原著論文 Original Paper

Transformerと自己教師あり学習を用いたシーン解釈手法の提案

Scene Interpretation Method using Transformer and Self-supervised Learning

小林 由弥 KOBAYASHI Yuya
東京大学大学院 工学系研究科
Graduate School of Engineering, The University of Tokyo
u-kobayashi@weblab.t.u-tokyo.ac.jp

鈴木 雅大 SUZUKI Masahiro
（同　上）
masa@weblab.t.u-tokyo.ac.jp

松尾 豊 MATSUO Yutaka
（同　上）
matsuo@weblab.t.u-tokyo.ac.jp

keywords: Deep Generative Models, Representation Learning, Scene Interpretation, Object Recognition, Self-supervised Learning, World Model

Summary

Ability to understand surrounding environment based on its components, namely objects, is one of the most important cognitive ability for intelligent agents. Human beings are able to decompose sensory input, i.e. visual stimulation, into some components based on its meaning or relationships between entities, and are able to recognize those components as “object ”. It is often said that this kind of compositional recognition ability is essential for resolving so called Binding Problem, and thus important for many tasks such as planning, decision making and reasoning. Recently, researches about obtaining object level representation in unsupervised manner using deep generative models have been gaining much attention, and they are called ”Scene Interpretation models”. Scene Interpretation models are able to decompose input scenes into symbolic entities such as objects, and represent them in a compositional way. The objective of our research is to point out the weakness of existing scene interpretation methods and propose some methods to improve them. Scene Interpretation models are trained in fully-unsupervised manner in contrast to latest methods in computer vision which are based on massive labeled data. Due to this problem setting, scene interpretation models lack inductive biases to recognize objects. Therefore, the application of these models are restricted to relatively simple toy datasets. It is widely known that introducing inductive biases to machine learning models is sometimes very useful like convolutional neural networks, but how to introduce them via training depends on the models and is not always obvious. In this research, we propose to incorporate self-supervised learning to scene interpretation models for introducing additional inductive bias to the models, and we also propose a model architecture using Transformer which is considered to be suitable for scene interpretation when combined with self-supervised learning. We show proposed methods outperforms previous methods, and is able to adopt to Multi-MNIST dataset which previous methods could not deal with well.

1. は じ め に

我々が住む実世界は，様々な種類の物体の組み合わせによって構成されている．こうした構成的な環境では，観測として得られる情報は一つの構成要素が変化するだけで大きく変化し，構成要素の数に応じて実現されうる組み合わせは指数的に増加する．環境の構成性を考慮しない場合，知的エージェントは膨大な組み合わせを個別に認識する必要があるが，これは汎化性能や計算効率，サンプル効率等の観点で望ましくない．また，構成性を考慮せずに複数の物体の認識を試みた場合には，個々の物体の性質が入り混じって正しく認識できなくなる可能性がある．これは神経科学や認知科学の分野で提起されてきた，バインディング問題として広く知られているものである [Revonsuo 99, Greff 16, Greff 20]．そのため環境の構成性を捉えることは，物体同士の関係性や自らと物体の相互作用を正しく把握することにつながり，ひいては意思決定や因果関係の推論（reasoning），異なる環境や未知の観測への汎化（分布外汎化）に役立つ．こうした情報処理は高度な知的エージェントの構築に不可欠であり，ロボティクスや自動運転，強化学習などの分野への貢献が期待される [Devin 18, Veerapaneni 19].

ところで我々人間の認識や思考も構成的であり，物体や何らかの概念などのシンボリックな対象を基本単位として多くの事象を捉えている．視野全体にわたって複数の物体の詳細な性質を認識することはできず，主に視野

の中央において部分的に認識を行い，思考することが知られている [Chen 12]．こうした構成的な認識や思考は人間レベルの情報処理を人工知能で実現するために考慮すべき特徴であるし，構成的な処理を行うエージェントを人工的に構築することが生物の認知機能の理解のための構成論的なアプローチとなることも期待できる．

近年，知的エージェントが環境を認識する仕組みは深層生成モデルによって実現されており，代表例として Variational Autoencoder（VAE）や，Generative Adversarial Network がある [Kingma 13, Goodfellow 20]．深層生成モデルとは，生成モデルにおいて確率分布のモデル化に深層ニューラルネットワークを用いる手法の総称である．生成モデルは観測データが未知の確率分布に従っていると考え，その生成過程を確率モデルとして表すものである．また，生成の要因として潜在変数と呼ばれる確率変数を設定することもある．深層生成モデル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，潜在変数のサンプリングによって観測には含まれない新たなデータを生成することが実際に可能となった．これにより，環境の未観測の部分や未来の状況を予測したり，経験にない未知の環境を想像することも可能となっている．これは環境についてのモデル（シミュレータ）をエージェントの内部に持つことに相当し，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 [Johnson-Laird 83, Kawato 99, Ha 18, Hafner 19].

知的エージェントの観点では，観測から生成要因（潜在変数）を推論することが認識に相当し，事前分布からの潜在変数のサンプリングによって観測に含まれないデータを生成することは予測や想像に相当する．そのため生成モデルの観点では構成的な認識とは，観測情報から系全体に対して一つの生成要因を推論するのではなく，物体ごとにその位置情報や性質といった生成要因を推論することと捉えることができる．深層生成モデルにおける潜在変数の推論は，潜在変数の事後分布を近似する推論用のニューラルネットワークを用意する手法（VAE）か，生成モデルを可逆な関数によって近似し，逆変換として推論を行う手法（flow-based モデル）が一般的である [Dinh 14]．したがって，深層生成モデルの枠組みにおいて構成的な認識を実現する一つの手段は，個々の構成要素（物体など）に対応する複数の潜在変数を仮定し，観測情報からそれらを推論することによって個々の物体の情報を得ることである．

複数の潜在変数を仮定し，それぞれが個別の物体を生成する深層生成モデルで，特に潜在変数の推論が可能な手法は，シーン解釈モデル（Scene Interpretation Model）と呼ばれている．[Greff 17, Steenkiste 18, Burgess 19, Greff 19, Engelcke 20, Eslami 16, Crawford 19a, Lin 20, Jiang 20]．ただし，複数の潜在変数を仮定し物体を個別に生成する手法でも，生成のみを目的として推論を行う機構を持たないものも存在する [Nguyen-Phuoc 20, Ehrhardt 20]．これらはシーン解釈モデルとは呼ばれず，シーン解釈は認識（推論）に主眼が置かれたものである．シーン解釈モデルは，セグメンテーションマスクやバウンディングボックスといった物体の位置情報と，視覚的な特徴など意味的な情報が各潜在変数から物体の数だけ生成され，それらの組み合わせで全体が表現される構造となっている．各潜在変数には適当な確率分布が仮定されており，上記のような各物体の意味的・空間的情報が潜在変数空間上に学習されることになる．したがって，逆に観測から潜在変数の推論を行うことで物体の情報が獲得され，物体を基本単位として環境を構成的に認識することができる．

シーン解釈には主に二つの方式があり，そのうちの一つとして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ものがある（Scene Mixture モデル）[Greff 17, Steenkiste 18, Burgess 19, Greff 19, Engelcke 20]．これはもう一つのバウンディングボックスを用いる方式（Neurosymbolic モデル）[Eslami 16, Crawford 19a, Lin 20, Jiang 20] に対して，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点で利点がある．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はなく，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．この場合，表現の自由度の代償として推論がより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を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，目的関数を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に意味のない分割となってしまう [Greff 19].

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 [Luo 16, Araujo 19]．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．

本研究では，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，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．

モデルに帰納バイアスを与える方法は自明ではなく，しばしば機械学習研究の貢献となる [Goyal 21]．基本的な方法としては，以下の三点が挙げられる．

(1) モデルアーキテクチャに制約を与えること

(2) 与えるデータの種類や量を増やすこと

(3) 目的関数への正則化項の導入やカリキュラム学習など学習の工夫を行うこと

Neurosymbolic モデルのように変数に特定の役割を与えることはアーキテクチャの制約に相当するが，必然的に手法の適用範囲は制限されることになる（上記 1 番）．また，与えるデータの種類を変更することは，問題設定を変更することに相当する（上記 2 番）．そこで本研究では，自己教師あり学習の利用による物体の帰納バイアスの導入を提案する．これは上記の分類では 3 番目に相当する．

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 [Jaiswal 21]．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，これまで大域的な特徴を捉えるために様々な工夫が行われてきた [Dai 17, Wang 18]．これに対し，近年は畳み込み処理を用いずに Transformer [Vaswani 17] 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 [Dosovitskiy 20, Touvron 20, Carion 20, Gulati 20, Chen 21]．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 [Dosovitskiy 20]．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，既存手法 [Engelcke 20] では物体間の関係性をモデル化するために LSTM 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 [Hinton 15] によって学習を補助する必要がある [Touvron 20]．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 [Kabra 19]，より複雑な背景やテクスチャを含む ShapeStacks データセット [Groth 18]，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 [Eslami 16] 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．

2. 関 連 研 究

本章では，シーン解釈の位置付けと関連研究について説明する．また，自己教師あり学習や Transformer による画像処理（Vision Transformer）について関連研究の説明を行う．

2·1 シ ー ン 解 釈

シーン解釈モデル（Scene Interpretation model）は複数の潜在変数を持つ深層生成モデルであり，それぞれの潜在変数が個別の物体を表現することでシーン全体を生成する手法である．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手法 [Nguyen-Phuoc 20, Ehrhardt 20] 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．

ところで，個別の物体に関する特徴量が得られる学習手法一般は，物体中心表現学習（object-centric representation learning）と呼ばれている．シーン解釈はその中でも深層生成モデルによる確率的なモデル化を行った手法であり，モデルによっては事前分布からの潜在変数のサンプリングにより，新たなシーンの生成も可能となっている [Engelcke 20, Jiang 20]．これに対して物体の特徴量を得ることのみを目的とした決定論的な定式化を行う研究も行われており，研究や文脈により用語の使用方法は若干の差異があるが，これらは基本的にシーン解釈とは呼ばない [Kipf 19, Locatello 20]．また，シーン解釈やこれまでに触れた研究については教師なしで特徴量を獲得する取り組みであるが，教師データを用いる研究も存在する [Devin 18]．目的や手法は異なるが，これも物体中心表現学習の一種である．

一般にコンピュータビジョンの分野で主流となっている物体認識の手法も教師あり学習に基づくもので，これは大量のデータを用いて入力画像と正解（物体の座標など）の関係を直接近似する関数を学習している [Girshick 14]．追加でセグメンテーションを同時に行うモデルなども存在するが，物体に関する特徴量の学習を目的とはしておらず，あくまで教師データに基づいた物体の位置情報の推定を行うことを主眼に置いている [He 17]．

深層生成モデルを用いることで，観測から環境をモデリングし，推論によって系の状態を把握したり，生成によって未知の状況を予測することが可能となる．このような環境（世界）のモデルは分野によって「世界モデル」や「メンタルモデル」,「内部モデル」などと呼ばれている [Johnson-Laird 83, Kawato 99, Ha 18, Hafner 19]．特にその系に関する新たなデータの生成が可能な場合は，生成結果を用いて知的エージェントが仮想的な訓練を行うことも可能で，サンプル効率や汎化性能の向上に貢献することが期待される [Ha 18]．シーン解釈モデルは深層生成モデルで実現されており，推論や生成が可能であるため，世界モデルとして利用することが可能である．特にシーン解釈で得られるような物体単位の表現は，物体同士の相互作用の推定や，因果関係の推論，未知の組み合わせの予測（組み合わせ汎化）など，高度な認知機能の実現に重要な役割を果たすと考えられる．これはロボティクスの分野での活用が期待され，[Veerapaneni 19] ではシーン解釈の枠組みにおいて更に物体との相互作用を考慮し，エージェントの認識やプランニングを向上させている．

シーン解釈には様々な方式が存在しているが，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．一つはバウンディングボックスによるもの（Neurosymbolic モデル）[Eslami 16, Crawford 19b, Jiang 19, Jiang 20]，もう一つはセグメンテーションによるもの（Scene Mixture モデル）[Greff 17, Steenkiste 18, Burgess 19, Greff 19, Engelcke 20] であり，もしくはこれらを併用したもの [Lin 20] である．以下の項ではそれぞれの方式について説明する．

§1 バウンディングボックスを用いたシーン解釈手法（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（presence）を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．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，拘束しない場合に対して表現の自由度が下がる代わりに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 [Eslami 16] 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 [Crawford 19b] 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 [Lin 20] 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，[Lin 20] や [Jiang 19] 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．

§2 空間的混合ガウス分布を用いたシーン解釈手法（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．推論の難しさは 1 章で述べ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で行われる．先行研究 [Greff 19] では反復的な推論（Iterative Amortized Inference（IAI））[Marino 18] 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 [Burgess 19] 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 [Engelcke 20]．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．

本研究では，複雑な形状の物体や背景などを含む場合でも，統一的に扱うことができる Scene Mixture モデルを扱うことを考える．また，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 [Jiang 20] 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，3 章で詳述する．

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 [Noroozi 17, Gidaris 18]．[Noroozi 17] 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，[Gidaris 18] 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 [Chen 20, He 20]．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，一つの画像を一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）[Grill 20] 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 θ のネットワーク A と，同じ構造で異なるパラメータ ξ 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 ξ はネットワーク A のパラメータ θ 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 θ 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 3·2·3 節にて述べる．

2·3 Vision Transformer

Transformer [Vaswani 17] 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 [Dosovitskiy 20, Touvron 20]．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 [Carion 20, Zhu 20, Liu 21].

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 [Dosovitskiy 20] 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 [Touvron 20, Li 21].

Vision Transformer の研究 [Dosovitskiy 20] 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 [Carion 20] 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．

3. 手　　法

1 章において，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 [Engelcke 20] 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章（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 [Wang 18]．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 [Dai 17]，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 [Wang 18]．これに対し，近年は畳み込みを用いずに，Transformer [Vaswani 17] 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 [Dosovitskiy 20, Touvron 20, Carion 20, Gulati 20, Chen 21]．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 [Dosovitskiy 20]．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 [Hinton 15] によって学習を補助する必要がある [Touvron 20]．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．一つ目は，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．

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 2 章で紹介した通り，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．

$$\hat{\mathbf{x}} = \sum_{k=1}^{K} \mathbf{x}_k \otimes \mathbf{m}_k \tag{1}$$

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．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 (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 [Burgess 19] 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．

$$p_\xi(\mathbf{x})$$
$$=\int\int p_\theta(\mathbf{x}|\mathbf{z}^c,\mathbf{z}^m)\,p_\phi(\mathbf{z}^c\mid\mathbf{z}^m)\,p_\psi(\mathbf{z}^m)\,d\mathbf{z}^m d\mathbf{z}^c \quad (2)$$

ただし

$$p_\phi(\mathbf{z}^c\mid\mathbf{z}^m)=\prod_{k=1}^{K}p_\phi(\mathbf{z}_k^c\mid\mathbf{z}_k^m) \quad (3)$$

$$p_\psi(\mathbf{z}^m)=p(\mathbf{z}_1^m)\prod_{k=1}^{K-1}p_\psi(\mathbf{z}_{k+1}^m\mid\mathbf{z}_{1:k}^m) \quad (4)$$

とする．また，$\xi=\{\theta,\phi,\psi\}$ とまとめたものとする．式 (2) の右辺第二項 $p_\phi(\mathbf{z}^c|\mathbf{z}^m)$ はパラメータ ϕ 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 [Hochreiter 97] 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 [Kingma 13] と同様に，Amortized Inference によって行われる．上記式 (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．

$$\mathbf{z}_k^m\sim q_\mu(\mathbf{z}_k^m|\mathbf{x},\mathbf{z}_{1:k-1}^m)\qquad(2\le k\le K) \quad (5)$$
$$\mathbf{z}_1^m\sim q_\mu(\mathbf{z}_1^m|\mathbf{x})$$
$$\mathbf{z}_k^c\sim q_\eta(\mathbf{z}_k^c|\mathbf{x},\mathbf{z}_k^m)\qquad(1\le k\le K) \quad (6)$$

式 (5) は式 (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) 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 $\mathcal{L}(\mathbf{x})$ は以下のようになる．

$$\mathcal{L}(\mathbf{x})=\mathbb{E}_{q_\rho(\mathbf{z}^c,\mathbf{z}^m|\mathbf{x})}\left[p_\theta(\mathbf{x}|\mathbf{z}^c,\mathbf{z}^m)\right]$$
$$-\beta\{KL\left[q_\mu(\mathbf{z}^m|\mathbf{x})\,||\,p_\psi(\mathbf{z}^m)\right]$$
$$+\mathbb{E}_{q_\mu(\mathbf{z}^m|\mathbf{x})}[KL\left[q_\rho(\mathbf{z}^c,\mathbf{z}^m|\mathbf{x})\,||\,p_\phi(\mathbf{z}^c|\mathbf{z}^m)p_\psi(\mathbf{z}^m)\right]]\} \quad (7)$$

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，β は KL 項の影響力を決定する係数である．β 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 [Rezende 18] 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，β 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 β の代わりに再構成誤差（式 (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一つ設定すれば良いので，学習ステップに応じた最適な β 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．

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．

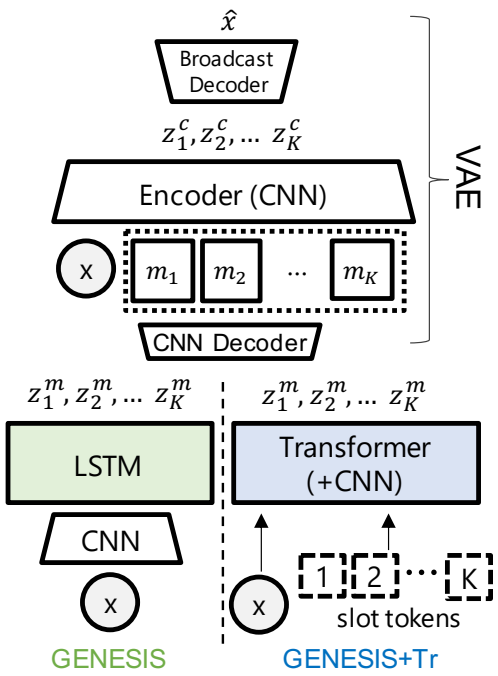

図 1 GENESIS と Transformer を用いたモデル構造（GENESIS+Tr）．後段の VAE 部分は共通した構造となっている．

§1 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 [Greff 19] 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 $\mathbf{x}$ 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 $\mathbf{z}_k$ を得る．ここまでの過程は式 (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，全結合層によって一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 (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．図 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，以後本項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 [Watters 19]．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 [Burgess 19]．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．

§2 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 (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図 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 [Dosovitskiy 20] 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 [Dosovitskiy 20] の一部や，[Zhu 20] 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 [Kingma 13] 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 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（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 [Vaswani 17] 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 [Hendrycks 16]．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 [Devlin 19] などで採用されている方法である．

§3 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 2·2 節で述べた通り，Bootstrap Your Own Latent（BYOL）[Grill 20] 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 [Deng 09] 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，64×64 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 64×64 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．

§4 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（4 章にて詳細を説明）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 4 章の図 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，$i, j \in \{1, 2, ..., K\}$，$i \neq j$ 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．現状ではこの制約の導入に関する定量的な規準はなく，通常の目的関数で適切な分割が行われなかった場合に導入を試みることを考えている．

4. 実　　験

本章では，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 [Kingma 17] 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 [Paszke 19]．

検証においては，Objects Room データセット [Kabra 19] と ShapeStacks データセット [Groth 18]，そして Multi-MNIST [Eslami 16] 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 [LeCun 10] 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 [Eslami 16]．

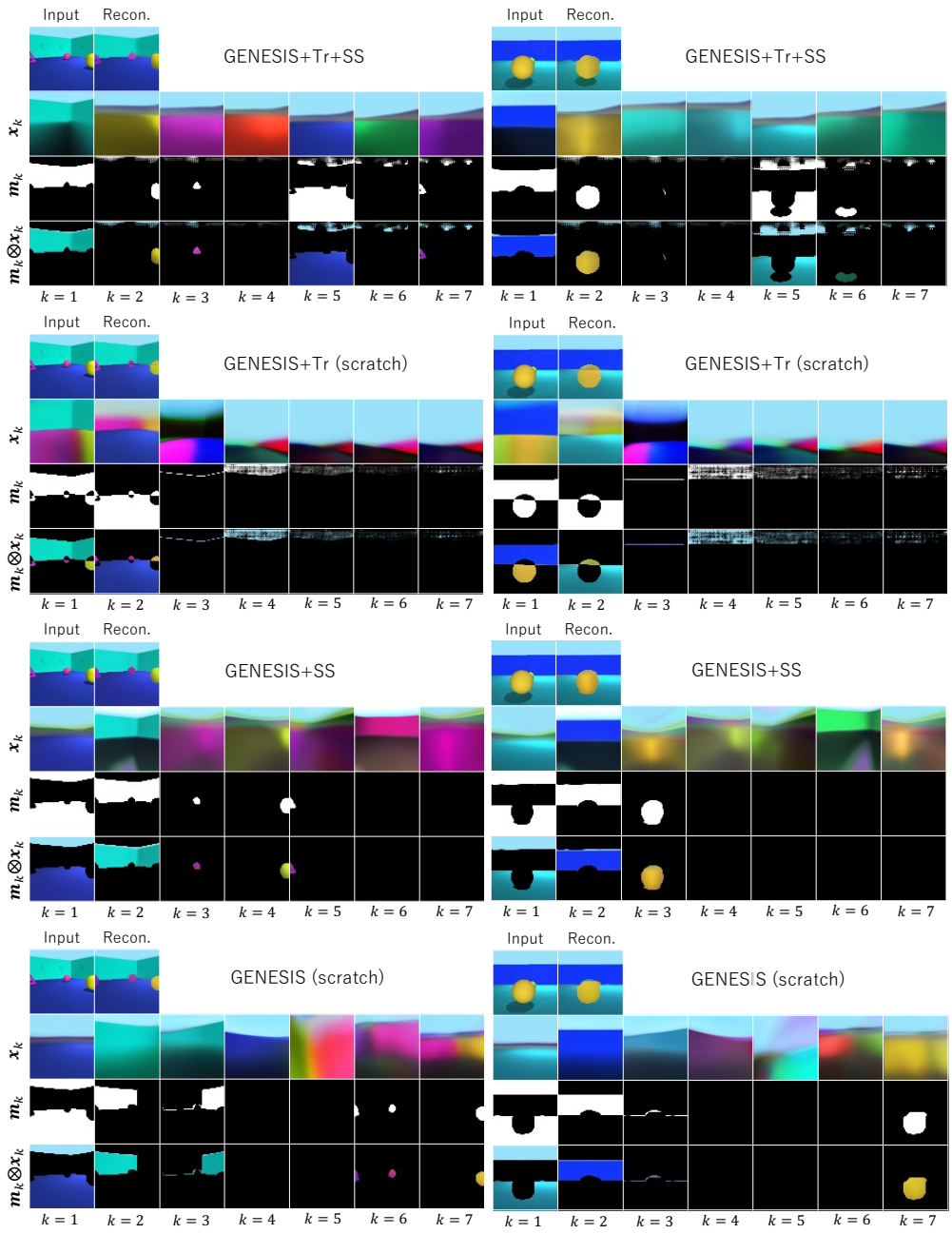

図 2 各モデルの Objects Room における推論結果．k はスロットの番号である．
Tr は Transformer，SS は自己教師あり学習を指す．また，scratch は自己教師あり学習を用いない通常の学習を指す．

変更点として画像解像度は 64×64 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，一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 (7) における KL 項の係数 β 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，

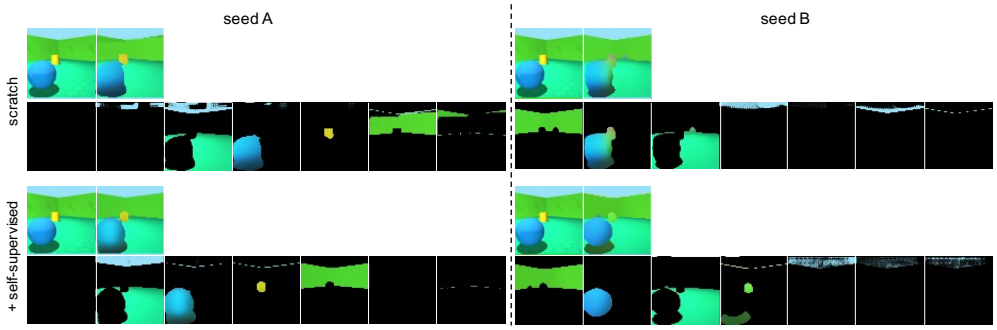

図 3 自己教師あり学習による学習の安定化を複数の乱数シード（学習時）について比較した結果．ここで利用したモデルは GENESIS+Tr である．上段が自己教師あり学習を用いない場合，下段が自己教師あり学習を用いた場合となっている．左右の列 (seed A, seed B) は異なる乱数シードに対応している．

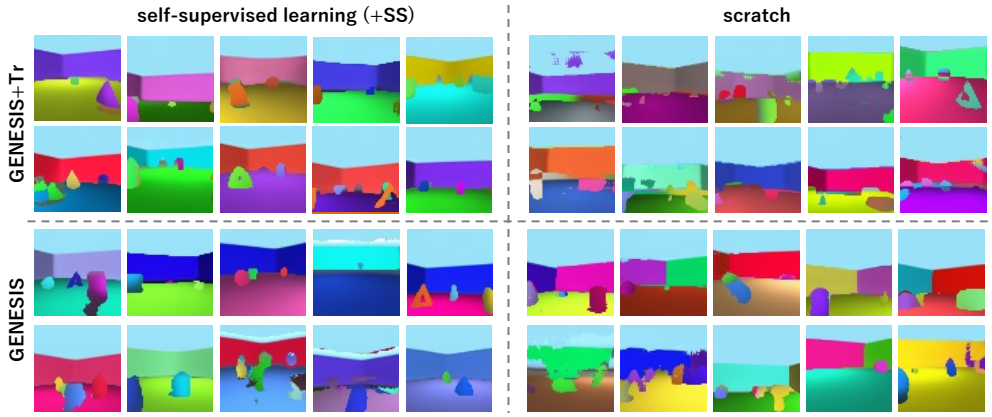

図 4 Objects Room データセットでの生成結果
左列は自己教師あり学習の，右列は通常の学習の結果を示している．

Multi-MNIST データセットにおいては β 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．

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．以下の項では，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．

§1 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．推論結果を図 2 に，生成結果を図 4 に示した．また，図 3 に自己教師あり学習の導入による変化を複数の学習時の乱数シードについて検証した結果も示した．図中の scratch は自己教師あり学習を用いない，通常の学習であることを意味する．

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．図 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 $k = 5$ の二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対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，図 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．図 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項にて定量評価を行う．

§2 定量評価

本項では各モデルでの 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）の値を表 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．表の値は三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．各条件での値を表 2 に示した．これは三つの乱数 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 $\mathbf{z}_m^k$ 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の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．改めて図 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かかわらず，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，

表 1 Objects Room データセットにおける mSC

model	GENESIS	GENESIS+Tr
scratch	0.55 ± 0.04	0.49 ± 0.14
self-supervised	0.57 ± 0.03	**0.59** ± 0.04

表 2 Objects Room データセットにおける FID

model	GENESIS	GENESIS+Tr
scratch	65.0 ± 7.1	96.29 ± 8.7
self-supervised	**59.2** ± 3.3	77.5 ± 2.1

表 3 ShapeStacks データセットにおける mSC

model	GENESIS	GENESIS+Tr
scratch	0.57 ± 0.07	0.53 ± 0.05
self-supervised	**0.58** ± 0.06	**0.58** ± 0.07

表 4 ShapeStacks データセットにおける FID

model	GENESIS	GENESIS+Tr
scratch	197.9 ± 4.0	231.5 ± 3.1
self-supervised	**191.5** ± 2.1	226.4 ± 3.2

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．一つは単純に，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が期待できる．もう一つは，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．

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．

表 3 と表 4 にそれぞれ ShapeStacks データセットでの mSC と FID を示した．GENESIS，GENESIS+Tr 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．

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ないという点でより難しい課題となっている．定性的な結果を図 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表 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 3·2·4 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（図 5: 右）では数字は分割されず，一つのスロット（$k = 1$）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，ここまで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

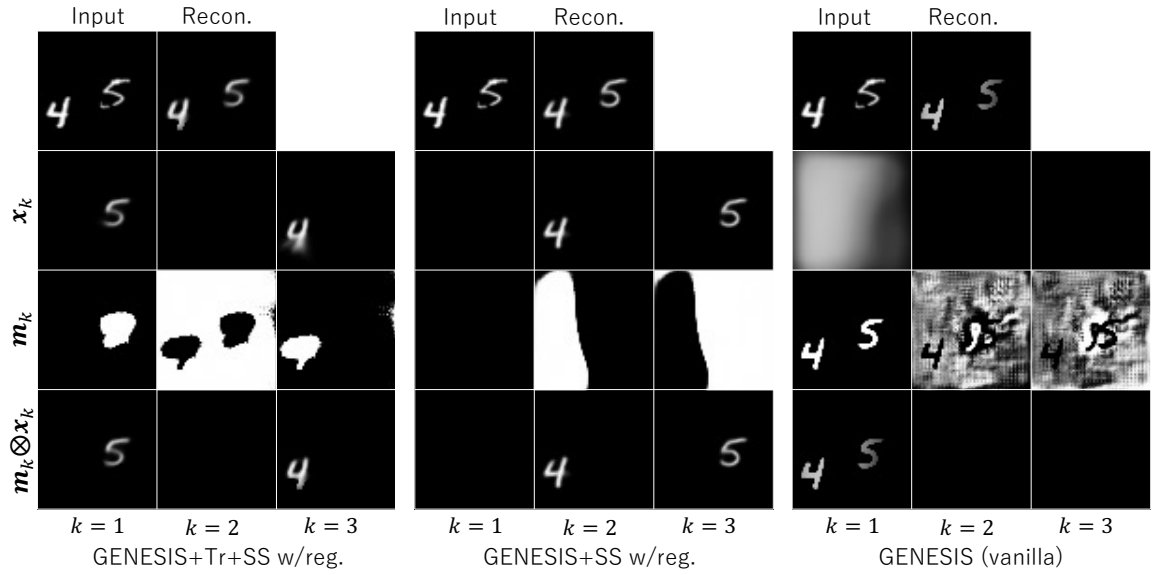

図 5 Multi-MNIST データセットでの推論結果．w/reg. は制約項の導入を意味する．

用いて，三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．表 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，図 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一つの場合，セグメンテーション自体は成功していることになり，数字が二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図 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図 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．

表 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 ± 0.04)	386.5 ± 14.7
+SS+Reg.	0.07 ± 0.01	342.5 ± 8.8
+Tr+SS+Reg.	$\mathbf{0.17} \pm 0.19$	$\mathbf{330.7} \pm 1.5$

5. 結　　論

本研究では，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の探索は行わなかった．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．

◇ 参 考 文 献 ◇

[Araujo 19] Araujo, A., Norris, W., and Sim, J.: Computing receptive fields of convolutional neural networks, *Distill*, Vol. 4, No. 11, p. e21 (2019)

[Burgess 19] Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., and Lerchner, A.: Monet: Unsupervised scene decomposition and representation, *arXiv preprint arXiv:1901.11390* (2019)

[Carion 20] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S.: End-to-end object detection with transformers, in *European Conference on Computer Vision*, pp. 213–229Springer (2020)

[Chen 12] Chen, Z.: Object-based attention: A tutorial review, *Attention, Perception, & Psychophysics*, Vol. 74, No. 5, pp. 784–802 (2012)

[Chen 20] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G.: A simple framework for contrastive learning of visual representations, in *International conference on machine learning*, pp. 1597–1607PMLR (2020)

[Chen 21] Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I.: Decision Transformer: reinforcement learning via sequence modeling, *arXiv preprint arXiv:2106.01345* (2021)

[Crawford 19a] Crawford, E. and Pineau, J.: Spatially invariant unsupervised object detection with convolutional neural networks, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 3412–3420 (2019)

[Crawford 19b] Crawford, E. and Pineau, J.: Spatially invariant unsupervised object detection with convolutional neural networks, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 3412–3420 (2019)

[Dai 17] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y.: Deformable convolutional networks, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 764–773 (2017)

[Deng 09] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255Ieee (2009)

[Devin 18] Devin, C., Abbeel, P., Darrell, T., and Levine, S.: Deep object-centric representations for generalizable robot learning, in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7111–7118 (2018)

[Devlin 19] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K.: 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)*, pp. 4171–4186, Minneapolis, Minnesota (2019), Association for Computational Linguistics

[Dinh 14] Dinh, L., Krueger, D., and Bengio, Y.: Nice: Non-linear independent components estimation, *arXiv preprint arXiv:1410.8516* (2014)

[Dosovitskiy 20] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* (2020)

[Ehrhardt 20] Ehrhardt, S., Groth, O., Monszpart, A., Engelcke, M., Posner, I., Mitra, N., and Vedaldi, A.: RELATE: Physically plausible multi-object scene synthesis using structured latent spaces, in Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. eds., *Advances in Neural Information Processing Systems*, Vol. 33, pp. 11202–11213, Curran Associates, Inc. (2020)

[Engelcke 20] Engelcke, M., Kosiorek, A. R., Jones, O. P., and Posner, I.: GENESIS: Generative scene inference and sampling with object-centric latent representations, in *International Conference on Learning Representations* (2020)

[Eslami 16] Eslami, S. A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Hinton, G. E., et al.: Attend, infer, repeat: Fast scene understanding with generative models, in *Advances in Neural Information Processing Systems*, pp. 3225–3233 (2016)

[Gidaris 18] Gidaris, S., Singh, P., and Komodakis, N.: Unsupervised representation learning by predicting image rotations, *arXiv preprint arXiv:1803.07728* (2018)

[Girshick 14] Girshick, R., Donahue, J., Darrell, T., and Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587 (2014)

[Goodfellow 20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative adversarial networks, *Communications of the ACM*, Vol. 63, No. 11, pp. 139–144 (2020)

[Goyal 21] Goyal, A. and Bengio, Y.: Inductive biases for deep learning of higher-level cognition, *arXiv preprint arXiv:2011.15091* (2021)

[Greff 16] Greff, K., Srivastava, R. K., and Schmidhuber, J.: Binding via reconstruction clustering, *arXiv preprint arXiv:1511.06418* (2016)

[Greff 17] Greff, K., Steenkiste, van S., and Schmidhuber, J.: Neural expectation maximization, in Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. eds., *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc. (2017)

[Greff 19] Greff, K., Kaufmann, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., and Lerchner, A.: Multi-object representation learning with iterative variational inference, *arXiv preprint arXiv:1903.00450* (2019)

[Greff 20] Greff, K., Steenkiste, van S., and Schmidhuber, J.: On the binding problem in artificial neural networks, *arXiv preprint arXiv:2012.05208* (2020)

[Grill 20] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., et al.: Bootstrap your own latent: A new approach to self-supervised learning, *arXiv preprint arXiv:2006.07733* (2020)

[Groth 18] Groth, O., Fuchs, F. B., Posner, I., and Vedaldi, A.: ShapeStacks: Learning vision-based physical intuition for generalised object stacking, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018)

[Gulati 20] Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R.: Conformer: Convolution-augmented transformer for speech recognition, in *Proc. Interspeech 2020*, pp. 5036–5040 (2020)

[Ha 18] Ha, D. and Schmidhuber, J.: World models, *arXiv preprint arXiv:1803.10122* (2018)

[Hafner 19] Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M.: Dream to control: learning behaviors by latent imagination, in *International Conference on Learning Representations* (2019)

[He 17] He, K., Gkioxari, G., Dollár, P., and Girshick, R.: Mask R-CNN, in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988 (2017)

[He 20] He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R.: Momentum contrast for unsupervised visual representation learning, in *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9729–9738 (2020)
[Hendrycks 16] Hendrycks, D. and Gimpel, K.: Gaussian error linear units (GELUs), *arXiv preprint arXiv:1606.08415* (2016)
[Hinton 15] Hinton, G., Vinyals, O., and Dean, J.: Distilling the knowledge in a neural network, *arXiv preprint arXiv:1503.02531* (2015)
[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural Comput.*, Vol. 9, No. 8, p. 1735–1780 (1997)
[Jaiswal 21] Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F.: A Survey on contrastive self-supervised learning, *Technologies*, Vol. 9, No. 1 (2021)
[Jiang 19] Jiang, J., Janghorbani, S., De Melo, G., and Ahn, S.: SCALOR: Generative world models with scalable object representations, in *International Conference on Learning Representations* (2019)
[Jiang 20] Jiang, J. and Ahn, S.: Generative neurosymbolic machines, *Advances in Neural Information Processing Systems*, Vol. 33, (2020)
[Johnson-Laird 83] Johnson-Laird, P.: *Mental Models: Towards a cognitive science of language, inference, and consciousness*, Cognitive science series, Harvard University Press (1983)
[Kabra 19] Kabra, R., Burgess, C., Matthey, L., Kaufman, R. L., Greff, K., Reynolds, M., and Lerchner, A.: Multi-object datasets, https://github.com/deepmind/multi-object-datasets/ (2019)
[Kawato 99] Kawato, M.: Internal models for motor control and trajectory planning, *Current Opinion in Neurobiology*, Vol. 9, No. 6, pp. 718–727 (1999)
[Kingma 13] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013)
[Kingma 17] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2017)
[Kipf 19] Kipf, T., Pol, van der E., and Welling, M.: Contrastive learning of structured world models, *arXiv preprint arXiv:1911.12247* (2019)
[LeCun 10] LeCun, Y. and Cortes, C.: MNIST handwritten digit database (2010)
[Li 21] Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., and Gao, J.: Efficient self-supervised vision transformers for representation learning, *arXiv preprint arXiv:2106.09785* (2021)
[Lin 20] Lin, Z., Wu, Y.-F., Peri, S. V., Sun, W., Singh, G., Deng, F., Jiang, J., and Ahn, S.: SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition, *arXiv preprint arXiv:2001.02407* (2020)
[Liu 21] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows, *arXiv preprint arXiv:2103.14030* (2021)
[Locatello 20] Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., and Kipf, T.: Object-centric learning with slot attention, *arXiv preprint arXiv:2006.15055* (2020)
[Luo 16] Luo, W., Li, Y., Urtasun, R., and Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4905–4913 (2016)
[Marino 18] Marino, J., Yue, Y., and Mandt, S.: Iterative amortized inference, *arXiv preprint arXiv:1807.09356* (2018)
[Nguyen-Phuoc 20] Nguyen-Phuoc, T., Richardt, C., Mai, L., Yang, Y.-L., and Mitra, N.: BlockGAN: Learning 3D object-aware scene representations from unlabelled images, in *Advances in Neural Information Processing Systems 33* (2020)
[Noroozi 17] Noroozi, M., Pirsiavash, H., and Favaro, P.: Representation learning by learning to count, in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5899–5907 (2017)
[Paszke 19] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An imperative style, high-performance deep learning library, in Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, d'F., Fox, E., and Garnett, R. eds., *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc. (2019)
[Revonsuo 99] Revonsuo, A. and Newman, J.: Binding and consciousness, *Consciousness and Cognition*, Vol. 8, No. 2, pp. 123–127 (1999)
[Rezende 18] Rezende, D. J. and Viola, F.: Taming VAEs, *arXiv preprint arXiv:1810.00597* (2018)
[Steenkiste 18] Steenkiste, van S., Chang, M., Greff, K., and Schmidhuber, J.: Relational neural expectation maximization: unsupervised discovery of objects and their interactions, in *International Conference on Learning Representations* (2018)
[Touvron 20] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H.: Training data-efficient image transformers & distillation through attention, *arXiv preprint arXiv:2012.12877* (2020)
[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is all you need, *arXiv preprint arXiv:1706.03762* (2017)
[Veerapaneni 19] Veerapaneni, R., Co-Reyes, J. D., Chang, M., Janner, M., Finn, C., Wu, J., Tenenbaum, J. B., and Levine, S.: Entity abstraction in visual model-based reinforcement learning, in *CoRL* (2019)
[Wang 18] Wang, X., Girshick, R., Gupta, A., and He, K.: Non-local neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803 (2018)
[Watters 19] Watters, N., Matthey, L., Burgess, C. P., and Lerchner, A.: Spatial broadcast decoder: a simple architecture for learning disentangled representations in VAEs, *CoRR*, Vol. abs/1901.07017, (2019)
[Zhu 20] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection, *arXiv preprint arXiv:2010.04159* (2020)

〔担当委員：高橋 茶子〕

2021 年 7 月 29 日　受理

著　者　紹　介

小林　由弥

2017 年 東京大学工学部卒業．2022 年同大学院工学系研究科博士課程修了．博士（工学）．深層生成モデル，計算論的神経科学，生体計測などの研究に従事．

鈴木　雅大（正会員）

2013 年北海道大学工学部卒業．2015 年同大学院修士課程修了．2018 年東京大学大学院工学系研究科博士課程修了．博士（工学）．2020 年まで東京大学大学院工学系研究科技術経営戦略学専攻 特任研究員．同年より同大学特任助教．人工知能，深層学習の研究に従事．

松尾　　豊（正会員）

1997 年東京大学工学部卒業．2002 年同大学院博士課程修了．博士（工学）．2019 年より東京大学大学院人工物工学研究センター／技術経営戦略学専攻 教授．2014 年より 2018 年まで人工知能学会倫理委員長．2017 年より日本ディープラーニング協会理事長．人工知能学会論文賞，情報処理学会長尾真記念特別賞，ドコモモバイルサイエンス賞など受賞．専門は，人工知能，深層学習，Web 工学．