クエリベースのアンカーを用いた人間と物体のインタラクション検出

PRMU2022

電気通信大学 大学院 情報学専攻
陳 俊文　　柳井 啓司

Human-Object Interaction (HOI) Detection

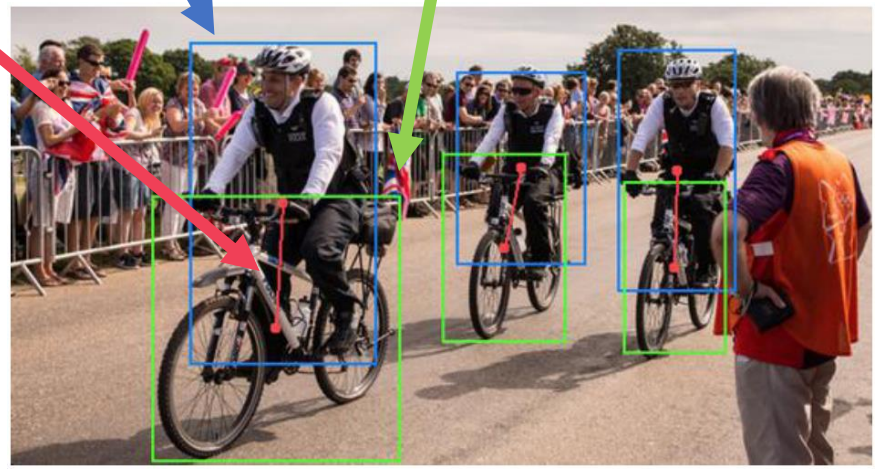
HOI 検出

- 一枚の画像から＜human, object, interaction＞の組合せを検出する

HOI インスタンス

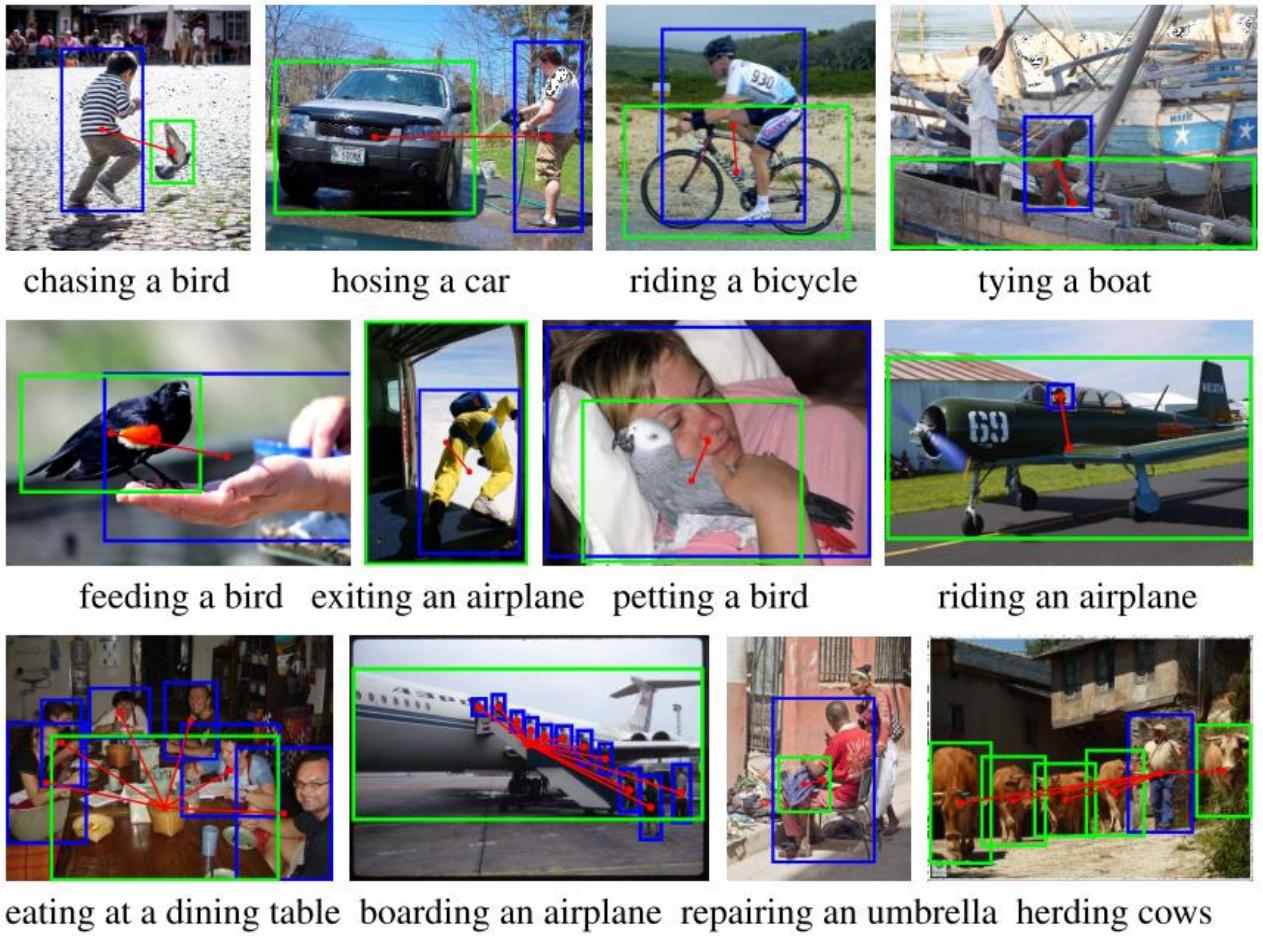
$$\left\{\left[x_1^{\text{human}}, y_1^{\text{human}}, x_2^{\text{human}}, y_2^{\text{human}}\right], \left[x_1^{\text{obj}}, y_1^{\text{obj}}, x_2^{\text{obj}}, y_2^{\text{obj}}\right], c_{\text{HOI}}\right\}$$

$$c_{\text{HOI}} : \left[c_{\text{obj}}, c_{\text{action}}\right]$$



HICO-DET

- HICO-DET [1] は HOI 検出に最もよく使われるデータセット
- トレーニングセット：38,118 枚，テストセット：9,658 枚
- HOI クラス：117 個の動詞と 80 個の物体から構成される 600 種類



[1] Chao, Yu-Wei, et al. "Learning to detect human-object interactions." *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2018.

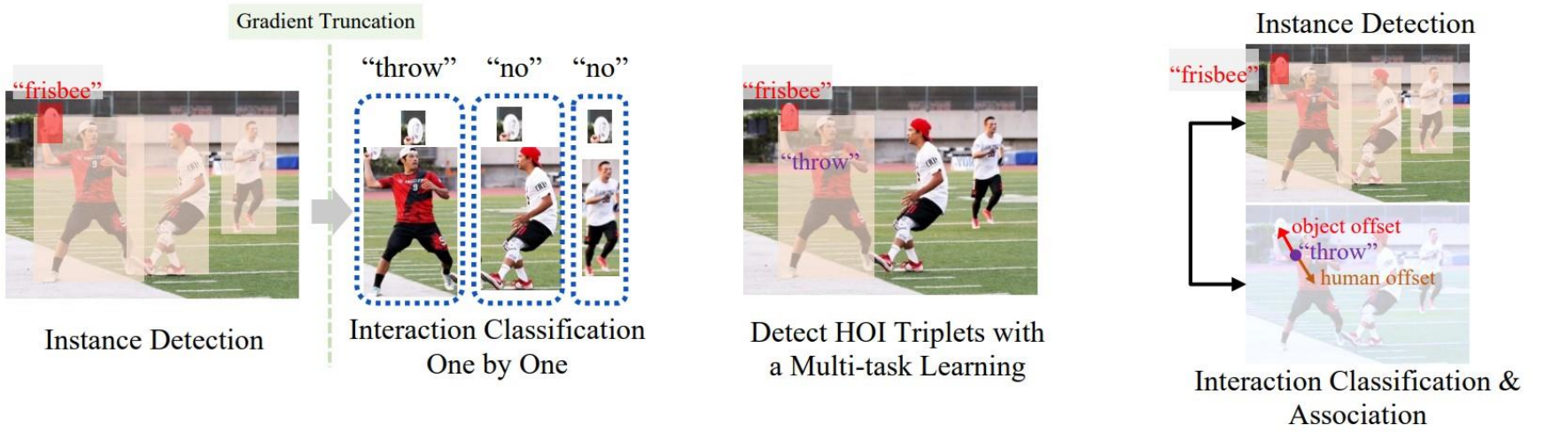
HOI 検出手法

☐ Two-stage

- 事前学習済の物体検出器が必要
- 2つのステップで物体検出とインタラクション認識を行う

☐ One-stage

- 1つのステップで検出と認識を行う（プロポーザルの必要がない）



(a). Two-stage framework

(b). One-stage end-to-end framework

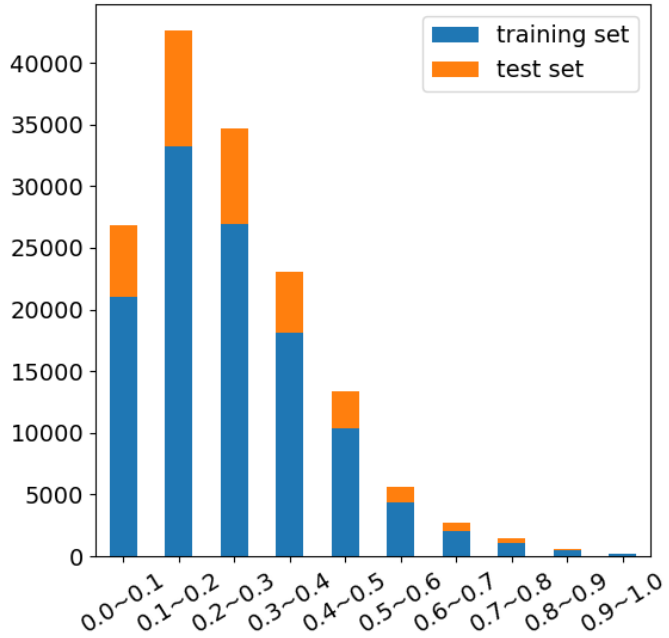
(c). One-stage framework with parallel architecture

モチベーション

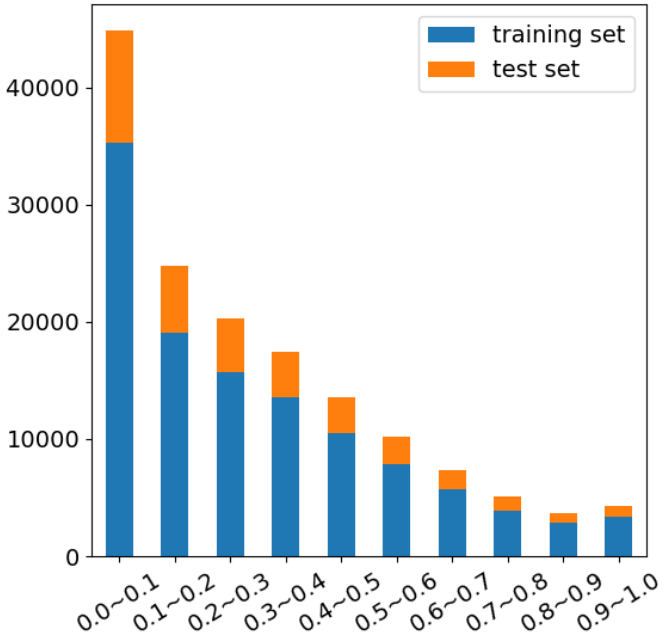
◘ HICO-DET における人間と物体の空間分布

- 人間と物体が離れている
- 小さなターゲット

◘ One-stage のマルチスケールアーキテクチャがない



(a) 人間と物体の中心点の距離

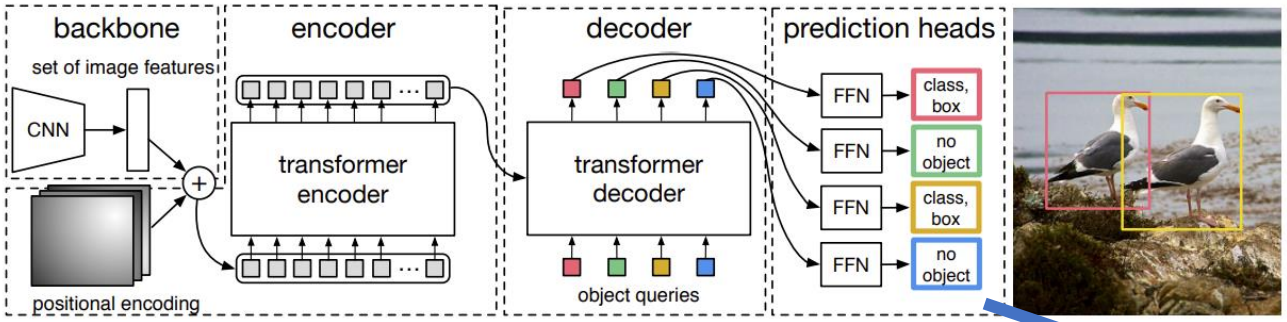


(b) 人間と物体のボックスのうち，
大きい方の領域

モチベーション

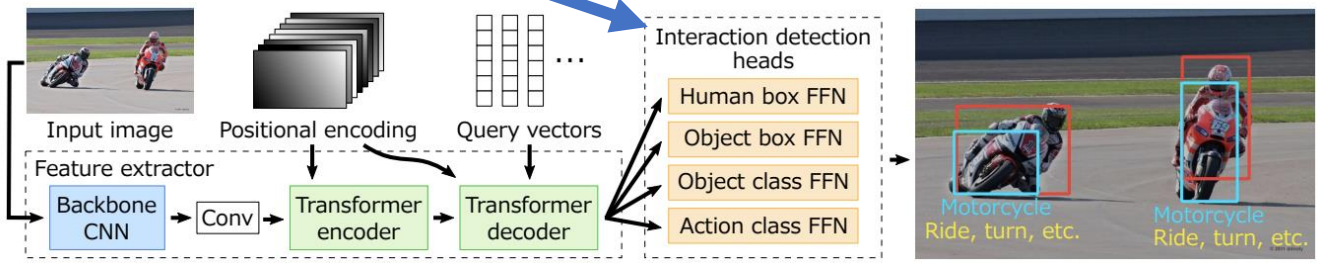
◘ 従来手法の問題点

- Transformer ベースの one-stage 手法 QPIC [2]
  - DETR [3] のアーキテクチャを HOI タスクに適用したものである
  - 学習の収束が遅く，アテンションの計算量は特徴マップの 2 乗に比例する



DETR

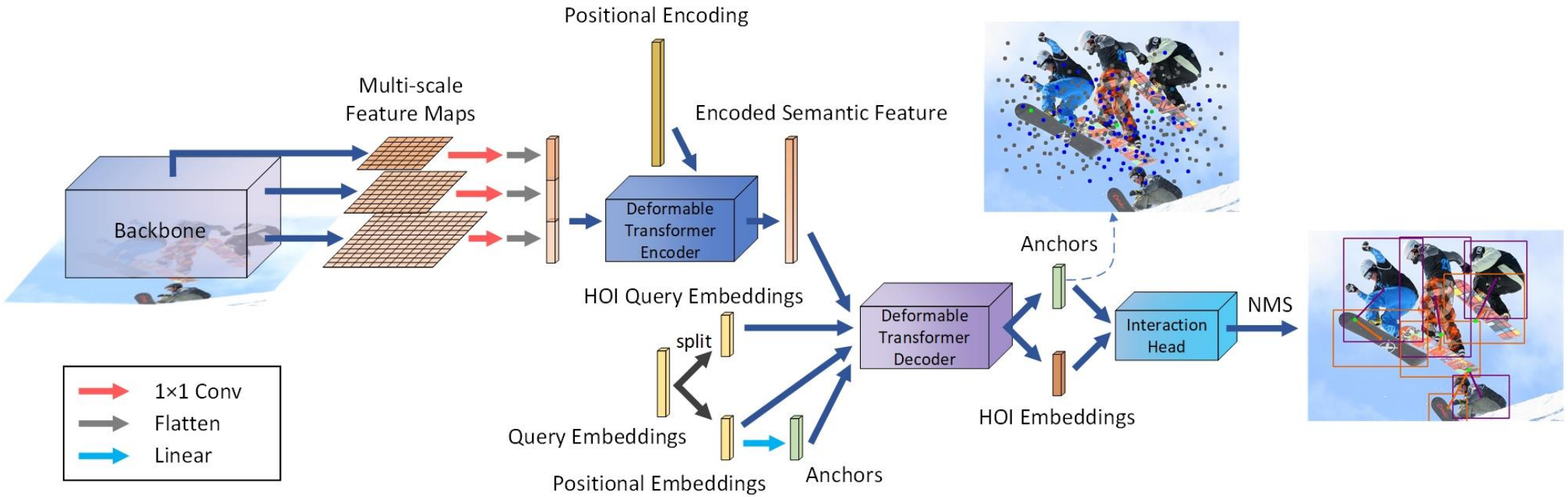
QPIC



[2] Tamura, Masato, Hiroki Ohashi, and Tomoaki Yoshinaga. "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[3] Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Springer, Cham, 2020.

QAHOI

- 小さな物体の検出と画像全体の意味的情報抽出の能力を向上させる
- DETR ⟶ Deformable DETR [4]
- Query-based Anchor + マルチスケールアーキテクチャ



[4] Chao, Yu-Wei, et al. "Learning to detect human-object interactions." *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2018.

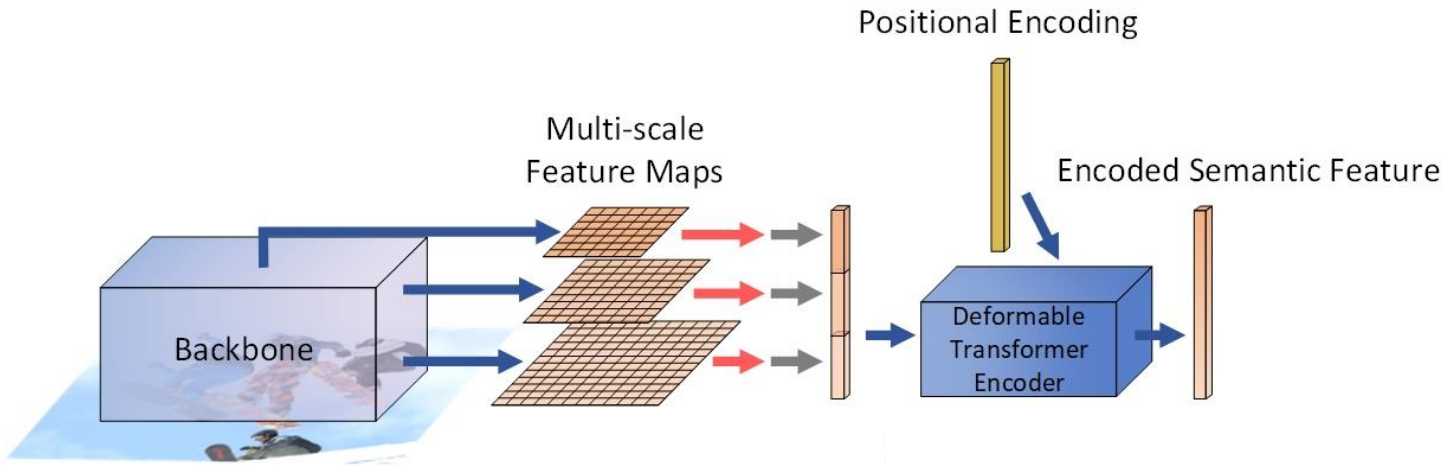
マルチスケール特徴抽出器

◘ QPIC の特徴抽出器

- CNN バックボーン + Transformer エンコーダ [5]
  - 計算コストが高いため，低解像度特徴マップを使用する

◘ QAHOI のマルチスケール特徴抽出器

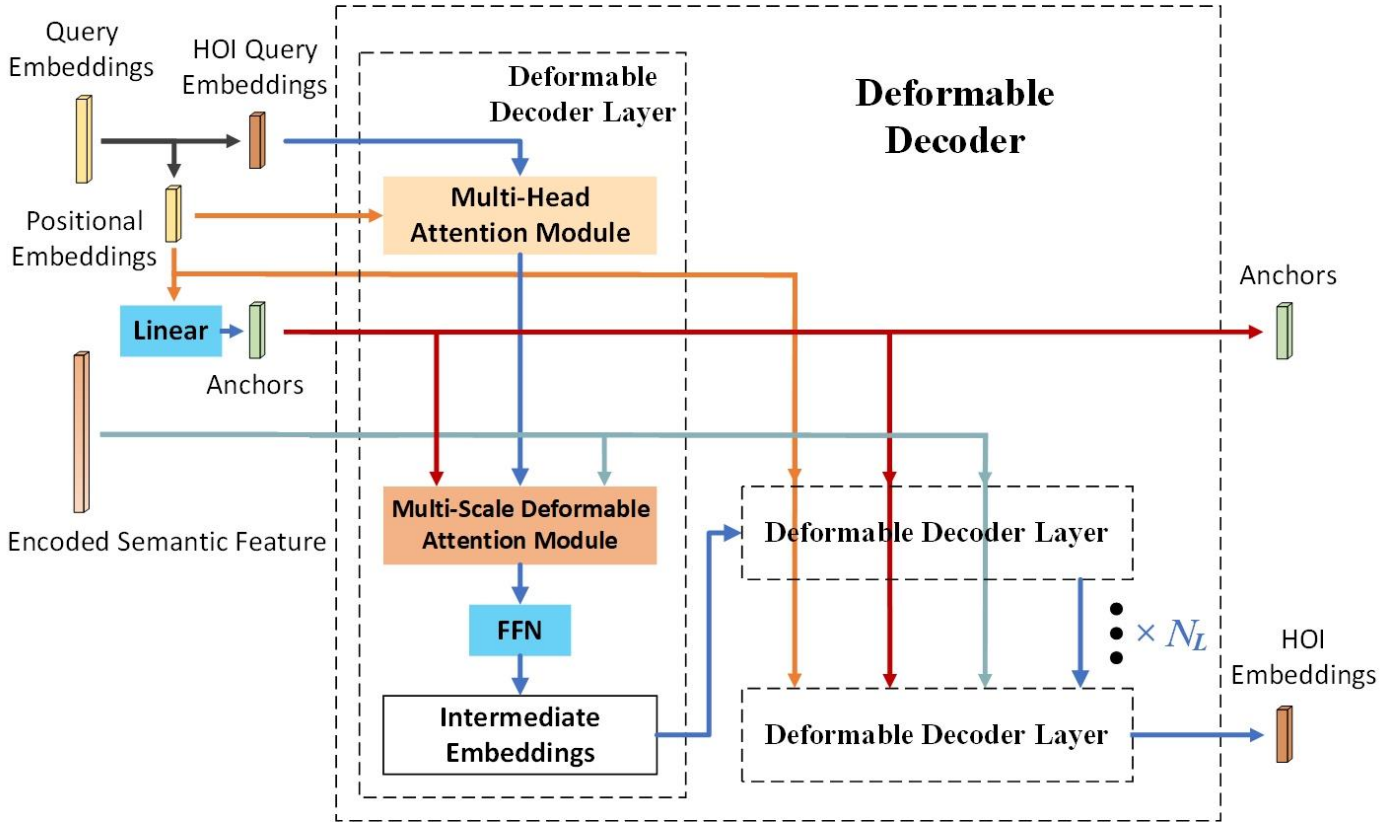
- 階層型バックボーン + Deformable Transformer エンコーダ
  - バックボーンの複数ステージの特徴マップを利用できる



[5] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.

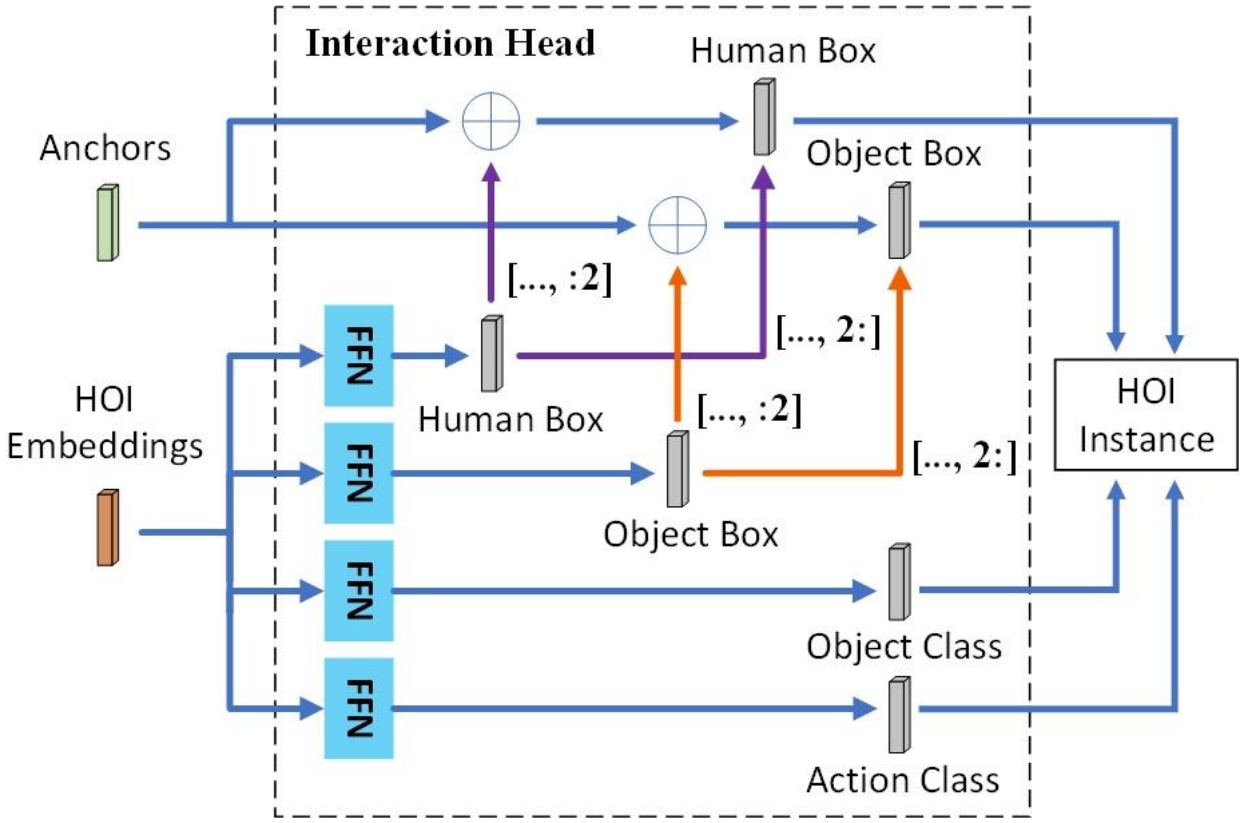
アンカーベースのデコーディング処理

- Deformable DETR に従い，クエリ埋め込みは，等しく 2 つの部分に分割される
  - HOI クエリ埋め込み
  - 位置埋め込み
- アンカーは位置埋め込みから線形層を介して生成されたものである



アンカーベースのインタラクション検出ヘッド

- QAHOI は，QPIC と同じくシンプルなインタラクション検出ヘッドを採用している
- アンカーをベースにして Feed-forward Network (FFN) を用いて HOI の全ての要素を予測する
- 物体のクラスとアクションのクラスを人間と物体のバウンディングボックスと組み合わせて，HOI インスタンスを構築する



Top K スコアと HOI NMS

◘ Top K スコア

- 物体クラスのスコアが上位 K の HOI インスタンスを選択する

◘ HOI NMS

- HOI インスタンス間の人間と物体の複合 IoU と，HOI スコアに基づいて算出する
  - HOI スコアは，物体スコアとアクションスコアを掛け合わせたものである
  - HOI インスタンス i と j の間の人間と物体の複合 IoU は，物体と物体の IoU 掛ける人間と人間の IoU の積である

$$\mathrm{IoU}(i,j) = \mathrm{IoU}(B_i^{(h)}, B_j^{(h)}) \cdot \mathrm{IoU}(B_j^{(o)}, B_j^{(o)})$$

モデルの学習と推論

- QPIC の学習手順に従って，ハンガリアン法 [6] を用いて，予測値を全ての ground-truth セットと一致させる
- Deformable DETR に従って，物体クラスの損失は Focal Loss [7] を使用する
- クエリ埋め込みは学習可能なパラメータであるため，クエリ埋め込みから得られたアンカーの位置は学習時に学習され，推論時に固定される

[6] Kuhn, Harold W. "The Hungarian method for the assignment problem." Naval research logistics quarterly 2.1-2 (1955): 83-97.
[7] Lin, Tsung-Yi, et al. "Focal loss for dense object detection." Proceedings of the IEEE international conference on computer vision. 2017.

実験

◘ データセット

- HICO-DET で実験を行う
- データセットに含まれる 600 個の HOI クラスのインスタンス数に基づいて，これら HOI クラスは3つのカテゴリに分類される
  - *Full*：全ての HOI クラス
  - *Rare*：インスタンスが 10 個未満の 138 個のクラス
  - *None-rare*：インスタンスが 10 個以上の 462 個のクラス

◘ 評価指標

- mAP (mean average precious) が使用される
- HICO-DET の Default 設定（未知物体あり）と Known Object 設定（未知物体なし）で *Full*，*Rare*, *Non-Rare* カテゴリに対する mAP を報告する

実験

◘ 最先端手法との比較

Arch.	Method	Backbone	Fine-tuned Detection	Default			Known Object		
				Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
Points	IP-Net [16]	ResNet-50-FPN	✗	19.56	12.79	21.58	22.05	15.77	23.92
	PPDM [9]	Hourglass-104	✓	21.73	13.78	24.10	24.58	16.65	26.84
	GGNet [18]	Hourglass-104	✓	23.47	16.48	25.60	27.36	20.23	29.48
Query	HOITrans [20]	ResNet-101	✓	26.61	19.15	28.84	29.13	20.98	31.57
	HOTR [7]	ResNet-50	✗	23.46	16.21	25.65	-	-	-
	HOTR [7]	ResNet-50	✓	25.10	17.34	27.42	-	-	-
	AS-Net [3]	ResNet-50	✗	24.40	22.39	25.01	27.41	25.44	28.00
	AS-Net [3]	ResNet-50	✓	28.87	24.25	30.25	31.74	27.07	33.14
	QPIC [15]	ResNet-101	✓	29.90	23.92	31.69	32.38	26.06	34.27
	QAHOI	**Swin-Tiny**	✗	28.47	22.44	30.27	30.99	24.83	32.84
	QAHOI	**Swin-Base**	✗	29.47	22.24	31.63	31.45	24.00	33.68
	QAHOI	**Swin-Base**$^{*+}$	✗	33.58	25.86	35.88	35.34	27.24	37.76
	QAHOI	**Swin-Large**$^{*+}$	✗	**35.78**	**29.80**	**37.56**	**37.59**	**31.66**	**39.36**

+5.88 (19.7%)

+4.1 (13.9%)

実験

◘ アブレーション実験

Arch.	Model	Backbone	Fine-tuned Detection	Multi-scale	Default *Full*	Default *Rare*	Default *Non-Rare*
QPIC	(1)	ResNet-50	✗	x_3	24.21	17.51	26.21
	(2)	ResNet-50	✓	x_3	29.07	21.85	31.23
	(3)	Swin-Tiny	✗	x_3	27.19	21.32	28.95
QAHOI	(4)	ResNet-50	✗	x_1, x_2, x_3, x_4	24.35	16.18	26.80
	(5)	ResNet-50	✓	x_1, x_2, x_3, x_4	26.18	18.06	28.61
	(6)	Swin-Tiny	✗	x_1, x_2, x_3, x_4	28.09	21.65	30.01
	(7)	Swin-Tiny	✗	x_1, x_2, x_3	28.47	22.44	30.27
	(8)	Swin-Tiny	✗	x_2, x_3	28.12	20.43	30.41
	(9)	Swin-Tiny	✗	x_3	26.65	19.13	28.89

物体検出器の学習を行わない場合，QAHOI は QPIC より優れている

マルチスケール特徴マップの効果

アーキテクチャについてのアブレーション実験

実験

◘ 後処理のアブレーション実験

- QAHOI-Swin-Tiny を用いて実験を行った

Top K と NMS のステップが重要である

method	Default		
	Full	*Rare*	*Non-Rare*
base	26.64	20.62	28.44
+ topk scores (N_t = 100)	26.70	20.89	28.43
+ NMS (δ = 0.5)	28.47	22.44	30.27

(a) フィルタリングステップのアブレーション実験

topk scores	N_t		
	50	100	150
c_a	26.63	26.63	26.63
c_o	26.69	**26.70**	26.64
$c_a \cdot c_o$	26.63	26.63	26.63

(b) Top K スコアのアブレーション実験

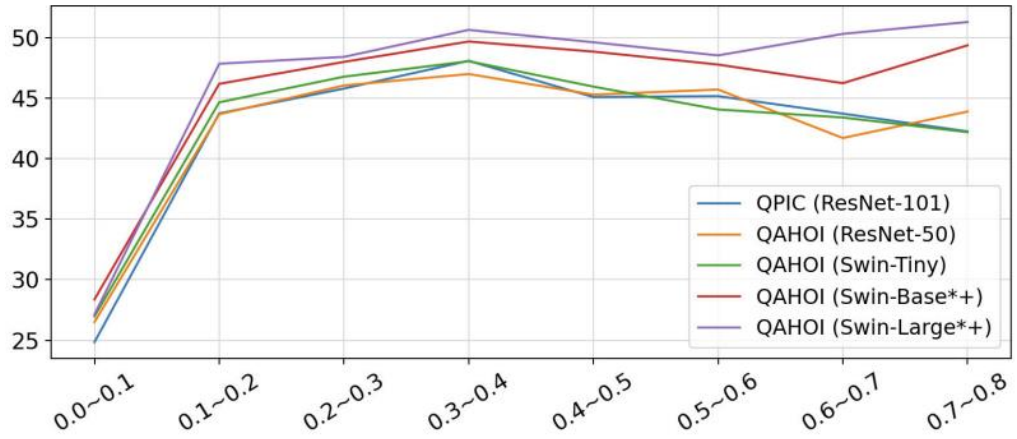
IoU	IoU threshold			
	0.4	0.5	0.6	0.7
IoU^h	27.85	27.93	27.96	27.93
IoU^o	26.69	26.77	26.84	26.85
$\text{IoU}^h \cdot \text{IoU}^o$	28.41	**28.47**	28.37	28.07

(c) NMS 処理のアブレーション実験

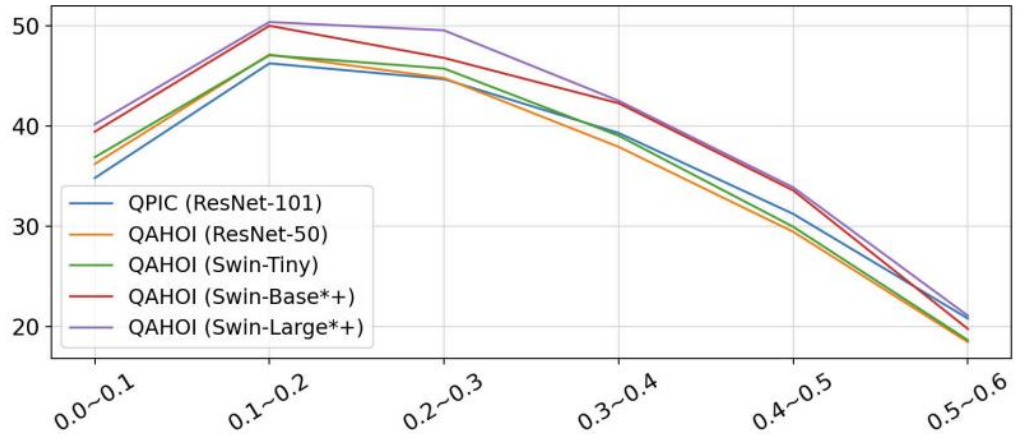
実験

◘ 異なる空間スケールでの評価

- 異なるスケールでの HOI インスタンスの異なる中心距離と大きな領域の両方を評価した
- 10 個のビンに分割し，インスタンス数が 1000 以上のビンを選択して AP 結果を表示させた



(a) 人間と物体のボックスのうち，大きい方の面積

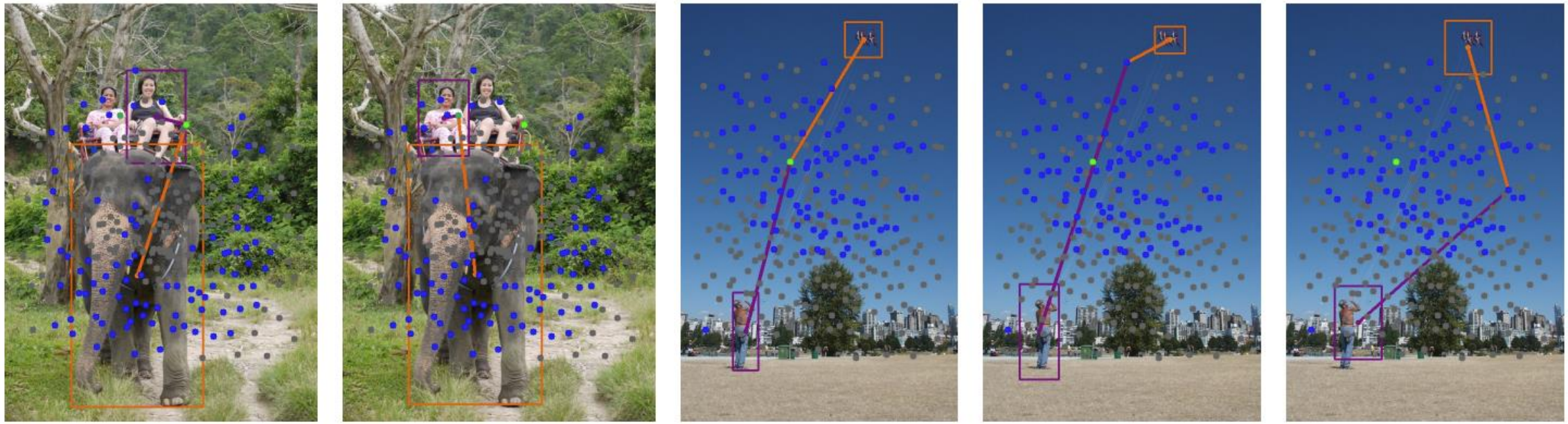


(b) 人間と物体の中心点の距離

定量的な結果

◘ クエリベースのアンカーの柔軟性

- アンカーは位置に関係なく HOI インスタンスを検出することが可能である
- 人間と物体が離れている場合も検出できる



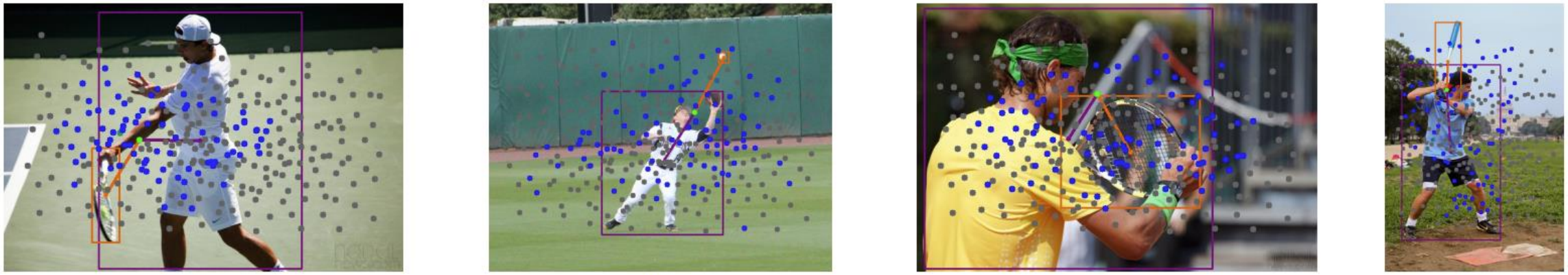
(a) ride, elephant (b) ride, elephant (c) fly, kite (d) fly, kite (e) fly, kite

アンカーの柔軟性

定量的な結果

◘ アンカーの分布

- 物体スコアの高いアンカーは，人間と物体の中心に近い位置にある
- HOI スコアが高いアンカーは人間と物体の中心に限定されていない



(a) (b) (c) (d)

アンカーの分布

定量的な結果

◘ アテンションの可視化

Top 1 スコア HOI インスタンスのアテンションの可視化

まとめ

- 本論文では，マルチスケール特徴を活用できるアンカーベースの新たな one-stage HOI 検出フレームワーク QAHOI を提案した
- アテンションメカニズムを持つ Transformer ベースのバックボーンは HOI 検出において大きつ Transformer ベースのバックボーンは HOI 検出において大きを柔軟に検出できることが実験で示された
- 提案手法はマルチスケールアーキテクチャとアンカーベースの設計を持ち，いくつかの改善点を追加することが可能
- QAHOI が，最新の物体検出器で使用されている技術でさらに発展し，将来の研究において HOI 検出タスクの強力なベースラインとして使用されることを期待している


UEC
TOKYO