分離されたデコーダとノイズ除去学習を用いた HOI 検出

PRMU2023

電気通信大学 大学院 情報学専攻

陳 俊文　　王 瀛成　　柳井 啓司

Human-Object Interaction (HOI) Detection

ロ HOI 検出

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

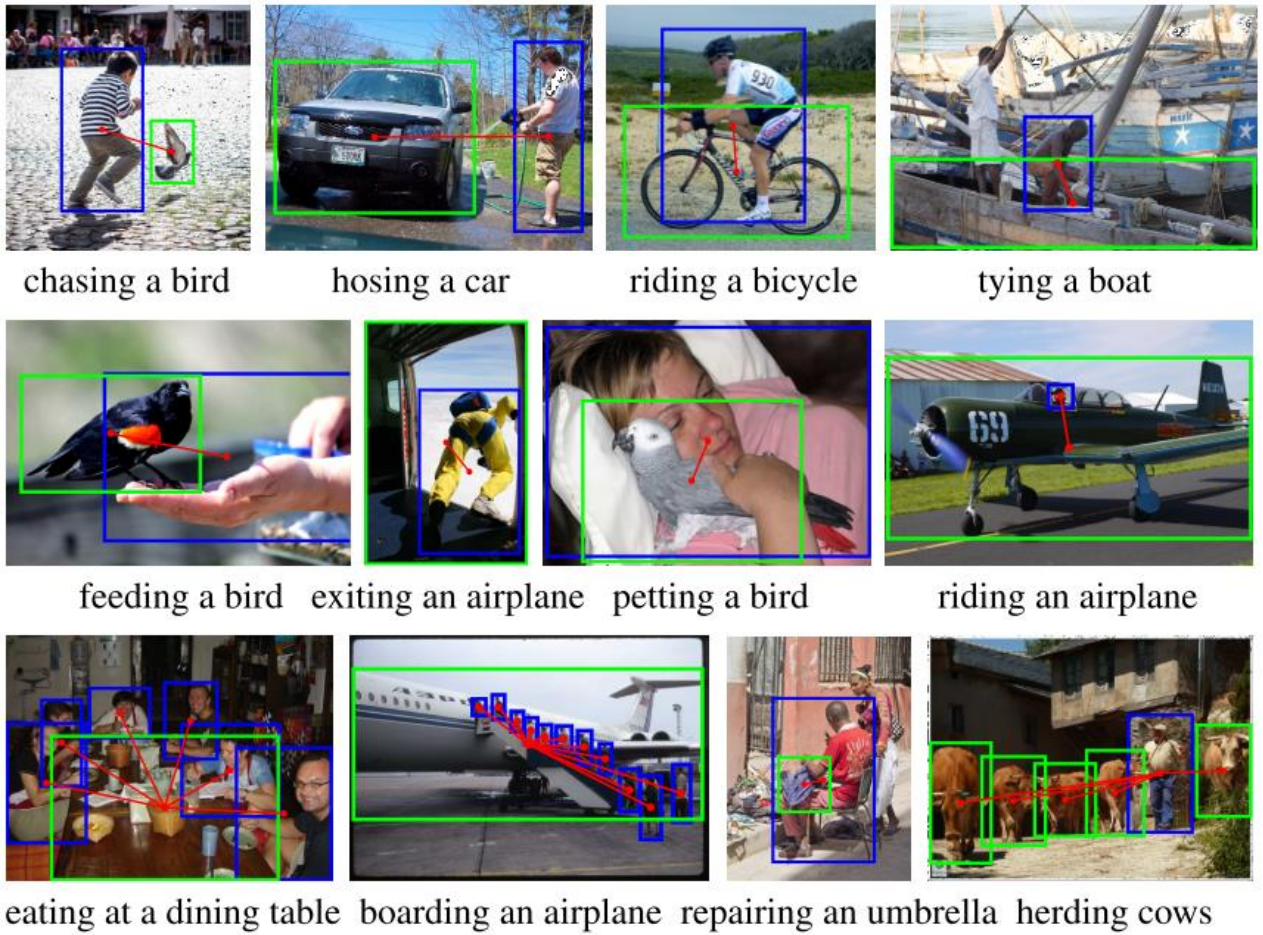
ロ HOI インスタンス

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

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

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)*, 2018.

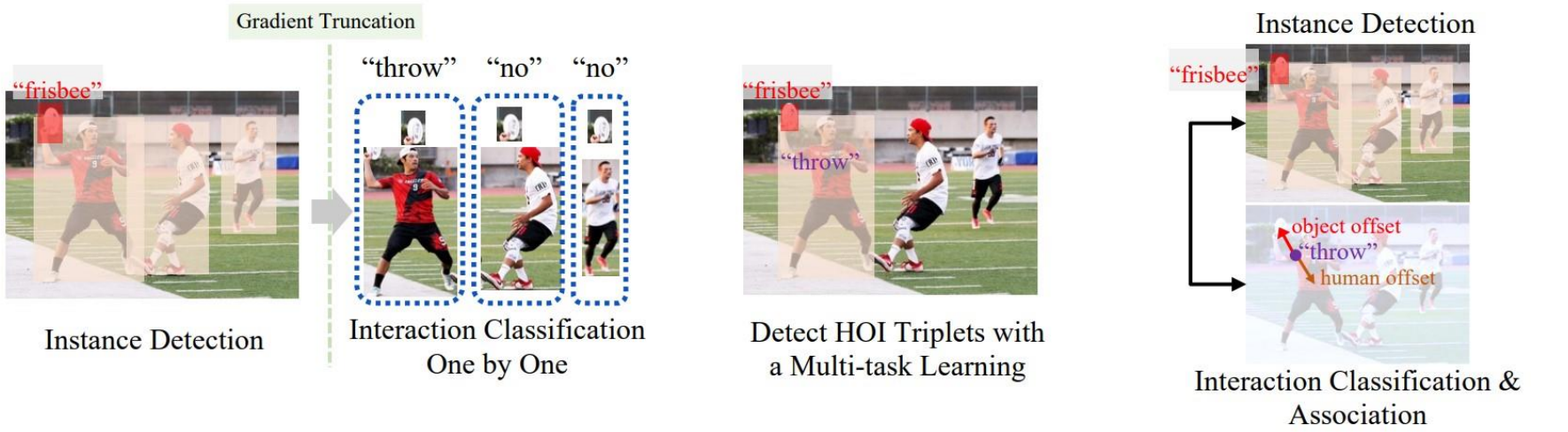
HOI 検出手法

□ 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

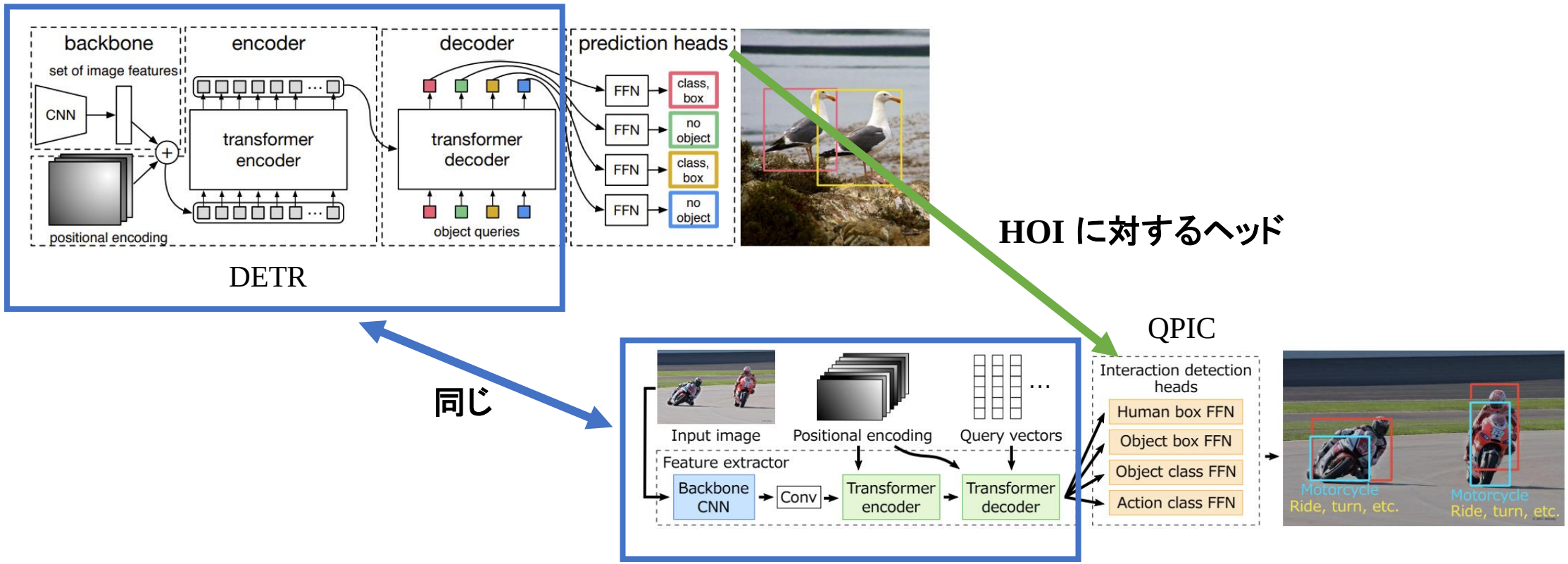
(b). One-stage end-to-end framework

(c). One-stage framework with parallel architecture

HOI 検出手法

□ Transformer-based HOI 検出手法

- Transformer-based 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



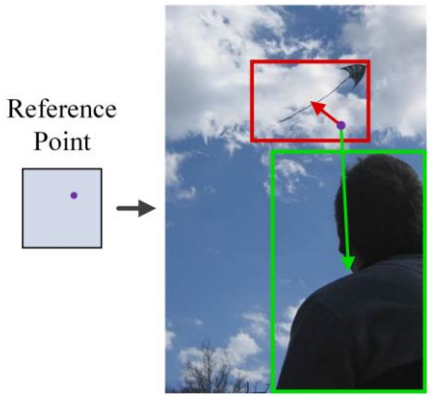
[2] Tamura, Masato, Hiroki Ohashi, and Tomoaki Yoshinaga. "QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
[3] Carion, Nicolas, et al. "End-to-end object detection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

モチベーション

□ Deformable DETR [4] ベースの one-stage 手法

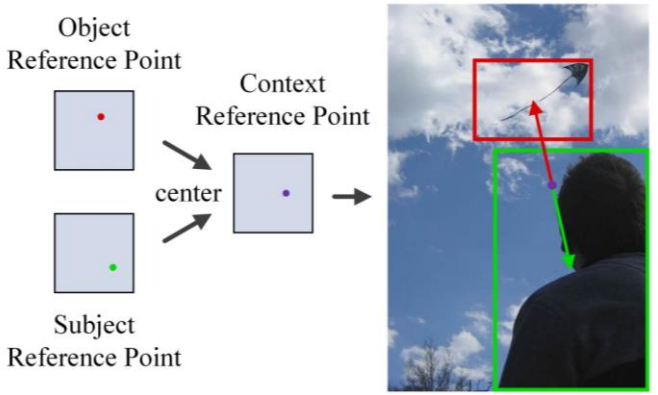
- Deformable Transformer デコーダの参照点をアンカーとして使用
- しかし，アンカーまたクエリ埋め込みは各 HOI 要素の予測に共用されている

すべての要素が同じアンカーによって予測される



(a) QAHOI [5]

同じクエリ表現は複数のタスクに共有される



(b) MSTR [6]

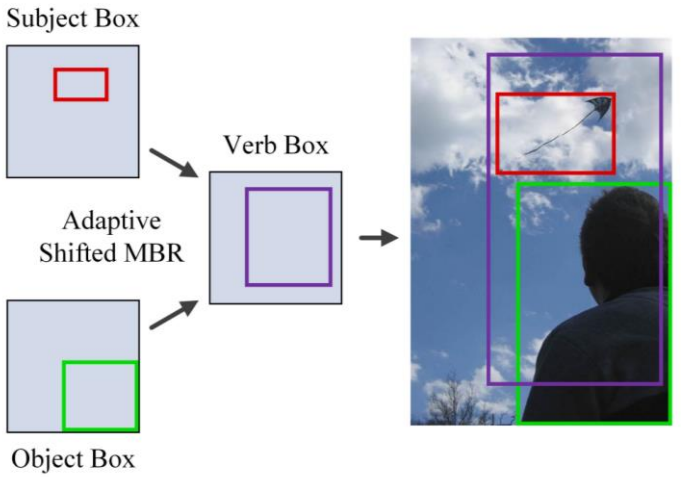
[4] Zhu, Xizhou, et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection." *International Conference on Learning Representations*, 2021.
[5] Junwen Chen and Keiji Yanai. QAHOI: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021.
[6] Kim, Bumsoo, et al. "MSTR: Multi-scale transformer for end-to-end human-object interaction detection." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

提案手法

□ Subject Object Verb – Split Target Guided (SOV-STG)

1. 人間，物体，動詞の予測を分離した Subject Object Verb（SOV）フレームワークを提案
2. アンカーボックスで HOI を表現するパイプラインを提案
3. 事前知識を学習導入する新たな Split Target Guided（STG）Denoising 学習方法を提案

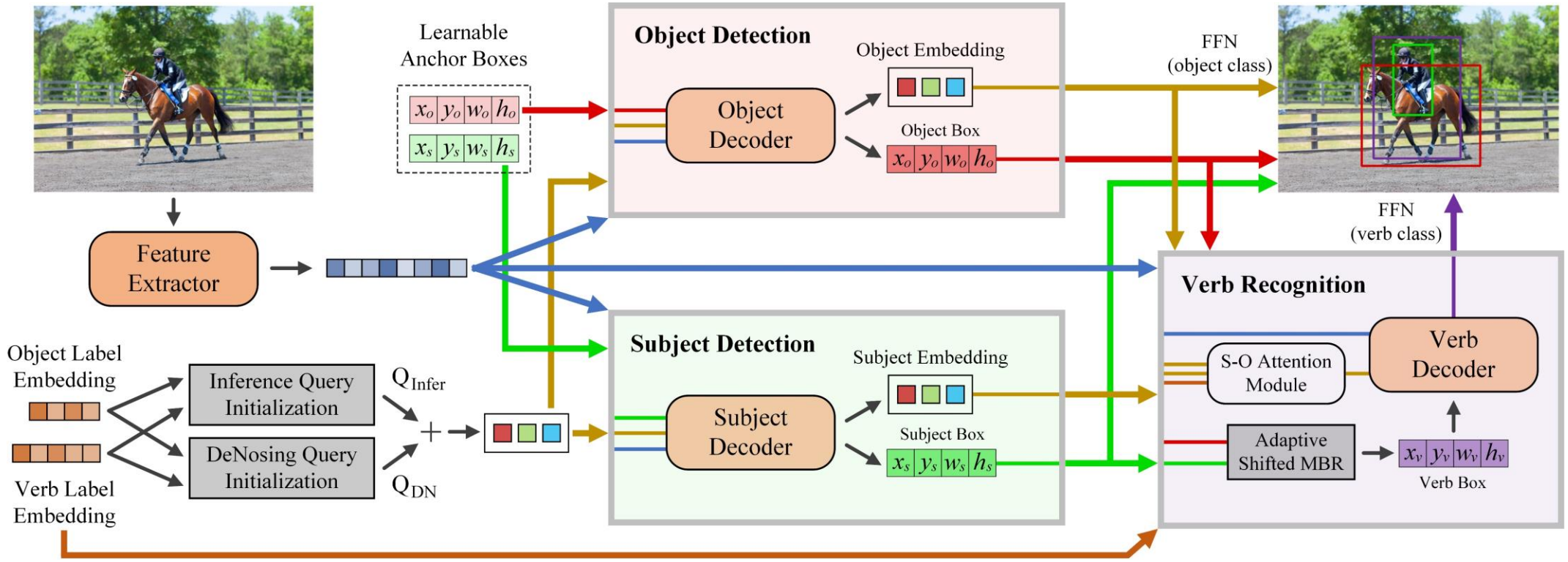
クエリ埋め込みを分離した
アンカーボックスで HOI 要素を予測



(c) SOV

SOV-STG

- DAB-Deformable-DETR [7] をベースにして，位置情報をコンテキストクエリから分離された
- 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に事前知識を提供する



SOV-STG フレームワークのパイプライン

[7] Liu, Shilong, et al. "DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR." *International Conference on Learning Representations*, 2022.

1 アンカーボックスによる HOI インスタンスの予測

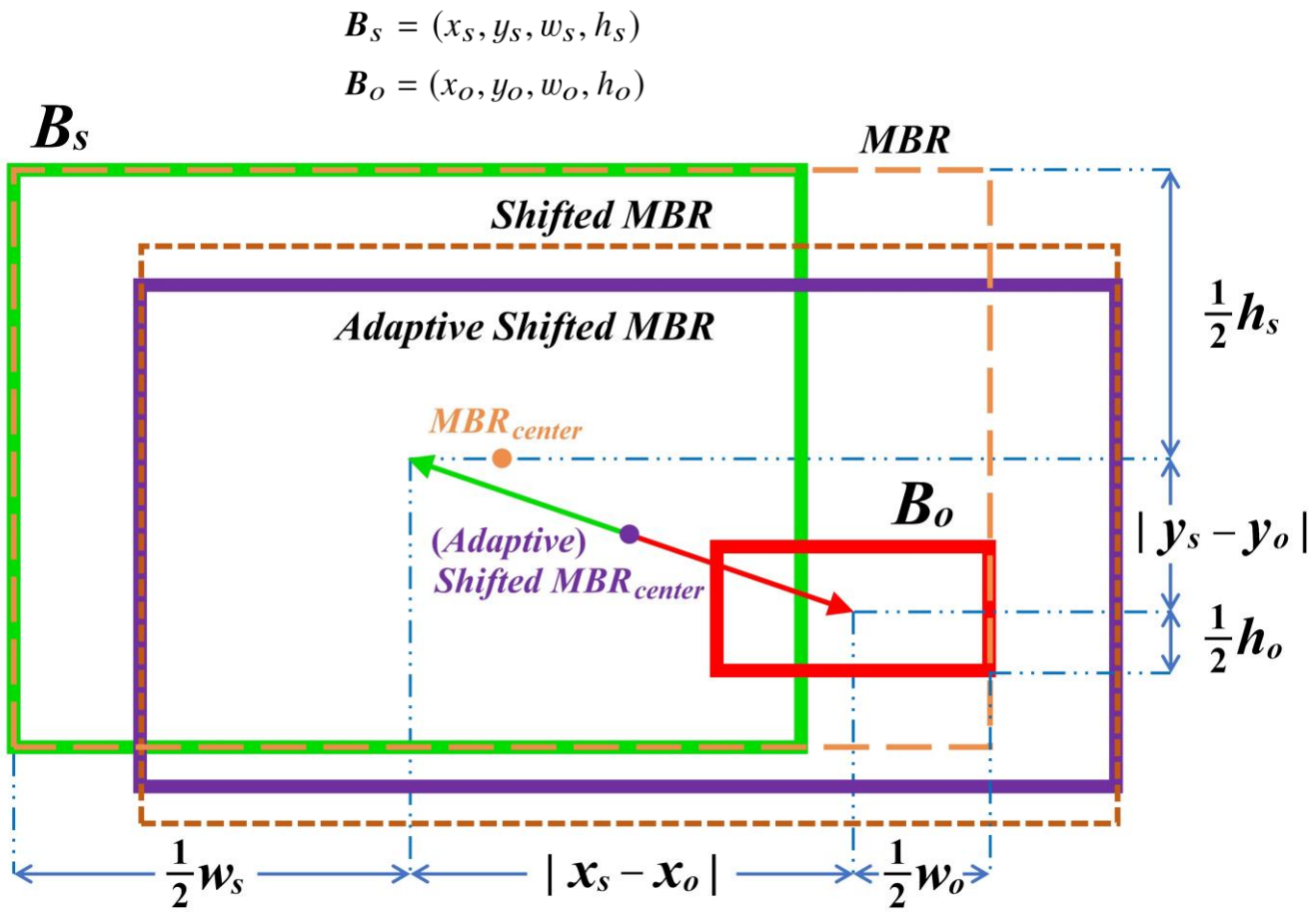
☐ Adaptive-shifted Minimum Bounding Rectangle (ASMBR)

- 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

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

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

- Adaptive: 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く
- Shift: 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

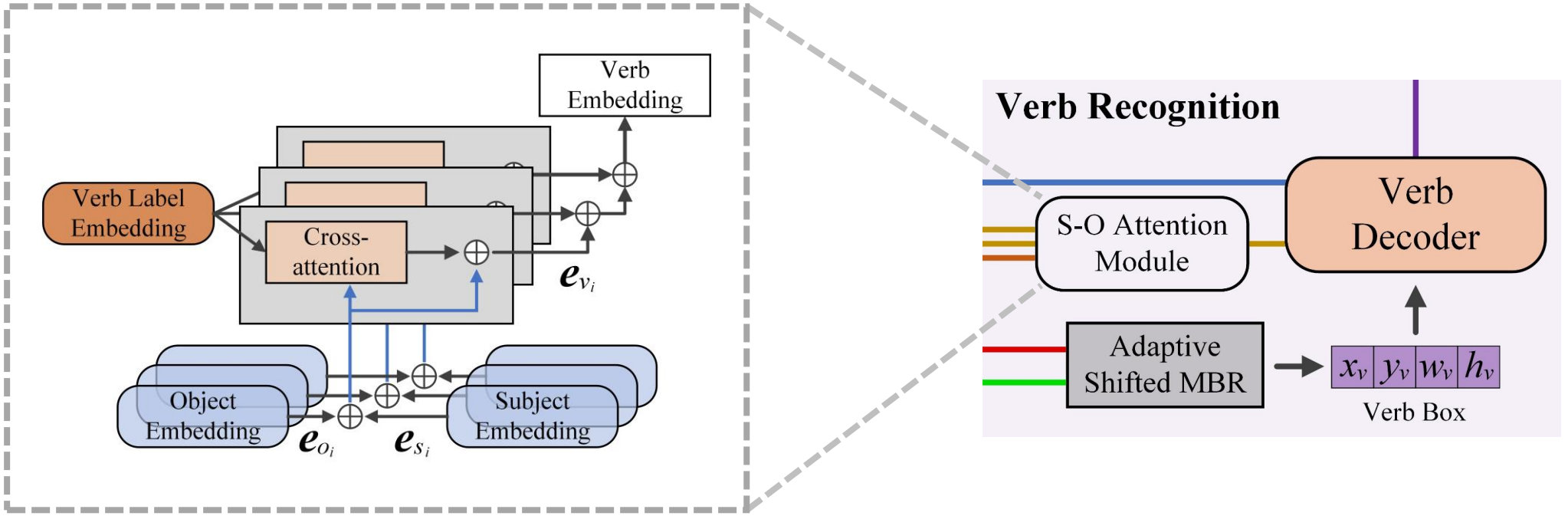


ASMBR のデザイン

2 動詞デコーダと S-O アテンションモジュール

ロ 動詞認識

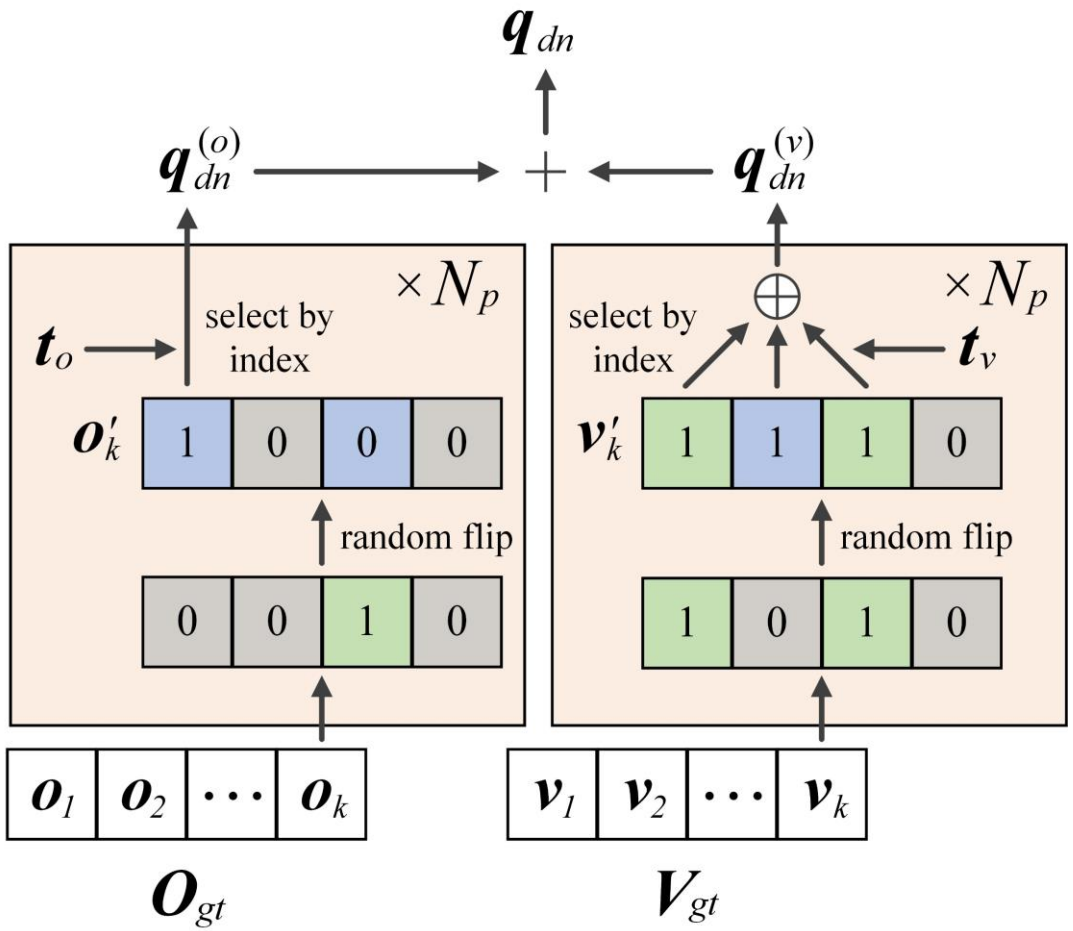
- S-O アテンションモジュールと動詞デコーダ
- 動詞ラベルの事前知識を融合させる
- Bottom-up path：下層から上層への情報を強化させる



3 ノイズ除去学習 Split Target Guided（STG）

□ DN クエリの初期化

- Ground-truth HOI インスタンスにノイズを追加
- 物体ラベル ground-truth のインデックスを他の物体クラスに反転させ，ノイズ物体ラベルを得る
- 物体 DN クエリが物体ラベル埋め込みから収集
- 物体 DN クエリと動詞 DN クエリを連結し，DN クエリを構成



DN クエリの生成

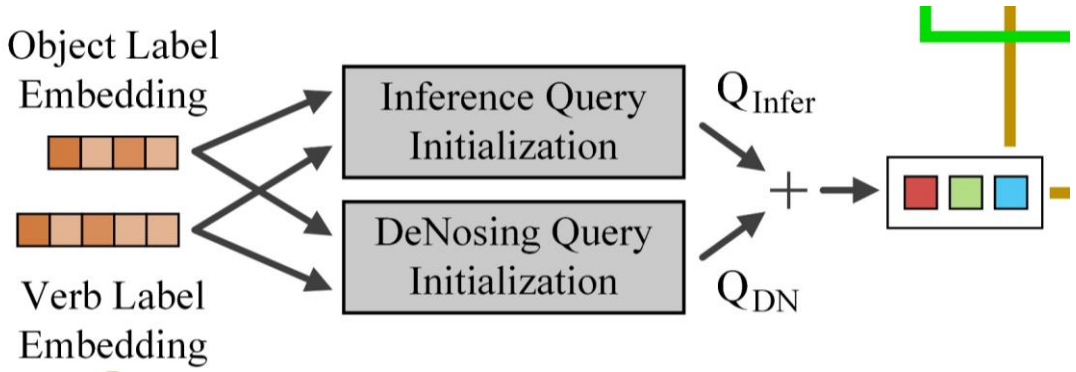
学習と推論

ロ 学習

- End-to-end で学習
- Hungarian algorithm [8] 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，損失を計算
- ノイズ除去部分と推論部分は同じ損失

ロ 推論

- ラベルの事前知識を入力クエリとして使用
- 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が得られる



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

実験

ロ データセット

- HICO-DET（トレーニングセット 38,118 枚，テストセット 9,658 枚）
  - 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
- V-COCO（トレーニングセット 5,400 枚，テストセット 4,946 枚）
  - 80 個の物体クラス
  - 29 個の動詞クラス

ロ 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

最先端手法との比較

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**

150 training epochs

150 epochs
100 epochs
50 epochs

90 epochs
80 epochs
30 epochs

+2.03
(6.39%)

表 1　HICO-DET での結果

Ablation 実験

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

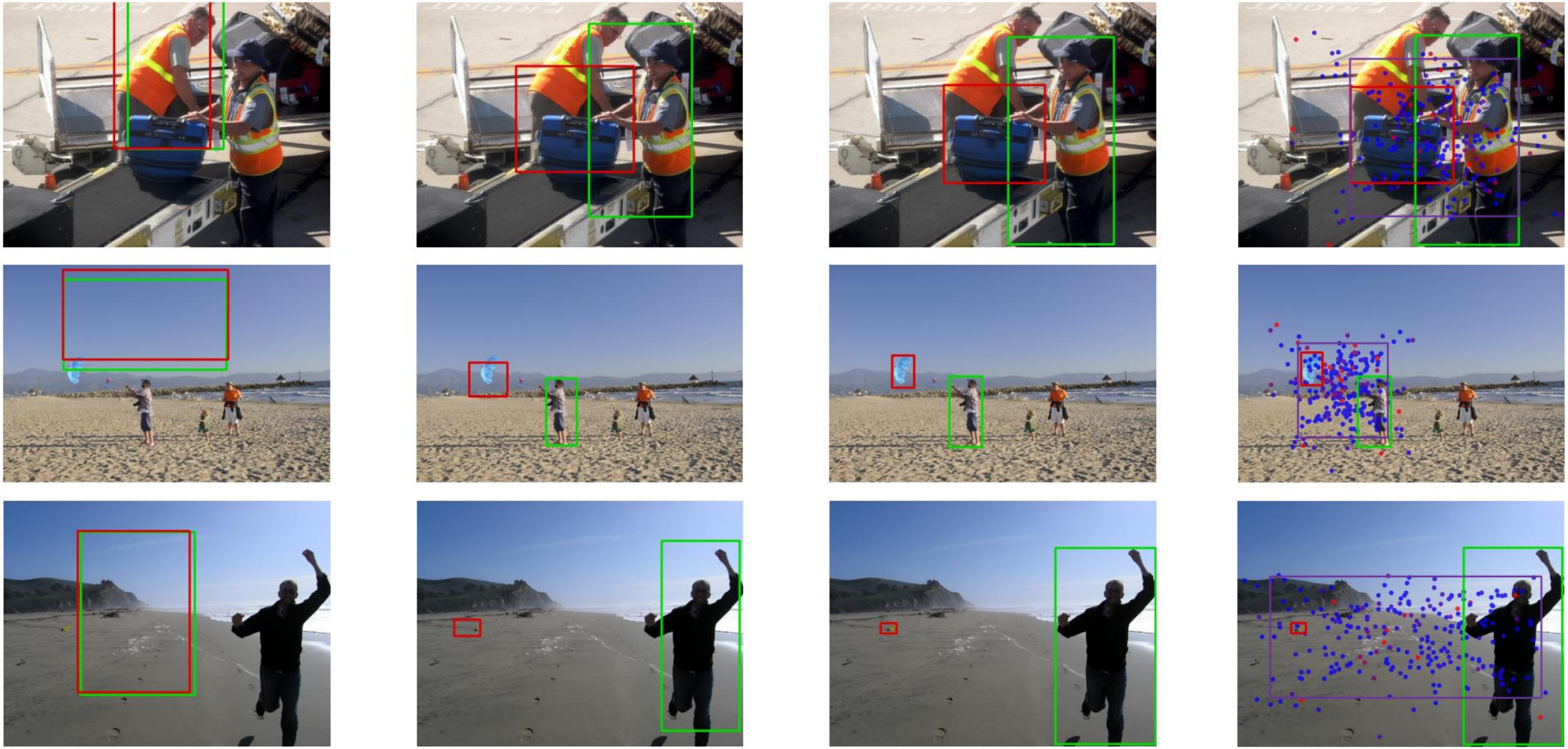
Method	Default		
	Full	*Rare*	*Non-Rare*
SOV-STG-S	32.97	29.28	34.07
-STG	32.44	27.30	33.98
-DN	30.61	24.91	33.32
-Subject Decoder	29.87	24.92	31.35
-Verb Decoder	28.74	22.63	30.57

表 5　各モジュールの貢献

#	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

定性的な結果



(a) box priors (b) layer 1 (c) layer 2 (e) last layer

まとめ

- 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

- SOV-STG は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

ロ 今後の課題

- 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


UEC
TOKYO