DINOSAURによる鏡面反射物体の教師なしセグメンテーション —ViTバックボーンの比較と構造的限界の検証—

Unsupervised Segmentation of Specular Objects via DINOSAUR: A Comparative Study of ViT Backbones and Structural Limitations

坂口 健 [*1]　　中世 大雄 [*2]
Ken Sakaguchi[*1]　　Hirooki Nakase

[*1]電気通信大学　　[*2]東京大学
The University of Electro-Communications　　The University of Tokyo

特徴再構成型の物体中心学習モデル DINOSAUR は，凍結 ViT バックボーンの特徴空間を活用することで教師なし物体セグメンテーションの精度を向上させた．しかし，鏡面反射（金属光沢）物体への適用可能性や，バックボーン選択がもたらす影響は十分に検証されていない．本研究では，DINOSAUR アーキテクチャに 3 種の ViT バックボーン（DINOv2, DINOv1, CLIP）を適用し，MOVi-A の金属物体サブセット（60 サンプル）を用いて定量的・定性的に評価した．実験の結果，DINOv2 が最も高い FG-ARI（0.165）を達成したが，16×16 パッチ解像度によるマスク境界のにじみや，鏡面反射によるスロット混同など，現行アーキテクチャの構造的限界が明らかになった．また，バックボーン間の特徴量スケール差が損失関数の設計に深刻な影響を与えることを示し，チャンネル正規化損失による対処法を提案した．

1. はじめに

物体中心学習（Object-Centric Learning, OCL）は，教師なしで画像中の物体を個別のスロットに分離・表現する枠組みであり，Slot Attention[Locatello 20] の提案以降，活発に研究されている．初期の Slot Attention はピクセル空間での再構成を目的関数としていたが，この手法は色やテクスチャの変化に脆弱である．DINOSAUR[Seitzer 23] は，事前学習済み ViT の凍結された特徴空間を再構成ターゲットとすることでこの限界を克服し，実画像への適用を可能にした．

しかし，DINOSAUR が対象としてきたシーンは主に Lambert 反射を仮定できる拡散反射物体であり，鏡面反射（金属光沢）物体に対する適用可能性は十分に検証されていない．金属物体は視点・照明に応じて外観が劇的に変化し，物体表面に周囲環境が映り込むため，特徴ベースのセグメンテーションにとって本質的に困難な対象である．

さらに，DINOSAUR の既存研究は DINOv2 バックボーンを標準的に採用しており，異なる ViT バックボーンを用いた場合に生じる特徴空間の幾何学的差異とその影響は体系的に分析されていない [Lüddecke 25]．

本研究では以下の 3 つの問いに取り組む：

1. DINOSAUR は鏡面反射物体をどの程度分離できるか
2. バックボーンの選択（DINOv2, DINOv1, CLIP）は性能にどう影響するか
3. 現行アーキテクチャの構造的限界はどこにあるか

本研究の貢献は新規手法の提案ではなく，既存の最先端モデルを困難な対象に適用した際の構造的限界を定量的・定性的に明らかにすることにある．

2. 関連研究

2.1 物体中心学習の発展

Slot Attention[Locatello 20] は，入力特徴から K 個のスロットへの競合的バインディングにより教師なしの物体分離を実現した．SAVi[Kipf 22] は GRU ベースの Slot Predictor を導入し，動画における時間的一貫性を確保した．SAVi++[Elsayed 22] はさらに深度情報と Transformer ベースの予測器を採用し，より複雑なシーンへの対応を図っている．

2.2 特徴再構成型アプローチ

DINOSAUR[Seitzer 23] は，ピクセル再構成に代えて凍結 DINOv2[Oquab 24] の特徴マップを再構成ターゲットとし，実画像でのロバストな物体発見を実現した．VideoSAUR[Zadaianchuk 23] はその動画拡張であり，optical flow ベースの対応損失を併用する．いずれの手法も DINOv2 が標準バックボーンとして採用されている．

2.3 バックボーンの空間弁別能力

Lüddecke & Ecker[Lüddecke 25] は，密な予測タスクに対するバックボーンの適性を体系的に評価し，DINOv2 がインスタンス認識において他のバックボーンを凌駕することを示した．CLIP はグローバルな対照学習で訓練されるため，空間的弁別能力が DINOv2 の約 1/10 に留まる [Yang 24]．DINOv2 には Register Tokens に関連する高ノルムアーティファクトの問題が指摘されているが [Darcet 24]，物体中心学習には最も適したバックボーンとされている．

3. 手法

3.1 モデル構成

本研究のモデルは，DINOSAUR[Seitzer 23] を SAVi[Kipf 22] の動画拡張フレームワークに統合したものである（図 1）．構成は以下の 4 段階からなる：(1) 凍結 ViT バックボーンからの特徴抽出（$F \in \mathbb{R}^{B\times 384\times 16\times 16}$），(2) 2 層 MLP（LayerNorm–Linear–ReLU–Linear, 384→64 次元）による射影，(3) Slot Attention（K=5, 64 次元, 3 回反復），(4) Broadcast Decoder による特徴再構成．

デコーダの出力マスクは温度 τ 付き Softmax で正規化される：

$$\alpha_k = \mathrm{Softmax}(m_k/\tau), \quad \hat{F} = \sum_{k=1}^{K} \alpha_k \cdot \hat{F}_k \tag{1}$$

3.2 損失関数

バックボーン間の特徴量スケール差に対処するため，標準的な MSE 損失に加え，チャンネル正規化損失を導入した．各空

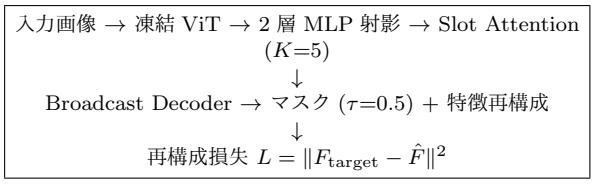

図 1: SAVi-DINOSAUR の構成．凍結 ViT で抽出した特徴を 2 層 MLP で射影し，Slot Attention でスロットに分解する．デコーダは各スロットから特徴とマスクを生成し，温度付き Softmax で統合する．

表 1: バックボーン比較．DINOv1 は標準 MSE では発散するため，チャンネル正規化損失を使用した．CLIP は射影バグ修正後の結果．

モデル	損失種別	FG-ARI(↑)	Full-ARI(↑)
DINOv2 ViT-S/14	MSE	**0.165**	**0.073**
DINOv1 ViT-S/16	ch-norm	0.153	0.047
CLIP ViT-B/16	MSE+det.	0.041	−0.026

間位置の特徴ベクトルに LayerNorm を適用してから MSE を計算する：

$$L_{\text{cn}} = \frac{1}{HW} \sum_i \|\text{LN}(\hat{y}_i) - \text{LN}(y_i)\|^2 \tag{2}$$

これにより，特徴の方向情報を保存しつつスケール差を吸収できる．コサイン類似度損失 [Lüddecke 25] と類似するが，空間構造をより直接的に保存する点が異なる．

3.3 実験設定

データセット：MOVi-A[Greff 22] のサブセット（60 サンプル：金属物体 20 + 混合 40），各 24 フレーム ×224×224 ピクセル．

バックボーン：DINOv2 ViT-S/14（384 次元），DINOv1 ViT-S/16（384 次元），CLIP ViT-B/16（768→384 次元に射影）．いずれもバックボーンは凍結し，射影以降のみ学習する．

訓練：バッチサイズ 2，学習率 0.0004（5 エポックウォームアップ + コサイン減衰），Adam 最適化，200 エポック（CLIP のみ 50 エポック）．

評価指標：FG-ARI（Foreground Adjusted Rand Index），Full-ARI，Mask Similarity（スロット間コサイン類似度; 低いほど分化が進む）．

ハードウェア：Intel Core Ultra 285K, NVIDIA RTX 5090 (32GB), 128GB RAM.

4. 実験結果

4.1 バックボーン間の分離性能

表 1 に 3 種のバックボーンの比較結果を示す．DINOv2 が最も高い FG-ARI（0.165）を達成し，チャンネル正規化損失を適用した DINOv1（0.153）が続いた．CLIP はバグ修正（4.4 節）後も 0.041 にとどまった．

DINOv1 は標準 MSE 損失では訓練損失 4.692 で発散したため，チャンネル正規化損失を適用した．特徴ベクトルの標準偏差は DINOv1: 4.0，DINOv2: 2.4 であり，チャンネル正規化損失適用後の訓練損失は 0.163 まで低下した．

表 2 に材質別 FG-ARI を示す．DINOv1 の Metal/Non-metal 差（0.002）は DINOv2（0.039）よりはるかに小さい．

表 2: 材質別 FG-ARI．DINOv1 は Metal/Non-metal 差が最小であり，材質によらない均一な分離を示す．

モデル	Metal	Non-metal	差
DINOv2	0.185	0.146	+0.039
DINOv1 (ch-norm)	0.154	0.152	+0.002
CLIP (det.)	0.048	0.034	+0.014

表 3: 温度 τ によるスロットマスクの分化度合い．τ=0.5 で Mask Similarity が最も改善されるが，τ=0.3 では過剰な先鋭化により逆効果となる．

τ	Mask Sim.(↓)	改善率
1.0（デフォルト）	0.723	—
0.7	0.642	11%
0.5	0.558	23%
0.3	0.621	14%

4.2 温度スケーリングによるマスク分化

デコーダのマスク Logits の標準偏差は 0.031 であり，Softmax 出力はほぼ一様分布（$\approx 1/K = 0.2$）であった．温度 τ によるスケーリングの結果を表 3 に示す．τ=0.5 で最良の Mask Similarity（0.558, 23%改善）を達成した．τ=0.3 では性能が悪化した．

4.3 射影層構成とスロット崩壊

特徴射影に Single Linear 層（384→64）を用いた場合，射影後の特徴分散が 0.001 まで圧縮され，全 5 スロットのマスクが同一に崩壊した（スロット崩壊）．2 層 MLP 構成では分散は 0.15（150 倍）に改善され，Mask Similarity も 0.866 から 0.723 へ低下した（表 4）．

4.4 CLIP 射影層における勾配漏洩

CLIP バックボーンでは 768 次元を 384 次元に変換する学習可能な射影層が必要となる．この射影出力がそのまま再構成ターゲットに使用されるため，MSE 損失の勾配がターゲット自体を崩壊させる問題が発生した（修正前損失: −0.005）．`target_feat.detach()` の適用により勾配漏洩を遮断し，訓練損失は 0.036 に改善した．なお，CLIP の空間弁別能力（Spatial std: 0.17）は DINOv2 の約 1/10 であり，FG-ARI は 0.041 にとどまった．

5. 定性分析：構造的限界の可視化

定量指標だけでは捉えきれない DINOSAUR の構造的限界を，出力マスクの可視化により分析する．本節が本研究の中心的な貢献である．

5.1 パッチ解像度によるマスク境界のにじみ

図 2 に DINOv2（最良モデル）のスロットマスク出力を示す．DINOv2 ViT-S/14 は 224×224 の入力を 14×14 のパッチに分割するため，マスクの実効解像度は 16×16 ピクセルである．Ground Truth のピクセル単位の物体輪郭に対し，出力マスクではパッチ境界にスナップされた粗い領域となり，境界が「にじむ」現象が確認できる．また，16×16 マスクを 224×224 に Bilinear 補間でアップサンプリングする過程で，物体の鋭い輪郭が失われている．

表 4: 射影層構成の影響．Single Linear では特徴分散がほぼ消失し，スロット崩壊が発生する．

射影層	特徴分散	Mask Sim.(↓)
Single Linear (384→64)	0.001	0.866
2 層 MLP (384→384→64)	0.15	0.723

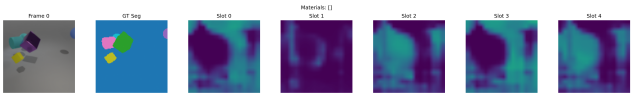

図 2: DINOv2 によるスロットマスクの出力例．物体の大まかな位置は捉えているが，マスク境界は 16×16 パッチ単位でしか区切れず，GT の鮮明な輪郭に対してにじみが顕著である．

5.2 鏡面反射によるスロット混同のメカニズム

図 3 に 3 種バックボーンの出力を並べて示す．DINOv2 は物体の概形を捉えているが，金属物体の表面ではスロット割り当てに混乱が観察される．具体的には，(1) ハイライト領域が背景と同じスロットに割り当てられる**過少分割**と，(2) 同一物体がハイライトの有無で複数スロットに分割される**過剰分割**の 2 種類が確認された．DINOv1 も類似の傾向を示し，CLIP では分離自体が不成立であった．

5.3 スロットマスクの分化状況の分析

図 4 に DINOv2 モデルのスロットマスクの詳細を示す．上段は各スロットのマスクを入力画像に重畳したオーバーレイ，下段はマスク値そのものを可視化している．各スロットの平均マスク値（μ）と標準偏差（σ）が極めて近い値を示しており，スロット間の分化が不十分であることが確認できる．

6. 考察

6.1 バックボーン性能差の解釈

DINOv2 ＞ DINOv1 ＞ CLIP という序列は先行研究 [Lüddecke 25] の予測と整合的である．DINOv1 がチャンネル正規化損失により DINOv2 に肉薄する FG-ARI（0.153）を達成した事実は，特徴量スケール差が MSE の二乗特性で増幅されることが性能差の主因であったことを示す．また，DINOv1 の材質間差（0.002）が DINOv2（0.039）よりはるかに小さい点は注目に値する．DINOv1 の空間弁別能力（Spatial/Channel 比: 0.740 vs DINOv2: 0.599）が材質に依存しない均一な分離に寄与していると考えられる．

6.2 設計要素の知見

表 5 に各検証項目と先行研究との関係を整理する．2 層 MLP 射影の必要性は DINOSAUR[Seitzer 23] の実装に暗黙的に含まれるが，Single Linear との定量的比較（分散 150 倍差）は報告されておらず，実装時に見落としやすい重要な設計要素である．温度 τ=0.5 の選定も，Logits 標準偏差（0.031）の実測に基づく知見である．CLIP 射影の勾配漏洩は，学習可能な射影層の出力をそのまま再構成ターゲットに使用する際に，MSE 損失の勾配がターゲット自体をゼロに崩壊させる構造的問題であり，DINOv1/v2 ではバックボーン凍結により顕在化しない．これらはいずれも既存文献で明示的に報告されていない．

6.3 構造的限界の考察

§ 5 で観察された 2 つの構造的限界について考察する．

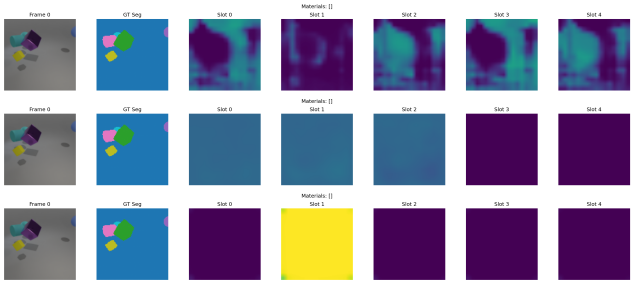

図 3: 3 種バックボーンのスロットマスク比較（MOVi-A 金属物体）．DINOv2（上段）は物体構造を概ね捉えるが，金属反射面でスロットの混同が見られる．DINOv1（中段, ch-norm 損失）も類似の傾向を示す．CLIP（下段）は空間弁別能力の不足により分離がほぼ不成立である．各行は左から入力フレーム，GT，スロット 1–5 のマスクを示す．

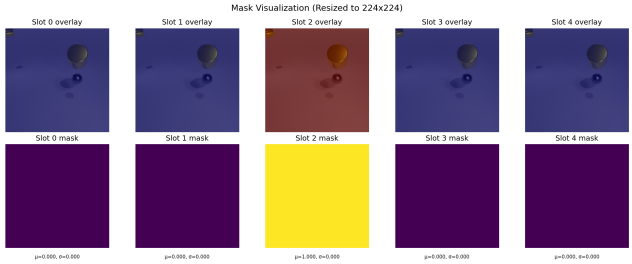

図 4: DINOv2 のスロットマスク詳細（MOVi-A 金属物体）．上段：各スロットマスクの入力画像への重畳．下段：マスク値の可視化と統計（μ, σ）．全スロットのマスク値がほぼ均一であり，スロット間の分化が不十分であることを示す．

パッチ解像度の壁:16×16 のマスク解像度は ViT のパッチ分割に起因する本質的制約であり，Bilinear 補間によるアップサンプリングでは物体の鋭い輪郭を回復できない．CRF（Conditional Random Field）等の後処理や SAM[Kirillov 23] との統合が必要である．

鏡面反射によるスロット混同：Slot Attention は各パッチの特徴とスロットの内積で排他的割り当てを決定するため，物体表面の特徴が空間的に一貫している（Lambert 反射的な拡散反射）場合には有効に機能する．しかし金属物体では BRDF の Specular 成分が視点依存の鋭いハイライトを生じさせ，同一物体上で全く異なる特徴ベクトルが生成される．この Lambert 反射の暗黙的前提がスロット混同の根本原因であり，View-Dependent Decoder[Smith 24] のような反射モデルの明示的導入が必要と考えられる．

マスク Logits の分散の小ささ：Logits の標準偏差が 0.031 と極めて小さく，Softmax 後のマスクが一様分布に近い状態にある．温度 τ=0.5 は部分的に緩和するが，Logits の分散自体を拡大するにはより深いデコーダ構成や Transformer Decoder[Elsayed 22] の採用が考えられる．

6.4 限界

データ規模：60 サンプルでの実験であり，結果の一般化には留意が必要である（DINOv2 FG-ARI 標準偏差: ±0.186）．**解像度制約**：16×16 パッチに起因するマスクのにじみは，現行 ViT アーキテクチャの本質的制約であり，後処理なしでは高精度なセグメンテーションは困難である．**反射モデルの欠如**：Lambert 反射を暗黙に仮定する特徴再構成損失は，鏡面反射

表 5: 各検証項目の位置づけ．

検証項目	先行研究との関係
2 層 MLP 射影の必要性	暗黙の実装詳細 [Seitzer 23]．定量比較は本研究
温度 τ=0.5	一般的手法．Logits 統計に基づく選定は本研究
CLIP 勾配漏洩	既存報告なし
ch-norm 損失による DINOv1 改善	既存報告なし

物体でスロット混同を引き起こす．

7. まとめ

本研究では，DINOSAUR アーキテクチャを 3 種の ViT バックボーンで鏡面反射物体に適用し，その性能と構造的限界を検証した．以下に主要な知見を整理する．

(1) 何を検証したか：DINOv2, DINOv1, CLIP の 3 種の ViT バックボーンを用いた DINOSAUR の鏡面反射物体（MOVi-A 金属サブセット, 60 サンプル）への適用可能性を，FG-ARI と出力マスクの可視化により定量的・定性的に評価した．

(2) 何が分かったか：DINOv2 が最も高い FG-ARI（0.165）を達成し，物体中心学習のバックボーンとしての優位性を確認した．一方，バックボーン間の特徴量スケール差が損失関数を通じて学習の安定性に深刻な影響を与えることも明らかになった．チャンネル正規化損失の導入により DINOv1 の FG-ARI は DINOv2 に肉薄する 0.153 に到達した．

(3) 今後の課題：本研究で明らかになった構造的限界は大きく 2 点に集約される．第一に，**16×16 パッチ解像度の壁**であり，CRF 後処理や SAM[Kirillov 23] 等の高解像度モデルとの統合が必要である．第二に，**鏡面反射への対応**であり，Lambert 反射を暗黙に前提とする現行の特徴再構成損失では金属物体のスロット混同を解消できず，View-Dependent Decoder[Smith 24] や 3 次元反射モデルの導入が求められる．

参考文献

[Locatello 20] Locatello, F., Weissenborn, D., Unterthiner, T., et al.: Object-Centric Learning with Slot Attention, in *Proc. NeurIPS 2020* (2020).

[Kipf 22] Kipf, T., Elsayed, G. F., Mahendran, A., et al.: Conditional Object-Centric Learning from Video, in *Proc. ICLR 2022* (2022).

[Elsayed 22] Elsayed, G. F., Mahendran, A., van Steenkiste, S., et al.: SAVi++: Towards End-to-End Object-Centric Learning from Real-World Videos, in *Proc. NeurIPS 2022* (2022).

[Seitzer 23] Seitzer, M., Horn, M., Zadaianchuk, A., et al.: Bridging the Gap to Real-World Object-Centric Learning, in *Proc. ICLR 2023* (2023).

[Oquab 24] Oquab, M., Darcet, T., Moutakanni, T., et al.: DINOv2: Learning Robust Visual Features without Supervision, *TMLR* (2024).

[Darcet 24] Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision Transformers Need Registers, in *Proc. ICLR 2024* (2024).

[Lüddecke 25] Lüddecke, T. and Ecker, A. S.: Characterizing Vision Backbones for Dense Prediction with Dense Attentive Probing, *TMLR* (2025).

[Yang 24] Yang, J., Luo, K. Z., Li, J., et al.: Denoising Vision Transformers, in *Proc. ECCV 2024* (2024).

[Greff 22] Greff, K., Belletti, F., Beyer, L., et al.: Kubric: A Scalable Dataset Generator, in *Proc. CVPR 2022* (2022).

[Zadaianchuk 23] Zadaianchuk, A., Seitzer, M., Martius, G.: Object-Centric Learning for Real-World Videos by Predicting Temporal Feature Similarities, in *Proc. NeurIPS 2023* (2023).

[Smith 24] Smith, C., et al.: Unsupervised Object-Centric Fields, *arXiv preprint* (2024).

[Kirillov 23] Kirillov, A., Mintun, E., Ravi, N., et al.: Segment Anything, in *Proc. ICCV 2023* (2023).