姿勢を表現する多様体に基づく GANs を用いた特定クラス物体姿勢推定の検討

川西 康友[1,a)]　出口 大輔[1,b)]　井手 一郎[1,c)]　村瀬 洋[1,d)]

概要

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，提案手法により，同一クラスに属する，学習データに含まれない物体の姿勢が補間できることを確認し，また，生成した画像を用いて姿勢推定器を学習することにより，高精度な姿勢推定ができることを示す．

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．ロボットが物体を把持することは，ロボットの様々な応用における共通の課題であり，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，物体の姿勢推定を行なう必要がある．本研究では，対象物体として，マグカップなどの，ある物体クラスを想定し，そのクラスに属する様々な形状の物体の姿勢を精度よく推定することを目指す．

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，Xtion や RealSense などに代表される 3 次元ビジョンセンサが一般的である．近年，こうしたセンサは小型化・低価格化が進んでおり，様々なところで利用されつつある．

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．

[1] 名古屋大学
[a)] kawanishi@i.nagoya-u.ac.jp
[b)] ddeguchi@nagoya-u.jp
[c)] ide@i.nagoya-u.ac.jp
[d)] murase@i.nagoya-u.ac.jp

本研究では，この GANs を用いて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ことを考える．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 [2] 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 により画像を生成する手法と，それを用いて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

1 枚の画像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 [3]，回帰モデルに基づく手法 [4]，テンプレートマッチングに基づく手法 [2] などがある．通常，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が，精度良く姿勢推定をするためにはあらゆる姿勢のテンプレートを保持しておく必要がある．この問題に対し，Murase ら [2] 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ている．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，その低次元空間へ投影したテンプレートの系列を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．そのため，保持すべきテンプレート数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 [5] 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．この手法では，姿勢を教師信号として学習した Deep Convolutional Neural Network（DCNN）[6] 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．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 2 つの姿勢の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対であり，Generator は Discriminator を騙せるように，Discriminator は騙されないように，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ことができるようになる．

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）[7] 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．また，与えられた条件に応じた画像生成を行う Conditional GANs（CGANs）[8] や，クラス分類を同時に学習する Auxiliary Classifier GANs（ACGANs）[9] など，様々な拡張が提案されている．CGANs 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，応用として画像から画像への変換 [10] などがある．

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 [11]，Wan ら [12] 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する．特に，Wan ら [12] 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．

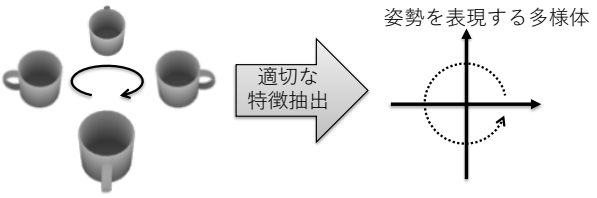

図 1 物体の回転と，特徴空間中での多様体．

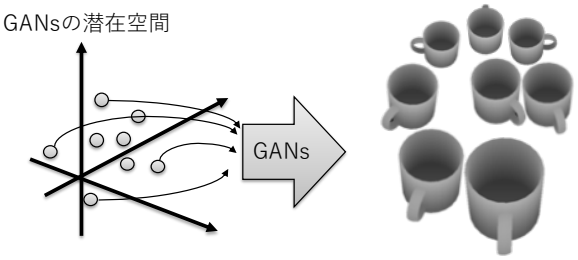

図 2 通常の GANs による画像生成．

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，ある特定クラスに属する物体に対する姿勢推定手法について考える．パラメトリック固有空間法 [2] の考え方に基づくと，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できれば，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回転角に応じて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確率分布からデータをサンプリングし，画像を生成する GANs を考える．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）．

ある 1 つの特定物体の姿勢推定であれば，この多様体を用

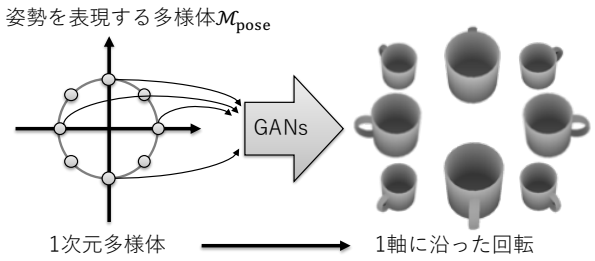

図 3 多様体上からのサンプリングに基づく GANs.

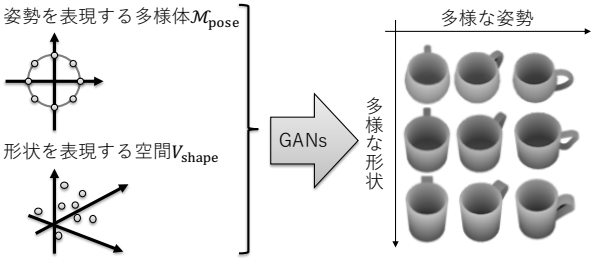

図 4 姿勢と形状の空間からサンプリングする GANs.

いて姿勢変化を表現するだけで十分であるが，ある特定クラスに属する多様な形状の物体を対象とするには不十分である．そこで，そのクラスに属する異なる形状の物体を表現するため，形状を表現する空間 V_{shape} と，姿勢を表現する多様体 $\mathcal{M}_{\text{pose}}$ の直積で表せる多様体 $\mathcal{M} = V_{\text{shape}} \otimes \mathcal{M}_{\text{pose}}$ を設計し，その多様体 $\mathcal{M}$ からサンプリングすることにより，多様な形状の物体の，多様な姿勢の画像の生成を実現する（図 4）.

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 が本物の学習データと区別できないようなデータ $\mathbf{x} = G(\mathbf{z})$ を出力するように学習する．

本研究では，形状を表す空間と，姿勢を表す多様体の直積空間 $\mathcal{M}$ 上に限定し，サンプリングをする（図 4）．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 $\mathbf{z}_{\text{pose}} = (z_1, z_2)$ は

$$z_1^2 + z_2^2 = 1, (z_1, z_2 \in [0, 1]) \tag{1}$$

と表せることから，この単位円を $\mathcal{M}_{\text{pose}}$ とする．具体的には，$[0, 2\pi)$ の区間で一様乱数 θ を生成し，$\mathbf{z}_{\text{pose}} = (\cos\theta, \sin\theta)$ とすることで，多様体 $\mathcal{M}_{\text{pose}}$ からのサンプリングを実現する．一方，形状については，その変化を明示的に表現することは難しいため，単純に N 次元空間 $V = \mathbb{R}^N$ 中の Gaussian 分布からサンプリング（$\mathbf{z}_{\text{shape}} \in V$）を行う．結果的に，$\mathbf{z} = (\mathbf{z}_{\text{shape}}, \mathbf{z}_{\text{pose}})$ を，Generator G の入力とする．

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．また，$\mathbf{z}_{\text{shape}}$ を変化させることにより，様々な形状の物体の画像を生成することができる．

3.3 姿勢に応じた画像生成

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・形状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，物体の姿勢推定をする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 $\mathbf{z}_{\text{pose}}$ を教師信号とみなす．姿勢推定をするネットワークは，二宮ら [5] 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，

- 生成した画像がどの方向の画像としての実物か，生成物かの $P \times 2$ 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，姿勢に応じた画像を生成することができるようになると期待できる．

3.4 姿勢の推定

Discriminator D の学習ができれば，得られた D に対して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．しかし，この姿勢推定器は GANs の学習途中の画像も学習に用いているため，高精度な推定は難しい．

そこで，向き推定だけをする推定器を別途用意し，学習データに加えて Generator G が生成する画像を用いて姿勢推定器を学習する．

4. 実験

4.1 データセット

CAD モデルをもとに，データセットを構築した．CAD モデルとして，3 次元物体のデータセットである ShapeNet [13] から，133 個のマグカップを選んだ．そのうち，100 個を学習，残りの 33 個を評価に用いた．

CAD モデルから深度画像を生成するため，レンダリングを行った．このとき，学習データについては，Z 軸ま

表 1　姿勢推定結果（平均絶対誤差）

仰角	Baseline	DA	Proposed
0°	9.39°	8.78°	**1.56°**
30°	10.29°	7.86°	**2.37°**
45°	7.95°	5.00°	**2.68°**
60°	4.75°	4.66°	**4.25°**

わりに $0, 10, \ldots, 350°$，評価データについては，Z 軸まわりに $5, 15, \ldots, 355°$ 回転させながらレンダリングを行った．結果として，学習データ 36 枚 × 100 物体，評価データ 36 枚 × 33 物体 を得た．物体に対する仰角を $0, 30, 45, 60°$ の 4 種類設定し，それぞれデータセットを作成した．ただし，仰角 0° は物体を真横から観測したものであり，値が大きくなるほど上から見下ろすような観測となる．

4.2　ネットワークの構成

GANs の構成は Radfod ら [7] の DCGANs を参考に，データセットの画像サイズに合わせて 1 層追加し，Generator G の入力は $\mathcal{M}$ 上の確率分布からサンプリングした $100 + 2$ 次元の $\mathbf{z}$ とした．また，Discriminator D の出力を，$P(= 36) \times 2$ クラスの分類問題とした．

さらに，姿勢推定を行なう場合のネットワークとして，上記モデルに加え，Discriminator D の Convolution 層を共有し，角度の余弦，正弦を出力する二宮ら [5] の TriNetR と同様の全結合層を追加した．

4.3　比較手法

二宮ら [5] の TriNetR による姿勢推定器を，学習画像のみを用いて学習したもの（Baseline），画像を移動・拡縮してデータ拡張を行った画像を用いて学習したもの（DA），学習画像と GAN によって生成した画像を用いて学習したもの（Proposed）を比較した．

4.3.1　実験結果

姿勢推定結果の平均絶対誤差を表 1 にまとめる．

仰角が小さいデータセットでは，ほぼ真横から観測しており，姿勢の区別がつきにくい姿勢が存在する．そのため，全体的に姿勢推定精度は悪い．しかし，どの仰角においても，提案手法である，姿勢変化を表現する多様体に基づく GANs により生成した画像を用いて姿勢推定器を学習した結果，高精度に姿勢推定ができることが確認できた．

5.　おわりに

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，それを用いて生成した画像を用いて姿勢推定器を学習する手法を提案した．

実験では，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，生成した画像を追加して用いて姿勢推定器を別途学習することにより，もとの学習データのみで姿勢推定器を学習するよりも，姿勢推定精度が大幅に向上することが確認できた．

本報告では 1 軸回転を表現する 2 次元空間中の 1 次元多様体を対象としたが，今後の課題として，姿勢を表現する多様体を超球面にすることにより，2 軸や 3 軸の回転による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．

参考文献

[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] 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.

[3] 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.

[4] M. Torki and A. Elgammal, “Regression from local features for viewpoint and pose estimation,” Proc. 13th Int. Conf. on Comput. Vision, pp.2603–2610, Nov. 2011.

[5] 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，PRMU2016-39，pp.25–30，Jan. 2016.

[6] 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.

[7] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” arXiv preprint arXiv:1511.06434, Nov. 2015.

[8] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[9] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier GANs,” arXiv preprint arXiv:1610.09585, 2016.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A.A. Efros, “Image-to-image translation with conditional adversarial networks,” arXiv preprint arXiv:1611.07004, Nov. 2016.

[11] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, and R. Webb, “Learning from simulated and unsupervised images through adversarial training,” Proc. 30th IEEE Conf. on Comput. Vision and Pattern Recognit., pp.2107–2116, July 2017.

[12] 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. 30th IEEE Conf. on Comput. Vision and Pattern Recognit., pp.680–689, July 2017.

[13] A.X. Chang, T.A. Funkhouser, L.J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, “Shapenet: An information-rich 3D model repository,” arXiv preprint arXiv:1512.03012, 2015.