, Daisuke DEGUCHI††, Ichiro IDE†, and Hiroshi MURASE†

† Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
†† Information Strategy Office, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan
E-mail: †kawanishi@i.nagoya-u.ac.jp

Abstract　Generative Adversarial Nets (GANs) is a pair of neural networks which can learn data distribution and generate various data from the distribution. In this research, by focusing on the fact that pose variation of a rigid object can be expressed on a manifold in a latent space, we introduce a GANs model which generates data from a distribution defined over a manifold. We also propose a pose estimation method which trains a pose estimator while interpolating training images using the GANs. We evaluated the interpolation capability of the proposed model using a public dataset, and also evaluated pose estimation accuracy of the proposed model.

Key words　Manifold, Generative Adversarial Nets, parametric eigenspace

1. はじめに

産業や生活支援など様々な分野でロボットが導入されつつある．産業分野では，ロボットの導入による生産現場での部品のピッキングに関する自動化の需要が高まっており，自動ピッキングロボットの性能を競う，Amazon Picking Challange [2] と呼ばれるコンペティションが開催されている．また，生活支援においても，少子高齢化の進展に伴う介護・福祉，家事等の労働力不足の懸念から，日常的な生活支援を目的としたロボットが開発されている [3]．生活支援におけるロボットの主な役割は，人の指示に従って，物体を受け渡しすることである．これらの分野とも，ロボットが物体を把持することが共通の課題であり，物体をうまく把持するための技術が求められている．ロボットが物体をうまく把持するには，対象物体をうまく把持できる方向を知るため，物体の姿勢推定を行なう必要がある．

物体の姿勢推定を実現するためにロボットに搭載するセンサとしては，Kinect や Xtion などに代表される 3 次元ビジョンセンサなどが考えられる．しかし，価格が高い，サイズが大きい，使用環境やデータ取得可能な対象物体が限定的，キャリブレーションが煩雑など，実用上の課題も多く，3 次元ビジョンセンサではなく従来の単眼カメラからの物体の姿勢推定技術が求められている．

一方，深層学習を用い，乱数から様々なデータを生成でき

るネットワークである敵対的生成ネットワーク（Generative Adversarial Nets; GANs）[1] が最近注目を集めている．GANs を用いることで，実際のデータと非常に類似した画像を大量に生成することが可能である．

本研究では，この GANs をもとに姿勢推定の学習データを補間し，精度の良い姿勢推定器を学習する方法論を採用する．一般の GANs では，データの生成に一様乱数などが用いられるが，本研究では，物体の姿勢変化は潜在的には多様体で表現可能であること [4] に着目し，多様体上に定義した確率分布から画像を生成する GANs の枠組みと，それを用いて学習画像を補間しつつ姿勢推定器を学習する手法を提案する．

以下，2 節で本研究に関係する，姿勢推定及び GANs に関する関連研究を述べる．次に，3 節で姿勢変化を表現できる多様体と，その多様体に基づく GANs を導入し，その GANs により画像を生成しつつ姿勢推定器を学習する方法について述べる．4 節で提案手法の有効性について評価する実験を行なう．最後に，5 節でまとめと今後の展望について述べる．

2. 関 連 研 究

2.1 物体姿勢推定

単眼カメラから物体の姿勢を推定する手法として，3 次元モデルをフィッティングする手法 [5]，回帰モデルに基づく手法 [6]，テンプレートマッチングに基づく手法 [4] などがある．テンプレートマッチングに基づく手法は，あらかじめ対象物体を様々な角度から撮影した画像とのテンプレートマッチングを行ない，最も類似した画像が対応する姿勢を推定結果として出力する，直観的な方法である．しかし，精度良く姿勢推定をするためには，膨大な数のテンプレートを記憶しておく必要がある．

この問題に対し，Murase ら [4] は，姿勢変化による 2 次元画像上での見えの変化を，主成分分析によって導出した低次元空間における多様体で表現するパラメトリック固有空間法を提案した．主成分分析により，画像の見え方の違いを最大化する低次元空間を得ることができる．その低次元空間へ投影したテンプレートの系列を 3 次スプライン曲線等で補間することにより，学習データに存在しない未知の姿勢にも対応できるため，記憶しておくべきテンプレートを減らすことができる．

パラメトリック固有空間法では，画像の見え方にのみ注目して低次元空間を得るため，姿勢変化による見え方の違いが小さいような物体には不向きである．これに対し二宮ら [7] は，画像の見え方の違いではなく，姿勢の分離性に着目した特徴量による多様体構築手法を提案し，パラメトリック固有空間法を拡張した．姿勢を教師信号として学習した Deep Convolutional Neural Network（DCNN）[8] の中間層から，姿勢の分離性が高い特徴を抽出し，この特徴量の空間中で補間を行なって多様体を構築する．DCNN は学習過程において識別に適した特徴量を自動的に獲得することができ，一般物体認識やシーン認識など様々なベンチマークで高い性能を示している [9]．そのため，画像全体の見えの分散を最大化する教師なし学習である主成分分析では区別できない見えの変化が小さい姿勢の違いであっても，姿勢を教師信号として教師あり学習を行なった DCNN を用いて抽出した姿勢の分離性の高い特徴量を用いることで区別することができる．しかし，深層学習による特徴抽出は非線形の変換であるため，単純に補間を行なうことで，その間の姿勢に相当する特徴量を得ることができる保証はない．

2.2 敵対的生成ネットワーク

敵対的生成ネットワーク（Generative Adversarial Nets; GANs）[1] は，データを生成する Generator と，データが実物か生成物かを分類する Discriminator の 2 つのネットワークの組であり，Generator は Discriminator を騙せるように，Discriminator は騙されないように，敵対的に学習が行われる．その結果，Generator は実物と区別がつかないようなデータを生成する分布を学習することができる．GANs は様々な拡張がなされているが，特に DCNN と組み合わせた Deep Convolutional Generative Adversarial Networks（DCGANs）[10] は，学習データの画像に近い画像を生成できることが知られている．Conditional GANs [11] は，生成モデルに用いる確率分布を，条件付き確率分布に変えることにより，条件に応じたデータ生成が可能とするモデルであり，例えば画像から画像への変換 [12] などが可能である．

近年，画像の生成だけでなく，それを分類器の学習に用いる手法として Shrivastava ら [13]，Wan ら [14] の研究がある．これらの研究は，シミュレーションによって生成した画像などから，実物に近い画像を生成する GANs を学習して画像を生成し，その画像を用いて視線推定や骨格モデルの推定器を学習している．特に，Wan ら [14] の手法では，Variational AutoEncoder によって得られる潜在空間から GANs の Generator への入力をサンプリングすることにより，シミュレーションによって得た骨格モデルの姿勢と，生成される画像とを結びつけ，学習に用いている．このように，GANs により生成した画像を用いて分類器を学習する方法は，学習データに含まれていない画像を生成しつつ学習ができるため，注目を集めている．

3. 姿勢を表現する多様体に基づく GANs

3.1 潜在空間における姿勢変化と GANs

本研究では，剛体である物体に対する，多様体に基づく姿勢推定手法について考える．ある剛体から一定の距離離れた地点から，その剛体の回転を観測した画像を撮影すると，その変化は剛体の回転，つまりその剛体に対するカメラの相対的な角度にのみ依存する．ここで，剛体の回転を 1 軸まわりの回転のみに限定すると，観測画像の変化はその回転軸まわりの回転角度にのみ依存する．そのため，適切な特徴空間を設計することが出来れば，ある軸の周りに回転した物体を観測した画像の系列は，その特徴空間中である閉曲線（1 次元多様体）上に分布すると考えられる（図 1）．同様に，剛体の回転が 2 軸の回転であれば，ある閉曲面（2 次元多様体）上に分布すると考えられる．

精度良く姿勢推定をするためには，ある軸回りに回転した物体を観測した画像の系列が，特徴空間の多様体上に均等に分布していることが望ましい．しかし，そのような特徴空間を直

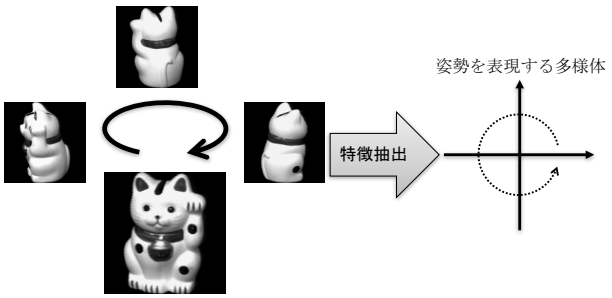

図 1　物体の回転と，特徴空間中での多様体.

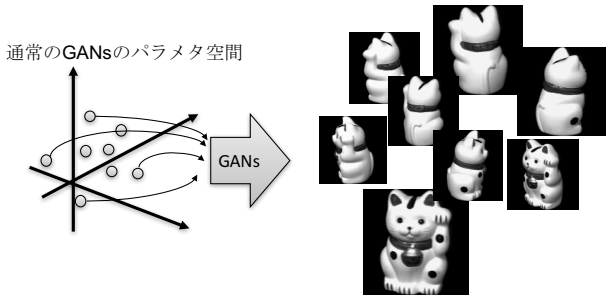

図 2　通常の GANs による画像生成.

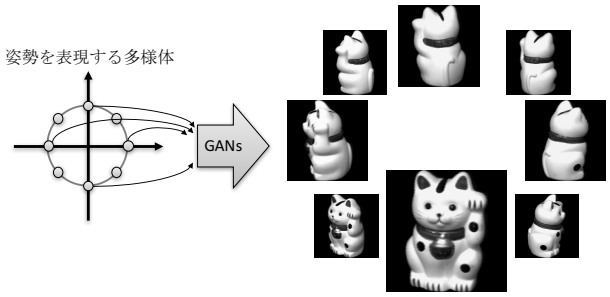

図 3　多様体上からのサンプリングに基づく GANs.

接学習することは難しい．そこで本研究では，多様体上に定義した一様分布からデータをサンプリングし，画像を生成する GANs を提案する．従来の GANs の Generator（図 2）とは異なり，サンプリングする変数を円周や球面といった剛体の姿勢変化を表現可能な多様体上に限定することで，サンプリングする乱数に物理的な意味を持たせることができる．

そのような GANs の Generator が学習できれば，多様体に沿ってデータをサンプリングすることにより，ある軸回りに剛体を回転させた画像を順番に生成することが可能になると考えられる（図 3）.

3.2　多様体上の確率分布に基づく画像生成

一般に GANs では，Generator G と Discriminator D を用意し，ある分布からサンプリングした $\mathbf{z} \sim p_z$ を G の入力として，D が区別出来ないようなデータ $\mathbf{x} = G(\mathbf{z})$ を出力するように学習する．

通常，変数 $\mathbf{z}$ は図 2 に示すように n 次元のパラメタ空間中の一様分布などからサンプリングするが，本研究では，図 3 の様に姿勢変化の多様体上に限定する．例えば，1 軸まわりの回転による姿勢変化を想定し，最も単純な 1 次元多様体である単位円でその姿勢変化を表現することを考えると，単位円周上の点 $\mathbf{z} = (z_1, z_2)$ は

$$z_1^2 + z_2^2 = 1 \tag{1}$$

と表せることから，これを制約としてサンプリングを行なう．具体的には，$[0, 2\pi)$ の区間で一様乱数 θ を生成し，$\mathbf{z} = (\cos\theta, \sin\theta)$ とする．

生成の際には，パラメタ θ を $[0, 2\pi)$ の区間で連続的に変化させることにより，物体の回転に対応したような，連続的に変化する画像を生成することができる．ただし，パラメタ θ と実際の物体の姿勢を紐付ける情報は与えられていないため，θ の値が必ずしも実際の物体の姿勢に対応するとは限らない．

3.3　姿勢推定への応用

3.1 節で述べた GANs を用いることで学習画像に含まれない姿勢の画像を補間して生成できるため，学習画像に加えて補間した画像で姿勢推定器を学習することにより，精度が高い姿勢推定器を得ることが期待できる．そこで，GANs において，学習用に用意した実画像か，Generator が生成した画像かを判定する本来のネットワークに加えて，物体の姿勢推定を行なうネットワークを追加し，それらを同時に学習するモデルを考える．

姿勢推定をするためには，姿勢の教師信号が付与された学習用画像が必要である．また，GANs により生成した画像についても姿勢の教師信号を与える必要がある．ここでは，GANs により画像を生成する際にサンプリングするパラメタ θ を教師信号とみなす．姿勢推定を行なうネットワークは，二宮ら [7] の TriNetR と同様に，周期性をもつ姿勢を表現するため，基準となる姿勢からの物体の回転角度の余弦，正弦（$\cos\theta, \sin\theta$）を出力するネットワークとする．

提案するネットワークでは，

- 生成した画像が実物か，生成物かの 2 クラス分類の誤差
- 推定した姿勢と教師信号との誤差

の両方を最小化するよう，マルチタスク学習により Discriminator $D(\mathbf{x})$ を学習する．この GANs を学習させることにより，GANs による画像の補間を行ないつつ，姿勢推定器を学習させられるようになることが期待される．

学習ができれば，得られた Discriminator D に対し，画像を入力することにより，姿勢推定結果を得ることができる．

4.　実　　験

4.1　データセット

物体を様々な方向から撮影したデータセットである COIL-20 データセット [15] を利用し，評価を行なった．COIL-20 データセットは，20 種類の物体を 5 度刻みで時計回りに回転させ，撮影したものであり，各画像の解像度は 128 画素四方である．図 4 にデータセットに含まれる画像の例を示す.

4.2　ネットワークの構成

GANs の構成は Radfod ら [10] の DCGANs を参考に，COIL-20 の画像サイズに合わせて 1 層追加し，100 次元の一様分布の代わりに 3.1 節で述べた 2 次元の $\mathbf{z}$ を入力とした．実験に用いたネットワーク構造を表 1，2 に示す.

図 4　COIL-20 データセットに含まれる画像の例.

表 1　Generator のネットワーク構造

Input	2
Fully-connect	32,768 units
Convolution g1	Kernel: 5×5 Channels: 512 Up sampling: 2×2
Convolution g2	Kernel: 5×5 Channels: 256 Up sampling: 2×2
Convolution g3	Kernel: 5×5 Channels: 128 Up sampling: 2×2
Convolution g4	Kernel: 5×5 Channels: 64 Up sampling: 2×2
Convolution g5	Kernel: 5×5 Channels: 1
Output	$128 \times 128 \times 1$

姿勢推定を行なう場合のネットワーク構成は，上記モデルに加え，Discriminator の Convolution 層を共有し，角度の余弦，正弦を出力する二宮ら [7] の TriNetR と同様の全結合層を追加した．追加した部分のネットワーク構造を表 3 に示す．

4.3　画像生成実験

まず，3.1 節で提案した，姿勢を表現する多様体に基づく GANs と，3.3 節で提案した，姿勢推定器を同時に学習するために姿勢を教師信号として与える GANs について，学習データに含まれない姿勢に対する画像の補間能力を評価した．

4.3.1　実 験 方 法

COIL-20 データセットに含まれる各物体について，0 度の姿勢から 10 度刻みで 36 枚の画像を選択し，GANs を学習した．次に，姿勢を表現する多様体上の $[0, 2\pi)$ の区間で均等に 144 点を選択し，画像を生成した．

4.3.2　結　　果

結果の一部を図 5 に示す．図 5 は，姿勢の教師信号を与えず生成したものであり，図 6 は，姿勢の教師信号を与えて生成したものである．これらの図は，最も左端の列が，角度 0 度（正面），90 度（左向き），180 度（後ろ向き），270 度（右向き）に対応する．教師信号を与えずに学習したものは，連続的に変化する画像を生成出来ているが，本来の物体の姿勢変化の方向（時計回り）とは異なる変化をしている．一方で，教師信号を与えることにより，物体の姿勢変化に応じた画像を生成出来ていることが確認できる．しかし，図 6 の左端の列を見ると，本来あるべき姿勢（図 7）よりも少しずれが生じていることが分かる．

表 2　Discriminator のネットワーク構造

Input	$128 \times 128 \times 1$
Convolution d1	Kernel: 3×3 Channels: 32 Max Pooling: 3×3
Convolution d2	Kernel: 3×3 Channels: 64 Max Pooling: 3×3
Convolution d3	Kernel: 3×3 Channels: 128 Max Pooling: 3×3
Convolution d4	Kernel: 3×3 Channels: 256 Max Pooling: 3×3
Fully-connect	256 units
Fully-connect	256 units
Output	2 units

表 3　追加のネットワーク構造（姿勢推定）

Input	$128 \times 128 \times 1$
Convolution d1–d4	
Fully-connect	256 units 重みに L_2 正則化
Fully-connect	256 units 重みに L_2 正則化
Output	2 units $(\cos\theta, \sin\theta)$

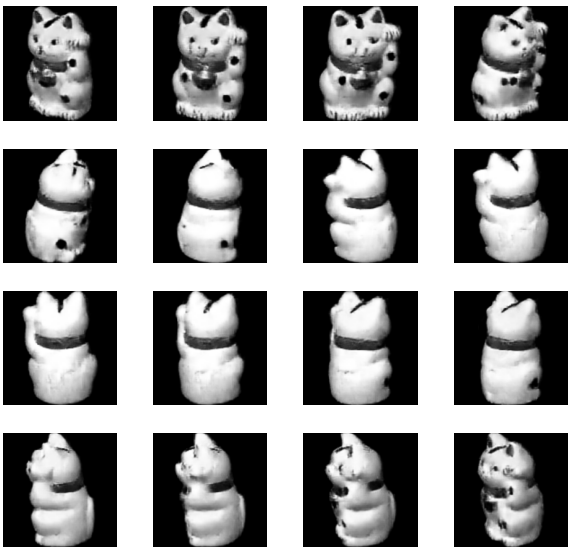

図 5　姿勢の教師信号を与えず学習した GANs により生成したもの.

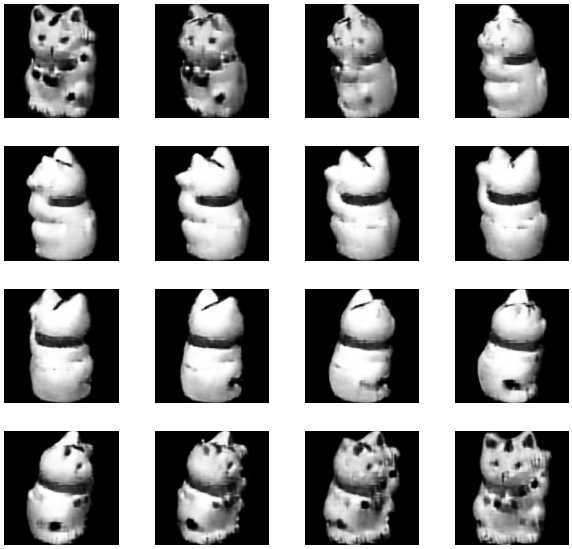

図 6　姿勢の教師信号を与えて学習した GANs により生成したもの.

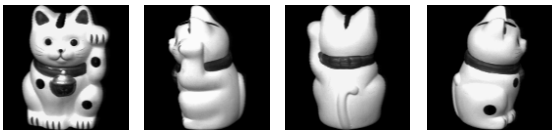

図 7　データセットから選択した画像.

表 4　姿勢推定結果

	10 度刻み	20 度刻み	40 度刻み
提案手法	11.33 度	12.77 度	27.19 度
比較手法	9.77 度	11.57 度	32.97 度

4.4　姿勢推定実験

次に，3.3 節で提案した，姿勢推定器を同時に学習する GANs について，学習データの数に応じた姿勢推定精度の変化を調べた.

4.4.1　実 験 方 法

COIL-20 データセットに含まれる各物体に対し，0 度の角度から 10 度刻み，20 度刻み，40 度刻みに画像を学習用画像として選択し，5 度の角度から 10 度刻みに選択した画像を評価に用いて姿勢推定の精度を平均絶対誤差で評価した.

比較手法として，GANs による画像生成をせず，提案手法で用いた Discriminator の姿勢推定部分からなる CNN を用いた．このネットワークは，入力となる対象物体の実画像のみを用いて学習した.

4.4.2　結　　果

まず，10 度刻みで学習した時の，学習の様子を図 8 に示す．判例中の，Classification loss 及び Orientation loss は，実画像か生成画像かの判定及び，物体の姿勢推定の，Discriminator と Generator のそれぞれにおける誤差を表す．生成した画像が本物と見分けがつかなくなり，姿勢推定誤差も小さくなることが分かる.

次に，姿勢推定結果を表 4 にまとめる.

提案手法が比較手法である通常の CNN による推定よりも精

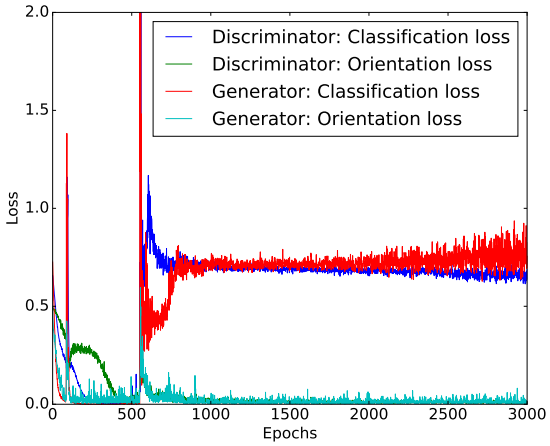

図 8　学習の過程.

度が低いのは，4.3.2 節で述べたように，生成した画像が本来の向きから少しずれることがあるため，提案手法は誤った姿勢を学習したためと考えられる．一方，学習画像の数が少なくなっても，比較手法に比べて姿勢推定精度の低下が小さいのは，GANs によって画像を補間することにより，学習画像の不足を補いながら推定器を学習出来ているからだと考えられる.

5.　お わ り に

本報告では，物体の姿勢変化は潜在的には多様体で表現可能であることに着目し，多様体上に定義した確率分布から画像を生成する GANs の枠組みと，それを用いて学習画像を補間しつつ姿勢推定器を学習する手法について検討を行なった.

実験では，教師信号を与えて学習することにより，物体の連続的な姿勢変化に応じた画像を生成できることを確認したが，真値からずれた画像が生成されてしまうことがあった．これは，画像の生成と姿勢推定を同時に学習させたことにより，まだうまく生成出来ていないデータを用いて姿勢推定を学習したことが原因の 1 つであると考えられる．今後の課題として，学習の方法及び，教師信号の与え方について検討する必要がある.

本報告では特定物体の姿勢推定について扱ったが，今後の課題として，例えば姿勢を表現する多様体を 2 次元平面中の円周から 3 次元以上の空間中の円環に変更することにより，同一クラスに属する様々な物体の姿勢変化を扱えるような拡張が考えられる.

謝辞　本研究の一部は，科学研究費補助金による.

文　　献

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems 27, pp.2672–2680, Dec. 2014.

[2] N. Correll, K.E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J.M. Romano, and P.R. Wurman, “Lessons from the Amazon picking challenge,” arXiv preprint arXiv:1601.05484, pp.1–14, Jan. 2016.

[3] J. Broekens, M. Heerink, and H. Rosendal, “Assistive social robots in elderly care: A review,” Gerontechnology, vol.8, no.2, pp.94–103, April 2009.

[4] H. Murase and S.K. Nayar, “Visual learning and recognition of 3-D objects from appearance,” Int. J. of Comput. Vision, vol.14, no.1, pp.5–24, Jan. 1995.
[5] S. Gupta, P. Arbelaez, R. Girshick, and J. Malik, “Aligning 3D models to RGB-D images of cluttered scenes,” Proc. IEEE Conf. on Comput. Vision and Pattern Recognit., pp.4731–4740, June 2015.
[6] M. Torki and A. Elgammal, “Regression from local features for viewpoint and pose estimation,” Proc. 2011 Int. Conf. on Comput. Vision, pp.2603–2610, Nov. 2011.
[7] 二宮宏史，川西康友，出口大輔，井手一郎，村瀬　洋，小堀訓成，橋本国松，“深層学習を用いた多様体構築による 3 次元物体の姿勢推定に関する予備検討，” 信学技報，第 116 巻，pp.25–30，Jan. 2016.
[8] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems 25, pp.1097–1105, Dec. 2012.
[9] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi, “Deep filter banks for texture recognition, description, and segmentation,” Int. J. of Comput. Vision, vol.118, no.1, pp.65–94, May 2016.
[10] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, pp.1–16, Nov. 2015.
[11] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, pp.1–7, Nov. 2014.
[12] P. Isola, J.-Y. Zhu, T. Zhou, and A.A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint arXiv:1611.07004, pp.1–16, Nov. 2016.
[13] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, “Learning from simulated and unsupervised images through adversarial training,” Proc. IEEE Conf. on Comput. Vision and Pattern Recognit., pp.2107–2116, July 2017.
[14] C. Wan, T. Probst, L. Van Gool, and A. Yao, “Crossing nets: Combining GANs and VAEs with a shared latent space for hand pose estimation,” Proc. IEEE Conf. on Comput. Vision and Pattern Recognit., pp.680–689, July 2017.
[15] S.A. Nene, S.K. Nayar, and H. Murase, “Columbia object image library (COIL-20),” Technical Report CUCS-005-96, Department of Computer Science, Columbia University, Feb. 1996.