情報処理学会論文誌 Vol.0 No.0 1–7 (??? 1959)

生成的特徴量の角度依存性に着目した視点角度推定の精度向上

チェン マーク[1,a)] 川上 玲[2,3,b)] 佐藤 育郎[2,3,c)] 苗村 健[1,d)]

受付日 xxxx年0月xx日, 採録日 xxxx年0月xx日

概要：画像の視点角度推定は，ロボットの行動決定や物体認識精度の向上に役立つ．視点角度推定の際，教師データを用いて角度の回帰を学習すると同時に，生成タスクを補助的に使用し，推定精度を向上させる従来法が存在する．これは，ある画像を入力した際，そのインスタンスを別の視点から見た画像として再構成するタスクを補助的に使うことで，視点角度によって見え方が変化する幾何構造の学習が進み，より正確な視点角度推定が可能になることに基づく．本稿では，視点角度推定の回帰に必要な情報のみを効率よく逆伝播させるため，上記の生成的なデコーダに入力する潜在特徴を，視点角度によって変化するものと不変のものに分割する．また，用いる画像を左右反転させたものも学習に用いるデータ拡張と自己教師あり学習を導入する．これらの工夫により，特に教師ラベルが少ない場合に，既存手法の角度推定の精度を大幅に改善できることを実験的に示す．

キーワード：視点角度推定，オートエンコーダ

Improving the Accuracy of Viewpoint Angle Estimation by Focusing on the Angle Dependence of Generative Features

Mark Chen[1,a)] Rei Kawakami[2,3,b)] Ikuro Sato[2,3,c)] Takeshi Naemura[1,d)]

Received: xx xx, xxxx, Accepted: xx xx, xxxx

Abstract: Estimating the viewpoint angle of an image is useful for determining robot actions and improving the accuracy of object recognition. An existing method improves the accuracy of viewpoint-angle estimation by learning angle regression with supervised data and using a generation task as an auxiliary. The generation task, which reconstructs an instance of a given image as an image viewed from a different viewpoint, enhances the learning of the geometric structure caused by the viewpoint changes. In this paper, we improve the method by dividing the latent features into those that are variant and invariant to the viewpoint changes. This can efficiently back-propagate only the information necessary for the regression of viewpoint angle estimation. The latent features input to the generative decoder for the generation task are divided into those that vary with viewpoint angle and those that are invariant. In addition, we introduce a data augmentation that exploits the symmetric characteristics of the objects in the images. We experimentally show that these ideas can significantly improve the accuracy of angle estimation of the existing method, particularly when there is little supervision in training.

Keywords: View Point Estimation, AutoEncoder

1. はじめに

画像におけるカメラの視点角度推定はコンピュータビジョンの分野の中でも重要なタスクの一つである．例えば，産業用ロボットを用いるとき，物体を正しく把持する

[1] 東京大学大学院情報理工学系研究科
[2] 東京工業大学
[3] デンソー IT ラボラトリ
[a)] mchen@nae-lab.org
[b)] reikawa@c.titech.ac.jp
[c)] isato@c.titech.ac.jp
[d)] naemura@nae-lab.org

ためには対象となる物体の向き，つまりカメラが物体をどの角度から撮っているかを正しく推定する必要がある．また，自動運転の分野では，事故を防ぐために他車両の運動予測が必須であるが，他車両の進行を予測するには車の向きを正しく推定する必要がある．

このように視点角度推定には様々な応用があり，古くから研究されている．初期では物体にマーカーをつけることで，カメラに映るマーカーの位置関係から画像内での物体の位置と角度，つまりカメラの視点角度を求めていた．その後，Structure from Motion (SfM) のように，マーカーなどがなくとも複数の視点から取られた画像を用いて 3 次元構造とそれぞれの画像の視点角度を推定する方法が現れた．CNN (Convolutional Neural Networks) の発展以降は，視点角度推定をパターン認識問題と捉え学習ベースで視点角度推定を行う方法が台頭し，単一の画像からでも視点角度を推定できるようになりつつある．視点角度の教師データを極力減らした半教師あり学習の設定に関しても研究が行われている [1].

本稿では，単一の画像に対して視点角度を推定できる，学習ベースの手法を提案する．つまり，テスト時には複数の画像を入力するのではなく，一枚の画像のみを入力し，その画像に対しての視点角度を推定する．視点角度は図 1 に示すようにクラスごとに定義された座標系により記述することとする．具体的には視点角度は，カメラの光学中心の方位角 θ ならびに仰角 ϕ の 2 つのスカラーによって表す．提案手法の入力は RGB 画像とする．ここでは簡単のため一枚の画像には一個の物体が存在することとする．学習データには多種類の物体がそれぞれ複数の視点角度から撮像された画像が含まれることとする．出力の真値として真の視点角度 (θ, ϕ) を与える．学習時，真値が全部のデータに対して与えられているケースと，一部のデータに対して与えられているケースを評価する．

単一画像からの視点推定問題に対して，生成的タスクを補助的に用いることで視点角度推定の精度を改善できることを示した Mariotti らの研究 [1] がある．この研究では，ある画像の視点角度を CNN と全結合層によって回帰するタスクに加えて，生成的タスクを補助的に導入する．この生成タスクは，回帰ネットワークへの入力画像 A にある物体を別の視点角度から撮像した画像 A' を入力として，画像 A を生成するものである．ただし，生成ネットワークには回帰ネットワークが出力する角度情報を入力することとする．このような視点角度依存の生成タスクの利用により，特に半教師あり学習の設定において回帰精度を改善させられることが示されている．

一般的に，画像生成タスクにおいて有効な潜在表現は 3 次元的な位置や角度に依存した成分と，それらに非依存な成分の 2 種類に分けられる．前者には物体の三次元構造に関する情報が，後者には大域的な色情報やクラス情報などが含まれる．ところが，Mariotti らの手法では生成的タスクにおいて前者の 3 次元の依存性をもつ特徴量のみが使用されている．

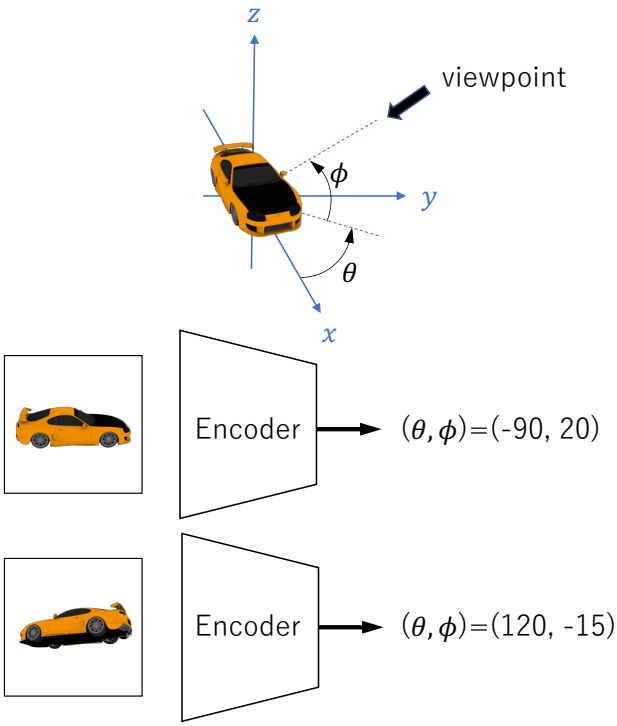


図 1: 視点角度推定の概要．クラスに対して標準となる座標の位置を定め，その座標におけるカメラ位置を推定する．図では方位角 θ と仰角 ϕ を推定している．

Fig. 1 Overview of viewpoint angle estimation. Define the position of the standard coordinates for the class and estimate the camera position at those coordinates. In the figure, the azimuth angle θ and elevation angle ϕ are estimated.

そこで，本研究では視点角度に非依存の特徴も潜在表現に明示的に加えることで，生成ネットワークの記述能力が改善されるとともに，回帰タスクの精度向上が期待できると考えた．また，車や椅子などおおよその左右対称性を持つデータにおいては，左右反転画像も同じインスタンスとして見ることができる．そこで，学習に用いる画像の左右反転画像も用いるデータ拡張と自己教師あり学習を導入した．このように潜在表現の分割，データ拡張の導入をしたところ，視点角度推定の精度を向上させることができた．

2. 関連研究

2.1 視点角度推定

視点角度推定には，SfM [2], [3], [4] を用いるなどテスト時に同じインスタンスについて複数の画像を入力し計測的に行う方法と，テスト時に一つずつ画像を入力しそれぞれについて推定し認識的にシングルショットで行う方法が存在する．さらに，連続的な視点角度を推定する回帰問題として視点角度推定を扱う方法 [1], [5] と，視点をいくつかの

ビンに分けて分類問題として扱う方法 [6], [7] がある．近年では特に教師あり学習での視点角度推定において十分に高い精度が得られるようになってきている．教師あり学習では，例えば視点角度・バウンディングボックス・特徴点のラベルを用いた特徴点検出タスクと併せて行う方法 [8] や RGB 値に加えて深度情報も用いる方法 [9]，一回転で元に戻る視点角度のビンに対応してシリンダー状の畳み込み層を用いる方法 [10] などがある．

確かにこれらの教師ありの手法では高い精度が実現されているものの，視点角度などのラベル付けには莫大なコストがかかり，ラベルがついているデータも限られているため，より少ない情報を学習に用いる方法に対する需要が大きい．そこで，視点角度のラベルをなるべく使わない方法や，視点角度以外の情報の使用を抑えるような方法が研究されている．より少ない視点角度のラベルを用いる研究では，例えば [6] は視点角度のラベルは完全に用いずに，予め定めた複数の視点からの画像を予め定めた順番で入力し，クラス尤度を表すベクトルがワンホットに近くなるように学習する．また，[1] は条件付きオートエンコーダによる生成的タスクを補助的に用いることで視点角度に対する半強師あり学習の精度を向上させている．一方，視点角度以外の学習情報の削減では，本物の画像に近い合成画像を生成し，生成した画像のみによる学習によって本物の画像を用いたテストで比較的高い精度を実現することでラベル付き画像のコストの問題を解決する [7], [11] ことや，画像の深度情報を使わず RGB 値のみを用いる [12] ことが考えられる．

また，視点角度推定問題と他の認識問題に相互作用が期待されるとして，他の問題とのマルチタスクで解く研究もなされている．視点角度推定でのマルチタスクの導入の例としては，物体のクラス分類 [10], [13]，特徴点抽出 [8], [14]，物体検出 [10], [15]，画像再構成 [1], [16]，3D モデル再構成 [17], [18] が挙げられる．本研究はこの中でも画像再構成と併せて視点角度推定問題をマルチタスクで解く方法にあたる．

2.2 生成的タスクを用いた視点角度推定

近年，CNN は 3 次元構造をうまく理解しておらず，無理やり 2 次元で解釈しているのではないかという議論がなされている．具体的には，認識において CNN は形状よりもテクスチャ情報によって物体のクラス問題を解いているのではないかと考えられている [19]．そこで，人間の脳のように 2 次元から 3 次元を知覚する生成的なタスクと認識的なタスクという本来同時に解くと性能が下がるものをうまく相互に補助させて性能を上げる研究がなされている [20]．視点角度推定でもそのような動機の研究 [1], [21] がなされており，その中でも条件付きオートエンコーダを用いて視点角度の回帰を補助する研究 [1] を紹介する．

この研究では，3D モデルからレンダリングされた画像を学習に用い，視点角度の真値を教師あり，ないしは半教師あり学習の要領で利用する．学習の際，視点角度推定のネットワークに入力する画像 I と条件付きオートエンコーダに入力する画像 I' の 2 枚の画像を用いる．ここで，I と I' は同じインスタンスの異なる視点角度からの画像である．図 2b のようなネットワークにおいて上の Encoder では $f_v(I) = v$ によって画像 I における視点角度 v を回帰する．つまり，ラベル付きデータとして真値を用いる画像にのみ生じる視点角度推定の損失 $\mathcal{L}_r$ は

$$\mathcal{L}_r = \|f_v(I) - v\|^2 \tag{1}$$

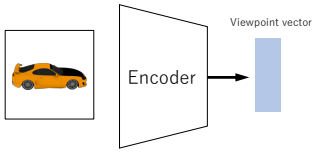
となる．再構成側では，下の Encoder によって得られた $3 \times n$ の潜在表現 $f_e(I')$ に回帰された視点角度 v から得られる 3×3 の回転行列 $R(f_v(I))$ をかけ，デコーダ f_d に入力することで視点角度推定のエンコーダに入力した画像 I を再構成する．つまり，再構成側の損失 $\mathcal{L}_g$ は

$$\mathcal{L}_g = \|f_d(R(f_v(I))f_e(I')) - I'\|^2 \tag{2}$$

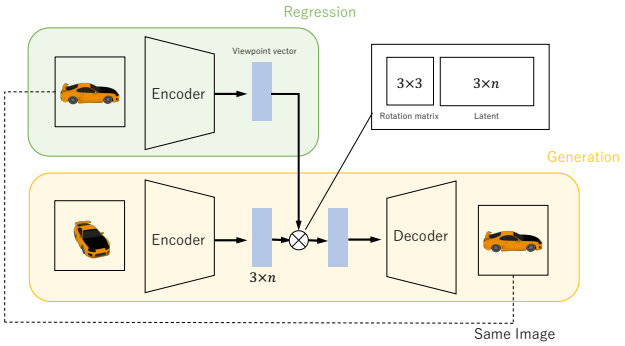
となる．よって全体の損失は式 1 と式 2 をあわせたものとなるので，真値 v がラベルとして与えられるときに 1，与えられないときに 0 となる変数 $[v \neq \emptyset]$ を用いて

$$\mathcal{L} = [v \neq \emptyset]\mathcal{L}r + \mathcal{L}_g \tag{3}$$

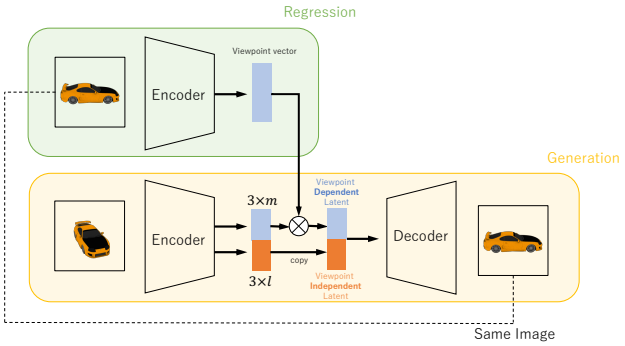
と表される．これによって視点角度推定の精度は式 1 のような視点角度の回帰のみを行うときよりも大幅に改善された．また，用いるラベル付き画像の数が少ない場合であるほど画像再構成の生成タスクによる精度上昇の幅が大きくなり，かつ全ての画像の視点角度のラベルを用いる場合でも精度が上昇することも確認された．



(a) 視点角度推定における回帰モデル



(b) 生成的タスクを導入した視点角度推定モデル



(c) 提案手法のモデル

図 2: **視点角度推定に用いるネットワークの概要．**図 2a は視点角度推定を回帰のみで行うネットワーク，図 2b は回帰に生成的タスクを補助的に加えたネットワーク，図 2c はオートエンコーダの潜在表現を視点角度依存性の有無によって分割したネットワークを示す．図 2b，2c では回帰で得られた視点角度を 3×3 の回転行列に変形し，回転行列を生成タスクでのエンコーダから得られた潜在変数にかけることで 3 次元空間での潜在表現を回転させている．回転させた潜在表現を用いて画像再構成を行うことで，画像再構成の損失を回帰のネットワークに伝播させている．

Fig. 2 Overview of the networks used for viewpoint angle estimation. Figure 2a shows a network where the viewpoint angle estimation is performed by regression alone. Figure 2b shows a network where regression is supplemented by a generative task. Figure 2c shows a network where the latent representation of the autoencoder is separated according to the presence or absence of 3D rotation dependency.In Figures 2b and 2c, the viewpoint angles obtained from the regression are transformed into a 3×3 rotation matrix, and the latent representation in 3D space is rotated by multiplying the rotation matrix by the latent variable obtained from the encoder in the generation task. By performing image generation using the rotated latent representation, the loss of the image generation is propagated to the network of regression.

3. 提案手法

第 2.2 節の手法では生成タスクを補助的に使用することで推定精度を向上させているが，生成的なデコーダに入力する潜在表現は回転によって全て変化を受けてしまう．そこで，本研究では潜在表現において視点角度に依存する特徴と視点角度に非依存の特徴に分割する．具体的には，図 2c のように潜在特徴の一部にのみ回転をかける手法を提案する．つまり，回転をかける Shape を表現するような潜在特徴の一部を $\mathrm{L}_s \in \mathbb{R}^{3m}$，回転をかけない Appearance を表すような潜在特徴の一部を $\mathrm{L}_a \in \mathbb{R}^{3l}$ とするとこれらは

$$\begin{pmatrix} \mathrm{L_s} \\ \mathrm{L_a} \end{pmatrix} = f_e(I') \tag{4}$$

で得られるので，画像再構成の損失は

$$\mathcal{L}_g = \|f_d(R(f_v(I))\mathrm{L}_s^{(3\times m)}, \mathrm{L}_a^{(3\times l)}) - I'\|^2 \tag{5}$$

と表される．なお，ここで $\mathrm{L}^{(\mathrm{a}\times\mathrm{b})}$ は $\mathbf{L}$ を $a \times b$ の行列に変形 (reshape) したものを表す．

また，今回用いるデータに含まれるインスタンスはそのほとんどがおおよその左右対称性を持つ物体であるため，左右反転した画像も同一のインスタンスとして扱うことができる．ここで左右対称性とは，図 1 に示す y 軸の符号の入れ換えに対して形状がほぼ変化しないことを表す．そこで，左右反転画像も学習に用いるデータ拡張と自己教師あり学習も潜在表現の分割に加えて導入する．左右反転画像も学習に用いることで，それぞれのインスタンスにおいて異なる視点角度からの画像の数を 2 倍に増やすことができる．一部の画像には，搭載している装備が左右どちらかに偏っている車両など左右対称でない物体も存在するが，それが左右反転していても視点角度を推定できるべきであるので影響は少ないと考えられる．視点角度推定のネットワークでは，元の画像がラベル付きデータとして用いられる場合は元の画像真値を左右反転したものを真値として用い，ラベルなしデータとして用いられる場合は視点角度の推定結果が互いに左右反転するようになるように自己教師ありの損失をかける．画像 I の左右反転画像を I_{inv}，視点角度 v を左右反転させる操作を $T(v)$ と表すと，操作 $T(v)$ は v の方位角が θ，仰角が ϕ のとき仰角は変えずに方位角の符号を反転させる操作，つまり $(\theta, \phi) \to (-\theta, \phi)$ を表す．すると，画像 I, I_{inv} の視点角度推定の損失の合計は，視点角度推定における教師あり学習と自己教師あり学習の重みを調整するパラメータ λ を用いて

$$\begin{aligned}\mathcal{L}_r = [v \neq \emptyset](\|f_v(I) - v\|^2 + \|f_v(I_{inv}) - T(v)\|^2) \\ +\lambda[v = \emptyset]\|f_v(I) - T(f_v(I_{inv}))\|^2\end{aligned} \tag{6}$$

と表される．また，画像 I_{inv} についても画像 I と同じ様に，式 5 と同様に表されるような画像 I' を生成的なエンコーダの入力に用いた再構成損失を用いる．

4. 実験

視点角度推定の実験の詳細と結果を以下に記す．評価指標は，視点角度の推定値と真値の差の絶対値が 30° 以下であることを正答と見なしたときの正答率と置く．これは視点角度推定の分野においてよく用いられている指標である [1], [10], [17].

4.1 データセット

実験には ShapeNet [22] を用いる．ShapeNet は，55 のクラスを含み合計 51,300 の 3D CAD モデルからなるが，本実験では 3,676 のインスタンスを含むクラス car のデータを用いる．各画像のレンダリングに用いるカメラ視点の方位角は −180°〜180°，仰角は −20°〜40° からランダムにそれぞれ数値を選び，各インスタンスに付き 10 枚の画像をレンダリングする．また，学習・検証・テスト用にそれぞれインスタンスの 70, 10, 20% を用いる．さらに，半教師あり学習での結果を得るため，レンダリング画像をラベルつきデータとして扱うインスタンスの割合を 1, 10, 100% と変化させて実験する．つまり，ラベルの割合が 10% のとき，学習に用いる 2,573 のインスタンスのうち 257 のインスタンスからレンダリングされた 2,570 枚の画像についてはラベルを用いて学習し，残りの 2,316 のインスタンスからレンダリングされた 2,3160 枚の画像についてはラベルを用いずに学習する．

4.2 実装

まず視点角度推定のネットワークのみを学習し，次に視点角度推定のネットワークは固定した状態で再構成のネットワークを学習する．最後に全体を fine-tuning する．視点角度推定用のエンコーダは conv 層と maxpooling 層からなるブロックが 6 つと，その後ろの 2 層の conv 層から構成されている．また，視点角度の損失には 3 次元空間内でのカメラ位置の座標に対する MSE を用いている．再構成用のエンコーダは 5 つのブロックからなり，それぞれのブロックには 2 層の conv 層が含まれ，2 層目のストライドを 2 にすることで空間解像度を下げている．再構成用のデコーダは，再構成用のエンコーダを逆にした構成である．また，画像再構成の損失には Parceptual loss [23] を用いている．

4.3 実験結果

潜在表現の分割と，左右反転画像を利用したデータ拡張を導入した実験結果を表 1 に示す．左右反転画像を利用したデータ拡張を用いた場合も従来手法より精度が高くなり，分割する手法を併用した場合が最も精度が高いことがわかる．

また，潜在表現を 3D 回転をさせる部分，させない部分に分けるときの分割の比率による精度の変化を表 2 に示す．潜在変数全体に 3D 回転をかける従来手法と比べて，半分のみに 3D 回転をかける場合のほうが全てのラベル率，全ての分割比率のもとで精度が高くなっていることが分かる．特に 1:1 で分割させる場合が全てのラベル率において最も精度が高くなることが分かる．この結果から，表 1 の実験では潜在表現の分割は 1:1 の比率で行っている．

表 1: **視点角度推定精度．**従来法に比べて，潜在表現の分割 (Separete) とデータ拡張 (Data Augmentation) をそれぞれ導入した場合の両方で精度が上がっており，潜在表現の分割とデータ拡張を同時に導入した場合 (Separate & Data Augmentation) が最も精度が高いことが分かる．*著者らによる再現.

Table 1 Accuracies of viewpoint angle estimation. Compared to the conventional method, the accuracy is improved in both cases where latent representation separation and data augmentation are introduced respectively, and the highest accuracy is obtained when latent representation separation and data augmentation are introduced simultaneously. *Reproduction by the authors.

	100%	10%	1%
Mariotti+[1]*	93.6	89.3	77.9
Separate	93.7	89.8	79.2
Data Augmentation	95.3	91.1	82.8
Separate & Data Augmentation	**95.3**	**91.4**	**83.3**

表 2: **潜在表現の構成比率と視点角度推定精度との関係．**比率は，（3D 回転を適用するベクトルの要素数：3D 回転を適用しないベクトルの要素数）を表す．1:1 の比率のときに一番精度が良くなっていることが分かる．なお，ここでの比較では，左右反転画像を用いたデータ拡張と自己教師あり学習は適用せず，潜在表現の分割のみを適用させている．*著者らによる再現.

Table 2 The relationship between the compositions of the latent representations and accuracies of viewpoint angle estimation. It can be seen that the accuracy is best when the ratio is 1:1. Note that in this comparison, data expansion and self-supervised learning using left-right reversed images are not applied, only latent representation segmentation is applied. *Reproduction by the authors.

	100%	10%	1%
Mariotti+[1]*	93.3	89.3	77.9
Ours (3:1)	**93.7**	**89.8**	78.5
Ours (1:1)	**93.7**	**89.8**	**79.2**
Ours (1:3)	**93.7**	89.3	78.5

4.4　考察

3D 回転を一部のみにかけることによって精度が改善したことは，視点角度に非依存の特徴も画像再構成に用いることができ，さらに視点角度に依存する，つまり視点角度推定に用いる特徴での損失のみを効率的に視点角度推定のネットワークに伝播させることができるからと説明できる．図 3 にテストでのいくつかの再構成の例を示す．それぞれの行の 5 つの画像は，左からそれぞれ f_v の入力，f_e の入力，f_d の出力，視点角度に非依存の潜在変数を 0 で置き換え視点角度に依存する潜在変数のみを用いたときの f_d の出力，視点角度に依存する潜在変数を 0 で置き換え視点角度に非依存の潜在変数のみを用いたときの f_d の出力を表す．例えば上から 1,3 行目の例がわかりやすいが，視点角度に依存する潜在変数のみを用いた場合は不適切な色かつ車のデティールが異なる画像が出力され，視点角度に非依存の潜在変数のみを用いたときは色は適切であるが形が全く本来とは異なる画像が出力されることがわかる．ここで，たしかに視点角度に依存する潜在変数は視点角度推定に必要な特徴を，視点角度に非依存の潜在変数は色や細部の特徴など視点角度推定に不要な特徴を保持していることが分かる．よって，[1] の手法では視点角度に依存する特徴と非依存の特徴が分離できず，誤差伝播時に回帰ネットワークが生成的ネットワークから視点角度に依存する情報のみを効率的に獲得することが困難であったが，提案手法ではその問題を解決したため精度が改善したといえる．

5.　おわりに

本稿では，生成的タスクを補助的に用いた視点角度推定に工夫を施すことで精度を向上させる手法を述べた．具体的には，生成的なデコーダに入力する潜在特徴を視点角度によって変化するものと不変のものに分割し，さらにデータセットのおおよその左右対称性を持つという特性を考慮し左右反転画像を用いたデータ拡張と自己教師あり学習を導入した．適切な損失の導入などによって，視点角度推定における特徴の必要性の有無による潜在特徴の分割を実現すればさらなる精度の向上が期待される．また，並進ベクトルを現在の回帰・生成タスクの中に適切に導入することで，カメラの視点が物体中心を向いているという制約を取り払い，カメラ位置の並進も含んだ視点推定も実現したい．

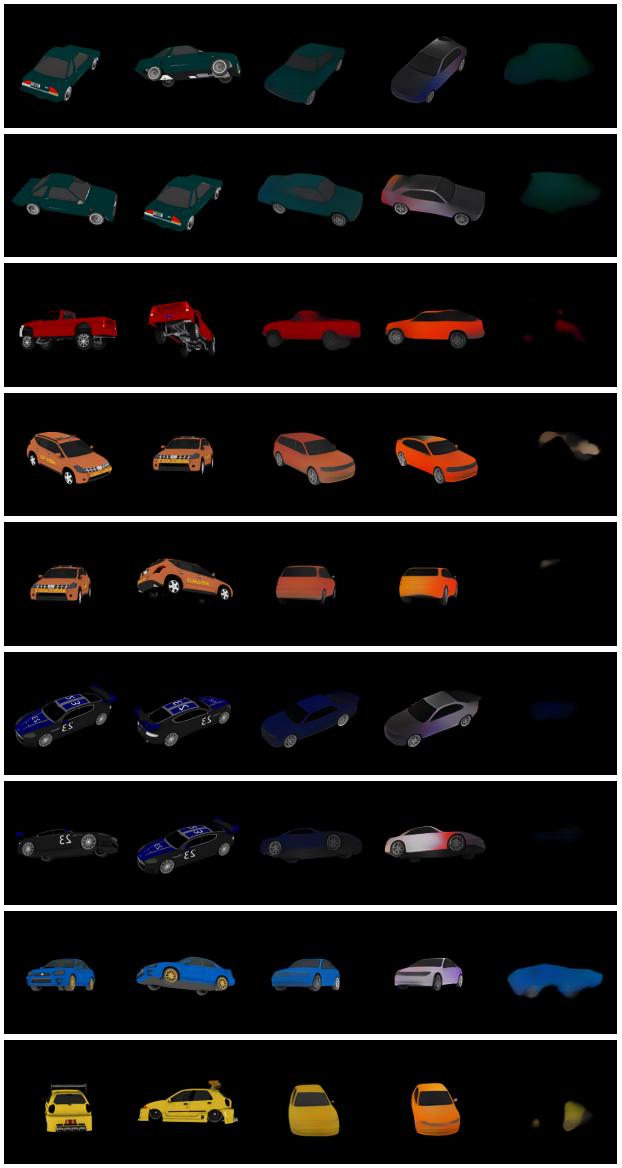


図 3: **再構成画像の分析．**それぞれの行の 5 つの画像は，左からそれぞれ f_v の入力，f_e の入力，f_d の出力，視点角度に非依存の潜在変数を 0 で置き換え視点角度に依存する潜在変数のみを用いたときの f_d の出力，視点角度に依存する潜在変数を 0 で置き換え視点角度に非依存の潜在変数のみを用いたときの f_d の出力を表す．

Fig. 3　Analysis of reconstructed images. The five images in each row represent, from left to right, the input of f_v, the input of f_e, the output of f_d, the output of f_d when the 3D rotation-invariant latent variable is zero-filled and only the 3D rotation-variant latent variable is used, and the output of f_d when the 3D rotation-variant latent variable is zero-filled and only the 3D rotation-invariant latent variable is used.

参考文献

[1] Mariotti, O. and Bilen, H.: Semi-supervised Viewpoint Estimation with Geometry-aware Conditional Generation, *European Conference on Computer Vision*, Springer, pp. 631–647 (2020).

[2] Schaffalitzky, F. and Zisserman, A.: Multi-view matching for unordered image sets, or “How do I organize my holiday snaps?”, *European conference on computer vision*, Springer, pp. 414–431 (2002).

[3] Rolin, P., Berger, M.-O. and Sur, F.: Viewpoint simulation for camera pose estimation from an unstructured scene model, *2015 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, pp. 6320–6327 (2015).

[4] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Cur-

less, B., Seitz, S. M. and Szeliski, R.: Building rome in a day, *Communications of the ACM*, Vol. 54, No. 10, pp. 105–112 (2011).

[5] Massa, F., Marlet, R. and Aubry, M.: Crafting a multi-task CNN for viewpoint estimation, *arXiv preprint arXiv:1609.03894* (2016).

[6] Kanezaki, A., Matsushita, Y. and Nishida, Y.: Rotationnet for joint object categorization and unsupervised pose estimation from multi-view images, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 43, No. 1, pp. 269–283 (2019).

[7] Su, H., Qi, C. R., Li, Y. and Guibas, L. J.: Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2686–2694 (2015).

[8] Tulsiani, S. and Malik, J.: Viewpoints and keypoints, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1510–1519 (2015).

[9] Krull, A., Brachmann, E., Michel, F., Yang, M. Y., Gumhold, S. and Rother, C.: Learning analysis-by-synthesis for 6D pose estimation in RGB-D images, *Proceedings of the IEEE international conference on computer vision*, pp. 954–962 (2015).

[10] Joung, S., Kim, S., Kim, H., Kim, M., Kim, I.-J., Cho, J. and Sohn, K.: Cylindrical convolutional networks for joint object detection and viewpoint estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14163–14172 (2020).

[11] Wang, Y., Li, S., Jia, M. and Liang, W.: Viewpoint estimation for objects with convolutional neural network trained on synthetic images, *Pacific Rim Conference on Multimedia*, Springer, pp. 169–179 (2016).

[12] Rad, M. and Lepetit, V.: Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth, *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3828–3836 (2017).

[13] Penedones, H., Collobert, R., Fleuret, F. and Grangier, D.: Improving object classification using pose information, Technical report, Idiap (2012).

[14] Pavlakos, G., Zhou, X., Chan, A., Derpanis, K. G. and Daniilidis, K.: 6-dof object pose from semantic keypoints, *2017 IEEE international conference on robotics and automation (ICRA)*, IEEE, pp. 2011–2018 (2017).

[15] Divon, G. and Tal, A.: Viewpoint Estimation—Insights & Model, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 252–268 (2018).

[16] Mustikovela, S. K., Jampani, V., Mello, S. D., Liu, S., Iqbal, U., Rother, C. and Kautz, J.: Self-supervised viewpoint learning from image collections, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3971–3981 (2020).

[17] Banani, M. E., Corso, J. J. and Fouhey, D. F.: Novel object viewpoint estimation through reconstruction alignment, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3113–3122 (2020).

[18] Tulsiani, S., Efros, A. A. and Malik, J.: Multi-view consistency as supervisory signal for learning shape and pose prediction, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2897–2905 (2018).

[19] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A. and Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness., *International Conference on Learning Representations*, (online), available from ⟨https://openreview.net/forum?id=Bygh9j09KX⟩ (2019).

[20] Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M. and Naemura, T.: Classification-reconstruction learning for open-set recognition, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4016–4025 (2019).

[21] Rhodin, H., Salzmann, M. and Fua, P.: Unsupervised geometry-aware representation for 3d human pose estimation, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 750–767 (2018).

[22] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H. et al.: Shapenet: An information-rich 3d model repository, *arXiv preprint arXiv:1512.03012* (2015).

[23] Johnson, J., Alahi, A. and Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution, *European conference on computer vision*, Springer, pp. 694–711 (2016).