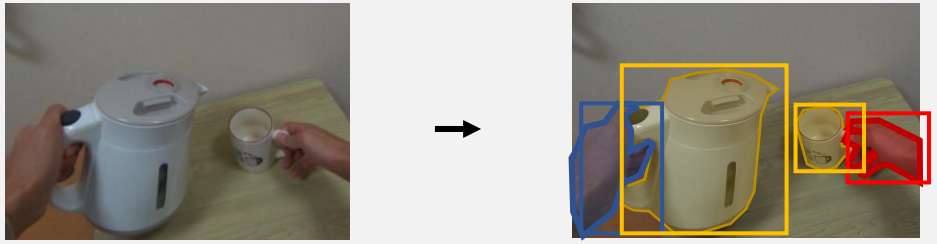
[IS2-1-17] 運動情報を用いた手およびその接触物体の弱教師ありセグメンテーション

八木拓真（東大） 佐藤洋一（東大）

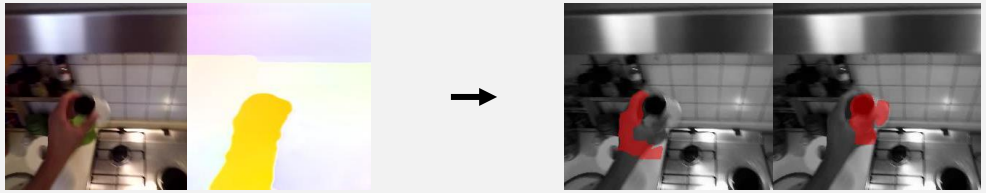
問題：手と接触物体のセグメンテーション

- 手操作：環境の状態を変化させる行動
- 接触判定や形状復元のため接触物体領域が欲しい
- 課題：アノテーションコストが高く、
  物体種類が多いため手動アノテーションは難しい



アプローチ

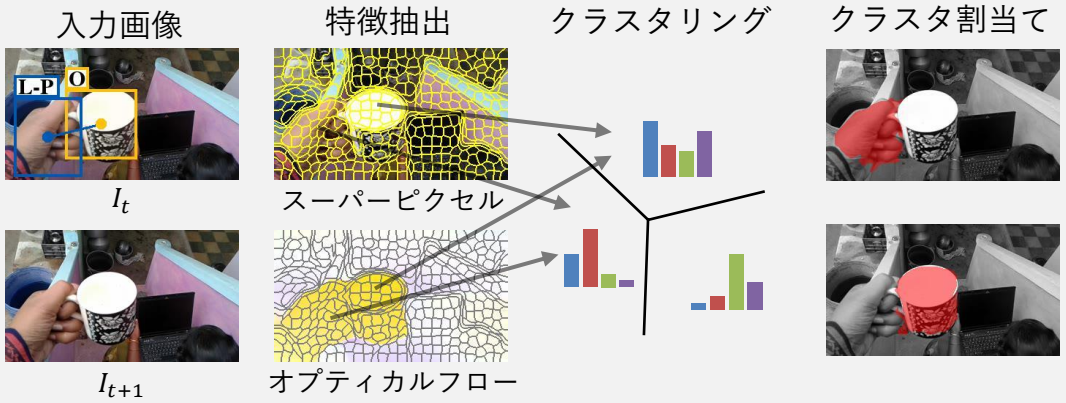
- 映像中の手運動が前景抽出の手掛かりになる
- バウンディングボックス単位のアノテーションから
  疑似教師セグメンテーションを生成し訓練に利用



画像+オプティカルフロー　　疑似教師マスク

アルゴリズム

色とフローのヒストグラムをsuperpixel単位で分割



実験結果（抜粋）

	手		物体	
	AP	AP50	AP	AP50
GrabCut	35.9±0.4	89.0±0.4	26.4±0.2	53.9±0.5
Color	40.0±0.4	96.5±0.2	25.4±0.7	56.4±0.5
提案手法	46.8±0.6	97.3±0.4	26.5±0.4	57.6±0.5

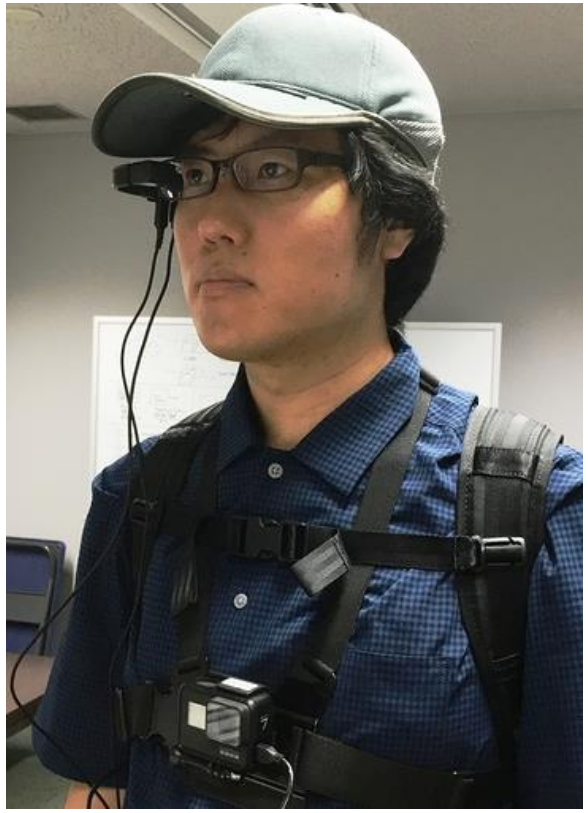
±は3試行の絶対平均偏差を表す

今後の展開

- 信頼する教師情報の動的な選択
- 得られたセグメンテーションからの接触判定

背景：一人称ビジョン

- 身体にウェアラブルカメラを装着（頭、首、胸、手首など）
- 撮影された一人称視点映像は装着者の行動や興味を反映
- 手操作を高解像度で記録可能



一人称視点映像における手解析

- 手操作は環境と相互作用する主要な手段の1つ
- これまで手の位置の検出や手に関わる行動分類などが取り組まれてきた

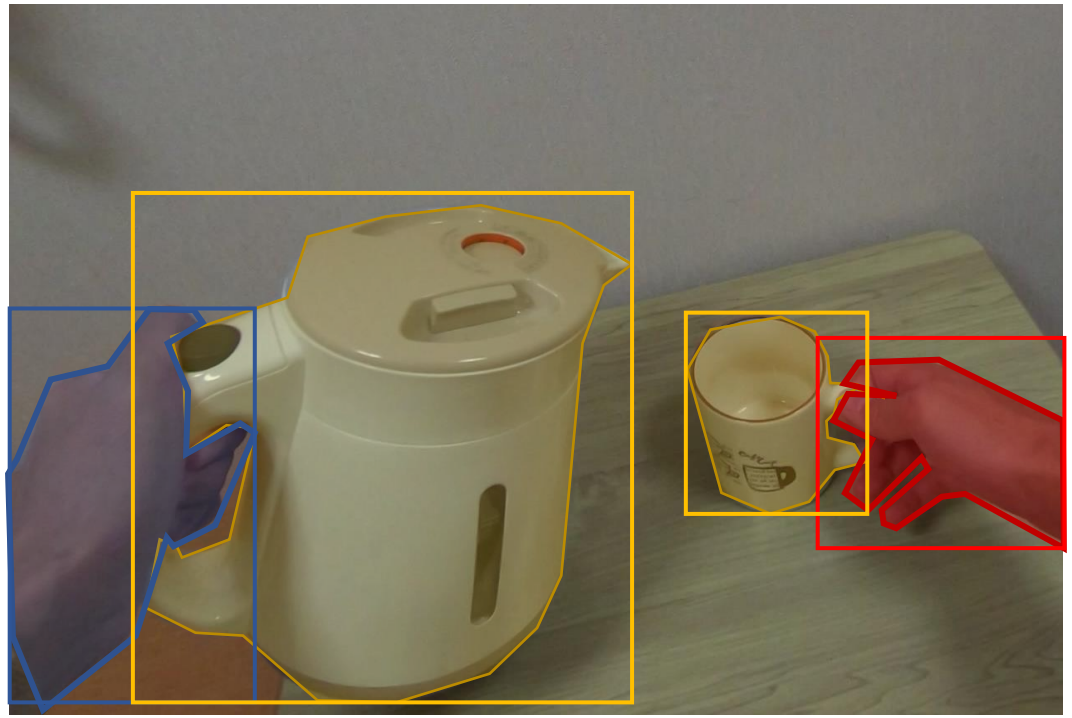
手位置検出

手-物体インタラクションの認識

問題設定：手と接触物体の検出およびセグメンテーション

- 一人称視点の画像から手および接触物体を検出する
- 接触状態等を判定するためには検出のみならず領域の推論も必要
  - 例：把持しているか否か？

入力：一人称視点画像



出力：手と接触物体の
位置およびセグメンテーション

課題

- **アノテーションコスト：**セグメンテーションの収集は容易でない
- **多種類の物体：**カテゴリーを事前に定義できない



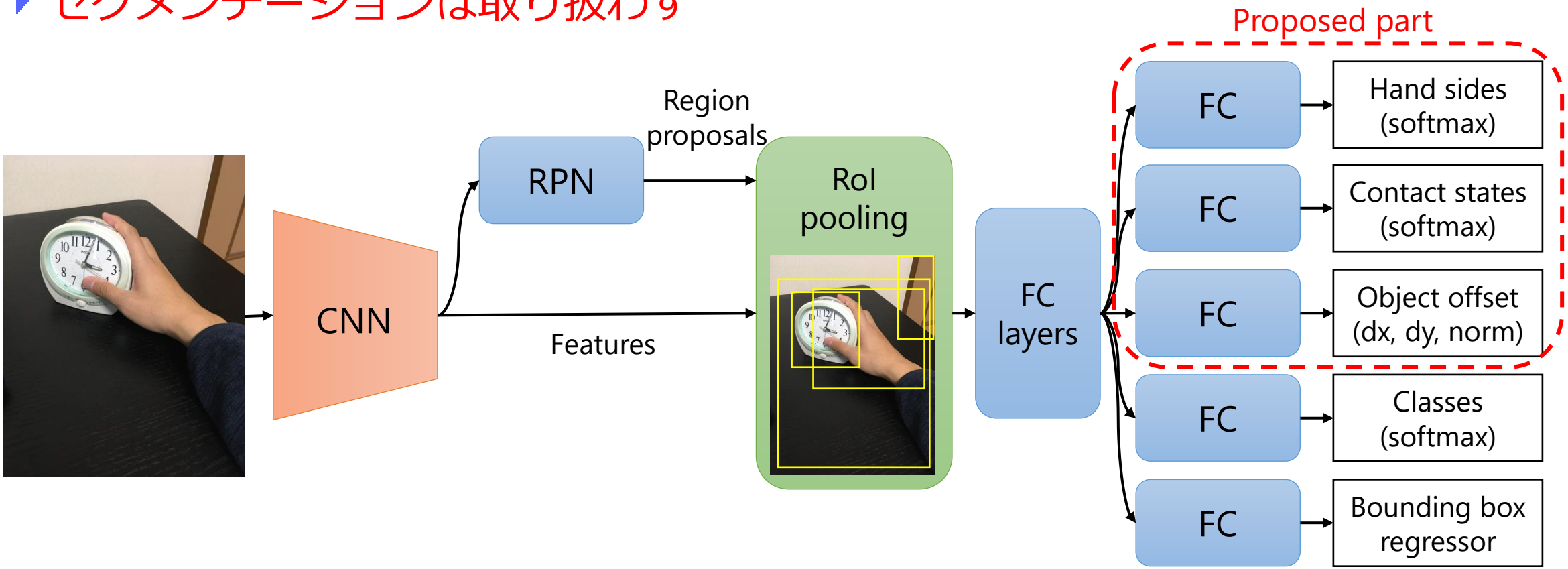


スケーラブルかつクラス非依存のアプローチが必要

関連研究

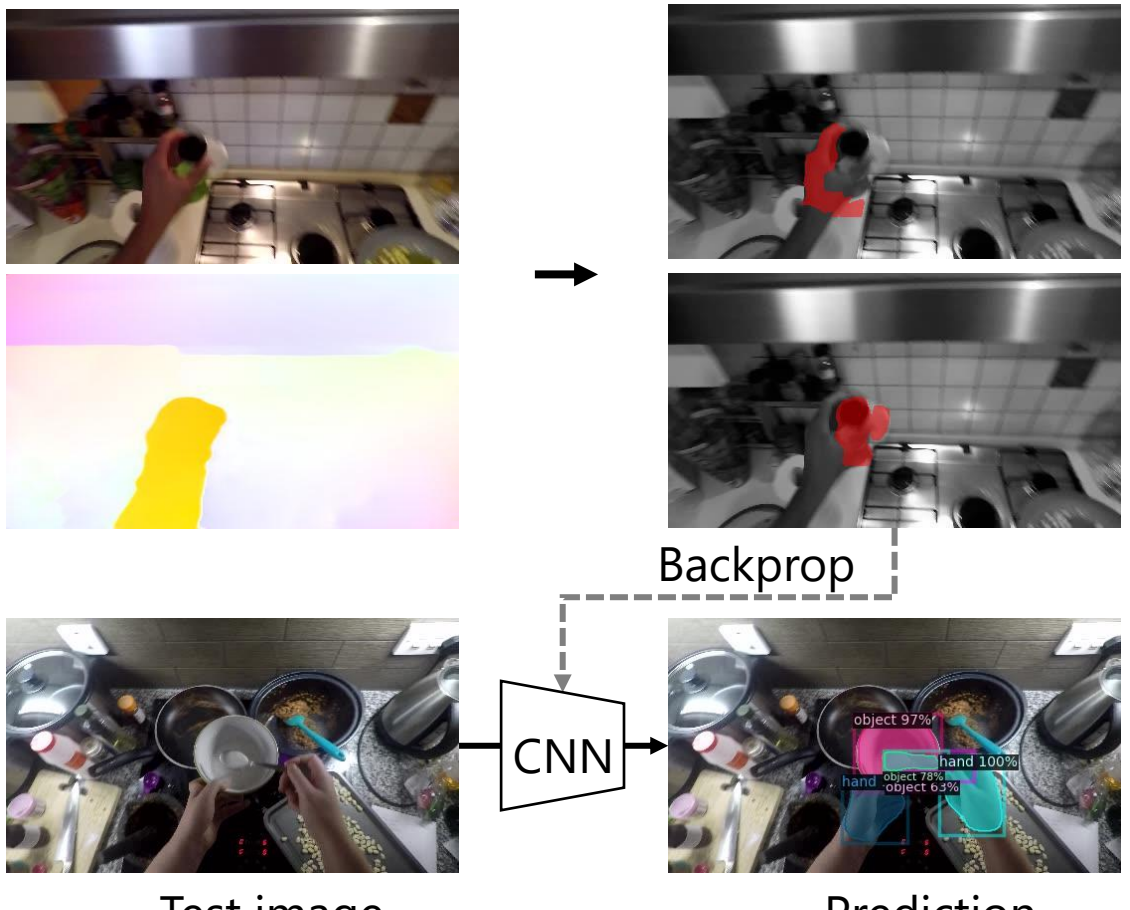
Detecting hands and objects in contact in-the-wild [Shan+, CVPR'20]

- 10万枚の画像に手と接触物体の位置/手の左右/接触状態のアノテーションを付与
- Faster R-CNN に新規のブランチを追加しマルチタスク学習
- セグメンテーションは取り扱わず



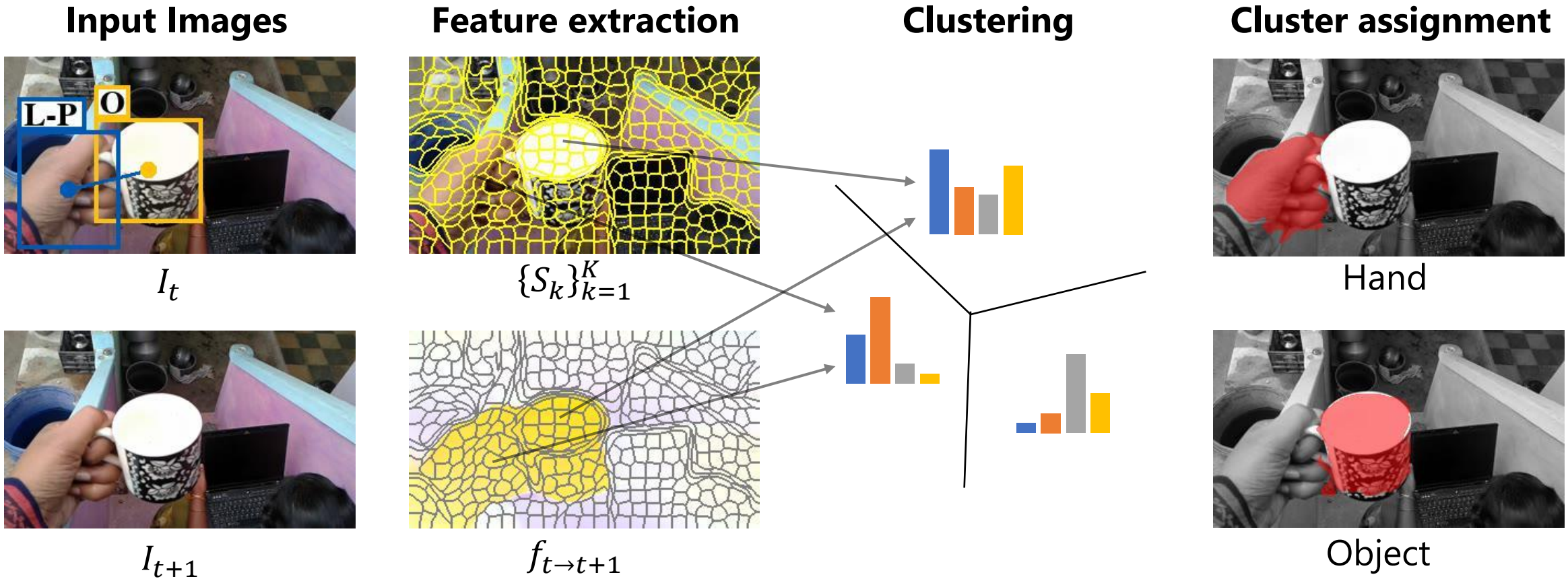
アプローチ：疑似マスクによる訓練

- 観察：手と把持物体は併せて移動する（場合がある）
- 低レベルの手掛かりから手と物体のおおよその領域を推論し疑似教師として与えられるのではないか？



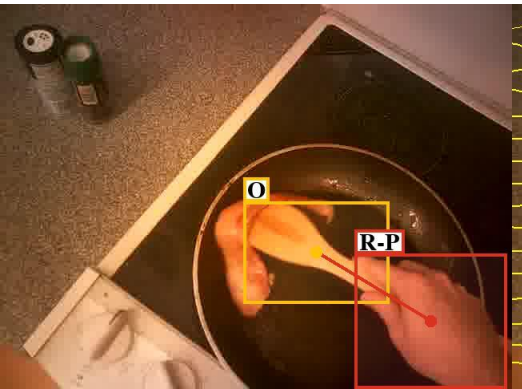
アルゴリズム

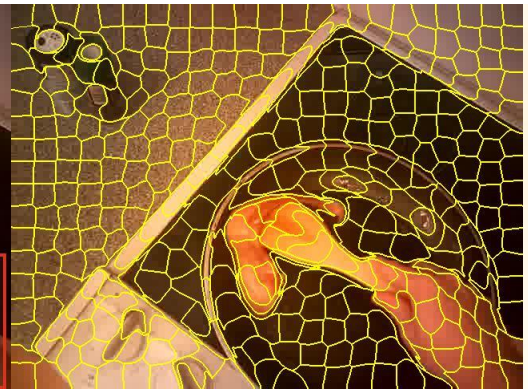
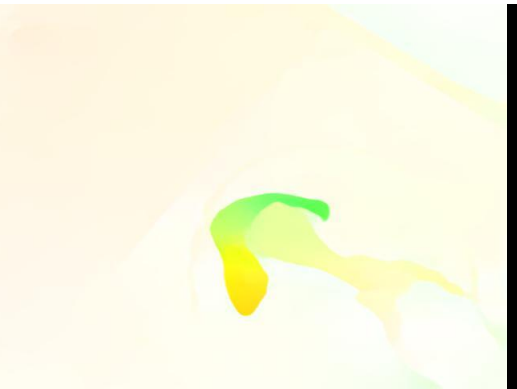
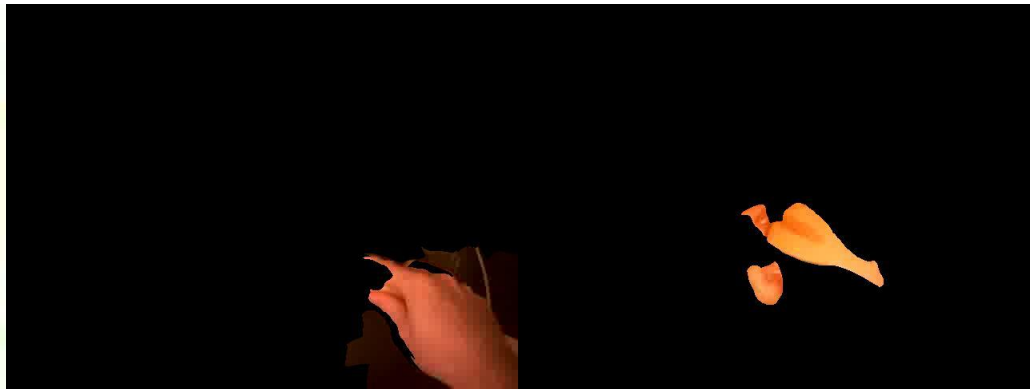
- 入力：2枚の連続画像とバウンディングボックス
- スーパーピクセル単位でのヒストグラムを構成し手・物体・背景の3クラスに分割



疑似マスク生成（前景-背景分離）の詳細

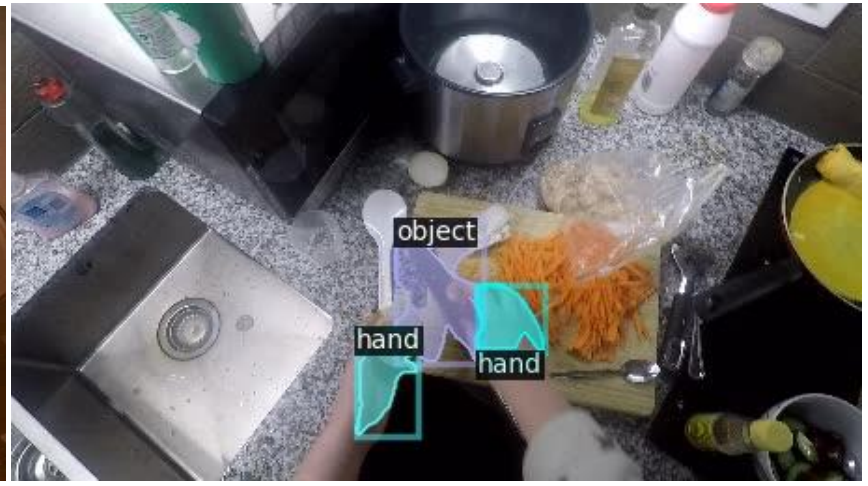
- 前提：スーパーピクセル（SLIC）とオプティカルフロー（FlowNet2）
- Step 1: オプティカルフローのヒストグラムによって2-meansまたは3-means
- もし手と物体が両方前景の場合
  - Step 2：カラー特徴も加えて3-means（前景（手）1、背景（物体）2クラスタ）
- もし手のみが前景の場合
  - 前景/背景について、バウンディングボックス周辺のsuperpixelを背景として除去
- K-means法のcentroidはバウンディングボックス情報を使って初期化
  - バウンディングボックスに完全に含まれるsuperpixelは信頼できる
- 物体がない場合は、素朴に2-meansを行う



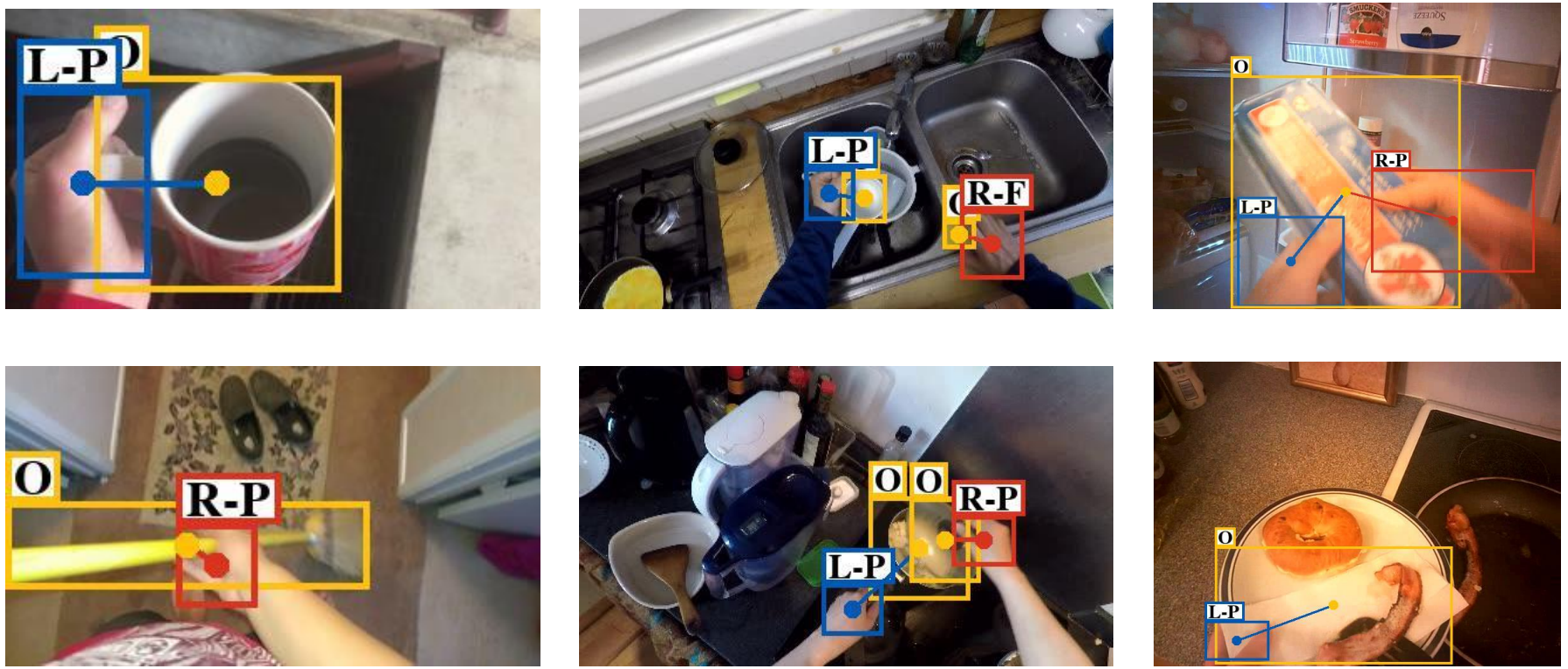
疑似マスクの生成結果

- 前景が複数色ある場合等に苦戦



データセット

- DOH-egoデータセット[Shan+, CVPR'20]を使用
  - Charades-ego [Siguardsson+, CVPR'18]、 EPIC-KITCHENS [Damen+, ECCV'18]、 EGTEA [Li+, ECCV'18] の各データセットに手と物体の検出アノテーションを付与したもの
- セグメンテーションの評価のため著者で500枚独自にアノテーションを実施



Charade-ego **EPIC-KITCHENS** **EGTEA**

実験設定

バックボーンネットワーク

- Mask R-CNN [He+, ICCV'17] を使用

比較手法

- Box：検出した矩形領域を予測とする
- GrabCut：カラー情報に基づく前景抽出手法
- Color：運動情報を除外した提案手法

評価指標

- Mask AP（IoUの閾値を0.50-0.95まで変化させた際の平均AP）
- Mask AP50（IoUを閾値が0.50以上の場合のAP）

https://cocodataset.org/#detection-eval

実験結果

- 手セグメンテーションにおいて提案手法が有効であることを確認
  - 教師あり学習と比べると未だ差が大きい
- 物体セグメンテーションにおいては定量的には大きな差が出ず

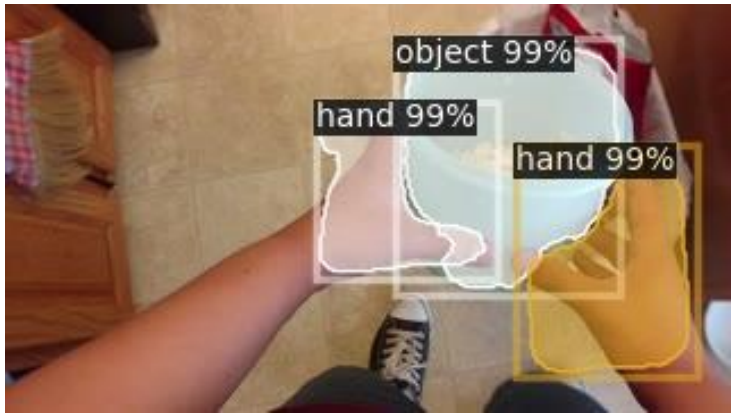
	手		物体	
	AP	AP50	AP	AP50
		検出+後処理		
Box	14.6	66.1	11.0	37.9
GrabCut	27.2	73.6	19.7	44.7
		疑似マスクによる訓練		
GrabCut	35.9±0.4	89.0±0.4	26.4±0.2	53.9±0.5
Color	40.0±0.4	96.5±0.2	25.4±0.7	56.4±0.5
提案手法	46.8±0.6	97.3±0.4	26.5±0.4	57.6±0.5
		教師あり学習（EGTEAデータセットの手マスク1.4万枚を追加使用）		
DOH-ego+EGTEA	64.7	98.2	N/A	N/A

±は3試行の絶対平均偏差を表す

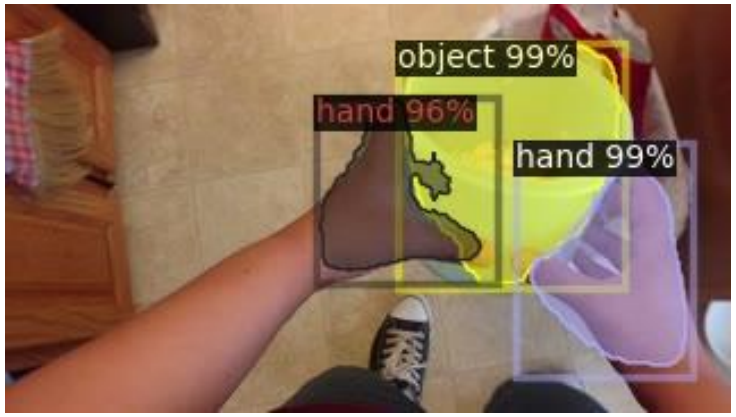
予測例の比較

- 提案手法の方が境界領域の推論が優れていることを確認

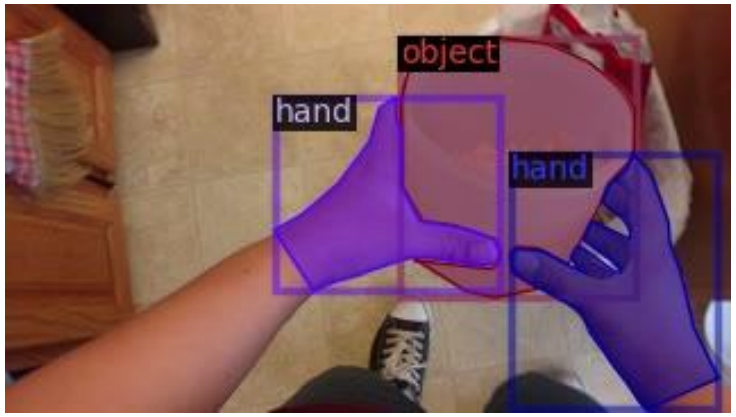
GrabCut



Color



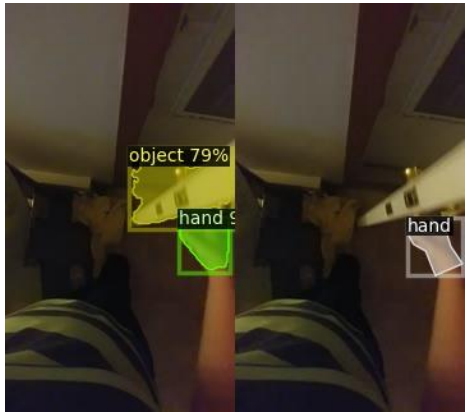
提案手法

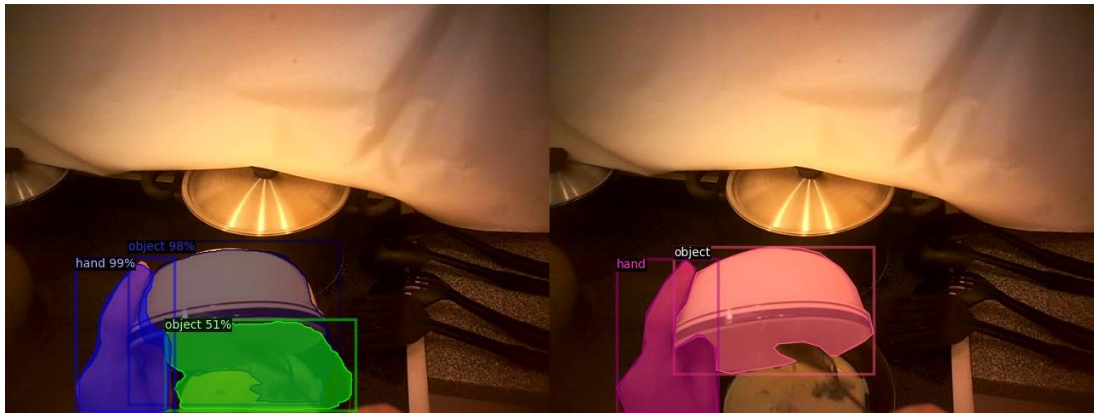


正解

失敗例（左：予測、右：正解）

- 手で分断される物体、領域推論等に課題





失敗例（左：予測、右：正解）

- 特に、細長い物体に脆弱