マルチスケール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

OS4B-S1

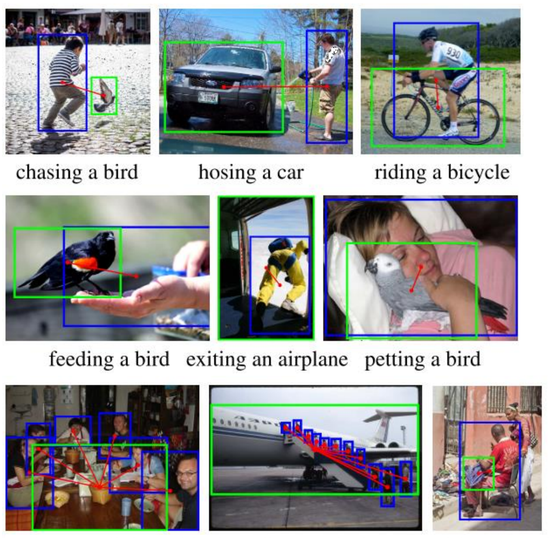
陳　俊文、王　瀛成、柳井　啓司

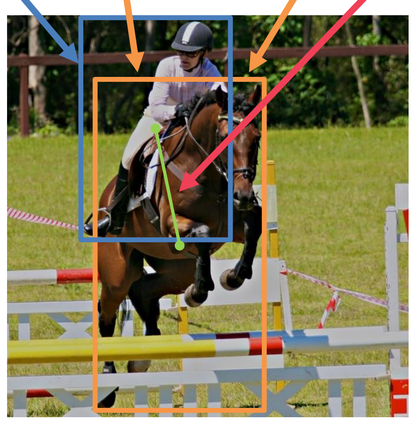
電気通信大学

HOI 検出

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

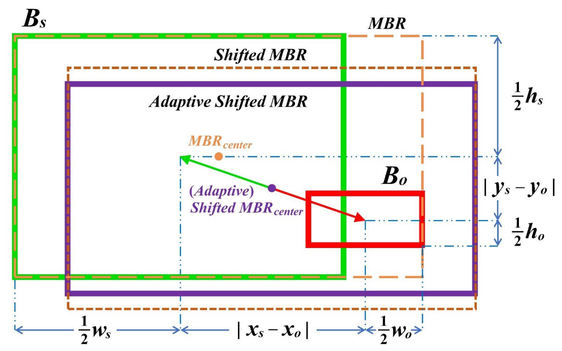
$$c_{\text{HOI}} : \left[c_{\text{obj}}, c_{\text{action}}\right]$$





- 近年、人間と物体のインタラクション検出、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
- 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 [1] は HOI 検出に最もよく使われるデータセット
  - HOI クラスは、**117** 個の動詞クラスと **80** 個の物体クラスから構成される **600** 種類がある

動作アンカーボックス
Adaptive Shifted MBR

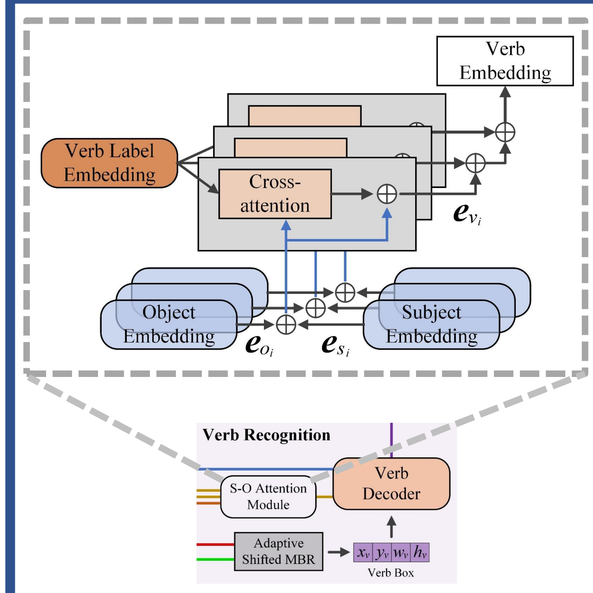


$$B_v = \left(\frac{x_s + x_o}{2}, \frac{y_s + y_o}{2}, w_v, h_v\right)$$

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

- MBR: Minimal Bounding Rectangle
- 空間的な関係を考慮しながら動作ボックスを生成
- **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

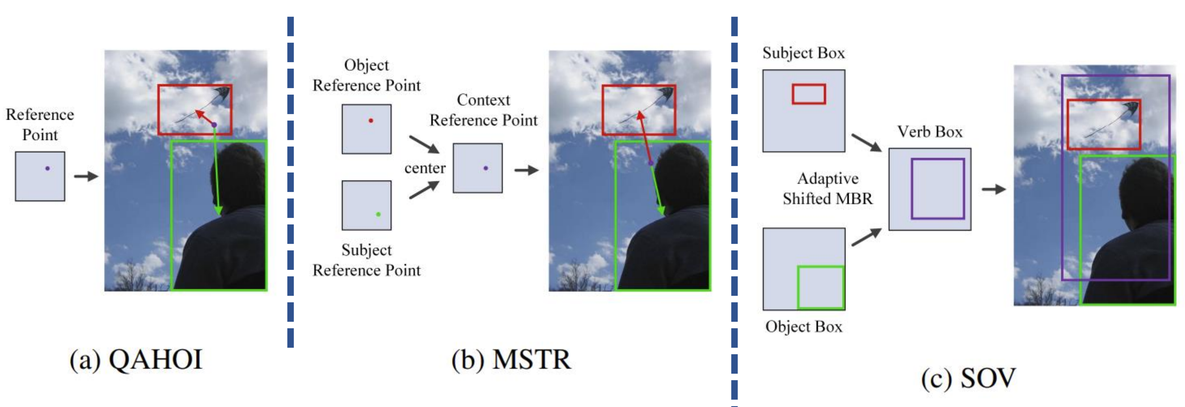
動作認識



$$e_{v_i} = ((\text{CrossAttn}(e_{so_{i-1}}, t_v) + e_{so_{i-1}}) + (\text{CrossAttn}(e_{so_i}, t_v) + e_{so_i}))/2$$

- S-O アテンションモジュールと動作デコーダ
- 動作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

モチベーション



- すべての要素が **同じアンカー** によって予測される
- **同じクエリ表現** は複数のタスクに共有される
- Deformable Transformer デコーダ [2] のサンプリングをアンカーとして使用
- **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**
- **動作ボックスを導入**

- 人物、物体、動作の予測を分離した Subject Object Verb（SOV）フレームワークを提案
- アンカーボックスで HOI を表現するパイプラインを提案
- 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

実験結果と定性評価

Method	Epoch	Backbone	Default			Known Object		
			Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
Two-stage								
CATN [5]	12	ResNet-50	31.86	25.15	33.84	34.44	27.69	36.45
UPT [20]	20	ResNet-101-DC5	32.62	28.62	33.81	36.08	31.41	37.47
Liu *et al.* [14]	129	ResNet-50	33.51	30.30	34.46	36.28	33.16	37.21
One-stage								
QAHOI [3]	150	ResNet-50	26.18	18.06	28.61	-	-	-
AS-Net [4]	90	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [17]	150	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
MSTR [10]	50	ResNet-50	31.17	25.31	32.92	34.02	28.83	35.57
Zhou *et al.* [21]	80	ResNet-50	31.75	27.45	33.03	34.50	30.13	35.81
CDN-B [19]	100	ResNet-50	31.78	27.55	33.05	34.53	29.73	35.96
GEN-VLKT-S [11]	90	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
GEN-VLKT-M [11]	90	ResNet-101	34.78	31.50	35.77	38.07	34.94	39.01
GEN-VLKT-L [11]	90	ResNet-101	34.95	**31.18**	36.08	**38.22**	**34.36**	**39.37**
QAHOI-Swin-L [3]	150	Swin-Large-22K	35.78	29.80	37.56	37.59	31.36	39.36
FGAHOI-Swin-L [16]	150	Swin-Large-22K	37.18	30.71	39.11	38.93	31.93	41.02
SOV-STG-S	30	ResNet-50	33.80	29.28	35.15	36.22	30.99	37.78
SOV-STG-M	30	ResNet-101	34.87	30.41	36.20	37.35	32.46	38.81
SOV-STG-L	30	ResNet-101	**35.01**	30.63	**36.32**	37.60	32.77	39.05
SOV-STG-Swin-L	30	Swin-Large-22K	**43.35**	**42.25**	**43.69**	**45.53**	**43.62**	**46.11**

表1　HICO-DET での結果

Method	Backbone	AP^{S1}_{role}	AP^{S2}_{role}
QPIC [17]	ResNet-50	58.8	61.0
UPT [20]	ResNet-101	61.3	67.1
MSTR [10]	ResNet-50	62.0	65.2
Liu *et al.* [14]	ResNet-50	63.0	65.2
CDN-L [19]	ResNet-101	63.9	65.9
GEN-VLKT-M [11]	ResNet-101	63.3	65.6
GEN-VLKT-L [11]	ResNet-101	63.6	65.9
SOV-STG-M	ResNet-101	63.7	65.2
SOV-STG-L	ResNet-101	63.9	65.4

表2　V-COCO での結果

#	Verb Box	Default		
		Full	*Rare*	*Non-Rare*
(1)	Object Box	33.16	27.21	34.94
(2)	Subject Box	32.78	28.01	34.21
(3)	MBR	33.44	27.84	35.11
(4)	SMBR	33.41	28.22	34.97
(5)	ASMBR	**33.80**	**29.28**	**35.15**

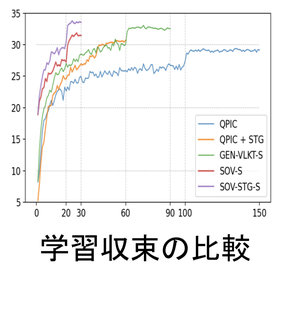
表3　動作ボックスデザイン

Denoising Strategies				Default		
#	Box	Obj	Verb	*Full*	*Rare*	*Non-Rare*
(1)				32.99	28.28	34.40
(2)	✓			33.27	29.07	34.53
(3)	✓		✓	33.28	28.57	34.69
(4)		✓	✓	33.39	28.82	34.76
(5)	✓	✓		33.51	29.05	34.84
(6)	✓	✓	✓	**33.80**	**29.28**	**35.15**

表4　ノイズ除去実験

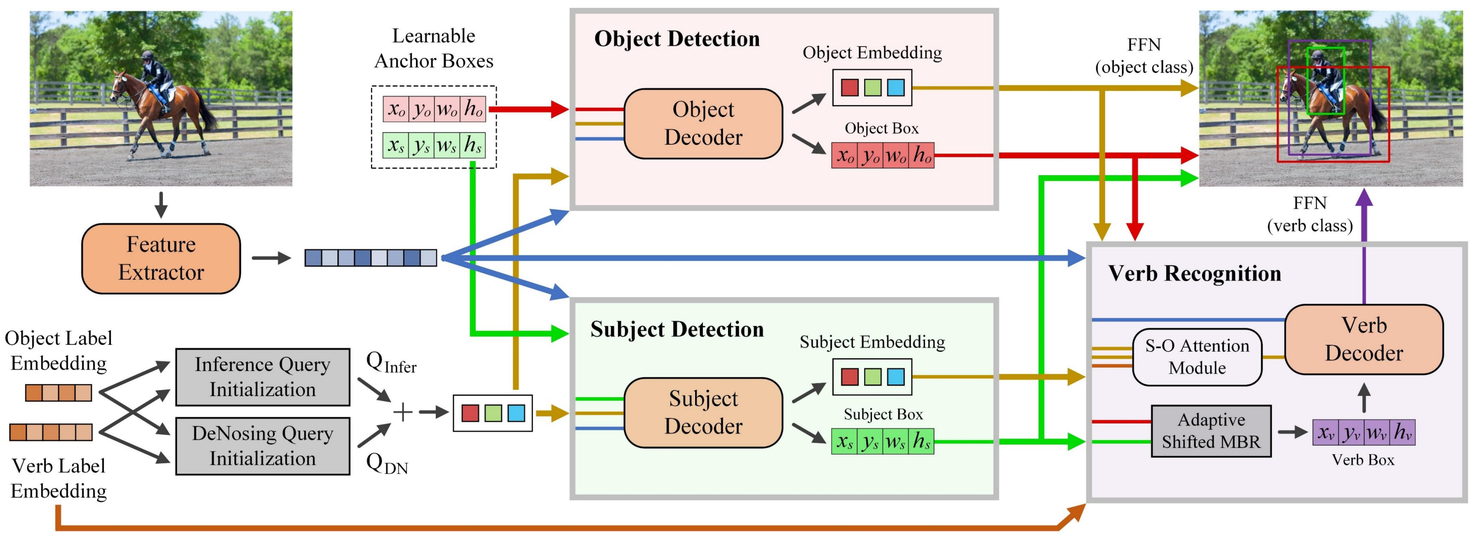
#	oDec	sDec	vDec	STG	Default		
					Full	*Rare*	*Non-Rare*
(1)	✓		✓	✓	32.68	28.21	34.02
(2)	✓	✓		✓	32.35	27.64	33.63
(3)	✓				30.14	22.82	32.32
(4)	✓	✓			30.62	24.60	32.42
(5)	✓	✓	✓		31.90	25.92	33.69
(6)	✓	✓	✓	✓	**33.80**	**29.28**	**35.15**

表5　各モジュールの貢献



学習収束の比較

SOV-STG: Focusing on what to decode and what to train



- DAB-Deformable-DETR [2] 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

参考文献

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

[2] 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.

[3] 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.