

博士論文

深層生成モデルによる  
環境の構成的な認識と生成

指導教員 松尾 豊 教授

東京大学大学院 工学系研究科 技術経営戦略学専攻

小林 由弥



# 要旨

深層学習の登場により、人工知能で実現可能な情報処理の種類や精度は飛躍的に向上した。しかし現状ではこれは適切に設計された狭い範囲の課題に限られており、人間のように広範な課題に対して柔軟に対応する人工知能の実現は未だ遠いものとなっている。より柔軟で汎用的な情報処理を行うために考慮すべき要素の一つとして、構成性があると考えている。構成性とは、ある系や意味表現がそれ自体よりも単純な構成要素の組み合わせによって規定されるような性質のことで、言語や我々の住む実世界に見られる特徴である。通常、人間は全体を一度に認識するのではなく、個々の要素に分解して認識を行っている。ここでの分解の単位となる要素は主に物体であるが、抽象的な概念である場合もある。現実世界はこうした単純な構成要素の組み合わせによって無数の異なる状態が実現され得るが、組み合わせを考慮しない場合には無数のパターンを個々に認識・理解する必要がある、汎化や転移能力の面で非効率的であると考えられる。また、物体は系の発展の単位となることが多く、個々の物体を認識することは因果推論（causal inference）や物体間の関係性の推論（relational inference）などに関して有利になることが期待される。

このような背景から本研究では、系を複数の物体の組み合わせとして捉えるような構成的な認識に関する研究を行うことを考え、特に画像を対象として教師なしで物体認識と各物体の表現学習を行う深層生成モデルを扱う。このような手法は物体中心表現学習（object-centric representation learning）と呼ばれるが、現状では深層生成モデルを基にした物体中心表現学習の手法は学習の不安定性や処理の精度、適用範囲が簡単なデータに限定されていると言った課題を抱えている。本研究でこうした問題が根本的には教師なし学習によって物体に関する前提知識（帰納バイアス）が十分に得られていないことによると考え、適切な帰納バイアスを導入する方法について考えるものである。

本研究は上記のテーマに関する3つの研究によって構成されている。研究1は背景に関する補助情報を導入することで、物体を認識するために必要な帰納バイアスを与えるものである。また、研究1では背景の情報を与えるために新たなデータを用いているが、研究2では新たなデータなしに、学習やネットワーク構造の工夫によって帰納バイアスを導入する方法を提案する。研究1と2では静止画一枚のみを与え、物体を認識するという二次元空間での課題となっ

ていたが，研究 3 では多視点の入力に拡張した問題設定に取り組んでいる．これは，複数の物体を含む，ある 3 次元的な空間に関して複数視点からの画像を与え，観測していない任意の視点からの画像の予測と物体認識を行うという課題である．また，研究 1 と研究 2 ではデータの与え方や学習の工夫による帰納バイアスの導入を試みたが，研究 3 ではモデル構造に前提知識を組み込む形の導入を行う．また，8 章ではこれらの研究をまとめ，本研究の発展や今後の研究の方向について考察を行う．



# 目次

<b>第 1 章</b>	<b>序論</b>	<b>1</b>
1.1	研究背景 . . . . .	1
1.1.1	人工知能研究の現状と課題 . . . . .	1
1.1.2	構成性の原理 . . . . .	4
1.1.3	深層学習と表現学習 . . . . .	6
1.1.4	深層生成モデルによる認識 . . . . .	7
1.1.5	深層生成モデルによる構成的な認識 . . . . .	8
1.1.6	概念的な構成性と空間的な構成性 . . . . .	9
1.1.7	シーン解釈手法の課題 . . . . .	10
1.2	本研究の目的 . . . . .	11
1.3	本論文の構成 . . . . .	12
<b>第 2 章</b>	<b>深層生成モデルに関する前提知識</b>	<b>14</b>
2.1	生成モデル . . . . .	14
2.1.1	生成モデルの学習 . . . . .	15
2.2	深層生成モデル . . . . .	16
2.2.1	変分自己符号化器 (Variational Autoencoder: VAE) . . . . .	17
2.2.2	自己符号化器 (Autoencoder: AE) . . . . .	18
<b>第 3 章</b>	<b>構成的な認識とその関連研究</b>	<b>19</b>
3.1	物体中心表現学習 (Object-centric Representation Learning) . . . . .	19
3.2	生成モデルを用いた Object-centric モデル . . . . .	19
3.2.1	VAE-based モデル . . . . .	20
3.2.2	AE-based モデル . . . . .	22
3.2.3	Expectation Maximization(EM) ベースのシーン解釈手法 . . . . .	22
3.3	識別モデルによる object-centric モデル . . . . .	22
3.4	Object-centric Representation の応用利用 . . . . .	23

3.5	教師なし画像処理手法に対する位置付け . . . . .	24
<b>第 4 章</b>	<b>深層生成モデルによる環境の構成的な認識と生成</b>	<b>25</b>
4.1	関連研究を踏まえた本論文の目標 . . . . .	25
4.2	本章以降の位置付け . . . . .	26
<b>第 5 章</b>	<b>深層生成モデルによる背景情報を利用したシーン解釈</b>	<b>29</b>
5.1	背景 . . . . .	29
5.2	関連研究 . . . . .	31
5.2.1	シーン解釈の位置付け . . . . .	31
5.2.2	シーン解釈の関連研究 . . . . .	32
5.2.3	データ分布の対照による特徴抽出 . . . . .	34
5.3	手法 . . . . .	35
5.3.1	問題設定 . . . . .	35
5.3.2	提案手法 . . . . .	36
5.4	実験結果 . . . . .	42
5.4.1	データセットについて . . . . .	42
5.4.2	実験結果 . . . . .	43
5.5	多スロットへの拡張 . . . . .	52
5.5.1	手法 . . . . .	52
5.5.2	多スロット版の実験 . . . . .	52
5.6	結論と考察 . . . . .	53
<b>第 6 章</b>	<b>自己教師あり学習と Transformer を用いたシーン解釈手法の提案</b>	<b>55</b>
6.1	背景 . . . . .	55
6.2	関連研究 . . . . .	58
6.2.1	シーン解釈手法とランダム生成 . . . . .	58
6.2.2	自己教師あり学習 . . . . .	59
6.2.3	Vision Transformer . . . . .	60
6.3	手法 . . . . .	61
6.3.1	確率モデル . . . . .	63
6.3.2	モデル構造 . . . . .	65
6.4	実験 . . . . .	69
6.4.1	Objects Room データセットでの実験結果 . . . . .	71
6.4.2	ShapeStacks での実験結果 . . . . .	74
6.4.3	Multi-MNIST での実験結果 . . . . .	75

6.5	結論 . . . . .	77
第 7 章	大域的な空間情報を持つ多視点シーン解釈モデルの提案	81
7.1	はじめに . . . . .	81
7.2	関連研究 . . . . .	83
7.2.1	多視点物体中心表現学習 (Multi-view Object-centric Representation Learning) . . . . .	83
7.2.2	Generative Query Network(GQN) の関連研究 . . . . .	84
7.3	手法 . . . . .	85
7.3.1	問題設定 . . . . .	86
7.3.2	確率モデル . . . . .	86
7.4	実験 . . . . .	91
7.4.1	Novel View Synthesis と Segmentation . . . . .	92
7.4.2	Novel Scene Generation . . . . .	95
7.4.3	大域的な潜在変数の表現について . . . . .	97
7.4.4	Downstream Task について . . . . .	98
7.5	結論 . . . . .	98
第 8 章	考察	100
8.1	各研究の整理 . . . . .	100
8.2	提案手法の貢献と今後の課題 . . . . .	101
8.3	構成的な認識技術の応用可能性について . . . . .	104
8.4	人間の認知機構と物体認識技術の今後の方向性 . . . . .	105
第 9 章	結論	108
	謝辞	110
	参考文献	111
付録 A	研究業績一覧	125

# 目次

1.1	構成性を考慮することが重要だと考えられる, relational inference の課題例 [1]. 複数の物体が含まれる系の動画に基づいて, 質問への回答や説明を行う課題となっている. . . . .	6
1.2	Superposition Catastrophe について示した図 [2]. . . . .	7
5.1	提案手法の構成図. Enc.A, Enc.B: エンコーダ, Disc.: 識別器 (ディスクリミネータ), Decoder: デコーダ を意味する. . . . .	39
5.2	生成 (再構成) の過程を示した模式図. . . . .	41
5.3	Textured MNIST データセットに対する提案手法の適用結果. 5 サンプルに対する結果を示している. 左列から, 入力画像, 再構成された画像, 前景 (数字) の抽出, 背景の抽出, 前景のマスク, マスクで切り出す前の画像, となっている. . . . .	44
5.4	本研究で新たに導入したデータセット 3 種 (Textured MNIST, Multi-Textured MNIST, NS-MNIST SMALL) に対する, 背景情報を用いない, 既存手法 (MONet) での実験結果. . . . .	46
5.5	提案手法の一部の機構を無効化した場合と既存手法の, ARI スコアによる定量評価. エラーバーは標準偏差を表している. 左から順に, 通常時, TC 項の除去, 背景利用なし, TC 項・背景ともに用いない場合, 先行研究 (MONet) による結果を示している. . . . .	47
5.6	Multi-Textured MNIST データセットに対する提案手法の適用結果. . . . .	48
5.7	Natural Scene MNIST データセットに対する提案手法の適用結果. . . . .	50
5.8	Objects Room データセットに対する提案手法の適用結果. 最左列の入力画像以外は提案手法の出力結果である. . . . .	51
5.9	Textured Multi-dSprites データセットに対する, 多スロット版の提案手法の適用結果. . . . .	53

6.1	GENESIS と Transformer を用いたモデル構造 (GENESIS+Tr). 後段の VAE 部分は共通した構造となっている. . . . .	66
6.2	各モデルの Objects Room における推論結果. $k$ はスロットの番号である. Tr は Transformer, SS は自己教師あり学習を指す. また, scratch は自己教師あり学習を用いない通常の学習を指す. . . . .	78
6.3	自己教師あり学習による学習の安定化を複数の乱数シード (学習時) について比較した結果. ここで利用したモデルは GENESIS+Tr である. 上段が自己教師あり学習を用いない場合, 下段が自己教師あり学習を用いた場合となっている. 左右の列 (seed A, seed B) は異なる乱数シードに対応している. . .	79
6.4	Objects Room データセットでの生成結果 左列は自己教師あり学習の, 右列は通常の学習の結果を示している. . . . .	79
6.5	Multi-MNIST データセットでの推論結果. w/reg. は制約項の導入を意味する.	80
7.1	提案手法 WeLIS の概要. このモデルではいくつかの視点 ( $\mathbf{x}^1$ - $\mathbf{x}^3$ ) を観測し, 任意の指定された方向 $\mathbf{v}^q$ からの画像を予測し, セグメンテーションを行うものである. $\mathbf{z}^g$ は空間的な情報をモデル化する大域的な潜在変数で, $\mathbf{z}_k$ は物体ごとの表現となっている. 図示した画像は実際に提案手法の入出力として得られた結果である. . . . .	85
7.2	クエリに関する推論の定量評価. RMSE は小さな方が, mIoU は高い方が良いスコアである. GQN Jaco は正解のセグメンテーションが含まれていないため, mIoU が未評価となっている. . . . .	90
7.3	観測に対する定量評価. 入力画像に対する再構成とセグメンテーションの品質を評価したものとなっている. . . . .	90
7.4	提案手法の各機構について除去実験を行った結果. . . . .	91
7.5	GQN Jaco における novel view synthesis の結果. これらの画像は未観測の視点の予測結果であり, 観測として与えた視点の数は $N_{obs} = 3$ である. . . . .	91
7.6	新たなシーンの生成を物体ごとに行った結果について, WeLIS とベースラインの比較結果. 各行は異なるクエリの視点に相当する. 最初の列は生成結果の画像を, そして次の 7 つの列は各スロット ( $\mathbf{x}_k, k \in \{1, \dots, K\}$ ) を, そして最後の列は生成されたセグメンテーションマスクを示している. . . . .	92

7.7	新たなシーンの生成結果. 上の行は WeLIS と MulMON の結果を示している. 下の行は WeLIS から Structured Prior を取り除いたものと, Normalizing Flow を取り除いたものを示している. これらの画像は, WeLIS については 8 個の異なる $\mathbf{z}_g$ をサンプリングすることで生成し, MulMON については各スロットの潜在変数をそれぞれサンプリングした結果である. なお WeLIS の完全なモデル (左上) と Structured Prior を除去した結果 (左下) については, 同じ $\mathbf{z}_g$ の組を用いている. . . . .	93
7.8	novel view synthesis の定性的結果と, その定量評価. 観測数は 3 で, 各グループの一行目は正解の画像 (GT) を, 二行目は予測結果 (Pred.) を, 三行目はセグメンテーション結果 (Seg.) を示している. . . . .	95
7.9	異なる 4 つの $\mathbf{z}_g$ (A–D) から生成した結果と, その近傍 $\mathbf{z}_g + \epsilon$ からの生成結果. ここで $\epsilon$ はランダムな摂動としてガウスノイズを利用している. 各グループの最も左の列は元の $\mathbf{z}_g$ から生成されたもので, それ以外の列は異なる摂動 $\epsilon$ を与えて生成した結果である. 各グループ (A–D) が似た空間的な構成を保持していることが分かる. 右側の潜在空間は概念図であり, 実際の数値を反映したものではない. . . . .	97
7.10	異なる 2 つの $\mathbf{z}_g$ の間を直線で補間し, 潜在変数空間を横断した場合の結果. 各行は異なる補間の例であり, 計 3 つ組の $\mathbf{z}_g$ についての結果を示している. . . . .	97
8.1	チョムスキー階層. regular が有限状態文法に, context-free が文脈自由文法に相当する. CC BY-SA 3.0 J. Finkelstein <a href="https://en.wikipedia.org/wiki/Chomsky_hierarchy#/media/File:Chomsky-hierarchy.svg">https://en.wikipedia.org/wiki/Chomsky_hierarchy#/media/File:Chomsky-hierarchy.svg</a> . . . . .	106

# 表目次

6.1	Objects Room データセットにおける mSC . . . . .	74
6.2	Objects Room データセットにおける FID . . . . .	74
6.3	ShapeStacks データセットにおける mSC . . . . .	75
6.4	ShapeStacks データセットにおける FID . . . . .	75
6.5	Multi-MNIST データセットにおける定量評価. Reg. は制約項の有無を意味する. 誤差項は標準誤差である. 括弧内は比較不可であるが, 参考値として記載した. . . . .	77
7.1	multi-object-multi-view (MOMV) の問題設定に関連する手法の比較. . . . .	83
7.2	FID スコアの比較. WeLIS, WeLIS から Normalizing Flow を除去したもの, MulMON について比較を行っている. なお, FID スコアは低いほど良い結果であることを示す. . . . .	94
7.3	Structured Prior の実装として異なるネットワークアーキテクチャを用いた場合の FID. 小さい方が良いスコアである. . . . .	96
7.4	シーンに含まれる物体の個数を推定する downstream task の精度. Observations は観測として与えた視点の数を意味する. . . . .	99





# 第 1 章

## 序論

### 1.1 研究背景

#### 1.1.1 人工知能研究の現状と課題

深層学習 [3] の登場により、人工知能で実現される情報処理の種類や性能は 2010 年頃を境に飛躍的に向上した。性能面について、画像分類<sup>\*1</sup>や自然言語処理、音声認識など、課題の範囲が適切に限定されたものであれば人間を超える場合も多くなっている [4, 5]。機能面についても様々な処理が実現されており、例えば画像の生成とその性質の変換や [6]、テキストからの画像生成、画像の補完や編集などを一つのモデルで実現可能な手法も登場している [7]。また、大規模データの学習により高品質な言語モデルが構築され、翻訳や質問への応答を自然に行うことが可能となっている [8, 9]。このような言語モデルを用いて文章からの画像生成を行った研究では、一般的な人間の想像力を超えるような結果も示されている [10]。

このような深層学習の発展により、産業応用も進んでいる。具体例としては、生簀内の養殖マグロの数を深層学習を用いた物体検出によって自動で数え、給餌量の最適化を行ったプロジェクトがある<sup>\*2</sup>。また、機械翻訳や音声認識による自動字幕生成<sup>\*3</sup>、音声合成による読み上げ<sup>\*4</sup>は主に Web サービスとして実装され、多くの人に利用されている。

しかし、これはあくまで適切に課題を設定し、特定の範囲の処理に限定した場合である。人間のように汎用的で柔軟な処理を行う人工知能は汎用人工知能 (Artificial General Intelligence:

---

<sup>\*1</sup> What I learned from competing against a ConvNet on ImageNet (<http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/>, 2018 年 8 月アクセス.)

<sup>\*2</sup> [https://www.isid.co.jp/library/special/2018\\_tunafarm\\_1.html](https://www.isid.co.jp/library/special/2018_tunafarm_1.html)

<sup>\*3</sup> Youtube など <https://www.youtube.com/>, 2021 年 10 月時点の機能。

<sup>\*4</sup> NHK 報道資料 <https://www.nhk.or.jp/info/pr/toptalk/assets/pdf/soukyoku/2020/10/002.pdf>, 2021 年 8 月アクセス。

AGI) と呼ばれるが、このように課題を細かく設定せずに適切な処理を行うことは未だ困難である。小売店・飲食店等の省人化・無人化や、介護ロボット、完全自動運転のような産業応用や自動化を行うためには、特定の課題における高い精度ではなく、汎用的で柔軟な課題解決能力が必要であるため、AGI に近い人工知能の実現が必要であると考えられる。

また、汎用的な課題でなくとも、複数の簡単な作業を組み合わせて行うような処理（手続き的な処理）や、数式操作のような記号的な処理についても現状では難しい課題となっている。例えば、大規模な言語モデルである GPT-3[9] を用いて素数を出力させた結果が話題になった<sup>\*5</sup> が、この結果にはいくつかの誤りが含まれており、モデルが入力データの単なる丸暗記をしているのではないことを示すと同時に、数学的な法則に則った記号的な処理を苦手としている例として捉えることができる。

このように、適切に設計された課題については高性能を発揮しているが、汎化能力や転移能力<sup>\*6</sup>の面では未だ人間に遠く及ばず [11]、記号処理や手続き的な処理についても課題がある。前者のように未知の課題に対して性能が出ないことは、与えられたデータの範囲内では十分な性能を発揮するが、その範囲外（分布外データ）にうまく適応できていないのだと解釈できる。また、後者のような記号処理についても膨大な組み合わせが実現可能な系については、学習時に見たデータに適応するだけでは不十分であるため、同様の問題が生じているのではないかと考えられる。つまりいずれの問題も、学習データの範囲内では高性能を発揮するものの、背後のより一般的な法則を捉えることができていないということに帰着する。

この問題は膨大な学習データがあれば緩和されるはずであり、実際に GPT-3 のような言語モデルは膨大な量の学習データを用いているし、近年は画像においても数億枚という膨大な学習データを用いて学習させたモデルが高性能を発揮している [12]。しかし、このようなアプローチによって精度の向上は実現できているものの、これまで困難であった情報処理（数式の処理など）が実現できるようになっているわけではない。また、根本的な問題意識として、知的エージェントにとって観測可能なデータは常に制限されており、有限の学習データから可能な限り汎用的な法則を発見する必要がある。実際に、人間の幼児はインターネット上の膨大な量のテキストを集めて言語を学習する必要はないし、数億枚もの画像から学ぶ必要もない。

これを実現するには、有限の学習データから可能な限り汎用的な法則を発見・抽出する必要がある。生物はこれを直観的に行っていると考えられるし、人間はそれに加えて言語や方程式

<sup>\*5</sup> [https://twitter.com/AravSrinivas/status/1287284785870602246?ref\\_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1287284785870602246%7Ctwgr%5E7Ctwcon%5Es1\\_c10&ref\\_url=https%3A%2F%2Fledge.ai%2Fchugai-digital-day%2F](https://twitter.com/AravSrinivas/status/1287284785870602246?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1287284785870602246%7Ctwgr%5E7Ctwcon%5Es1_c10&ref_url=https%3A%2F%2Fledge.ai%2Fchugai-digital-day%2F), 2021 年 10 月アクセス

<sup>\*6</sup> ここでは新たに与えられた課題に適応する能力を指す。

といった記号的な論理によって世界を捉えている。汎用的な法則を発見するための唯一の方法は、現象や入力情報（知覚）が生成される過程に何らかの仮定を置くことである。そしてそれが将来観測する全ての情報に対して当てはまること、つまり世界の法則を反映したものになっていなければならない。こうした議論は Goyal らによって詳しく行われている [13]。

このような生成過程に対する仮定は、例えば集合に対する可換性や、移動や回転に関する物体の同変性であったり、場合によっては「リンゴは赤い」というナイーブな仮定でも良い。これは普遍的な物理のルールではなく、場合によっては明らかに間違っているだろうが、グレースケールの画像から実際の様子を想像する場合にはこの仮定は多くの場合で有効である。こうした仮定、もしくは解に対する制約は学習アルゴリズムの補助となるものであり、**帰納バイアス (inductive bias)** と呼ばれている<sup>\*7</sup>。上記の例以外にも、「物事の説明は可能な限り単純であるべきだ」とする有名な指針であるオッカムの剃刀もこの帰納バイアスの一種として捉えることができる。

理想的にはこのような帰納バイアスとなるルールや前提知識、何らかの仮定は外部から与えるのではなく、学習アルゴリズムが学習データから自動で獲得することが望ましい。しかし学習データのみから未知のデータに対しても当てはまる普遍的なルールを獲得することは現実的には難しい場合も多く、結局は学習データが説明すべき対象をきちんと網羅しているかどうかや、与えるデータ量の多寡によって決まる。この観点から言えば先に述べたような巨大な言語モデルや画像認識モデルは、膨大な学習データによってほとんどの状況を事実上網羅的に経験しているために汎化性能が高くなっていると捉えることができる。これは Hasson らが *direct-fit* と呼ぶものに近い [14]。また、先に述べたような記号処理（素数の抽出など）や手続き的な処理が未だ十分に実現できていないのは、こうした課題では生じるパターンが組み合わせによって幾何級数的に増えるため、膨大な学習データを用いるアプローチでは対応しきれないためだと考えられる。

そのため、これらの点、つまり記号処理の性能や学習効率<sup>\*8</sup>を向上させるためには適切な帰納バイアスを機械学習モデルに対して導入してやることが重要となる。自然知能（人間や動物）がどのように帰納バイアスを獲得し、深層学習よりも効率的な学習を行っているのかは不明だが、進化の過程で時間をかけて生得的に適切な帰納バイアスを持つようなハードウェア（脳など）を獲得したり、人間については先の世代から継承される言語によっても学習に適切

---

<sup>\*7</sup> この単語の原義としては、何らかの帰納 (induction) によって得られる解に対する選好 (bias) を与えるものである。

<sup>\*8</sup> 目標の性能を実現するために必要な学習データの量

な制約がかかっているものと考えられる。つまり想定される選択肢としては膨大な学習データと時間を用いて自動で適切な帰納バイアスを獲得するか、人間における言語のように外部的に与えられた帰納バイアスを活用することになる。前者については生物の知能の創発という観点で興味深いものであるが、人工知能の改良という観点ではこれは遠回りとなる恐れがある。また、全くの仮定なしに学習を行った場合、仮に性能が発揮できたとしても我々が期待したものとは別の解を獲得する可能性がある。例えば移動（歩行）の仕方の学習を試みる場合、坂では歩かずに転がり落ちる方が速いため、そちらを解として選んでしまうことが起こり得る。このような点も踏まえ、先に述べたように適切な帰納バイアスを機械学習モデルに対して導入することが重要だと考えている。

帰納バイアスと分類できるものは多岐に渡り、様々な粒度のものが考えられる。例えば回帰における L1 正則化のような具体的な数学的制約もその一種であるが、畳み込みニューラルネットワークにおける並進不変性<sup>\*9</sup>のようなこの世界の性質を反映した抽象的でマクロな法則もまたその一種である。本研究では、この世界の重要な性質として、**構成性 (compositionality)** に注目した。我々は構成性を考慮することが、機械学習モデルが高い汎化能力を実現する上で重要な帰納バイアスの一つになると考えている。以降の節では構成性に関して詳しく述べ、なぜこれが重要な帰納バイアスとなり得るのかについて議論を行う。

### 1.1.2 構成性の原理

**構成性の原理 (the principle of compositionality)** は言語学や数理論理学の分野で考案された概念で、「複雑な表現の意味は、その構造と構成要素の持つ意味によって規定される」というものである<sup>\*10</sup>。これは自然言語や数学などの記号体系に見られる性質であり、例えば「赤いリンゴ」という表現は「赤い」と「リンゴ」という構成要素の組み合わせによって実現されている。要素の組み合わせによる表現力は非常に強力である。ある  $N$  個の要素を持つ集合  $\chi = \{c_1, c_2, \dots, c_N\}$  を仮定し、 $\chi$  の部分集合がそれぞれ別の意味を持つとすると、これは冪集合で  $2^N$  個の異なる表現が得られる<sup>\*11</sup>。構成性を考慮しない場合はこの全ての組み合わせを 1 つ 1 つ全く別の表現として捉える必要があるが、考慮できる場合は  $N$  個の要素の意味を把握するだけで済む。そして前者の場合は経験したことのある組み合わせしか意味を理解できない

<sup>\*9</sup> 被写体が視野内（画像内）で平行移動してもその意味や特性は変化しないという性質

<sup>\*10</sup> Stanford Encyclopedia of Philosophy: Compositionality <https://plato.stanford.edu/entries/compositionality/>, 2021 年 10 月アクセス

<sup>\*11</sup> 自然言語の場合は順序による意味の変化も起きるため、得られる表現はさらに多くなる。

が、後者の場合は知っている概念の個数に応じて、これまでに経験したことのない場合でも理解可能となる。これはつまり未経験の観測に対する汎化能力である。

この構成性の原理は自然言語や数学等の記号的な体系に限定的に当てはまるものではなく、我々の住むこの世界自体もまた構成的である。順序としてはむしろ、まずこの世界が構成的であり、その中で可能な限り多くの事象を表現するために、記号体系が構成性を獲得したものと解釈できる。上述の記号体系における構成性と同様に、任意の系はいくつかの「要素」の組み合わせとして解釈することが可能で、その組み合わせ方が変われば別の系が実現される。この「要素」は物体のような実体を持ったものだけでなく概念的な存在でも構わず、その粒度についても大小様々なものが想定される。以降は Greff らの議論 [2] に従い、これをオブジェクト (object) と総称することにする。

人間は知覚した情報を直接扱うのではなく、オブジェクトのような既知の概念を切り出し、その組み合わせとして全体を認識している。これは視覚の場合に顕著で、視野の全体に渡って詳細を認識することはできず、視野の中央のみでオブジェクトの意味的な情報を認識し、これを繰り返すことで全体（複数のオブジェクトが組み合わさった系全体）を認識している [15]。このような構成的な認識を行うことは、膨大な数の実現され得る系を少数の概念に基づいて認識したり、因果関係の推論を行うために有効な手段である。また、組み合わせの単位となるオブジェクトは効率性の観点から、観測の中に頻繁に登場することや、組み合わせるための物理的・概念的な独立性を持つことが重要である。こうした意味で、オブジェクトとして扱われる対象は何らかの概念と結びついた、記号的な存在である。

このような背景から、構成性を考慮した認識や思考の過程を実現することは人工知能においても高い汎化性能の実現や、記号的な処理を実現するために重要である。より具体的な利用例としては物体や概念間の関係性を考慮し、因果関係や系の発展の推測を行う relational inference<sup>\*12</sup>のような課題にとって重要である。relational inference の例を図 1.1 に示した [1]。このような処理は系全体を一つの対象として認識しては難しく、個々の構成要素の物理的性質や、それらの相互作用を捉えていなければ正しい処理ができない。この relational inference のような情報処理は様々な応用の基礎となるはずで、構成的な認識の実現はロボティクスやゲーム AI、自動運転など、知的エージェントが利用される様々な分野の発展の鍵になると考えている。

---

<sup>\*12</sup> 日本語において一般的な訳語が存在しないため、英語表現とした。

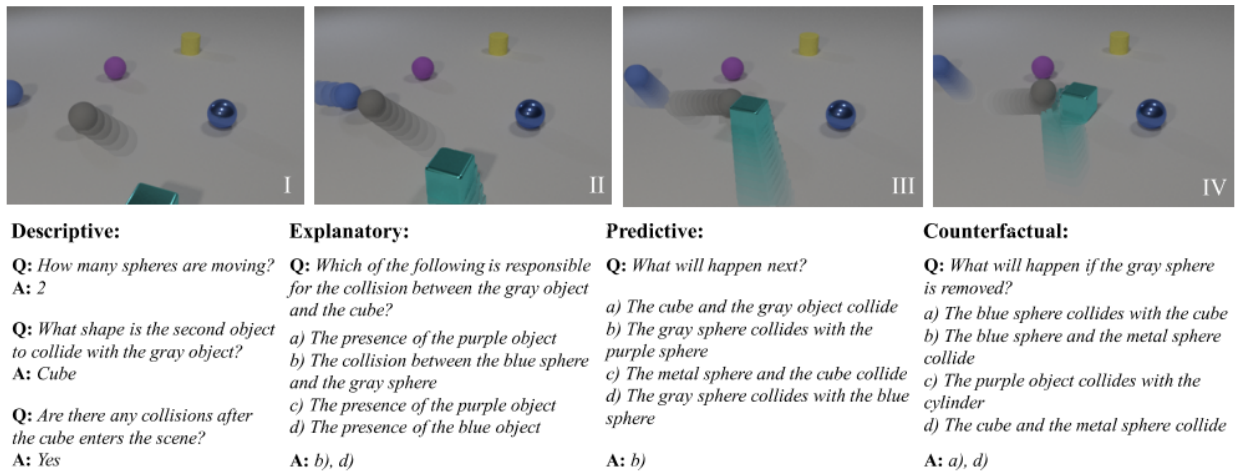


Fig.1.1 構成性を考慮することが重要だと考えられる，relational inference の課題例 [1]. 複数の物体が含まれる系の動画に基づいて，質問への回答や説明を行う課題となっている．

### 1.1.3 深層学習と表現学習

深層学習は学習データから入出力関係を学ぶことによって，自動でその課題に適した特徴量を獲得することが可能である．深層学習以前の情報処理では専門家が作成した特徴量を利用することが主流であったが，現在は深層学習が多くの分野でより高い性能を発揮している．深層学習のようにデータから必要な特徴ベクトル（特徴量）を自動で獲得することは**表現学習 (representation learning)**と呼ばれている．学習する入出力関係は様々で，教師あり学習の場合は入力データと予測したい教師ラベルの関係性を学習することになり，教師なし学習であれば何らかの課題，例えば入力画像とできるだけ同じ画像を出力させる再構成課題に基づいてモデルの学習を行うことになる．

表現学習は入力全体を一つの対象として扱い，全体に対応する特徴ベクトルを獲得するものが基本的である．しかしこれは先に述べた構成性とも関連する問題を生じる．ここで示すのは Superposition Catastrophe という問題で，これは古くから提起されている Binding 問題に関連するものである．この概要を図 1.2 に示した [2]．図中の a や b のように単体の対象のみを扱う場合は全体を一つの特徴ベクトルで表現しても問題ない．しかし c のように複数の物体が含まれる場合，同様の一つの特徴ベクトルで表現した場合には特徴どうしが衝突し，「赤であり緑でもある」という不明瞭な表現になってしまう．これは観測した特徴をどのように変数と結びつける (bind) するかという問題であり，Binding 問題の一種となっている．これを解決す

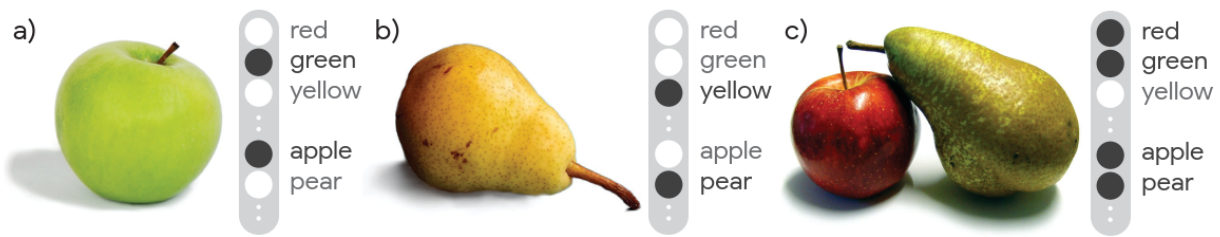


Fig.1.2 Superposition Catastrophe について示した図 [2].

るには、画像中のリンゴと洋梨がそれぞれ別の物体であることを正しく認識し、それぞれについて個別に特徴ベクトルを獲得する必要がある。

本研究では、このような表現学習の課題を構成性という観点から取り組む。特に本研究では深層生成モデルによる構成的な表現学習手法を扱うことを考えており、以降の節では深層生成モデルと構成的な表現学習について述べる。

#### 1.1.4 深層生成モデルによる認識

近年、知的エージェントが環境を認識する枠組みは深層生成モデルによって実現されている。生成モデルは、観測データが未知の確率分布に従って生成されていると仮定し、その生成過程を確率モデルによって表現する手法である。ここで、確率分布の近似を深層学習によって行う場合に、特に深層生成モデルと呼ばれている [16, 17]。観測データの生成要因として外部からは観測できない確率変数である、潜在変数を仮定することもある。この潜在変数を観測から推論することによって、観測データの性質を捉えることができる。生成モデルは従来から存在するものであるが、深層学習の登場によって高次元の確率分布を扱うことができるようになり、実際に画像や音声のような実データを扱うことができるようになった。潜在変数として入力データの特徴を得るだけでなく、画像や音声などの生成過程をモデル化し、潜在変数をサンプリングすることで、観測には含まれない新たなデータを生成することも可能になっている。新たなデータを生成することは、環境中の未観測の部分を予測したり、何らかの変化が加わった後の状況を予測することを可能とする。これは知的エージェントが自らの中に環境のモデル（シミュレータ）を持つことに相当し、そのような環境のモデルは分野によって「世界モデル（World Model）」や「内部モデル」、「メンタルモデル」などと呼ばれている [18, 19, 20, 21]。環境のモデルを持つことの利点としては、モデルから生成したデータを学習に利用し、行動を

学習する際のサンプル効率<sup>\*13</sup>を向上させることや、その環境内での汎化性能の向上などが挙げられる。生成したデータを学習に利用することは、いわゆるイメージトレーニングと解釈することが可能で、効率的な学習を実現するものとなっている。

### 1.1.5 深層生成モデルによる構成的な認識

構成的な認識とは、ある系を何らかの構成要素（オブジェクト）の組み合わせとして認識することである。深層生成モデルにおいてこれを実現する 1 つの方法は、複数の潜在変数を仮定し、それぞれが個々の構成要素を表現することである。深層生成モデルの多くの研究では観測全体を単体の潜在変数によって表現しているが、近年シーン解釈モデル（Scene Interpretation Models）と呼ばれる手法が登場している [22, 23, 24, 25, 26, 27, 28, 29, 30]。シーン解釈モデルでは複数の潜在変数を仮定し、それぞれが観測に含まれる個々の物体を表現するように学習を行う。各々の潜在変数が個々の物体に対応するため、入力となる複数の物体を含むシーンに対して潜在変数の推論を行うことで、個々の物体の認識を行うことができる。シーン解釈においては潜在変数が物体の位置情報や意味的な情報を保持しており、位置情報はバウンディングボックスやセグメンテーションマスクなどによって表される。そのため、潜在変数の推論によって教師なし学習でこうした物体検出を行うことが可能である。また意味的な情報については、潜在変数が物体に関する幅広い情報を保持することになるため、これを特徴ベクトルとして、様々な downstream task<sup>\*14</sup>に利用することが可能である。downstream task の例としては先に述べたような relational inference が挙げられる [31]。

シーン解釈の応用例としてはロボティクス分野が考えられる。ロボットアームの制御を人間によるティーチングではなく、学習によって獲得する場合を考える。Pick-and-Place 課題<sup>\*15</sup>を解く場合、ロボットは物体を認識し、対象となる物体の把持点を計算し、アームを動かして物体を把持し、目的地に移動させる必要がある。これに対して現在取られている方法は大きく 2 通りあり、1 つは物体検出や把持点の計算、アームの制御など別々に構築した処理をパイプラインとして繋げ、一連の動作を実現する方式、もう 1 つは一連の過程を end-to-end な学習<sup>\*16</sup>によって行う方式である [32, 33]。前者の方式の欠点として、一連の処理のどこかで失敗してし

<sup>\*13</sup> 与えたデータ量に対して、実現できた性能の比率を指す

<sup>\*14</sup> 事前に学習されたモデルや特徴量を用いて解く様々な課題のこと。特に追加の学習を行わずに解く場合にこう呼ばれる。

<sup>\*15</sup> ロボットのベンチマークとしてしばしば用いられる、机や棚などにある物体を持ち上げ、目的地に移動させる課題。

<sup>\*16</sup> 複数のプロセスを個別に学習するのではなく、全体を一度に学習すること



まうと、やり直しや立て直しが困難なことが挙げられる。また、物体検出には教師ラベルの作成が高コストな instance segmentation を用いる必要がある。後者の方式は、やり直したり細かい動作を行うことが可能な点でより望ましいが、現状では end-to-end な学習を行う場合、物体の認識に課題がある。前者の方式のように instance segmentation を利用することもできるが、それができない場合は、認識部分の学習効率の低さや、視点が少し変わると認識に失敗してしまうという問題が生じる。シーン解釈モデルのような物体に関する表現学習手法の発展は、こうした課題の解決に貢献することが期待される。

### 1.1.6 概念的な構成性と空間的な構成性

空間的な構成性に着目し、ある系を物体の組み合わせとして扱う手法は一般に物体中心表現学習 (object-centric representation learning) と呼ばれている。物体中心表現学習といった場合は深層生成モデルを用いているとは限らず、環境のモデル化を行わない手法も存在する [34, 35, 36, 37]。ただし、こうした手法では新たなデータの生成や環境のモデル化は行うことができず、利用目的や思想がシーン解釈とは少々異なる。そして、物体中心表現学習の中で、深層生成モデルを用いて認識を行うものをシーン解釈と分類することができる。また本来、組み合わせの単位としては実体のある物体だけでなく、概念的なものでも構わない。その中で、上述のシーン解釈は特に空間的な構成性に注目し、実体のある存在のみを対象としたものとなっている。

概念的な構成性について考える研究も存在し、これは Disentangled Representation Learning<sup>\*17</sup>もしくは Disentanglement と呼ばれている [38, 39]。これは、異なる性質を持つ独立した概念をそれぞれ別の変数や特徴ベクトルの異なる次元で表現することである。Disentangled Representation が深層生成モデルにおいて実現されることが確認された当初<sup>\*18</sup>は正確な定義が行われておらず、後に Higgins らによって形式的な定義が試みられている [40]。

Higgins らの研究においては物理学での議論を基に、自然界でしばしば見られる「環境中の特定の対象 (object) に変化を与えるが、それ以外には変化を及ぼさないような介入」を対称変換と定義している。そして、Disentangled Representation はこれを用いて、「ベクトル表現がいくつかの部分空間に分割でき、それぞれの部分空間が対称変換と対応しているか、もしくはある対称変換によって独立に変換され得ること」と定義している。この定義におい

---

<sup>\*17</sup> 「もつれない表現学習」と日本語訳される場合もあるが、英語表記がそのまま用いられることが多い。そのため本論文内でも英語表記を採用した。

<sup>\*18</sup> 深層生成モデルとして実現した最初の手法は [38] であるが、Disentanglement の概念自体の初出は諸説ある。

て対称変換が適用される単位は“object”となっているが、これも実体を持つ物体だけでなく概念的な対象も含めたものである。具体的な例として次のようなものが考えられる。ある物体について色や形状、テクスチャといった性質を考えた場合に、色のみが変化し、その他の性質は変わらないという変化が想定される。このようにある性質の変化（対称変換）がベクトル表現の部分空間の変化のみにとどまり、他には変化を与えないようになっている表現が Disentangled Representation である。こうした概念的な構成性は枠組みとしては空間的な構成性も内包するため、シーン解釈もこの1つと考えることもできる。しかし、現状では Disentangled Representation の研究においては空間的な構成性は十分に獲得されていない。また大規模実験の結果を根拠に、適切な帰納バイアスなしには Disentangled Representation は実現できないとする研究も存在している [41]。そのため、この観点からも帰納バイアスとしての空間的な構成性を考慮することは重要であると考えられる。

### 1.1.7 シーン解釈手法の課題

シーン解釈の利点や期待される応用例について述べてきたが、教師なしで物体に関する汎用的な表現を獲得するという問題設定は難しいものである。現状では CG で作成されたトイデータや簡単な図形の組み合わせなど適用範囲が限られており、実画像のように複雑な対象はうまく扱えない。また、局所解に陥って目的とする表現が得られない場合があり、学習の難しさや不安定性に関する問題もある。

我々は先に表現学習における構成性の帰納バイアスとしての重要性を論じ、シーン解釈手法を研究対象とすることを述べたが、こうしたシーン解釈の問題もまた帰納バイアスの観点から考えることができる。組み合わせの単位となる対象<sup>\*19</sup>をどのように選ぶかは自明ではなく、将来の観測に対する説明性が高くなるような組み合わせ方を経験的に設定する必要があるが、与えられたデータセットや限定的な仮想環境の範囲内では限界があり、それが人間の基準と一致する保障もない。人間の場合も事前情報なしにこうした基準を獲得しているわけではなく、自然言語や数学などの既に確立された記号体系の補助によって学習していると考えられる。このように物体の概念は経験的な基準に基づいたものであり、何らかの前提知識を与えない限り人間と同様の基準を獲得することは必ずしも期待できない。そのため、シーン解釈のように教師なしで物体の概念を獲得するためには構成性だけでなく、物体を認識するためのもう少し具体的な帰納バイアスを導入する必要があると考えている。

---

\*19 シーン解釈においては主に物理的な実体としての物体

## 1.2 本研究の目的

ここまで構成的な認識の重要性を述べ、それが深層生成モデルを用いたシーン解釈手法によって実現されていることや、現状の課題について説明してきた。本研究は、学習の不安定性や適用範囲の狭さといった現状のシーン解釈手法の課題を解決することや、物体の認識に関する性能向上を主な目的とするものである。本研究の論点の根本である構成性に加え、さらに物体を認識するために必要な帰納バイアスをモデルに導入することにより、これらを実現することを考える。

1. 背景情報を利用したシーン解釈手法の提案
2. 自己教師あり学習と Transformer を用いたシーン解釈手法の提案
3. 階層的な表現学習による 3 次元空間の認識と生成を行う手法の提案

各論はいずれも深層生成モデルによる構成的な認識を実現するために手法を提案するものである。各研究は帰納バイアスの導入によってシーン解釈手法における課題を解決するという目的に加え、どのような帰納バイアスを導入すれば良いのかを探る目的もある。

研究 1 は物体を含まない場合、つまり背景の画像集合を補助情報として利用することで、物体に着目する手がかりを与えるものである。ただし、古典的な画像処理のように処理の対象となる画像と一対一で対応する背景の画像を与えるのではなく、物体を含む画像集合（主な処理の対象）と物体を含まない画像集合（背景）という独立な 2 つの画像集合を用いるものである。この研究の意義として、物体部分に注目させるような帰納バイアスを明示的に導入した場合、どのような改善が見られるかを確認する意味がある。

次に、研究 2 においては対照学習に基づく自己教師あり学習を利用することを考えた。このような学習を行った場合、モデルは画像中の特徴的な部分に注目することになり、それが結果的に物体の認識に必要な特徴量を獲得することにつながることを期待される。研究 1 では新たなデータを与え、明示的に物体に注目させるような解に誘導したが、研究 2 では新たなデータは用いずに学習の工夫によって物体の認識に必要な特徴量を獲得することを期待している。

研究 1 と 2 では静止画一枚のみを与え、物体を認識するという二次元空間での課題となっていたが、研究 3 では多視点の入力に拡張した問題設定に取り組んでいる。これは、複数の物体を含む、ある 3 次元的な空間に関して複数視点からの画像を与え、観測していない任意の視点からの画像の予測と物体認識を行うという課題である。また、研究 1 と研究 2 ではデータの与

え方や学習の工夫による帰納バイアスの導入を試みたが、研究 3 ではモデル構造に前提知識を組み込む形の導入を行う。具体的には、ある 3 次元的な空間は、そこに含まれる物体それぞれの表現と、その配置に分割することが可能であると考え、これを考慮した階層的な確率モデルを導入し、より適切な環境のモデル化を可能とする手法を提案するものである。

### 1.3 本論文の構成

本論文の構成は以下の通りである。

第 2 章では生成モデルと深層学習を組み合わせた技術である、深層生成モデルに関する前提知識を述べる。

第 3 章では、深層生成モデルによる構成的な認識を行う手法である、シーン解釈モデル (Scene Interpretation Model) と、その関連領域について説明する。

第 4 章では、研究の背景や関連研究を踏まえ、本研究の新規性や意義について述べる。

第 5 章では、背景情報を利用したシーン解釈手法を提案する。既存のシーン解釈手法は複雑なテクスチャを含む画像や実画像を適切に扱うことができず、物体ごとの表現を獲得することができないという問題がある。本章では学習時に背景の集合 (物体を含まないシーンの集合) を利用することによって、物体として認識すべき対象についての帰納バイアスを与え、シーン解釈手法の適用範囲を拡張する手法を提案する。まず、背景と数字に対して複雑なテクスチャを適用した手書き数字のデータセットである Textured-MNIST データセットなどを利用し、背景と物体を 1 つずつ含む場合の有効性を検証した。そして次に、背景と複数物体を含む場合について、既存のシーン解釈手法を拡張する手法として提案手法が利用可能であることを確認する。

第 6 章では、自己教師あり学習と Transformer を用いたシーン解釈手法を提案する。これまでに述べた通り、シーン解釈では物体として認識すべき対象に関する情報が十分に与えられておらず、何らかの形で必要な帰納バイアスを与える必要がある。5 章の研究 1 では背景についての情報を新たに与えていたが、理想的には新たなデータを使わずに性能を向上させたい。そこで、本章では自己教師あり学習の利用によって追加データを用いずに物体に関する帰納バイアスを与えることを考える。自己教師あり学習は、クラス分類やセグメンテーションといった、目的とする課題そのものではなく、適切に設計された何らかの事前課題を解くことによって様々な課題に汎用的な特徴量を獲得する手法である。この事前課題は例えば回転対称性を考慮したものや、画像変換に対する被写体の不変性など、物理的な前提知識に基づくものとなっ

ている [42, 43]. 本章では, これによって物体に関する帰納バイアスが獲得されることを期待し, シーン解釈への自己教師あり学習の導入を行う. また, 既存手法のモデル構造はシーン解釈に適していない側面があり, これはモデル構造への過剰な制約, つまり逆方向の帰納バイアスと解釈できる. そこで本章では, これを解消するために Transformer[44] を用いたネットワーク構造の提案も行う.

第 7 章では, 複数視点の画像を入力として与える問題設定について考える. これは 3 次元空間を仮定していることに相当し, Neural Scene Rendering[45] と呼ばれる問題設定の一種である. 既存手法ではシーン中の物体は潜在変数が従う確率分布から独立に生成されるものと仮定されており, 物体間の関係性やシーン全体の空間的な構成についての表現学習は行われていない. しかし 3 次元的な空間において各々の物体の性質と, その空間的な配置は独立したものである. 同じ物体の組み合わせでも配置が変われば別のシーンが構築されるし, 同じ配置でも異なる物体に置き換わることが想定される. そこで本章では, シーン全体の空間的な構成と, 物体の表現を分け, それぞれについて表現学習を行う手法を提案する. これにより提案手法では, 既存手法において困難であった新たなシーンの生成が可能となり, シーンの認識精度も向上することを確認する. 特にシーン全体の構成に関する表現学習を行い, 環境のモデル化を行ったことで, 観測として与えられる視点の数が少ない場合の性能低下が小さいことを確認する.

第 8 章では, 3 章から 7 章までの研究内容を踏まえて, 今後の課題や展望について述べる. 最後に, 第 9 章では, 本論文のまとめを述べる.

## 第 2 章

# 深層生成モデルに関する前提知識

### 2.1 生成モデル

あるデータ  $\mathbf{x}$  について、それが従う確率分布  $p(\mathbf{x})$  を考える。この確率分布は仮想的なもので、現実的にはこのようなデータを完全に説明できる「真の分布」が存在する保証はない。そのような真の分布自体を計算することはデータ数が無限に必要となり、不可能である。そこで、真の分布をパラメータ  $\theta$  を持つ確率分布  $p_\theta(\mathbf{x})$  で近似することを考えたとき、この分布は**生成モデル (Generative Models)** と呼ばれる。

生成モデルについて、データの生成要因となる隠れ変数  $\mathbf{z}$  を仮定することが可能で、これは潜在変数 (Latent Variable) と呼ばれる。潜在変数を導入した生成モデルは、 $\mathbf{x}$  と  $\mathbf{z}$  の同時分布を  $p(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$  と表すことができる。ただし  $p_\theta(\mathbf{z})$  は実際には学習パラメータを持たず、サンプリングが簡単な分布を事前分布として仮定することが多い。その場合、 $\mathbf{z} \sim p(\mathbf{z}), \mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{z})$  とサンプリングすることで、新たなデータ  $\mathbf{x}$  を生成することが可能である。

この潜在変数  $\mathbf{z}$  はクラスラベルのような離散変数を仮定することも、実数値ベクトルとして分散表現とすることも可能である。分類問題を解く際に一般的に用いられる識別モデルとは  $p_\theta(\mathbf{x}|\mathbf{z})$  の部分に相当し、データ分布のモデル化を行わずに、この条件付き分布を決定論的な手法で直接モデル化したものと解釈できる。そのため一般に識別モデルの方が学習が容易であるが、データ分布をモデル化する利点として、訓練データ集合に含まれない新たなデータを生成することが可能であることや、観測データの欠損した部分を補完することが可能であるといった点が挙げられる。

潜在変数の設計については様々な方法が考えられ、例えば複数の潜在変数  $\mathbf{z}_1, \dots, \mathbf{z}_m$  を仮

定し,

$$p(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_m) = p_\theta(\mathbf{x}|\mathbf{z}_1) \prod_{i=1}^{m-1} p_\theta(\mathbf{z}_i|\mathbf{z}_{i+1})$$

という階層的な確率モデルを構築することができる。また、生成モデルにガウス分布のような再生性を持つ分布を仮定した場合、その和もまたガウス分布となるため、以下のような定式化も可能である。

$$p(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_m) = \sum_{i=1} p_\theta(\mathbf{x}|\mathbf{z}_i)$$

これは混合ガウスモデル (Gaussian Mixture Model) と呼ばれるものである。このように様々な確率モデルを仮定することが可能であり、データに対する仮説に基づいて設計されることになる。

### 2.1.1 生成モデルの学習

生成モデルのパラメータを学習するためには、仮定した生成モデルを真の分布に近づければ良い。つまり、距離の指標を  $D$  として

$$\operatorname{argmin}_{\theta} D(p(\mathbf{x}), p_\theta(\mathbf{x}))$$

として学習を行えばよいが、真の分布は不明であり、実際には訓練データの分布  $p_{data}(\mathbf{z})$  (= 経験分布) との距離を最小化することになる。また、この確率分布間の距離の指標として一般的にはカルバックライブラーダイバージェンス (Kullback-Leibler Divergence) が用いられる。つまり、

$$\operatorname{argmin}_{\theta} KL[p_{data}(\mathbf{x}) \parallel p_\theta(\mathbf{x})]$$

を行うことになる。この目的関数は対数尤度の最大化と等価で、訓練データを  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_N\}$  とした時、以下のように表される。

$$\begin{aligned} \operatorname{argmax}_{\theta} \frac{1}{N} \sum_{i=1}^N \log p_\theta(\mathbf{x}_i) &= \frac{1}{N} \sum_{i=1}^N \int \log p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \\ &= \frac{1}{N} \sum_{i=1}^N \int \log p_\theta(\mathbf{x} | \mathbf{z}) p_\theta(\mathbf{z}) d\mathbf{z} \end{aligned}$$

上記のようにある変数を積分によって取り除くことを周辺化 (marginalization) というが、一般に  $\mathbf{z}$  は高次元であるため、単純に積分を行うことは困難 (intractable) である。この解決策としては、マルコフ連鎖モンテカルロ法 (MCMC) などのサンプリング法を用いて分布からサンプリングを行うか、以下に述べる変分推論という方法を用いることが一般的である。

ここで、 $q_\phi(\mathbf{z})$  という生成モデルとは異なるパラメータを持った任意の分布を考える。これを用いると対数尤度は

$$\begin{aligned}\log p_\theta(\mathbf{x}) &= \int p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \\ &= \log \int \frac{q_\phi(\mathbf{z})}{q_\phi(\mathbf{z})} p_\theta(\mathbf{x}, \mathbf{z}) d\mathbf{z} \\ &\geq \int q_\phi(\mathbf{z}) \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z})} d\mathbf{z}\end{aligned}$$

と表せる。この不等式はイェンセンの不等式によるもので、凸関数 (ここでは対数関数) の性質を用いている。これは  $q_\phi(\mathbf{z})$  で期待値を取っていると見なすことが可能で、以下のように書き換えられる

$$\log p_\theta(\mathbf{x}) \geq \int q_\phi(\mathbf{z}) \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z})} d\mathbf{z} = \mathbb{E}_{q_\phi(\mathbf{z})} \left[ \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z})} \right] = \mathcal{L}_{ELBO}$$

これは対数尤度  $\log p_\theta(\mathbf{x})$  (evidence) の下界であるため、Evidence Lower Bound (ELBO) と呼ばれる。ここで ELBO と対数尤度の等号が成り立つのは、真の事後分布  $p_\theta(\mathbf{z} | \mathbf{x})$  と  $q_\phi(\mathbf{z})$  が一致する場合 (両者の KL ダイバージェンスが 0 となる時) である。ELBO の最大化は生成モデルのパラメータ  $\theta$  と近似事後分布のパラメータ  $\phi$  を交互に更新すること (EM アルゴリズム) で行うことができる。しかし真の事後分布  $p(\mathbf{z} | \mathbf{x})$  は基本的には解析的に定まらない。そこで

$$q_\phi(\mathbf{z}) = \prod_i q_\phi^i(\mathbf{z}_i)$$

と  $q_\phi$  が積の形に分解できると仮定し、単純化する方法がある。これは平均場近似 (mean-field approximation) と呼ばれるものである。これを用いて最適化を行うことを変分推論 (Variational Inference) と呼ぶ。

## 2.2 深層生成モデル

生成モデルについて前節で説明したが、高次元の確率分布をモデル化することは困難であった。確率分布の近似に深層ニューラルネットワークを用いることでこれを可能にしたのが、深層生成モデルである。深層生成モデルとしては Variational Autoencoder (VAE)[16] や Generative Adversarial Network (GAN)[17], Flow-based models[46], Energy-based models[47, 48] などがある。以下では本研究で利用している VAE について説明する。



### 2.2.1 変分自己符号化器 (Variational Autoencoder: VAE)

上述した変分推論において、確率分布の近似に深層ニューラルネットワークを持ちいたものが Variational Autoencoder (VAE) である。潜在変数の事前分布として  $\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mu, \sigma)$  を仮定し、生成モデルは  $\mathbf{x} \sim p_\theta(\mathbf{x} | \mathbf{z})$  とする。また、近似事後分布として  $q_\phi(\mathbf{z} | \mathbf{x})$  と、パラメータ  $\phi$  を持つ学習可能な確率的な写像を考える。

このとき ELBO は

$$\mathcal{L}_{ELBO} = \int q_\phi(\mathbf{z} | \mathbf{x}) \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} | \mathbf{x})} d\mathbf{z} = \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} \left[ \log \frac{p_\theta(\mathbf{x}, \mathbf{z})}{q_\phi(\mathbf{z} | \mathbf{x})} \right]$$

となる。これはさらに以下のように書き換えられる。

$$\mathcal{L}_{ELBO} = -KL[q_\phi(\mathbf{z} | \mathbf{x}) \parallel p(\mathbf{z})] + \mathbb{E}_{q_\phi(\mathbf{z} | \mathbf{x})} [\log p_\theta(\mathbf{x} | \mathbf{z})]$$

このとき、近似事後分布  $q_\phi(\mathbf{z} | \mathbf{x})$  と生成モデル  $p_\theta(\mathbf{x} | \mathbf{z})$  をそれぞれ別のニューラルネットワークでモデル化することが可能である。前者は入力から潜在変数への確率的な写像を行うため、エンコーダと呼ばれる。後者は潜在変数から入力の再構成（生成）を行うため、デコーダと呼ばれる。

ニューラルネットワークによってこれらをモデル化する上での問題点は、単に  $\mathbf{z} \sim q(\mathbf{z} | \mathbf{x})$  とすると微分を行うことができず、誤差逆伝播法による最適化ができなくなってしまうことである。そこで VAE においては固定の確率分布  $p(\epsilon)$ （ノイズ）と微分可能な関数  $f$  を用いて、 $\mathbf{z} = f(\epsilon, \phi)$ ,  $\epsilon \sim p(\epsilon)$  と計算を行う工夫（Reparameterization Trick）を用いている。エンコーダに仮定する分布をこのように表せるもの、例えばガウス分布とすることで微分不可能な操作を行わずに計算を行っている。ガウス分布の場合は、平均のパラメータ  $\mu$  と分散  $\sigma$  について、以下のようにモデル化することができる。

$$\begin{aligned} \mathbf{z} &= \mu + \sigma \otimes \epsilon, \quad \epsilon \sim p(\epsilon) \\ \mu &= f_\mu(f_\phi(\mathbf{x})), \\ \sigma &= f_\sigma(f_\phi(\mathbf{x})) \end{aligned}$$

ただし、 $f_\phi, f_\mu, f_\sigma$  はいずれもニューラルネットワークとする。デコーダは出力する画素値に応じて連続値ならばガウス分布が、二値ならばベルヌーイ分布などが仮定される。

事前分布には正規分布を仮定したため、 $\mathbf{z}$  をサンプリングすることで生成モデル（デコーダ）を用いてランダムに新しいデータを生成することが可能である。

### 2.2.2 自己符号化器 (Autoencoder: AE)

自己符号化器 (AE) は確率的なモデルではないが、VAE との関連性からここで簡単に説明を行う。自己符号化器も VAE と同様に入力から潜在変数への写像  $f_{enc}(\mathbf{z} | \mathbf{x})$  と潜在変数から入力への写像  $f_{dec}(\mathbf{x} | \mathbf{z})$  を持つ手法となっているが、いずれも確率分布ではなく決定論的なネットワークである。学習は再構成によって行われるが、いずれも決定論的なモデルであるため目的関数は単に  $\mathcal{L} = D(f_{dec}(f_{enc}(\mathbf{x})), \mathbf{x})$  と入力と再構成の距離を最小化することになる。ただし  $D$  は二乗誤差などで、任意の距離が適用される。

自己符号化器は元々深層学習の黎明期に重みの初期化を行う手法として利用されたり、入力データの次元削減を行って特徴量を獲得する手法として用いられてきたものである。潜在変数のサンプリングができないため、画像の生成は行うことができないし、VAE のような類似の入力が潜在変数空間でも近い位置に写像されるといった機能もない。とはいえ生成や潜在変数空間の表現が重要でない応用事例もあるため、現在でも様々な用途で利用されている。本研究の対象であるシーン解釈手法に関連する手法にも AE ベースのものが存在し、3 章において紹介する。

## 第 3 章

# 構成的な認識とその関連研究

### 3.1 物体中心表現学習 (Object-centric Representation Learning)

物体中心表現学習 (object-centric representation learning) とは主に視覚的な入力を対象として、そこに含まれる各物体の表現や、その位置情報を獲得するものである。基本的な定義は、画像  $\mathbf{x}$  を入力として、そこに含まれる各物体の表現  $\mathbf{z}_1, \dots, \mathbf{z}_K$  を推論し、セグメンテーションやバウンディングボックスを獲得することである。ただし、そのような決まった形式で物体の位置情報を獲得しない場合や [49]、物体ごとの表現を明示的に獲得しない場合もある [50]。文脈や研究分野によって用語の使い方や細かな定義は変化するが、概して物体についての認識を構築するための手法である。ここでは object-centric な手法を、まず生成モデルによるものと、それ以外に分類し、各節にて代表的な手法を紹介する。

### 3.2 生成モデルを用いた Object-centric モデル

本節では生成モデルに基づいた Object-centric モデルを紹介する。これらは主に VAE ベースの手法、AE ベースの手法、EM アルゴリズムベースの手法に分けることができる。以下では各手法とそれに属する研究について述べる。特に、物体の表現の推論を行うことができる手法はシーン解釈モデル (Scene Interpretation model) と呼ばれる。生成モデルを用いている場合は基本的にこれを行うことが可能である。また、複数の潜在変数を仮定した上で新しいシーンの生成のみに焦点を置いた GAN ベースの手法 [51, 52] も存在するが、シーン解釈モデルと呼ばれるのは特に潜在変数の推論によって物体の情報が得られる手法のみを指す。

### 3.2.1 VAE-based モデル

VAE を用いたシーン解釈モデルにはいくつかの方式が存在し、特に物体の位置の表現方法について大きく2種類に分類できる。1つはバウンディングボックスによるもの (Neurosymbolic モデル) [27, 53, 54, 30], もう1つはセグメンテーションによるもの (Scene Mixture モデル) [22, 23, 24, 25, 26] であり、もしくはこれらを併用したもの [29] もある。以下の節ではそれぞれの方式について説明する。

#### Bounding Box ベース (Neurosymbolic モデル)

バウンディングボックスを用いるシーン解釈の手法では、潜在変数をアフィン変換のパラメータとして拘束することで、矩形領域で物体の位置の位置を指定している。この手法では潜在変数を物体の位置 (where), 見た目 (what), 存在するかどうか (presence) など表現するよう構造化し、指定した矩形領域をそれぞれの潜在変数が担当し、表現を行う形になっている<sup>\*1</sup>。このように潜在変数の役割を拘束 (binding) した場合、推論するのはその値だけで済むため、表現の自由度が下がる代わりに拘束しない場合よりも推論や学習が容易になる。このように潜在変数に記号的な役割を明示的に割り当てたシーン解釈モデルは、Neurosymbolic モデルと呼ばれている。矩形で領域を指定するため、対象となる物体の形状が不規則であったり、大きさが多様な場合は扱いが困難である。そのため、例えば背景のような観測全体に広がるような対象については特別な扱いが別途必要となってしまうが、同程度の大きさの細かな物体が多数含まれる場合には有効な手法となっている。

この方式を最初に提案した Attend, Infer, Repeat (AIR)[27] では物体の数だけ反復的に自己回帰による推論を行う必要があり、物体の数に応じて計算量が増加してしまう。そのため、実質的には検出可能な物体の数が制限されていた。これを解決するために [53] では画像を予め格子状の矩形領域に区切り、その領域ごとに並列に物体検出を行う方式を採用した。しかし、このような方法は格子のサイズが敏感なハイパーパラメータとなってしまうことが [29] で示されており、物体の大きさが均一でなく、大きなものと小さなものを同時に含むような状況を苦手としている。特に背景のような入力全体に広がりを持つような対象についてはバウンディングボックスでの処理が困難であるため、[29] や [54] では別途背景を表現する機構を追加することで対処している。

---

<sup>\*1</sup> 存在するかどうかというのは、attention をかけた領域に物体が存在しない場合もあり、その際に不要な処理を避けるために用意されている。

また、物体ごとの構造化された表現しか持たないため、新たなシーンの生成（サンプリング）ができないという問題があった。Generative Neurosymbolic Machine (GNM)[30] はそれを解決するために、大域的な潜在変数を導入し、サンプリングが可能な neurosymbolic モデルを構築した。ただし上記のような手法の問題点は残っており、研究の余地があるものと考えられる。

### 空間的混合ガウス分布ベース（Scene Mixture モデル）

Scene Mixture モデルでは、それぞれの潜在変数がセグメンテーションマスクによって物体の領域を指定し、その領域ごとに個別に表現する方式となっている。Neurosymbolic モデルで潜在変数がアフィン変換のパラメータとして拘束されていたのに対し、こちらはそのような構造化は行われておらず、連続値のベクトル（分散表現）として潜在変数がモデル化されている。物体の形状の複雑さや大きさの多様性、背景の取り扱いといった点で自由度が高い代わりに、推論が難しくなる傾向がある。ここで、推論の難しさは 1 章でも触れたように、学習の不安定性や、各潜在変数が物体ごとの表現にならないような局所解の存在という形で現れる。

この手法ではセグメンテーションマスクによって切り出した領域を足し合わせることで全体を表現することになるが、これは混合ガウス分布でモデル化される。特に、ピクセルレベルで足し合わせを行うため、これは空間的混合ガウス分布モデル（Spatial Gaussian Mixture Models）と呼ばれ、シーン解釈の方式としては Scene Mixture モデルと呼ばれることが多い。

混合ガウス分布の確率変数の推論は典型的には EM アルゴリズムのような反復的な手法が用いられる。IODINE[25] はこの部分に Iterative Amortized Inference (IAI) [55] という反復的な推論を行う手法を用いている。これは物体の表現として推論した  $K$  個の潜在変数をそれぞれ更新ネットワーク（refinement network）で反復的に更新していくもので、更新するにつれ個々の物体を分担して表現するようになる。しかしこれは反復する回数だけ生成を行う必要があり、計算コストが非常に高くなってしまう。この点について、計算量の問題を緩和することを試みた研究も存在する [56]。また、IAI によって潜在変数は事前分布から離れてしまうため、事前分布からのサンプリングによる新たなシーンの生成も困難となる。

MONet[24] では反復的な最適化を用いる代わりに、マスクを先に決定論的なネットワークで近似し、それを用いて空間的混合ガウス分布の計算、つまり生成を行っている。しかし、マスクの近似が確率モデルとして定式化されていないため、これについても新たなシーンの生成が困難である。マスクの計算も確率モデルに組み込み、かつ潜在変数同士の関係を自己回帰によってモデル化することで新たなシーンの生成を可能にした手法が GENESIS である [26]。各

潜在変数を独立にサンプリングしてしまうと、物体同士の位置関係や影の方向などを考慮しないことになり、シーン全体として物理的な一貫性のない生成結果になってしまう。こうした問題を自己回帰による推論モデルと、事前分布の導入により解決している。

また、IODINE を制御に応用した例として OP3[57] がある。これは IODINE のような反復的な推論で物体を認識することに加え、物体同士の相互作用を明示的にモデル化し時間方向のダイナミクスを考慮した手法となっている。物体の相互作用を考慮したことにより、各フレームで IODINE を適用するよりも安定した物体の認識が可能になることを主張している。

### 3.2.2 AE-based モデル

潜在変数を確率変数ではなく、決定論的な変数としたもの、つまり Autoencoder ベースの手法も存在する。Slot Attention はそのような手法の1つで、 $K$  個のスロットに対応する表現を抽出したのち、反復的に更新を行い、それぞれのスロットが各々の物体を分担して表現するようにするものである。確率モデルではなくなっているため、潜在空間での表現学習や disentanglement については期待できないが、学習や推論は容易になる可能性がある。

### 3.2.3 Expectation Maximization(EM) ベースのシーン解釈手法

Tagger[58] や Neural Expectation Maximization (NEM)[22], Relational NEM[23] などの手法は EM アルゴリズムのような反復的な更新によって、物体ごとに表現が分割されるように学習を行う手法である。R-NEM は NEM を時系列に拡張し、物体同士の相互作用を考慮したものである。EM アルゴリズムは K-means を用いて画像のピクセルレベルでのクラスタリングに利用されることがあったが、これらはその延長となっている。これらの手法も上述の VAE ベースの手法と同様に混合ガウス分布を用いており、順序としてはこれらの研究の定式化が継承されたものと考えられる。しかし Tagger や NEM はグレースケール画像のみを入力としており、RGB 画像にスケールしないことが指摘されている。

## 3.3 識別モデルによる object-centric モデル

物体認識に関連する手法のうち生成モデルを用いないものは物体の表現を獲得しないものが多いが、ここではいくつかの関連研究を紹介する。ただし、ここで紹介するのは何らかの形で物体の表現や特徴ベクトルを獲得し、強化学習や制御への応用を試みた手法に限った。

Devin らの研究 [34] は、object-centric な手法をロボットアームの制御に活用したものであ

る。物体認識自体には一般的な教師あり学習の手法を用いているが、これにより与えられた課題と関係なく物体が含まれる領域に注目（meta-attention）し、さらに与えられた課題に合わせて各物体に attention するものとなっている。学習に教師データが必要であるため上述の手法とは少々異なるものだが、物体ごとに認識や方策の学習を行うことの利点を示した研究となっている。

Transporter[50] は制御や強化学習を目的とした手法であり、物体ごとの表現を獲得するわけではないものの、画像中の重要と考えられる点（=キーポイント）を抜き出し、画像特徴量と組み合わせることでサンプル効率の良い強化学習が可能となることを示している。

また、SCOFF[59] は物体の状態を記述する “*object files*” とその状態の更新の仕方についての外部的な知識 “*schemata*” を用意し、attention によってそれぞれを選択するという手法を構築している。この手法も物体についての記述的な（declarative）知識と、手続き的な（procedural）知識が分解できるという仮定に基づいて構築されたものであり、LSTM[60] や GRU[61] などの時系列を扱う手法の置き換えとして利用可能であるとしている。

### 3.4 Object-centric Representation の応用利用

物体ごとの表現を獲得する手法について主要なものを紹介したが、これらを認識モデルとして固定し、得られた表現を何らかの課題に利用する応用も行われている。Ding らの研究 [31] は、学習済みの MONet で獲得した表現と、文章入力をともに Transformer[44] に入力し、入力された質問に答えるような推論課題<sup>\*2</sup>を行うものである。このように物体認識と言語処理を組み合わせる方法が、全てを1つのネットワークでモデル化する方法よりも良い結果を出すことが報告されている。この研究ではこうした物体認識と言語処理を分けるアプローチを *neurosymbolic approach* と呼んでいる。上述の潜在変数を bind する *neurosymbolic* モデルとは少々異なるものだが、いずれも知覚的な処理と言語処理の折衷を試みるという点で共通している。

他にも強化学習の認識部分に *object-centric* なエンコーダを用いてサンプル効率を向上させる研究 COBRA[62] や、物体の配置の法則のような抽象的な関係を推論することを目指した研究 Constellation[63] などが存在する。

COBRA については認識部分の学習も行っているが、本節で紹介した他の2つの手法は学習

---

<sup>\*2</sup> 日本語では区別がなくなってしまうが、生成モデルの文脈での推論が *inference* の訳であるのに対し、この文脈での推論は *relation* の訳語である。*relation* については推理や理由付けと表現する方が直感的にはニュアンスが近いかもしれない。

済みの認識モデルを固定して利用している。より複雑な物体を認識するためには静止画から事前に学習するだけでなく、物体との相互作用や時間方向のダイナミクスを踏まえることが重要であると考えられる。その点において COBRA は興味深いものであるが、扱っているデータセットは単純な図形の組み合わせによる簡単なものに留まっている。獲得した物体の表現を利用するだけでなく、物体の認識を高度化するアプローチとして、このような推論やダイナミクスを考慮する研究を行うことも必要だと考えている。

### 3.5 教師なし画像処理手法に対する位置付け

その他に識別モデルを用いた類似の研究として、教師なしセグメンテーション [64] や salient feature extraction[65] などの教師なし画像処理が挙げられる。こうした手法は実画像にも適用されており、より応用可能性が高いようにも見える。しかし、こうした手法と上述の object-centric representation learning 手法の最も大きな差は、物体の表現や特徴量を獲得するかどうかにある。教師なしセグメンテーションのような手法は、あくまで入力画像からセグメンテーション結果を出力するものであり、分割した各領域の表現や特徴量は得られない。そのため、獲得した表現の強化学習や制御への利用を目指す 3.3 で紹介したような手法とは利用目的が異なるものとなっている。ただし教師なし画像処理手法の技術を object-centric モデルに活用することや、その逆についても可能であり、1つの研究の方向性として念頭に置く必要がある。



## 第 4 章

# 深層生成モデルによる環境の構成的な認識と生成

### 4.1 関連研究を踏まえた本論文の目標

2 章では深層生成モデルについての前提知識を説明し、3 章では物体中心表現学習の関連研究、特に本研究の主な対象となっているシーン解釈の既存研究について述べた。

1 章で構成的な認識の重要性と、それを実現するシーン解釈について適切な帰納バイアスの導入が手法の高度化に必要であるという主張を述べた。機械学習において帰納バイアスは様々な形で導入されており、意図的な場合もそうでない場合もある。シーン解釈において必要なのは何を物体として認識すべきかという前提知識であるが、具体的にどのような形で実現するかは自明ではない。一般に学習の際にどのように、どんな帰納バイアスを導入すべきかは自明ではなく、しばしば機械学習研究の貢献となる [13]。深層学習における帰納バイアスの導入手段を簡単に分類するならば、以下のようになる。

1. モデルの構造や、それを実現するネットワーク構造に制約を与えること
2. 与えるデータの種類や量を変化させること
3. 目的関数への正則化項の導入やカリキュラム学習、オプティマイザーの選択など、学習や最適化に関する工夫を行うこと

1 番目の方法はニューラルネットワーク自体の構造やその組み合わせ方など様々な粒度があり、深層生成モデルの場合は確率モデルの設計も含まれる。2 番目の方法は基本的には問題設定を変えていることになるが、簡単に入手可能な情報であれば補助的な情報として導入することも考えられる。また、目的関数の設計は機械学習において大きな役割を持っており、3 番目

の方法は目的関数自体やその最適化方法の操作によって望む解が得られるように工夫するものである。

整理すると、帰納バイアスとしてどのような情報を与えるかと、どのような方法で与えるかはある程度独立しており、両者を考える必要がある。そこで、本研究では以下の2つの目標を設定する。

1. シーン解釈モデルの学習において導入すべき、帰納バイアスの種類や有効性について考える。
2. 上記の帰納バイアスを手法の改良のために組み込む方法を具体的に提案し、学習安定性の向上や適用範囲の拡大などの観点において、有効性を検証する。

## 4.2 本章以降の位置付け

以下の章は、いずれも上記の目標の達成に向けた内容となっている。5章と6章では基本的なシーン解釈の問題設定に対して、異なる帰納バイアスの導入方法を考案し、それを検証する内容となっている。また7章については、多視点の情報を入力とする問題設定を扱い、同様の目標に取り組んでいる。

以降、各章の研究内容と上記の目標の関係性を整理し、位置付けを明確にする。

5章では、物体認識の汎化能力を向上させるための帰納バイアスとして、背景の情報を与えることを提案している。既存のシーン解釈手法では物体や背景のテクスチャや形状が複雑になった場合、うまく認識することができなかった。未知の物体に対する認識を構築するための帰納バイアスとして、対象となる物体自体の知識を与えたのでは意味がなく、与えた知識の範囲で対応できない物体への汎化は期待できない。そこで直接的に物体に関する情報ではなく、かつ知的エージェントにとって定常的に得やすいデータとして、物体がない場合の系の情報、つまり背景の情報をを用いることを提案する（目標1）。静止画を入力とする問題設定の場合、物体がある場合とない場合で同位置・同時点の画像を用意できれば背景差分などの古典的な画像処理手法で処理できるが、実際にそのような構造的なデータが得られるのは限られた環境のみである。ここで与える背景の集合はあくまで物体を含む場合と類似の環境からサンプリングされた、位置や画角が必ずしも対応しない画像集合である。そして、帰納バイアスとして与えられたこの背景の情報を利用し、背景と物体の情報が異なる潜在変数に保持されるような手法を提案する（目標2）。実験においては、まず提案する帰納バイアスの有効性を検証するため、

背景と前景（物体の集合）に分けた表現を獲得できることを検証し、次に物体と背景全てを分割して表現することが可能であることを確認する．これにより既存手法では対応できなかったデータセットに対しても提案手法が有効であることを確認する．

6章では、新たなデータを用いる5章に対し、学習やネットワーク構造の工夫による帰納バイアスの導入を試みる．この研究では自己教師あり学習によって必要な帰納バイアスを与えることを提案する．また、シーン解釈に適したネットワーク構造を提案する．後者についても帰納バイアスの観点から重要だと考えられたものであり、いずれも（目標1）に対応している．自己教師あり学習は事前課題をモデルに解かせることで、様々な課題に適した特徴量抽出器を獲得する手法である．この事前課題は物理的な前提知識に基づいたものであり、事前課題を解くことで物体に関する帰納バイアスが導入できると考えた．この導入については、シーン解釈モデルの入力部分のネットワークに対して、自己教師あり学習による事前学習によって行う（目標2）．また、ネットワーク構造について、既存手法で用いている畳み込みニューラルネットワークは群対称性や画像における近接する領域の情報を重視する帰納バイアスを導入するものである．しかしシーン解釈においてこうした性質はむしろ悪影響で、過剰な制約になってしまっていると考えられる．そこで Transformer[44] を用いた構造を提案し、この解決を試みる（目標2）．

7章では、5章や6章とは少し問題設定が変わり、複数視点からの画像と、その視線の方向が得られる設定を扱う．ただし、与えられた画像に対するシーン解釈だけでなく、観測にない他の視点からの画像の予測（novel view synthesis）も課題となる．知的エージェントへの応用を考えた場合、異なる視点からの画像の入手や簡単な自己位置推定は容易であり、その観点ではこちらがより一般的な問題設定であると言える．三次元的な環境に複数の物体が設置されているような系においては、ある視点からの画像では一部の物体が隠れる（occlusion）場合があり、隠れた部分は経験に基づいて予測する必要がある．また、物体間の相互作用による物理的な制約も生じる．例えば、同じ場所に2つの物体は存在できないし、床の上にしか物体は存在できない．このような系において、部分的な観測から全体を予測するためにはシーン全体についてのモデル化（表現学習）が重要となる．ここで我々の経験的な知識として、シーン全体の表現は空間的な配置と、配置された各物体の性質という2つの要素に分解することが可能であり、これらは概ね独立して変更することが可能である．7章ではこれを確率モデルに反映することを試みる（目標1）．具体的には、シーン全体の空間的な構成を表現するグローバルな潜在変数と、各物体を表現するローカルな潜在変数に分割することでこれを実現する（目標2）．これによる改善点は2つある．1つは、物体の表現学習しか行っていなかった既存手法に対

し、シーン全体の表現学習を明示的に行うことで観測にない新たなシーンの生成が可能となることである。もう1つは、汎化性能の向上や獲得される表現の向上、特に観測として得られる視点の数が少ない場合の性能向上が期待される。

## 第 5 章

# 深層生成モデルによる背景情報を利用したシーン解釈

### 5.1 背景

本章では、補助情報の利用によってシーン解釈に必要な帰納バイアスをモデルに導入し、手法の適用範囲を広げる方法を提案する。シーン解釈の既存研究では [45, 66] のような人工的なデータセットにおいて一定の成功を収めているが、実画像や複雑なテクスチャを含むデータセットでは意図した通りに物体単位の表現を得られないという問題がある [25]。これは既存研究が完全な教師なし学習の枠組みで行われており、対象として認識すべき物体に関する基準や制約がないことが原因の一つだと考えられる。IODINE や MONet[24, 25] といったシーン解釈手法は深層生成モデルの一つである Variational AutoEncoder(VAE)[16] に基づいているが、この場合目的関数は入力と再構成画像の誤差と正則化項のみであり、必ずしも物体に関する性質を反映した表現を獲得できるとは限らない。そのため、この問題を解決するためには何らかの帰納バイアスや前提知識の導入が必要であるが、その方法は自明ではない。また、既存手法の別の問題点として、デコーダの表現力を意図的に落とすことで物体ごとの分割へと促進していることが挙げられる。これは単体のオブジェクトが単純なデータセットを対象とする限りは問題にならないが、対象が複雑になれば適切な処理が難しくなり、手法の適用範囲に関する制約となり得る。

ところで物体の概念、つまり我々がある対象を一つの物体と識別する基準は物理法則によって客観的に定まるものではなく、行動上の便宜や知識・経験に基づいて変化する恣意的なものであるため、何らかの明確な基準を自動で獲得することは困難である。これを踏まえた上で、物体と認識される条件を考えるには symmetry-based disentangled representation の研

究 [40, 67] が参考になる。この研究では物理学での議論を基に、自然界でよく見られる「環境中の特定の対象 (object) に変化を与えるが、それ以外には変化を及ぼさないような介入」を対称変換と定義し、対称変換に基づいた構成的な表現学習について扱っている。この定義において対称変換が反映される単位は“object”と呼ばれているが、この単位は空間的・性質的な独立性を持つことになるため、物体の概念と非常に近いものである。ここで、disentangled representation とは潜在変数の特定の要素が対称変換による変化に対応するような場合と定義されており、disentangled representation の研究において、何らかの帰納バイアス無しに目的通りの表現を獲得することは困難であることが確認されている [41]。これを踏まえ、シーン解釈においても何らかの方法で帰納バイアスを導入することが重要であると考えられる。

既存手法の問題点はこれまでに述べた通り、必ずしも物体に関する、目的通りの表現を獲得するような学習が行われないことである。そこで本研究では、「シーン解釈において我々が意図するような物体の認識を獲得するためには、外部知識や帰納バイアスを導入することが重要である」という仮説を基に、学習の補助となるような補助情報を導入することを提案する。この補助情報に求められる条件として、以下の三点を考えている。一つは認識の対象となる、未知の物体自体に関する情報でないこと。次に認識の対象となる物体を特定するために十分な情報量があること。最後に、上記の条件を満たす範囲で可能な限り情報量が小さく、かつエージェントにとって入手が容易であることである。

これらの条件を満たす補助情報として、本研究では背景に関する知識を利用することを提案する。背景の情報であれば物体自体に関する情報ではなく（第一の条件）、後に観測する物体が含まれる状況との比較を行うことで背景とそれ以外の生成要因が異なることを認識するために活用できる（第二の条件）。また、任意の知的エージェントを想定した場合、物体は種類が多く、未経験の観測（分布外データ）となりやすい。一方で背景の情報は定常的に取得可能であり、物体と比較して経験する機会が多い（第三の条件）。応用を考える場合も、例えば工場でのロボットアームの制御のような管理された環境であれば、物体が無い状態でのデータを取得することは比較的容易であると考えられる。

他に考えられる補助情報の候補としては、言語情報 (画像の説明文) や、他の条件を全て固定して一つの物体のみを少し動かした画像の組を与えることが考えられる。しかし、これらはいずれも認識対象となる物体の情報を含むため、第一の条件に反し、未知の物体に対する認識を構築する補助にはならない。また、前者はエージェントにとって入手が容易であるとは言えない。後者の候補は入手が動画像である場合には自然に得られる情報であるため、別の問題設定として、動画を入力とする深層生成モデルによる教師なし物体認識の研究も進められている

[68].

以上の議論を踏まえ、本研究では複雑なテクスチャを含む画像に対しても適用可能なシーン解釈の手法を提案する。具体的には背景についての補助情報を学習に利用することで、シーン解釈手法の改良を行うことを目指す。提案する手法はシーン解釈の先行研究を基本とし、背景に関する情報と物体に関する情報をそれぞれ別の潜在変数で表現する構造的な深層生成モデルである。提案手法は、背景と物体1つのみを含むデータを対象とする場合は単体で、背景と物体複数を含むデータを対象とする場合は既存のシーン解釈手法と組み合わせて用いることができる。つまり、後者の課題に関して提案手法は、シーン解釈の学習を補助する付加的な手法として位置付けられる。この提案手法によって、既存手法では扱いが困難であったデータセットに対しても有効な、シーン解釈手法の構築を試みる。

本研究の貢献は以下の通りである

- 深層生成モデルによるシーン解釈において、何らかの帰納バイアスや補助情報の導入が必要であることを述べ、背景の情報を利用することを提案した。
- 既存のシーン解釈手法を元に、2つのデータ分布の対照を行う手法を導入し、背景の情報を活用する方法を提案した。
- 複雑なテクスチャを含む画像や実画像などは既存手法で適切に扱えないことが知られているが、既存手法との比較を行いつつ提案手法の有効性を実験的に確認した。
- 既存手法では意図的に表現力の弱いデコーダを用いることで物体ごとに表現を分離させていたが、補助情報を用いることによって提案手法ではこの制約が不要であることを実験的に示した。

## 5.2 関連研究

### 5.2.1 シーン解釈の位置付け

シーン解釈は深層生成モデルを用いて画像単体から、そこに含まれる構成要素（物体など）ごとの表現を教師なしで獲得する課題である。

関連する課題としては物体検出 [69] や、セグメンテーション [70, 71], Saliency Detection[65] などが挙げられる。こうした課題とシーン解釈の違いとして、これらが基本的に教師あり学習で行われること、また教師なし学習であったとしても物体単位の表現の獲得は想定していないことがある。例えば Mask R-CNN[70] ではバウンディングボックスによる物体検出、部分的

なセグメンテーション、さらにクラス分類まで行っている。しかしこれらは入力画像と教師ラベルの関係を直接表現する関数を学習することで実現されており、得られる特徴ベクトルはラベルの識別に特化したもので、生成モデルで得られるような汎用的な情報を含む特徴量とは異なるものである。またこうした特徴から、観測データにおいて見えていない部分（オクルージョン）などの補完については生成モデルのアプローチが有利となる可能性がある。

シーン解釈は、深層生成モデルである Variational AutoEncoder (VAE) [16] を基本としている。まず通常の AutoEncoder (AE) は、入力データから潜在変数(入力よりも低次元な1次元ベクトル)を推論し、その潜在変数から入力を再現するような出力を学習する(再構成)ものである。これにより、潜在変数が入力データの視覚的な情報をできる限り保持するように学習を行っている。VAEではさらに、潜在変数にサンプリング可能な任意の確率分布、例えば標準正規分布を事前分布として仮定している。これにより、サンプリングによって潜在変数を獲得し、観測に含まれない新たなデータを生成することが可能である。また、潜在変数の各次元が何らかの独立した属性を表すような表現学習は *disentangled representation* と呼ばれている。ここで独立した属性とは、例えば入力の被写体の色や形状などであり、各要素を変更しても他の要素には影響がないものである。

シーン解釈は VAE に基づいたものであり、特に潜在変数が複数仮定されている。それぞれの潜在変数が物体に対応した表現を獲得するように学習を行う。ただし、各潜在変数が物体の表現を獲得するためには帰納バイアスの導入など、何らかの学習の工夫が必要となる。シーン解釈手法の詳細については、本章の以降の節で述べる。

次の 5.2.2 節ではシーン解釈の関連研究について2つの方式を説明する。5.2.2 節では提案手法の基礎となっている Spatial Gaussian Mixture モデルによるシーン解釈について、5.2.2 節では Attention Window を用いるシーン解釈の手法について述べる。また、5.2.3 節ではシーン解釈の研究ではなく、*disentangled representation* の文脈において2つのデータ集合間の変換や特徴の抜き出しを考えた研究について説明する。提案手法においてはこれらの研究を元に背景の情報を利用する機構を導入しているため、関連研究としてここに記した。

## 5.2.2 シーン解釈の関連研究

### Spatial Gaussian Mixture を用いたシーン解釈

生成モデルによる教師なしシーン解釈では、画像の次元のセグメンテーションマスクと混合ガウス分布を用いる、Spatial Gaussian Mixture モデルによって入力画像を物体ごとに分割するモデルが多数提案されている [24, 25, 26]。一般的な設定の場合、このモデルの尤度は次式



のようになる.

$$p(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^D \sum_{k=1}^K m_i^{(k)} \mathcal{N}(x_i; \mu_i^{(k)}, \sigma^2) \quad (5.1)$$

ここで  $D$  は入力データの次元で,  $K$  は混合する分布の数であり, 何個の構成要素に分割するか (スロット数) を決定する.  $\mathbf{x}$  は入力画像,  $\mathbf{z}$  は潜在変数,  $m_i^{(k)}$  は混合率,  $\mathcal{N}$  は正規分布を意味し,  $\mu_i^{(k)}, \sigma^2$  はそれぞれ正規分布の平均と分散と意味するパラメータである. ただし, 深層生成モデルにおいて  $\sigma^2$  は基本的に定数のハイパーパラメータとして扱われる. 潜在変数  $\mathbf{z}^{(k)}$  からは  $\mu^{(k)}$  (物体の見た目) と  $\mathbf{m}^{(k)}$  (物体の領域を示すセグメンテーションマスク) という2つの確率変数が計算される. この設定では, ある添字  $k$  についての一組,  $\mu^{(k)}$  と  $\mathbf{m}^{(k)}$  が物体一つを表現することで, 入力画像の物体ごとへの分割が行われる.

上記の混合ガウス分布のような, 複数の確率変数  $\mathbf{m}, \mu$  を一つの確率変数から同時に推論する場合, 一般的には EM アルゴリズムのような反復的な手法が用いられる. [24] で提案されている Multi-Object Network(MONet) では, 計算量や必要なメモリに関して高コストな反復的な手法を回避するため, 決定論的なネットワーク [72] を用いてマスク  $\mathbf{m}$  を近似している. しかしこの近似を用いる場合, 潜在変数をサンプリングしてもマスク  $\mathbf{m}$  を得ることはできず, 新たな画像の生成ができなくなってしまう. そのため MONet では, 決定論的なネットワークで得た  $\mathbf{m}$  を入力画像と共にエンコーダに入力し,  $\mathbf{m}$  の再構成も同時に行うことでこの問題を解決している. 一方で [25] で提案されているモデルでは Iterative Amortized Inference[55] という反復的な手法を用いて直接2つの確率変数を推論している.

これらの手法では [25] で述べられているように, 実画像や複雑なテクスチャを含む画像などに対して適切なセグメンテーションが行えないことが知られている. これらのモデルは入力画像のみを用いる教師なし学習の枠組みで学習されるため, 正則化項を除けば再構成誤差のみに従って学習が行われる. 複雑な画像の場合は特に, 混合するそれぞれのガウス分布が一つの物体を表現することが必ずしも再構成誤差を最小にする分割であるとは限らないため, このような問題が起こりやすいと考えられる. また, デコーダの表現力が高いと一つのスロット (混合される一つのガウス分布) が入力の全てを説明するような局所解に陥ってしまうため, 先行研究では broadcast decoder[73] と呼ばれる比較的表現力が低い機構を用いている. 表現力の低いデコーダを用いる場合, 再構成や生成の精度を低下させる恐れがあり, 対象の物体が複雑になればデコーダでの表現が困難となることが想定される. 本研究で提案する手法は, Spatial Gaussian Mixture モデルを基本として, 背景の補助情報を導入して対象となる物体を分割するように誘導したものであり, これらの問題点の解決・緩和を試みている.

### Attention Window を用いたシーン解釈

マスクによるセグメンテーションではなく、バウンディングボックスによって物体の位置情報を扱うシーン解釈の手法として、[27] を始めとした研究が多数行われている。この手法では物体の位置を矩形で指定して切り出し、その領域について個別に再構成を行う。この手法は明示的に物体を切り出す矩形を得ることから、混合ガウス分布を用いる手法に対していくつかの利点がある。具体的には、学習された結果の解釈性の高さ、物体同士の位置関係が明らかになること、サンプル効率の良さなどが挙げられる。一方で、物体の切り出しが矩形に限定されるため、Spatial Gaussian Model を用いた手法と比べて複雑な形状の扱いを苦手としている。また、基本的にこれらの手法では物体のみを含む入力データを想定しており、背景を含むようなデータは扱っていない。そのため、本研究では Spatial Gaussian Mixture によるシーン解釈モデルを基本としている。[27] を発展させた研究としては、多数の物体を含む場合の認識精度を向上させたものや [53]、動画を入力として物体の認識を構築する [68] などがある。

### 5.2.3 データ分布の対照による特徴抽出

深層生成モデルにおいて、潜在変数の一部の要素の変化が入力画像の特定の属性のみを変化させるような表現について研究が盛んに行われており、このような表現は *disentangled representation* と呼ばれる。その研究の中でもあるデータ集合  $\mathcal{A}$  と、何らかの属性に関して少し変化が加わったデータ集合  $\mathcal{A}'$  を考え、集合  $\mathcal{A}$  と  $\mathcal{A}'$  の間で変換を行ったり特徴的な部分の抽出を行うことで任意の属性に関する *disentangled representation* を獲得させるという研究がいくつか存在している [74, 75, 76]。タスクの具体例としては [74] のように、サングラスを着用していない顔画像の集合 (集合  $\mathcal{A}$ ) と着用している集合 (集合  $\mathcal{A}'$ ) の2つを用意し、それらの変換を行うことが挙げられる。

これらの研究では2つのエンコーダを用意し、片方は集合  $\mathcal{A}$  と集合  $\mathcal{A}'$  に共通の特徴 (顔に関する一般的な情報) を、もう一方は集合  $\mathcal{A}'$  のみに含まれる特徴 (サングラス) に関する潜在変数を推論するようなモデルを構築している。この学習のために [74] では、集合  $\mathcal{A}$  から得た潜在変数と集合  $\mathcal{A}'$  から得た潜在変数の見分けがつかないようにする制約 (domain confusion loss) を導入している。

また、[75, 76] では Contrastive Variational AutoEncoder (CVAE) というモデルを提案しており、潜在変数の推論はほぼ同様の枠組みであるが、デコーダを潜在変数間で共有させることを提案している。特に [76] では集合  $\mathcal{A}'$  から推論される2つの潜在変数が独立になるように、

全相関 (total correlation) に関する制約を導入している [39]. [39] では潜在変数の各次元の独立性を高め、disentangled representation を獲得することを目的として導入した制約であるが、CVAE では 2 つの潜在変数間での独立性の制約として導入されている.

本研究では、シーン解釈の補助情報として背景を利用する機構を構築するために、これらの手法を利用した. 物体を含む集合と背景の集合はそれぞれ上記のデータ集合  $\mathcal{A}'$  と  $\mathcal{A}$  に相当し、物体のみに関する情報と背景に関する情報をそれぞれ異なる潜在変数で表すために利用している.

## 5.3 手法

本章では最初に問題設定を述べ、次に提案手法の詳細について説明する.

### 5.3.1 問題設定

本研究で扱うシーン解釈の問題設定について述べる. シーン解釈は深層生成モデルを用いて教師なしで学習が行われる. 入力単体の静止画のみで、データセットは異なるシーンからサンプリングした画像の集合で構成される. モデルに求められる課題としてはまず、複数の潜在変数を仮定し、それぞれが個別の物体についての表現を獲得すること. つまり任意のシーンを物体の組み合わせとして表現すること. そしてそれらの潜在変数を入力画像から推論することや、潜在変数の推論によってセグメンテーションやバウンディングボックス等の形式で物体の位置情報を得ることが挙げられる. また、手法によっては潜在変数のサンプリングによって新たなシーンを生成することも可能である. 既存手法では、実画像や複雑なテクスチャを含む画像に関しては物体として扱うべき対象がうまく認識できず、意味のない分割を行ってしまうという問題がある. そこで提案手法では補助情報の導入によって、何を物体として認識すべきかという基準を獲得できるようにする. 提案手法単体では物体の領域と背景を分割することが可能で、物体単位の分割も行う場合は既存のシーン解釈手法に提案手法を組み込み、学習の補助として用いることができる. 本論文では課題の簡潔さと結果の解釈性の観点から、まず前者の物体と背景を分離する設定 (2 スロットの場合) にて検証を行う. そして次に、提案手法を既存のシーン解釈手法に組み込み、複数物体を分割する場合にも利用可能であることを示す.

本研究では、学習の補助として背景情報を導入する. この動機については 1 節で述べた通りである. 提案手法の入力は処理の主な対象である物体と背景を含む画像の集合,

$$\mathcal{D}_x = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\} (\mathbf{x}_i \in \mathbb{R}^{W \times H})$$

と、補助情報となる背景画像の集合

$$\mathcal{D}_b = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_M\} (\mathbf{b}_i \in \mathbb{R}^{W \times H})$$

の2つであり、 $\mathcal{D}_b$  は訓練時のみ利用する。ただし  $W, H$  は画像サイズで、整数値を取る。ここで、背景とは認識したい対象の物体を含まず、かつ集合  $\mathcal{D}_x$  と同じドメインに属する画像集合と定義する。ここで、同じドメインに属するとは、同じ環境からサンプリングされ定性的な特徴が類似している画像集合のことを指すものとする。例えば本研究で利用する Objects Room データセットでは、様々な種類の物体を含む部屋の画像集合が  $\mathcal{D}_x$  であり、同じデータセットで物体を含まない場合が背景  $\mathcal{D}_b$  である。そして、学習は  $\mathcal{D}_x$  や  $\mathcal{D}_b$  の要素の再構成によって行われるが、その過程で物体のセグメンテーションマスク  $\mathbf{m} \in \mathbb{R}^2$  や、入力から物体を抽出した画像  $\mathbf{x}_{fg} \in \mathbb{R}^2$ 、背景を抽出した画像  $\mathbf{x}_{bg} \in \mathbb{R}^2$  を獲得すること、そしてこれらに対応する潜在変数  $\mathbf{s}_x, \mathbf{z}_x$  による表現学習が行われることが重要な目的となっている。訓練時には入力の集合  $\mathcal{D}_x, \mathcal{D}_b$  から任意に同数のデータを抽出して利用する。本研究では  $\mathcal{D}_b$  のある要素  $\mathbf{b}_i$  が  $\mathcal{D}_x$  の要素  $\mathbf{x}_i$  に対応した背景になっている必要はなく、あくまで同ドメインの背景の集合として  $\mathcal{D}_x$  の背景とは異なるもので構わない。そのため、これらの集合の要素数が同数である必要もなく、訓練時にはより要素数が少ないことが想定される背景画像の集合  $\mathcal{D}_b$  から重複して取り出せばよい。また、テスト集合として想定しているのは  $\mathcal{D}_x$  に対応する未知の入力であり、背景の情報は不要である。そのため、基礎的な画像処理手法に、物体を含む画像とそれに対応する背景を利用して物体を切り出す背景差分があるが、本研究の問題設定はこれとは異なるものであることに注意されたい。

ところで、一般的なシーン解釈モデルは入力画像を複数の構成要素へと分割するものであるが、ここではまず背景と物体の間の分割に課題を限定した場合について述べる。含まれる物体が複数の場合は通常のシーン解釈の枠組みを用いて個々の物体に分割することも可能であり、複数の物体を分割する場合については後半の節 5.5 節にて述べる。

### 5.3.2 提案手法

1 章や 5.2.2 節で述べたように、Gaussian Mixture モデルを用いた既存のシーン解釈にはいくつかの問題点が存在する。特に本研究では、実画像や複雑なテクスチャを含む画像に対して適切なセグメンテーションを行うことができないという問題や、分割のために表現力の低いデコーダを用いる必要があるという問題について解決を試みる。既存研究は再構成誤差のみを目的関数とした完全な教師なし学習を行っているために、物体を分割するための学習が行われ

ない．この問題の解決には何らかの教師情報や帰納バイアス，補助情報などの導入が必要であると考えられる．そこで本研究では Spatial Gaussian Mixture モデルによるシーン解釈手法 (MONet) を基本の構造として，5.2.3 節で説明したデータを対照する手法を導入し，背景の集合を補助情報として利用する．また，こうした先行研究に関する問題については 5.4 節以降で実験的に確認する．

提案手法の構成図を図 5.1 節に示す．Spatial Gaussian Mixture 部分は概ね MONet の構造に従っており，マスク  $\mathbf{m}$  の推論は U-Net[72] によって行われている．

既存手法と提案手法の差分について，一つは複数のエンコーダを用いたことである．MONet を含めた Spatial Gaussian Mixture によるシーン解釈では全ての入力を一つの共有エンコーダで処理するが，提案手法では 5.2.3 節の手法に基づいて 2 つのエンコーダを利用し，背景を利用する機構 (図上側) を追加している．これは 2 つの潜在変数に異なる情報を獲得することを目的としたものである．また，Abid らの研究 [76] で利用されている，全相関による潜在変数間の独立性に関する制約を潜在変数  $\mathbf{s}_x, \mathbf{z}_x$  に対して適用している．これは上記の機構に加え，各潜在変数が獲得する表現がさらに分離されることを目的としたものである．この制約項の計算は識別器 (ディスクリミネータ，図中央下) によって行われる．

ネットワークの構造について，エンコーダとディスクリミネータはいずれも畳み込みニューラルネットワークで実装されている．先行研究ではデコーダとして表現力の弱い Broadcast Decoder を用いることで分割を促進しているが，本研究では補助情報の利用によりこの制約を導入する必要がなく，一般的な逆畳み込みニューラルネットを使用している．

### 生成モデルについて

5.1 に示すように，提案手法では入力画像  $\mathbf{x}$  に関する潜在変数として  $\mathbf{s}_x, \mathbf{z}_x$  を，背景の画像  $\mathbf{b}$  に関する潜在変数として  $\mathbf{z}_b$  を仮定する． $\mathbf{s}_x$  は前景（物体）に対応し， $\mathbf{z}_x$  と  $\mathbf{z}_b$  は背景に対応した表現が獲得されることを想定している．そして画像  $\mathbf{x}, \mathbf{b}$  はいずれも正規分布が仮定されており，以下の生成モデルに従うものとする．

$$\mathbf{z} \sim p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (\mathbf{z} = \{\mathbf{s}_x, \mathbf{z}_x, \mathbf{z}_b\})$$

$$\mathbf{x} \sim p_\theta(\mathbf{x}|\mathbf{s}_x, \mathbf{z}_x, \mathbf{m}), \quad \mathbf{b} \sim p_\theta(\mathbf{b}|\mathbf{z}_b) \quad (5.2)$$

ただし，5.1 中の  $\mu_x, \mu_b, \mu_{bg}$  はそれぞれの正規分布のパラメータ（平均）であり，分散は任意の固定の値が設定される．ここで，生成モデル（デコーダ）は上式の示した通り共通のパラ

メータ  $\theta$  を持っており、各潜在変数  $\mathbf{s}_x$ ,  $\mathbf{z}_x$ ,  $\mathbf{z}_b$  を入力としてそれぞれ  $\mu_x$ ,  $\mu_b$ ,  $\mu_{bg}$  を出力するものとなっている。 $\mathbf{x}$  については背景と前景の両方が必要なため、2つの潜在変数が条件付けられている。

$\mathbf{x}$  の生成には Spatial Gaussian Mixture モデルが仮定されており、具体的な生成過程はマスク  $\mathbf{m}$  を用いて以下の式で表される。

$$p_{\theta}(\mathbf{x}|\mathbf{s}_x, \mathbf{z}_x, \mathbf{m}) = \mathbf{m} \otimes \mathcal{N}(\mathbf{x}_{fg}|\mu_x, \sigma) + (\mathbf{1} - \mathbf{m}) \otimes \mathcal{N}(\mathbf{x}_{bg}|\mu_b, \sigma) \quad (5.3)$$

第一項が  $\mathbf{x}$  から抽出された前景 (オブジェクト) で、第二項は背景に相当する。 $\otimes$  は要素ごとの積を意味する。 $\mathbf{x}_{fg}$  と  $\mathbf{x}_{bg}$  はそれぞれパラメータ  $\mu_x$ ,  $\mu_b$  を用いてガウス分布からサンプリングされた結果の画像であり、マスクで切り出す前の全体が補完された状態となっている。上記の式 (5.3) の生成過程に関して、具体例を 5.2 に示した。また、訓練時には背景  $\mathbf{b}$  とマスク  $\mathbf{m}$  の再構成も行うが、その生成過程は以下の通りである。

$$\mathbf{b} \sim \mathcal{N}(\mathbf{b}|\mu_{bg}, \sigma) \quad \mathbf{m} \sim \mathcal{N}(\mathbf{m}|\mu_m, \sigma) \quad (5.4)$$

ただし、 $\mu_m$  は  $\mu_x$  とともに  $\mathbf{s}_x$  から得られる。

### 推論モデルについて

推論は一般的な Variational Auto-Encoder(VAE) と同様に amortized variational inference で行われる。入力から潜在変数を推論するための近似事後分布をガウス分布で仮定し、これをエンコーダと呼ぶ。提案手法では 5.1 に示されているように、2つのエンコーダと3つの潜在変数を考えている。エンコーダ A は入力  $\mathbf{x}$  とマスク  $\mathbf{m}$  を入力として潜在変数  $\mathbf{s}_x$  を推論するものであり、パラメータを  $\phi$  として  $q_{\phi}(\mathbf{s}_x|\mathbf{x}, \mathbf{m})$  と表される。ここで  $\mathbf{m}$  を入力に取っているのは  $\mathbf{s}_x$  にマスクの情報も学習させ、再構成を行うためである。

また、エンコーダ B は入力画像  $\mathbf{x}$  と、背景に関する入力画像  $\mathbf{b}$  に共通のエンコーダであり、それぞれの入力から潜在変数  $\mathbf{z}_x, \mathbf{z}_b$  が推論される。つまりエンコーダ B は2つのガウス分布をモデル化しており、エンコーダのパラメータを  $\psi$  とすると、それぞれ  $q_{\psi}(\mathbf{z}_x | \mathbf{x})$ ,  $q_{\psi}(\mathbf{z}_b | \mathbf{b})$  と表される。エンコーダはいずれも畳み込みニューラルネットワークで実装されており、reparameterization trick によってガウス分布に従う潜在変数を得る。

### 全相関による制約

提案手法では CVAE で導入されている全相関 (Total Correlation) の制約を用いている [76, 39]。これは潜在変数の独立性についての制約であり、提案手法では潜在変数  $\mathbf{s}_x$  と  $\mathbf{z}_x$  を

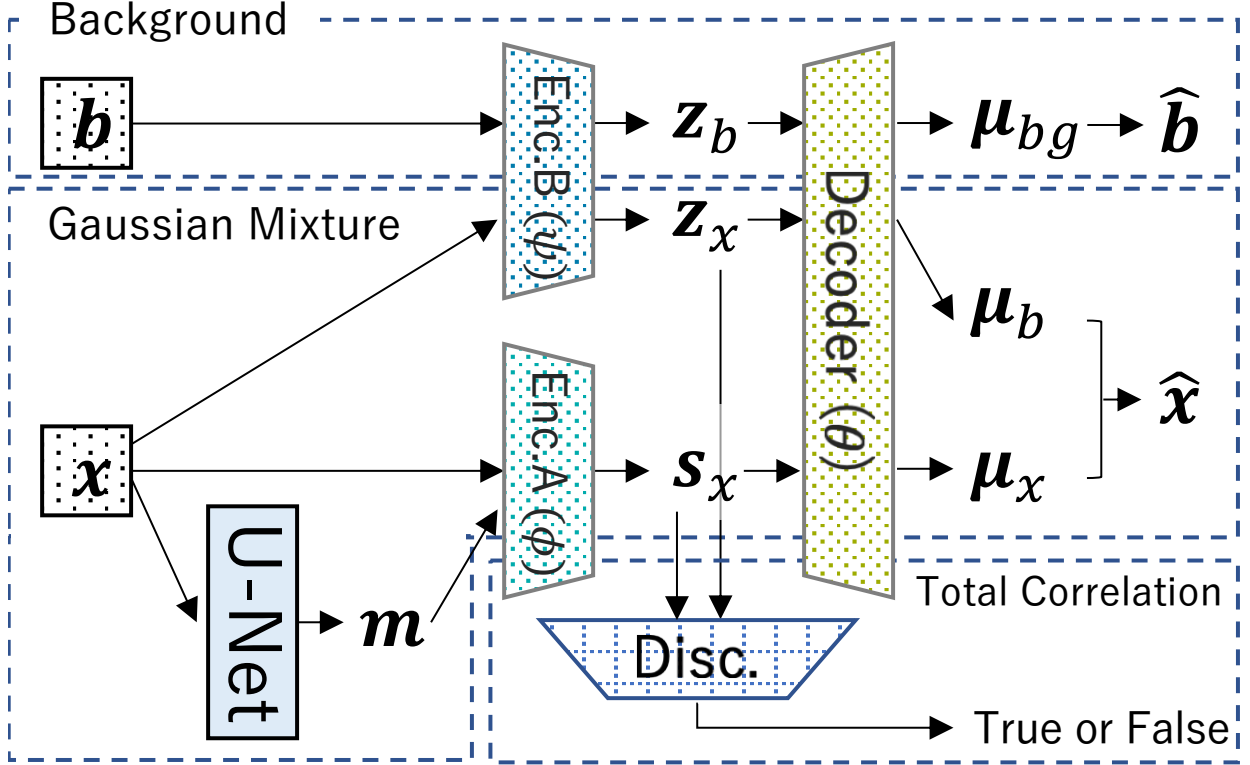


Fig.5.1 提案手法の構成図. Enc.A, Enc.B: エンコーダ,  
Disc.: 識別器 (ディスクリミネータ),  
Decoder: デコーダ を意味する.

対象として適用することで、両者が異なる情報を保持するよう促す効果がある． $q_\phi(\mathbf{s}_x|\mathbf{x}, \mathbf{m})$ ,  $q_\psi(\mathbf{z}_x|\mathbf{x})$  の各分布に対し， $q_{\phi,\psi}(\mathbf{s}_x, \mathbf{z}_x|\mathbf{x})$  という同時分布を考える．この時，全相関の値は以下のように表される．

$$-KL \left[ q_\phi(\mathbf{s}_x | \mathbf{x}, \mathbf{m}) \cdot q_\psi(\mathbf{z}_x | \mathbf{x}) \parallel q_{\phi,\psi}(\mathbf{s}_x, \mathbf{z}_x | \mathbf{x}) \right] \quad (5.5)$$

ここで，KL はカルバックライブラーダイバージェンスである．この項は  $q_\phi$  と  $q_\psi$  が独立であれば 0 となるため，目的関数に足し合わせることで独立性の制約項として利用可能である．実際にはこの項は，識別器  $D_\omega(\mathbf{s}_x, \mathbf{z}_x)$  を用いて計算することが可能であり，パラメータを  $\omega$  とする．これは 5.1 中で Disc. と表されている部分である．この識別器は元の分布から推論された潜在変数と，上記の同時分布から得られたものを区別するように分類問題の学習を行うことで，その予測確率  $D_\omega(\mathbf{s}_x, \mathbf{z}_x)$  が式 (5.3) の推定に利用可能となる．これは density-ratio trick[77]

と呼ばれるもので、全相関の値は以下のように計算される．

$$TC(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{s}_x|\mathbf{x}, \mathbf{m}), q_\psi(\mathbf{z}_x|\mathbf{x})} \left[ \log \frac{D_\omega(\mathbf{s}_x, \mathbf{z}_x)}{1 - D_\omega(\mathbf{s}_x, \mathbf{z}_x)} \right] \quad (5.6)$$

既存研究ではデコーダに表現力の低い Broadcast Decoder[73] を用いることで、一つの潜在変数だけでは入力画像を十分に表現できないようにしており、これによって複数の潜在変数が分担して入力画像を表現するよう誘導している．提案手法では補助情報の導入や全相関による独立性の制約によって、このようなアーキテクチャ的な帰納バイアスを用いずに適切な表現を学習することを試みている．

### 学習について

提案手法の目的関数である、対数尤度  $\log p(\mathbf{x} | \mathbf{s}_x, \mathbf{z}_x)$  と  $\log p(\mathbf{b}|\mathbf{z}_b)$  の変分下界は以下のようになる．

$$\begin{aligned} \mathcal{L}_{fg}(\mathbf{x}) = & \mathbb{E}_{q_\xi(\mathbf{s}_x, \mathbf{z}_x, \mathbf{m}|\mathbf{x})} [\log p_\theta(\mathbf{x} | \mathbf{s}_x, \mathbf{z}_x, \mathbf{m})] \\ & - \beta_1 (KL[q_\phi(\mathbf{s}_x|\mathbf{x}, \mathbf{m}) \parallel p(\mathbf{z})] \\ & + KL[q_\psi(\mathbf{z}_x|\mathbf{x}) \parallel p(\mathbf{z})]) \\ & - \gamma KL[q_\phi(\mathbf{s}_x|\mathbf{x}, \mathbf{m}) \cdot q_\psi(\mathbf{z}_x|\mathbf{x}) \parallel q_{\phi, \psi}(\mathbf{s}_x, \mathbf{z}_x|\mathbf{x})] \\ & - \delta KL[q_m(\mathbf{m}|\mathbf{x}) \parallel p_\theta(\mathbf{m}|\mathbf{s}_x)] \end{aligned} \quad (5.7)$$

$$\mathcal{L}_{bg}(\mathbf{b}) = \mathbb{E}_{q_\psi(\mathbf{z}_b|\mathbf{b})} [\log p_\theta(\mathbf{b} | \mathbf{z}_b)] - \beta_2 KL[q_\phi(\mathbf{z}_b|\mathbf{b}) \parallel p(\mathbf{z})] \quad (5.8)$$

$$\mathcal{L}_{disc.}(\mathbf{s}_x, \mathbf{z}_x) = \log(D_\omega(\mathbf{s}_x, \mathbf{z}_x)(1 - D_\omega(\mathbf{s}_x, \mathbf{z}_x))) \quad (5.9)$$

ただし  $q_\xi(\mathbf{s}_x, \mathbf{z}_x, \mathbf{m}|\mathbf{x}) = q_\phi(\mathbf{s}_x|\mathbf{x}, \mathbf{m})q_\psi(\mathbf{z}_x|\mathbf{x})q_m(\mathbf{m}|\mathbf{x})$  と推論モデルをまとめたものであり、 $\xi = \{\phi, \psi, m\}$  とする． $q_m$  はマスク  $\mathbf{m}$  の分布であり、提案手法では U-Net で決定論的なネットワークとして実装している．ここで、 $\mathcal{L}_{fg}$ 、 $\mathcal{L}_{bg}$ 、 $\mathcal{L}_{disc.}$  はそれぞれ  $\mathbf{x}$  を入力とする Spatial Gaussian Mixture モデルの部分、 $\mathbf{b}$  の再構成を行う背景を利用する機構、識別器の目的関数（全相関）である． $\mathcal{L}_{fg}$  と  $\mathcal{L}_{bg}$  は同時に最適化が可能であり、これらを合わせた  $\mathcal{L} = \mathcal{L}_{fg} + \mathcal{L}_{bg}$  がモデル全体の目的関数となる．また、 $\mathcal{L}_{disc.}$  はモデル本体とは別に学習を行う必要があり、 $\mathcal{L}$  の最適化と交互に行われる．学習はいずれも確率的勾配降下法によって行われる．



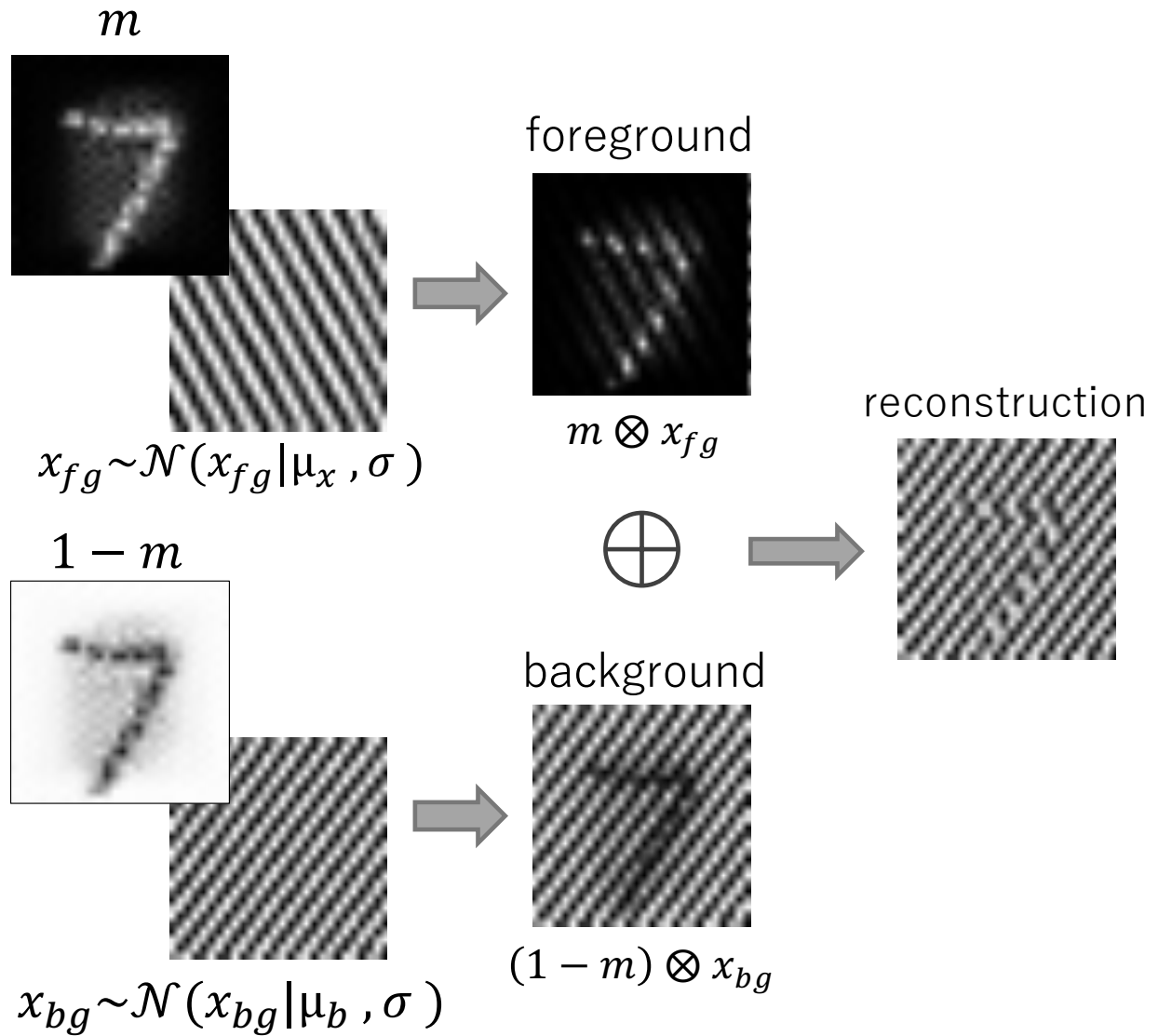


Fig.5.2 生成 (再構成) の過程を示した模式図.

$\mathcal{L}_{fg}$  の第 1 項は再構成誤差, 第 2 項はカルバックライブラーダイバージェンスによる正則化項, 第 3 項は全相関の項で, 最後の第 4 項はマスクの再構成誤差である. これらはそれぞれ式の通り, 定数係数  $\beta_1, \gamma, \delta$  が適用されている.  $\mathcal{L}_{bg}$  は背景  $\mathbf{b}$  に関して通常の VAE と同様であり, 第一項が再構成誤差, 第二項が正則化項となっている.  $\mathcal{L}_{disc.}$  は識別器自体の目的関数であり, 潜在変数に関する分類問題の誤差を表している.

また MONet では再構成誤差についてマスクによる重み付けを行っているが, 提案手法ではこれは行わず, 画像全体に対して通常の再構成誤差を使用した.

## 5.4 実験結果

### 5.4.1 データセットについて

本節では本研究で実験に用いた全 4 種のデータセット, Textured MNIST, Multi-Textured MNIST, Natural Scene MNIST, Objects-Room Dataset について説明する.

#### Textured MNIST

このデータセットは手書き数字文字データセット MNIST の背景と文字にそれぞれテクスチャを適用したものである. 本研究では [58] の実装で公開されている, 画像中に 1 文字のみを含む設定のものをを用いた. シーン解釈の既存研究ではこのデータセットの適切な処理が難しいことが知られており, 提案手法の検証のために採用した. 枚数やテスト集合の割合はオリジナルの MNIST に準拠しており, 画像は  $28 \times 28$  ピクセルのグレースケールである.

#### Multi-Textured MNIST

このデータセットは Multi-MNIST と呼ばれる, 画像中に複数の手書き数字を含むデータセットを拡張し, 背景にのみ上述の Textured-MNIST と同様のテクスチャを適用したものである. [58] では背景と数字 2 つにテクスチャを適用し, 数字同士の重なりを許容した非常に難しい設定を同様に Textured-MNIST と呼んでいるが, 本研究は複数物体の分割を主な対象とはしておらず, 物体の重なりを含む場合についてはシーン解釈の既存手法においても扱っていない. そのため本研究ではこの設定は採用せず, 単に前景が複雑になった場合の検証という目的で利用することを考え, 数字同士の重なりと数字側のテクスチャを除いた設定を用いた. また, 本実験では数字をランダムに 1 個または 2 個含み,  $50 \times 50$  ピクセルのグレースケールの画像を利用した.

#### Natural Scene MNIST

これは MNIST の背景に自然風景 (実画像) を適用したデータセットである. 既存のシーン解釈研究では実画像についても扱いが難しいことが知られており, 実画像への有効性の検証のために本データセットを作成した. 背景に用いたのは実画像のデータセットである Cifar100 の中の, trees という集合である [78]. これは 100 クラス中で関連する 5 クラスを統合したもので, 2500 枚の訓練集合と 500 枚のテスト集合を含む. 本データセットでも同様に, 数字と背景を分離することを課題としている. Cifar100 の画像サイズは  $32 \times 32$  であるが, MNIST に合

わせ  $28 \times 28$  にリサイズして利用した。また、背景の多様性を減らし、難易度を落とす目的で上記の 5 クラスの内の一つ「oak\_tree」のみを用いたデータセットも用意した。本データセットは Natural Scene MNIST(NS-MNIST) と呼ぶことにし、5 クラスを用いたものを NS-MNIST LARGE, 1 クラスのみのものを NS-MNIST SMALL と区別する。

### Objects-Room Dataset

これは Multi Object Datasets[66] に含まれる、部屋に 3 つの物体 (プリミティブ) が設置されたデータセットである。先の 3 種のデータセットとは異なり、シーン解釈の研究で頻繁に用いられている。提案手法では背景 (部屋) と物体を分割することを目的として実験を行う。このデータセットによる実験の目的としては 2 点あり、一つはカラー画像でも問題なく提案手法が適用できるかという検証として、もう一点は物体と背景を適切に区別することが可能であるかという検証である。既存手法では物体と背景を区別せず、背景もいくつかの要素に分解しているが、提案手法では物体と背景を明確に区別することが可能であることを確認する。画像は 3 チャンネルのカラー画像であり、 $50 \times 50$  ピクセルにリサイズして用いている。背景のデータとしては、テスト用集合として提供されている画像集合 (Empty Room) を用いた。この集合は訓練用データと同ドメインであるが物体を含まず、床と壁と空だけを含む集合となっている。また、色や形状、画角などについて任意に組み合わせた集合となっているため、データ数は訓練集合と一致しない。

### 5.4.2 実験結果

実験の設定のうち全データセットに共通のものとして、バッチサイズを 128, 潜在変数の次元をそれぞれ 128 とした。学習率は  $10^{-3}$  で最適化アルゴリズムは Adam を使用し、学習結果は 100 エポック時点のものである。実験によって異なるハイパーパラメータを利用しているが、これは検証データを用いて調整したものである。また、実装には Pytorch[79] をベースとした深層生成モデルのライブラリ, Pixyz[80] を用いた。

### Textured MNIST

このデータセットによる実験では、複雑なテクスチャを含む画像に対する手法の性能評価を行うことを目的としている。また、既存手法との比較実験や、提案手法の機構を部分的に取り除いた実験を行うことで、提案手法の全般的な有効性を確認する。本実験で用いたパラメータは  $\beta_1 = \beta_2 = 0.5, \gamma = 1.0, \delta = 1.0$  で、 $\beta$  については初期値を 0.1 とし、10 エポック時点で 0.5

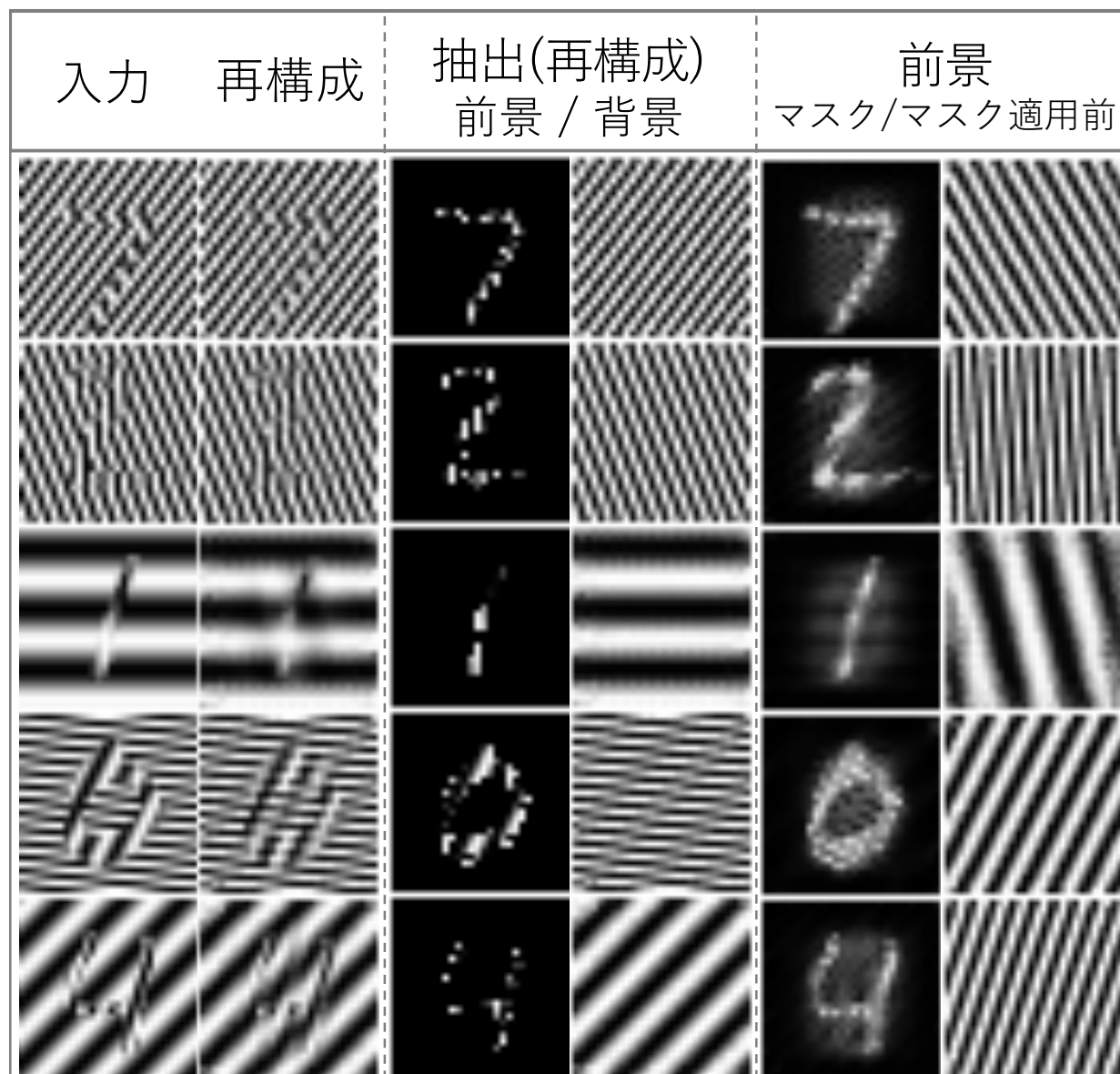


Fig.5.3 Textured MNIST データセットに対する提案手法の適用結果. 5 サンプルに対する結果を示している. 左列から, 入力画像, 再構成された画像, 前景 (数字) の抽出, 背景の抽出, 前景のマスク, マスクで切り出す前の画像, となっている.

となるように線形にアニーリングを行った.

図 5.3 に, Textured MNIST に対する提案手法の適用結果を示した. 提案手法は物体と背景のために 2 つの画像を生成し, それをマスクで切り出す形で再構成を行うため, 適切に機能した場合はマスクが数字の形状を捉え, 物体側の画像から切り出す形になることが期待される. 実験結果では, いずれのサンプルにおいてもマスクが数字の形状を適切に捉えていること

が確認できる．オブジェクトと背景の分割や，入力画像の再構成についても目的通りに行うことができおり，提案手法が Textured-MNIST に対して有効であることが確認された．図 5.3 における前景・背景とは，それぞれ式 (5.3) の第一項，第二項に対応するものであるが，抽出の処理として，前景は二値化したマスクを適用した結果を，背景はマスク適用前のもの  $\mathbf{x}_{bg} \sim \mathcal{N}(\mathbf{x}_{bg}|\boldsymbol{\mu}_b, \boldsymbol{\sigma})$  を示している．ただしこの処理は可視化を目的としたものであり，学習時にはマスクの二値化は行っていない．以降の図 5.6，図 5.7，図 5.8 についても同様の形式で図示している．特筆すべき点として，マスクで切り出す前の前景の画像  $\mathbf{x}_{fg} \sim \mathcal{N}(\mathbf{x}_{fg}|\boldsymbol{\mu}_x, \boldsymbol{\sigma})$  において，僅かな部分しか見えていない数字側のテクスチャを補完し，再現していることが挙げられる．これはテクスチャのパターンがある程度決まっており，生成モデルがテクスチャの性質を十分に学習することができたためだと考えられる．

また，既存手法を用いた場合との比較のため，提案手法の基本となっている MONet[24] の実験を行った．ただし MONet では背景の情報は利用していない．今回新たに導入したデータセット 3 種 (Textured-MNIST, Multi-Textured MNIST, NS-MNIST) における MONet での実験結果を図 5.4 に示した．MONet は最大分割数，つまり混合するガウス分布の数 (スロット数) を予めハイパーパラメータとして設定する必要があるが，今回は比較のため，数字と背景で 2 スロットとしている．実験結果では片方のスロットが全てを説明してしまっており，数字と背景の分割には失敗していることが分かる．

さらに，提案手法で導入した各機構の有効性について検証するため，モデルの一部を機構を取り除いた状況での実験を行った．実験の条件としては，通常時，TC 項を除いた場合，背景を利用しない場合，TC 項も背景も利用しない場合の 4 条件で行い，定量評価による比較を図 5.5 に示した．また，参考として MONet のスコアも記した．評価に用いた指標は Adjusted Rand Index (ARI) と呼ばれるもので，クラスタリングやセグメンテーションの定量的な評価に用いられるものである．これは正解と予測の 2 つのベクトルを入力とし，出力は  $-1$  から  $1$  までの値を取る指標である．出力  $1$  は正解と一致する完全な状態であるが，ラベルの順序には影響されず，パーミュテーション不変である．正解と関係なくランダムなクラスタリングが行われた場合，つまり期待値の場合は出力が  $0$  となる．また，期待値を下回るラベル付けに対しては  $-1$  までの負の値が出力される．Textured MNIST データセットでは元の MNIST の数字データをバイナリ化したものを正解のセグメンテーションマスクと考えることができるため，これを正解データとして，モデルが推論したマスクを二値化 (softmax) したものと正解の間の ARI スコアを算出した．ここでは実装に用いたライブラリである Pytorch の乱数シードとして，5 つの異なる値を設定してモデルの学習を行い，この結果の全てのサンプルを用いてス

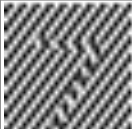
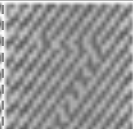

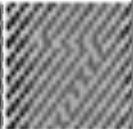

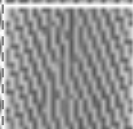

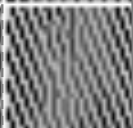

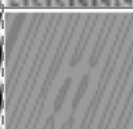

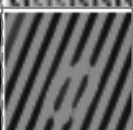


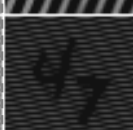
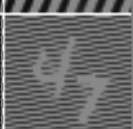








	入力	再構成	slot1	slot2
Textured MNIST				
				
Multi-Textured MNIST				
				
NS-MNIST (SMALL)				
				

Fig.5.4 本研究で新たに導入したデータセット 3 種 (Textured MNIST, Multi-Textured MNIST, NS-MNIST SMALL) に対する, 背景情報を用いない, 既存手法 (MONet) での実験結果.

コアの平均と標準偏差を計算した. ただし ARI スコアの計算自体は決定論的なものであり, この乱数は学習時にネットワークの初期値や確率分布からのサンプリングに影響を与えるものである.

定性的に確認したところ, 通常時がほぼ全ての入力データに対して適切なセグメンテーションを出力しているのに対し, 他の条件ではシードや入力データによって成功と失敗が混在しているという状況であった. そのため, 定量的にはいずれの条件でも通常時と比べて平均スコアが下がり, 標準誤差が大きくなるという結果になっている. 特筆すべき結果として, TC 項と背景を両方用いない場合はむしろ背景を利用しない場合 (TC 項は残っている場合) よりも良い結果となったことが挙げられる. 2 つの潜在変数間の独立性を増加させる制約項としての役割は, 背景の補助情報が無い状況では目的通りの効果を発揮せず, むしろ学習に悪影響を及ぼしたものと考えられる. つまり, 本実験によると背景の補助情報が提案手法の処理に最も貢献

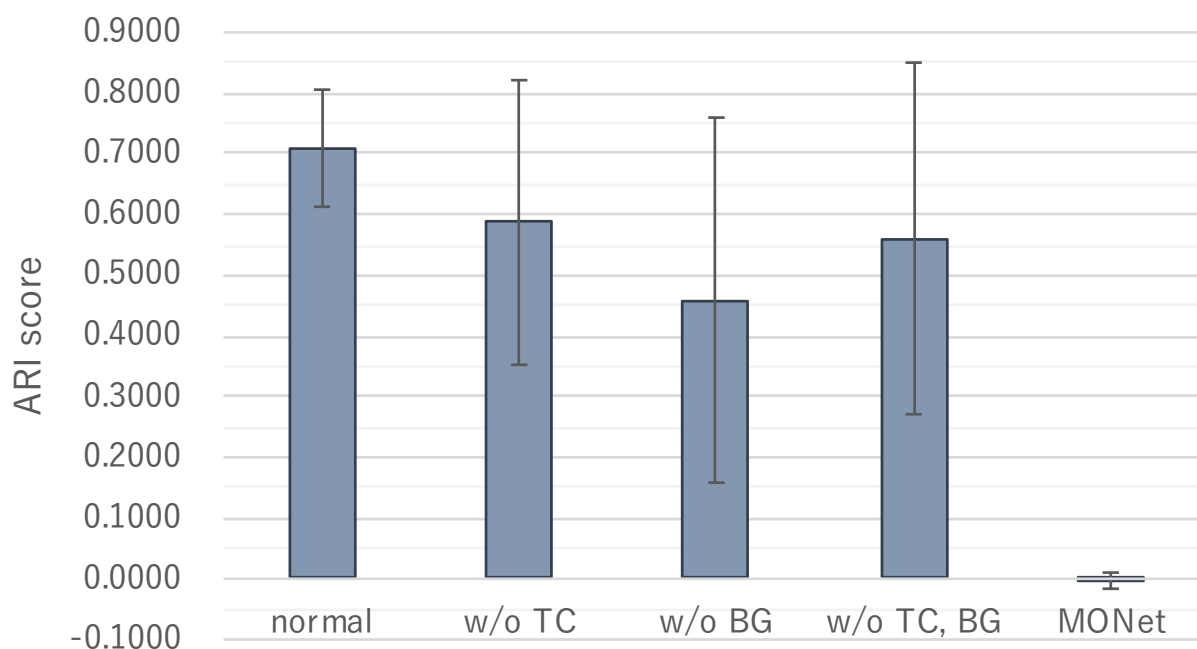


Fig.5.5 提案手法の一部の機構を無効化した場合と既存手法の、ARI スコアによる定量評価。エラーバーは標準偏差を表している。左から順に、通常時、TC 項の除去、背景利用なし、TC 項・背景ともに用いない場合、先行研究 (MONet) による結果を示している。

しており、TC 項はその補助となるが単体では機能しないという結果になった。MONet については適切なセグメンテーションができていないため、いずれの試行でも ARI は 0 付近の微小な値を取っていた。

### Multi-Textured MNIST

Textured MNIST の実験では、分割の対象となるオブジェクト (数字) が単体で、画像の中央にのみ存在する設定であった。オブジェクトが画像内のあらゆる場所に存在し、個数もランダムとなる条件での検証を行うことが本データセットによる実験の目的である。ただし Textured MNIST の場合とは異なり、このデータセットでは数字側にテクスチャを適用していない。本実験で用いたパラメータは  $\beta_1 = \beta_2 = 1.0, \gamma = 1.0, \delta = 2.0$  で、 $\beta$  については初期値を 0.1 とし、10 エポック時点で 1.0 となるように線形にアニーリングを行った。

提案手法での実験結果を図 5.6 に示した。複数の数字が様々な場所に現れる条件となっているが、マスクは正しく数字の形状を捉えており、背景の抽出にも成功している。マスク切り出し前の前景の画像は Textured MNIST の場合とは異なり、背景のテクスチャを残しつつ数字の

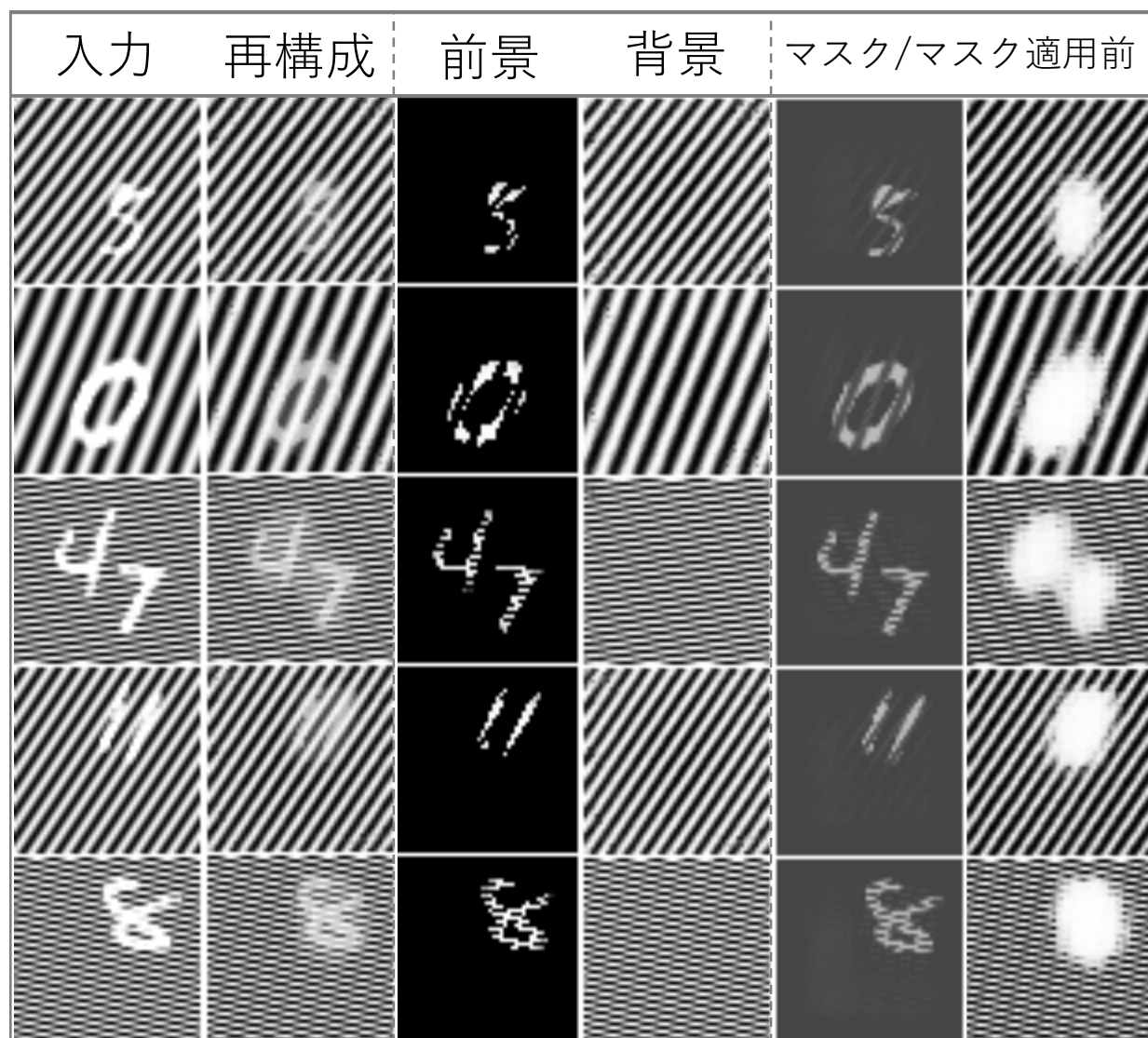


Fig.5.6 Multi-Textured MNIST データセットに対する提案手法の適用結果.

概形を表している。抽出された数字が背景のテクスチャに応じて穴あきになっているのは、マスクが背景のテクスチャの白地部分を取り除く形になっているためである。これは物体(数字)についての知識が無く、背景を基準として物体を考えるならば正しい結果であると言える。数字としての形を表現させるためには、どの物体が同じカテゴリなのか、つまり数字のラベルを与える必要があると考えられる。



### Natural Scene MNIST

提案手法の実画像への有効性を検証するために、実画像を背景とした NS-MNIST データセットによる実験を行った。本実験で用いたパラメータは  $\beta_1 = \beta_2 = 0.5, \gamma = 1.0, \delta = 5.0$  で、 $\beta$  については初期値を 0.1 とし、10 エポック時点で 0.5 となるように線形にアニーリングを行った。

結果を図 5.7 に示した。計 6 サンプルに対する結果を示しており、上から 2 つずつ、NS-MNIST SMALL に対して 10 エポック学習したもの、100 エポック学習したもの、NS-MNIST LARGE での実験結果となっている。本データセットによる実験では他のデータセットと異なり、学習段階によって定性的な結果の変化が見られたため、学習の初期段階と、十分に学習が進み定性的な変化が起きなくなった段階としてこれらのエポックを選択し、結果を示した。いずれの条件でも入力画像の再構成には成功しているが、前景と背景の分割については最初の条件 (SMALL での 10 エポック時点) でしか成功していない。SMALL の 100 エポック時点では背景の木の部分は黒塗りになり、マスク側が代わりにその部分を表現することで再構成を行っている。LARGE 条件ではこの傾向がより顕著で、単なる白黒の矩形から切り出す形で全てをマスクが表現してしまっている。これは、提案手法を構成している生成モデル側 (エンコーダ・デコーダの構造) がマスクを担当する U-Net よりも表現力が低いことが原因だと考えられる。ただし、これまでに実験を行った 3 つのデータセットについては、このような現象は確認されていない。本データセットのような実画像を用いた複雑な入力画像に対しては、生成モデル側で入力画像の詳細を表現するよりも、U-Net が作成するマスクによって切り出す形で表現した方が再構成誤差をより小さく抑えられると考えられる。そのため、画像の多様性が増して再構成がより難しくなる NS-MNIST LARGE ではこの傾向が顕著になり、潜在変数  $\mathbf{z}_b$  を介した背景単体での再構成自体には概ね成功しているにも関わらず、生成モデル側が全く画像を表現しなくなってしまっている。このような傾向が確認されたため、NS-MNIST に関する実験ではエンコーダ・デコーダの畳み込みのフィルタ数をこれまでの 3 つのデータセットでの実験の 4 倍に増やしているが、LARGE 条件では改善が見られなかった。この問題を根本的に解決して LARGE 条件でも適切に処理を行う方針としては、生成モデル部分により表現力の高いアーキテクチャを用いることや、マスクの近似方法を変更すること、もしくは背景の学習データ量を増やすことなどが挙げられる。

	入力	再構成	前景	背景	マスク/マスク適用前	
SMALL (ep10)						
SMALL (ep100)						
LARGE						

Fig.5.7 Natural Scene MNIST データセットに対する提案手法の適用結果.

### Objects Room Dataset

このデータセットはシーン解釈で一般的に用いられる，CG データでの提案手法の有効性を検証する目的で実験を行った．また，これまでのデータセットはグレースケールであり，カラー画像での提案手法の有効性を検証する目的もある．本実験で用いたパラメータは  $\beta_1 = \beta_2 = 1.0, \gamma = 1.0, \delta = 5.0$  で， $\beta$  については初期値を 0.1 とし，10 エポック時点で 1.0 と

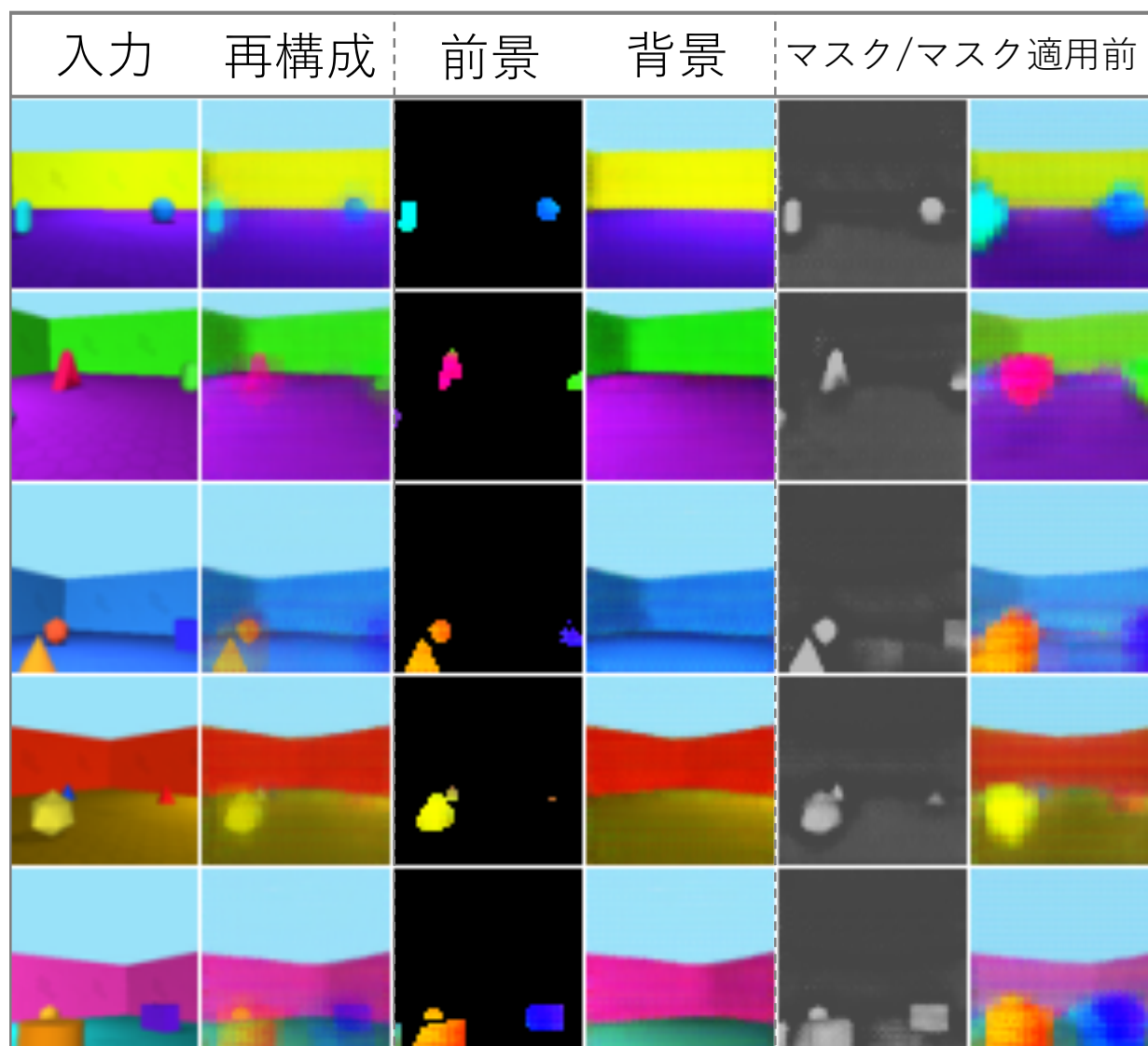


Fig.5.8 Objects Room データセットに対する提案手法の適用結果. 最左列の入力画像以外は提案手法の出力結果である.

るように線形にアニーリングを行った.

結果を図 5.8 に示した. 小さな物体を無視してしまったり, 色が変わっているケースが散見されるが, マスクは正しく物体の形状を捉え, 物体と背景の切り出しにも概ね成功している. マスク切り出し前の前景は Multi-Textured MNIST の結果に近く, マスクで切り出す前提で大雑把にオブジェクトの色を反映させたものとなっている. この実験結果から, 本データセットのような複数の異なる形状の物体を含む CG データでも, 目的通りに物体だけを切り出すことが可能であることが確認できた. また, 本実験では背景の再構成誤差の係数である  $\delta$  の影響が

大きく、小さな値を用いると背景の抽出結果に物体が混じるような形で失敗するケースが多くなり、全体の挙動が不安定になった。

## 5.5 多スロットへの拡張

ここまで、背景と前景のみの二分割を行う設定で手法の妥当性を検証してきたが、任意のシーン解釈モデルに提案手法を導入することで、複数物体の分割が可能となる。そこで本節以降では MONet に提案手法を組み込み、手法の有効性を検証する。

### 5.5.1 手法

本節では、MONet の持つ複数のスロットの中の 1 つを、提案手法と同じ方法で背景を表現するように学習を行う。MONet では以下のような反復的な処理を行うことで、マスク  $\mathbf{m}_k$  を推論している。

$$\mathbf{m}_k = \mathbf{s}_{k-1} \alpha(\mathbf{x}, \mathbf{s}_{k-1}) \quad (5.10)$$

ここで、 $\mathbf{s}_k$  は、スコープと呼ばれ、現在までのマスクで説明されていない残りの領域を指している。 $\alpha(\mathbf{x}, \mathbf{s}_k)$  は入力画像とスコープを入力に取り、次のマスクを出力する更新用の関数で、U-Net で実装されている。スコープは初期値を  $\mathbf{s}_0 = 1$  として以下のように更新する。

$$\mathbf{s}_{k+1} = \mathbf{s}_k(1 - \alpha(\mathbf{x}, \mathbf{s}_k)) \quad (5.11)$$

最後のスロットはその時点までのマスクで説明されていない部分がスコープによって表されているため、そのままマスクとなる。つまりスロット数を  $K$  として  $\mathbf{s}_K = \mathbf{m}_K$  である。

提案手法を MONet に組み込む方法は様々なものが考えられるが、ここでは最初のスロットが背景を表現するものと仮定し、MONet と同様の反復的な推論を行いつつ、最初のスロットのエンコーダには背景用のものを用いた。つまり、最初に背景を生成し、その後のイテレーションで順次物体を生成していくことになる。背景用のデコーダは、提案手法の図 6.1 の  $\text{Enc.B}(\psi)$  に相当するものであり、学習についても先に説明した提案手法と同様に行う。

### 5.5.2 多スロット版の実験

本実験では、Multi-dSprites データセット [66] にテクスチャ状の背景を適用した、Textured Multi-sSprites データセットを作成し、検証を行った。 $\beta, \delta$  は 0.5、 $\gamma$  は 2.0 とした。最適化手

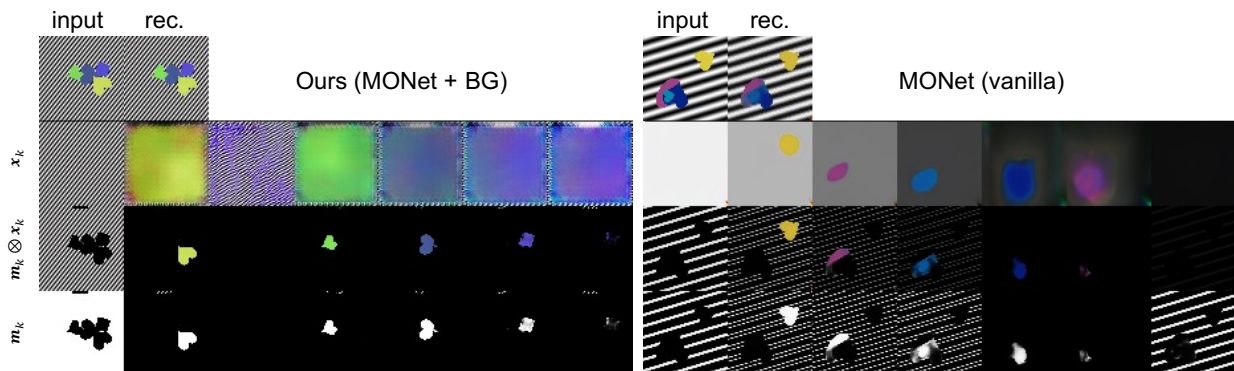


Fig.5.9 Textured Multi-dSprites データセットに対する，多スロット版の提案手法の適用結果．

法は同様だが，学習率は 0.001 とした．図 5.9 にスロットごとの再構成結果を示した．提案手法を導入していない，オリジナルの MONet (vanilla) では再構成自体には成功しているものの， $x_k$  は背景のテクスチャを捉えることができず，適切な物体のセグメンテーションも得られていない．一方，提案手法を導入したもの（Ours）では物体の形状を適切に捉えたセグメンテーションマスクが獲得されている．このように，背景の情報をを用いることで，物体として認識すべき対象に関する帰納バイアスが獲得され，既存手法では対応できなかったデータセットにも適用範囲を拡張できることが確認できた．

本実験では Textured Multi-MNIST ではなく，Textured Multi-sSprites を用いたが，これはそもそも MONet が Multi-MNIST を適切に物体ごとに分割することができないためである．本研究で導入した背景の帰納バイアスは「何が物体であるか」に関する基準を与えているが，物体を分割するための帰納バイアスを導入することはできていない．この物体を分割するための帰納バイアスを考案し，シーン解釈モデルに導入していくことは今後の課題である．

## 5.6 結論と考察

本研究では，深層生成モデルによる物体認識手法の改良に取り組み，補助情報を導入することによって目的通りのセグメンテーションを可能とする手法を提案した．実験ではまず Textured MNIST データセットにおいて，既存手法に対する提案手法の有効性や，提案手法に導入された機構の有効性についての検証を行った．次に Multi-Textured MNIST, Objects Room, Natural Scene MNIST で実験を行い，提案手法の適用可能な範囲についての検証を行った．Textured MNIST データセットや Multi-Textured MNIST のような複雑なテクスチャを含

む画像は既存手法で扱うことが難しかったが、提案手法では適切に数字と背景を分離可能であることが確認された。また、Objects Room データセットは既存手法でも実験で扱われているものだが、提案手法では背景の補助情報を用いることで、指定した通り物体と背景を分割して表現することができた。Natural Scene MNIST データセットでは実画像に対する提案手法の適用可能性を検証した。これは背景のクラスを制限した NS-MNIST SMALL では成功したものの、背景の多様性が上がる NS-MNIST LARGE の条件では前景と背景の適切な切り出しができなかった。

本研究は、既存手法では扱いが難しいことが知られている Textured MNIST データセットのような複雑なテクスチャを含む画像を対象として、適切な認識を行うための手法を提案した。まず物体が単体の場合について検証を行い、有効性を示した。次に、複数の物体を含む場合について、提案手法を用いて既存のシーン解釈手法を拡張することで適切な物体ごとの分割が可能になることを確認した。また、提案手法で導入したのは物体とそれ以外を見分ける補助情報であり、複数の物体を区別するための補助情報ではない。後者に関する研究は今後の課題とする。

また、その他にも今後の課題として NS-MNIST LARGE のようなデータセットに対しても適切な処理を行うために、より表現力の高いアーキテクチャを用いて実画像への有効性を向上させることが考えられる。また、既存手法・提案手法ともに現状では画像生成の品質についてあまり考慮されていないが、モデルベース強化学習や世界モデルと呼ばれる技術への応用を考えると重要な機能であり、改善の余地がある。さらに、本研究のようなシーン解釈モデルでは学習した表現を別の課題に利用することが重要な目的の一つであるため、downstream task<sup>\*1</sup>についての有効性を検証することも重要となる。

---

<sup>\*1</sup> 例えば入力画像に含まれる物体の種類や数を推論された潜在変数から推測するといった課題。

## 第 6 章

# 自己教師あり学習と Transformer を用いたシーン解釈手法の提案

### 6.1 背景

本章では、帰納バイアスの導入によってシーン解釈手法の学習安定性や手法の適用範囲を広げることを目的として研究を行う。5 章においては、背景のデータ集合という新たなデータを用いることで物体に関する帰納バイアスを導入することを試みたが、本章では新たなデータを利用せずに実現可能な手法を考える。3 章で紹介したように、シーン解釈には主に 2 つの方式がある。一つはセグメンテーションを行い、その領域ごとに対応する潜在変数によって表現する Scene Mixture モデル [22, 23, 24, 25, 26] で、もう一つは矩形領域ごとにその位置や物体の情報を構造化された潜在変数を用いて表現する Neurosymbolic モデルである。Scene Mixture モデルは背景を別途モデリングする必要がないことや、複雑な形状の物体でも表現できるという利点がある。その一方で潜在変数が明示的に座標情報や物体の有無といった形で制約されている Neurosymbolic モデルに対し、Scene Mixture モデルではそうした制約がなく、連続値のベクトル（分散表現）として潜在変数がモデル化されている。つまり、潜在変数の役割自体を定める（Binding）必要がある Scene Mixture モデルと、潜在変数にシンボリックな役割が定められた Neurosymbolic モデルという構図で、前者は表現の自由度の代わりに学習や推論が Neurosymbolic モデルよりも難しくなる傾向がある。

推論の難しさは、具体的には学習の不安定性や、適用可能なデータセットが限定されるという形で現れている。まず前者について、Scene Mixture モデルでは複数の潜在変数を仮定しているが、いずれか一つ、もしくは少数の潜在変数を用いるだけでも目的関数がある程度最適化することが可能であり、他の潜在変数は何の情報も保持しない局所解が発生し得る。これによ

り、各潜在変数が物体に対応した表現にならなかったり、初期条件の違いなどによって最終的な学習結果が大きく変化することになる。後者については、現状では出現する物体の種類が少なく、物体のテクスチャや物体以外の背景が単純な場合に限って物体ごとに分割する表現が得られており、主に CG で作成されたトイデータに適用範囲が限定されている。物体の種類が多かったり、例えば手書き数字のように一つの種類（数字）の中でも多様性がある場合は処理が困難であり、目的関数がある程度最適化できても、物体に対応した潜在表現にはならず意味のない分割となってしまう [25]。

また、モデルアーキテクチャの観点からも同様の問題が生じる可能性がある。シーン解釈モデルは畳み込みニューラルネットワーク（CNN）を用いて実現されているが、CNN はその構造上受容野が限られる。そのため、物体の形状や全体的な特徴よりもテクスチャなど、より局所的な特徴に注目する傾向がある [81, 82]。畳み込み処理は画像において近接する領域が強い関係性を持つことや、並進対称性を帰納バイアスとして導入したアーキテクチャであり、この工夫により計算コストと学習難易度を下げている。しかし、受容野が制限されることによって、例えば大きな物体を認識したり、離れた箇所の関係性を捉えることが難しくなる可能性がある。こうした局所的な特徴に過度に注目してしまう性質はシーン解釈に望ましいものではなく、物体が適切に分割されない局所解を招いたり、一部の潜在変数しか情報を保持しなくなるなど、上述の推論・学習の問題と同様の結果をもたらす恐れがある。

そこで本章では、シーン解釈の手法の一つである Scene Mixture モデルについて、上述の問題点を解決する手法を提案する。まず推論の難しさについてはここまでの章でも述べてきたように、物体に関する事前知識、つまり帰納バイアスの不足という観点から解釈することができると考えている。物体とは物理法則やデータの性質のみから定義可能なものではなく、人間が便宜上の都合から構築してきた記号的な概念である。そのため、人間が期待した通りに物体認識が行われるためには何らかの形でモデルに物体の定義を教えなければならない。しかし一般的な深層生成モデルと同様に、シーン解釈モデルの学習は教師なしで行われる。そのため、物体についての十分な帰納バイアスが与えられておらず、上述のような学習・推論における問題や、適用範囲の限界が生じるものと考えられる。5 章では背景のデータ集合を補助情報として利用することで帰納バイアスを導入したが、本章では新たなデータを用いずに可能な手法として、自己教師あり学習の利用を提案する。

自己教師あり学習とは目的とする課題そのものではなく、何らかの事前課題（pretext task）によってモデルを事前学習する手法である。事前課題を解くことで、様々なタスクに有効な特徴抽出器を学習することが可能であり、近年では分類問題やセグメンテーション課題におい



て、教師あり学習に迫る性能を発揮することが知られている [83]. 特徴抽出に適した事前課題を与えることによって、モデルに何らかの物理的な前提知識、つまり帰納バイアスが与えられることになる. そのため自己教師あり学習によって、教師なし学習では不足していた物体についての帰納バイアスがモデルに導入されることになり、本研究の目的である推論や学習の安定化に寄与することが期待できる. 本研究では、対照学習を用いた自己教師あり学習を事前学習に用いることで、Scene Mixture モデルの学習の安定化や、局所解の防止に役立つことを実験によって示す.

次に、モデルアーキテクチャの問題として挙げた CNN の性質について、大域的な特徴を捉えるためにこれまで様々な工夫が行われてきた [84, 85]. これに対し、近年は畳み込み処理を用いずに Transformer[44] を用いて画像処理を行う研究 (Vision Transformer) が盛んに行われている [12, 86, 87, 88, 89]. Transformer はその構造上、離れた場所に注目するために層を重ねる必要がなく、入力に近い層でも広い受容野を確保することが可能である [12]. これは大域的な特徴や関係性を表現するために望ましい性質である.

こうした利点はシーン解釈においても有効に働くと考えられ、本研究では Transformer を用いたシーン解釈のアーキテクチャを提案する. 大域的な特徴を捉える能力は物体を捉え、認識精度の向上に寄与することが期待できる. 具体的には、大きな物体の認識や、離れた箇所の関係を捉えることに優位であると期待される. また、GENESIS[26] では物体間の関係性をモデル化するために LSTM[60] を用いているが、提案手法では Transformer は CNN の代替として特徴抽出を行うだけでなく、この物体間の関係性を表現する役割も同時に担う. これによって訓練速度の向上や、物体同士の関係性の表現力の向上が期待される.

一方で、上述の利点に対して Vision Transformer は学習が難しいことも知られている. CNN と同等以上の性能を発揮するには、大量のデータを用いたり、CNN を教師ネットワークとした蒸留 [90] によって学習を補助する必要がある [86]. そのため、上述した自己教師あり学習については帰納バイアスの導入に加えて、Transformer を用いたアーキテクチャの学習を補助する役割も期待される. 本研究で提案する、自己教師あり学習の利用と Transformer を用いた新たなアーキテクチャの両者を導入することで、物体の表現を獲得する能力の向上、ひいてはセグメンテーションや生成の品質などシーン解釈手法としての性能向上が期待される.

本研究ではシーン解釈の既存手法で広く扱われている Objects Room データセット [66], より複雑な背景やテクスチャを含む ShapeStacks データセット [91], そして既存手法では適切に扱えていない Multi-MNIST データセット [27] において提案手法を検証する. また、Multi-MNIST データセットにおいては安定して物体の表現を獲得するために、潜在変数に対

する Cosine 距離の制約を用いることを新たに提案する。

本研究の貢献は以下の通りである

- 自己教師あり学習によってシーン解釈モデルに帰納バイアスを与える方法を提案し，それによって学習の不安定性や局所解に陥る問題が緩和されることを示した。
- Transformer を用いたシーン解釈モデルを提案し，特に自己教師あり学習と併用することで物体の表現を獲得する能力が向上することを Objects Room データセット，ShapeStacks データセット，Multi-MNIST データセットにて確認した。
- 新たに潜在変数に対する Cosine 距離の制約を提案し，この導入によって Multi-MNIST データセットにおいても安定して物体の認識を行うことが可能となることを確認した。

## 6.2 関連研究

### 6.2.1 シーン解釈手法とランダム生成

本研究は研究 1 (5 章) と同様に，混合ガウス分布を用いたシーン解釈手法を扱う。混合ガウス分布の確率変数の推論は典型的には EM アルゴリズムのような反復的な手法で行われる。先行研究 [25] では反復的な推論 (Iterative Amortized Inference (IAI)) [55] を用いているが，これは反復する回数だけ生成を行う必要があり，計算コストが非常に高くなってしまう。また，IAI によって潜在変数は事前分布から離れてしまうため，事前分布からのサンプリングによる新たなシーンの生成も困難となる。MONet[24] では反復的な最適化を用いる代わりに，マスクを先に決定論的なネットワークで近似し，それを用いて空間的混合ガウス分布の計算，つまり生成を行っている。しかし，マスクの近似が確率モデルとして定式化されていないため，新たなシーンの生成が困難である。

マスクの計算も確率モデルに組み込み，かつ潜在変数同士の関係を自己回帰によってモデル化することで新たなシーンの生成を可能にした手法が GENESIS である [26]。各潜在変数を独立にサンプリングしてしまうと，物体同士の位置関係や影の方向などを考慮しないことになり，シーン全体として物理的な一貫性のない生成結果になってしまう。こうした問題を自己回帰による推論モデルと，事前分布の導入により解決している。

シーン解釈において新たなシーンを物理的な一貫性を担保して生成できるのは Scene Mixture モデルでは GENESIS, Neurosymbolic モデルでは Generative Neurosymbolic Machines[30] のみであるため，ベースの先行研究として GENESIS を選択した。GENESIS の手法について

は、6.3 節で詳述する。

### 6.2.2 自己教師あり学習

目的とする課題と正解データを用いて学習する教師あり学習に対し、自己教師あり学習は目的とする課題とは別の事前課題 (pretext task) を用いて、様々な課題に有効な特徴量を教師なしで獲得する枠組みである。目的とする課題は分類課題や物体検出、セグメンテーションタスクなど任意であり、事前学習した重みを固定し、出力部分の分類器のみを学習することによってこれらの課題を解くというのが一般的な設定である。自己教師あり学習における事前課題は、データに関する何らかの帰納バイアス (前提知識) を導入するものになっており、これまでに様々な手法が提案されている。例えば画像を対象として、物理的な前提知識を利用した課題を提案している研究がある [92, 42]。[92] では「視覚的な特徴は画像をいくつかに区切って数えてから足し合わせても、全体で数えても同一になるはず」という仮定に基づき、特徴を数え上げる事前課題 (counting) を構築している。具体的にどのような特徴を数えているかは分からず、恐らく人間に理解できないものではあるが、様々な課題で有効であることが示されている。また、[42] では入力画像を回転させ、その回転角度を識別する課題を提案している。これは画像や、画像に含まれる被写体の回転不変性を仮定して作成した課題であり、シンプルだが実験的に有効性が示されている。

一方で近年は Instance Discrimination と対照学習を組み合わせた手法が高い性能を発揮している [43, 93]。ある入力画像に対してカラーノイズや幾何的な変換など任意のデータ拡張を行ったものを用意し、そこから得られた特徴量を  $\mathbf{h}$  とする。そして、同じ画像に対して異なるデータ拡張を施したものを Positive Sample、異なる画像に対して任意のデータ拡張を行ったものを Negative Sample として、Positive Sample から得られた特徴量と  $\mathbf{h}$  との距離は近づけ、Negative Sample から得られた特徴量と  $\mathbf{h}$  との距離は遠ざけるという学習を行う方法である。ここで Instance Discrimination とは、1つの画像を1つのクラスと見立て、被写体 (Instance) レベルの識別を行うことを指している。また、対照学習は Positive Sample と Negative Sample を用いた学習のことであり、上記はこれらの組み合わせとなっている。

異なる変換が加えられた場合に同一の対象かどうかを識別するためには、画像中の被写体の性質を捉える必要があるため、特徴量抽出器の獲得に有効な課題となっている。この事前課題も画像の変換に対して被写体の性質は変化しないという物理的な前提知識を与えることに相当し、上述の回転や数え上げ課題と根本的な発想は同じである。

本研究ではモデルに物体を認識するための帰納バイアスを導入する方法として、自己教師あり学習の利用を試みる。具体的には、入力画像の特徴抽出を行う畳み込みニューラルネットワークの事前学習を Bootstrap Your Own Latent (BYOL) [94] によって行う。BYOL 以前の手法は Positive Sample に対し、大量の Negative Sample を用意することで学習を行っていたが、BYOL では学習方法の工夫により Negative Sample を用いずに学習することが可能となっている。BYOL では、パラメータ  $\theta$  のネットワーク A と、同じ構造で異なるパラメータ  $\xi$  を持つ学習対象のネットワーク B を考える。学習は同じ入力画像に対して 2 通りのデータ拡張を行ったものをそれぞれのネットワークに入力し、得られた特徴量の距離を近づけることで学習を行う。ただし、ネットワーク B のパラメータ  $\xi$  はネットワーク A のパラメータ  $\theta$  の移動平均になっており、学習によって直接更新されるのは  $\theta$  のみである。この工夫によって Negative Sample が不要になり、BYOL ではそれ以前の手法に対してバッチサイズに対するロバスト性が得られたり、計算量が軽減されるという利点がある。さらに、性能もそれ以前の手法よりも高いことが経験的に確認されている。

本研究では、こうした学習の安定性や性能の高さから BYOL を自己教師あり学習の手法として選択した。具体的な導入方法については手法の 6.3.2 節にて述べる。

### 6.2.3 Vision Transformer

Transformer[44] は自然言語処理の分野で提案された手法で、同分野において高い汎化性能を実現している。2020 年以降、自然言語処理だけでなく画像処理に関しても Transformer が有効であることが示されており、画像処理のための Transformer は Vision Transformer と呼ばれている [12, 86]。また、画像分類だけでなく、物体検出やセマンティックセグメンテーションに応用した研究も存在している [87, 95, 96]。

画像処理に Transformer を用いる利点として、一度に注目することができる画像の領域の広さ、つまり受容野の広さが挙げられる。従来、深層学習における画像処理では畳み込みニューラルネットワーク (CNN) を用いることが一般的であったが、CNN では局所的な特徴抽出を繰り返すことで処理を行うため非常に多くの層を重ねなければ受容野を広げることができなかった。それに対し Transformer では self-attention 機構により、各層で画像全体に注目することが可能である。そのため、離れた場所の関係性を捉えることや、画像中の重要な場所に柔軟に注目 (attention) をかけることが可能であり、これは CNN が苦手とする処理である。

このような利点の一方、Transformer は CNN に比べて学習の難しさが知られており、大量

のデータを用いなければ良い精度が出ない傾向がある．例えば [12] では 3 億枚もの画像を用いて学習を行っており，データ量が少ない場合は精度が落ちている．その後，モデル構造の最適化や蒸留（distillation）によってデータ効率の向上を試みる研究が盛んに行われている [86, 97]．

Vision Transformer の研究 [12] では入力画像をいくつかの格子状の領域（パッチ）に分割し，それらを Transformer に入力する形でモデル全体を Transformer のみで構成することが多い．また，予め CNN を用いてある程度の特徴抽出と次元削減を行った上で Transformer に入力するという方式も考えられ，物体検出の研究 [87] ではそのような方式を用いている．この場合は全て Transformer を用いるよりも学習に必要なデータ量が少なくなることが期待でき，本研究においても CNN と Transformer を併用する方式を採用している．

## 6.3 手法

6.1 節において，現在の Scene Mixture モデルが持つ問題点について述べた．本研究で指摘する問題点は主に二点で，一つは推論と学習の難しさの問題，もう一点はアーキテクチャの問題である．本章ではこれらの問題を解決する方法を提案する．また，今回 Scene Mixture モデルの先行研究としては GENESIS[26] を想定している．現状では GENESIS がサンプリングによって新たなシーンの生成を行うことが可能な唯一の Scene Mixture モデルであるため，ベースの手法として選択した．

まず前者について整理する．既存手法では，各潜在変数が物体に対応しない表現を獲得する局所解に陥ってしまったり，初期条件の違いによって試行ごとに結果が変わり，期待した精度が得られない場合がある．本研究では，こうした推論や学習の難しさを帰納バイアスの不足という観点から考える．物体を認識し，対応する潜在表現を得るためには物体に関する事前知識や適切な制約等が必要であるが，教師データが与えられないシーン解釈の問題設定では，別の何らかの方法でモデルに帰納バイアスを与えなければならない．そこで，本研究では自己教師あり学習の利用によって物体認識に適した特徴抽出を行うことを提案する．自己教師あり学習は適切に設計された事前課題によって事前学習を行い，様々な課題に対して有効な特徴抽出器の学習を試みる手法である．この事前課題については関連研究の 6.2.2 節で詳しく述べたが，物理的な前提知識や物体についての知識を元に設計されたものである．事前学習によってそのような前提知識，つまり物体に関する帰納バイアスがモデルに導入されることを期待している．

また、アーキテクチャについて、畳み込みニューラルネットワーク (CNN) はその構造上、入力に近い層では画像の広い領域に注目することができない。そのため、形状や全体的な特徴よりもテクスチャのような局所的な特徴に注目する傾向があり、空間的に離れた場所の関係性を捉えることを苦手としている [85]。これは大きな物体を捉えることが難しくなったり、離れた物体同士の関係性を捉えられないといった問題が生じる可能性があり、シーン解釈モデルにとって望ましくない性質である。ひいては認識精度の低下をはじめとして、物体を認識していない局所解に陥ることや学習の不安定化など、結果的に上記の帰納バイアスの不足と同様の問題を引き起こす可能性がある。アーキテクチャについても帰納バイアスの観点から捉えるならば、学習効率のために過剰な制約を導入していると解釈することが可能であり、アーキテクチャの観点からも問題を解決する必要がある。

CNN において、大域的な特徴を捉えるために様々な研究が行われてきた。具体的には、畳み込みの範囲を学習によって動的に変更したり [84]、大域的な特徴を捉えるための attention 機構を導入する研究がある [85]。これに対し、近年は畳み込みを用いずに、Transformer[44] を用いて画像処理を行う研究 (Vision Transformer) が盛んに行われている [12, 86, 87, 88, 89]。入力側の層では受容野が限られ局所的な特徴しか抽出できない CNN に対し、Vision Transformer は最初の層でも受容野の制限がないため、局所的な特徴と大域的な特徴の両方を捉えやすくなっている [12]。これは上述の CNN に関する問題点を克服し得る性質であり、シーン解釈においても大きな物体の認識や、物体同士の関係性を捉える上で有効となることが期待できる。

よって本研究では、物体認識の精度や手法の適用範囲の拡大を期待して、Transformer を用いたシーン解釈のアーキテクチャを提案する。ただし、Transformer は学習が難しいことが知られており、CNN と同等以上の性能を発揮するには、大量のデータを用いたり、CNN を教師ネットワークとした蒸留 [90] によって学習を補助する必要がある [86]。そこで、上述した自己教師あり学習については、帰納バイアスの導入だけでなく、Transformer を用いた際に学習を補助する役割も兼ねたものとなっている。

本研究における提案をまとめる。1つ目は、追加の帰納バイアスを導入し、推論・学習の安定化をする目的で、シーン解釈モデルへの自己教師あり学習の導入を行うことである。もう一つは、Transformer の導入によって、シーン解釈に必要と考えられる大域的な特徴を捉えやすいアーキテクチャを提案することである。これにより、物体の表現学習に関する能力の向上や、前者の提案と同様に学習の安定化が期待される。

自己教師あり学習の有効性については、GENESIS と Transformer を用いた提案アーキテクチャの両者に対して検証を行う。Transformer を用いたアーキテクチャについては確率モデル

は GENESIS と同様で、アーキテクチャのみを変更して比較する。今後、単に GENESIS と表記した場合は自己教師あり学習や Transformer を利用していない、オリジナルの手法を指すものとする。Transformer を用いたアーキテクチャについては GENESIS+Tr, 自己教師あり学習 (Self-Supervised 学習) を用いた場合は GENESIS+SS, GENESIS+Tr+SS と表記する。以降の節では、まず確率モデルの説明を行い、次に GENESIS と、GENESIS+Tr のモデル構造について説明する。最後に、これらのモデルに対してどのように自己教師あり学習を適用するのかを説明する。

### 6.3.1 確率モデル

まず本研究のベースとなる、GENESIS の確率モデルについて説明する。シーン解釈モデルでは、複数の潜在変数  $\mathbf{z}_k$  を仮定し、それぞれが物体等の構成要素を生成する。こうして生成された構成要素を組み合わせて全体を生成するが、その組み合わせ方には 6.2 で紹介した通り、いくつかの方式がある。特に GENESIS を含む Scene Mixture モデルでは、各構成要素 (物体) に相当する画像  $\mathbf{x}_k \in \mathbb{R}^{C \times H \times W}$  に対して、セグメンテーションマスク  $\mathbf{m}_k \in \mathbb{R}^{1 \times H \times W}$  を適用して物体に相当する部分を切り出し、それらをピクセルレベルで足し合わせることで画像全体  $\hat{\mathbf{x}}$  を生成する。この過程は以下のように表される。

$$\hat{\mathbf{x}} = \sum_{k=1}^K \mathbf{x}_k \otimes \mathbf{m}_k \quad (6.1)$$

ただし  $K \in \mathbb{N}$  はスロット数と呼ばれ、分割数の上限となるハイパーパラメータである。正しく学習が行われた場合、各スロットがそれぞれ別の物体を分担して表現することになる。 $K$  は必ずしもデータに含まれる物体の数と同一の値に設定する必要はなく、物体数よりも大きな値を設定しておけば余分なスロットは何も表現しなくなることが経験的に知られている。また、 $\otimes$  は要素ごとの積を意味している。ここでは  $\mathbf{x}_k, \mathbf{m}_k$  にいずれもガウス分布が仮定されているため、ガウス混合分布で画像がモデル化されていることになる。これは空間的混合モデル (Spatial Mixture Model) と呼ばれるものであるが、シーン解釈の文脈ではシーン混合モデル (Scene Mixture Model) や、単に混合モデル (Mixture Model) などと呼ばれることが多い。

生成モデルの尤度については、以下のようになる。

$$p_{\theta}(\mathbf{x}|\mathbf{z}^c, \mathbf{z}^m) = \prod_{i=1}^D \sum_{k=1}^K m_{ik} \otimes p_{\theta}(x_{ik}|\mathbf{z}_k^c) \\ \mathbf{m}_k \sim p_{\theta}(\mathbf{m}_k|\mathbf{z}_k^m, \mathbf{m}_{1:k-1})$$

ただし、 $D$  は画像の次元（ピクセル数）、 $K$  は式 (6.1) と同様である． $\mathbf{z}_k^m \in \mathbb{R}^{d_m}$  はマスク  $\mathbf{m}_k$  に対応する潜在変数で、 $\mathbf{m}_k \in [0, 1]^{1 \times H \times W}$  とする．また、 $\mathbf{z}_k^c \in \mathbb{R}^{d_c}$  は各スロットの画像  $\mathbf{x}_k$  に対応する潜在変数である．また、マスク  $\mathbf{m}_k$  の生成については MONet[24] で提案された、Stick Breaking Process と逆畳み込みを利用した機構が用いられている．

これらを踏まえると、生成モデルの周辺尤度は以下のように表される．

$$p_\xi(\mathbf{x}) = \int \int p_\theta(\mathbf{x} | \mathbf{z}^c, \mathbf{z}^m) p_\phi(\mathbf{z}^c | \mathbf{z}^m) p_\psi(\mathbf{z}^m) d\mathbf{z}^m d\mathbf{z}^c \quad (6.2)$$

ただし

$$p_\phi(\mathbf{z}^c | \mathbf{z}^m) = \prod_{k=1}^K p_\phi(\mathbf{z}_k^c | \mathbf{z}_k^m) \quad (6.3)$$

$$p_\psi(\mathbf{z}^m) = p(\mathbf{z}_1^m) \prod_{k=1}^{K-1} p_\psi(\mathbf{z}_{k+1}^m | \mathbf{z}_{1:k}^m) \quad (6.4)$$

とする．また、 $\xi = \{\theta, \phi, \psi\}$  とまとめたものとする．式 (6.2) の右辺第二項  $p_\phi(\mathbf{z}^c | \mathbf{z}^m)$  はパラメータ  $\phi$  を持つ多層パーセプトロンでモデル化される．また、右辺第三項の  $p_\psi(\mathbf{z}^m)$  は、マスクの潜在変数  $\mathbf{z}_k^m$  の事前分布である．この分布は自己回帰によって各スロットの潜在変数  $\mathbf{z}_k^m$  間の関係をモデル化しているため、自己回帰事前分布（auto-regressive prior）と呼ばれ、ここでは LSTM[60] によって実装されている．ただし  $\mathbf{z}_{1:k}^m$  は  $\mathbf{z}_1^m$  から  $\mathbf{z}_k^m$  までの  $k$  個の変数を意味する．また、 $p(\mathbf{z}_1^m)$  にはガウス分布  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  など、サンプリング可能な任意の分布が設定される．実際に  $\mathbf{z}_k^m$  を事前分布からサンプリングする際は  $p(\mathbf{z}_1^m)$  から自己回帰的に  $\mathbf{z}_k^m$  を得ることになる．

GENESIS 以前の手法では各  $\mathbf{z}_k^m$  は独立した確率変数としてモデル化されているが、実際には各物体は物理的な制約で拘束されるため独立ではない．そのため、潜在変数を事前分布からサンプリングして新たなシーンを生成する際には物体同士の関係性が考慮されず、物体の配置や大きさの面で一貫性のある生成ができない．例えば、各潜在変数を独立にサンプリングした結果、物体同士や物体と床の位置関係が適切に考慮されず、物理的にあり得ない場所に物体が生成されたり、適切な大きさにならない可能性がある．また、本研究において潜在変数が従う分布はいずれもガウス分布が仮定されている．

潜在変数の推論は一般的な VAE[16] と同様に、Amortized Inference によって行われる．上記式 (6.2) の周辺尤度を解析的に最大化することはできないため、Amortized Inference では潜在変数の近似事後分布（推論モデル）を導入し、これを周辺尤度の変分下界の最大化によって学習することによって推論を実行する．各潜在変数における推論モデルは以下のように表さ



れる.

$$\mathbf{z}_k^m \sim q_\mu(\mathbf{z}_k^m | \mathbf{x}, \mathbf{z}_{1:k-1}^m) \quad (2 \leq k \leq K) \quad (6.5)$$

$$\mathbf{z}_1^m \sim q_\mu(\mathbf{z}_1^m | \mathbf{x})$$

$$\mathbf{z}_k^c \sim q_\eta(\mathbf{z}_k^c | \mathbf{x}, \mathbf{z}_k^m) \quad (1 \leq k \leq K) \quad (6.6)$$

式 (6.5) は式 (6.4) に対応する推論モデルであり, これは入力  $\mathbf{x}$  のエンコーダ (CNN) と, 潜在変数間の関係性を自己回帰的に表現するための RNN などを組み合わせて実現される. また, 式 (6.6) はマスク  $\mathbf{m}$  と  $\mathbf{x}$  を入力として潜在変数  $\mathbf{z}^c$  を推論するもので, 後段の VAE のエンコーダに相当する.

学習は周辺対数尤度  $\log p_\theta(\mathbf{x})$  の変分下界 (ELBO) を最大化することによって行われる. 上記の確率モデルの定式化を踏まえると, 目的関数  $\mathcal{L}(\mathbf{x})$  は以下ようになる.

$$\begin{aligned} \mathcal{L}(\mathbf{x}) = & \mathbb{E}_{q_\rho(\mathbf{z}^c, \mathbf{z}^m | \mathbf{x})} [p_\theta(\mathbf{x} | \mathbf{z}^c, \mathbf{z}^m)] \\ & - \beta \left\{ KL[q_\mu(\mathbf{z}^m | \mathbf{x}) \parallel p_\psi(\mathbf{z}^m)] + \mathbb{E}_{q_\mu(\mathbf{z}^m | \mathbf{x})} \left[ KL[q_\rho(\mathbf{z}^c, \mathbf{z}^m | \mathbf{x}) \parallel p_\phi(\mathbf{z}^c | \mathbf{z}^m) p_\psi(\mathbf{z}^m)] \right] \right\} \end{aligned} \quad (6.7)$$

ただし, 上式の  $q_\rho(\mathbf{z}^c, \mathbf{z}^m | \mathbf{x})$  については  $\rho = \{\eta, \mu\}$  であり, 以下の通りである.

$$q_\rho(\mathbf{z}^c, \mathbf{z}^m | \mathbf{x}) = q_\eta(\mathbf{z}^c | \mathbf{z}^m, \mathbf{x}) q_\mu(\mathbf{z}^m | \mathbf{x})$$

上記の式において, スロットのインデックス  $k$  については表記が煩雑になるため省略している. ここで KL は Kullback-Leibler 情報量を意味し,  $\beta$  は KL 項の影響力を決定する係数である.  $\beta$  は一般にはハイパーパラメータで, 学習段階によって変化させるアニーリングも行われる. この係数を自動決定する最適化手法として, GECO[98] がある. これはラグランジュの未定乗数法で再構成誤差の目標値を制約条件に,  $\beta$  を未定乗数として自動決定する最適化手法である. つまり  $\beta$  の代わりに再構成誤差 (式 (6.7) 右辺第一項) の目標値がハイパーパラメータとなるが, これは最終的な値を 1 つ設定すれば良いので, 学習ステップに応じた最適な  $\beta$  を選択するよりも容易である. GENESIS では GECO を用いて最適化を行っており, 本研究でもこれに従って GECO を利用した.

### 6.3.2 モデル構造

本節では GENESIS と, 今回提案する Transformer を用いたモデル (GENESIS+Tr) の構造について述べる. 確率モデルはいずれの手法にも共通で, 前節で説明したものを用いている.

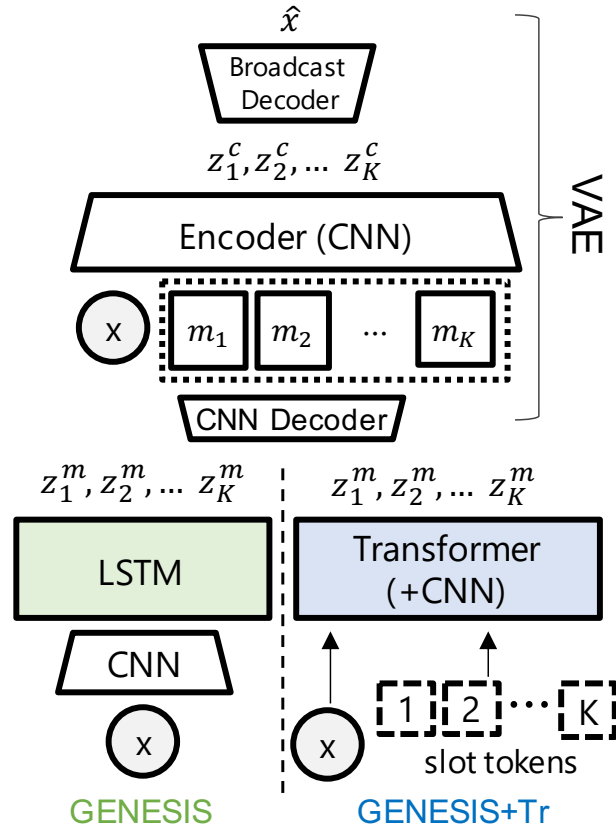


Fig.6.1 GENESIS と Transformer を用いたモデル構造 (GENESIS+Tr). 後段の VAE 部分は共通した構造となっている.

### GENESIS のモデル構造

まず GENESIS のモデル構造について説明する. GENESIS は, 物体のマスクを出力する前段の機構 (first stage) と, そのマスクで指定された領域を表現する後段の機構 (second stage) に大きく分けられる. このようなモデルが 2-stage モデルと呼ばれるのに対し, IODINE[25] のように 1 段階のモデル構造を反復的な最適化によって学習するモデルは 1-stage モデルと呼ばれる.

first stage では, CNN で入力  $x$  を特徴抽出した後, 自己回帰を用いた推論モデルによってマスクの潜在変数  $z_k$  を得る. ここまでの過程は式 6.5 に相当し, GENESIS ではこの自己回帰モデルは LSTM を用いて実現されている. CNN で得られた特徴マップは, 全結合層によって次元のベクトルに圧縮され, LSTM へと入力される. また, この推論モデルに対応する, 式 6.4 の学習可能な事前分布を用いることで新たなシーンの生成を可能としている. この事前分布も自己回帰モデルであり, LSTM によって実現されている.

second stage は VAE の枠組みになっており、入力  $\mathbf{x}$  とマスク  $\mathbf{m}_k$  を入力として潜在変数  $\mathbf{z}_k^c$  を推論するエンコーダと、各マスクの領域ごとの画像  $\mathbf{x}_k$  を生成するデコーダによって構成されている。Fig.6.1 は提案手法の概要図であるが、second stage の VAE については GENESIS と提案手法で同一であるため、以後本節で説明する内容と対応している。デコーダ  $p_\theta(\mathbf{x}|\mathbf{z}^c, \mathbf{z}^m)$  全体は前節で示した通り、混合ガウス分布となっており、各構成要素  $\mathbf{x}_k$  の生成は Spatial Broadcast Decoder で行われる [73]。また、エンコーダ  $q_\eta(\mathbf{z}_k^c|\mathbf{x}, \mathbf{z}_k^m)$  は CNN であり、 $\mathbf{x}$  と前段で得られたマスク  $\mathbf{m}_k$  を入力とする。ただし潜在変数  $\mathbf{z}_k^m$  からマスク  $\mathbf{m}_k$  を生成する過程は、CNN と Stick Breaking Process (SBP) を用いた機構が用いられている [24]。これについては SBP を用いずに、単に softmax 関数によって  $\sum_{k=1}^K \mathbf{m}_k$  を 1 にすることも考えられる。しかし我々が検証した限りでは、MONet や GENESIS 等の 2-stage モデルにおいて SBP を用いない場合は全てのスロットが同じマスク（画像全体が  $1/K$  の値になる）になってしまった。そのため、適切な学習のために現状ではこの機構が必要となっている。

### Transformer を用いたアーキテクチャ

提案アーキテクチャ (GENESIS+Tr) でも 2-stage モデルを採用しており、後段 (second stage) の構造と確率モデルについては GENESIS と同様である。GENESIS+Tr では式 6.5 の first stage の推論モデルに Transformer を用いることを新たに提案する。つまり GENESIS との差分は first stage のアーキテクチャである。提案手法のモデル構造を Fig.6.1 に示した。

GENESIS+Tr において、first stage の推論モデルは CNN のエンコーダと Transformer によって構成される。CNN のエンコーダによって得られた特徴マップ  $\mathbf{f} \in \mathbb{R}^{C \times D}$  は全結合層によって次元削減するのではなく、チャンネル数  $C$ 、次元  $D$  の系列として Transformer に入力する。VisionTransformer では画像を矩形に区切ってパッチ状にしたものを直接 Transformer に入力しているが、この方式では学習に大量のデータが必要となる傾向がある。本研究で扱うデータセットの規模は高々 10 万データのオーダーだが、VisionTransformer を十分に訓練するために [12] では 3 億枚もの画像を利用している。そのため、本研究では Transformer に画像を直接入力するのではなく、CNN によってある程度次元を落としてから特徴量を入力することで、必要なデータ量やモデルの次元を抑える。これは [12] の一部や、[95] において採用されている方式である。つまり提案アーキテクチャにおいて Transformer は特徴抽出と、各潜在変数の推論を兼ねた機構として用いられている。

また、GENESIS+Tr では CNN が出力した特徴マップと同時に、 $K$  個 (スロット数) の学習可能パラメータ、Slot Token を Transformer に入力する。具体的には、特徴マップ  $\mathbf{f} \in \mathbb{R}^{C \times D}$  と、

$K$  個の Slot Token  $\mathbf{t} \in \mathbb{R}^{K \times D}$  を結合し、長さ  $C + K$  で  $D$  次元の系列として Transformer に入力することになる。潜在変数  $\mathbf{z}_k$  の推論については、Slot Token に対する Transformer の出力に、reparameterization trick[16] を適用することで行われる。学習可能パラメータを Transformer の入力として用いるのは関連研究 6.2.3 節で述べた CLS token でも行われているが、CLS token がクラス分類のため一つだけ用いられるのに対し、Slot Token はスロットの数  $K$  だけ入力する。Slot Token の次元は特徴マップに合わせて  $D$  としている。

関連研究 (6.2.3 節) で紹介したように Transformer の構造については様々なものが提案されている。しかし本研究では Transformer 自体の構造についてはオリジナルの基本的な構造 [44] で固定し、入出力の設定や Slot Tokens の利用など、シーン解釈モデルにどのように Transformer を組み込むかという観点の探索を重視した。Transformer 自体の構造の最適化によって精度を向上させることも可能と思われるが、これについては今後の課題とする。

オリジナルの Transformer は Transformer Encoder と、Transformer Decoder に構造が分かれているが、本研究で用いたのは Transformer Encoder のみである。通常 Transformer Encoder 単体では出力の系列長が入力の系列長に固定されるが、上述の Slot Token の導入によって特徴マップのチャンネル数とスロット数  $K$  を独立に決定することが可能となっている。

Transformer モデルの次元は 256、線形層の次元は 2048、attention head の数は 4 で、活性化関数には Gaussian Error Linear Unit (GELU) を利用した [99]。Transformer の層数については、ハイパーパラメータとしてデータセットによって変更した。また、位置埋め込み (Positional Embedding) は三角関数を用いた埋め込み方法が一般的だが、学習可能パラメータを入力する方式が経験的に最も学習の安定性等の面で優れていた。これは BERT[8] などで採用されている方法である。

### 自己教師あり学習

自己教師あり学習は LSTM もしくは Transformer の前段の、入力画像の特徴抽出を行っている CNN に対して適用する。この CNN は初期の特徴抽出に関わり、Transformer への入力となるため、事前学習が学習の安定性にも寄与すると考えられる。

自己教師あり学習の手法は 6.2.2 節で述べた通り、Bootstrap Your Own Latent (BYOL) [94] を採用した。これは BYOL 以前の手法とは異なり、学習時に Negative Sample を必要としないという特徴がある。そのため、バッチサイズに対するロバスト性や計算量の軽減、学習の安定化といった利点がある。これらの理由から本研究では BYOL を自己教師あり学習の手法として選択した。

自己教師あり学習による事前学習では、Tiny ImageNet と呼ばれる ImageNet データセット [100] の小規模版を利用した。この Tiny ImageNet は被写体の切り出しやアスペクト比の調整を施され、64x64 の解像度の画像で統一されている。本研究で利用するデータセットはいずれも 64x64 であるため、これを採用した。自己教師あり学習による CNN の事前学習は 60epoch 行った。

自己教師あり学習によって事前学習したネットワークは重みを固定して用いることが一般的だが、本研究では固定せずに用いた。自己教師あり学習が頻繁に利用される課題としては分類問題や物体認識、セグメンテーションなどがある。これに対し、生成モデルのエンコーダ部分に自己教師あり学習で得られた特徴抽出器をそのまま用いてしまうと、生成や再構成に必要な視覚的な特徴量が十分に得られない場合があり、本研究では学習可能な形で導入した。

### コサイン距離による制約の導入

予備実験において GENESIS 以前の手法は Multi-MNIST データセット（6.4 節にて詳細を説明）にて物体の分割が正しく行えないことを確認していた。これは実験の章の Fig.6.5 右側において確認できる。物体を分割する解を得るために、本研究ではコサイン距離による潜在変数への制約の導入を行った。これはマスクの潜在変数  $\mathbf{z}_k^m$  に対し、コサイン距離が大きくなるように制約を課すものである。全ての  $\mathbf{z}_i^m$  と  $\mathbf{z}_j^m$  の組み合わせでコサイン距離を計算し、その和を目的関数に加える形で制約項を導入した。ただし、 $i, j \in \{1, 2, \dots, K\}$ ,  $i \neq j$  とする。制約項の強度は目的関数に加算する際の係数によって調整可能である。この制約は Multi-MNIST データセットの実験においてのみ利用した。Objects Room データセットにおいては大きな変化がないことや、一般的な制約項と同様に係数の値を大きくした場合に学習を阻害する可能性があることから、採用しなかった。

## 6.4 実験

本節では、自己教師あり学習の導入と、Transformer を用いた新たなアーキテクチャについて、どのように性能が変化するかを検証した。従って比較する対象は、ベースラインとしてオリジナルの GENESIS、自己教師あり学習を導入した GENESIS+SS、Transformer を導入した GENESIS+Tr、その両方を導入した GENESIS+Tr+SS の全 4 種となる。モデルの学習は確率的勾配降下法で行われ、バッチサイズは 32、最適化アルゴリズムは Adam[101] を学習率 0.0001 に設定して使用した。マスクの潜在変数  $\mathbf{z}_m^k$  の次元は 64、 $\mathbf{z}_k^c$  の次元は 16 とした。ま

た、いずれの実験においてもモデルの実装には Pytorch を利用した [79].

検証においては、Objects Room データセット [66] と ShapeStacks データセット [91], そして Multi-MNIST データセットを用いた. Objects Room は CG で作成されたデータセットで、複数の物体が設置された部屋を様々な組み合わせで作成し、それを任意の角度から撮影した静止画の集合となっている. ShapeStacks は同様に CG で作成されているが、より複雑な背景を含み、いくつかの物体が塔のように中央に積まれたものとなっている. Multi-MNIST は手書き数字データセットの MNIST[102] をランダムに画像中に複数並べて作成したデータセットである. 様々な研究で利用され、設定も様々であるが、ここでは Eslami らの設定に従った [27]. 変更点として画像解像度は  $64 \times 64$  に、画像中に含む数字の個数は 1 ~ 2 個でランダムに変化させている. ただし、予め乱数シードを固定して生成したデータセットを利用しているため、本研究で行った全ての実験で完全に同一の画像集合となっている.

Objects Room や ShapeStacks は視点や背景にある程度の多様性がある状況で、複数の物体を分割する必要がある. ただし、物体の種類や色は限定的であり、シーン解釈モデルの基本的なベンチマーク課題として利用されている. 一方で Multi-MNIST は手書き数字の組み合わせによる二次元的なグレースケール画像による環境で、背景や視点は固定されているが、1 種類の数字を取っても同じ形状のものは存在せず、物体の形状に関して多様である. このように Objects Room や ShapeStacks と、Multi-MNIST データセットでは難しさの性質が異なるが、既存の Scene Mixture モデルは後者のような多様な物体が含まれる環境を苦手としているため、本実験で採用した.

Objects Room と ShapeStacks の実験においては、最適化に GECO を用いた. GECO の目標値（再構成誤差の値）は、6950 とした. ただしこれは画素と RGB チャンネルで和を取った値であり、これは GENESIS の実験設定に合わせたものである. Multi-MNIST データセットの実験においては、式 (6.7) における KL 項の係数  $\beta$  を 0 から 0.5 まで 10 エポックで最大となるよう線形に変化させた. GECO を用いた場合、再構成誤差が下がらない場合は KL 項が設定した最小値を取り続けることになるが、Multi-MNIST データセットにおいては  $\beta$  の値がある程度大きくないと学習が進みづらい場合があった. そのため、各種パラメータを適切に設定して GECO を利用するか、学習の進行に応じて決定論的に変化させる方法が考えられ、本実験では後者を選んだ.

### 6.4.1 Objects Room データセットでの実験結果

本節では Objects Room データセットでの実験結果を確認する．GENESIS+Tr の Transformer は 7 層に設定した．いずれのモデルにおいても，60epoch 時点の学習結果を用いた．以下の節では，推論と新たなシーンの生成品質の定性評価，それらの定量評価，学習と推論の実行速度について検証した結果を記す．

#### 定性評価

本節では各モデルでの Objects Room データセットにおける推論結果と，生成の結果を定性的に比較する．推論結果を Fig.6.2 に，生成結果を Fig.6.4 に示した．

まず推論，つまり物体の表現学習の結果について確認する．Fig.6.2 に示した結果は，同じモデル内では同一の乱数シードが適用されている．つまり，ネットワークの重みの初期値や，データセットからの画像の抽出の順番などが一致している．モデル自体や最適化にも確率的な要素が含まれているため，完全に条件を揃えることは不可能であるが，可能な範囲での統一を試みた．図は上の行から GENESIS+Tr+SS，GENESIS+Tr，GENESIS+SS，GENESIS（ベースライン）となっており，左右の列で異なる入力データについての結果を示している．

まず自己教師あり学習の導入による変化を確認する．GENESIS と GENESIS+SS を比較すると，GENESIS では複数の物体を一つのスロットで表現してしまっている場合（左列の最下段， $k = 6$ ）があるが，GENESIS+SS ではこれらは  $k = 3$  と  $k = 5$  の 2 つに正しく分割されている．自己教師あり学習なしの GENESIS でも乱数シード次第でスロットと物体が一对一に対応する場合もあるが，図に示したような局所解に陥ってしまう場合が頻繁に見られる．

また，GENESIS+Tr と GENESIS+Tr+SS を比較すると，GENESIS+Tr が物体を無視し，意味のない分割を行ってしまっているのに対し，GENESIS+Tr+SS では物体ごとの分割を正しく行うことができている．確認した限り，GENESIS+Tr はほとんどの乱数シードで適切に物体を分割する解が得られておらず，Transformer を用いたアーキテクチャについては自己教師あり学習の導入が重要となっていた．GENESIS+Tr に関して，6.3 に異なる乱数シードを用いた場合の結果も示した．セグメンテーションの失敗の仕方は様々だが，それぞれについて結果の改善が確認できる．

次に，Transformer の導入による変化を確認する．Transformer は上述の通り，自己教師あり学習と合わせることで性能を発揮している．自己教師あり学習を用いた場合について，つまり GENESIS+Tr+SS と GENESIS+SS を比較する．右列のサンプルでの結果を確認す

ると、GENESIS+Tr+SS のみが物体の影や、大きな球体の後ろに隠れた小さな物体（球体の右上）を考慮できている。また、左列のサンプルの左端の中空の三角形の物体について、GENESIS+Tr+SS のみが中央の穴を表現できている。これは CNN よりも大域的な特徴や形状を捉えることが得意な Transformer の優位性が表れたものと考えられる。一方で、Transformer を用いたモデルでは、空（上部の水色の領域）の一部が複数のスロットに分散してしまう傾向がある。これは近接する領域に注目する CNN に対し、離れた場所も同様に扱う Transformer の特性が表れたものではないかと考えられるが、本実験のみからは判断できず今後の検証が必要である。

次に事前分布からの潜在変数のサンプリングによって、新たなシーンの生成を行った結果について確認する。Fig.6.4 では GENESIS, GENESIS+Tr, GENESIS+SS, GENESIS+Tr+SS の、4 条件の結果を示している。左の列は自己教師あり学習を用いた場合（+SS）、右の列は用いずに初期値から学習した場合（scratch）で、上の行は Transformer を用いた場合（GENESIS+Tr+SS / GENESIS+Tr）、下の行は用いない場合（GENESIS+SS / GENESIS）である。

自己教師あり学習を用いた場合（左列）の方がシーン全体としての一貫性が確保されているように見える。また、自己教師あり学習を用いた場合の GENESIS+SS と GENESIS+Tr+SS で Transformer の有無による比較を行うと、GENESIS+Tr+SS の方が異常な形状の物体や、空のアーティファクトが少ないように見える一方で、壁に穴が空いているだけで物体が生成されていない箇所が散見される。生成の品質についても次節にて定量評価を行う。

## 定量評価

本節では各モデルでの Objects Room データセットにおけるセグメンテーションと、生成の品質を定量的に比較する。

まず、セグメンテーションの精度の指標として GENESIS の論文で提案されている mean Segmentation Covering (mSC) の値を Table6.1 に示す。これは値が大きい方が正解に近いセグメンテーションとなる指標である。表の値は 3 つの乱数シードで学習したモデルから、それぞれ 500 枚のテスト集合に対するセグメンテーションを行い、その結果を評価したものである。誤差は標準誤差を示している。なお、GENESIS の論文と同様に定量評価は物体のセグメンテーションのみについて行っており、背景については考慮していない。これは背景の分割方法は任意であり、必ずしも正解データに近づく必要がないためである。

表の上段は自己教師あり学習なしで一から学習した場合、下段は自己教師あり学習を用いた場



合 (+SS) の結果である。自己教師あり学習によって、いずれのモデルでもスコアの平均値が向上し、分散も減少していることが分かる。モデル間の比較では、GENESIS+Tr+SS が最も良い結果となった。

次に、生成品質の指標として Frechét Inception Distance (FID) を使用した。FID は本来のデータ集合と、比較対象の集合の距離を示す指標であり、小さいほど生成の品質が高いことを意味する。各条件での値を Table 6.2 に示した。これは 3 つの乱数 seed で学習したモデルで各々 10000 枚の画像を生成し、この画像集合を用いて FID を評価した結果である。誤差は標準誤差を示している。

いずれのモデルでも自己教師あり学習によって平均値が向上し、分散が大幅に減少していることが確認できる。最も良い結果を出したのは GENESIS+SS となった。GENESIS+Tr+SS では自己教師あり学習を用いない場合 (GENESIS+Tr) に対して大きく品質が向上しているが、ベースラインの GENESIS よりも低いスコアとなった。

GENESIS+Tr+SS の FID がベースラインよりも良い値にならなかった理由について考察する。GENESIS では推論時に自己回帰的に潜在変数  $z_m^k$  間の関係を表現し、それに対応する自己回帰事前分布からのサンプリングによって新たなシーンの生成を行っている。GENESIS では事前分布と事後分布 (推論モデル) のいずれにも LSTM を用いているのに対し、GENESIS+Tr の場合は事後分布には Transformer、事前分布には LSTM を用いている。学習時にこれらの分布を KL ダイバージェンスによって近づけており、基本的にはこの KL ダイバージェンスが小さければ生成の品質も高くなることが期待できるが、本実験の GENESIS+Tr+SS の結果を確認すると GENESIS と同程度か、それよりも小さい値になっていた。これを踏まえると、GENESIS のような潜在変数間関係のモデル化を行った場合、各スロットの事前分布と事後分布が近づくことによって、各スロットの個別の生成の品質のみが保証され、必ずしもシーン全体としての一貫性は得られない可能性がある。改めて Fig. 6.4 を確認すると、GENESIS+Tr+SS では物体を描くために開けておいた壁の部分に、最終的に何の物体も描かれていないというケースが散見される。また、GENESIS では床や空にアーティファクトが出ていたり、不自然な形状の物体がよく見られるが、GENESIS+Tr+SS では空や床を含む、個別の物体ごとの生成結果については自然なものが多く、上記の考察と合致する特徴である。これにより FID の差が生じた可能性がある。

また、他の可能性として FID 自体の問題も考えられる。FID は学習済みの CNN を用いた指標であるため、物理的な一貫性や視覚的な自然さに関わらず、視覚的に大きな領域が元のデータ集合に含まれないような場合には、抽出される特徴量に大きな差が生じスコアが下がると考

Table6.1 Objects Room データセットにおける mSC

model	GENESIS	GENESIS+Tr
scratch	0.55 $\pm$ 0.04	0.49 $\pm$ 0.14
self-supervised	0.57 $\pm$ 0.03	<b>0.59</b> $\pm$ 0.04

Table6.2 Objects Room データセットにおける FID

model	GENESIS	GENESIS+Tr
scratch	65.0 $\pm$ 7.1	96.29 $\pm$ 8.7
self-supervised	<b>59.2</b> $\pm$ 3.3	77.5 $\pm$ 2.1

えられる。FID は広く用いられている指標であり、シーン解釈の生成品質の評価でも FID は一般に用いられているが、実画像で学習されているために Objects Room のような CG のデータセットに適用した場合は必ずしも官能評価と一致しない可能性がある。

生成に関する問題を解決する方法として、モデルの改良の観点では2通り考えられる。1つは単純に、事前分布にも Transformer を用いることである。事前分布と事後分布に同じアーキテクチャを採用することで、各スロットの KL ダイバージェンスの最小化が結果的に類似した系列モデルの学習につながることを期待できる。もう1つは、シーン全体を表現するグローバルな潜在変数を追加で用意することである。生成の品質に関する問題について、本質的な解決のためには現在行っているような物体レベルの表現学習に加え、シーン全体の構成についての表現学習が必要である。グローバルな潜在変数  $\mathbf{z}_g$  を用意し、 $p(\mathbf{z}_k^m | \mathbf{z}_g)$  と階層化された確率モデルを仮定するか、 $p(x | \mathbf{z}_k^m, \mathbf{z}_g)$  とシーン全体の表現と物体の表現を分担して表現することが考えられる。

#### 6.4.2 ShapeStacks での実験結果

本節では ShapeStacks データセットにおける推論結果を確認する。これは前節の Objects Room に加え、異なるデータセットでの結果を確認することで手法の有効性を検証するものである。ShapeStacks は複数のブロックが縦に積まれた系を撮影したデータセットで、Objects Room よりも複雑な背景を含んでいることや、重なりによる物体のオクルージョンが頻繁に発生することから、難易度がより高くなっている。実験設定は Objects Room と同様となっている。

6.3 と 6.4 にそれぞれ ShapeStacks データセットでの mSC と FID を示した。GENESIS,

Table6.3 ShapeStacks データセットにおける mSC

model	GENESIS	GENESIS+Tr
scratch	$0.57 \pm 0.07$	$0.53 \pm 0.05$
self-supervised	<b><math>0.58 \pm 0.06</math></b>	<b><math>0.58 \pm 0.07</math></b>

Table6.4 ShapeStacks データセットにおける FID

model	GENESIS	GENESIS+Tr
scratch	$197.9 \pm 4.0$	$231.5 \pm 3.1$
self-supervised	<b><math>191.5 \pm 2.1</math></b>	$226.4 \pm 3.2$

GENESIS+Tr とともに自己教師あり学習の導入による改善傾向が見られた。mSC については本データセットでは自己教師あり学習を用いた GENESIS と GENESIS+Tr がほぼ同等のスコアとなっており、FID に関しては Objects Room の場合と同様 GENESIS+SS が優位となった。本データセットで物体は必ず画像の中央に配置されており、Objects Room データセットのようにシーン全体に分散してはいない。そのため、Transformer の離れた場所への注目が得意であるという利点が薄くなっている可能性が考えられる。

### 6.4.3 Multi-MNIST での実験結果

本節では、Multi-MNIST データセットにおける推論結果を確認する。Multi-MNIST データセットについては物体（数字）の種類や形状が Objects Room データセットよりも多様であり、色の違いによる手がかりも無いという点でより難しい課題となっている。定性的な結果を Fig.6.5 に示した。また、mSC・FID によるセグメンテーションと生成品質の定量評価を Table6.5 に示した。図はいずれも 50epoch 時点の結果である。左から、コサイン距離の制約を導入した GENESIS+Tr+SS と GENESIS+SS、制約を導入していないオリジナルの GENESIS の結果を示した。コサイン距離の制約については 6.3.2 節で述べたものである。

まず定性的な結果について考える。オリジナルの GENESIS (Fig.6.5:右) では数字は分割されず、1つのスロット ( $k = 1$ ) が全てを表現してしまう局所解となっている。また、マスクで数字の形状を表現し、VAE の出力  $\mathbf{x}_k$  は単に全体的に白い画像を出力するだけになっている。つまりいずれの潜在変数  $\mathbf{z}_k^m, \mathbf{z}_k^c$  においても数字単体に関する表現が獲得されていない。自己教師あり学習と制約を導入した GENESIS+SS (図:中央) では物体の分割には成功し、 $\mathbf{z}_k^c$  が数字の表現を獲得できている。しかし、セグメンテーションマスクは数字の形状を無視し、領域を

全体的に分割するだけとなっている。GENESIS+Tr+SS (図:左) ではセグメンテーションマスクが数字の存在する領域をより細かく捉えており、両方の潜在変数が数字についての情報を保持している。GENESIS+Tr+SS は学習の初期ではより細かく数字の形状に沿ったセグメンテーションを行っていたが、学習後半になるにつれ図のような解になることが確認された。これは KL 項の係数のアニーリングによって学習後半により強い正則化が働くため、数字の形状に沿わずに一般的な形状に近づいていったものと考えられる。そのため、潜在変数に仮定する分布をより複雑なものにすることで、数字の形状を捉えやすくなる可能性がある。

次に定量評価について、Objects Room データセットと同様にセグメンテーションの指標として mSC を、生成の指標として FID を用いて、3 つの異なるランダムな乱数シードで学習した場合の平均と標準誤差を示している。Table 6.5 中の Reg. という表記はコサイン距離の制約の有無を意味しており、表の上から GENESIS, GENESIS+SS に制約項を加えたもの、GENESIS+Tr+SS に制約項を加えたもの、となっている。事例ごとの指標を見ると、Fig. 6.5 の中央、GENESIS+SS+Reg. で見られたようなセグメンテーションでは mSC がほぼ 0 となった。一方、同図右の GENESIS の場合に見られるような、分割に失敗した場合のセグメンテーションがスコアとしてはより高くなっている。これは画像に含まれる数字が 1 つの場合、セグメンテーション自体は成功していることになり、数字が 2 つの場合もいずれか片方の数字のセグメンテーションとして評価され、ある程度のスコアが出るためである。そのため、mSC のスコアは分割に成功している場合についてのみ比較することとし、GENESIS における mSC の値は参考までに記載した。Transformer を用いた場合 (GENESIS+Tr+SS+Reg.) は、学習時の乱数シードによって Fig. 6.5 のような数字に沿ったセグメンテーションが得られる場合と、同図中央の GENESIS+SS+Reg. と同様のセグメンテーションになる場合の両方が確認された。数字の形状を捉えるのに成功した場合は高いスコアが得られたが、GENESIS+SS+Reg. のようなセグメンテーションになった場合にはスコアがほぼ 0 となるため、結果的に標準誤差が比較的大きな値となった。

Transformer を用いた場合にのみ Fig. 6.5 左のような数字に沿ったセグメンテーションマスクが得られたが、これは受容野が制限されず、離れた場所の関係を表現しやすい Transformer の利点が現れたものと考えられる。数字を分割するためには、離れた位置にある数字が独立して生成されたものであることを認識しなければならず、Transformer の Attention 機構はこの点で有利に働いているものと考えられる。

Table6.5 Multi-MNIST データセットにおける定量評価. Reg. は制約項の有無を意味する. 誤差項は標準誤差である. 括弧内は比較不可であるが, 参考値として記載した.

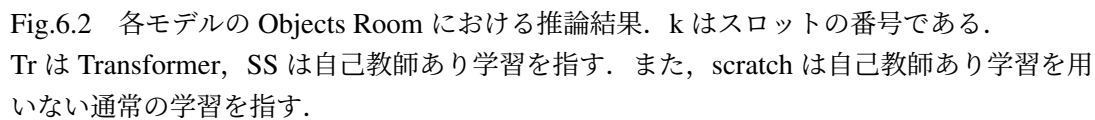
model	mSC	FID
GENESIS	(0.43 $\pm$ 0.04)	386.5 $\pm$ 14.7
+SS+Reg.	0.07 $\pm$ 0.01	342.5 $\pm$ 8.8
+Tr+SS+Reg.	<b>0.17 <math>\pm</math> 0.19</b>	<b>330.7 <math>\pm</math> 1.5</b>

## 6.5 結論

本章では, シーン解釈において帰納バイアスの不足が学習の不安定性や, 物体認識に失敗する局所解に陥る原因であると考え, それを自己教師あり学習により緩和することを提案した (GENESIS+SS). また, CNN アーキテクチャの局所的な特徴を重視する性質がシーン解釈に望ましくないことを指摘し, シーン解釈における Transformer を用いたモデル構造 (GENESIS+Tr) を提案した.

実験においては, GENESIS, GENESIS+Tr とともに自己教師あり学習の利用によって学習の安定化や定量評価が向上することが確認された (GENESIS+SS, GENESIS+Tr+SS). 特に GENESIS+Tr については, スクラッチでの学習では局所解に陥ってしまうことがほとんどだったが, 自己教師あり学習の利用によって安定した結果が得られるようになった. Objects Room データセットのセグメンテーションにおいては GENESIS+Tr+SS が最も良い結果となり, 生成品質 (FID) については GENESIS+SS が最も良い結果となった. ShapeStacks についても Objects Room と同様の傾向であったが, GENESIS+SS と GENESIS+Tr+SS のセグメンテーションはほぼ同等となった. Multi-MNIST データセットにおいてはいずれも GENESIS+Tr+SS が最も良い結果となった.

本研究では既に確立された既存の自己教師あり学習手法を用いたが, 今後の研究としてシーン解釈に適した事前課題を考案し, 性能を向上させることが考えられる. また, Slot Tokens の利用や Transformer の層数・モデルの次元のようなハイパーパラメータなど, Transformer をシーン解釈手法に組み込むための探索は行ったが, self-attention の方式や内部の結合, 入力方法といった, Transformer 内部の構造については行っていない. 近年 Vision Transformer の構造については盛んに研究が行われており, 構造の最適化によって認識性能やサンプル効率の向上が期待できるが, これは今後の課題とする.



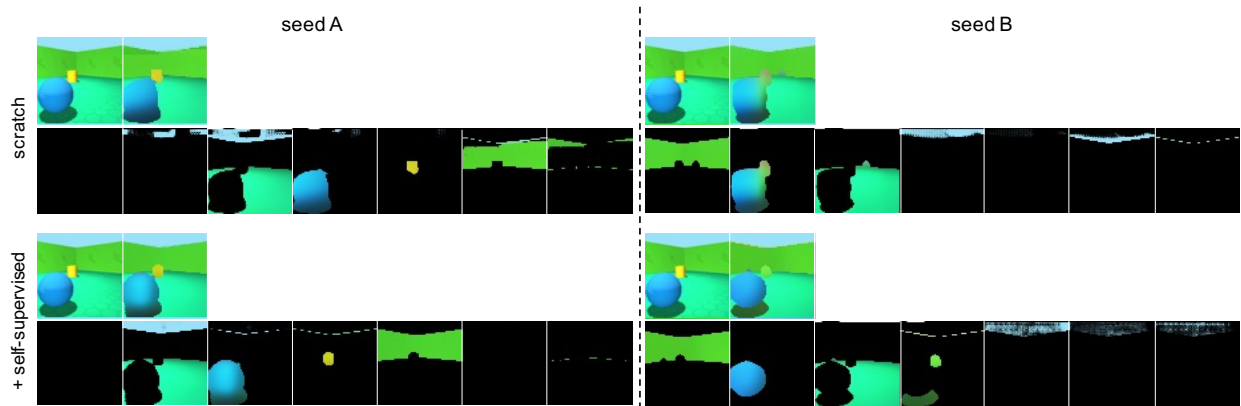


Fig.6.3 自己教師あり学習による学習の安定化を複数の乱数シード（学習時）について比較した結果．ここで利用したモデルは GENESIS+Tr である．上段が自己教師あり学習を用いない場合，下段が自己教師あり学習を用いた場合となっている．左右の列 (seed A, seed B) は異なる乱数シードに対応している．

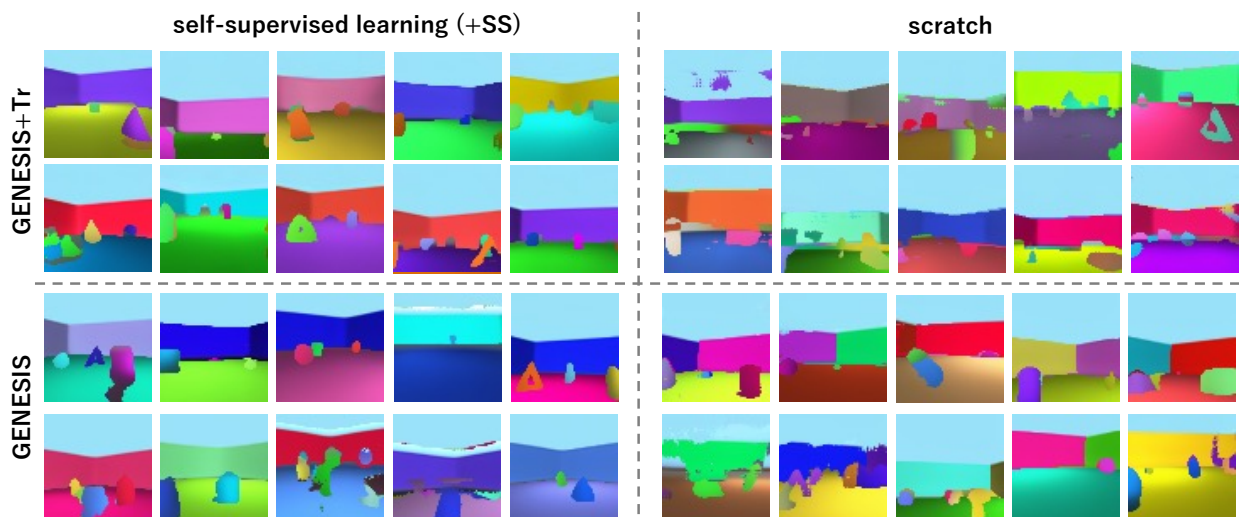


Fig.6.4 Objects Room データセットでの生成結果  
左列は自己教師あり学習の，右列は通常の学習の結果を示している．

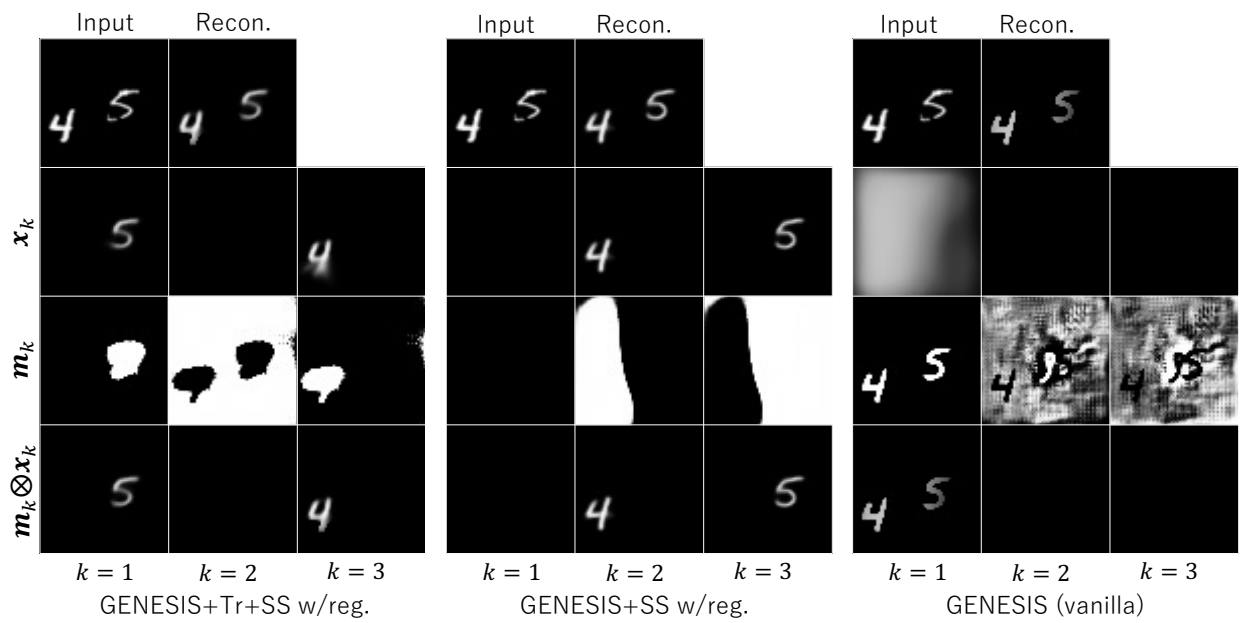


Fig.6.5 Multi-MNIST データセットでの推論結果. w/reg. は制約項の導入を意味する.



## 第 7 章

# 大域的な空間情報を持つ多視点シーン解釈モデルの提案

### 7.1 はじめに

本章では、複数の物体を含む三次元的な環境を想定し、その認識とモデル化について考える。5 章と 6 章では、静止画一枚を入力として物体の表現を獲得するシーン解釈の最も基本的な設定に取り組んできた。しかしロボットのような知的エージェントへの応用を考えた場合、環境中で行動することによって複数視点からの視覚的な情報は容易に得られる。また、簡単な自己位置推定であれば多くの状況で行うことができる。そこで、本章では複数視点からの画像と、その視線のベクトル情報を入力として与える設定に取り組む。モデルに与えられる課題としては、あるシーンについてのいくつかの視点からの画像から空間全体の情報を推論し、観測外のものも含む任意の視点からの画像の予測を行うこと (Novel View Synthesis) に加え、物体単位の表現を獲得し、任意の視点で物体の位置を検出 (セグメンテーション・バウンディングボックス等) することになる。

Novel View Synthesis については生成モデルで取り組んでいる Generative Query Network(GQN) と呼ばれる研究や [45, 103, 104], Neural Implicit Representation[105, 106] によって取り組んでいる研究がある。前者の研究は環境のモデルを構築し、任意の視点からの画像を予測することである。また、後者は 3D モデルをニューラルネットワークによって近似することが目的であり、ボリュームレンダリングや符号付き距離場 (SDF) などコンピュータグラフィックスの技術を取り込んでいる。これらの手法においては、物体ごとの表現学習は行われない。2020 年頃から、GQN を基にシーン解釈手法を組み合わせ、物体ごとの表現も獲得できるようにした手法 [107, 108] が登場している。本研究ではこれらを多視点シーン解釈モ

デルと呼ぶことにする。

ここで、複数物体を含む3次元的な空間は、各物体についての情報と、それらをどう配置するかという空間的な情報の両者が揃うことで一意に定まる。しかし、既存の多視点シーン解釈モデルは物体単位の表現しかモデル化していない。これは環境のモデル化が不完全であることに相当し、前後に隠れた物体の位置関係を予測することや、新たなシーンの生成を行う上で悪影響になっていると考えられる。というのも環境のモデル化が適切に行われていなければ、部分的な観測から物体の前後関係や隠れている部分の推論を空間についての前提知識なしで行わなければならない、生成についても物体の適切な配置が行えなくなってしまうためである。

空間全体についての情報は、物体を表す潜在変数に加え、追加の潜在変数を導入することでモデル化することが可能である。1つの方法として、物体の表現と独立な別の潜在変数を用意することが考えられる。また、もう一つは大域的な潜在変数を追加し、階層的な確率モデルを構築する方法である。しかし、前者の方法は物体の表現と空間的な配置の情報を各潜在変数が分担して表現する保証がなく、学習の工夫によってそれを実現しなければならない (disentanglement)。これは一般に簡単ではなく、シーン解釈のために物体単位の表現を学習することと同時にを行うのは更に困難であると考えられる。後者の方法では各物体の情報と空間的な配置の情報が完全には分離されず、大域的な潜在変数に物体に関する情報も混じってしまう懸念があるが、前者に比べて学習が容易である。また、別の観点として認知科学の領域では、人間は全体と部分を分けて認識しているという仮説が存在する [109]。特徴統合理論 [110] においては視覚的な特徴マップは、物体の視覚的な位置情報と組み合わせられて “master map of location” を構成するというモデルが提案されている。そのため、空間的な情報と物体の特徴を階層的にモデル化することは、人間の認知を考慮してもより自然な定式化であると考えられる。

そこで本研究では、物体ごとの表現の上にシーン全体の空間的な情報を表す大域的な潜在変数を仮定したモデルを構築する。大域的な潜在変数を導入すること自体は様々な分野の研究で既に行われている [111, 112]。特に Generative Neurosymbolic Machines (GNM) [30] では新たなデータの生成のために大域的な潜在変数を導入しているが、本研究で提案している階層的な確率モデルを持った多視点シーン解釈モデルについて、安定した学習や推論を実現する方法は自明ではなく、本研究の目的はこれを実現するモデル構造や学習方法を示すことである。提案手法はシーン全体の空間的な配置に関する情報と、物体ごとの個別の情報をモデル化したものであり、これを Whole-part Representation Learning Model for Object-centric Scene Inference and Sampling (**WeLIS**) と呼ぶことにする。

Model	Object-Centric	Novel View Synthesis	Inference	Sampling (Novel Scene Generation)
GQN	✗	✓	✓	✓ [104]
ObSuRF, uORF <sup>a</sup>	✓	✓	✓	✗
ROOTS	✓	✓	✓	✗
MuMON	✓	✓	✓	✗ <sup>b</sup>
WeLIS (Ours)	✓	✓	✓	✓

Table 7.1 multi-object-multi-view (MOMV) の問題設定に関連する手法の比較.

<sup>a</sup> これらの手法は生成モデルではない.

<sup>b</sup> 物体を個々に生成することは可能だが、位置関係が考慮されず、シーンとしての一貫性が保証されない.

本章の貢献は以下の通りである.

- 大域的な潜在変数によって空間全体の配置を表現する、多視点シーン解釈を構築した.
- 上記のモデルにおいて、安定的な学習と推論を行うための、いくつかの重要なモジュールを提案した.
- 空間的な構成についての表現学習を行うことが novel view synthesis と、それに対応するセグメンテーションの性能を向上させることを実験によって示し、さらに訓練データに含まれない新たなシーンの生成 (ランダム生成) が可能となることを示した.

## 7.2 関連研究

### 7.2.1 多視点物体中心表現学習 (Multi-view Object-centric Representation Learning)

近年、いくつかの研究が多視点のシーン解釈に取り組んでいる [107, 108]. これらの手法は 3 次元空間をいくつかの視点から観測し、それを基に任意の視点からの画像を予測することに加え、物体ごとの表現を獲得することを行っている. シーン解釈手法として、特にこれらは物体ごとの表現を獲得した上でセグメンテーションやバウンディングボックスによって物体の位置情報も獲得している. これらの手法は Generative Query Network (GQN)[45, 113, 104] の物

体中心表現学習への拡張であると解釈することもできる。MulMON[107]において、静止画を対象とする物体中心表現学習の手法は multi-object-single-view (MOSV)、ここで述べた複数視点を扱う手法は multi-view setting is defined as multi-object-multi-view (MOMV) と定義されており、本研究でも必要に応じてこの分類を用いる。

表 7.1 は MOMV に関連する手法を比較したものである。ここで、Novel View Synthesis は任意の視点の画像を予測することで、Novel Scene Generation は訓練データに含まれない新たなシーンを生成することを意味する。これらの手法の中で、ROOTS[108] と MulMON のみが生成モデルで、かつ MOMV の問題設定に取り組む手法となっている。また、O3V [114] も類似の課題に取り組んでいるが、入力の見点数が固定されている。uORF[36] と ObSuRF[37] は MOMV に取り組んでいるが、これらは生成モデルではなく、Slot Attention[35] と NeRF[106] に基づいた手法となっている。そのため、これらの手法は出力される画像の品質は高いが、新たなデータを生成することや、Variational Autoencoder (VAE) [16] のように潜在表現として低次元の特徴量を獲得することはできない。

本研究では、大域的な潜在変数を持ち、それによってシーン全体の空間的な情報を表現する MOMV の手法を構築する。提案手法は新たなシーンの生成が可能な唯一の手法である。ここでは以下の理由から MulMON をベースの手法として設定する。1つは、ROOTS がバウンディングボックスに基づく手法であるのに対し、MulMON はセグメンテーションに基づく手法となっているためである。Attend, Infer, Repeat[27] をはじめとするバウンディングボックスの手法は物体の位置情報や属性を明示的に異なる潜在変数として獲得することが可能なものの、物体の大きさや形状の変化に繊細であるという弱点がある [115]。次に、MulMON が ROOTS よりも必要な観測数の面で優位であることが挙げられる。本研究では MulMON に基づいて手法を構築しているが、類似の手法が ROOTS や他の手法にも適用できると考えている。

## 7.2.2 Generative Query Network(GQN) の関連研究

GQN は先に触れた通り、3次元空間をいくつかの視点から観測し、それを基に任意の視点からの画像を予測するというものである。これはつまり部分的な観測から3次元空間を把握することである。関連する技術としては Visual SLAM[116] や Structure from Motion(SFM)[?] などがあり、これらは深層学習の登場以前から存在していた技術であり、GQN とは目的も機能的にも異なる部分がある。最も大きな違いは GQN が潜在表現として低次元に圧縮された

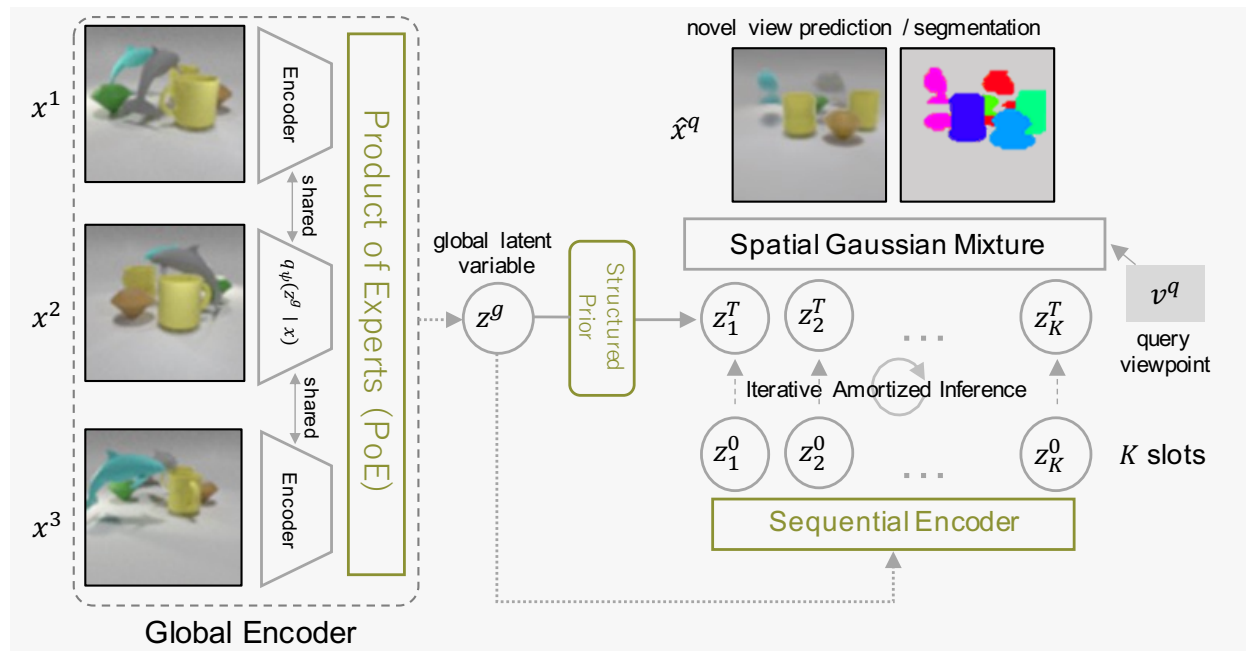


Fig.7.1 提案手法 WeLIS の概要. このモデルではいくつかの視点 ( $\mathbf{x}^1$ - $\mathbf{x}^3$ ) を観測し, 任意の指定された方向  $\mathbf{v}^q$  からの画像を予測し, セグメンテーションを行うものである.  $\mathbf{z}^g$  は空間的な情報をモデル化する大域的な潜在変数で,  $\mathbf{z}_k$  は物体ごとの表現となっている. 図示した画像は実際に提案手法の入出力として得られた結果である.

空間の表現を獲得した上で3次元空間の再構成を行うことが目的であるのに対し, SLAM や SFM は表現の獲得は行わず, 立体形状の正確な把握に重点が置かれている. GQN は基本的に3次元幾何に関する仮定(帰納バイアス)をほとんど利用しておらず, 簡単なトイデータでしか機能しないのに対し, SLAM や SFM は三次元幾何やカメラに関する知識を用いて実世界で利用されている. また, GQN については入力に基づくカメラポジションの推定などは行わず, 指定された視点の映像を予測するものであるという差異もある. 現在, 実用的には SLAM のような手法が広く用いられている. しかし GQN やその拡張である MOMV のモデルのような表現学習も行う手法が実世界で利用できるようなになれば, 入力のシーンや物体ごとの表現が獲得されるため, 情報処理のパイプラインを繋がなくとも意味的な状況認識や物体の識別まで一貫して行うことが期待できる.

## 7.3 手法

改めて整理すると, 本研究の目的は大域的な潜在変数を持つ MOMV モデルを構築し, 空間的な配置に関する情報もモデル化することで, より良い novel view synthesis やセグメンテー

ションの精度を得ることや、新たなシーンの生成を可能にすることである。大域的な潜在変数は物体の潜在変数と合わせて階層的な確率モデルとして実現する。本研究では、このモデルにおいて推論や学習を安定して行うために必要ないくつかの機構を提案する。手法の概要を図7.3に示した。

本節では、はじめに問題設定について述べ、その次に提案手法の確率モデルについて説明する。そして最後に学習方法と、導入した機構の詳細について説明する。

### 7.3.1 問題設定

本研究で対象とするのは複数の物体を含む3次元的な空間であり、データセットとしては  $\mathcal{D} = \{\mathbf{X}_1, \dots, \mathbf{X}_N\}$  と表せる。ただし  $N$  は含まれるシーンの数で、各シーンは  $\mathbf{X}_i = \{(\mathbf{x}_i^1, \mathbf{v}_i^1), \dots, (\mathbf{x}_i^M, \mathbf{v}_i^M)\}$  と、 $M$  個の異なる視点からの画像と視点ベクトルを含むものとなっている。 $\mathbf{v}$  は視線の方向を示すベクトルであり、基本的には3次元である。

課題はあるシーンについていくつかの観測  $\{(\mathbf{x}_i^1, \mathbf{v}_i^1), \dots, (\mathbf{x}_i^{N_{obs}}, \mathbf{v}_i^{N_{obs}})\}$  から任意の指定された視点(クエリ)の画像を予測し、同時に各物体に対応する  $K$  個の表現  $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_K\}$  を獲得することである。実際の学習時には、データセットに含まれる  $M$  個の視点を  $N_{obs}$  個の観測用と  $N_{qry}$  個のクエリ用に分割して利用することになる。汎化性能を向上させる目的で、この分割の割合はシーンごとに変更されることが多い。

### 7.3.2 確率モデル

#### 生成モデル

先に述べた通り、本研究では階層的な確率モデルにより、物体の表現  $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_K\}$  に加え、シーン全体の空間的な情報を表現するための大域的な潜在変数  $\mathbf{z}^g$  を導入することを考える。以降の節では物体の潜在変数は必要に応じて  $\mathbf{z}$  と略記する。この確率モデルの周辺尤度は、以下のようなになる

$$p_\gamma(\mathbf{x} | \mathbf{v}) = \int \int p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{v}) p_\phi(\mathbf{z} | \mathbf{z}^g) p(\mathbf{z}^g) d\mathbf{z} d\mathbf{z}^g, \quad (7.1)$$

ここで、 $\gamma = \{\theta, \phi\}$  である。右辺第一項の生成モデル  $p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{v})$  は空間的混合ガウスモデル (Spatial Gaussian Mixture model) と呼ばれるもので、シーン解釈の先行研究で広く利用されているものである [107, 24]。

$$p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{v}) = \prod_{i=1}^D \sum_{k=1}^K p_\theta(C_i = k | \mathbf{z}_k, \mathbf{v}) \otimes p_\theta(x_{ik} | \mathbf{z}_k, \mathbf{v}), \quad (7.2)$$

ここで  $D \in \mathbb{N}$  は画像のピクセル数である． $K$  は混合ガウス分布の混合数で，しばしば「スロット数」とも呼ばれ，物体の個数の上限に相当する．デコーダは broadcast decoder で実装されており，パラメータ  $\theta$  を持つ [73]．はじめに  $\mathbf{z}_k$  と  $\mathbf{v}$  が結合され，多層パーセプトロン (MLP) によって単一のベクトル  $\mathbf{f}_k$  となったあと，デコーダに入力される．つまりデコーダは  $K$  個の構成要素  $\{\mathbf{x}_1, \dots, \mathbf{x}_K\}$  とそれに対応するセグメンテーションを  $\{\mathbf{f}_1, \dots, \mathbf{f}_K\}$  から生成することになる．

式 7.1 の右辺第二項  $p_\phi(\mathbf{z} | \mathbf{z}^g)$  は学習可能な事前分布で，これを Structured Prior と呼ぶことにする．そして，第三項は正規分布が仮定されている．Structured Prior については節 7.3.2 にて詳述する．

### 推論モデル

本研究では Amortized Variational Inference [16] を用いて，近似変分事後分布を計算することで，計算が困難な式 7.1 の周辺尤度を扱う．ここでは変分事後分布，つまり推論モデルを以下のように分解する．

$$q_\psi(\mathbf{z}, \mathbf{z}^g | \mathbf{X}) = q_\psi(\mathbf{z}_1, \dots, \mathbf{z}_K | \mathbf{z}^g) q_\psi(\mathbf{z}^g | \mathbf{X}), \quad (7.3)$$

ここで  $\mathbf{X} = \{(\mathbf{x}_1, \mathbf{v}_1), \dots, (\mathbf{x}_{N_{obs}}, \mathbf{v}_{N_{obs}})\}$  を意味する．右辺第一項と第二項はそれぞれ  $\mathbf{z}$  と  $\mathbf{z}^g$  の近似事後分布である．これらのエンコーダ (推論モデル) をそれぞれ Sequential Encoder と Global Encoder と呼ぶことにする．

Global Encoder は可変個  $N_{obs}$  の観測視点を扱うため PoE[117] を用いた．ここでは特に，Prior Expert と呼ばれる事前分布を導入したものを採用した [118]．これに加え，Global Encoder には Normalizing Flow (NF)[46] を導入した．これはモデルの構築において必須ではないが，複雑な 3 次元空間の空間的な配置を表現するためには十分な事後分布の表現力が必要であり，学習の安定性や精度に貢献する．これらを踏まえると，近似事後分布は以下のようになる．

$$\mathbf{z}^g \sim q_\psi(\mathbf{z}^g | \mathbf{X}) = \left( p(\mathbf{z}_1^g) \prod_{v=1}^{N_{obs}} q_\psi(\mathbf{z}_1^g | \mathbf{x}_v, \mathbf{v}_v) \right) \prod_{t=1}^T \left| \det \frac{\partial f_\psi}{\partial \mathbf{z}_t^g} \right|^{-1} \quad (7.4)$$

ここで  $T$  は NF の変換数である．本研究では手法の簡潔さから Planar Flow を選択したが，他の後続の手法を用いることも可能で，探索の余地がある．上記の変換  $f_\psi$  は以下ようになる．

$$f_\psi(\mathbf{z}_{i+1}) = \mathbf{z}_i + \mathbf{u}_i \cdot h(\mathbf{w}^T \mathbf{z}_i + \mathbf{b}_i).$$

Sequential Encoder はここでは LSTM[60] を用いて自己回帰モデルとして実装した．他にもいくつかの構造を試したが，検証した限りでは現在の実装以外は安定した学習が困難であった．このエンコーダでは，まず自己回帰部分が  $K$  個の潜在変数  $\mathbf{z}_1^0, \dots, \mathbf{z}_K^0$  を推論し，それらが Iterative Amortized Inference (IAI)[55] によって更新される．これも MulMON や IODINE[25] で採用されている方法を踏襲したものである．ただし，提案手法では EfficientMORL[56] にならない，更新ネットワーク (Refinement Network) への画像サイズの入力を省略した．Sequential Encoder は以下のように定義される．

$$\mathbf{z}_1^0, \dots, \mathbf{z}_K^0 \sim \prod_{k=1}^K q_\psi(\mathbf{z}_k^0 | \mathbf{z}_{1:k}^0, \mathbf{z}^s), \quad (7.5)$$

$$\mathbf{z}_k^{i+1} = \mathbf{z}_k^i + f_\psi(\mathbf{z}_k^i, \mathbf{z}^s, \nabla_{\mathbf{z}_k^i} \mathcal{L}_{IAI}^i), \quad (7.6)$$

ここで， $\nabla_{\mathbf{z}_k^i} \mathcal{L}_{IAI}^i$  はそのイテレーションにおける負の対数尤度の勾配である．また， $f_\psi$  は MLP と LSTM で実装された更新ネットワークである．事後分布は任意の回数更新することが可能だが，本研究では常に  $L = 5$  とした． $\mathcal{L}_{IAI}^i$  の詳細は次節にて述べる．

## 学習

WeLIS の全パラメータは周辺対数尤度  $\log p(\mathbf{x}_1, \dots, \mathbf{x}_N)$  の Evidence Lower BOund(ELBO) の最大化によって end-to-end に学習を行うことが可能である．ELBO は以下のように表せる．

$$\begin{aligned} \mathcal{L}_{ELBO} &= E_{q_\psi(\mathbf{z}, \mathbf{z}^s | \mathbf{X})} \left[ \log \frac{p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{v}) p_\phi(\mathbf{z} | \mathbf{z}^s) p(\mathbf{z}^s)}{q_\psi(\mathbf{z} | \mathbf{z}^s) q_\psi(\mathbf{z}^s | \mathbf{X})} \right] \\ &= E_{q_\psi(\mathbf{z}, \mathbf{z}^s | \mathbf{X})} [\log p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{v})] \\ &\quad - KL[q_\psi(\mathbf{z}^s | \mathbf{X}) \parallel p(\mathbf{z}^s)] - E_{q_\psi(\mathbf{z}^s | \mathbf{X})} [KL[q_\psi(\mathbf{z} | \mathbf{z}^s) \parallel p_\phi(\mathbf{z} | \mathbf{z}^s)]] . \end{aligned} \quad (7.7)$$

ここで， $KL$  はカルバックライブラーダイバージェンスを意味する．そして， $KL$  項の影響を調整する係数  $\beta_1$  と  $\beta_2$  を導入し，目的関数は以下ようになる．

$$\begin{aligned} \mathcal{L} &= E_{q_\psi(\mathbf{z}, \mathbf{z}^s | \mathbf{X})} [\log p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{v})] \\ &\quad - \beta_1 KL[q_\psi(\mathbf{z}^s | \mathbf{X}) \parallel p(\mathbf{z}^s)] - \beta_2 E_{q_\psi(\mathbf{z}^s | \mathbf{X})} [KL[q_\psi(\mathbf{z} | \mathbf{z}^s) \parallel p_\phi(\mathbf{z} | \mathbf{z}^s)]] . \end{aligned} \quad (7.8)$$

さらに，IAI の反復的な処理により，対数尤度は以下のように計算される．

$$E_{q_\psi(\mathbf{z}, \mathbf{z}^s | \mathbf{X})} [\log p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{v})] = \frac{2}{L(L+1)} \sum_{i=0}^{L-1} (i+1) \mathcal{L}_{IAI}^i, \quad (7.9)$$

$$\mathcal{L}_{IAI}^i = \mathbb{E}_{\mathbf{z}^i \sim q_\psi(\mathbf{z}^i, \mathbf{z}^s | \mathbf{X})} [\log p_\theta(\mathbf{x} | \mathbf{z}^i, \mathbf{v})], \quad (7.10)$$



ここで、 $L$  は IAI のイテレーション数である．式 7.9 に示したように、IAI による反復的な更新において、後半の尤度がより重視されるように重みがかけている．これも MulMON と IODINE に従ったものである．なお、EfficientMORL は逆に更新の前半の尤度が重視されるように重みをかけているが、我々が実験した限りでは提案手法ではこの方法は学習に悪影響があった．EfficientMORL の場合は非常に表現力の高い高階層の推論モデルを用いているため、提案手法とは状況が異なるものと考えられる．

さらに、ここでは研究 2 と同様に GECO[98] を用いた．改めて GECO について述べると、これは ELBO を制約付き最適化問題と再解釈し、ラグランジュの未定係数法により対数尤度が目標値を満たすことを制約として KL 項を最適化するものである．GECO は必須ではないが、適切な KL 項の強度の調整は新たなシーンの生成をさらに改善することが期待される．

GECO によって再解釈された目的関数は以下ようになる

$$\operatorname{argmin}_{\psi, \phi} KL[q_{\psi}(\mathbf{z}^g | \mathbf{x}_{obs}, \mathbf{v}_{obs}) \parallel p(\mathbf{z}^g)] + E_{q_{\psi}(\mathbf{z}^g | \mathbf{x})} [KL[q_{\psi}(\mathbf{z} | \mathbf{z}^g) \parallel p_{\phi}(\mathbf{z} | \mathbf{z}^g)]] \quad (7.11)$$

$$\text{such that } -E_{q_{\psi}(\mathbf{z}, \mathbf{z}^g | \mathbf{x})} [\log p_{\theta}(\mathbf{x} | \mathbf{z}, \mathbf{v})] \leq R, \quad (7.12)$$

ここで  $R \in \mathbb{R}$  は KL 項の係数の代わりにハイパーパラメータとなる制約項の強度で、負の対数尤度の上限値である．WeLIS では全てのパラメータを end-to-end に学習することもできるが、これは Structured Prior のパラメータ  $\phi$  の過学習を起こす傾向があった．そこで我々は、Structured Prior の訓練をモデル本体と分離することを考えた．この方法の詳細は次の節で述べる．

### Structured Prior

本節では Structured Prior の詳細と、その学習方法について説明する．まず、Structured Prior は  $p_{\phi}(\mathbf{z}_1^L, \dots, \mathbf{z}_K^L | \mathbf{z}^g)$  と定義されるものであった．式 7.5 にて先に述べたように、各スロットの潜在変数は IAI によって反復的に更新されるが、これにより更新された事後分布  $\mathbf{z}_1^L, \dots, \mathbf{z}_K^L$  からのサンプリングは困難となる．IAI は潜在変数を再構成誤差やその他の入力から更新するが、これは推論時のみ可能であり、生成モデルは IAI の行った過程を入力情報なしに盲目的に実行する必要がある．

そこで我々はこの IAI の過程を Transformer[44] によってモデル化することを考えた．まず、初期の  $K$  個の変数  $\mathbf{z}_k^0$  を Sequential Encoder を通して  $\mathbf{z}^g$  から得る．そして、その  $K$  個の変数は Transformer に入力され、IAI 後の潜在変数  $\mathbf{z}_k^L$  を近似するように学習を行う．ここで、Sequential Encoder のパラメータは推論時と共有することとした．ただし Structured Prior の実装として必ずしも Transformer を利用する必要はなく、精度は落ちるが多層パーセプトロン

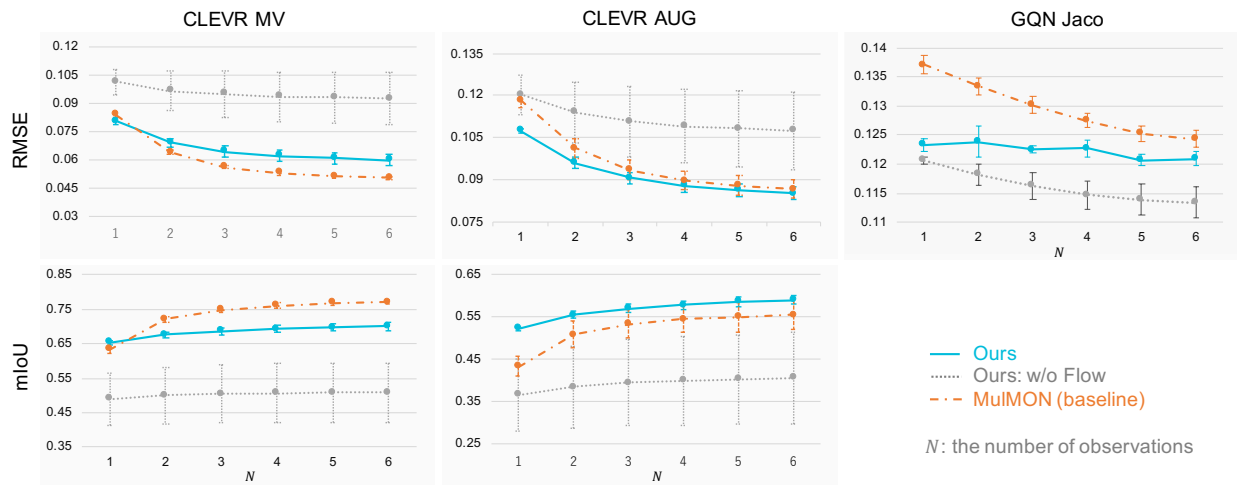


Fig.7.2 クエリに関する推論の定量評価. RMSE は小さな方が, mIoU は高い方が良いスコアである. GQN Jaco は正解のセグメンテーションが含まれていないため, mIoU が未評価となっている.

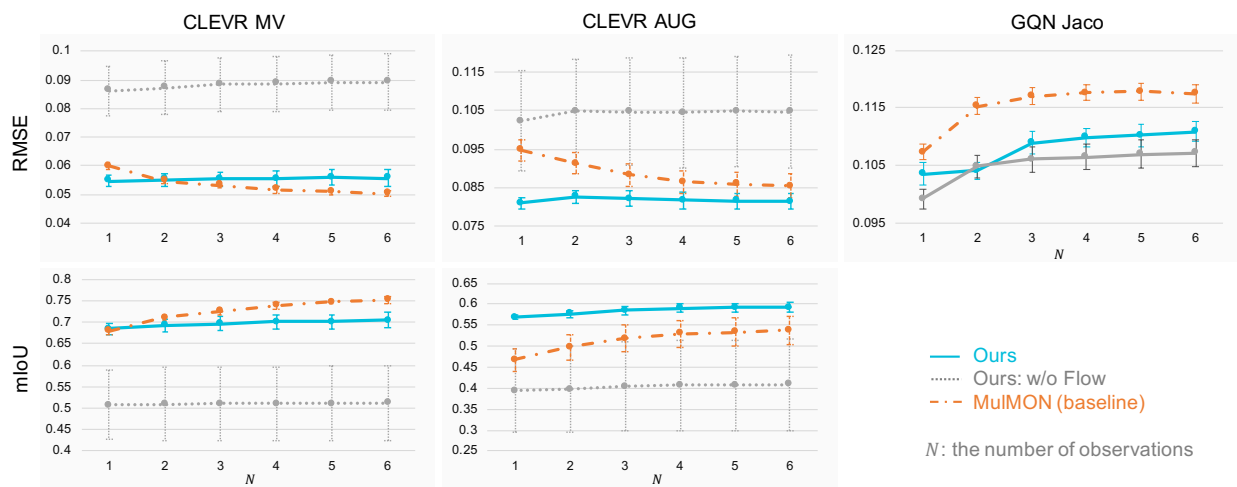


Fig.7.3 観測に対する定量評価. 入力画像に対する再構成とセグメンテーションの品質を評価したものとなっている.

(MLP) などを用いることも可能である. これについては 7.4.2 節にて定量評価を示す.

**Separate Training:** 先に述べた通り, 我々は Structured Prior のパラメータ  $\phi$  をモデル本体と分離して訓練することを提案する. 第1段階では WeLIS のモデル本体の学習時には Structured Prior のパラメータ以外を更新する. つまり,  $\phi$  は固定され, 式 7.7 の  $\beta_2$  は 0 に設定される. そして第2弾階として, 他の全てのパラメータを固定し  $\phi$  のみを更新する.

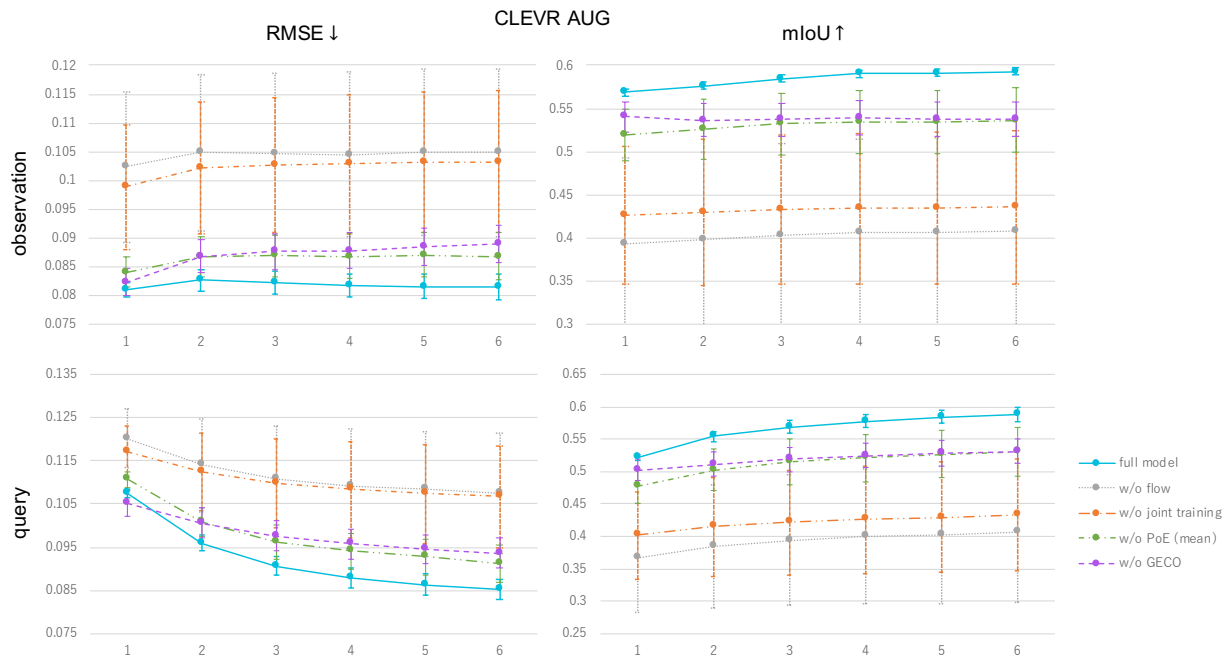
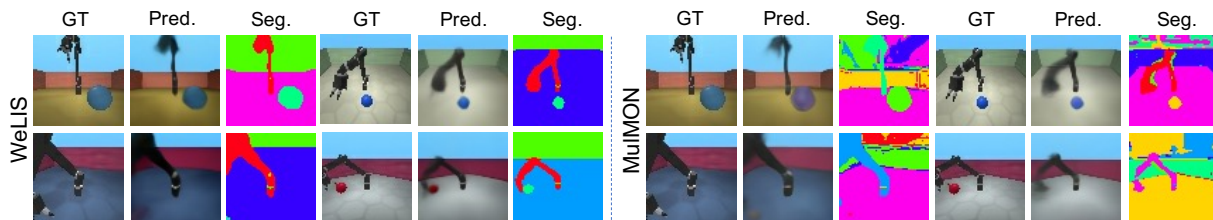


Fig.7.4 提案手法の各機構について除去実験を行った結果.

Fig.7.5 GQN Jaco における novel view synthesis の結果. これらの画像は未観測の視点の予測結果であり, 観測として与えた視点の数は  $N_{obs} = 3$  である.

## 7.4 実験

実験には3つのデータセットを利用した. これらはいずれも MulMON の研究にて利用しているものである. CLEVR Multi-View (CLEVR MV) は CLEVR データセット [119] の多視点への拡張版である. CLEVR Augmented (CLEVR AUG) は物体の種類と複雑さを CLEVR MV よりも上げ, 難易度を拡張したものとなっている. また, GQN Jaco は GQN で提案されたロボットアームのデータセットである [45]. いずれのデータセットも解像度は  $64 \times 64$  で利用し

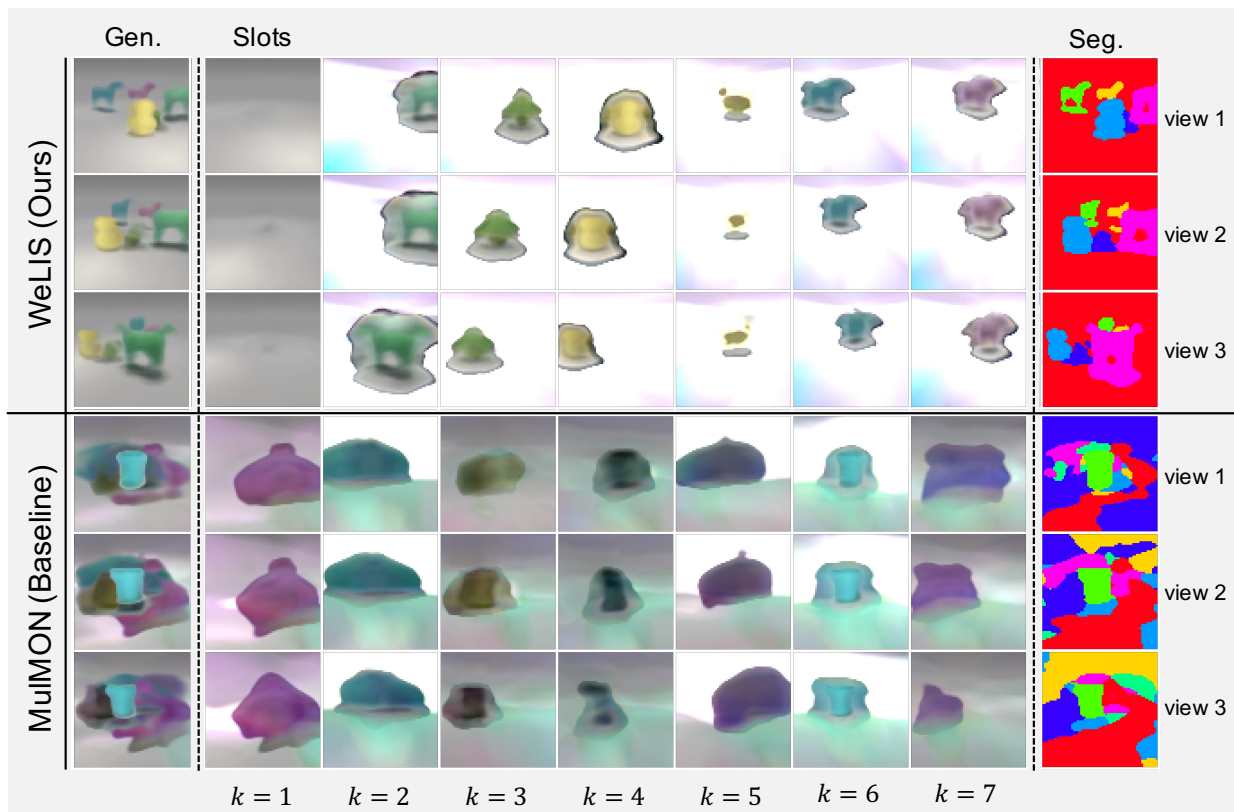


Fig.7.6 新たなシーンの生成を物体ごとに行った結果について，WeLIS とベースラインの比較結果．各行は異なるクエリの視点に相当する．最初の列は生成結果の画像を，そして次の7つの列は各スロット ( $\mathbf{x}_k, k \in \{1, \dots, K\}$ ) を，そして最後の列は生成されたセグメンテーションマスクを示している．

た<sup>\*1</sup> 学習には Adam[101] を利用し，30 万イテレーションの更新を行った．観測として与える視点の数は最大 6 とした．

本節では推論の性能について確認し，次に新たなシーンの生成能力について検証する．そして，各モジュールを除外した場合の検証 (ablation study) と，大域的な潜在変数がどのような表現を獲得しているかを確認するための実験を行った．

#### 7.4.1 Novel View Synthesis と Segmentation

本節では，novel view synthesis とセグメンテーションの品質について検証した．図 7.2 はクエリに関する各手法の定量評価の結果を，図 7.3 示している．novel view synthesis については

<sup>\*1</sup> MulMON は CLEVR AUG のみ  $128 \times 128$  で利用しているが，本研究では統一的な比較のために全て同じ解像度で利用した

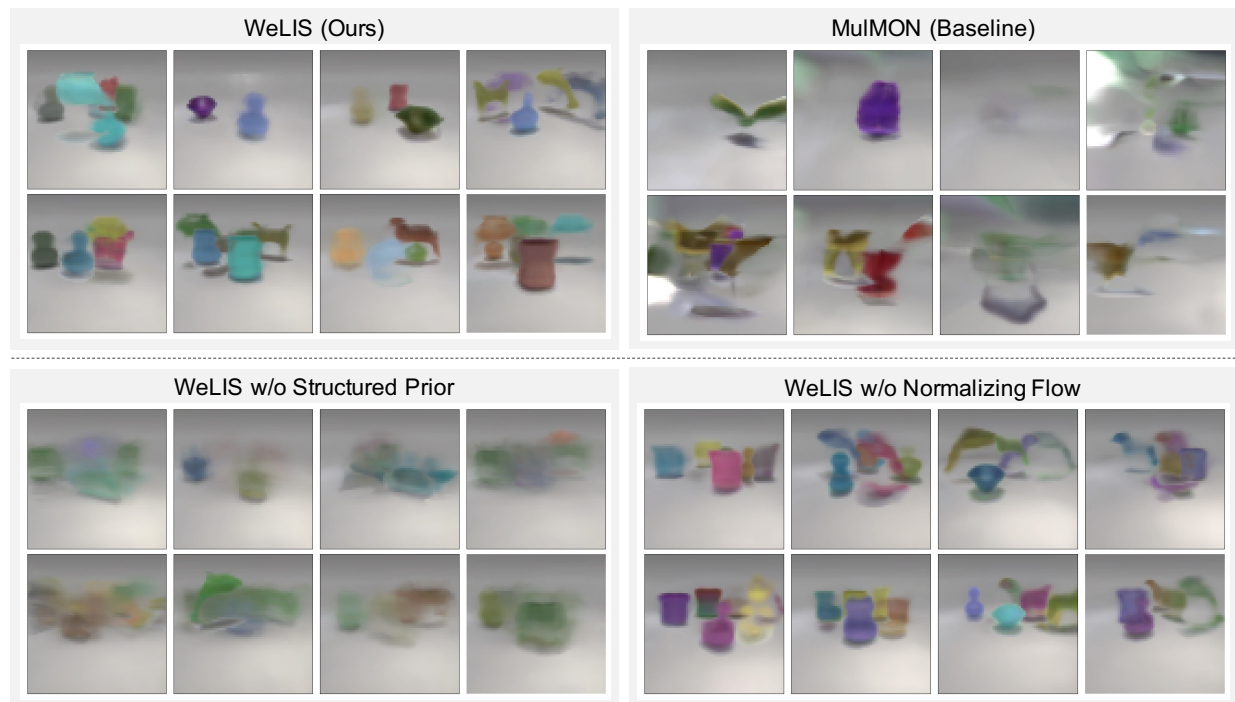


Fig.7.7 新たなシーンの生成結果. 上の行は WeLIS と MulMON の結果を示している. 下の行は WeLIS から Structured Prior を取り除いたものと, Normalizing Flow を取り除いたものを示している. これらの画像は, WeLIS については8個の異なる  $\mathbf{z}_g$  をサンプリングすることで生成し, MulMON については各スロットの潜在変数をそれぞれサンプリングした結果である. なお WeLIS の完全なモデル (左上) と Structured Prior を除去した結果 (左下) については, 同じ  $\mathbf{z}_g$  の組を用いている.

平均最小二乗誤差 (RMSE) で, セグメンテーションは mean Intersection over Union (mIoU) によって評価を行った. この結果は4つの異なる乱数シードを用いて学習を行い, それらの結果の平均と標準誤差を用いている. GQN Jaco については正解のセグメンテーションマスクが提供されていないため, mIoU の代わりに図 7.5 に定性的な結果を示した.

定量評価の結果を確認すると, WeLIS は CLEVR AUG と GQN Jaco, CLEVR MV の  $N$  が小さい場合でベースラインの MulMON に優位となっている. また, この結果から WeLIS は観測数  $N$  に関して比較的ロバストであることがわかる. これは WeLIS が限られた観測情報から全体を推測することが得意であることを意味している. これは大域的な潜在変数によって空間的な情報を扱い, 環境のモデル化が正確になったことにより, 隠れている物体の前後関係や見づらい物体の位置の推論が向上したためだと考えている. GQN Jaco の結果について確認すると, WeLIS ではセグメンテーションの品質や, ボールの認識の一貫性が改善している様子が分かる.

Datasets	WeLIS (Ours)	WeLIS w/o NF	MuLMON
CLEVR MV	125.8 $\pm$ 1.7	<b>117.2</b> $\pm$ 5.9	157.9 $\pm$ 8.4
CLEVR AUG	<b>83.3</b> $\pm$ 1.7	134.2 $\pm$ 25.8 <sup>*1</sup>	168.6 $\pm$ 5.0
GQN Jaco	<b>251.7</b> $\pm$ 2.1	262.1 $\pm$ 11.0	273.7 $\pm$ 10.4

<sup>a</sup> 学習が収束しなかった場合を除くと、スコアは  $90.0 \pm 1.3$  となる。

Table7.2 FID スコアの比較. WeLIS, WeLIS から Normalizing Flow を除去したもの, MuLMON について比較を行っている. なお, FID スコアは低いほど良い結果であることを示す.

また, この定量評価においては NF を除去した場合の結果についても検証を行っている (図 7.2 中, 灰色の破線). NF は推論や学習の安定化に重要であり, NF なしでは物体を適切に認識しない局所解に陥りがちであった. これは NF を用いない場合の標準誤差の大きさからも見て取れる. ただし, 成功率は下がるものの, 学習が適切に行われた場合は NF を用いた場合と遜色ない品質を実現できることも確認された.

さらに, 図 7.4 では提案手法で導入した各機構の除去実験を行い, 定量評価をまとめている. 凡例の上から, 提案手法, NF を除去したもの, Structured Prior の学習の分離を行わない場合, PoE の代わりに平均を取ったもの, GECO を用いなかった場合となっている. なお PoE の代わりに平均を取る処理は, GQN や NerfVAE で利用されているもので, 可変個の入力を扱うための基本的な手法である. この結果からは, それぞれの機構が精度の向上に寄与していることが確認できる.

### 定量指標の差と定性的な品質の関係について

図 7.8 は novel view synthesis について, 定量指標の差分がどの程度定性的な結果に影響を与えるのかを確認するために示したものである. また, WeLIS は背景を常に同じスロット (セグメンテーションの色に相当) で示していることが分かる. これは Sequential Encoder を自己回帰で実装したことによる影響で, 各スロットが permutation invariant でなくなったためだと考えられる. 背景が常に同じスロットで表現されるのは利点もあるが, disentanglement には悪影響があるものと考えられる. どちらが望ましい性質であるかは一概には言えず, 目的とする downstream タスクによって変わるものと考えられる.



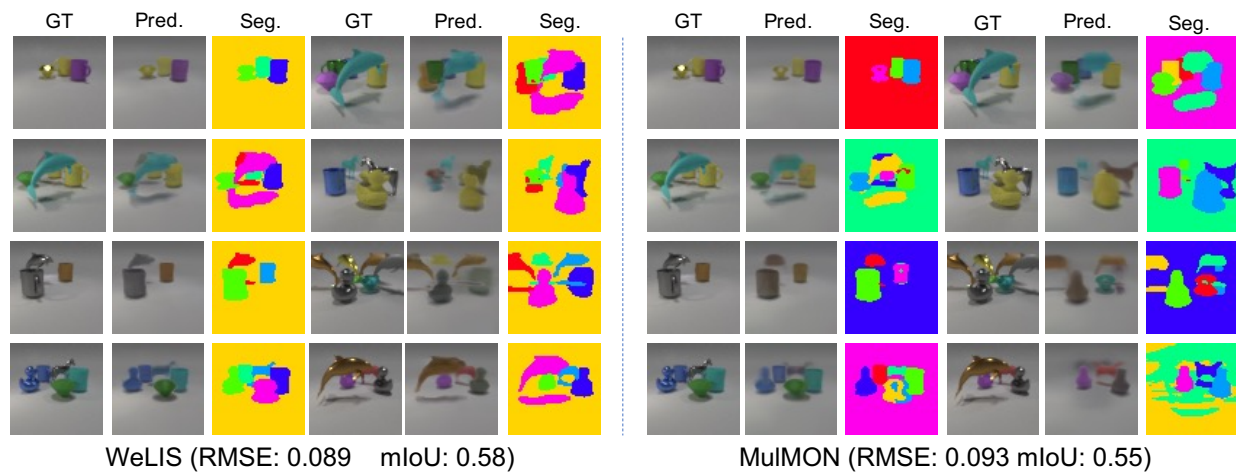


Fig.7.8 novel view synthesis の定性的結果と、その定量評価。観測数は3で、各グループの一行目は正解の画像 (GT) を、二列目は予測結果 (Pred.) を、三列目はセグメンテーション結果 (Seg.) を示している。

### 7.4.2 Novel Scene Generation

この節では、新たなシーンの生成結果についての評価を行う。図 7.6 に示した通り、WeLIS は各物体とセグメンテーションマスクを適切に生成できており、結果としてシーン全体の生成に成功していることがわかる。一方、MulMON は物体や配置が曖昧になってしまっており、シーン全体として適切に生成されていない。図 7.7 の上の行において複数の生成結果を示した。

#### 除去実験 (ablation studies)

図 7.7 の下の行は Structured Prior と NF を除去した場合の生成結果を示している。まず Structured Prior の結果について、これはフルモデル (図左上) と同じ  $\mathbf{z}_g$  の組から生成されている。Structured Prior なしで生成した結果は、ほぼ物体が識別できないほどにぼやけたものになっている。しかしフルモデルの場合と比較すると、物体の配置はほぼ一致していることが見て取れる。これは Structured Prior が貢献しているのは各物体の生成品質であり、空間的な配置によって物理的な一貫性を担保するのは大域的な潜在変数によるものと言える。

次に NF を除去した結果について確認する。生成結果はフルモデルとほぼ一致しており、NF 自体は生成品質にほぼ影響がないことが確認できる。ただし、上述したように学習の安定性に

	Transformer	MLP (w/o interaction)	MLP (w/ interaction)
FID	83.3 $\pm$ 1.7	84.35 $\pm$ 2.33	84.18 $\pm$ 2.61

Table 7.3 Structured Prior の実装として異なるネットワークアーキテクチャを用いた場合の FID. 小さい方が良いスコアである.

寄与するため、NF なしでは学習が失敗し、生成品質も低くなる場合が多々ある。そのため、実用的には生成についても NF が必要であると言える。

### 定量評価

生成品質について、Fréchet Inception Distance(FID) スコアによる定量評価を行った (表 7.4.1). WeLIS は MulMON を全てのデータセットで上回っていることが分かる。また、NF なしの場合でも、学習が成功しさえすればフルモデルとほぼ変わらない性能を発揮することが定量的にも確認できた。CLEVR AUG や GQN Jaco では生成結果が悪化しているが、これはデータセットが難しくなるにつれ、NF なしでの学習の成功確率が落ちるためである。

MulMON の生成品質が低いのは主に 2 つの理由による。1 つは環境のモデル化が不十分であること、もう一つは IAI によって更新された事後分布をサンプリングすることができないためである。WeLIS では前者は大域的な潜在変数を導入することで解決を試みており、後者については Structured Prior の導入によって解決を試みている。Structured Prior の有効性については既に図 7.7 の ablation study によって確認しており、大域的な潜在変数が獲得した表現の詳細については次の節 (7.4.3) にて確認する。

また、表 7.3 に Structured Prior の構造を変えた場合の FID を示した。Transformer がこれまで説明した標準のモデルで、MLP(w/o interaction) は MLP で、各スロットの更新を独立に行ったもの、MLP(w/ interaction) は Transformer と同様にスロット間の更新に相互作用がある場合、つまり更新に他のスロットの情報をを用いる場合、となっている。Transformer が最終的な精度は良かったが、MLP によるスロット間の相互作用の有無の比較はほとんど同じ結果となった。これは大域的な潜在変数が物体の配置や位置関係を十分にモデル化しており、単に各スロットを独立に更新すれば良いということを示している。





Fig.7.9 異なる4つの  $\mathbf{z}_g$  (A-D) から生成した結果と、その近傍  $\mathbf{z}_g + \epsilon$  からの生成結果。ここで  $\epsilon$  はランダムな摂動としてガウスノイズを利用している。各グループの最も左の列は元の  $\mathbf{z}_g$  から生成されたもので、それ以外の列は異なる摂動  $\epsilon$  を与えて生成した結果である。各グループ (A-D) が似た空間的な構成を保持していることが分かる。右側の潜在空間は概念図であり、実際の数値を反映したものではない。

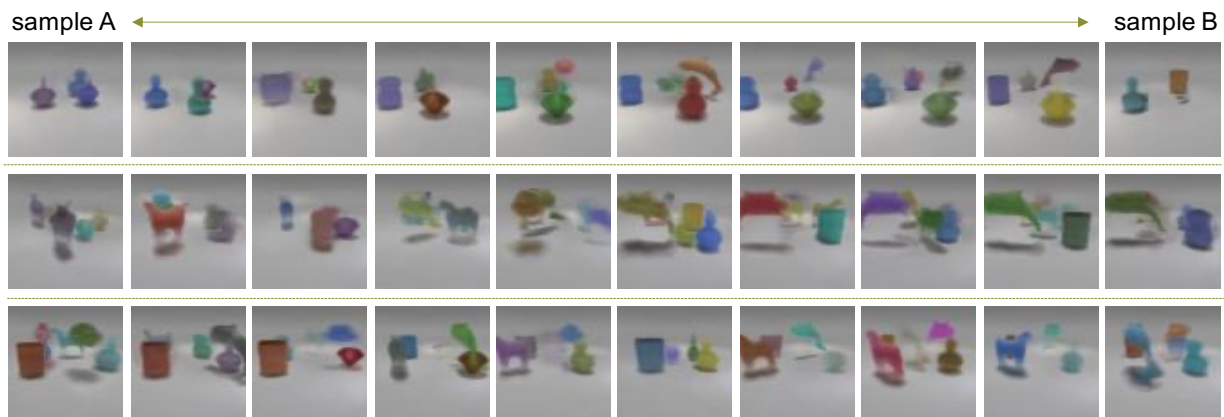


Fig.7.10 異なる2つの  $\mathbf{z}_g$  の間を直線で補間し、潜在変数空間を横断した場合の結果。各行は異なる補間の例であり、計3つ組の  $\mathbf{z}_g$  についての結果を示している。

### 7.4.3 大域的な潜在変数の表現について

大域的な潜在変数  $\mathbf{z}_g$  によって獲得された表現を確認するため、4つの異なる  $\mathbf{z}_g$  (A-D) から生成した結果と、その近傍  $\mathbf{z}_g + \epsilon$  からの生成結果を図 7.9 に示した。ここで、 $\epsilon$  は正規分布からサンプリングされた微小な摂動である。最初の列は元の  $\mathbf{z}_g$  から生成されたもので、それ以外の列は異なる摂動  $\epsilon$  を与えて生成した結果である。そのため、各グループは  $\mathbf{z}_g$  の周辺のクラスタからの生成結果となっている。

この結果から、同じ  $\mathbf{z}_g$  から生成された結果は、物体の種類は変化するものの、空間的な構成は類似していることが見て取れる。これは我々が期待した通り、 $\mathbf{z}_g$  が空間的な構成を獲得し

ていることを確認できる結果である。

また、図 7.10 には異なる 2 つの  $\mathbf{z}_g$  の間を直線で補間し、潜在変数空間を横断した場合の結果を示した。図には 3 組の例が示されており、いずれも物体の配置が連続的に変化するが、物体の形状や色、向きなどについては不連続に変化していることが見て取れる。これも図 7.9 と同様に  $\mathbf{z}_g$  が空間的な配置についての情報を獲得していることを確認できる結果となっている。

#### 7.4.4 Downstream Task について

本節では、獲得した表現を用いて何らかの課題を解く、downstream task について実験を行った。課題としては、シーンに含まれる物体の数を推測するものとなっている。結果を表 7.4 に示した。表現から結果を出す際には、線形分類器を用いた。

WeLIS は  $\mathbf{z}$  と  $\mathbf{z}^g$  の両方を用いることが可能であるため、これらを組み合わせた 3 パターン全てについて検証を行った。結果として、WeLIS で  $\mathbf{z}$  と  $\mathbf{z}^g$  の両方を用いた場合がほとんどの場合で最も高い精度となった。 $\mathbf{z}^g$  または  $\mathbf{z}$  のみからの分類では CLEVR MV の場合と CLEVR AUG の場合で優位なものが入れ替わっており、得られた表現の性質が両方で異なっている可能性が示唆される。CLEVR MV で観測を増やした場合については WeLIS は MulMON よりも低いスコアとなっていたことも関連がある可能性があり、今後より詳細な検証が必要である。

### 7.5 結論

本研究では新たな多視点シーン解釈モデルである WeLIS を提案した。提案手法は novel view synthesis を object-centric に実行することが可能であり、新たなシーンの生成も行うことができる。WeLIS は物体レベルの表現に加え、大域的な潜在変数が階層的な確率モデルとして導入されているが、このモデルにおいて安定した学習や推論を行うために必要ないくつかの機構を提案した。大域的な潜在変数は特に観測数が少ない場合に推論の精度を向上させ (図 7.2)、物理的な一貫性を保った新たなシーンの生成も可能とした (図 7.7)。また、いくつかの ablation study や獲得された表現を確認する実験を行った (図 7.9)。

今後の研究の方向性として、推論をより高度にすることが考えられる。本研究では Normalizing Flow や IAI を用いたが、Normalizing Flow については基本的な Planar Flow を用いている。最新の手法を用いることで精度が向上する可能性がある。また、IAI について提案手法では画像サイズの入力を用いなかった。これは計算量の削減に大きく貢献する一方で、視覚的に似通った物体やコントラストの低い物体の識別に問題が生じる可能性がある。この点について

Models	Datasets	Observations	Accuracy
MulMON	CLEVR MV	3	64.6
WeLIS( $\mathbf{z}^g$ )			73.0
WeLIS( $\mathbf{z}$ )			63.5
WeLIS( $\mathbf{z}^g + \mathbf{z}$ )			71.5
MulMON	CLEVR MV	4	75.0
WeLIS( $\mathbf{z}^g$ )			76.5
WeLIS( $\mathbf{z}$ )			53.5
WeLIS( $\mathbf{z}^g + \mathbf{z}$ )			76
MulMON	CLEVR MV	5	77.5
WeLIS( $\mathbf{z}^g$ )			76.0
WeLIS( $\mathbf{z}$ )			58.0
WeLIS( $\mathbf{z}^g + \mathbf{z}$ )			76.0
MulMON	CLEVR AUG	3	65.4
WeLIS( $\mathbf{z}^g$ )			63.9
WeLIS( $\mathbf{z}$ )			74.5
WeLIS( $\mathbf{z}^g + \mathbf{z}$ )			77.5
MulMON	CLEVR AUG	4	70.7
WeLIS( $\mathbf{z}^g$ )			68.8
WeLIS( $\mathbf{z}$ )			78.7
WeLIS( $\mathbf{z}^g + \mathbf{z}$ )			78.3
MulMON	CLEVR AUG	5	71.8
WeLIS( $\mathbf{z}^g$ )			69.3
WeLIS( $\mathbf{z}$ )			79.3
WeLIS( $\mathbf{z}^g + \mathbf{z}$ )			79.6

Table7.4 シーンに含まれる物体の個数を推定する downstream task の精度. Observations は観測として与えた視点の数を意味する.

の悪影響が CLEVR MV データセットでの結果に表れたことも考えられる. これらの点について最適な手法を探ることは今後の課題の 1 つとなる.

## 第 8 章

# 考察

1 章で詳しく述べた通り，本研究はシーン解釈手法について，学習の不安定性や適用範囲の限界といった現状の諸課題を解決することや，物体の認識に関する性能向上を主な目的としたものである．特に，こうした課題を解決するためにはデータの構造に関する何らかの前提知識 (帰納バイアス) が重要であると考え，これをモデルに導入する方法を各研究にて提案した．本章では各研究の内容と結果をまとめ，貢献を整理する．また，本研究で提案した手法の限界と，今後行うべき研究の方向性について述べる．

### 8.1 各研究の整理

5 章では，背景に関する知識を用いることで，物体を認識するための補助とすることを考え，補助情報として背景情報を用いる手法を提案した．提案手法では既存手法で適切に扱うことができなかった，複雑なテクスチャを含む画像に対する有効性が確認され，補助情報を導入することの意義が示された．その一方で，実画像を背景とするデータについては改良の余地が残る結果となった．また，補助情報として用いた背景のデータは訓練データに対応した背景である必要がなく収集コストは高くないが，既存の全てのデータセットに対して利用できるわけではないという制約もある．

6 章では，自己教師あり学習を用いて帰納バイアスを導入することによって，シーン解釈手法の安定性や性能，適用範囲等の向上を試みた．また，シーン解釈のための Transformer を用いたネットワーク構造を提案した．Transformer を用いたネットワーク構造は通常の end-to-end な学習では安定しなかったが，自己教師あり学習を用いることで改善され，性能を発揮することが確認できた．また，既存のシーン解釈の手法においても，自己教師あり学習の導入によって学習の安定化や性能の向上が確認できた．

7 章では、研究 1・研究 2 で取り組んでいる静止画を対象とした問題設定を多視点 (multi-view) に拡張したものに取り組んだ。これは、複数の物体を含む、ある 3 次元的な空間に関して複数視点からの画像を与え、観測していない任意の視点からの画像の予測と物体認識を行うという課題である。複数の物体を含む 3 次元的な空間を一意に定めるためには、各物体についての情報と、それらの空間的な配置に関する情報の両者があれば良いが、この問題設定に取り組む既存研究は物体の表現しかモデル化していなかった。そこで、この研究は階層的な確率モデルによって物体の配置を表現する大域的な潜在変数を導入した。結果として、各種精度や与えられる観測数に対する安定性の向上に加え、既存手法で難しかった新たなシーンの生成も可能となることを示した。しかし、どのようなデータセットに対して有効であるかという手法の適用範囲の問題については今後の検証が必要である。

## 8.2 提案手法の貢献と今後の課題

本研究はシーン解釈において物体認識に必要な補助情報を明示的に、もしくは暗黙的に与えることで、手法の適用範囲を拡大したり認識の精度や安定性を向上させることを試みたものであり、提案手法それぞれについて一定の有効性があることを確認した。しかしシーン解釈手法の理想形としては人間のように、未経験の物体を含む実世界の画像一般に対して物体認識とその表現が獲得可能となることであり、提案手法を含む既存手法はこの目標の達成には未だ遠いものとなっている。以下では本研究の結果を踏まえて現在の限界とその原因を改めて考察し、これを解決するための今後の研究の方向性について検討する。

### シーン解釈手法の課題の整理

現状のシーン解釈手法の課題について、最も大きなものは適用範囲の限界であり、その他に学習の安定性や精度の問題などが挙げられる。適用範囲について、現状では含まれる物体の種類が少なく、物体や背景についても形状や模様が単純であることや、周囲からの識別が容易で視覚的なコントラストが高いものでないと適切に学習が行われない状況となっている。研究 1 では背景の情報を利用することでこの制限が緩和できたが、実画像の複雑な背景を含む場合は完全には成功しなかった。研究 2 では自己教師あり学習の利用により安定性や精度の向上が確認できたが、適用範囲の拡大については限定的であった。適用範囲の限界は、物体の概念を獲得するための情報が十分に与えられていないこと、つまり現在与えられている課題を解くために必要な情報 (物体についての supervision) が不足しているためと考えられる。研究 1・研究 2

においてもこの点を考慮して改良を試みたもので、効果は確認できたが、教師あり学習の物体認識のように実世界のデータに対応するためにはさらなる研究の余地があると言える。

### どのような帰納バイアスが有効なのか？

研究1で与えた背景の情報は、どの部分が物体で、どこが背景なのかを知る手がかりになっていると考えられる。本研究において物体と背景とは、組み合わせ性を持つかどうかという点で区別できる。物体の領域は個々の構成要素（物体）の組み合わせによって様々なパターンが実現されるが、背景については組み合わせによる幾何級数的なパターンの増加は生じない。組み合わせとして認識する部分と、そうでない部分がわかれば、構成要素への分割は通常よりも簡単になると考えられる。研究2で用いた自己教師あり学習についても、対照学習の枠組みによって画像中で特徴的な領域に注目することになり、それは複数の物体を含むようなデータセットでは結果的に物体の領域になると考えられる。そのため、自己教師あり学習によって獲得した特徴量も結果的には研究1と同様に、物体とそれ以外の部分、つまり組み合わせ性の高い部分とそうでない部分を見分けることに役立つものになっていると考えられる。これはつまり物体の認識を阻害するもの（distractors）を見分けることであるということもできる。distractorsを見分けたり、除去するなどして視覚的な品質を向上させたり画像処理の精度向上を試みる研究は以前から行われている [120, 121, 122]。これらは基本的には人間が作成した教師データに基づいて行われていることや、画像の美観を損ねる対象を取り除くことが目的であることから、物体の認識とは観点が異なるものである。そのため本研究との直接的な関連は薄い、こうした技術は今後の発展において大いに参考になるものと考えられる。

研究3で導入した階層的な確率モデルと大域的な潜在変数については、これも多くの系が全体と部分（構成要素・物体）に分けて考えることができるという点で構成性を捉えるために重要な帰納バイアスの一つになると考えられる。本研究で行ったような物理的な系全体とその構成要素という階層性だけでなく、複雑な物体はそれ自体がさらに部分から構成され、階層性を持つことになる。階層性は重要な概念であると同時に、本研究で提案したモデルのように、どのような階層性が存在するのかを人間が考え設計するには限界があることも事実である。具体的なモデル設計にとどまらず、本質的にはデータから自動で階層性を獲得できるような補助を行う帰納バイアスが必要である。階層性を効率的に捉える方法の一つとしては記号系との接地が考えられ、この点についても記号系との融合を考えていくことが重要となる可能性がある。

### 今後の発展の方向性について

研究1や2のように、静止画のみを入力とする問題設定では「何を物体として認識すべきか」、言い換えれば「どの部分を組み合わせとして認識すれば良いのか」に関する基準をデータのみから獲得することが難しい。そのため、上記のような物体に注目する特徴量が得られるような誘導が有効であったと考えられる。しかし、これだけでは物体の領域、つまり組み合わせによって実現されていると考えられる領域をどのように分割すれば良いかについての帰納バイアスは十分でなく、これが提案手法を含めた既存手法の適用範囲が簡単な物体に限定されていることの原因の一つだと考えられる。

生成モデルを用いた既存のシーン解釈手法は Autoencoder[123] もしくは Variational Autoencoder (VAE)[16] を基に構築されており、本研究で提案している手法も VAE に基づいたものである。これらは入力画像を出力で再現すること (再構成) によって教師データなしで学習を行っているが、これは再構成を課題に使った自己教師あり学習と解釈できる。再構成を行うことは生成を行う上で本質的であるが、大まかな視覚的な特徴に重きを置き、小さな物体やコントラストが低いものは軽視する傾向がある。そのため再構成のみを用いることは、物体認識のような課題には有効でない場合がある。一方で予測誤差の最小化 (predictive coding) や「周囲の環境が予測しやすくなるように行動する」という自由エネルギー原理に基づく active inference[124] のような予測性の向上を目指す枠組みは生物学的なモデルとして古くから神経科学の分野で提案されている。これを踏まえると、再構成誤差最小化のように入力の再現を行うだけでなく、時間方向を考慮した予測の最小化につながるような課題が自己教師あり学習の pretext task として生物学的な妥当性が高いと言えるのではないだろうか。研究2では対照学習による自己教師あり学習を導入したものの、学習自体は VAE を基に再構成のみを用いており、この点について改良の余地があると考えている。なお学習方法については、生成モデルを用いずに対照学習を用いる方法も存在するが [49]、再構成を行わないために画像の生成ができないという違いがある。

このような背景を考慮し、今後は時間発展も考慮したモデルを扱い、より複雑な物体にも対応可能なように手法を一般化させていくことが考えられる。その際、本研究で扱った静止画を対象とした問題設定での知見を活用していくことが可能である。例えば対照学習を用いた自己教師あり学習は時間方向への対照学習を考えることもできるし、大域的な潜在変数や階層的な確率モデルは空間的に大域的なものに限らず、時間発展に対して大域的であるようなものも考えることができる。

この方向性の最終的な到達点としては実世界や、それに近い環境で学習を行うことになる。生物であれば生存という目的のために様々な自己教師あり学習の課題が与えられているような

ものであり、それによって必要な認識が構築されていると考えている。しかし現実世界で行動・学習することは情報処理としてもハードウェア的にも現状では難しい。そのため、実際には適度に複雑な仮想環境を人工的に構築することになるが、そのような環境の構築は技術的なコストが高い。既存研究でも動画や物体との相互作用によって学習を行う研究が存在するが、静止画の場合と同様に比較的簡単な環境での実証に留まっている [57, 125]。

そこで、研究の次のステップ一つとしては個々の物体の認識、つまり構成性の考慮が重要であるような課題をモデルに課していくことが考えられる。これは大枠としては実世界で行動した場合に必要となるであろう課題を個別に設計し、与えていくというものである。例えば1章で触れた relational inference や、物体やシーンの特徴に関する分類問題・質問応答のような課題が挙げられる。relational inference については記号や自然言語が関わることになり、獲得した物体の表現と記号系のグラウンディングについて考える題材にもなる。また、既存研究では物体中心表現学習によって獲得された表現の優れた評価方法が存在せず、再構成誤差や物体検出の精度で評価するしかない現状となっている。表現の良さとは本来は目的とする何らかの課題に対する有効性であり、このような課題は獲得された表現の良さを評価する指標として利用することも期待できる。加えて、既存研究は物体ごとの表現獲得という段階で終わっているが、獲得された物体の表現間の関係性やシーン全体としての意味について考えていくことも実用上重要だと考えられる。これは簡単な例としてはシーン全体の大域的な潜在変数を獲得することも含まれるし、各物体を考慮したシーングラフのようなものを構築することも挙げられる。

このような構成性の考慮が必要な課題を与える中で、徐々に環境を複雑にしていき、手法の一般化を試みるというのが今後のシーン解釈研究の方向性になると考えている。その中で、仮想環境の学習結果を実世界に転移させる sim2real のような技術 [126, 127] を用いて徐々に工場や室内などの制御された実環境での利用が可能になっていくことや、得られた知見を直接実環境での利用に活用することなどが期待できる。

### 8.3 構成的な認識技術の応用可能性について

本研究で提案したシーン解釈手法を含む、物体中心表現学習 (object-centric representation learning) の手法の将来的な応用可能性について考察する。ここでは大きく分けて2つの応用が考えられ、一つは単純に教師なし物体認識技術としての利用、もう1つは強化学習の認識部分として用いることである。近年は強化学習の技術の1つとして、制御やプランニングを確率



モデルの推論と捉える Control as Inference と呼ばれる枠組みが存在する [128]. 特に POMDP 環境を仮定する場合, この枠組みでは方策の最適化と環境モデル (状態の推論モデル: 認識モデル) の学習を分離することが可能で, SLAC や VRM[129, 130] では認識部分に VAE を基にした手法を利用している. この認識モデルにシーン解釈のような物体の表現を獲得可能な認識技術を用いることで, 複数の物体を含むような環境での行動を改良することが期待できる. また, ロボティクスの分野では物体を認識する技術は重要であり, キーポイントを抽出して用いたり, 教師あり学習による物体認識を活用した研究が存在している [32, 34]. シーン解釈が実画像にも適用可能となり, ロボットを用いるような実環境での認識ができれば, 物体に関するより情報量の豊かな表現を獲得可能なことや, 生成したデータによって訓練を行ういわば“イメージトレーニング”によるサンプル効率の向上も可能であり, こうした技術を置き換えることが期待される. 大規模な教師データの作成を行わずに様々な物体を認識することが可能となれば, 例えば産業用ロボットの認識を高度化し, 導入コストの低下と相まって自動化を促進することが期待できる.

もしくは, 先に述べた自由エネルギー原理, または active inference の枠組みに物体中心表現学習を組み込むことで, 生物的な物体認識のモデルとして利用し, 知能や認知機構についての研究につなげることも可能かもしれない.

## 8.4 人間の認知機構と物体認識技術の今後の方向性

本研究は主にシーン解釈手法の改良を試みたものであるが, より広い視点では「物体認識を題材として構成性についてのより深い理解を目指す取り組み」であると考えている. 物体を基本単位とする構成性は身近であるし, 視覚的な性質や物体に関する認知にも関わることから, 深層学習研究の枠に留まらない興味深い取り組みであると考えている. ところで, 構成性は物体に限ったものではなく, 構成性が強く表れる他の題材として言語や数学などの記号系 (シンボル) がある. 人工知能の発展は推論や知識表現といった記号的表現に基づくシンボリックなものから, 現在の学習ベースの手法へと研究の流行が置き換わり発展してきた. しかし記号的表現に基づく人工知能技術が今後不要であるわけではなく, 学習ベースの手法との得意不得意がある. そのため, 近年は深層学習をはじめとする学習ベースの人工知能にシンボリックな処理を取り込むことが今後の人工知能の発展に重要であるという主張が見られる [131]. こうした人工知能は neuro-symbolic AI と呼ばれることがある [132, 31]. また, Bengio は Daniel Kahneman の提案した system1 と system2 という用語を再解釈し, 直感的な処理を

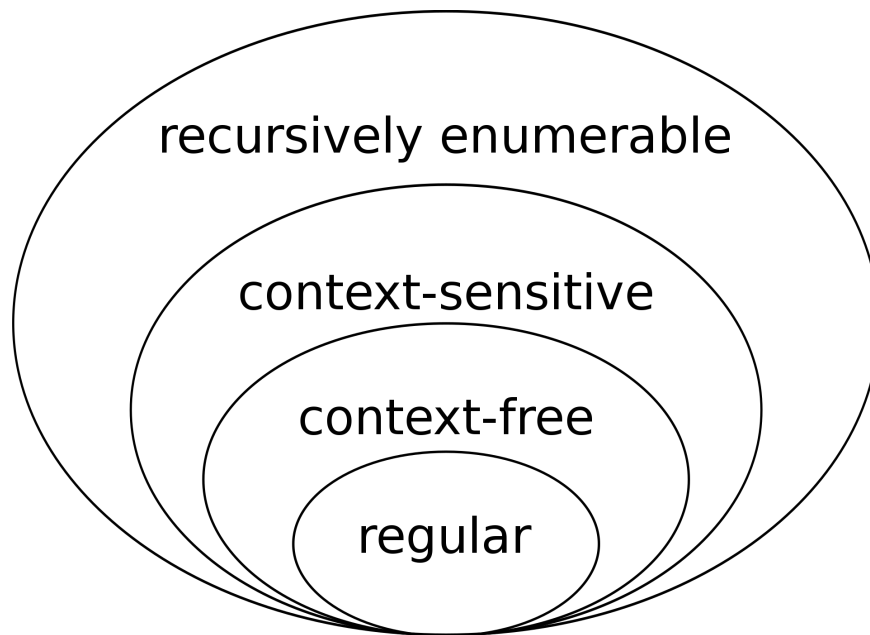


Fig.8.1 チョムスキー階層. regular が有限状態文法に, context-free が文脈自由文法に相当する. CC BY-SA 3.0 J. Finkelstein [https://en.wikipedia.org/wiki/Chomsky\\_hierarchy#/media/File:Chomsky-hierarchy.svg](https://en.wikipedia.org/wiki/Chomsky_hierarchy#/media/File:Chomsky-hierarchy.svg)

行う system1 に対し人間の持つような記号処理系のことを system2 と呼び, 現在の深層学習は system1 のみを実現しており, system2 の実現が今後重要だと主張している<sup>\*1</sup>.

言語などの記号体系が任意の対象を記述し, ひいては存在しない架空の存在や概念すらも表現可能なのは, 組み合わせによる膨大な表現力を持つためであり, これは構成性の性質に深く関係すると考えられる. 現在は膨大なデータを用いた統計的な言語モデル [8, 9] が高い性能を発揮しているが, 統計的な妥当性だけを考慮してこのような言語の表現力を獲得できるのか, つまり人間のような言語能力を獲得できるのかは未知数である. またこれらの手法は人間が必要とするよりも遥かに多くのテキストデータを学習に必要としており, 人間の言語処理とは少々異なるものであると思われる.

これまでシンボルの構成性を取り込むことが人工知能の発展に重要であるということを述べたが, そもそも我々が知る限り言語を扱うことが可能なのは人間のみである. 動物は種類によっては鳴き声によるコミュニケーションを取るが, これはあくまで固定されたパターンが複数あるのみで, 表現の組み合わせによって新たな表現ができるわけではない. 類人猿に人間の言語を模した構成性を持つ人工の記号体系を学習させることを試みても失敗するという研究結

<sup>\*1</sup> Deep Learning for System 2 Processing <http://www.iro.umontreal.ca/~bengioy/AAAI-9feb2020.pdf>  
2021 年 10 月アクセス.

果が存在する [133]. 一方で, 鳥類のジュウシマツは鳴き声 (歌) が組み合わせによって構成されることが知られている. しかしこれは順番で意味が変わったりはせず, 組み合わせによる表現のパターン増加は限定的である. これはチョムスキー階層 [134] の中で有限状態文法と区分されるものに相当する. 有限状態文法では, 組み合わせの最小単位が何か特定の意味を持つわけではなく, 順序に意味がない. 各構成要素を確率的に行き来するだけであり, 有限オートマトンで記述可能である. ジュウシマツについては単に鳴き声の複雑さで求愛の強さを表現しているものとされている [135]. チョムスキー階層においては, 人間の言語は有限状態文法よりも一般的な文脈自由文法, もしくはそれ以上でないと記述できないとされている.

このような背景から, 構成性を扱うことができるかどうかの差は人間と動物の知能を区別する境界になっていると言え, system1 と system2 の違いを決定づける鍵である可能性もある. そのため, 構成性の理解や人工知能への適用は汎用人工知能の実現や, 人間の知能や認知機構の理解に大きく貢献するものと考えている.

## 第 9 章

# 結論

本論文は主に帰納バイアスの導入という観点からシーン解釈手法の改善を試みた。

1 章では本研究の背景となる、構成性や物体を中心とした認識の必要性について述べた。2 章と 3 章では深層生成モデルや物体中心表現学習 (object-centric representation learning) の関連研究についてまとめ、4 章では本論文の目的や各章の位置付けを整理した。5 章から 7 章までは本論文を構成する 3 つの研究について述べており、それぞれシーン解釈手法の発展を目指したものである。

まず 5 章は、シーン解釈において物体を認識するための補助情報として背景の情報を活用することを提案した。既存のシーン解釈手法は複雑なテクスチャを含む画像や実画像を適切に扱うことができず、物体ごとの表現を獲得することができないという問題があった。この研究では学習時に背景の画像集合を利用することで、何を物体として認識すべきかという基準を暗黙的に与え、既存手法では扱えなかったデータセットに対しても有効となることを確認した。まず物体が 1 つの場合について検証を行い、複雑な背景があっても物体を認識できることを示した。そして、次に複数物体を含む場合について実験を行い、既存のシーン解釈手法を拡張する方法として提案手法が利用可能であることを確認した。

次に 6 章では、自己教師あり学習と Transformer を用いたシーン解釈手法を提案した。5 章の研究 1 では背景についての情報を新たに与えていたが、6 章では追加のデータを用いるのではなく、学習の工夫による改良を試みた。また、シーン解釈の課題には畳み込みニューラルネットワーク (CNN) の性質が望ましくないと考え、Transformer を用いた新たなネットワークアーキテクチャを提案した。自己教師あり学習は既存手法、Transformer を用いた提案モデルの両者において学習の安定性や精度に寄与し、特に後者は自己教師あり学習を利用することによって安定して学習可能となることを確認した。

そして7章では、複数視点の入力を扱う問題設定のシーン解釈手法として WeLIS を提案した。複数の物体を含む3次元的な空間においては、個々の物体の表現に加えてそれらをどのように配置するのかを定める必要があるが、後者は既存手法ではモデル化されていなかった。そこで提案手法では物体の潜在変数に加えて大域的な潜在変数を導入し、シーン全体の空間的な配置を表現することを試みた。これにより、多くの場合で既存手法の精度を上回り、加えて既存手法では難しかった新たなシーンの生成も行うことが可能となることを確認した。

8章では5章から7章までの各研究の結果と貢献を整理し、今後の研究の方向性や応用可能性について考察を行った。シーン解釈を始めとする、物体中心表現学習の手法は今後の人工知能の発展や産業応用、ひいては生物の認知機構の理解のような方向性にも関連する、重要な研究領域になっていくことが考えられる。本論文で得られた知見が今後の関連分野の研究を促進することを期待する。

## 謝辞

博士課程の研究において、私の研究活動に関わってくださった皆様に感謝を申し上げます。特に直接の指導を頂いた松尾豊先生と鈴木雅大先生には素晴らしい環境と、学術研究に留まらない広範な知見を与えて頂きました。博士課程の活動を経る前と後では、文字通り世界の見方が変わったと言っても過言ではありません。また、岩澤有祐先生は個別の研究会の運営や研究への助言を、山川宏先生にはコネクトームに関する研究グループに参加させて頂いたり、個別に相談させて頂くなど、様々な面でお世話になりました。加えて、副査の先生方には建設的な意見を頂き、本論文を改善する上での助けになりました。

松尾研究室で関わらせて頂いた学生・職員の皆様にも様々な面で助けられており、学術的な範囲に留まらず、皆様との交流が研究を行う上での精神的な支えになっていました。この場を借りてお礼させていただきます。

## 参考文献

- [1] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, and J. B. Tenenbaum, “Clevrer: Collision events for video representation and reasoning,” in *International Conference on Learning Representations*, 2019.
- [2] K. Greff, S. van Steenkiste, and J. Schmidhuber, “On the binding problem in artificial neural networks,” 2020.
- [3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [4] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, and Z. Žabokrtský, “Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals,” *Nature Communications*, vol. 11, no. 1, p. 4381, Sep 2020. [Online]. Available: <https://doi.org/10.1038/s41467-020-18073-9>
- [5] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, “Toward human parity in conversational speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.
- [6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in *Proc. CVPR*, 2020.
- [7] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, “Nüwa: Visual synthesis pre-training for neural visual world creation,” 2021.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp.

- 4171–4186. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423>
- [9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>
- [10] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8821–8831. [Online]. Available: <https://proceedings.mlr.press/v139/ramesh21a.html>
- [11] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, “Generalisation in humans and deep neural networks,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS’18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 7549–7561.
- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=YicbFdNTTy>
- [13] A. Goyal and Y. Bengio, “Inductive biases for deep learning of higher-level cognition,” 2021.
- [14] U. Hasson, S. A. Nastase, and A. Goldstein, “Direct fit to nature: An evolutionary perspective on biological and artificial neural networks,” *Neuron*, vol. 105, no. 3, pp. 416–434, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S089662731931044X>
- [15] Z. Chen, “Object-based attention: A tutorial review,” *Attention, Perception, & Psychophysics*, vol. 74, no. 5, pp. 784–802, 2012.



- 
- [16] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
  - [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
  - [18] P. Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*, ser. Cognitive science series. Harvard University Press, 1983. [Online]. Available: <https://books.google.co.jp/books?id=FS3zSKAfLGMC>
  - [19] M. Kawato, “Internal models for motor control and trajectory planning,” *Current Opinion in Neurobiology*, vol. 9, no. 6, pp. 718–727, 1999. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0959438899000288>
  - [20] D. Ha and J. Schmidhuber, “Recurrent world models facilitate policy evolution,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: <https://proceedings.neurips.cc/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf>
  - [21] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, “Dream to control: Learning behaviors by latent imagination,” in *International Conference on Learning Representations*, 2019.
  - [22] K. Greff, S. van Steenkiste, and J. Schmidhuber, “Neural expectation maximization,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/d2cd33e9c0236a8c2d8bd3fa91ad3acf-Paper.pdf>
  - [23] S. van Steenkiste, M. Chang, K. Greff, and J. Schmidhuber, “Relational neural expectation maximization: Unsupervised discovery of objects and their interactions,” in *International Conference on Learning Representations*, 2018.
  - [24] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, “Monet: Unsupervised scene decomposition and representation,” *arXiv preprint arXiv:1901.11390*, 2019.
  - [25] K. Greff, R. L. Kaufmann, R. Kabra, N. Watters, C. Burgess, D. Zoran, L. Matthey, M. Botvinick, and A. Lerchner, “Multi-object representation learning with iterative vari-

- ational inference,” *arXiv preprint arXiv:1903.00450*, 2019.
- [26] M. Engelcke, A. R. Kosiosek, O. P. Jones, and I. Posner, “Genesis: Generative scene inference and sampling with object-centric latent representations,” in *International Conference on Learning Representations*, 2020.
- [27] S. A. Eslami, N. Heess, T. Weber, Y. Tassa, D. Szepesvari, G. E. Hinton, *et al.*, “Attend, infer, repeat: Fast scene understanding with generative models,” in *Advances in Neural Information Processing Systems*, 2016, pp. 3225–3233.
- [28] E. Crawford and J. Pineau, “Spatially invariant unsupervised object detection with convolutional neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3412–3420.
- [29] Z. Lin, Y.-F. Wu, S. V. Peri, W. Sun, G. Singh, F. Deng, J. Jiang, and S. Ahn, “Space: Unsupervised object-oriented scene representation via spatial attention and decomposition,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=rkl03ySYDH>
- [30] J. Jiang and S. Ahn, “Generative neurosymbolic machines,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [31] D. Ding, F. Hill, A. Santoro, M. Reynolds, and M. Botvinick, “Attention over learned object embeddings enables complex visual reasoning,” in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. [Online]. Available: <https://openreview.net/forum?id=lHmhW2zmVN>
- [32] P. Florence, L. Manuelli, and R. Tedrake, “Self-supervised correspondence in visuomotor policy learning,” *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 492–499, 2019.
- [33] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee, “Transporter networks: Rearranging the visual world for robotic manipulation,” *Conference on Robot Learning (CoRL)*, 2020.
- [34] C. Devin, P. Abbeel, T. Darrell, and S. Levine, “Deep object-centric representations for generalizable robot learning,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7111–7118.
- [35] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, “Object-centric learning with slot attention,” *arXiv preprint arXiv:2006.15055*, 2020.

- 
- [36] H.-X. Yu, L. J. Guibas, and J. Wu, “Unsupervised discovery of object radiance fields,” 2021.
  - [37] K. Stelzner, K. Kersting, and A. R. Kosiorek, “Decomposing 3d scenes into objects via unsupervised volume segmentation,” *arXiv preprint arXiv:2104.01148*, 2021.
  - [38] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” in *ICLR*, 2017.
  - [39] H. Kim and A. Mnih, “Disentangling by factorising,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 2649–2658.
  - [40] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, “Towards a definition of disentangled representations,” *arXiv preprint arXiv:1812.02230*, 2018.
  - [41] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled representations,” in *international conference on machine learning*. PMLR, 2019, pp. 4114–4124.
  - [42] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=S1v4N2l0->
  - [43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
  - [44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
  - [45] S. A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, *et al.*, “Neural scene representation and rendering,” *Science*, vol. 360, no. 6394, pp. 1204–1210, 2018.
  - [46] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *Proceedings of the 32nd International Conference on Machine Learning*, ser.

- Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1530–1538. [Online]. Available: <http://proceedings.mlr.press/v37/rezende15.html>
- [47] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A Fast Learning Algorithm for Deep Belief Nets,” *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 07 2006. [Online]. Available: <https://doi.org/10.1162/neco.2006.18.7.1527>
- [48] Y. Du and I. Mordatch, “Implicit generation and modeling with energy based models,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/file/378a063b8fdb1db941e34f4bde584c7d-Paper.pdf>
- [49] T. Kipf, E. van der Pol, and M. Welling, “Contrastive learning of structured world models,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=H1gax6VtDB>
- [50] T. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, *Unsupervised Learning of Object Keypoints for Perception and Control*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [51] T. Nguyen-Phuoc, C. Richardt, L. Mai, Y.-L. Yang, and N. Mitra, “Blockgan: Learning 3d object-aware scene representations from unlabelled images,” in *Advances in Neural Information Processing Systems 33*, Nov 2020.
- [52] S. Ehrhardt, O. Groth, A. Monzpart, M. Engelcke, I. Posner, N. Mitra, and A. Vedaldi, “Relate: Physically plausible multi-object scene synthesis using structured latent spaces,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 11 202–11 213. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/806beafe154032a5b818e97b4420ad98-Paper.pdf>
- [53] E. Crawford and J. Pineau, “Spatially invariant unsupervised object detection with convolutional neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3412–3420.
- [54] J. Jiang, S. Janghorbani, G. De Melo, and S. Ahn, “Scalor: Generative world models with scalable object representations,” in *International Conference on Learning Representations*,

- 2019.
- [55] J. Marino, Y. Yue, and S. Mandt, “Iterative amortized inference,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 3403–3412.
  - [56] P. Emami, P. He, S. Ranka, and A. Rangarajan, “Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 2970–2981. [Online]. Available: <http://proceedings.mlr.press/v139/emami21a.html>
  - [57] R. Veerapaneni, J. D. Co-Reyes, M. Chang, M. Janner, C. Finn, J. Wu, J. B. Tenenbaum, and S. Levine, “Entity abstraction in visual model-based reinforcement learning,” in *CoRL*, 2019.
  - [58] K. Greff, A. Rasmus, M. Berglund, T. Hao, H. Valpola, and J. Schmidhuber, “Tagger: Deep unsupervised perceptual grouping,” in *Advances in Neural Information Processing Systems*, 2016, pp. 4484–4492.
  - [59] A. Goyal, A. Lamb, P. Gampa, P. Beaudoin, C. Blundell, S. Levine, Y. Bengio, and M. C. Mozer, “Factorizing declarative and procedural knowledge in structured, dynamical environments,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=VVdmjgu7pKM>
  - [60] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, Nov. 1997. [Online]. Available: <https://doi.org/10.1162/neco.1997.9.8.1735>
  - [61] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1724–1734. [Online]. Available: <https://aclanthology.org/D14-1179>
  - [62] N. Watters, L. Matthey, M. Bosnjak, C. P. Burgess, and A. Lerchner, “COBRA: data-efficient model-based RL through unsupervised object discovery and curiosity-driven exploration,” *CoRR*, vol. abs/1905.09275, 2019. [Online]. Available: <http://arxiv.org/abs/1905.09275>

- [63] J. C. Whittington, R. Kabra, L. Matthey, C. P. Burgess, and A. Lerchner, “Constellation: Learning relational abstractions over objects for compositional imagination,” *arXiv preprint arXiv:2107.11153*, 2021.
- [64] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, “U2-net: Going deeper with nested u-structure for salient object detection,” *Pattern Recognition*, vol. 106, p. 107404, Oct 2020. [Online]. Available: <http://dx.doi.org/10.1016/j.patcog.2020.107404>
- [65] D. Zhang, H. Tian, and J. Han, “Few-cost salient object detection with adversarial-paced learning,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 236–12 247. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/8fc687aa152e8199fe9e73304d407bca-Paper.pdf>
- [66] R. Kabra, C. Burgess, L. Matthey, R. L. Kaufman, K. Greff, M. Reynolds, and A. Lerchner, “Multi-object datasets,” <https://github.com/deepmind/multi-object-datasets/>, 2019.
- [67] H. Caselles-Dupré, M. G. Ortiz, and D. Filliat, “Symmetry-based disentangled representation learning requires interaction with environments,” in *Advances in Neural Information Processing Systems*, 2019, pp. 4608–4617.
- [68] A. Kosiorek, H. Kim, Y. W. Teh, and I. Posner, “Sequential attend, infer, repeat: Generative modelling of moving objects,” in *Advances in Neural Information Processing Systems*, 2018, pp. 8606–8616.
- [69] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [70] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [71] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [72] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [73] N. Watters, L. Matthey, C. P. Burgess, and A. Lerchner, “Spatial broadcast decoder: A sim-

- ple architecture for learning disentangled representations in vaes,” *The 2nd Learning from Limited Labeled Data (LLD) Workshop in International Conference on Learning Representations*, 2019.
- [74] O. Press, T. Galanti, S. Benaïm, and L. Wolf, “Emerging disentanglement in auto-encoder based unsupervised image content transfer,” in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=BylE1205Fm>
- [75] K. A. Severson, S. Ghosh, and K. Ng, “Unsupervised learning with contrastive latent variable models,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4862–4869.
- [76] A. Abid and J. Zou, “Contrastive variational autoencoder enhances salient features,” *arXiv preprint arXiv:1902.04601*, 2019.
- [77] M. Sugiyama, T. Suzuki, and T. Kanamori, “Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation,” *Annals of the Institute of Statistical Mathematics*, vol. 64, no. 5, pp. 1009–1044, 2012.
- [78] A. Krizhevsky, G. Hinton, *et al.*, “Learning multiple layers of features from tiny images,” 2009.
- [79] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [80] 鈴. 雅大, 金. 貴輝, 谷. 尚平, 松. 達也, and 松. 豊, “Pixyz: 複雑な深層生成モデル開発のためのフレームワーク,” *人工知能学会全国大会論文集*, vol. JSAI2019, pp. 1L2J1105–1L2J1105, 2019.
- [81] W. Luo, Y. Li, R. Urtasun, and R. Zemel, “Understanding the effective receptive field in deep convolutional neural networks,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 4905–4913.

- [82] A. Araujo, W. Norris, and J. Sim, “Computing receptive fields of convolutional neural networks,” *Distill*, vol. 4, no. 11, p. e21, 2019.
- [83] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” *Technologies*, vol. 9, no. 1, 2021. [Online]. Available: <https://www.mdpi.com/2227-7080/9/1/2>
- [84] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.
- [85] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
- [86] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers & distillation through attention,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 10 347–10 357. [Online]. Available: <https://proceedings.mlr.press/v139/touvron21a.html>
- [87] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.
- [88] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *Proc. Interspeech 2020*, 2020, pp. 5036–5040. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2020-3015>
- [89] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, “Decision transformer: Reinforcement learning via sequence modeling,” in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. [Online]. Available: <https://openreview.net/forum?id=a7APmM4B9d>
- [90] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [91] O. Groth, F. B. Fuchs, I. Posner, and A. Vedaldi, “Shapestacks: Learning vision-based physical intuition for generalised object stacking,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.



- 
- [92] M. Noroozi, H. Pirsiavash, and P. Favaro, “Representation learning by learning to count,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5899–5907.
  - [93] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
  - [94] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, “Bootstrap your own latent - a new approach to self-supervised learning,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 21 271–21 284. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf>
  - [95] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, “Deformable {detr}: Deformable transformers for end-to-end object detection,” in *International Conference on Learning Representations*, 2021. [Online]. Available: <https://openreview.net/forum?id=gZ9hCDWe6ke>
  - [96] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 10 012–10 022.
  - [97] C. Li, J. Yang, P. Zhang, M. Gao, B. Xiao, X. Dai, L. Yuan, and J. Gao, “Efficient self-supervised vision transformers for representation learning,” 2021.
  - [98] D. J. Rezende and F. Viola, “Taming vaes,” 2018.
  - [99] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
  - [100] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
  - [101] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2017.
  - [102] Y. LeCun and C. Cortes, “MNIST handwritten digit database,” 2010. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
  - [103] A. Kumar, S. Eslami, D. J. Rezende, M. Garnelo, F. Viola, E. Lockhart, and M. Shanahan,

- “Consistent generative query networks,” *arXiv preprint arXiv:1807.02033*, 2018.
- [104] A. R. Kosiosek, H. Strathmann, D. Zoran, P. Moreno, R. Schneider, S. Mokra, and D. J. Rezende, “Nerf-vae: A geometry aware 3d scene generative model,” in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 5742–5752. [Online]. Available: <https://proceedings.mlr.press/v139/kosiosek21a.html>
- [105] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, “Deepsdf: Learning continuous signed distance functions for shape representation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 165–174.
- [106] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “Nerf: Representing scenes as neural radiance fields for view synthesis,” in *ECCV*, 2020.
- [107] L. Nanbo, C. Eastwood, and R. B. Fisher, “Learning object-centric representations of multi-object scenes from multiple views,” in *34th Conference on Neural Information Processing Systems*, 2020.
- [108] C. Chen, F. Deng, and S. Ahn, “Roots: Object-centric representation and rendering of 3d scenes,” *Journal of Machine Learning Research*, vol. 22, no. 259, pp. 1–36, 2021.
- [109] T. Schmidt, “Perception: The binding problem and the coherence of perception,” in *Encyclopedia of Consciousness*, W. P. Banks, Ed. Oxford: Academic Press, 2009, pp. 147–158. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780123738738000591>
- [110] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0010028580900055>
- [111] M. Vasco, F. S. Melo, and A. Paiva, “Mhvae: a human-inspired deep hierarchical generative model for multimodal representation learning,” 2020.
- [112] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, “Information-theoretic regularization for learning global features by sequential vae,” *Machine Learning*, pp. 1–28, 2021.
- [113] J. Tobin, W. Zaremba, and P. Abbeel, “Geometry-aware neural rendering,” *Advances in Neural Information Processing Systems*, vol. 32, pp. 11 559–11 569, 2019.
- [114] P. Henderson and C. H. Lampert, “Unsupervised object-centric video generation and decomposition in 3D,” in *Advances in Neural Information Processing Systems (NeurIPS) 33*,

2020.

- [115] M. Engelcke, O. P. Jones, and I. Posner, “Genesis-v2: Inferring unordered object representations without iterative refinement,” *arXiv preprint arXiv:2104.09958*, 2021.
- [116] T. Taketomi, H. Uchiyama, and S. Ikeda, “Visual slam algorithms: A survey from 2010 to 2016,” *IPSN Transactions on Computer Vision and Applications*, vol. 9, 2017, publisher Copyright: © The Author(s).
- [117] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
- [118] M. Wu and N. Goodman, “Multimodal generative models for scalable weakly-supervised learning,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 5580–5590.
- [119] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, “Clevr: A diagnostic dataset for compositional language and elementary visual reasoning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.
- [120] O. Fried, E. Shechtman, D. B. Goldman, and A. Finkelstein, “Finding distractors in images,” in *Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [121] F.-L. Zhang, X. Wu, R.-L. Li, J. Wang, Z.-H. Zheng, and S.-M. Hu, “Detecting and removing visual distractors for video aesthetic enhancement,” *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 1987–1999, 2018.
- [122] K. Aberman, J. He, Y. Gandelsman, I. Mosseri, D. E. Jacobs, K. Kohlhoff, Y. Pritch, and M. Rubinstein, “Deep saliency prior for reducing visual distraction,” 2021.
- [123] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [124] K. Friston, J. Kilner, and L. Harrison, “A free energy principle for the brain,” *Journal of Physiology-Paris*, vol. 100, no. 1, pp. 70–87, 2006, theoretical and Computational Neuroscience: Understanding Brain Functions. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S092842570600060X>
- [125] M. Lohmann, J. Salvador, A. Kembhavi, and R. Mottaghi, “Learning about objects by learning to interact with them,” in *NeurIPS*, 2020.
- [126] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, “Domain random-

- ization for transferring deep neural networks from simulation to the real world,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 23–30.
- [127] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, “Sim-to-real transfer of robotic control with dynamics randomization,” in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3803–3810.
- [128] S. Levine, “Reinforcement learning and control as probabilistic inference: Tutorial and review,” 2018.
- [129] A. Lee, A. Nagabandi, P. Abbeel, and S. Levine, “Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [130] D. Han, K. Doya, and J. Tani, “Variational recurrent models for solving partially observable control tasks,” in *International Conference on Learning Representations*, 2020. [Online]. Available: <https://openreview.net/forum?id=r1lL4a4tDB>
- [131] A. d’Avila Garcez and L. C. Lamb, “Neurosymbolic ai: The 3rd wave,” 2020.
- [132] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, “The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision,” in *International Conference on Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=rJgMlhRctm>
- [133] 尾島司郎, 宮川繁, and 岡ノ谷一夫, “紙上討論 人間以外の動物に「文法」は使えるのか?” *Brain and nerve: 神経研究の進歩*, vol. 66, no. 3, pp. 273–281, 2014.
- [134] N. Chomsky, “Three models for the description of language,” *IRE Transactions on information theory*, vol. 2, no. 3, pp. 113–124, 1956.
- [135] R. C. Berwick, K. Okanoya, G. J. Beckers, and J. J. Bolhuis, “Songs to syntax: the linguistics of birdsong,” *Trends in cognitive sciences*, vol. 15, no. 3, pp. 113–121, 2011.

## 付録 A

# 研究業績一覧

### 論文誌

1. 小林 由弥, 鈴木 雅大, 松尾 豊, 深層生成モデルによる背景情報を利用したシーン解釈, 人工知能学会論文誌 (投稿中)
2. 小林 由弥, 鈴木 雅大, 松尾 豊, Transformer と自己教師あり学習を用いたシーン解釈手法の提案 (人工知能学会論文誌 第 37 巻 2 号 2022 年 3 月出版予定)
3. 小林由弥, 庭野恭彰, 田中敬, 赤尾旭彦, 小谷潔, 神保泰彦. (2018). 定量的な光刺激による一次視覚野の  $\alpha$  波の応答評価 . IEEJ Transactions on Electronics, Information and Systems Vol.138 No.7 pp.822-827 (2018 年電子・情報・システム部門誌 論文奨励賞)
4. 小林 由弥, 庭野恭彰, 田中敬, 赤尾旭彦, 小谷潔, 神保泰彦. (2017). 神経ネットワークの信号伝搬特性に関する数理的評価. IEEJ Transactions on Electronics, Information and Systems Vol.139 No.2 pp.154-160
5. **Kobayashi, Y**, Akao, A, Shirasaka, S, Kotani, K, Jimbo, Y. Mathematical analysis of the signal propagation characteristics of neuronal networks. Electron Comm Jpn. 2019;102:27-34. (上記論文の英語版)

### 国際学会

1. **Yuya Kobayashi**, Masahiro Suzuki, Yutaka Matsuo, Learning Global Spatial Information for Multi-View Object-Centric Models, ICLR2022 (投稿中)
2. **Kobayashi Yuya**, Gu Feng, Kotani Kiyoshi, Jimbo Yasuhiko, Multivariate Analysis Method Using Cross Recurrence Plot and Convolutional Neural Network 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology

Society (EMBC), USA, Hawaii, Honolulu, July 2018 (poster)

3. Gu Feng, **Kobayashi Yuya**, Kotani Kiyoshi, Jimbo Yasuhiko, Influence of Structures of Deep Neural Network on Classification of EEG During Motor Imagery Task 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), USA, Hawaii, Honolulu, July 2018 (poster)

## 国内学会発表(査読なし) ※筆頭著者以外のものは省略

1. 小林 由弥, 鈴木 雅大, 松尾 豊, Transformer を用いた深層生成モデルによる教師なし物体認識手法の提案, 人工知能学会全国大会 (2021)
2. 小林 由弥, 鈴木 雅大, 松尾 豊, データ分布の対照によるシーン認識モデルの改良, 人工知能学会全国大会 (2020)
3. 小林 由弥, 赤尾旭彦, 庭野恭彰, 小川雄太郎, 小谷潔, 神保泰彦 神経集団の内部状態とマクロな刺激応答特性に関する基礎的研究 平成 28 年度電気学会研究会「医用・生体工学会」2017 年 3 月 東京 MBE-17-039 (ポスター)
4. 小林 由弥, 庭野恭彰, 田中敬, 赤尾旭彦, 小谷潔, 神保泰彦 定量的な光刺激による一次視覚野の $\alpha$ 波の応答評価 平成 29 年電気学会 電子・情報・システム部門大会 2017 年 9 月 香川, サポートホール高松 1546-1547 (ポスター)

## 国際ワークショップ(査読なし)

1. Kobayashi Yuya Response of Alpha Wave and Pupil Diameter to Controlled Light Stimulus International workshop on analysis of phase dynamics and its application to biological phenomena 2017, June