Vol.2021-CVIM-225 No.1
2021/3/4

不完全な部分観測情報に基づく情報復元の曖昧さを表現する状況記述子の提案

福井 尚卿[†1]　松尾 直志[†1]　島田 伸敬[†1]

概要：人が状況を認識する時，不完全な情報しか観測できなかったとしても，欠けている情報も推定することができる．また，新しい情報を入手することができれば，その情報を用いて適応的に認識を更新することができる．昨今のディープラーニングとコンピュータビジョンの研究の成果により，大規模のデータセットを用いて精密なデータを用意しなくてもパフォーマンスが高い 3 次元再構成を行うことが可能になってきた，しかし，既存の手法では不完全な観測情報に対して確定的に推定を行うことが多く，情報が不足している箇所を推定した際にどのように形状が補完されたのかや観測から得られた情報がどれくらいの曖昧性を持っているのかを解釈できない．本研究では深層生成モデルの枠組みを用いて，部分観測から全体像を復元する際の曖昧性を考慮した解釈容易な記述子空間を設計する手法を提案する．この手法は 3 次元再構成のみならず，復元に曖昧さを伴う問題に用いることができ，また複数の部分観測が得られる場合はそれらを統合して，条件を満たした候補を明示的に生成することができる．

Situation Descriptor Expressing Ambiguity in Information Recovery Based on Incomplete Partial Observation

Abstract: When a person recognizes a situation 3D scene etc., even if only incomplete information can be observed, the missing part can be predicted. In addition, if new information is available, it can be used to adaptively update recognition. Recent results of deep learning and computer vision research have made it possible to perform high-performance 3D reconstruction using large datasets without strict parameters like camera intrinsics and extrinsics etc. In previous methods, the estimate from incomplete observation is often deterministically, it is not possible to interpret how the shape was complemented when estimating the part where the information was lacking, and how ambiguous the estimate obtained from the observation was. In this study, we propose a method for designing an easy-to-interpret descriptor space in consideration of ambiguity when restoring the whole image from imcomplete partial observations using the framework of the deep generative model. This method can be used not only for 3D reconstruction but also for problems with ambiguity in 2D image restoration, and when multiple partial observations are obtained, they can explicitly generate candidates meeted the conditions.

1. はじめに

1.1 研究背景

人間は初めて見る環境であっても，1 枚のシーンから状態の遷移を想起することができる．例えばテーブルのふちにペットボトルが置かれている状況があるとする．それを見た人はそれが不安定な状態であると理解するとともに指で少し押すと下に落ちてしまうということを想像できるだろう．このように人には力学方程式を解かずとも世界の状態をシミュレートする「視覚的想像力」が備わっている．また近年，ロボットに関して産業用のみならず，人と寄り添って生活をサポートするような用途でも研究が盛んである．ロボットがより人間らしい判断をするためにはこの「視覚的想像力」をコンピュータで表現することが不可欠であり，重要性が高まっている．

福井らの研究 [1] では深層学習を用いて剛体の 2D 画像内運動をモデリングし，運動を予測するための枠組みを提案した．また [2] では実世界を撮影した画像に適用するための先駆けとして，3D シミュレーションを使い作成したデータセットを用いて机の上に物体が乗っている状況下にてその物体が落ちそうかそうでないかを推定するモデルを

[†1] 現在，立命館大学
Presently with Ritsumeikan Uniersity

構築し，類似したものであれば実世界で撮影された画像においても有効であると示すことができた．これらの研究の課題として，単視点の奥行き情報が明示的ではない画像からではシーンの 3 次元的な構造に関わる情報を取得することが難しい点があげられる．よって人間のように単視点の情報だけでなく複数視点の情報を効率的に組み合わせて視覚情報を処理することが 3 次元的な構造を理解するために重要であると考えられる．

このような課題を解決するための 3 次元構造の理解に複数視点から撮影された画像を用いる研究は古くから行われている．Structure from Motion (SfM) [3] や Simultaneous Localization and Mapping (SLAM) [4] などの古典的な手法では幾何学的な制約を用いて 3 次元構造の復元を試みているが，多くの場合，十分にキャリブレーションされたカメラから撮影された画像と正確なパラメータが必要となりロバスト性に欠け，別視点から得られた画像同士の対応点をとることができないような観測が不十分な部分については 3 次元構造の復元を行うことができないという課題がある．

近年では深層学習の普及により，大規模なデータセットの可用性が高まり，ハードウェアのコストをかけずに撮影した画像からでも 3 次元再構成のパフォーマンスを向上させることが可能になった．深層学習を用いた手法の利点として，情報が不足する部分を類推によって補って推定することが可能である点が挙げられる．深層学習を用いた 3 次元再構成のステップとして，多くの手法では画像を 3 次元構造の記述子に変換する段階と記述子から 3 次元構造に復元する段階の 2 段階に分けることができる [5]．また，深層学習を用いて複数の視点画像から 3 次元構造を復元するいくつかの研究 [6] [7] [8] では，記述子や再構成された構造を学習可能な関数を用いて統合することで視点ごとの特徴を加味し 3 次元構造を復元することができている．しかし，これまでの深層学習ベースの手法において，画像から得られた 3 次元構造の記述子の解釈性には焦点を当てておらず，情報が不足している箇所を推定した際にどのように形状が補完されたのか明にわからないという課題があり，どの特徴がどれくらいの曖昧性を持っているのかを解釈することが難しい．

これらを踏まえて，3 次元復元問題は各視点から得られる不完全な部分情報から全体情報を復元する問題と言うことができる．このような問題には 1 つの観測からは復元先が 1 つに定まらないという性質があり，他に超解像やブレの補正なども同様の性質を持つ．本研究ではこれらの復元に際して曖昧性を含む問題に対して，深層生成モデルの枠組みを用いた手法を提案する．

1.2 研究概要

本研究では，これまでの深層学習ベースの手法に加え，観測から得られた特徴ベクトルを 1 つの記述子として扱うのではなく Fig. 1 のように一つの観測から，それと矛盾しない複数の状況を表す記述子集合を導くことを考える．記述子集合として扱うことで別の観測情報から得られた記述子集合と積集合をとって統合することが可能であり，定量的に曖昧性を評価できる記述子空間を形成するための手法について提案する．本手法を用いることで，不完全な観測が与えられた時，その観測が復元に際してどのくらい曖昧性を持っているか検討をつけることができ，それを補うような戦略を立てることでより効率的かつ高精度な状況復元を行うことができると考えられる．本研究で提案する手法について，3D モデルのデータセットである ShapeNet [9] を用いてレンダリング画像から 3D モデルを再構成する実験を行い，実験を通して本手法を用いた生成された 3 次元状況記述子集合の解釈性について評価を行う．また，3 次元再構成に限らず部分観測から状況を復元する問題として画像の復元問題においても実験を行い，本研究で提案する手法の有効性を確認する．

2. 提案手法

本研究では先に述べたように，部分観測情報を 1 つ 1 つが全体像に対応する記述子集合に変換し，他の部分情報に由来する記述子集合と統合することができる解釈が容易な記述子空間を形成するための手法について提案する．

まず復元したい状況 s を観測条件 θ で観測 T を行った結果得られる部分観測は $x = \mathrm{T}(s, \theta)$ と表せる．本研究で用いる深層学習モデルでは，入力される x に対して，x の由来である状況 s に対応するものを含む状況記述子 ξ の集合に変換する写像 Enc と状況記述子 ξ から全体像 s を再生する写像 Dec を最適化することを目指す．

2.1 用いる損失関数について

ここで，曖昧性を加味した再構成を行うための Enc と Dec は以下の 3 つの条件を満たすべきである．

(1) 同じ状況を撮影した観測群からは，共通の状況記述子を見つけられるべき (整合性)
(2) この条件は全ての状況記述子 $\xi \in \mathrm{Enc}(x)$ について復元した状況に対して適切な撮影パラメータを選べば x と同じ情報が得られるべき (無矛盾性)
(3) $\xi \in \mathrm{Enc}(x)$ は観測 x の元になった状況 s に対応する状況記述子を含むべき (再生性)

これらを定式化し，最適化のために条件を満たすような損失関数を設計する．

まず整合性に関する損失関数を考える．同じ状況を撮影した観測群は共通の状況記述子を必ず持つべきという条件を満たすため，ある状況 s 由来の n 個の観測群 $\{x_i | i = 1, ..., n\} = \{\mathrm{T}(s, \theta_i) | i = 1, ..., n\}$ に対して $\cap_i \mathrm{Enc}(x_i) \neq \emptyset$ が成り立つべきである．記述子集合間の

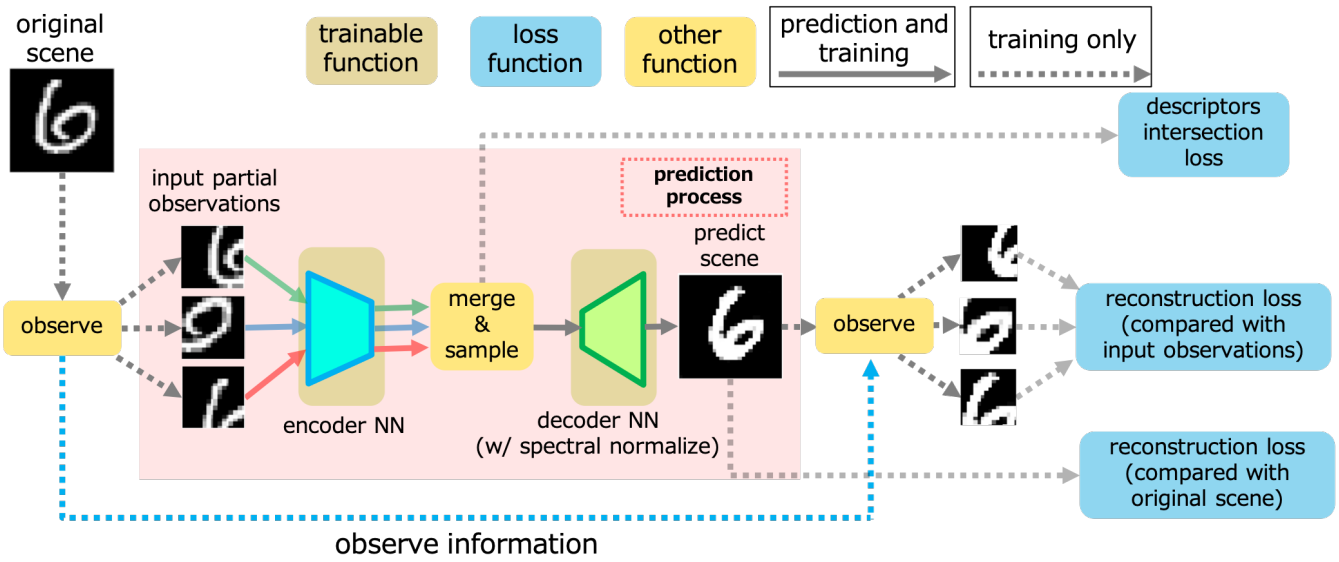


図 1 Overall of Our Model

最短距離が 0 以下になるような制約を加えることで条件を満たすようにする．よって損失関数 L_o として以下のように表すことができる．

$$L_o = \sum_{i \leq n, j \leq n, i \neq j} \text{mindist}(\text{Enc}(x_i), \text{Enc}(x_j))$$

ここで $\text{mindist}(X, Y)$ は集合 X, Y 間の最短距離で以下のように定義する．

$$\text{mindist}(X, Y) = \min\{\text{dist}(x, y) | x \in X, y \in Y\}$$

また $\text{dist}(x, y)$ は記述子 x, y 間のユークリッド距離である．この損失は図 1 中の descriptors intersection loss に対応する．

次に無矛盾性に関する損失関数を考える．ある状況 s 由来の 1 つの観測 $x = \text{T}(s, \theta)$ が与えられた時，$\forall \xi \in \text{Enc}(x)\ \exists \theta'\ s.t.\ x = \text{T}(\text{Dec}(\xi), \theta')$ が成り立つべきである．全ての $\forall \xi \in \text{Enc}(x)$ を評価に用いるのはできないので，近似計算として記述子集合 $\text{Enc}(\text{T}(s, \theta))$ から m 個取り出した集合 $\{\xi_i | i = 1, ..., m\}$ を用いて損失関数 L_c とする．

$$L_c = \sum_{i=1,...,m} \frac{||\text{T}(s, \theta) - \text{T}(\text{Dec}(\xi_i), \theta)||}{m}$$

最後に再生性に関する損失関数を考える．この条件はある状況 s 由来の 1 つの観測 $x = \text{T}(s, \theta)$ が与えられた時，$\exists \xi \in \text{Enc}(x)\ s.t.\ s = \text{Dec}(\xi)$ が成り立つべきである．よって損失関数 L_r で表すことができる．

$$L_r = \min\{\text{sdist}(s, \text{Dec}(\xi)) | \xi \in \text{Enc}(x)\}$$

$\xi \in \text{Enc}(x)$ の中で損失が最も小さいものを逆伝播することで曖昧性を許容することができる．また，ここでの sdist は状況同士の距離である．この損失は図 1 中の reconstruction loss(compared with original scene) に対応する．

ここでの $\text{T}(s, \theta)$ はある状況 s を θ で撮影する処理である．また，この損失は図 1 中の reconstruction loss(compared with input observations) に対応する．

結果，ある状況 s が存在し，入力として $\{x\} = \{\text{T}(s, \theta)\}$ が与えられた時，損失関数は以下のように表すことができる．W はそれぞれの損失に関する重みパラメータである．

$$L = W_o L_o + W_c L_c + W_r L_r$$

この損失関数を用いることで曖昧性を加味した復元が可能な深層学習モデルを訓練することができる．

2.2 Lipschitz 連続性に関する制約

しかし，ここまでの設定では状況記述子 ξ の空間の計量については制限することができていない．制限しないことで，狭い記述子空間にいくらでも多様性のある状況 s を埋め込むことができてしまうため，$\forall \xi \in \text{Enc}(x)$ から $\{\xi_i | i = 1, ..., n\}$ をサンプリングした時，それらから得られた復元結果 $\{s_i | i = 1, ..., n\}$ のいずれとも大きく異なる復元結果が存在することが否定できない．また，記述子空間上の距離が復元した状況のバリエーションと対応しないため，空間形状記述子の解釈が困難になる．そこで，状況記述子 ξ から全体像 s を再生する写像 Dec に対して，以下のような明示的な Lipschitz 定数 C についての Lipschitz 連続性を要求する．

$$\text{dist}(\text{Dec}(\xi_1), \text{Dec}(\xi_2)) \leq C\ \text{dist}(\xi_1, \xi_2)$$

Lipschitz 定数 C を定めることによって復元した状況の差を記述子空間上での座標の距離として定量的に解釈できるようになる．本研究では Spectral Normalization [10] を

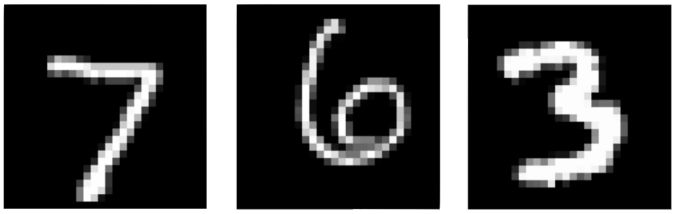


図 2 Example data of MNIST dataset

図 3 Example data of Fashion MNIST dataset

Dec の学習可能な重み行列に適用し，1-Lipschitz 連続性を実現する．

2.3 正則化による不自然な出力の抑制

2.1 で述べたとおり，損失関数の設計として，正解に矛盾しないように多様な出力を得るため，出力群の中に正解に近いものが含まれているならばそれ以外は観測操作を行った後の正解に一致する条件の元で人間から見て不自然な出力結果が含まれる可能性がある．そのため，出力結果を抑制するためのタスクごとに異なる正則化を用いなければいけない場合がある．詳細に関してはそれぞれの実装の章で述べる．

3. 実験

3.1 矩形で切り取られた部分画像からの全体画像の復元

本実験では，画像の一部を複数パターン矩形で切り取ったものを入力とし，画像の全体を復元するタスクを行う．提案手法に対して，本手法を用いて生成された全体像の記述子集合の解釈性について評価する．

3.1.1 用いるデータについて

本実験では MNIST データセット [11] と Fashion MNIST [12] を用いて実験する．MNIST データセットには図 2 のような 0 から 9 までの手書き文字の画像が含まれる．Fashion MNIST データセットには図 3 のような様々な服飾品に関する画像が含まれる．どちらのデータセットに含まれる画像も大きさは 28×28 で白黒画像である． MNIST・Fashion MNIST 共に学習データとテストデータの分割はデータセットの提供元に従い行った．

3.1.2 実装

ここではある画像 s から切り取られた複数の矩形画像 $\{x_i|i=1,...,n\} = \{\mathrm{T}(s,\theta_i)|i=1,...,n\}$ が存在するとする．学習可能な関数 Enc は入力として $\{x_i|i=1,...,n\}$ をとり，復元される画像を表現する記述子の集合 $\{\Xi_i|i=1,...,n\}$ を出力する．Enc は 3 層の CNN を重ねた後，CNN の出力を $\{c_i|i=1,...,n\}$ と $\{r_i|i=1,...,n\}$ に変換する 2 つの 1 層の線形結合層から構成される．それぞれ c_i と r_i は 64 次元のベクトルである．記述子集合 Ξ_i が存在する空間の区間は各次元 k の区間 $[(c_{i,k}-r_{i,k})/2, (c_{i,k}+r_{i,k})/2]$ の直積として表現する．ある矩形画像 x_i に対して，区間の中心 c_i と区間の広さ r_i を区間に持つ一様分布について考え，サンプリングすることで記述子 ξ_i を得る．また，重複領域 $\cap_n \Xi_n$ の区間を持つ一様分布を考えることで，複数の矩形画像に由来した記述子 ξ をサンプリングすることができる．

$\cap_n \Xi_i$ の区間を持つ一様分布からサンプリングされた記述子 ξ は学習可能な関数 Dec に入力することで矩形を切り取られる前の元画像を表す s_{pred} を出力する．一様分布からのサンプリングには Variational Auto Encoder [13] と同じく reparameterization trick を用いる．学習可能な関数 Dec は 3 層の逆 CNN 層を用いて実装した．

深層学習モデルの最適化には，本実験では 2 章で述べた reconstruction loss(compared with original scene) で用いる生成された状況と元になった状況との距離 sdict を平均二乗誤差で計算する．また，reconstruction loss(compared with input observations) も平均二乗誤差で計算する．

MNIST・FashionMNIST の両実験共に, 最適化には Adam を用い 200epoch 学習させ，学習率は $1e-4$，weight decay のパラメータは $1e-5$ を用いた．学習率は CosineAnnealingLR を用いて最終的に $1e-5$ まで減少させた．MNIST の実験では元々 28×28 の大きさだったのを矩形に切り取る際に 14×14 の大きさで切り取り，深層学習モデルに入力する際は 28×28 に戻して用いた．Fashion MNIST の実験では元々 28×28 の大きさだったのを $28/3 \times 28/3$ の大きさで切り取り，深層学習モデルに入力する際は 28×28 に戻して用いた．

3.2 実験結果

MNIST での複数の部分観測 $\{x_i|i=1,2,3\}$ を逐次 Enc に入力した時，新しい情報が得られるごとに推定する手書き文字画像のバリエーションが絞り込まれていくことを図 4 に示す．次に，Fashion MNIST でも同じく部分観測を逐次入力した時，復元される画像のバリエーションが絞り込まれていくことを図 5 に示す．それぞれ結果の可視化は test データ (学習時に用いていないデータ) で行った．図 4 や図 5 から，1 つ 1 つの入力からはわからなかった切り取られた矩形以外の領域について，他の入力情報を使って正解に近い画像を生成できている．また，1 つの入力からは確定できない復元領域は様々な形状を復元することができており，曖昧性を表現することができている．

3.3 Lipschitz 連続性についての分析

本研究で提案するモデルでは，Dec において入力と出力

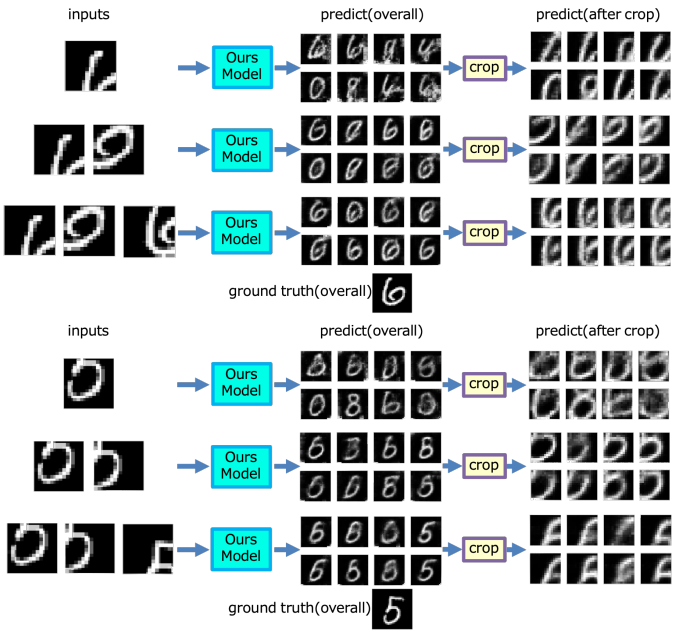


図 4 Output Variations for Input Images using MNIST

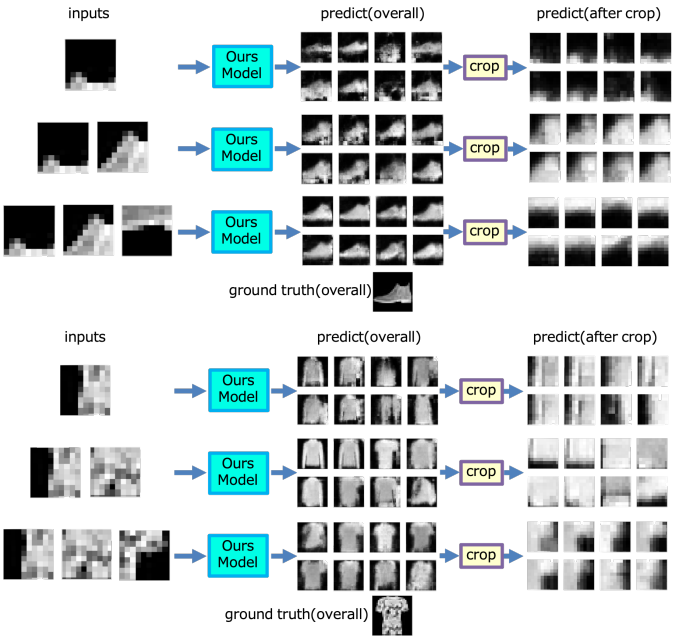


図 5 Output Variations for Input Images using Fashion MNIST

に 1-Lipschitz 連続性が成立するようにし，記述子空間の計量をモデルの出力空間の計量が離れすぎないように制限する．

本章では MNIST のデータを用いて Dec の入力と出力の関係が 1-Lipschitz 連続性に近似できているかどうかを分析した．2 つの記述子 ξ_1, ξ_2 間の距離 d_x とそれらを入力としたときの Dec の出力 $\mathrm{Dec}(\xi_1), \mathrm{Dec}(\xi_2)$ 間の距離 d_y の関係について，$d_y/d_x < 1$ が成立することが 1-Lipschitz 連続性が成立する条件である．図 6 は 200 パターンのデータごとに 50 サンプル中の全ての組み合わせについて，出力される画像間の距離と，記述子集合上の 2 サンプルの距離の商を計算した．このように，大多数のデータでは

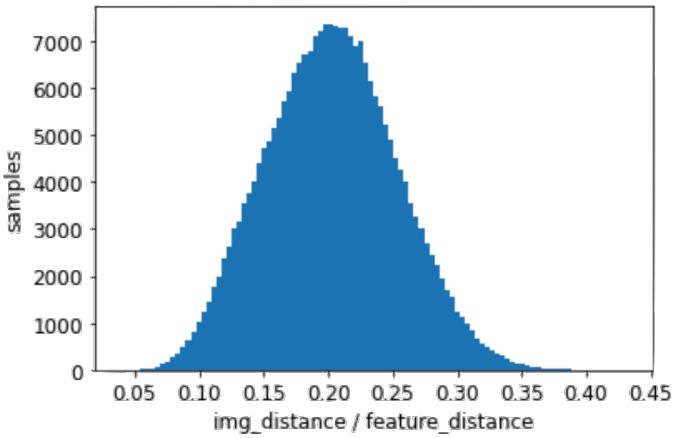


図 6 Visualization of 1-Lipschitz Continuous

図 7 Examples of ShapeNet Dataset

$d_y/d_x < 1$ が成立していることから Dec の入力と出力の関係が 1-Lipschitz 連続性が成立していることがわかる．

3.4 ShapeNet を用いた 3D 再構成の実験

3.4.1 用いるデータについて

使用するデータセットとして，ShapeNet の中の Car カテゴリを用いて実験を行う．Car カテゴリの中には Fig. 7 のような車型物体の 3D モデルが 10692 通りあり，それぞれの 3D モデルを 24 通りの角度で Blender を用いてレンダリングされた画像が含まれる．レンダリング画像は Choy らの研究 [6] で用いられた物と同様である．レンダリングされた画像はグレースケール画像に変換して用いられる．データセットは訓練用，検証用，test 用に三等分して使用した．

3.4.2 実装

ここではある 3 次元状況 s に関する複数の 2 次元の観測 $\{x_i | i = 1, ..., n\} = \{\mathrm{T}(s, \theta_i) | i = 1, ..., n\}$ が存在するとする．学習可能な関数 Enc は入力として $\{x_i | i = 1, ..., n\}$ をとり，3 次元状況記述子集合 $\{\Xi_i | i = 1, ..., n\}$ を出力する．本実験に用いる実装では Enc に resnet18 [14] と出力された特徴ベクトルを $\{c_i | i = 1, ..., n\}$ と $\{r_i | i = 1, ..., n\}$ に変

換する 2 つの 1 層の線形結合層を用いる．それぞれ c_i と r_i は 64 次元のベクトルである．観測を統合する部分の実装は 3.1.2 で述べたものと同様である．

$\cap_n \Xi_i$ の区間を持つ一様分布からサンプリングされた 3 次元状況記述子 ξ は学習可能な関数 Dec に入力することで 3 次元状況を表す s_{pred} を出力する．学習可能な関数 Dec には 3 層の線形結合層とその出力を決められた点数の単位球の面の結合情報を用いて変形させる関数が含まれている．

観測パラメータ θ で 3 次元状況 s を観測する操作 $\mathrm{T}(s, \theta)$ は微分可能レンダラーである DIB-Renderer [15] を用いて実装する．

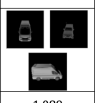
深層学習モデルの最適化には，本実験では 2 章で述べた reconstruction loss(compared with original scene) で用いる生成された状況と元になった状況との距離 sdict を Chamfer Distance [16] で計算する．また，reconstruction loss(compared with input observations) は 3.1.2 で述べたものと同様に平均二乗誤差で計算する．それに加えて 3 次元状況を生成する際に用いる正則化として，smoothness loss [17], laplacian loss [18], edge length loss [19] を用いる．

訓練について，最適化には Adam [20] を用い，学習率の初期値は 1e-5 に設定して 200epoch 学習させ，検証用データセットにおいて最も損失が小さい重みを保存した．学習率は CosineAnnealing [21] 法を用い，200epoch かけて学習率が 1e-8 になるように減衰させた．

3.4.3 実験結果

本実験では 2 で提案した手法を用いて生成した 3D 状況記述子集合の解釈性について，ある 3 次元状況 s に関する複数の不完全な観測 $\{x_i | i = 1, 2, 3\}$ を逐次 Enc に入力した時，新しい情報が得られるごとに推定する 3 次元状況のバリエーションが絞り込まれていくことを Fig. 8 に示す．また，Fig. 8 に対応して，入力画像に対して生成された Enc の出力である記述子集合の範囲を表す特徴ベクトルの平均値 r_{mean} と 3 次元状況のバリエーションを表す指標 v_{3d} の変化を表に示す．

表 1 Evaluation of Output Variations for Input Images

入力	x_1	x_1, x_2	x_1, x_2, x_3	x_1	x_1, x_2	x_1, x_2, x_3
入力される観測						
r_{mean}	5.522	4.768	0.8699	5.001	2.394	1.089
v_{3d}	0.5416	0.4379	0.01405	0.4380	0.1102	0.02195

v_{3d} は $v_{3d} = \mathrm{trace}(\mathrm{cov}(\{P_i | i = 1...n\}))$ と計算される．ここで P_i はある 3 次元状況 s_i を構成する頂点の 3 次元座標の集合であり，ここで cov は共分散行列をとる計算を，trace は共分散行列の対角要素の総和を表す．Fig. 8 の例を見ると，左から 2 つめのような同じような視点から得られた情報では記述子集合の重複領域は絞り込まれず，3 つめの観測のような違う視点から得られた情報を用いるとその画像で確定する領域に関するバリエーションが絞り込むことができている．また，Table 3.4.3 から定量的に見ても新しい情報が得られるごとに重複領域の大きさ・3 次元状況のバリエーションが絞り込まれていることがわかる．

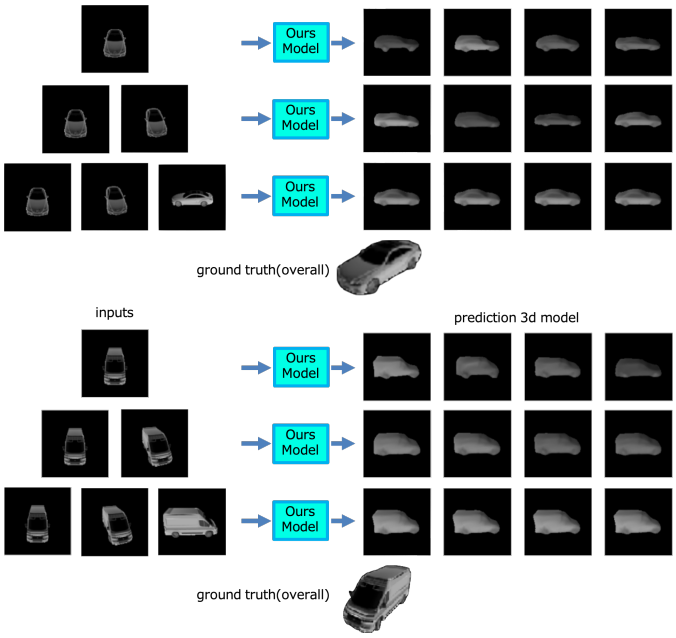


図 8 Output Variations for Input Images using ShapeNet

4. おわりに

本研究では深層生成モデルの枠組みを用いて，不完全な 2D 観測情報から曖昧性を加味した 3 次元再構成を行うための解釈容易な記述子空間を生成する手法を提案した．ShapeNet の Car カテゴリを用いた実験では，新たな視点情報が得られるごとに 3 次元状況記述子の空間を更新し，得られた情報に応じて生成する 3 次元状況のバリエーションを絞りこむことができていることを確認することができた．また，2D 画像における復元実験では使った実験では提案手法で述べたように 1-Lipschigtz 連続性が概ね成立していることを示し，3 次元再構成以外のタスクでも有効であることを示すことができた．今後の展望としては，他の様々なカテゴリのデータセットでの実験や，視点情報や拡大だけではなく，超解像のタスクや時間軸での情報復元など様々な種類の情報復元に関する実験を行いたい．また，既存手法との 3 次元再構成の精度の比較や 3 次元状況記述子を用いた強化学習などの別タスクへの応用など行うことを目指す．

参考文献

[1] 福井尚卿，島田伸敬，松尾直志：力入力に対する剛体群の運動応答予測と静力学的構造安定性の推定，ロボティクス・メカトロニクス講演会講演概要集 2019，一般社団法人 日本機械学会，pp. 2P2–I03 (2019).
[2] Tadashi, M., Fukui, T. and Shimada, N.: Detection of Unstable Objects by Using Deep Learning for Domes-

tic Environment, *The 15th Joint Workshop on Machine Perception and Robotics (MPR2019)*, pp. P2–19 (2019).
[3] Özyesil, O., Voroninski, V., Basri, R. and Singer, A.: A survey of structure from motion., *Acta Numerica*, Vol. 26, p. 305 (2017).
[4] Fuentes-Pacheco, J., Ruiz-Ascencio, J. and Rendón-Mancha, J. M.: Visual simultaneous localization and mapping: a survey, *Artificial intelligence review*, Vol. 43, No. 1, pp. 55–81 (2015).
[5] Han, X., Laga, H. and Bennamoun, M.: Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era, *IEEE transactions on pattern analysis and machine intelligence* (2019).
[6] Choy, C. B., Xu, D., Gwak, J., Chen, K. and Savarese, S.: 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction, *European conference on computer vision*, Springer, pp. 628–644 (2016).
[7] Kar, A., Häne, C. and Malik, J.: Learning a Multi-View Stereo Machine, *Advances in Neural Information Processing Systems 30* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Curran Associates, Inc., pp. 365–376 (online), available from ⟨http://papers.nips.cc/paper/6640-learning-a-multi-view-stereo-machine.pdf⟩ (2017).
[8] Xie, H., Yao, H., Sun, X., Zhou, S. and Zhang, S.: Pix2vox: Context-aware 3d reconstruction from single and multi-view images, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2690–2698 (2019).
[9] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L. and Yu, F.: ShapeNet: An Information-Rich 3D Model Repository, Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago (2015).
[10] Yoshida, Y. and Miyato, T.: Spectral norm regularization for improving the generalizability of deep learning, *arXiv preprint arXiv:1705.10941* (2017).
[11] LeCun, Y. and Cortes, C.: MNIST handwritten digit database, (online), available from ⟨http://yann.lecun.com/exdb/mnist/⟩ (2010).
[12] Xiao, H., Rasul, K. and Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms (2017).
[13] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes (2013).
[14] He, K., Zhang, X., Ren, S. and Sun, J.: Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (online), DOI: 10.1109/CVPR.2016.90 (2016).
[15] Chen, W., Gao, J., Ling, H., Smith, E., Lehtinen, J., Jacobson, A. and Fidler, S.: Learning to Predict 3D Objects with an Interpolation-based Differentiable Renderer, *Advances In Neural Information Processing Systems* (2019).
[16] Achlioptas, P., Diamanti, O., Mitliagkas, I. and Guibas, L.: Learning representations and generative models for 3d point clouds, pp. 40–49 (2018).
[17] Kato, H., Ushiku, Y. and Harada, T.: Neural 3d mesh renderer, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3907–3916 (2018).
[18] Liu, S., Li, T., Chen, W. and Li, H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning, pp. 7708–7717 (2019).
[19] Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W. and Jiang, Y.-G.: Pixel2mesh: Generating 3d mesh models from single rgb images, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 52–67 (2018).
[20] Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations* (2014).
[21] Loshchilov, I. and Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts (2016).