Reformulating HOI Detection as Adaptive Set Prediction

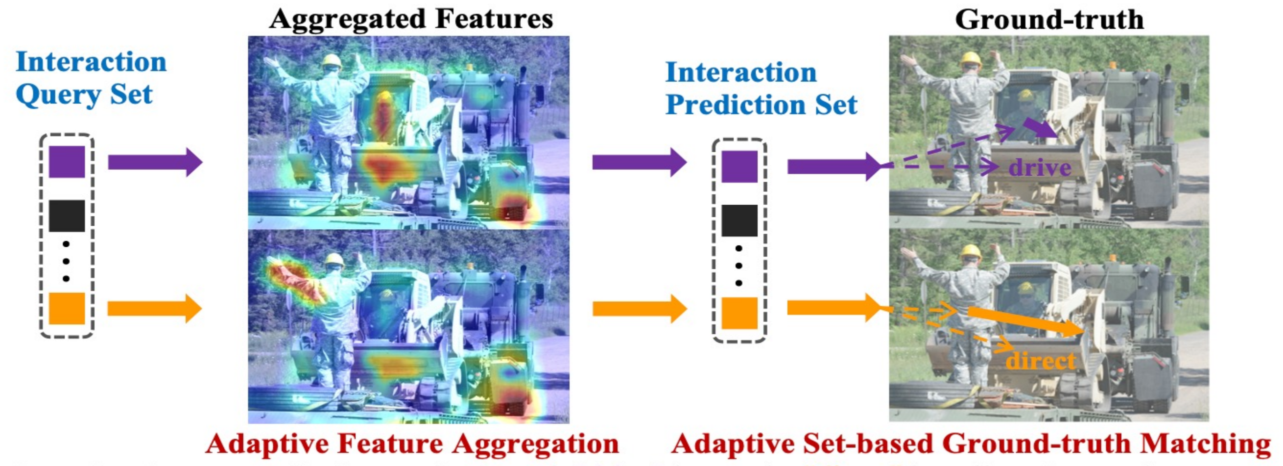
Mingfei Chen[1,3*] Yue Liao[2*] Si Liu[2†] Zhiyuan Chen[3] Fei Wang[3] Chen Qian[3]
[1] Huazhong University of Science and Technology
[2] Institute of Artificial Intelligence, Beihang University [3] SenseTime Research

CVPR VIRTUAL JUNE 19-25

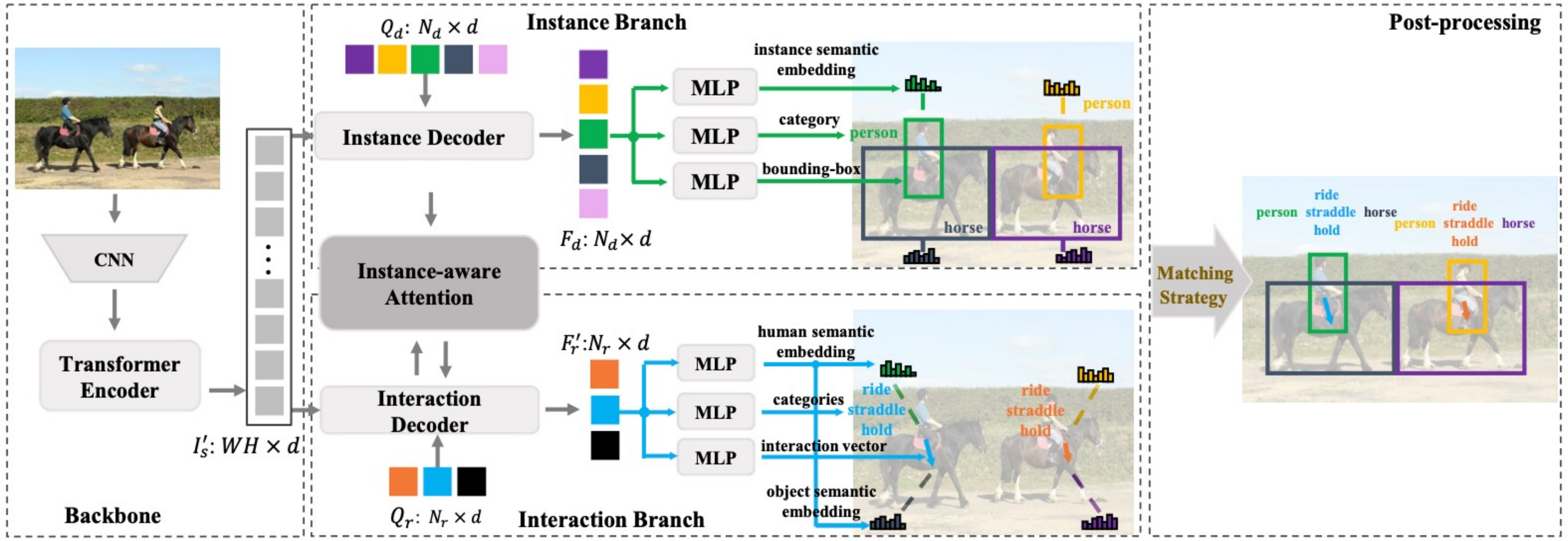
◆ Challenges

Interaction category prediction is limited by
- detection performance (previous two-stage methods)
- predefined interaction locations (union boxes or interaction midpoints of previous one-stage methods)

(a) **Union boxes:** verb "direct" in yellow, "drive" in purple, matched anchor in red. (b) **Interaction midpoints:** verb "direct" in yellow "drive" in purple, matched point in red.



(c) **Our adaptive set prediction method:** verb "drive" in purple, "direct" in yellow. Interaction vectors point from human centers to object centers. The features aggregated by queries are visualized at left.



◆ Contributions

- Reformulate HOI detection as set prediction problem, adaptively concentrate on the most suitable features to improve the predicting accuracy
- Propose a novel one-stage transformer-based HOI detection framework (AS-Net)
- Design instance-aware attention module to introduce instance information into the interaction branch
- Maintaining the high efficiency and without any extra features, our method gains 31% relative improvements on HICO-DET, especially 73% on rare HOI categories

◆ Solutions

- Applying co-attention to adaptively aggregate interaction-relevant features using learnable queries
- Adaptively matching the most suitable ground-truth considering both action categories and location distances

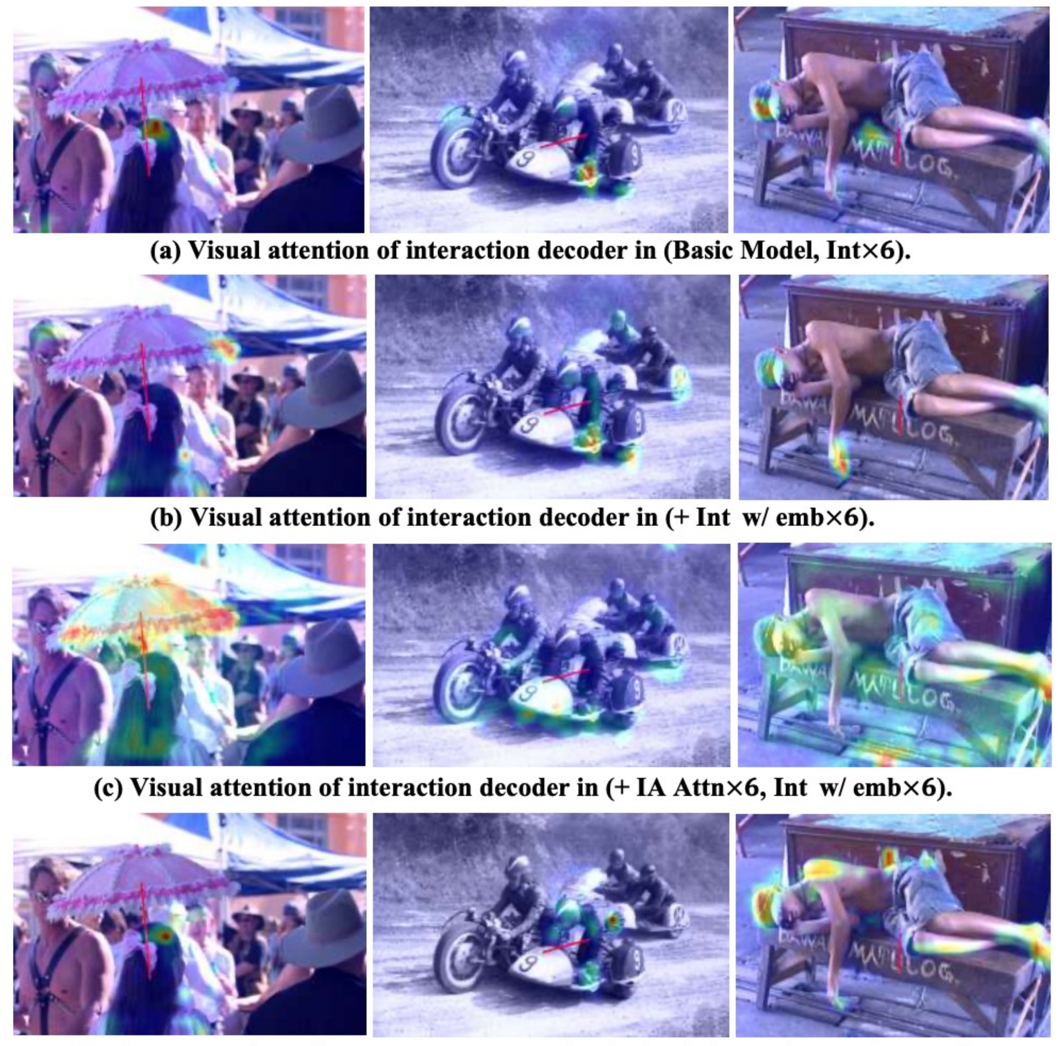
Instance-aware Attention

- Involve instructive instance features to interaction branch in co-attention manner

Semantic embedding

- Point to the specific instance more accurately for better matching
- Bridging instance branch and interaction branch implicitly

◆ Attention Visualization



(a) Visual attention of interaction decoder in (Basic Model, Int×6).

(b) Visual attention of interaction decoder in (+ Int w/ emb×6).

(c) Visual attention of interaction decoder in (+ IA Attn×6, Int w/ emb×6).

(d) Visual attention of instance-aware attention in (+ IA Attn×6, Int w/ emb×6).

◆ Experiments

Strategy	Full	Rare	Non-Rare
Vector	28.56	24.13	29.88
Embedding	28.65	23.95	30.05
Combined	**28.87**	**24.25**	**30.25**

(a) **Matching Strategy:** Analysis of different matching strategies, *i.e.*, interaction vector and semantic embeddings.

K	Full	Rare	Non-Rare	#Parameters
4	28.21	22.65	29.87	52.527 M
8	**28.87**	**24.25**	**30.25**	52.530 M
16	28.36	23.08	29.93	52.537 M
32	28.70	23.83	30.16	52.549 M

(b) **Dimension of Semantic Embeddings:** Choice of dimension of semantic embeddings.

λ_{emb}	Full	Rare	Non-Rare
0.05	28.31	23.65	29.70
0.1	**28.87**	**24.25**	**30.25**
0.5	27.84	21.71	29.67

(c) **Weight Coefficient λ_{emb}:** The effects of different settings of loss weight.

	Decoder Layers	Embeddings	IA Attention	Full	Rare	Non-Rare	#Parameters
Single Branch	6×	✗	-	25.91	17.88	28.31	41.44 M
Basic Model, Int×6	6×	✗	-	27.52	22.04	29.16	50.94 M
+ IA Attn×6, Int w/o emb×6	6×	✗	6×	27.96	23.01	29.44	52.13 M
+ Int w/ emb×6	6×	✓	-	27.75	22.71	29.25	51.34 M
+ IA Attn×3, Int w/ emb×6	6×	✓	3×	28.39	24.02	29.70	51.94 M
+ IA Attn×3, Int w/ emb×3	3×	✓	3×	28.63	23.61	30.13	47.20 M
+ IA Attn×6, Int w/ emb×6	6×	✓	6×	**28.87**	**24.25**	**30.25**	52.53 M

(d) **Component Analysis:** Results of the variants with various components, *i.e.*, interaction branches (Int), instance-aware attention module (IA Attn) and semantic embeddings (emb).

Table 4. Ablation studies of our proposed model on the HICO-DET test set.

Method	Backbone	Finetune Detection	Extra	Time (ms) / FPS	Default Full	Default Rare	Default Non-Rare	Know Object Full	Know Object Rare	Know Object Non-Rare
Two-stage Method:										
InteractNet [12]	ResNet-50-FPN	✗	✗	145 / 6.90	9.94	7.16	10.77	-	-	-
GPNN [34]	Res-DCN-152	✗	✗	-	13.11	9.34	14.23	-	-	-
iCAN [10]	ResNet-50	✗	✗	204 / 4.90	14.84	10.45	16.15	16.26	11.33	17.73
No-Frills [14]	ResNet-152	✗	*P*	494 / 2.02	17.18	12.17	18.68	-	-	-
PMFNet [40]	ResNet-50-FPN	✗	*P*	253 / 3.95	17.46	15.65	18.00	20.34	17.47	21.20
DRG [9]	ResNet-50-FPN	✗	*L*	200 / 5.00	19.26	17.74	19.71	23.40	21.75	23.89
IP-Net [42]	Hourglass-104	✗	✗	-	19.56	12.79	21.58	22.05	15.77	23.92
VSGNet [39]	ResNet-152	✗	✗	312 / 3.21	19.80	16.05	20.91	-	-	-
PD-Net [47]	ResNet-152-FPN	✗	*L*	-	20.81	15.90	22.28	24.78	18.88	26.54
DJ-RN [23]	ResNet-50	✗	*P*	-	21.34	18.53	22.18	23.69	20.64	24.60
One-stage Method:										
UnionDet [20]	ResNet-50-FPN	✓	✗	78 / 12.82	17.58	11.72	19.33	19.76	14.68	21.27
PPDM-Hourglass [27]	Hourglass-104	✓	✗	71 / 14.08	21.94	13.97	24.32	24.81	17.09	27.12
AS-Net*	ResNet-50	✗	✗	71 / 14.08	24.40	22.39	25.01	27.41	25.44	28.00
AS-Net	ResNet-50	✓	✗	**71 / 14.08**	**28.87**	**24.25**	**30.25**	**31.74**	**27.07**	**33.14**

Table 1. **Performance comparison on the HICO-DET test set.** The 'P', 'L' represent human pose information and the language feature, respectively. * denotes freezing the instance detection related parameters pretrained on the MS-COCO dataset. Our one-stage model with a high inference speed of 71 ms / 14.08 FPS outperforms all previous work by a large margin.

Method	Backbone	Extra	mAP_{role}
Two-stage Method:			
InteractNet [8]	ResNet-50-FPN	✗	40.0
GPNN *et al.* [26]	Res-DCN-152	✗	44.0
iCAN [6]	ResNet-50	✗	45.3
DRG [5]	ResNet-50-FPN	*L*	51.0
IP-Net [33]	Hourglass-104	✗	51.0
VSGNet [30]	ResNet-152	✗	51.8
PMFNet [31]	ResNet-50-FPN	*P*	52.0
PD-Net [36]	ResNet-152-FPN	*L*	52.6
FCMNet [22]	ResNet-50	✗	53.1
One-stage Method:			
UnionDet [13]	ResNet-50-FPN	✗	47.5
AS-Net*	ResNet-50	✗	**53.9**

Table 2. **Performance comparison on the V-COCO test set**. The 'P', 'L' represent the human pose information and the language feature, respectively. * denotes freezing the instance detection related parameters pretrained on the MS-COCO dataset.