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.

分離されたデコーダとノイズ除去学習を用いたHOI検出

陳　　俊文† 王　　瀛成† 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,wang-y}@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　現在の one-stage HOI 検出手法は，物体デコーダの検出ターゲットを変更し，ボックスターゲットがクエリ埋め込みから明示的に分離されていないため，学習収束が遅い．本研究では主語デコーダ，物体デコーダ，動詞デコーダからなる新しい one-stage フレームワークを提案する．さらに，学習効率を向上させるために，学習可能な物体と動詞のラベル埋め込みを用いたノイズ除去学習方法を提案する．HICO-DET で本手法は学習エポックの 3 分の 1 で最先端手法より 6.39% 高い精度を達成することを示した．

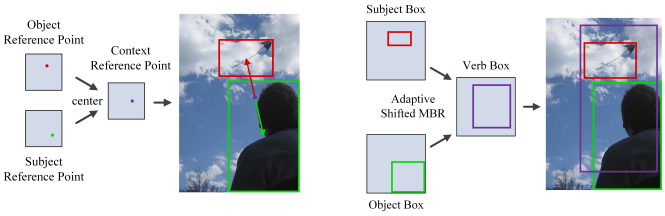
キーワード　HOI 検出，Transformer

1. はじめに

最近の HOI（Human-Object Interaction）検出の研究は，主に物体検出のフレームワークに基づいて構築されている．最も広く使用されるデータセットである HICO-DET [2] と V-COCO [6] は，MS-COCO データセット [16] と同じ物体カテゴリを共有している．HOI インスタンス $\{B_s, (B_o, O), V\}$ は，主語（人間）ボックス B_s，クラス O を持つ物体ボックス B_o，動詞クラス V のトリプレットの定義に従い，検出方法は one-stage と two-stage に分かれる．

One-stage アプローチでは，検出効率が高く，トレーニングが容易であるため，近年注目されている．最初に，CNN を用いた one-stage 手法 [14], [21], [24] は，インタラクションポイントを利用して，人間と物体間のインタラクションの確率を検出し，有望なパフォーマンスを達成した．

Transformer のアテンションメカニズム [5] は，特徴マップの異なる位置にある特徴の関係を扱い，グローバルなコンテキスト情報を抽出する上で，CNN アーキテクチャよりも柔軟である．DETR [1] は，Transformer を用いて，物体検出の one-stage フレームワークを提案した．最近，Transformer ベースの HOI 検出手法 [4], [10], [20], [27] は，DETR [1] を採用することにより，アテンションメカニズムのメリットを示した．QPIC [20] は，HOI 検出問題を集合予測問題として捉え，DETR と同様の学習パイプラインを使用している．QPIC は，one-stage および two-stage の CNN ベースの手法におけるマッチング処理を行わず，エンコーダ・デコーダのアーキテクチャを採用し，インタラクションヘッドを用いて HOI インスタンスを直接予測する．しかし，QPIC の単一のデコーダは，人間と物体の位置関係やインタラクション認識の特徴が混ざっているため，HOI の予測精度が低下する．単一のデコーダの設計を改善するため，物体検出とインタラクション認識をカスケード的に分離した one-stage 手法 [8], [15], [22], [23], [25] が提案された．カスケード型の設計は HOI の予測精度を向上させるが，インスタンスデコーダでは人間と物体の検出はまだ混ざっているため，物体検出タスクで事前学習したモデルの性能を活用していない．



(a) MSTR　　(b) SOV

図 1　Deformable DETR ベースの HOI 検出手法のパイプライン

DN-DETR [13] は，物体検出タスクの Denoising 学習方法を提案し，ground-truth を入力クエリとして利用することで，物体検出の性能を大きく向上させた．HOI 検出において，DOQ [19] はオラクルクエリを導入し，人間と物体ペアの ground-truth ボックスと物体ラベルを符号化し，ground-truth の HOI インスタンスを再構成するためにデコーダを学習するように誘導する．このように，DOQ は ground-truth に関する知識をデコーダに抽出し，モデルの性能と学習収束を向上させた．しかし，オラクルクエリは検出と認識に共有されるため，ground-truth 情報の利用による学習加速効果は制限されている．

本論文の貢献は主に以下の 2 点である．

- HOI 検出の要素を分離した新たな one-stage フレームワ (SOV) を提案する．
- Ground-truth から位置とラベルの事前知識を学習する新たな（Split Target Guided, STG）Denoising 学習方法を提案する．提案した手法 SOV-STG は，HOI 検出ベンチマークにおいて，現在の最先端手法よりも 3 倍少ない学習エポック数（HICO-DET では 30 エポック）でより高い精度を達成した．

2. 関連研究

最近の研究 [3], [11] では，Deformable-DETR [26] のアテンションメカニズムを利用して，DETR ベースの学習収束が遅い点を改善する．QAHOI [3] は，Deformable Transformer デコーダの

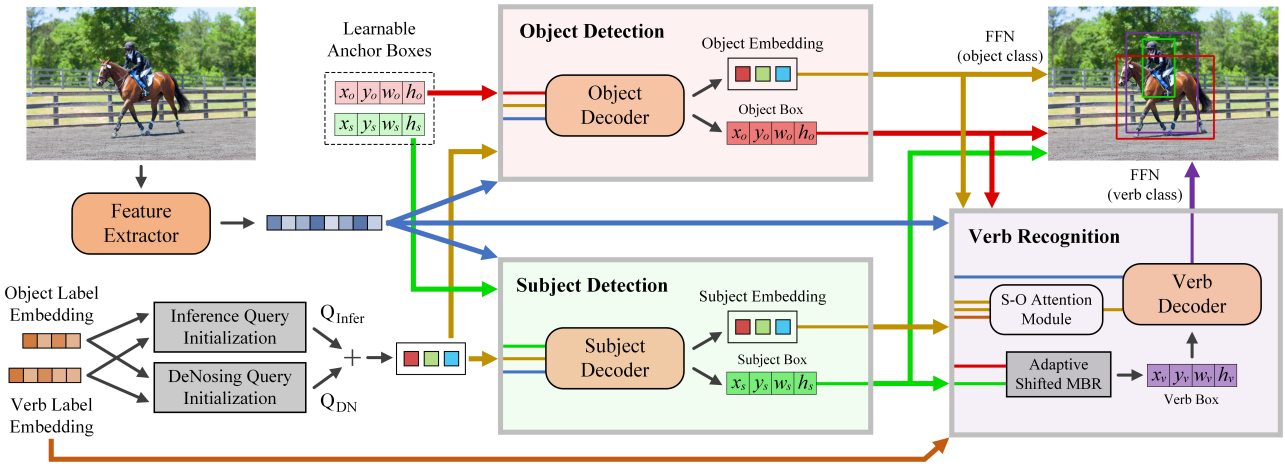


図 2　SOV-STG のフレームワークの全体図

参照点を HOI インスタンスのアンカーと見なし，アンカーを用いて人間と物体検出を誘導する．QAHOI は HOI クエリの埋め込みから参照点の埋め込みを分割しているが，QPIC と同じように，HOI クエリの埋め込みは HOI インスタンスの全要素の予測に使われることに変わらない．図 1(a) では，MSTR [11] が人間，物体，コンテキストの参照点を用いて HOI インスタンスを表現し，参照点に基づいて人間，物体，動詞を予測する．しかし，MSTR のクエリ埋め込みは，HOI インスタンスの最終的なボックスとラベルの予測に使用され，クエリ表現は複数のタスクに共有されるため，学習が遅くなる．特定の用途のためのクエリ埋め込みを明確にするために，本論文では，DAB-Deformable-DETR [17] のアテンションメカニズムを活用する SOV を提案した．図 1(b) に示すように，SOV は学習可能な人間（subject）と物体（object）のアンカーボックスを使用し，人間と物体のボックスを直接予測する．

3.　手　　法

図 2 は，SOV-STG のフレームワークを示している．SOV は特徴抽出器と SOV デコーダから構成される．学習可能なアンカーボックスとラベル埋め込みは，推論とノイズ除去学習のために HOI に特化した事前知識を提供する．ネットワーク全体はエンコーダ・デコーダの設計に従っており，end-to-end で学習できる．

3.1　アンカーボックスによる HOI インスタンスの予測

クエリ埋め込みのデコーディングターゲットを明確にするために，SOV フレームワークは DAB-Deformable-DETR [17] のアテンションメカニズムを活用し，学習可能な subject と object のアンカーボックスを直接使用して人間と物体のボックスを予測する．アンカーボックスは層ごとに更新され，最終層の subject と object のボックスが verb ボックスを形成するために使用される．図 1(b) に示すように，本研究では，adaptive shifted minimum bounding rectangle（ASMBR）を提案し，人間ボックスと物体ボックスの空間的な関係を考慮しながら動詞ボックスを生成している．図 3 に示すように，デコーダの最終層で予測された人間ボックス $\boldsymbol{B}_s = (x_s, y_s, w_s, h_s)$ と物体ボックス $\boldsymbol{B}_o = (x_o, y_o, w_o, h_o)$（$(x, y)$：ボックス中心）を与えると，ASMBR（動詞ボックス）は，次のように定義される：

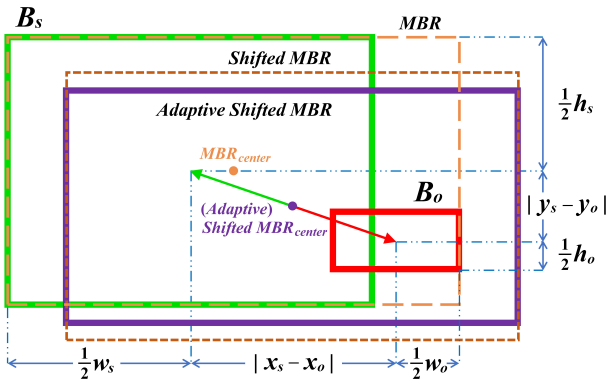


図 3　ASMBR のデザイン

$$\boldsymbol{B}_v = \left(\frac{x_s + x_o}{2}, \frac{y_s + y_o}{2}, w_v, h_v \right) \tag{1}$$

$$w_v = \frac{w_s + w_o}{2} + |x_s - x_o|, h_v = \frac{h_s + h_o}{2} + |y_s - y_o| \tag{2}$$

MBR の縮小 (adaptive) と移動 (shift) の目的は，インタラクション領域から遠い関連性が低い情報を取り除き，インタラクション領域周辺のコンテキスト情報をより多くカバーすることである．

3.2　SOV Decoders

デコーディングターゲットを明確にするために，分離されたデコーダの設計が重要である．QAHOI [3] や MSTR [11] と同様に，CNN バックボーンと Deformable Transformer エンコーダ [26] を利用して，マルチスケールグローバル特徴量 $\boldsymbol{f}_g \in \mathbb{R}^{N_g \times D}$ を抽出する．ここで，N_g はマルチスケール特徴マップの全画素数，D は全 Transformer アーキテクチャにおける埋め込みの潜在次元である．図 2 に示すように，グローバル特徴量は学習可能なアンカーボックスを持つ主語および物体デコーダに入力される．物体検出器の検出能力を維持するため，フィードフォワード（FFN）ヘッドを持つ物体デコーダが検出タスクで学習されたものと同じ重みから初期化される．

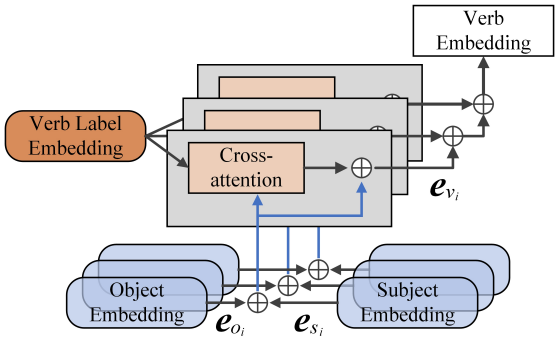


図 4　S-O アテンションモジュール

さらに，物体デコーダの重みを用いて，主語デコーダの初期化を行い，主語デコーダの学習負担を軽減させる．主語デコーダと物体デコーダは，人間アンカーボックス $\boldsymbol{B}_s$ と物体アンカーボックス $\boldsymbol{B}_o$ とクエリ埋め込み $\boldsymbol{e}$ を層ごとに並列に更新する．物体デコーダからの物体埋め込み $\boldsymbol{e}_o$ を用いて物体クラスを予測し，人間ボックスと物体ボックスを用いて動詞ボックス $\boldsymbol{B}_v$ を生成する．次に，物体と人間の埋め込みを S-O アテンションモジュール（セクション 3.3）に入れ，動詞の埋め込みを融合させる．最後に，人間ボックスと物体ボックスと動詞の埋め込みから生成された動詞ボックスを動詞デコーダに与え，動詞クラスを予測する．

3.3　動詞デコーダと S-O アテンションモジュール

提案した動詞ボックス（ASMBR）は人間と物体のボックスから直接生成されるため，動詞デコーダは動詞ボックスの予測学習をすることなく，動詞認識に集中することができる．図 2 に示すように，動詞認識部分は主に S-O アテンションモジュールと動詞デコーダの 2 つの部分から構成されている．特徴量融合時に動詞ラベルの知識を統合するために，S-O アテンションで動詞ラベルの事前分布を融合させる．さらに，S-O アテンションに bottom-up path を設計し，下層から上層への情報を強化させる．図 4 で，S-O アテンションモジュールの計算を示している．i 番目の層 ($i > 1$) から，人間の埋め込み $\boldsymbol{e}_{s_i} \in \mathbb{R}^{N_q \times D}$ と物体の埋め込み $\boldsymbol{e}_{o_i} \in \mathbb{R}^{N_q \times D}$ を与え，N_q はクエリ数であると仮定する．まず，GEN-VLKT [15] と同様に，S-O アテンシンモジュールも人間と物体の埋め込みを multi-layer sum で融合する．そして，融合した埋め込み量 $\boldsymbol{e}_{so_i} = (\boldsymbol{e}_{o_i} + \boldsymbol{e}_{s_i})/2$ を用いて，動詞ラベル埋め込み量 $\boldsymbol{t}_v$ とのクロスアテンションの計算を行う．動詞ラベルの基礎知識として学習可能な動詞ラベル埋め込み $\boldsymbol{t}_v \in \mathbb{R}^{N_q \times D}$ については，次のセクション 3.4 で紹介する．さらに，レイヤーの情報を強化するために，bottom-up path を追加する．最後に，bottom-up path を追加した後の動詞埋め込み $\boldsymbol{e}_{v_i}$ は次のように定義できる：

$$\begin{aligned}\boldsymbol{e}_{v_i} =&((\mathrm{CrossAttn}(\boldsymbol{e}_{so_{i-1}}, \boldsymbol{t}_v) + \boldsymbol{e}_{so_{i-1}}) \\ &+ (\mathrm{CrossAttn}(\boldsymbol{e}_{so_i}, \boldsymbol{t}_v) + \boldsymbol{e}_{so_i}))/2\end{aligned} \tag{3}$$

3.4　Split Label Embeddings

図 2 に示すように，SOV デコーダのクエリ埋め込みを初期化するために，2 種類の学習可能なラベル埋め込みを使用している．物体検出のノイズ除去学習方法 [13] とは異なり，本手法は学習と推論の両方でラベル埋め込みを使用する．これにより，推論するとき，最初からラベルの事前知識を入力クエリとして使用することができる．D 次元の C_o 個ベクトルからなる（C_o は物体クラス数，D は Transformer の潜在次元）Object label embedding $\boldsymbol{t}_o \in \mathbb{R}^{C_o \times D}$ は，物体ラベルの事前分布（知識）として定義する．同様に，verb label embedding $\boldsymbol{t}_v \in \mathbb{R}^{C_v \times D}$ を動詞ラベル事前分布として定義する．物体ラベルと動詞ラベルの事前分布を用いて，物体ラベルのクエリ埋め込み $\boldsymbol{q}_o \in \mathbb{R}^{N_q \times D}$ と動詞ラベルの埋め込み $\boldsymbol{q}_v \in \mathbb{R}^{N_q \times D}$ を線形結合により初期化する．次に，物体ラベルと動詞ラベルの埋め込みを加算して，推論クエリ埋め込み $\boldsymbol{q}_{ov} \in \mathbb{R}^{N_q \times D}$ が得られる．線形結合は 2 つの学習可能な行列 $\boldsymbol{A}_o \in \mathbb{R}^{N_q \times C_o}$ と $\boldsymbol{A}_v \in \mathbb{R}^{N_q \times C_v}$ を用いて以下のように定義される．

$$\boldsymbol{q}_o = \boldsymbol{A}_o \boldsymbol{t}_o, \quad \boldsymbol{q}_v = \boldsymbol{A}_v \boldsymbol{t}_v \tag{4}$$

$$\boldsymbol{q}_{ov} = \boldsymbol{q}_o + \boldsymbol{q}_v \tag{5}$$

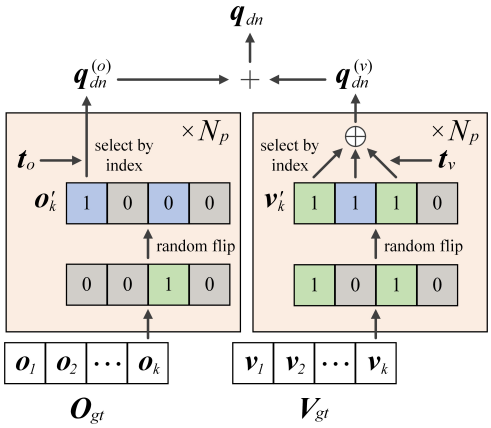


図 5　DN クエリの生成

3.5　Split Target Guided Denoising

物体ラベルと動詞ラベルは HOI 検出のターゲットであるため，2 つのラベル埋め込みは分割した事前分布と見なすことができる．ノイズ除去クエリ埋め込み（DN クエリ）は分割した事前分布から生成され，ノイズ除去学習の監督に用いられるので，提案したノイズ除去学習方法を **S**plit **T**arget **G**uided (STG) denoising と名付ける．図 5 では，DN クエリの初期化と，ground-truth HOI インスタンスにノイズを追加する過程を示している．Ground-truth の物体ラベル集合 $\boldsymbol{O}_{gt} = \{\boldsymbol{o}_i\}_{i=1}^k$ と動詞ラベル集合 $\boldsymbol{V}_{gt} = \{\boldsymbol{v}_i\}_{i=1}^k$ を与えると，2 種類のラベル DN クエリが初期化されている．ここで，$\boldsymbol{o}_i$ と $\boldsymbol{v}_i$ は物体クラスと動詞クラスの one-hot ラベルであり，k は ground-truth の HOI インスタンス数である．DN-DETR [13] に従い，k 番目の ground-truth の HOI インスタンスに対して，物体ラベル $\boldsymbol{o}_k$ の ground-truth のインデックスを他の物体クラスのインデックスにランダムに反転させ，ノイズ物体ラベル $\boldsymbol{o}_k'$ を得て，N_p グループのノイズラベルが生成される．次に，物体 DN クエリ $\boldsymbol{q}_{dn}^{(o)} \in \mathbb{R}^{N_p \cdot k \times D}$ が，物体ラベル埋め込み $\boldsymbol{t}_o$ から，ノイズ物体ラベル $\boldsymbol{O}_{gt}'$ のイ

Method	Backbone	Default			Known Object		
		Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
CNN-based							
UnionDet [9]	ResNet-50-FPN	17.58	11.72	19.33	19.76	14.68	21.27
IP-Net [21]	Hourglass-104	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [14]	Hourglass-104	21.73	13.78	24.10	24.58	16.65	26.84
GGNet [24]	Hourglass-104	23.47	16.48	25.60	27.36	20.23	29.48
Transformer-based							
QAHOI [3]	ResNet-50	26.18	18.06	28.61	-	-	-
AS-Net [4]	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [20]	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
CDN-S [23]	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42
MSTR [11]	ResNet-50	31.17	25.31	32.92	34.02	28.83	35.57
Zhou *et al.* [25]	ResNet-50	31.75	27.45	33.03	34.50	30.13	35.81
CDN-B [23]	ResNet-50	31.78	27.55	33.05	34.53	29.73	35.96
GEN-S [15]	ResNet-50	31.88	26.24	33.57	-	-	-
DOQ (CDN-S) [19]	ResNet-50	33.28	29.19	34.50	-	-	-
SOV-STG-S	ResNet-50	32.97	29.28	34.07	35.58	31.73	36.73
SOV-STG-S+CCS	ResNet-50	33.63	**30.40**	34.59	36.24	**32.09**	37.48
SOV-STG-B	ResNet-50	**33.81**	29.51	**35.09**	**36.44**	31.78	**37.83**
SOV-STG-Swin-L	Swin-Large-22k	**43.62**	**43.36**	**43.70**	**45.67**	**44.70**	**45.96**

表 1　HICO-DET での結果

Method	AP^{S1}_{role}	AP^{S2}_{role}
UnionDet [9]	47.5	56.2
IP-Net [21]	51.0	-
AS-Net [4]	53.9	-
GGNet [24]	54.7	-
HOTR [10]	55.2	64.4
QPIC [20]	58.8	61.0
MSTR [11]	62.0	65.2
SOV-STG-B	62.0	63.8

表 2　V-COCO での結果

Verb Box	Default		
	Full	*Rare*	*Non-Rare*
Object Box	31.93	27.08	33.38
Subject Box	32.15	26.97	33.70
MBR	32.20	27.55	33.59
SMBR	32.44	27.98	33.78
ASMBR	**32.97**	**29.28**	**34.07**

表 3　動詞のボックスのデザイン

Denoising Strategies				Default		
#	Box	Obj	Verb	*Full*	*Rare*	*Non-Rare*
(1)				32.48	27.76	33.89
(2)	✓			31.71	26.94	33.13
(3)	✓	✓		32.11	27.03	33.63
(4)	✓		✓	32.26	28.39	33.41
(5)		✓	✓	32.24	28.37	33.40
(6)	✓	✓	✓	**32.97**	**29.28**	**34.07**

表 4　ノイズ除去のアブレーション実験

Method	Default		
	Full	*Rare*	*Non-Rare*
SOV-STG-S	32.97	29.28	34.07
-STG	32.44	27.30	33.98
-DN	30.61	24.91	33.32
-Subject Decoder	29.87	24.92	31.35
-Verb Decoder	28.74	22.63	30.57

表 5　各モジュールの貢献

ンデックスによって収集される．動詞ラベルは co-occurrence ground-truth クラスがあるため，co-occurrence ground-truth インデックスがノイズ動詞ラベルに現れるように，ground-truth 動詞ラベルの他のインデックスをランダムに反転してノイズ動詞ラベル $\boldsymbol{v}'_k$ を生成する．物体 DN クエリと同じように，動詞ラベル DN クエリ $\boldsymbol{q}^{(v)}_{dn} \in \mathbb{R}^{N_p \cdot k \times D}$ は，動詞ラベル埋め込み $\boldsymbol{t}_v$ の中から，ノイズ動詞ラベル $\boldsymbol{V}'_{gt}$ のインデックスによって選択された動詞ラベル DN 埋め込みを合計したものである．最後に，物体 DN クエリと動詞 DN クエリを連結し，ノイズ除去学習用の DN クエリ $\boldsymbol{q}_{dn} \in \mathbb{R}^{2N_p \cdot k \times D}$ を形成する．このように，ノイズ除去学習により分割した事前分布を学習し，SOV の推論を誘導することができる．

3.6　学習と推論

提案するフレームワーク SOV-STG は，end-to-end で学習できる．推論クエリ埋め込みには，Hungarian algorithm [12] を用いて，ground-truth の HOI インスタンスと予測 HOI インスタンスをマッチングし，マッチングコストと学習損失は QAHOI [3] と同じである．さらに，ノイズ除去部分と推論部分は同じ損失関数で学習される．同じ ground-truth ラベルの反転率は，学習初期にはモデルにとってノイズ除去が困難であるが，学習中に容易できるようになるという考えに基づき，dynamic DN scale factor $\gamma \in (0,1)$ を導入した．物体ラベル反転率 $\eta_o \in (0,1)$ と動詞ラベルノイズ除去率 $\eta_v \in (0,1)$ を学習エポックに応じて制御することによってノイズ除去学習をさらに改善する．Dynamic DN scale factor では，ラベル反転率 η を学習開始時に $\gamma \cdot \eta$ に設定し，学習中に η まで線形に増加させる．また，動詞ラベル反転率 $\lambda_v \in (0,1)$ は，η_v で選択されたマルチホット動詞ラベル内の要素の反転率を制御するために使用される．ボックスノイズ除去は DN-DETR [13] と同じである．

4. 実　　験

4.1　実 験 設 定

データセット　HICO-DET [2] と V-COCO [6] データセットで実験を行った．HICO-DET は 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）から構成されており，データセットは，600 種類の HOI クラス（117 種類のアクションクラスと 80 種類の物体クラスの組合せ）のインスタンス数によって，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）に分けられる．V-COCO データセットには，トレーニング用の 5,400 画像とテスト用の 4,946 画像が含まれている．COCO [16] と同じ 80 種類の物体クラスと 29 種類の動詞クラスがアノテーションされており，29 種類の動詞クラスがあるシナリオ 1 と 25 種類の動詞クラスがあるシナリオ 2 の 2 つのシナリオ設定がある．

#	S-O Attention Designs			Default		
	last layer	multi-layer	Attention	*Full*	*Rare*	*Non-Rare*
(1)	✓		S-O Fuse	**32.97**	**29.28**	**34.07**
(2)		✓	S-O Fuse	30.80	25.16	32.48
(3)	✓		S-O w/o bottom-up	32.57	29.12	33.60
(4)	✓		Sum Fuse	32.54	27.61	34.01
(5)		✓	Sum Fuse	29.73	24.90	31.17

表 6　異なる S-O アテンション設計のアブレーション実験

評価指標　評価指標は mAP（mean average precious）を採用する．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 より高くなる必要がある．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　特徴抽出器，主語デコーダ，物体デコーダの重みを初期化するために，COCO データセットで学習した DAB-Deformable-DETR を採用する．特徴抽出器は，ResNet-50 [7] バックボーンと 6 層の Deformable Transformer エンコーダから構成される．CDN [23] と同様に，全てのデコーダの層数を調整することにより，SOV-STG の 2 つのバリエーションを実装し，3 層デコーダの SOV-STG-S と 6 層デコーダの SOV-STG-B

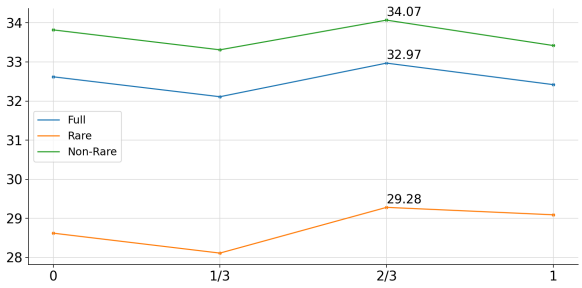


図 6　Dynamic DN scale factor γ の効果

と表記する．Transformer の潜在次元は $D = 256$，クエリ数は $N_q = 64$ とする．DN 部分では，各 ground-truth HOI インスタンスに対して，$2N_p = 6$ グループのノイズラベルを生成する．Dynamic DN scale factor を $\gamma = \frac{2}{3}$ とし，ボックスの反転率を $\delta_b = 0.4$，物体ラベルの反転率を $\eta_o = 0.3$，動詞ラベルノイズ除去率を $\eta_v = 0.6$，物体ラベル反転率を $\lambda_v = 0.6$ として学習開始時と同じノイズ除去レベルを維持する．HICO-DET データセットに対して，AdamW オプティマイザーで学習率 2e-4（バックボーンは 1e-5），重み減衰 1e-4 でモデルを学習する．バッチサイズは 32（GPU あたり 4 枚画像），学習エポックは 30（20 エポックで学習率減衰）とし，CDN [23] の 3 分の 1，QPIC [20] と QAHOI [3] の 5 分の 1 としている．V-COCO については，オーバーフィッティングを防ぐため，バックボーンをフリーズしている．全ての実験は 8 枚の NVIDIA A6000 GPU で行っている．

4.2　最先端手法との比較

表 1 では，HICO-DET データセットにおいて，提案した SOV-STG と最近の SOTA 手法を比較した．ResNet-50 をバックボーンとする SOV-STG-B は，Default 設定の *Full* カテゴリで 33.81mAP を達成した．Deformable Transformer を用いた one-stage の手法である QAHOI や MSTR と比較すると，アンカーポイントを用いた手法と比較して，SOV-STG はアンカーボックス事前分布とラベル事前分布の知識を受け，それぞれ 29.14% と 8.47% の mAP 改善を達成した．さらに，SOV-STG-B は CDN-B を 1/3 の学習エポックで 6.39% の mAP で上回る．また，SOV-STG-S+CCS は ground-truth 情報を明示的に利用するため，ground-truth を利用して学習を行う DOQ と比較して，DOQ と同じデータセット拡散方法（CCS）を利用して DOQ（80 エポック）より少ない学習エポック（30 エポック）で 1.05%mAP の改善を達成した．SOV-STG はデコーディングターゲットを完全に分離し，学習ターゲットを明確にすることで，SOV-STG-S は GEN-S [15] を 3.42% 上回る精度を達成した．V-COCO において，表 2 に示すように，SOV-STG-B は AP^{S1}_{role} で 62.0 mAP を達成し，QPIC を 5.44% で上回っていることが示されている．また，Swin Transformer [18] を用いてベストモデルの SOV-STG-Swin-L を学習した，43.62 mAP で新しい SOTA を達成した．

4.3　アブレーション実験

SOV-STG-S モデルを用いて，HICO-DET データセットでアブレーション実験を行った．

各モジュールの貢献　SOV-STG は柔軟なデコーディングアーキテクチャと学習方法で構成されている．各提案モジュールの貢献を明らかにするため，表 5 では，提案モジュールを一つずつ削除し，アブレーション実験を行った．2 行目は，STG を削除し，推論クエリ埋め込みを学習可能な埋め込みから初期化する一般的な DN 学習に置き換えた実験である．1 行目の結果と比べて，STG は *Full* カテゴリにおいて 1.63% 精度を向上させることが分った．次に，3 行目では，DN 学習を削除した．その結果，3 行目と 1 行目を比較すると，ground-truth から事前知識を学習しない場合，性能が 7.71% 低下することが分かった．しかし，DN 学習を用いない場合でも，提案したフレームワークは QPIC（ResNet-50）に対して，5 分の 1 の学習エポックで 5.30% の大幅な性能向上が得られる．次に，4 行目で主語デコーダと S-O アテンションモジュールを削除し，人間と物体ボックスの両方を物体デコーダで更新する．検出のデコーディング負担のバランスを取らない場合，3 行目と比較して，精度は 2.48% 低下する．最後に，動詞デコーダを削除した 5 行目と 4 行目を比較すると，精度は 3.93% 低下する．動詞デコーダから洗練された動詞表現は動詞予測に役立つことを示した．

S-O アテンションモジュール　S-O アテンションメカニズムの効果を調べるため，試したデザインの異なるバリエーションを表 6 に示した．1 行目は SOV-STG-S で使用されている S-O アテンションモジュールを示している．2 行目では，GEN [15] と同様に，全ての層の融合された埋め込みを動詞デコーダに入力する．しかし，その結果，*Full* カテゴリでは 7.05%，*Non-Rare* カテゴリでは 16.38% 精度が低下している．4 行目では，S-O アテンションのクロスアテンションを削除し，1 行目と比較して 1.32% 精度が低下している．クロスアテンションにより，動詞表現を抽出する能力が向上していることが分かった．同様に，5 行目は GEN と同じようなアーキテクチャを試したが，精度は低下した．精度低下の原因は，Deformable Transformer のアテンションが局所的なアテンションメカニズムであり，異なる層で元特徴の異なる部分に注目するためと考えられる．具体的には，動詞デコーダのサンプリング点は，物体デコーダや主語デコーダのサンプリング点とは関係がなく，グローバルな意味特徴の異なる位置に着目している．

ノイズ除去学習方法　表 4 では，ボックス座標，物体ラベル，動詞ラベルの 3 つの部分について，ノイズ除去学習方法を実験した．6 行目は SOV-STG-S の結果である．1 行目では，ground-truth のボックス座標，物体ラベル，動詞ラベルをノイズなしで直接モデルへ入力する．その結果，6 行目の完全なノイズ除去学習と比較して，精度が 1.51% 低下していることが分かった．1 行目と 2 行目，5 行目と 6 行目の結果を比較すると，ボックスのノイズ除去のみではパフォーマンスが低下するが，物体と動詞のラベルのノイズ除去とともにボックスのノイズ除去を使用すると，パフォーマンスを向上させることができる．3 行目と 4 行目の結果では，動詞ラベルとボックスのノイズ除去の方が，物体ラベルとボックスのノイズ除去より優れていることを示した．また，6 行目の *Rare* の結果は 3

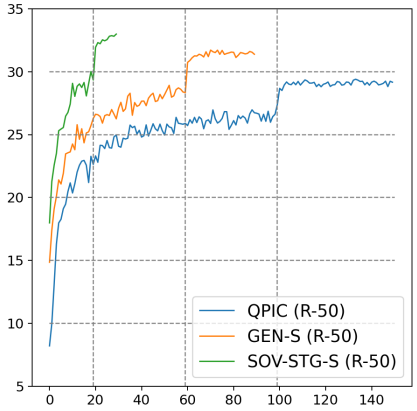


図 7　SOTA との学習中の精度の比較.

行目の結果より 8.32% 高く，動詞ラベルのノイズ除去が希少な HOI インスタンスの検出に有効であることを示した.

Dynamic DN　Dynamic DN scale factor は，ノイズ除去学習の難しさを調整するために使用される．図 6 では，dynamic DN scale factor γ を調整し，γ による効果を明らかにする．$\gamma = \frac{2}{3}$ としたとき，最良の性能が得られる．Dynamic DN を用いない $\gamma = 1$ と比較して，dynamic DN は主に *Full* と *Non-Rare* のカテゴリで性能を向上させることが分かった.

学習コストの軽減　提案手法が学習コストを軽減することを示すために，提案手法と SOTA モデル，QPIC と GEN の学習プロセスを可視化して比較した．図 7 に示すように，SOV-STG-S は，学習最初から高い AP を達成し，QPIC と GEN よりも早く収束することができる.

5. おわりに

本論文では，ターゲットに特化した分離されたデコーダ SOV とノイズ除去学習方法 STG を用いた新たな one-stage のフレームワークを提案する．提案したフレームワーク SOV-STG は，HOI インスタンスをボックスで表現する新しい形式を採用し，デコーディングに特化した事前知識を学習できる．また，設計されたアーキテクチャと効率的な学習方法により，より少ない学習コストで最先端の性能を達成することができる．SOV-STG は，HOI の検出を特定の事前分布とデコーダで分離しているため，それらのいずれかを改良することも容易である．今後は，言語モデルから初期化された物体ラベルや動詞ラベルの事前分布を導入し，性能向上をさらに向上させることを目指す.

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] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021.

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

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[6] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *CVPR*, 2022.

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

[10] 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.

[11] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *CVPR*, 2022.

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

[13] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022.

[14] 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.

[15] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*, 2022.

[16] 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.

[17] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *ICLR*, 2022.

[18] 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.

[19] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *CVPR*, 2022.

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

[21] 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.

[22] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. In *AAAI*, 2022.

[23] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. In *NeurIPS*, 2021.

[24] 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.

[25] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *CVPR*, 2022.

[26] 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.

[27] 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.