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 —

坂口 健（電気通信大学）　中世 大雄（東京大学）

1. 背景と目的

- **モデル**：DINOSAUR（凍結 ViT ＋ Slot Attention）
- **目的**：鏡面反射物体への適用可能性の検証
- **課題**：バックボーン選択の影響の解明
- **データ**：MOVi-A（金属 20 ＋混合 40 ＝ 60 件）

2. 3 種のバックボーン比較と定量結果

モデル	損失種別	FG-ARI (↑)	Metal FG-ARI
DINOv2 ViT-S/14	MSE	**0.165**	0.185
DINOv1 ViT-S/16	ch-norm	0.153	0.154
CLIP ViT-B/16	MSE+detach	0.041	0.048

⇒ **DINOv2 が最高精度**（FG-ARI 0.165）　⇒ DINOv1 はスケール差で破綻 → **ch-norm 損失**で解決（0.153）
⇒ CLIP は空間弁別能力の不足により 0.041

3. 構造的限界の可視化（定性分析）★本研究の中心的貢献

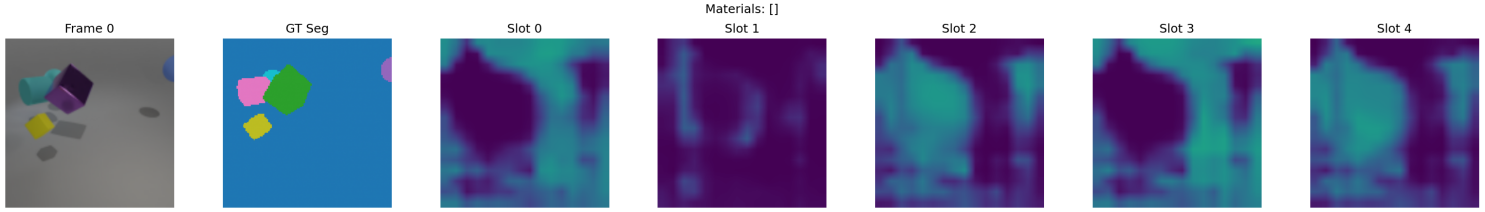

図 1: DINOv2 スロットマスク — 16×16 パッチによる境界のにじみ

限界 1
14px パッチ起因の
16×16 解像度の壁
→ CRF・SAM 等が必要

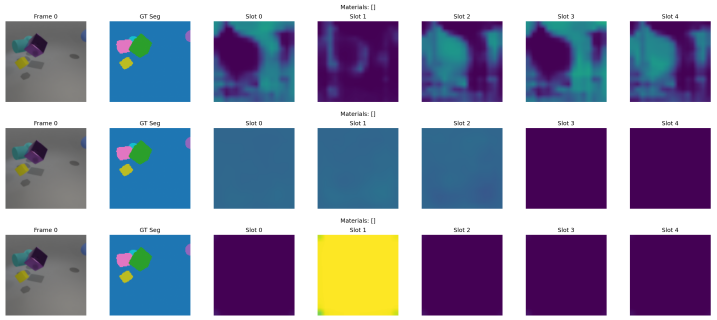

図 2: 3 種バックボーン比較（MOVi-A 金属物体）— 金属面での**スロット混同**

限界 2
BRDF Specular 成分が
スロット混同を誘発
→ 視点依存デコーダ必要

4. まとめと今後の課題

(1) DINOv2 のバックボーンとしての優位性を確認　**(2)** 16×16 解像度の限界 → SAM 等の高解像度モデルとの統合が必要　**(3)** 鏡面反射への対応 → 3 次元反射モデルの導入が必要