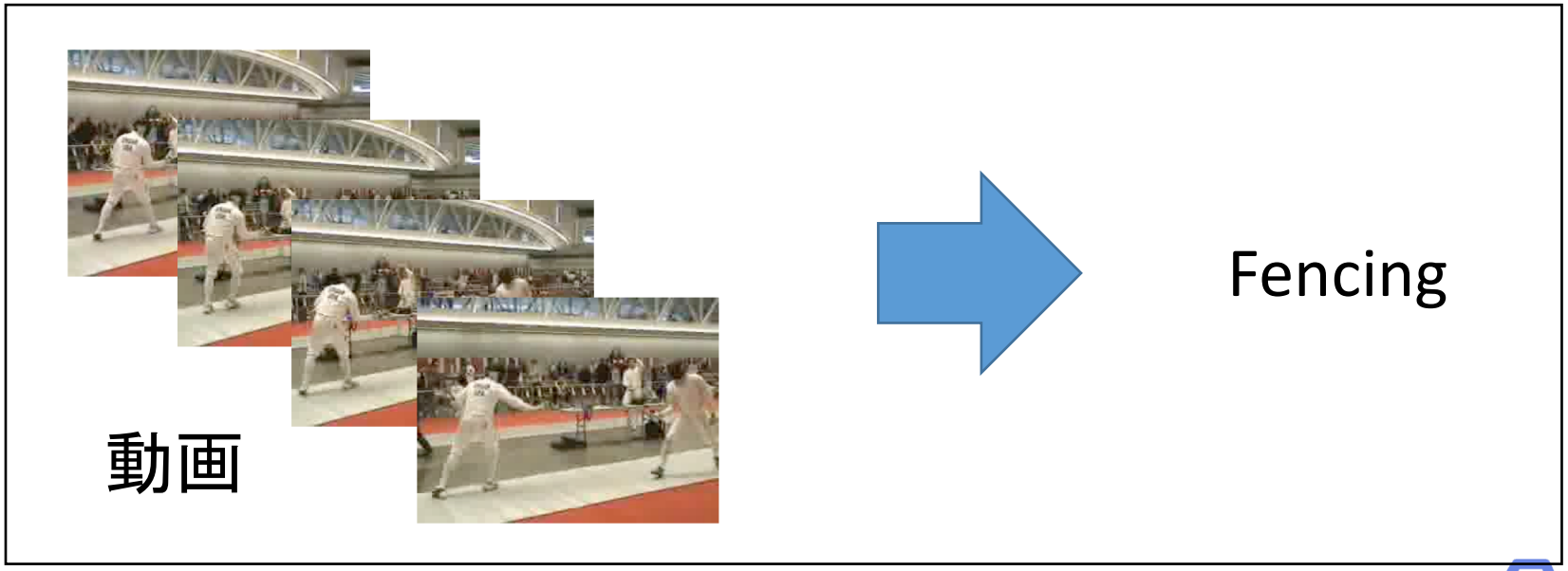
PRMU2022

Transformerを用いた人物行動検出

水野 颯介　柳井啓司
電気通信大学　大学院　情報学専攻

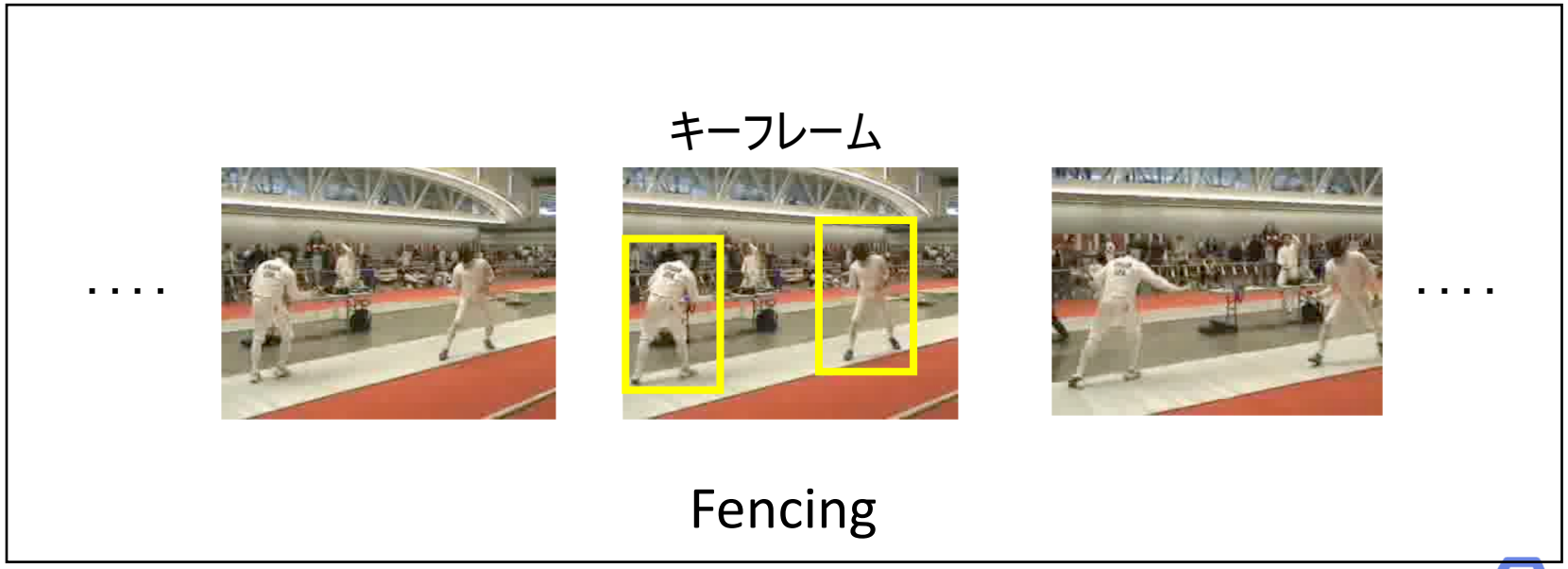
1. はじめに

- これまでに，深層学習を用いた「**行動認識**」の研究が広く行われてきた
  → 近年，技術発展と共に「**行動検出**」のタスクが注目されている



1. はじめに

- **行動検出(Action Detection)**
  - 行動クラスを認識
  - キーフレーム内のアクターの検出



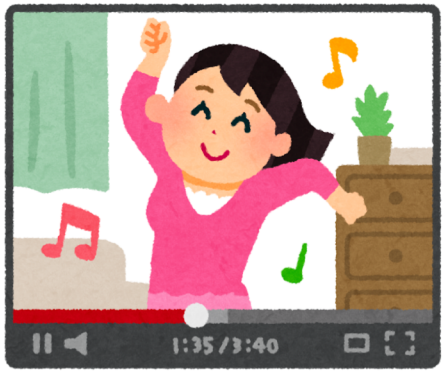
1. はじめに

行動検出

応用可能

異常検知

WEB動画の自動分類&収集



1. はじめに

- 既存の行動検出の研究では，CNNをベースにした手法がほとんど

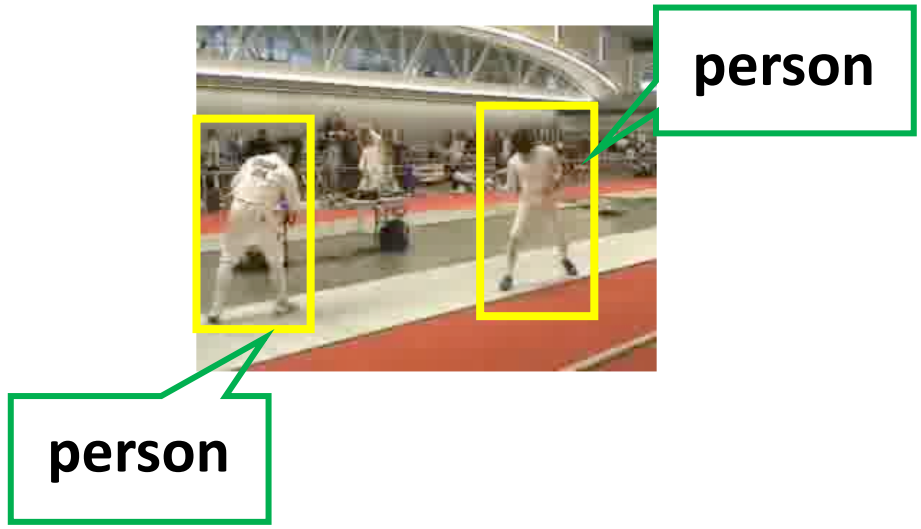
- 近年のコンピュータビジョン
  - 自然言語処理モデルのTransformerをベースにしたモデル

提案内容

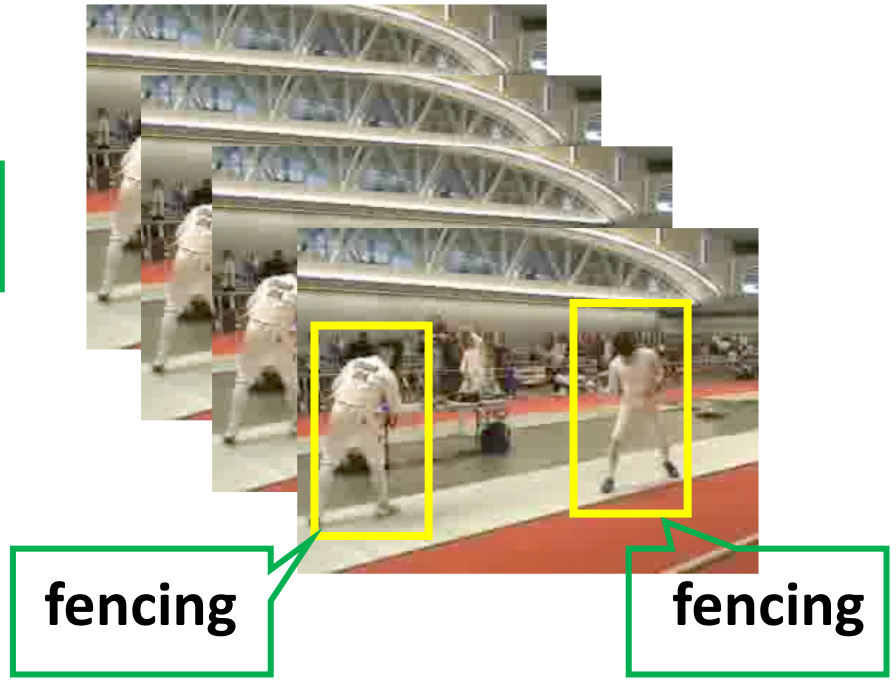
- Transformerをベースにした**行動検出**手法

2. 関連研究 – 物体検出

物体検出



行動検出



2. 関連研究 – 物体検出

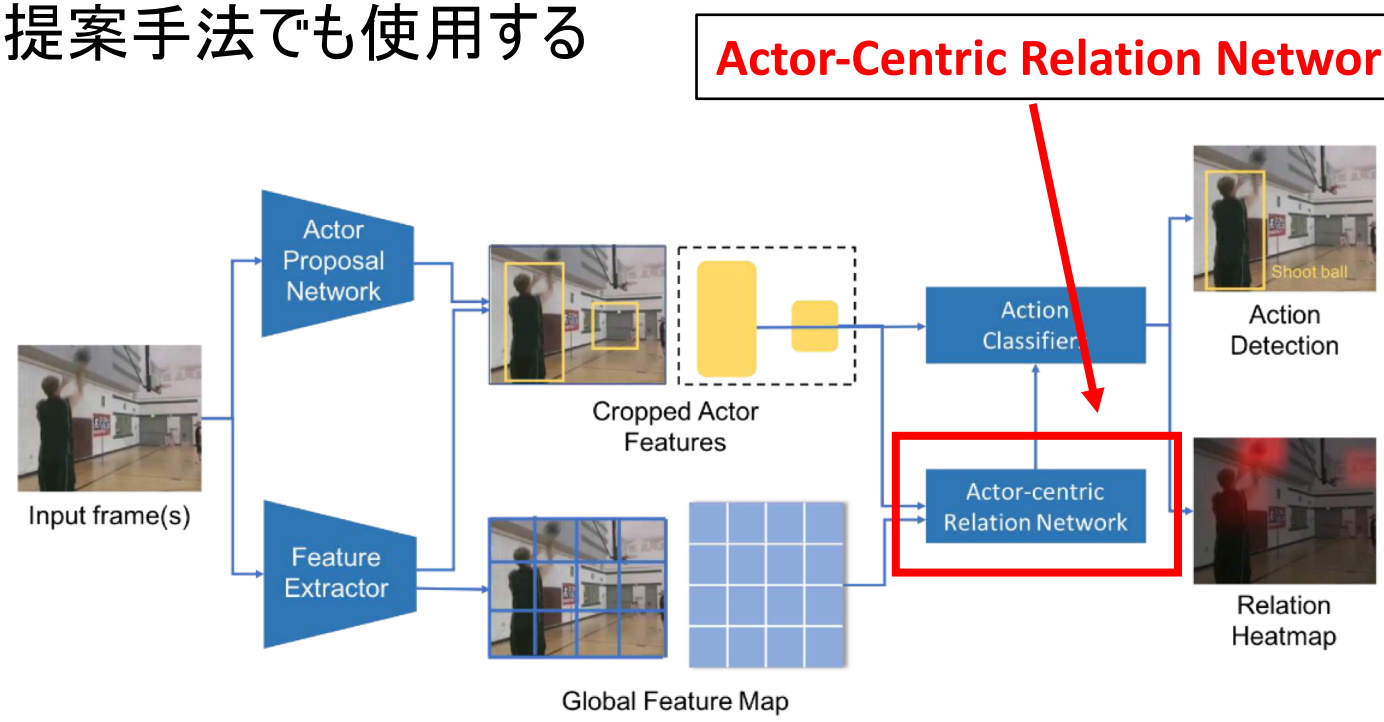
	2ステージ手法		1ステージ手法	
物体検出	・検出後に分類を行う ・高精度 ・処理が遅い	Faster R-CNN	・検出と分類を同時に行う ・処理が早い ・2ステージほど精度が出ない → Transformerを用いることで解消(DETR)	YOLO, SSD, DETR
行動検出		ACRN, ACAR		YOWO, MOC, WOO

2. 関連研究 – 行動検出

ACRN

[12] Sun et al. Actor-centric relation network. ECCV 2018

- アクターの検出後に行動の分類を行う2ステージ手法
- アクターとコンテキスト特徴の関係を学習するACRNモジュールを提案
  → 提案手法でも使用する

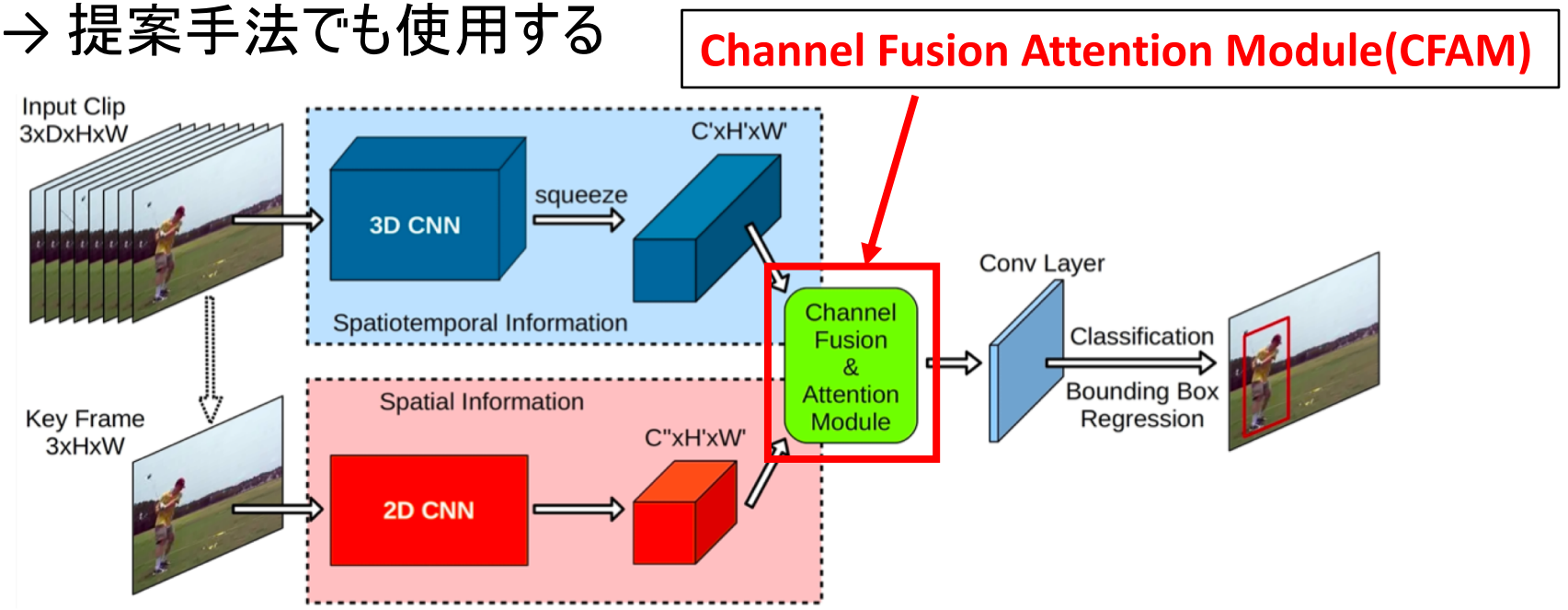


2. 関連研究 – 行動検出

YOWO

[13] K˙opu¨klu¨ et al. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. arXiv 2019.

- 2D/3DCNN特徴を効果的に融合するCFAMを提案
  - チャネル間の相互関係をモデリング
    → 提案手法でも使用する



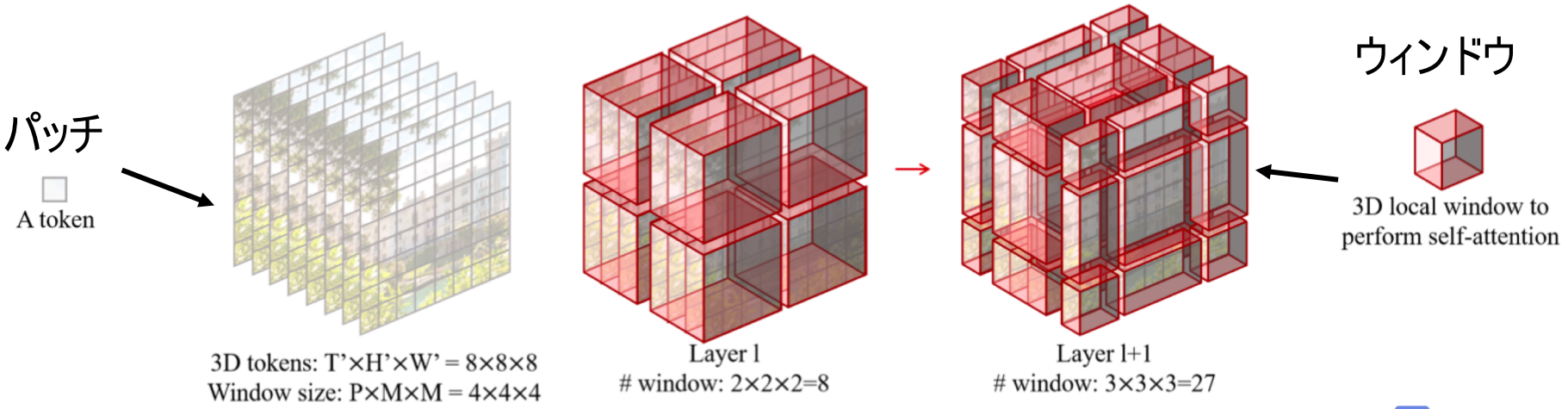
2. 関連研究 – 行動認識

Video Swin Transformer

[11] Liu et al. Video swin transformer. arXiv 2021

- 画像認識モデルSwin Transformerを行動認識に拡張
- ウィンドウ内でSelf-Attentionを計算することで局所的な関係を，
  ウィンドウをシフトすることでウィンドウ間の大域的な関係をモデリング

→初めて行動検出手法で使用



3. 提案手法

手法1

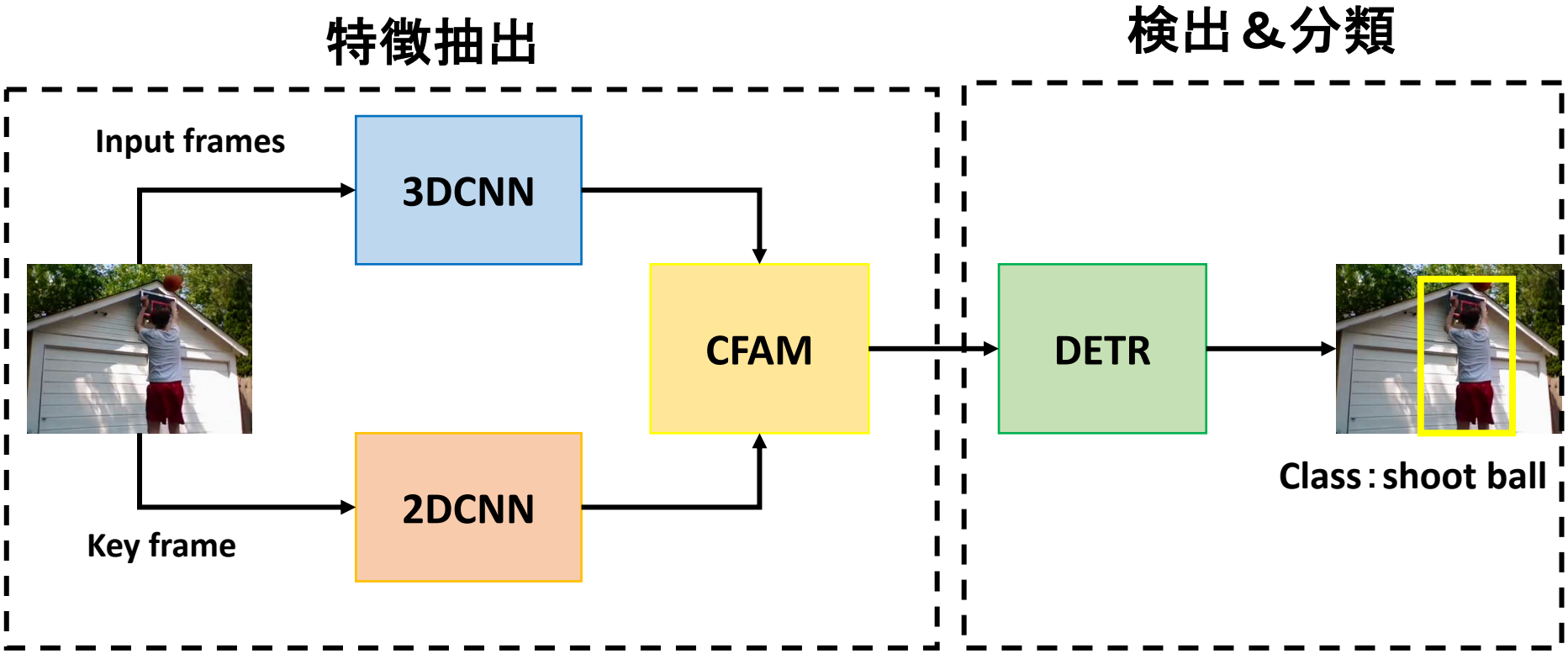
YOWOとDETRを組み合わせた1ステージ手法

手法2

最先端の行動認識モデルであるVideo Swin Transformerと,
ACRNやYOWOのCFAMを組み合わせた2ステージ手法

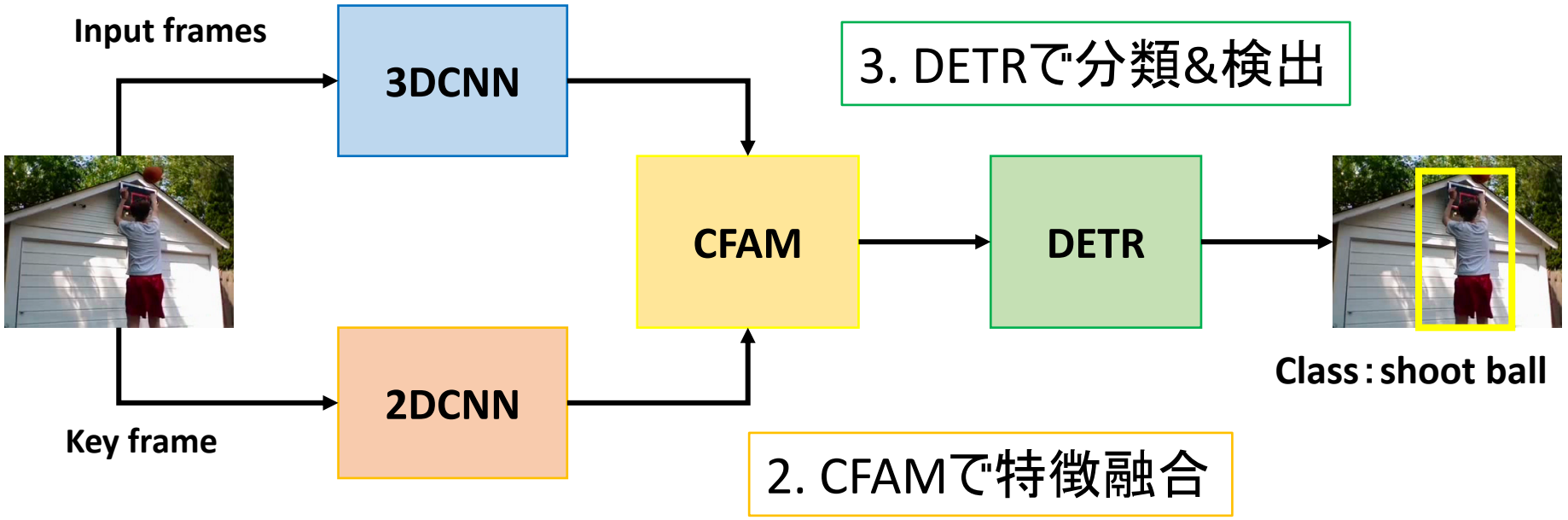
3.1 手法1

- YOWOとDETRを組み合わせた1ステージの行動検出手法



3.1 手法1

1. 特徴抽出
   (a) 3DCNNで時空間特徴を抽出
   (b) 2DCNNで空間特徴を抽出



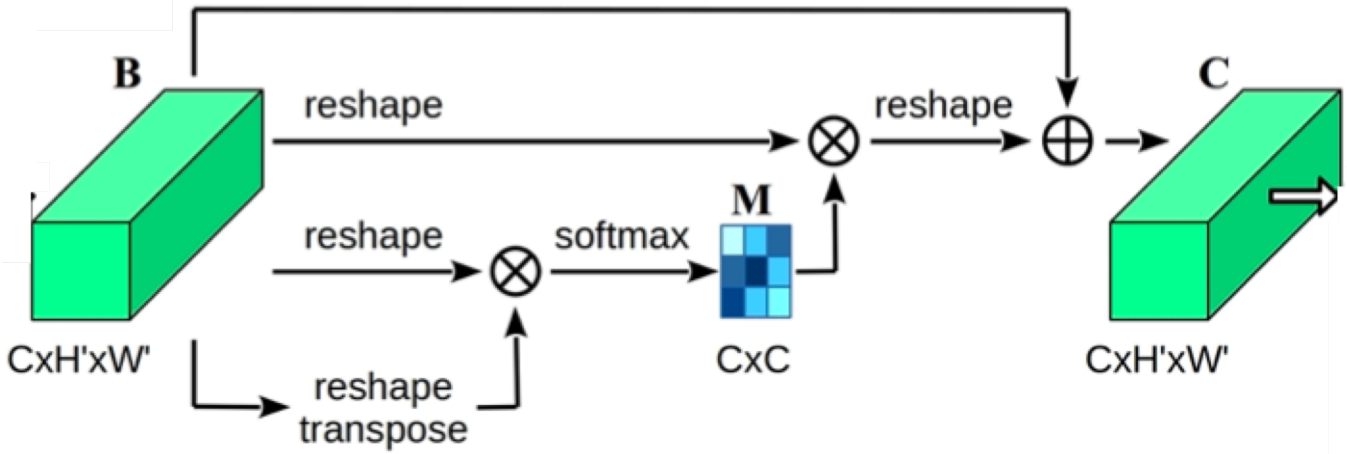
3.1 手法1

CFAM(Channel Fusion Attention Module)

- チャネル間のAttentionを計算するモジュール
- 手順

1. 特徴マップ$B \in \mathbb{R}^{C\times H\times W}$を$F \in \mathbb{R}^{C\times N}(N = HW)$に変形する
2. グラム行列$G \in \mathbb{R}^{C\times C}$を以下の式で求める

$$G = F \cdot F^T, G_{ij} = \sum_{k=1}^{N} F_{ik} \cdot F_{jk}$$



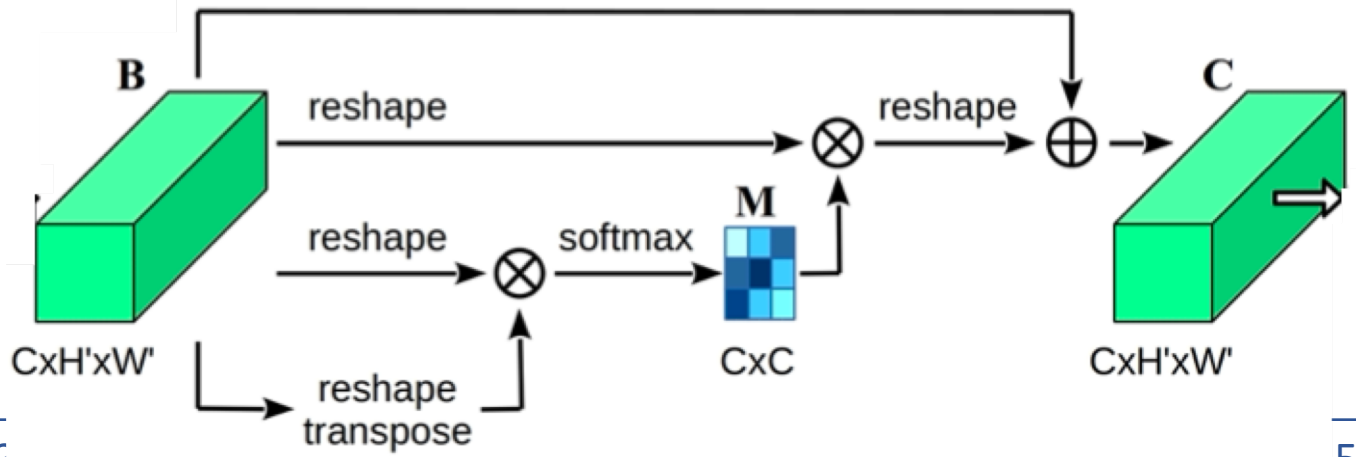
3.1 手法1

CFAM

3. Gにsoftmax層を適用しAttentionマップ$M \in \mathbb{R}^{C\times C}$を得る
4. 元の特徴マップ$F \in \mathbb{R}^{C\times N}$にAttentionマップ$M \in \mathbb{R}^{C\times C}$を乗算し，
   元の入力特徴Bと同じ形状に変形し$F' \in \mathbb{R}^{C\times H\times W}$を得る

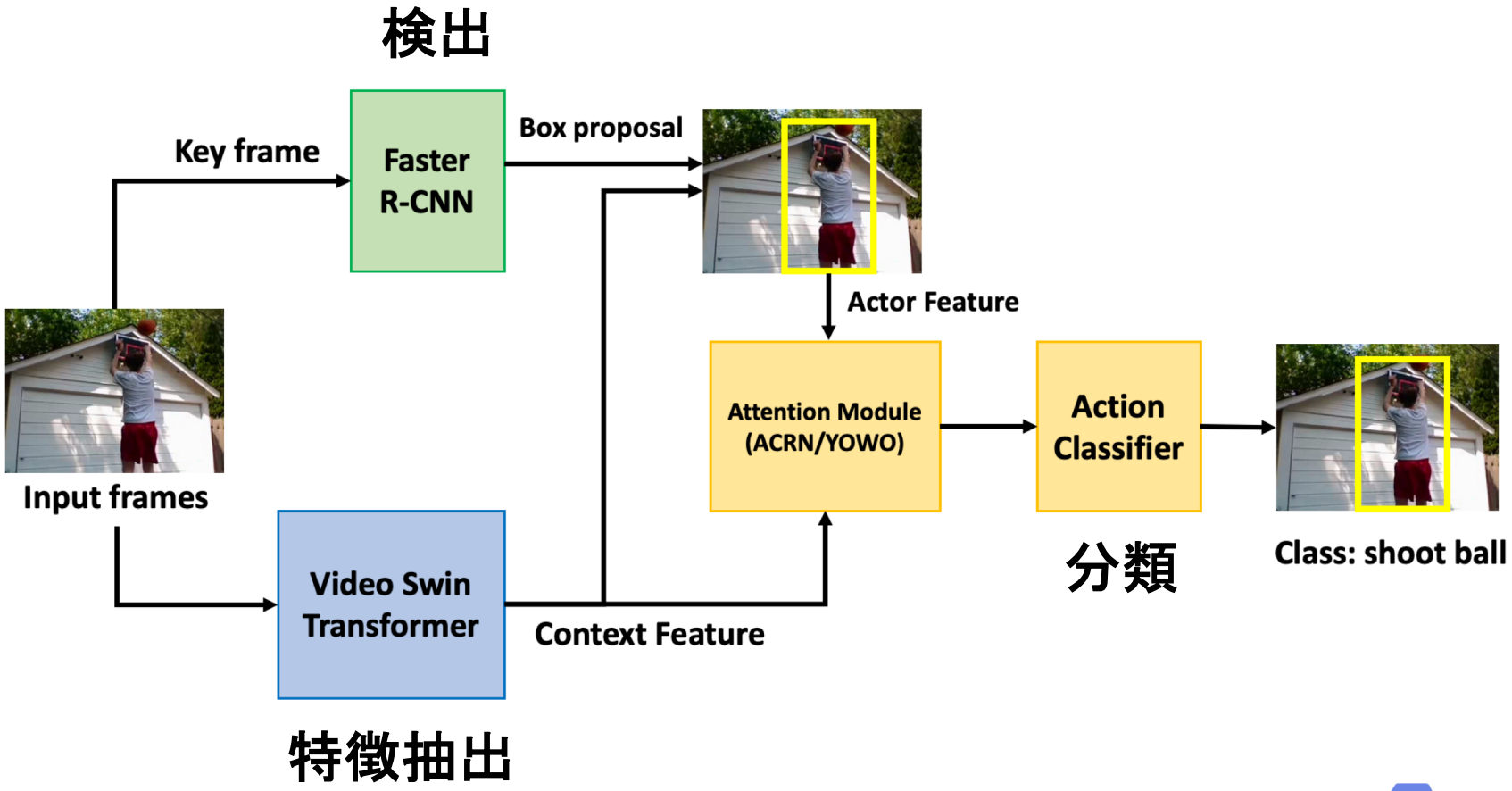
5. F'を元の入力特徴Bに要素和演算を用いて，
   学習可能なパラメータαで結合し特徴マップ$C \in \mathbb{R}^{C\times H\times W}$を得る

$$C = \alpha \cdot F' + B$$



3.2 手法2

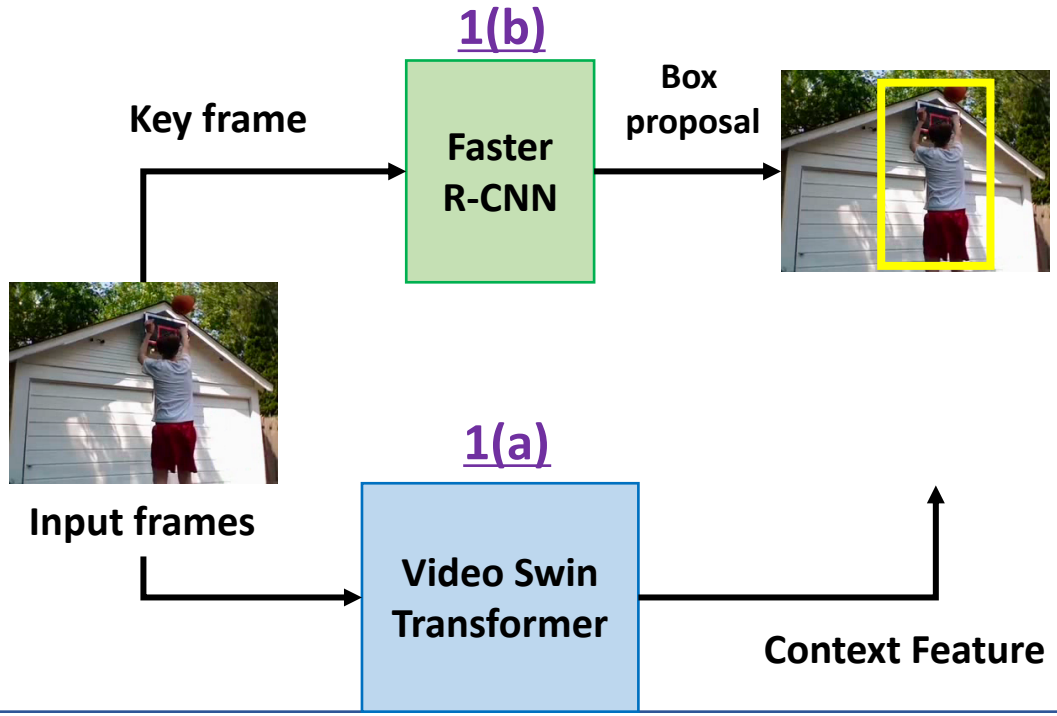
- Video Swin Transformerをベースにした2ステージの行動検出手法



3.2 手法2

処理の流れ

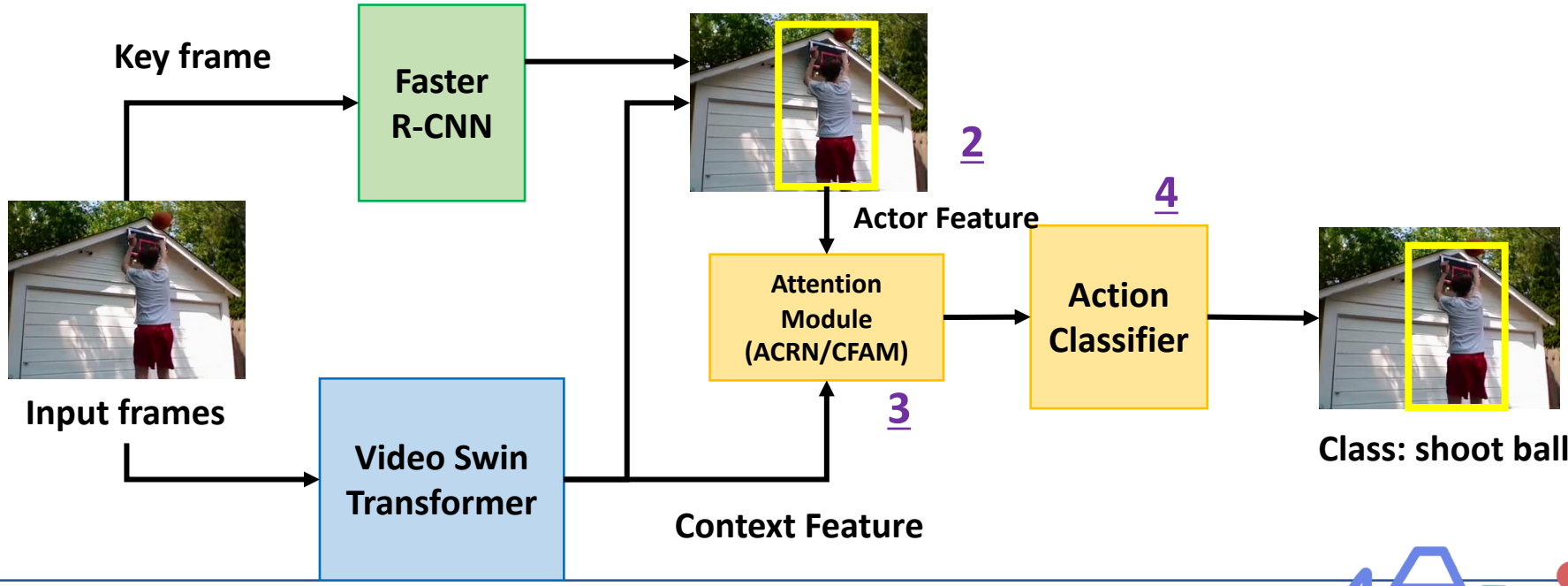
1. 特徴抽出&人物検出
   (a) Video Swin Transformerを用いてコンテキスト特徴を抽出
   (b) 検出器Faster R-CNNを用いて人物の領域候補を取得



3.2 手法2

処理の流れ

2. 各領域候補のアクター特徴を抽出する
3. ACRN及びCFAMを用いて，アクターとコンテキスト間や
チャネル間の関係をモデリング
4. FC層を適用し，分類スコアを得る



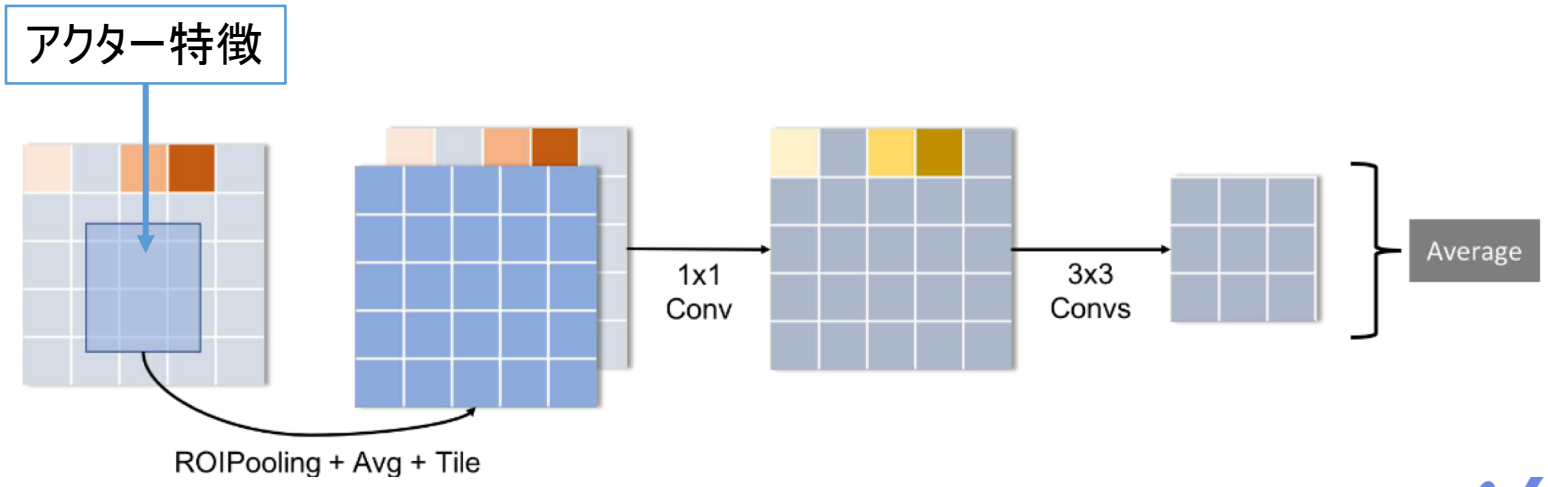
3.2 手法2

ACRN

•アクター特徴とコンテキスト特徴の関係をモデリング

•手順

1. アクター特徴にAverage Poolingを適用して1次元の特徴ベクトルを得る
2. この1次元特徴ベクトルをコンテキスト特徴の各位置に連結し，
新たな特徴マップを得る

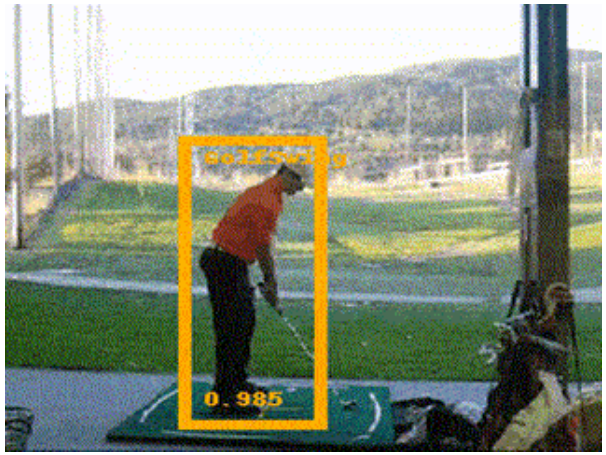


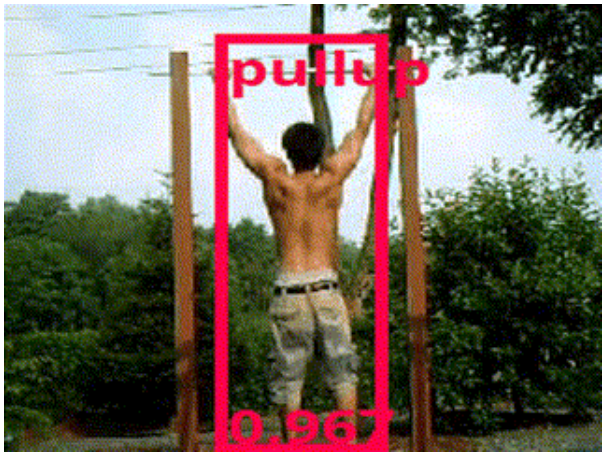
4. 実験 – データセット

UCF101-24

- 行動認識タスクで昔よく用いられていたUCF101のサブセット
- 全フレームにアノテーション(クラス、バウンディングボックス)が付与
- 全24クラス
  - 1つの動画に1クラス
  - 運動に関するクラスが多い

- 手法1で使用



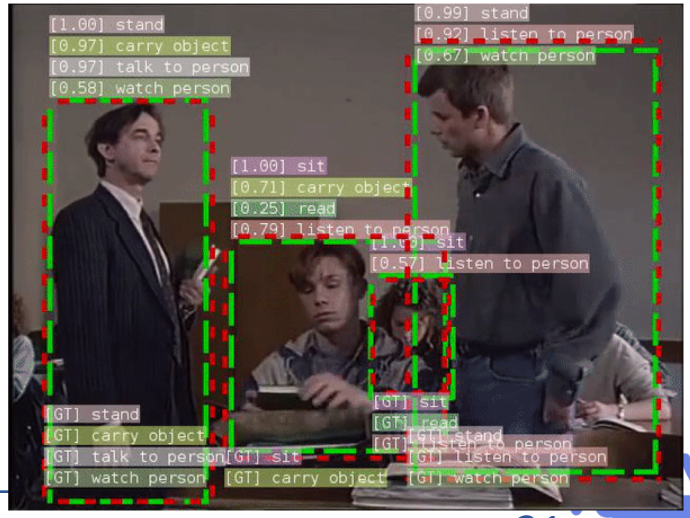


4. 実験 – データセット

AVA

- 行動認識及び行動検出タスクで広く用いられる
- 動画クリップ数：約5.7万本(約3秒/本)
- キーフレームにのみアノテーション(クラス、バウンディングボックス))を付与
- 全80クラス
  - 各動画に複数のクラスが存在する
  - 評価時には，サンプル数が25個以上存在する60クラスを使用

- 学習時間：10~20日間
- 手法2で使用



4.1 手法1の結果

分類精度 : 2D > 3D
→ UCF101-24は、空間情報が重要であることから妥当

手法		定位精度	分類精度
YOWO	2D	91.7	85.9
	3D	90.8	92.9
	2D+3D+CFAM	92.7	92.3
DETR	2D	58.6	75.5
	3D	75	72.8
YOWO+DETR(手法1)		66.9	78.9

DETRへの入力特徴の次元数が小さいことが原因？
- 計算コストの都合上、次元数を2473→512にした

4.2 手法2の結果

model	Reference	pretrain	mAP
AVA baseline	CVPR18	K400	15.6
ACRN	ECCV18	K400	17.4
YOWO	arXiv19	K400	19.2
SlowFast, R101	ICCV2019	K600	27.3
Context-Aware	ECCV20	K400	28.0
ACAR	CVPR21	K400	28.8
ACAR w/o ACFB	CVPR21	K400	27.8
WOO, SFR101	ICCV21	K600	28.0
Swin-T	-	K400	21.4
Swin-B	-	K600	26.5
Swin-B+ACRN	-	K600	28.3
Swin-B+CFAM	-	K600	26.5

同条件下では
提案手法が上回る

ACARに次いで2番目に高い28.3mAPを達成
-ウィンドウ内でSelf-Attentionを計算しており，
より識別性の高いコンテキスト特徴が得られた結果

4.2 手法2の結果

- AVAデータセットの全80クラスは以下の3つのカテゴリに分けられる

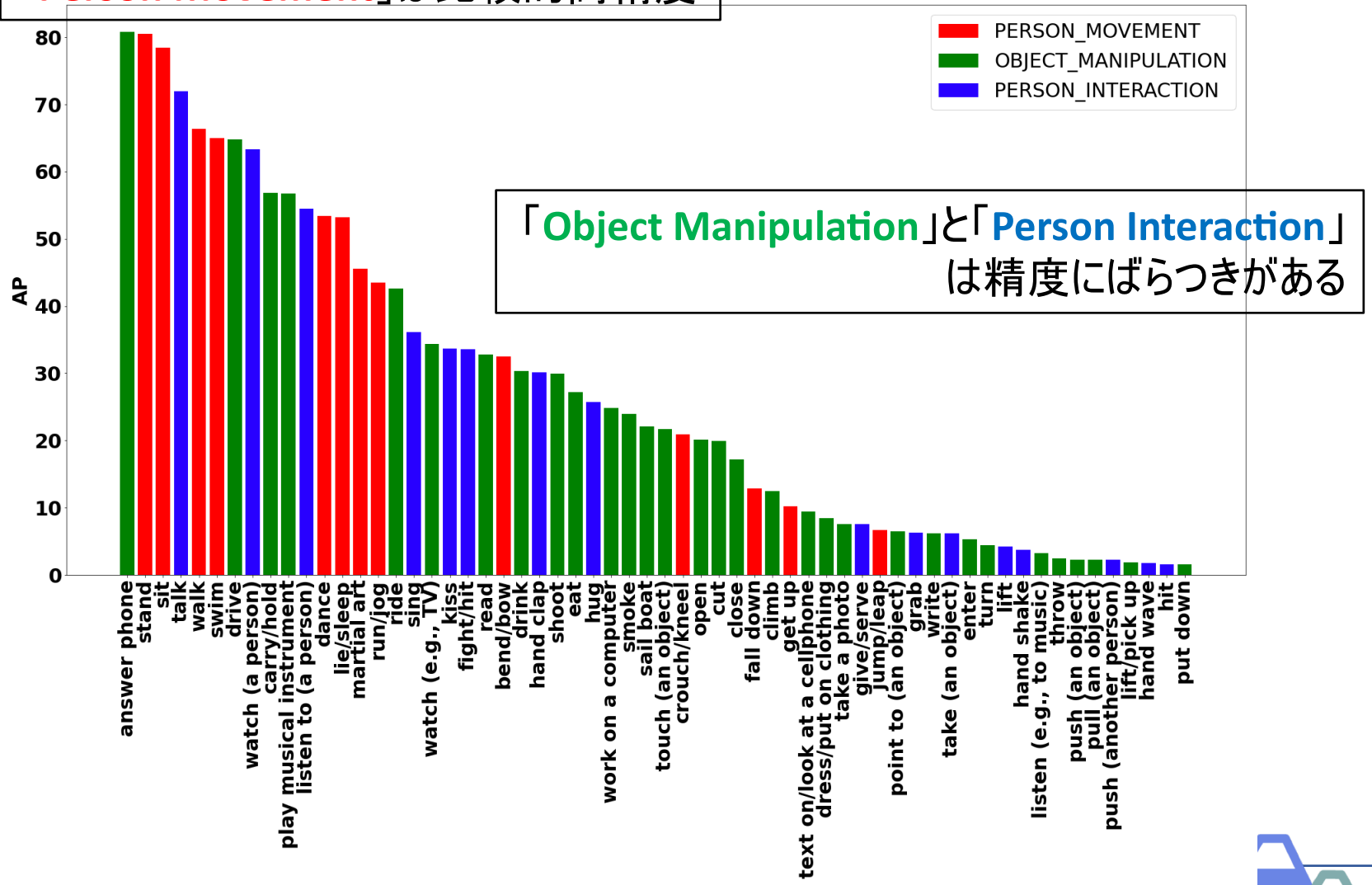
(1) **Person Movement**
  - 人間のみに関係する行動(jump, dance, runなど計13個)

(2) **Object Manipulation**
  - 人間が物体を操作する行動(answer phoneなど計31個)

(3) **Person Interaction**
  - 人間同士の行動(hand shake, hug, talkなど計16個)

4.2 手法2の結果

「Person Movement」は比較的高精度

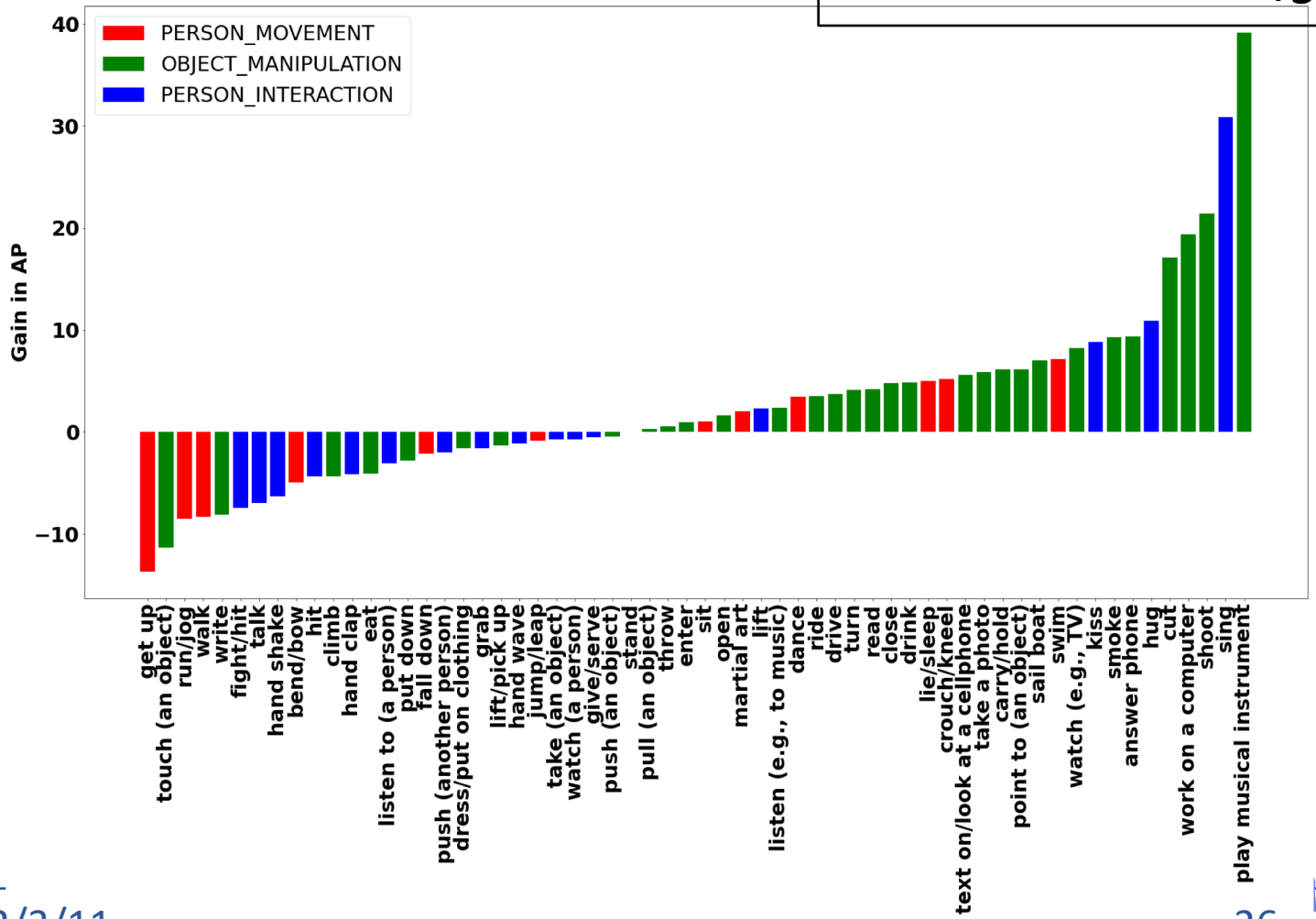


「Object Manipulation」と「Person Interaction」は精度にばらつきがある

4.2 手法2の結果

SlowFast vs. 提案手法2

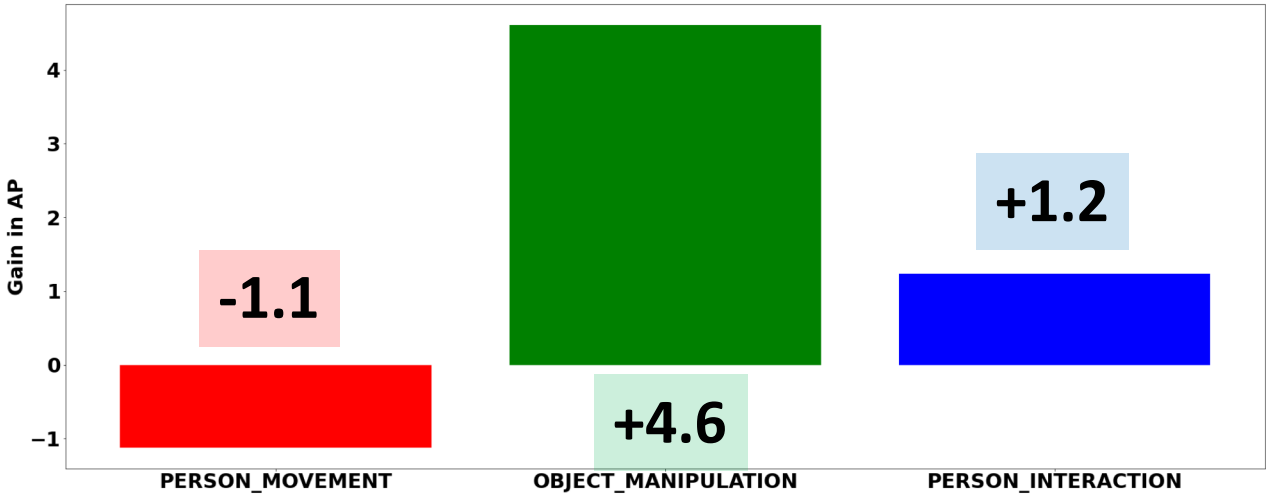
「play musical instrument」
+39%



4.2 手法2の結果

SlowFast vs. 提案手法2

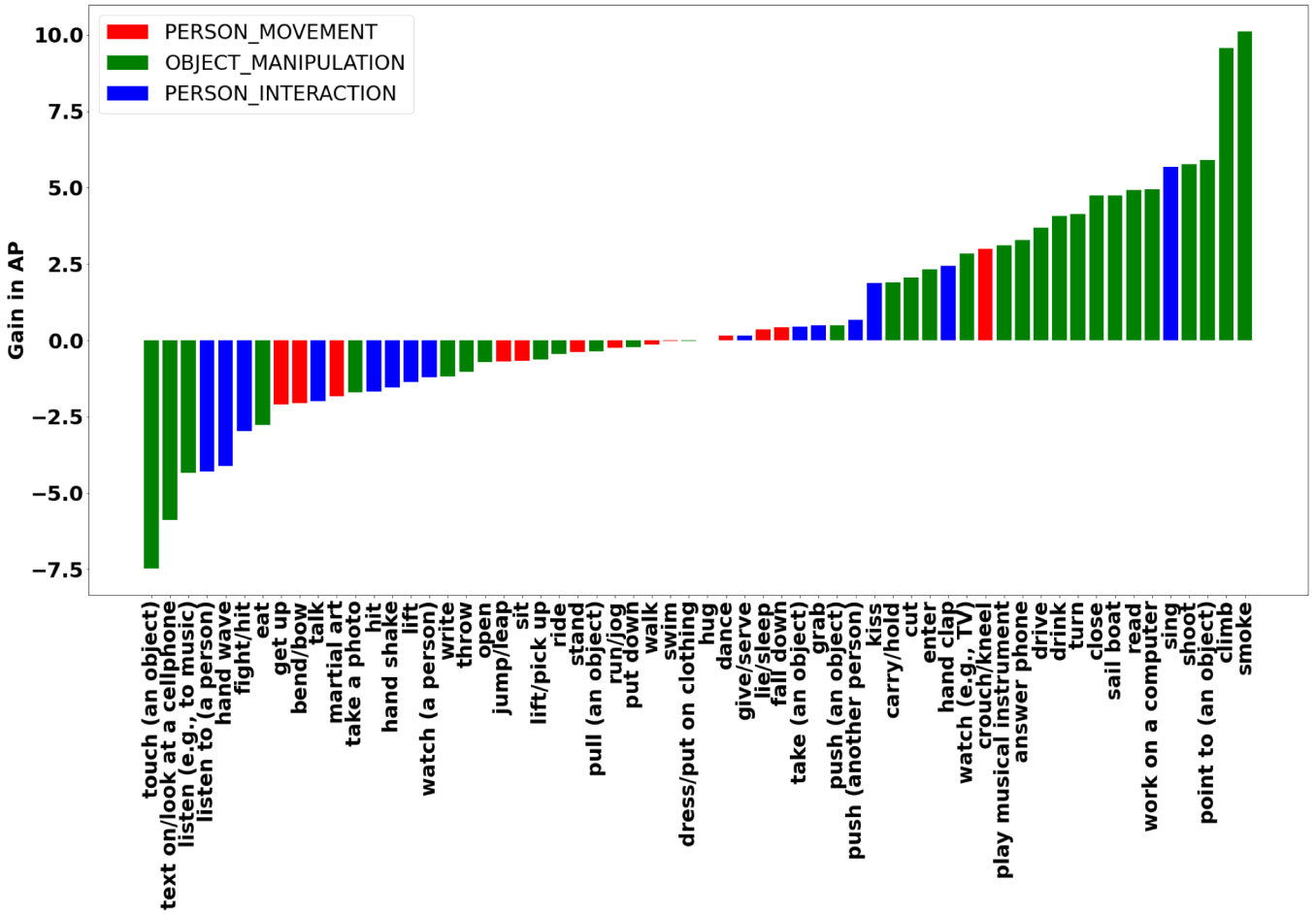
- 「**Object Manipulation**」: 4.6%向上
- 「**Person Interaction**」: 1.2%向上
  → 空間特徴を捉える能力を向上
- 「**Person Movement**」は1.1%低下
  → 時間特徴を捉える能力は少し低下



4.2 手法2の結果

提案手法2(w/o ACRN) vs. 提案手法2

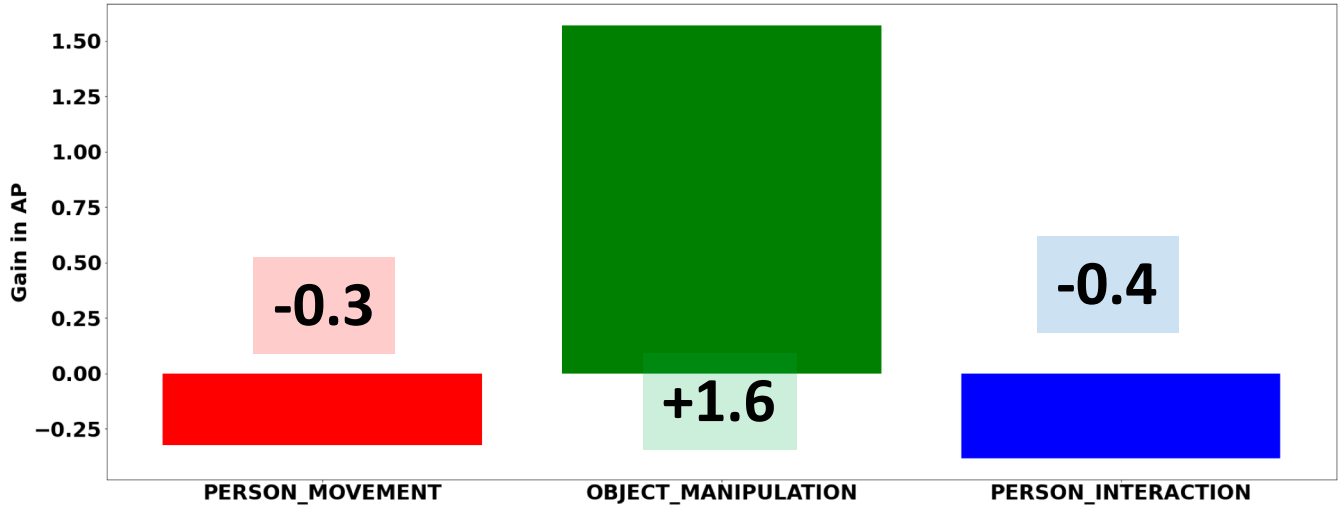
「smoke」:+10%



4.2 手法2の結果

提案手法2(w/o ACRN) vs. 提案手法2

- 「**Object Manipulation**」: 1.6%向上
  → 人間と物体の関係を捉える能力が向上
- 「**Person Movement**」は0.3%低下
- 「**Person Interaction**」: 0.4%向上



5. まとめ

- 本研究では，行動検出のタスクに取り組んだ
- Transformerベースの提案手法は，既存手法を上回った
- 「Object Manipulation」クラスにおいて精度が改善された

6. 今後の課題

- 物体の検出も行うことで更なる精度向上が期待できる
- また，AVAデータセットはロングテールなので，学習方法を工夫することで更なる精度向上が期待できる
- 推論時に特徴バンクを用いることで更なる精度向上が期待できる


UEC
TOKYO