社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

陳　　俊文†　柳井　啓司†

† 電気通信大学　大学院情報理工学研究科　情報学専攻
E-mail: †chen-j@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　人間と物体のインタラクション（HOI）検出では，人間と物体のペアをローカライズし，画像から人間と物体間の意味的関係を抽出する必要がある．既存の one-stage の手法は，可能なインタラクションポイントの検出や人間と物体のペアのフィルタリングに注目している．空間スケールにおける異なる物体の位置やサイズの違いを考慮していない．本研究では，Transformer を用いたマルチスケールアーキテクチャを採用し，クエリに基づくアンカーを用いて HOI インスタンスの全ての要素を予測する one-stage の手法を提案する．また，Transformer ベースのバックボーンを用いて，HICO-DET ベンチマークで提案手法が最高精度を達成したことを示した．

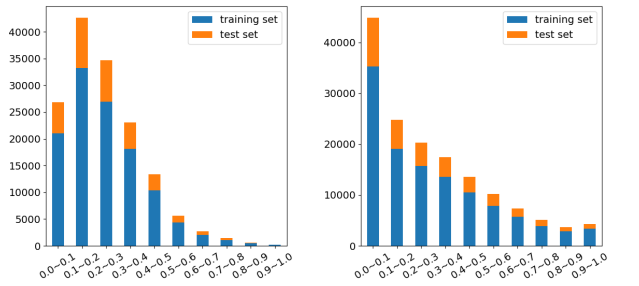
キーワード　HOI 検出，Transformer

1. はじめに

近年，人間と物体のインタラクション（HOI）検出は，大きな応用可能性を持つ分野として注目されている．HOI 検出アプローチでは，人間と物体の間の意味的関係を抽出し，画像内の ⟨human, object, action⟩ のトリプレットセットを予測する．具体的には，HOI インスタンスは人間と物体のバウンディングボックスのペアであり，対応するアクションクラスはそれらの間の関係を表す．HOI 検出は，物体検出と人間と物体のインタラクション認識の 2 つの部分の組み合わせとみなすことができる．

One-stage アプローチでは，インタラクションポイントベースの手法 [5], [6], [9], [16], [18] と Transformer [17] ベースの手法 [3], [7], [15], [20] がある．人間と物体のペアを検出し，対応するアクションクラスを並行して認識できる．最近の Transformer ベースの手法は良い結果を達成しているが，CNN バックボーンの低解像度特徴マップを利用するだけで，Transformer エンコーダに空間情報の抽出の負担を残している．また，Transformer の学習は収束が遅い問題点があり，物体検出タスクでの事前学習モデルが必要である．

図 1(a) に示すように，広く使われている HOI 検出データセット HICO-DET [2] では，人間と物体のボックスの中心距離が画像サイズの 3 分の 1 以上ある HOI インスタンスが普通に存在する．**図 1(b)** に示すように，ほとんどの HOI インスタンスは，人間と物体のボックスの面積は画像サイズの 0.1 倍未満である．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ために，本論文では，**Q**uery-based **A**chor を用いた新しい Transformer ベースの one-stage 手法 QAHOI を提案する．また，本論文は最初に one-stage の HOI 検出に関する Transformer ベースのバックボーンを研究し，HOI 検出タスクに対する大きな可能性を示す．



(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

図 1　HICO-DET [2] データセットにおける HOI インスタンスの空間分布

2. 関連研究

2.1 One-stage アプローチ

One-stage アプローチは特別なデザインで提案され，一般的に 2 ブランチアーキテクチャを採用している．PPDM [9] は，人間の中心点と物体の中心点の中点をインタラクションポイントと定義する．インタラクションポイントを介して人間と物体のインスタンスを一致させる．GGNet [18] は，特徴マップの各ピクセルの周囲に action-aware points (ActPoints) を推論することにより，インタラクションポイントのアイディアを拡張したものである．しかし，インタラクションポイントを用いる手法では，HOI インスタンスを明らかにするためのマッチングやクラスタリングのプロセスが必要である．

一方，Transformer の self-attention メカニズムを利用してコンテキスト情報を抽出し，埋め込みによって HOI インスタンスを表現する Transformer ベースの手法 [3], [7], [15], [20] は，HOI 検出タスクの新しいトレンドとなっている．Tamura ら [15] は，Transformer ベースの物体検出器 DETR [1] の物体検出ヘッドをインタラクション検出ヘッドに変換し，HOI インスタンス

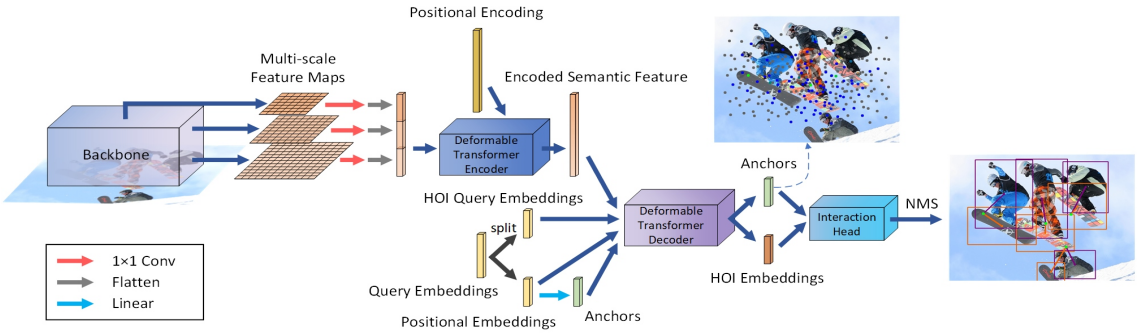


図 2　モデル全体図

の全ての要素を直接予測する QPIC を提案した．同様に，Zou ら [20] は CNN バックボーンと Transformer を組み合わせて，クエリ埋め込みから直接 HOI インスタンスを予測する．Chen ら [3] と Kim ら [7] は，HOI インスタンスのボックスとアクションクラスをデコードするために，インスタンスデコーダとインタラクションデコーダを並行して構築し，Transformer ベースの two-branch アーキテクチャを提案した．

3. 手　　法

Deformable DETR [19] は，Deformable マルチスケールアテンションモジュールを設計し，DETR におけるアテンションの複雑度を空間サイズに応じた線形複雑度に軽減し，マルチスケール Transformer ベースの物体検出を実現した．QPIC が DETR を HOI 検出タスクに拡張したのと同様に，QAHOI は，このアイディアに従い，Deformable Transformer を HOI 検出に適用した．QAHOI は，クエリ埋め込みを利用してアンカーを生成し，HOI 埋め込みをデコードすることで，Deformable Transformer デコーダを HOI インスタンス検出器に適応させる．QAHOI の全体的なアーキテクチャを**図 2** に示す．

3.1　マルチスケール特徴抽出器

Self-attention メカニズムを持つ Transformer は，画像から意味的情報を抽出することに優れているため，最近の one-stage の手法 [3], [7], [15], [20] は CNN バックボーンと Transformer エンコーダからなる特徴抽出器を構築する．これらの特徴抽出器はバックボーンからの低解像度の特徴マップを使用するため，小さなスケールの空間情報を抽出することが困難である．モデルのパフォーマンスを向上させるため，QAHOI は**図 2** に示すように，階層型バックボーンと Deformable Transformer エンコーダを組み合わせて，マルチスケール特徴抽出器を構築している．CNN ベース（ResNet）や Transformer ベースのバックボーン（Swin-Transformer [13]）を使用することができる．QAHOI は階層型バックボーンの 4 つのステージからマルチスケール特徴マップを抽出する．QAHOI は最後の 3 つのステージの特徴マップ $x_1 \in \mathbb{R}^{2C_s \times \frac{H}{8} \times \frac{W}{8}}$，$x_2 \in \mathbb{R}^{4C_s \times \frac{H}{16} \times \frac{W}{16}}$ と $x_3 \in \mathbb{R}^{8C_s \times \frac{H}{32} \times \frac{W}{32}}$ を使用する．1×1 の畳み込み層により，x_1, x_2, x_3 の特徴マップを C_s 次元から C_d 次元へ射影する．Deformable Transformer エンコーダは，意味的特徴量 $S \in \mathbb{R}^{N_S \times C_d}$ をマルチスケールで抽出し，Deformable Transformer デコーダが HOI インスタンスをデコードする際に提供する．N_S はバックボーンからの 3 つの特徴マップの画素数の和である．

3.2　アンカーベースのデコーディング処理

Deformable DETR に従い，QAHOI の Deformable Transformer デコーダのクエリ埋め込みは，等しく 2 つの部分に分割される．HOI クエリ埋め込み $Q_{HOI} \in \mathbb{R}^{N_q \times C_d}$ と位置埋め込み $Q_{Pos} \in \mathbb{R}^{N_q \times C_d}$ である．アンカー $P \in \mathbb{R}^{N_q \times 2}$ は位置埋め込み Q_{Pos} から線形層を介して生成されたものである．Deformable Transformer エンコーダからの意味特徴量 S，HOI クエリ埋め込み Q_{HOI} とアンカー P を用いて，HOI 埋め込み $E \in \mathbb{R}^{N_q \times C_d}$ を Deformable Transformer デコーダのアテンションメカニズムによりデコーディングする．Deformable Transformer デコーダのデコーディング処理を**図 3** に示す．HOI クエリ埋め込みの self-attention は，位置埋め込みを用いて，マルチヘッドアテンションモジュール [17] で計算する．アンカーは Deformable Transformer エンコーダの出力から意味的特徴を集約してマルチスケール Deformable Attention [19] を計算する．Self-attention とマルチスケール attention を N_L 個のデコード層で N_L 回計算し，最後の層で HOI インスタンスを予測するためにインタラクション検出ヘッド用の HOI 埋め込みを出力する．

3.3　アンカーベースのインタラクション検出ヘッド

QPIC に従い，QAHOI はシンプルなインタラクション検出ヘッドを設計しており，アンカーをベースにして HOI の全ての要素を予測する．QAHOI におけるインタラクションヘッドの予測処理を**図 4** に示す．アンカーセット $P \in \mathbb{R}^{N_q \times 2}$ の各アンカー (p_x, p_y) は，人間と物体のペアのボックスの基点として使用する．インタラクションヘッドにおいて Feed-foward Network（FFN）が予測する人間と物体のボックス要素 B^h，$B^o \in \mathbb{R}^{N_q \times 4}$ は，$\{d_x, d_y, w, h\}$ で構成される．ここで，d_x と d_y は，アンカーとボックスの中心とのオフセット，w と h は，ボックスの幅と高さを表す．最終的なバウンディングボックス $\hat{B}^h$，$\hat{B}^o$ は $\{d_x + p_x,\ d_y + p_y,\ w, h\}$ で構成される．最後に，物体ボックスの物体クラス $O \in \mathbb{R}^{N_q \times K_o}$ と HOI インスタンスのアクションクラス $A \in \mathbb{R}^{N_q \times K_a}$ を，人間と物体のバウンディングボックス $\hat{B}^h, \hat{B}^o$ と組み合わせて，出力 HOI インスタンスを構築する．

3.4　Top K スコアと HOI NMS

QAHOI は，マルチスケール特徴を抽出するために十分なアンカーが必要である．一般に，アンカー数は画像中の HOI イ

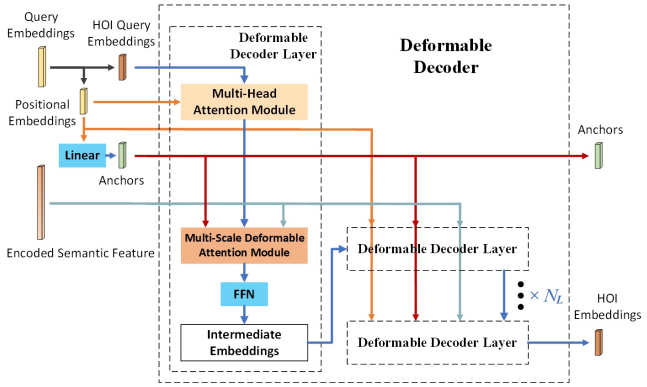


図 3　QAHOI のデコーディング処理

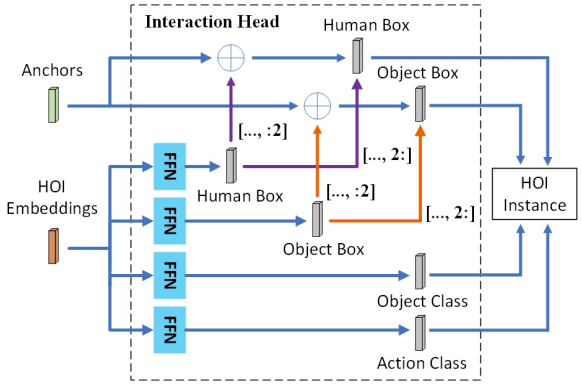


図 4　アンカーベースのインタラクション検出ヘッド

ンスタンス数を大きく上回っている．HICO-DET データセットでは，96% の画像に 10 個以下の HOI インスタンスしか含まれていない．QAHOI は，2 つのステップで結果をフィルタリングする．まず，物体クラスのスコアが上位 N_t の HOI インスタンスが選択される．その後，HOI Non-Maximal Suppression（NMS）を用いて，最終的な結果をフィルタリングする．HOI NMS は，HOI インスタンス間の人間と物体の Intersection of Union（IoU）と，HOI スコアに基づいて算出される．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，$c_{\text{HOI}} = c_o \cdot c_a$ となる．HOI インスタンス i と j の間の人間と物体の複合 IoU は，次のように計算される

$$\text{IoU}(i,j) = \text{IoU}(B_i^{(h)}, B_j^{(h)}) \cdot \text{IoU}(B_j^{(o)}, B_j^{(o)}) \qquad (1)$$

物体検出タスクと同様に，IoU に基づき，各アクションクラスのスコアが低い HOI インスタンスを閾値 δ で除去する．

3.5　モデルの学習と推論

人間と物体をペアで予測するため，インタラクションポイントに基づくアプローチで重要となるマッチング処理が不要になる．QPIC の学習手順に従って，ハンガリアン法 [8] を用いて，N_q 個の予測値を全て 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 [11] 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．

4.　実　　験

4.1　実 験 設 定

データセット　47,776 枚の画像（トレーニングセット 38,118 枚，テストセット 9,658 枚）を含む HICO-DET [2] データセットで実験を行った．HICO-DET には 117 のアクションクラスと 80 の物体クラス（物体クラスは MS-COCO [12] データセットと同じ）が含まれている．アクションクラスと物体クラスで 600 の HOI クラスが構成されている．データセットに含まれる 600 の HOI クラスのインスタンス数に基づいて，これらの HOI クラスは 3 つのカテゴリ *Full*（全ての HOI クラス），*Rare*（インスタンスが 10 個未満の 138 クラス），*Non-Rare*（インスタンスが 10 個以上の 462 クラス）に分類される．

評価指標　予測された HOI インスタンスの評価には，mAP（mean average precious）が使用される．True Positive の HOI インスタンスでは，予測された人間のボックスと ground-truth の人間のボックスの間の IoU が 0.5 より高く，予測された物体と ground-truth の物体のボックスの間の IoU も 0.5 より高くなっている．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．

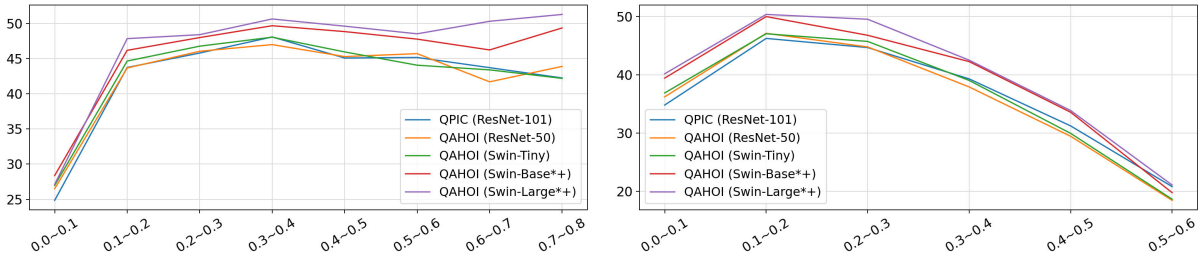
学習設定　バックボーンには，ImageNet [4] で事前に学習させた Swin-Transformer [13] をベストモデルとして QAHOI を学習させる．具体的には，ImageNet-1K で事前学習した Swin-Tiny と Swin-Base，ImageNet-22K で事前学習した Swin-Base と Swin-Large を使用している．Deformable Transformer エンコーダとデコーダはともに 6 層（$N_L = 6$）である．クエリ埋め込み数は $N_q = 300$ であり，物体スコアにより上位 $N_t = 100$ の HOI インスタンスがモデルの出力として選択される．NMS 処理では，$\delta = 0.5$ を使用する．Swin-Tiny，Swin-Base，Swin-Large をバックボーンとした場合，第 1 ステージの特徴マップの次元は $C_s = 96$，$C_s = 128$，$C_s = 192$ となる．Deformable Transformer の埋め込み次元は $C_d = 256$ である．バックボーンの学習率を 10^{-5}，その他を 10^{-4}，重みの減衰を 10^{-4} とした AdamW [14] オプティマイザを使用する．バッチサイズ 16（1GPU あたり 2 枚画像，8GPU）で 150 エポックの学習を行った．

4.2　最先端手法との比較

HICO-DET において最先端手法と比較した結果を**表 1** に示す．マルチスケール特徴マップとマルチスケール Deformable Attention を用いて，モデルの物体検出部分に有益な検出器を事前学習しない場合でも，Swin-Large バックボーンを用いた QAHOI は，最先端の one-stage 手法の QPIC と比べて，5.88 mAP（相対 19.7%）上回った．ImageNet-20K で事前学習した Swin-Base バックボーンを用いた QAHOI の mAP は，ImageNet-1K で事前学習した同じバックボーンより 4.1（相対 13.9%）高くなる．分類タスクで事前学習したバックボーンの性能が上がれば上がるほど，HOI 検出の精度がさらに向上することがわかった．

4.3　アブレーション実験

CNN ベースと Transformer ベースのバックボーンを用いてア



(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

図 5　異なる空間スケールでの評価結果

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**

表 1　この図は HICO-DET において最先端手法との比較である．* と $^+$ は，ImageNet-22K で 384×384 の入力解像度で事前学習したことを表す．

Arch.	Model	Backbone	Fine-tuned Detection	Multi-scale	Default		
					Full	*Rare*	*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

表 2　アーキテクチャについてのアブレーション実験

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

表 3　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

表 4　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

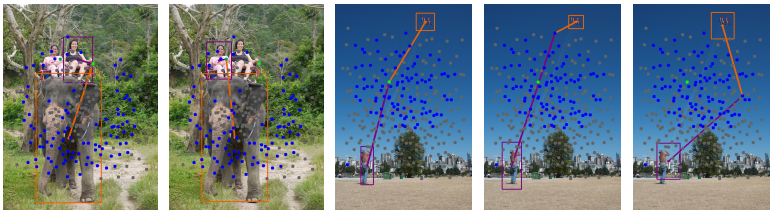
表 5　NMS 処理のアブレーション実験

ブレーション実験を行った．CNN ベースのバックボーンには ResNet-50 を用い，ゼロからスタートと検出器の重みの fine-tune という 2 つの学習方法の違いを調査した．

学習方法　QPIC と同様に MS-COCO データセットで学習した Deformable DETR の重みを用いて QAHOI を初期化し，その後 HICO-DET データセットで QAHOI を fine-tune する．Deformable DETR の実装に従い，特徴マップ x_3 に対して 3×3 の畳み込み層を用いて，低解像度特徴マップ $x_4 \in \mathbb{R}^{C_d \times \frac{H}{64} \times \frac{W}{64}}$ を追加する．また，QAHOI と QPIC をそれぞれ ResNet-50 と Swin-Tiny でゼロから学習する実験を行った．**表 2** の結果から，物体検出器の学習を行わない場合，(4) QAHOI-ResNet-50 または (7) Swin-Tiny は，(1) QPIC-ResNet-50 または (3) Swin-Tiny と比較して，*Full* および *Non-Rare* カテゴリで良い結果を達成し，QPIC を上回ることがわかった．

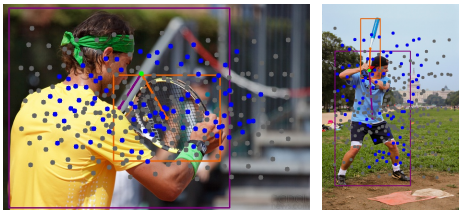
マルチスケール特徴マップ　Swin-Tiny バックボーンを用いて，特徴マップの組み合わせの違いによる提案手法の精度への影響を調査した．**表 2** (6)(7) の結果より，特徴マップを追加しても精度は向上しない．QAHOI の (7)(8)(9) のモデルでは，マルチスケール特徴マップの削除に伴い，精度が低下している．(9) と (7) を比較すると，3 つの特徴マップを用いることで，em Full カテゴリにおいて 1.82 mAP（相対 6.8%）の精度向上が得られている．

バックボーン　Swin-Tiny は ResNet-50 とモデルサイズや計算量が似ているが，ImageNet での精度は ResNet-50 より高い．物体検出器の学習を行わない場合，**表 2** (1)(4) の ResNet-50 で学習したモデルと比較して，Transformer ベースのバックボーン Swin-Tiny は，(3) QPIC（2.98 mAP，相対 12.3%）と (7) QAHOI (4.12 mAP，相対 16.9%) の精度を向上させることができる．(7) QAHOI-Swin-Tiny は精度，改善ともに (3) QPIC-Swin-Tiny より優れており，提案手法が優れた設計のバックボーンに基づいて大きな可能性を持っていることがわかった．**表 1** にある Swin-Base と Swin-Large で学習した QAHOI の結果からも，分類タスクでより精度の高いバックボーンを用いることで HOI 検出の精度を大幅に向上できることがわかった．(5) Deformable DETR から fine-tune した QAHOI の結果は，(2) DETR から fine-tune した QPIC より低い．QPIC は 500 エポックの学習を行った DETR を使用するのに対し，QAHOI は 50 エポックの学習を行った Deformable DETR を使用することである．Deformable DETR の学習エポック数が不足であるため，fine-tune するとき学習の収束が遅くなることが原因だと考えられる．



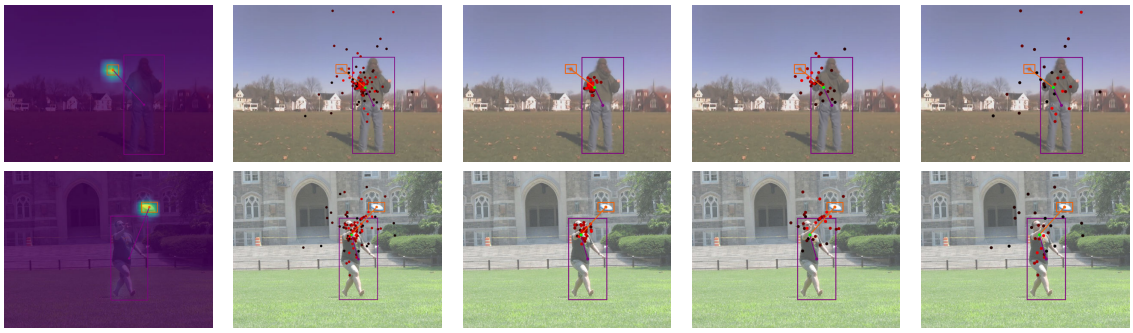
(a) ride, elephant (b) ride, elephant (c) fly, kite (d) fly, kite (e) fly, kite

図 6　アンカーの柔軟性.



(a) (b)

図 7　アンカーの分布



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

図 8　Top 1 スコア HOI インスタンスのアテンションの可視化．(a) は QPIC の Transformer デコーダの最終層におけるアテンションマップを示す．(b)~(e) は QAHOI の Deformable Transformer デコーダの最終層のサンプリング点のアテンションを示す．(b) は全解像度のサンプリング点，(c)~(e) はそれぞれ異なる解像度（特徴マップ x_1, x_2, x_3 に従って）のアテンションである．アンカーによりアテンションが高いサンプリング点は赤色で表示されている．

後処理のアブレーション実験　QAHOI ではフィルタリング処理が重要であり，**表 3** では，top K スコアステップと NMS ステップにより，*full* カテゴリで，QAHOI-Swin-Tiny の精度が 1.83 mAP 向上している．また，Top K スコアと HOI NMS のパラメータを調整してアブレーション実験を行った．Top K のステップを最適化するために，異なる種類のスコアと K の数値をテストした．**表 4** の結果から，アクションスコアを使用するより物体スコアを使用する方が適切だと考える．c_a と $c_a \cdot c_o$ の結果は同じであり，アクションスコアが Top K に影響を与えないことがわかった．物体スコアの上位 100 個の HOI インスタンスを出力とする条件で最良の結果が得られた．HOI NMS の IoU 計算と閾値をテストした結果を**表 5** に示す．IoU^h と IoU^o は，2 つの HOI インスタンス間の人間または物体ボックスを使用して IoU を計算することを表す．人間または物体ボックスのみを使用して重複する HOI インスタンスをフィルタリングする場合，人間ボックスの方が良い結果を得られることがわかった．2 つの HOI インスタンスの重なり程度を表す複合 IoU，$\text{IoU}^h \cdot \text{IoU}^o$ を用いると，IoU 閾値を $\delta = 0.5$ とすることで最適な結果が得られることがわかった．

4.4　異なる空間スケールでの評価

QAHOI のマルチスケールアーキテクチャは，小さな物体の検出には有利である．異なる空間スケールでの検出能力を調べ，QPIC の評価方法と同じ，最先端の方法と比較するために，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．結果は**図 5**(a) と**図 5**(b) に示す．実験において，HICO-DET テストセット中の ground-truth HOI インスタンスを 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．

人間や物体の領域が小さいうちは，低解像度の特徴マップではインタラクティブ情報を含む領域の特徴を抽出することは難しい．**図 5**(a) では，最初の 3 つのビンにおける小さな HOI インスタンスの検出において，Transformer ベースのバックボーンを持つ QAHOI が ResNet-101 を持つ QPIC より優れていることがわかった．Deformable DETR から fine-tune した ResNet-50 を用いた QAHOI は，0.0~0.1 のビンにおいて ResNet-101 を用いた QPIC を上回り，0.1~0.2 のビン，0.2~0.3 のビンにおいては同等の結果を得ることができた．また，Swin-Large と Swin-Base を用いた QAHOI は大きなインスタンスで良好な性能を発揮することができる．

HOI インスタンスの人間と物体の距離が短いほど，特徴を区別することが難しくなる．**図 5**(b) では，人間と物体のボックスの中心点の距離が画像サイズの 0.3 倍未満である場合，QAHOI は QPIC よりも精度が高いと示した．QPIC，QAHOI ともに，人間と物体の距離が離れると精度が低下するが，バックボーンを改善することでこの問題を緩和することができる．

4.5　定量的な結果

クエリベースのアンカーの柔軟性　クエリベースのアンカーは，マルチスケール特徴マップから特徴を抽出することができるため，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．アンカーの柔軟性を**図 6** に示す．青色と灰色の点は，物体クラススコアが上位 100 位までの選択アンカーと未選択アンカーを表し，緑色の点は，検出された各

HOI インスタンスの HOI スコアが最も高いアンカーを表す．**図 6** (a)(b) に示すように，女性 2 人が象に乗っており，インタラクションポイントに基づく手法とは異なり，最も信頼度の高いインタラクションを検出するアンカーは，人間と物体のペアの中心から遠くても，人に近くてもよいのである．**図 6** (c)(d)(e) は，結果を上位 3 つのアクションクラスのスコアで順に表示する．このシーンでは，人間と物体が離れているため，(c) に人間と物体のペアの真ん中にあるアンカーは，人間をよく検出するが，物体をうまく囲んでいない．(d) に，アンカー物体に近いが人間からは離れており，物体はよく検出するが，人間をうまく囲んでいない．また，(e) に，アンカーが人間からも物体からも離れており，人間と物体のペアを見つけることはできても，それぞれをうまく囲んでいない．この 3 つのアンカーは位置が異なるが，正確な人間と物体のペアに注目し，正解の HOI インスタンスを出力できる．

図 7 は，さらに HOI インスタンスが 1 つの画像に対する物体クラスのスコアが上位 100 のアンカー分布を示している．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が，HOI スコアが高いアンカーは，テニスラケットを持つ人のように，人間と物体の中心に限定されていない．以上の定量的な結果から，クエリベースアンカーは HOI インスタンスに対して，有効な手法であることがわかった．

図 8 では，QPIC-ResNet-101 と QAHOI-Swin-Large の両方のアテンションを可視化している．(a) により，QPIC は人間やインタラクションの領域より，物体に着目していることがわかった．(b) により，QAHOI はアンカー周辺に注目し，サンプリング点は人間や物体の周囲に柔軟に配置される．(c) に示すように，高解像度・低レベルのサンプリング点は，アンカーに近い特徴を抽出する．また，(d)(e) では，低解像度・高レベルのサンプリングポイントは，広い領域で特徴を抽出することができる．

5. おわりに

本論文では，マルチスケール特徴を抽出するための階層的バックボーンと Transformer エンコーダ，HOI 埋め込みをデコードするための Transformer デコーダ，HOI インスタンスを予測するためのインタラクション検出ヘッドで構築された one-stage の HOI 検出フレームワーク QAHOI を提案した．Transformer デコーダとインタラクションヘッドは，クエリに基づくアンカーを利用して HOI 埋め込みをデコーディングし，HOI インスタンスを予測する．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 検出において大きな改善を示し，クエリベースのアンカーも HOI インスタンス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を持ち，物体検出のようにアンカーを活用して HOI インスタンスを検出するため，いくつかの改善点を追加することが可能だと考える．例えば，マルチスケール特徴を強化するために，Feature Pyramid Networks（FPN）[10] などを追加することができる．アンカーの予測は HOI のプロポーザルとして利用でき，two-stage Deformable DETR のように 2 段階で精度をさらに上げることが可能だと考える．また，物体検出器を事前に学習させる必要がなく，提案手法は大規模なモデルをゼロから学習させることができ，最先端の結果を得ることができる．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．

文　献

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[3] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[6] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. UnionDet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020.

[7] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. HOTR: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021.

[8] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

[9] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. PPDM: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020.

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

[15] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.

[16] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020.

[17] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[18] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and Gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021.

[19] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.

[20] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021.