DEEP LEARNING JP
[DL Papers]

Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks (CoRL 2024)

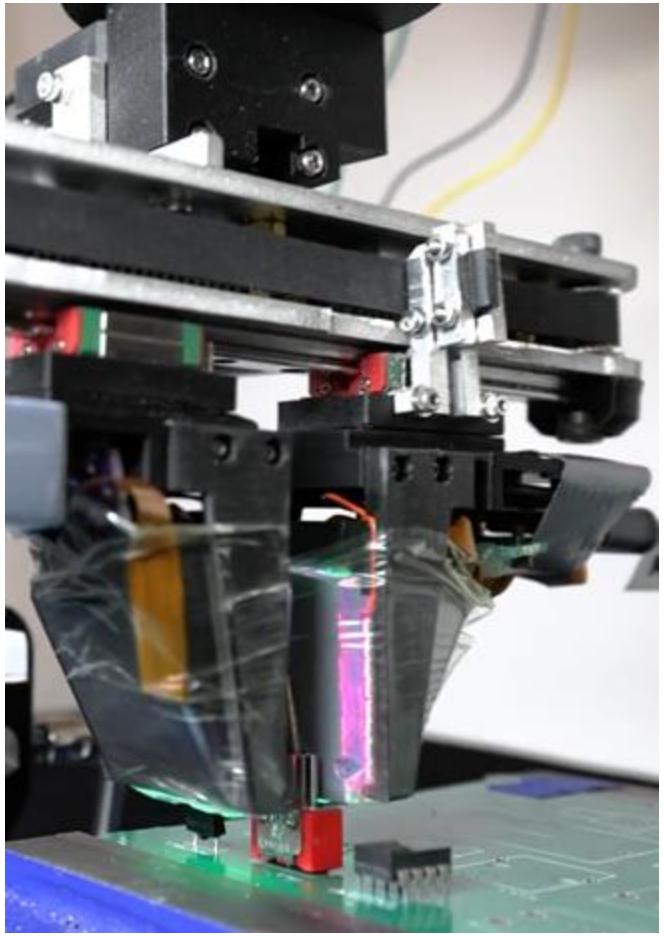
Tatsuya Kamijo, Matsuo-Iwasawa Lab, M1

書誌情報

題名	**Transferable Tactile Transformers for Representation Learning Across Diverse Sensors and Tasks**
著者	Jialiang Zhao, Yuxiang Ma, Lirui Wang, Edward H. Adelson
所属	MIT CSAIL
会議	CoRL 2024
概要	- 光学式触覚センサは形状や構造が多様，公開データも少ない - 13種類11タスク300万枚の触覚データセット「**FoTa**」を公開 - 異なるセンサーやタスク間で転移可能な共通触覚表現学習のフレームワーク「**T3**」を提案

概要

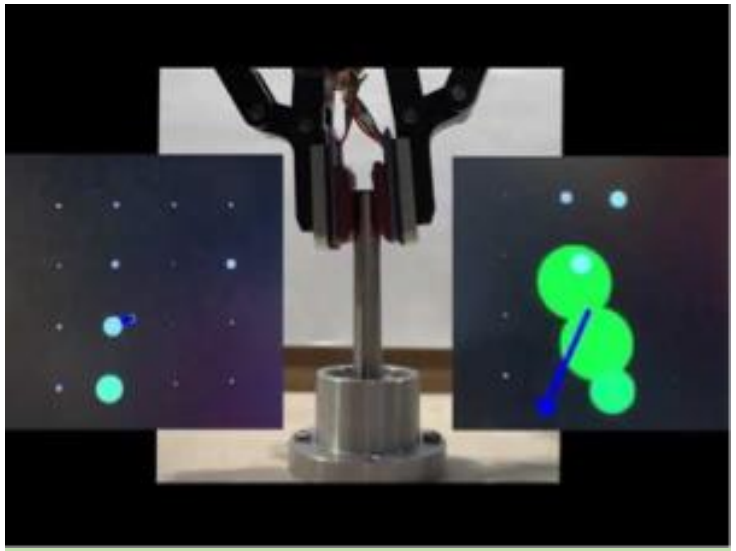
異種センサ・タスク間で共有可能な触覚の表現学習を提案



背景

視覚モダリティと異なり触覚は多様なデータ形式

従来は**一つのセンサ**で**単一のタスク**を**固有のEncoder / Decoder**で解くのが主流
学習時と違うセンサを使うと著しく性能が低下：**汎用的なエンコーダが欲しい**

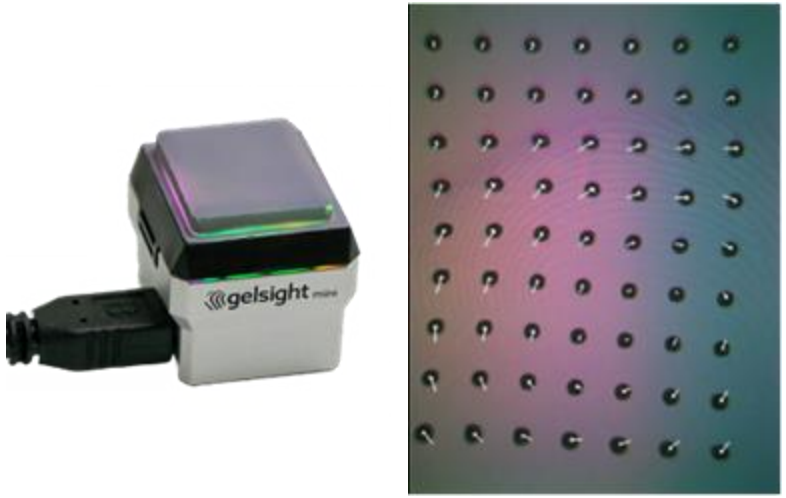


uSkin, XELA Robotics

3軸×16
(4, 4, 3)

DIGIT, Meta

RGB
(320, 240, 3)



GelSight Mini

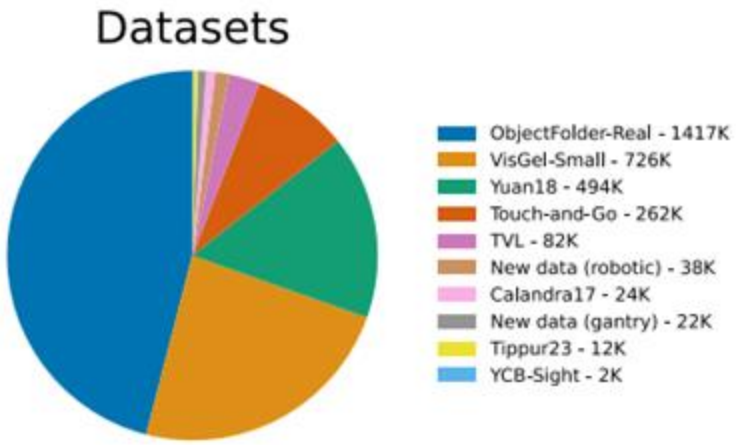
RGB / Flow
(320, 240, 3) / (h, w, 2)

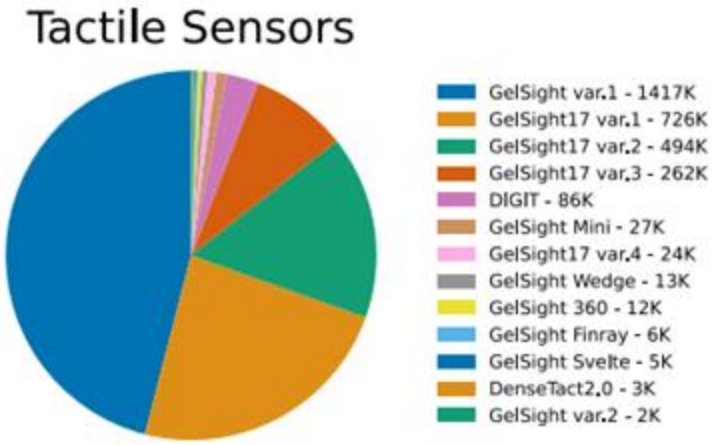
Foundation Tactile (FoTa)

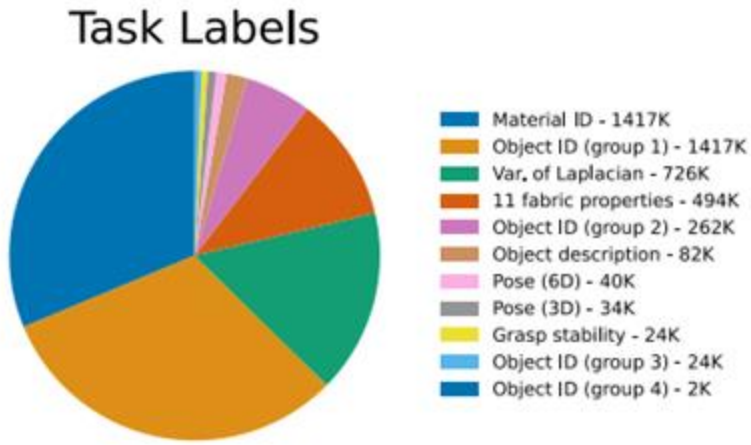
13センサ・11タスクからなる300万枚の触覚画像データセット

- 13種類の光学式触覚センサ，11タスクで収集
  - 既存データセットの組み合わせ＋独自収集
- 計3,083,452データ（最大）
- WebDatasetフォーマットで統一
  - jpg + json

FoTa Data





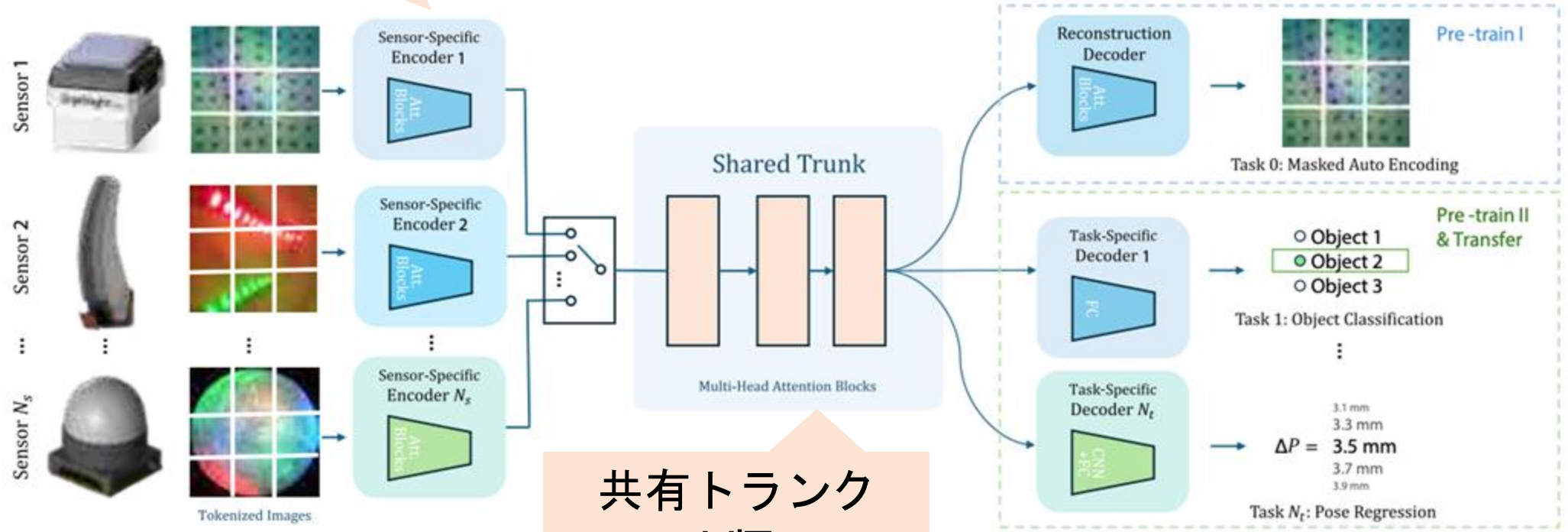


Transferable Tactile Transformers (T3)

センサ固有のエンコーダ＋共有トランクで共通の埋め込み表現を獲得

センサ固有のエンコーダ
ViT

タスク固有のデコーダ
ViT(再構成), MLP(回帰・分類)，ResNet + MLP(姿勢推定)

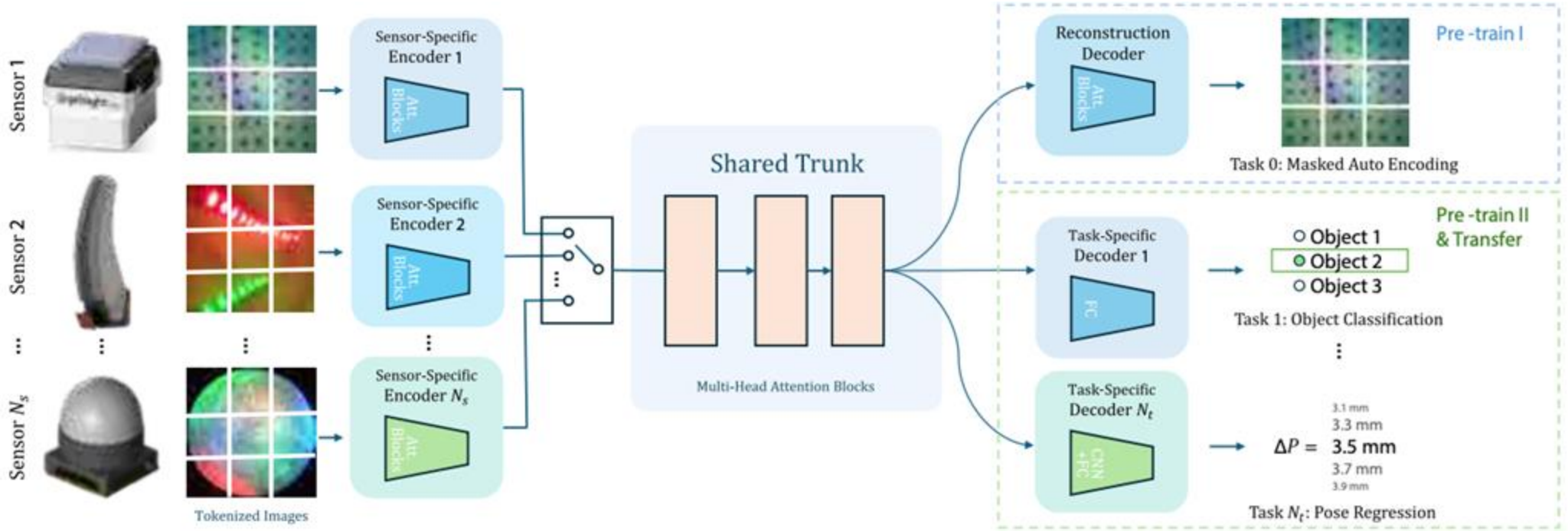


共有トランク
ViT

Transferable Tactile Transformers (T3)

二段階の事前学習とファインチューニング (optional)で学習

- Pre-training Ⅰ: MAEで自己教師あり学習．ピクセルレベルの理解が目的.
- Pre-training Ⅱ: 各タスクのラベルで教師あり学習．意味的な理解が目的.
- Fine-tuning: 特定のセンサ・タスクペアでfine-tune．手元環境での性能向上.



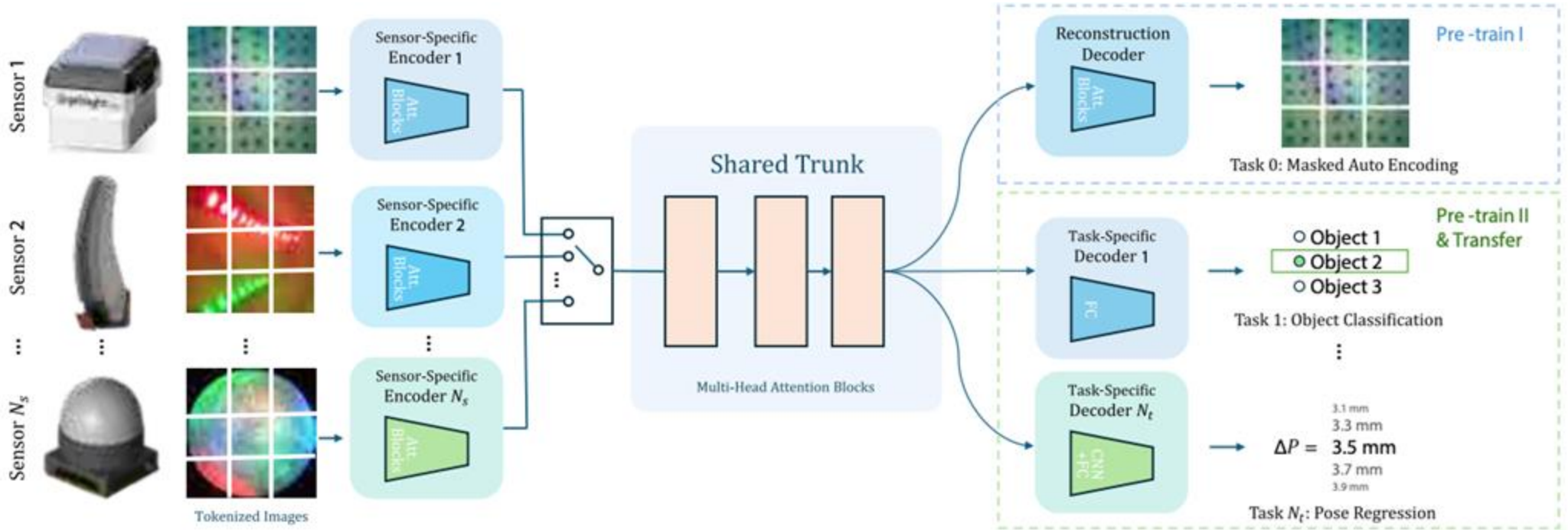
Transferable Tactile Transformers (T3)

姿勢推定タスクでは２つの触覚画像 [X^1, X^2] の埋め込み表現をconcat

目的関数

$$loss(X_i, Y_j) = L_j(Y_j, Dec_j(Trunk(Enc_i(X_i))))$$

$$loss([X_i^1, X_i^2], Y_j) = L_j(Y_j, Dec_j(Trunk(Enc_i(X_i^1)) \oplus Trunk(Enc_i(X_i^2))))$$



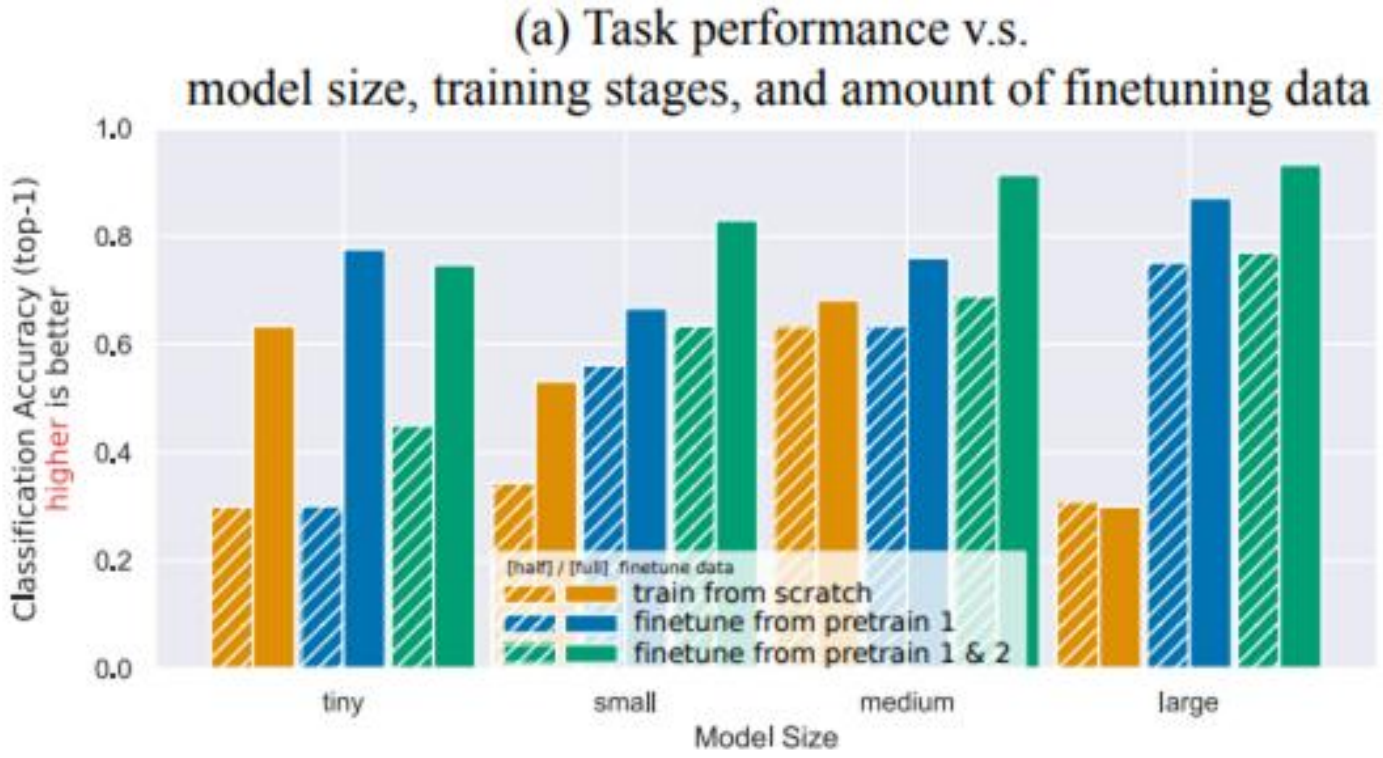
実験

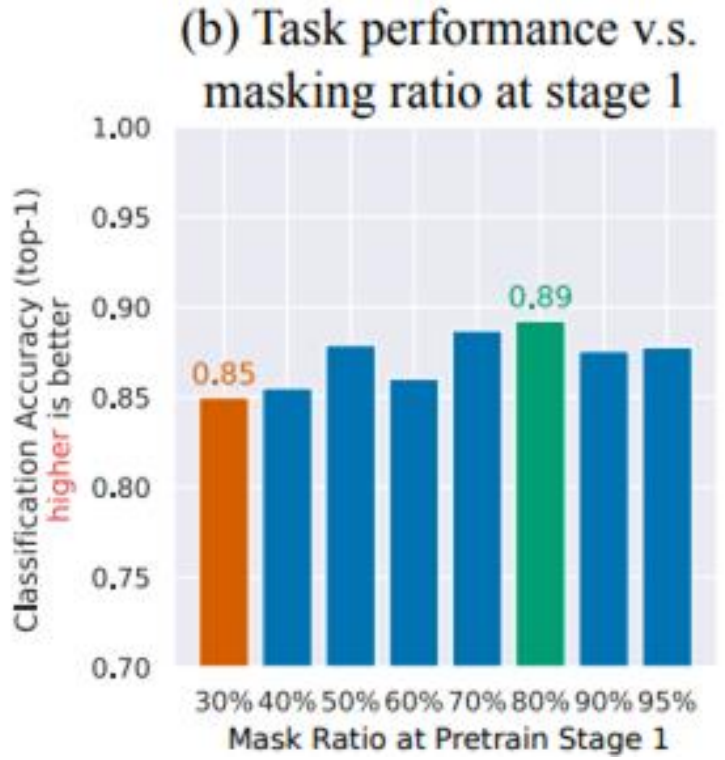
- T3事前学習の効果は？
  - **学習段階**（Scratch / Pre-training Ⅰ / Ⅰ＆Ⅱ），**モデルサイズ**，ファインチューニングに使う**データ数**，**マスク率**でablation
- T3の未知のセンサ・タスクに対するゼロショット転移性能は？
- T3の埋め込み表現は長期的なマニピュレーション（挿入作業等）で有用か？

実験：事前学習 vs Scratch

事前学習の効果は？： Scratch < Pre-training Ⅰ < Pre-training Ⅰ＆Ⅱ

- 6種類物体の分類タスク
  - 2種類の触覚センサ，各3300データで学習
- MAEのマスク率は80%で性能が最も高い

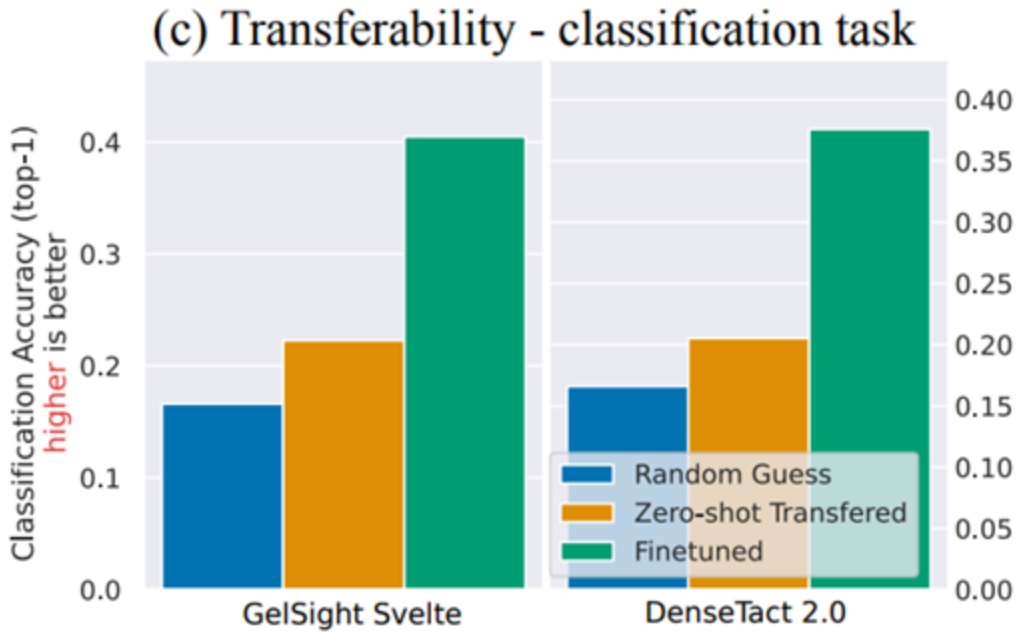


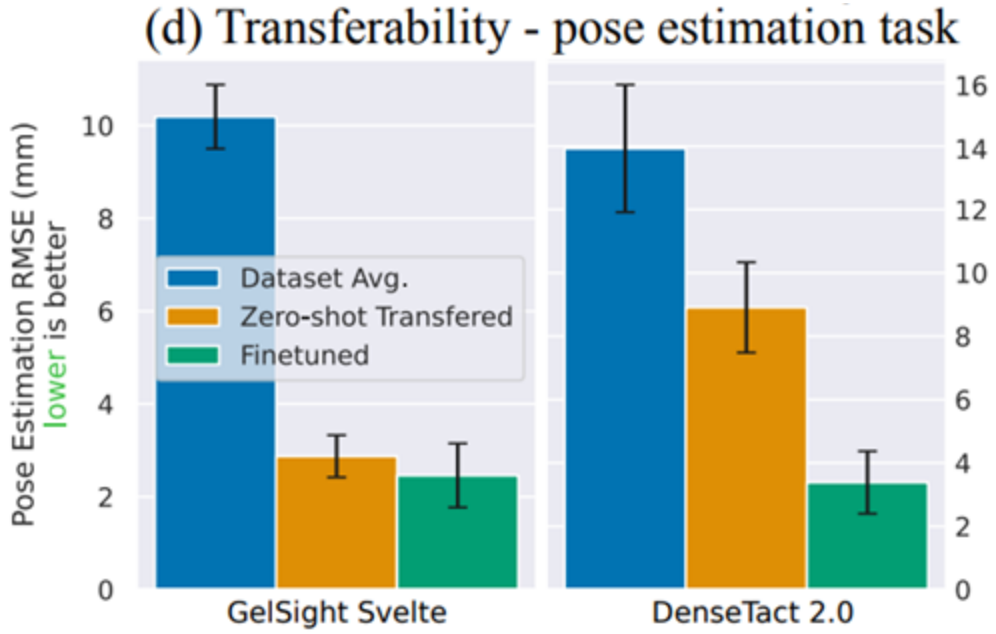


実験：T3ゼロショット転移性能

事前学習されたT3は新しいセンサに汎化する？：ゼロショットは△，fine-tuneで◯

- 物体分類タスク＆姿勢推定タスクの２つのタスクで検証
- これらのタスクはGelSight Wedge, GelSight Finray, GelSight Miniで事前学習
- 新しい（未知の）センサ：DenseTact2.0, GelSight Stelve
  - エンコーダにはそれぞれ似ているGelSight Mini，GelSight Wedgeの物を利用

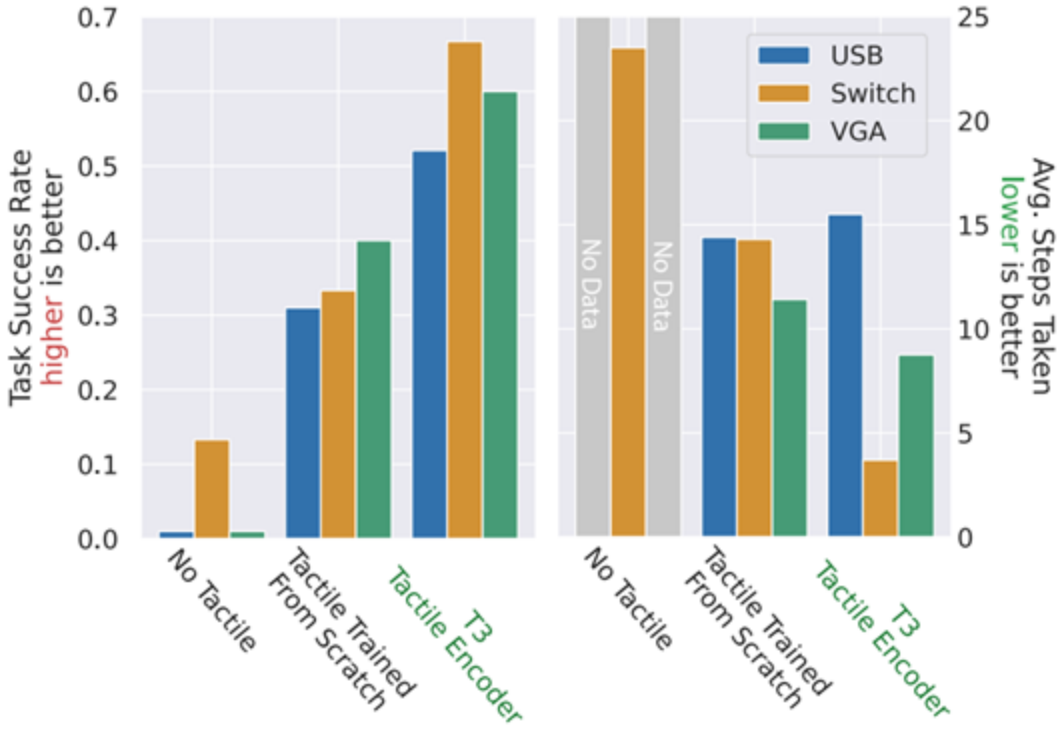




実験：挿入タスクへの応用

T3は長期タスクでの触覚エンコーダとして使えるか？

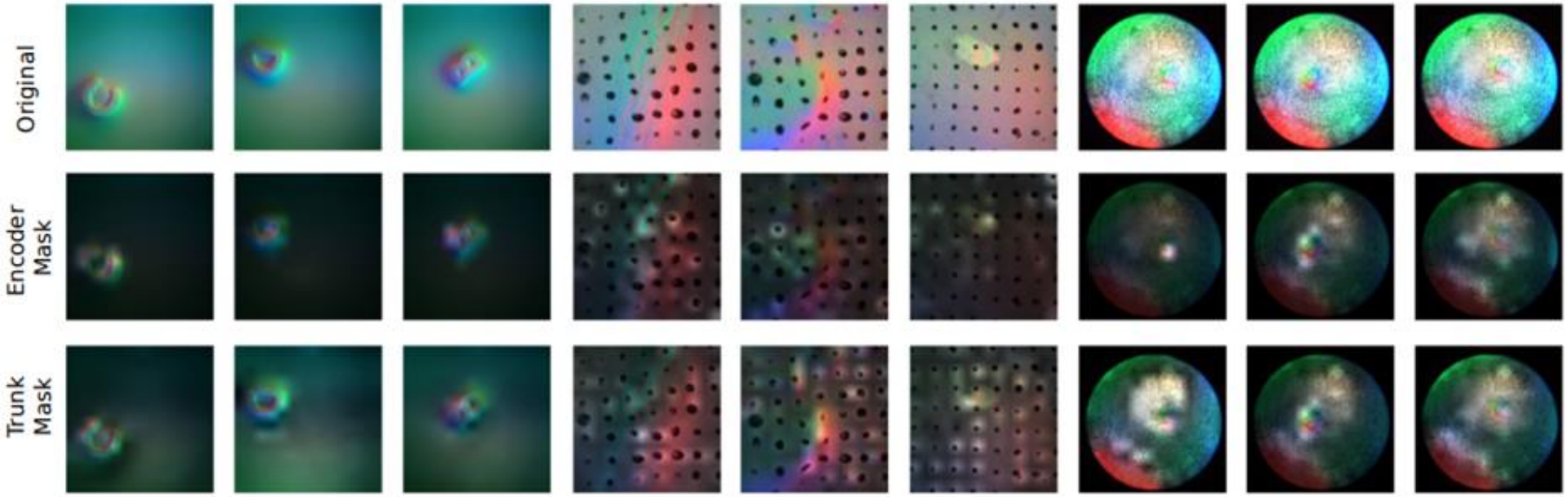
- 3つの部品（12-pin USB, 3-pin Toggle Switch, 17-pin VGA）の**挿入タスク**で検証
  - クリアランス（遊び）は**0.4mm**：視覚のみでは難しい**サブミリの精密挿入**
- 模倣学習（BC）で学習された3つの方策の性能を比較
  - 触覚**なし**
  - **スクラッチ学習NN**で触覚エンコード
  - **T3**で触覚エンコード
- 結果：
  - 触覚なしだとほぼ成功しない
  - T3がもっとも成功率が高い



Appendix: Attention Mapの可視化

Embeddingの定性的理解：Encoder・Trunkはそれぞれどこを見ている？

- Encoder Mask: **主に接触領域**にattentionがかかっている
- Trunk Mask: 接触領域に限らず**幅広く**attentionがかかっている
- 理解：Encoderは単に触覚画像の特徴的な部分を見ているが，**Trunkはセンサ固有の特徴（マーカー等）も処理**して共有表現にするため，幅広い範囲を見ている



まとめ

- 光学式触覚センサは多様で，従来センサ・タスクペアで固有の表現を得ていた
- 13種類11タスクで収集された300万枚の触覚データセット**FoTa**を公開
- センサ固有のエンコーダ・**共有トランク**・タスク固有のデコーダからなる**T3**は共有トランクにより**センサによらない共通埋め込み表現**を得る
- 実験によりFoTaで事前学習されたT3の有効性を確認

感想

- アイデアは直感的で正しそうだが，事前学習とScratchで性能大して変わらない
  - スライド10ページの(a), 論文のFigure 3.a.
- （Trunkの）モデルをでかくすると性能は上がる一方推論は遅くなる
  - 挿入タスクの制御周期は2Hz（！）
  - ダイナミックなコンタクトリッチマニピュレーションでは使えない遅さ
- 事前学習Ⅰだけの時とⅠとⅡ両方行った時の分類性能もあまり大差ない
  - 論文中ではⅠが局所的なピクセルレベルの理解，Ⅱが大局的な意味的な理解を担うと書いてある
  - それぞれの学習段階でのattention mapも見てみたい
- 主に物体分類タスクで実験している（FoTaもほぼ物体分類と姿勢推定）
  - 実際欲しいのは力に関する情報（force estimation, slip detection)
  - だが，ラベル付きデータを大量に集めるのが難しい
    - 力覚センサが必要・触覚センサ表面のエラストマーが破れやすい

感想

- CV分野由来の技術（ViT, MAE等）をそのまま触覚に適用
  - LLM，CV分野での知見をロボティクスに適用して成果が出るロボット学習分野の流れの一部
  - 非常に強力な一方，触覚が必要なコンタクトリッチタスクでは力のフィードバックによるダイナミックな制御が必要となることが多いが，準静的なタスクしか考慮していない・制御周期が遅すぎるなど，ロボティクスの観点では改善の余地が大きい
  - 良いロボット学習の研究をするためにはCV・LLM等の学習分野と伝統的なロボット制御分野両方の知見が必要となっている