ACM multimedia 2020 Seattle

ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection



Ye Liu [1] Junsong Yuan [2] Chang Wen Chen [2,3,4]

[1] Wuhan University
[2] State University of New York at Buffalo
[3] Peng Cheng Laboratory
[4] The Chinese University of Hong Kong, Shenzhen

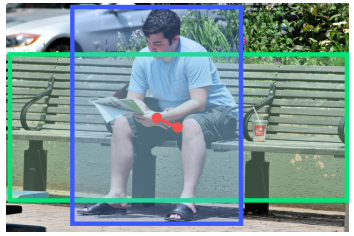
武漢大學 WUHAN UNIVERSITY

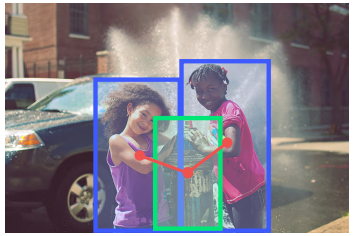
University at Buffalo The State University of New York

鹏城实验室 PENG CHENG LABORATORY

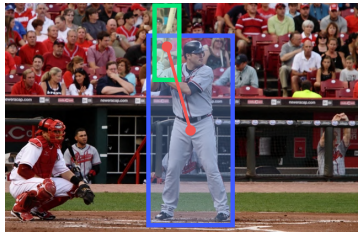
香港中文大學(深圳) The Chinese University of Hong Kong, Shenzhen

Human Object Interaction Detection

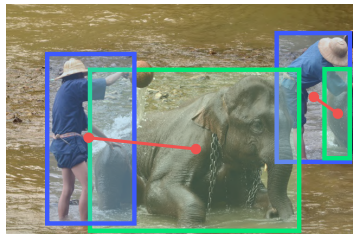

sit on bench


hug fire hydrant

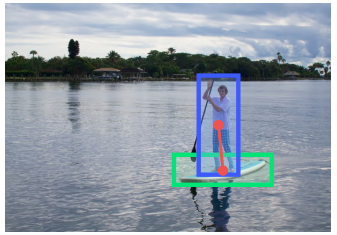

carry umbrella

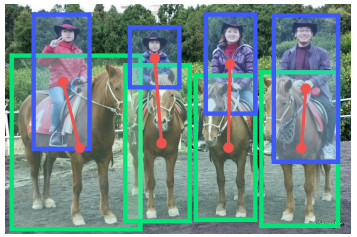

hold baseball bat

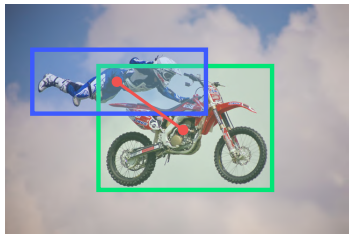

drive bus


wash elephant

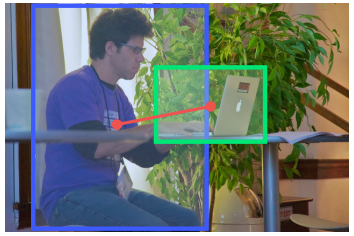

hose truck

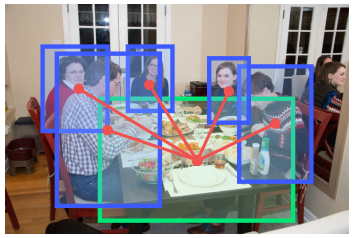

ride surfboard


straddle horse

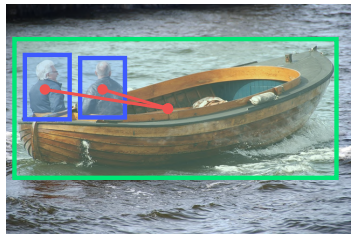

jump motorcycle


train dog


read laptop


eat at dinning table


catch frisbee

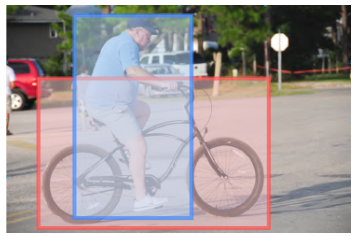

ride boat

Detecting <human, action, object> triplets in static images

A common pipeline of HOI Detection

input image

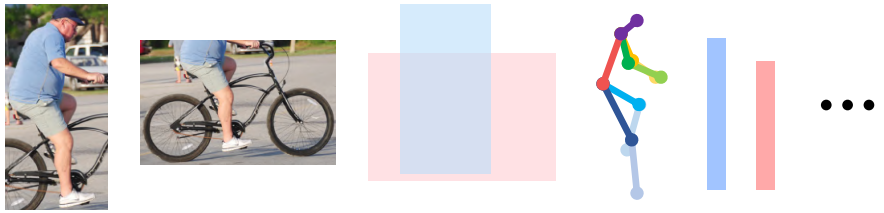
object detection



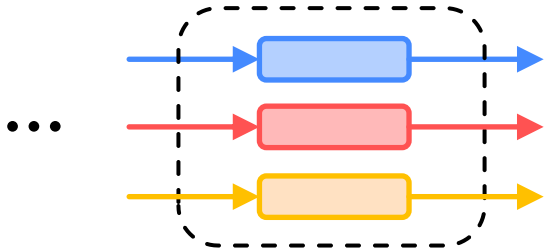
human-object pair proposals

feature extraction

(and alternative pose estimation)

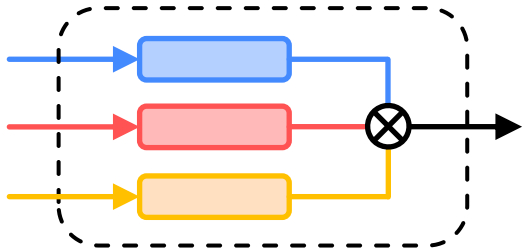


human and object features, layouts, poses, confidences, etc.

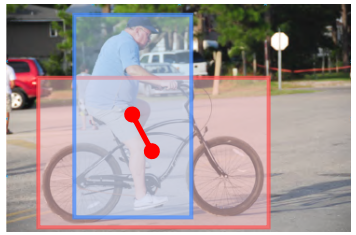


compositional methods

or



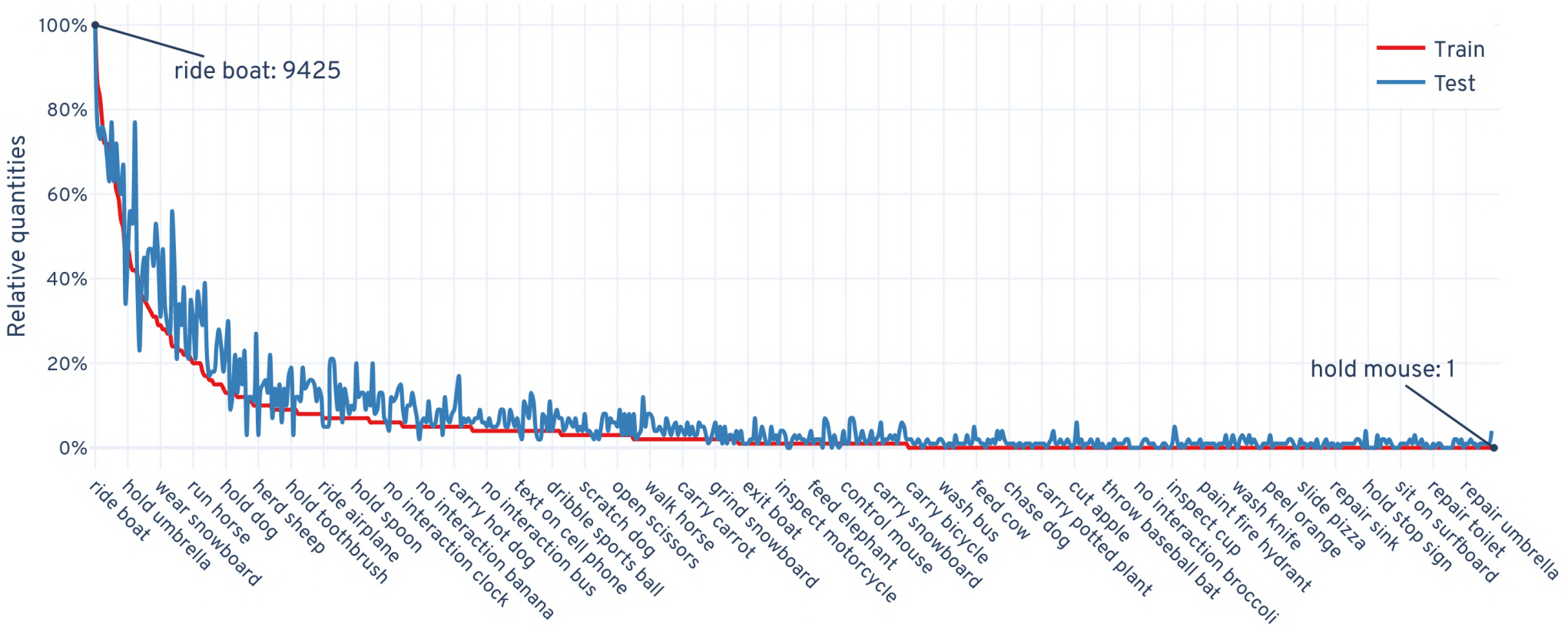
non-compositional methods



detection results

<human, ride, bicycle>	0.86 ✓
<human, jump, bicycle>	0.09 ✗
<human, sit on, bicycle>	0.90 ✓
<human, ride, horse>	0.16 ✗

Challenges



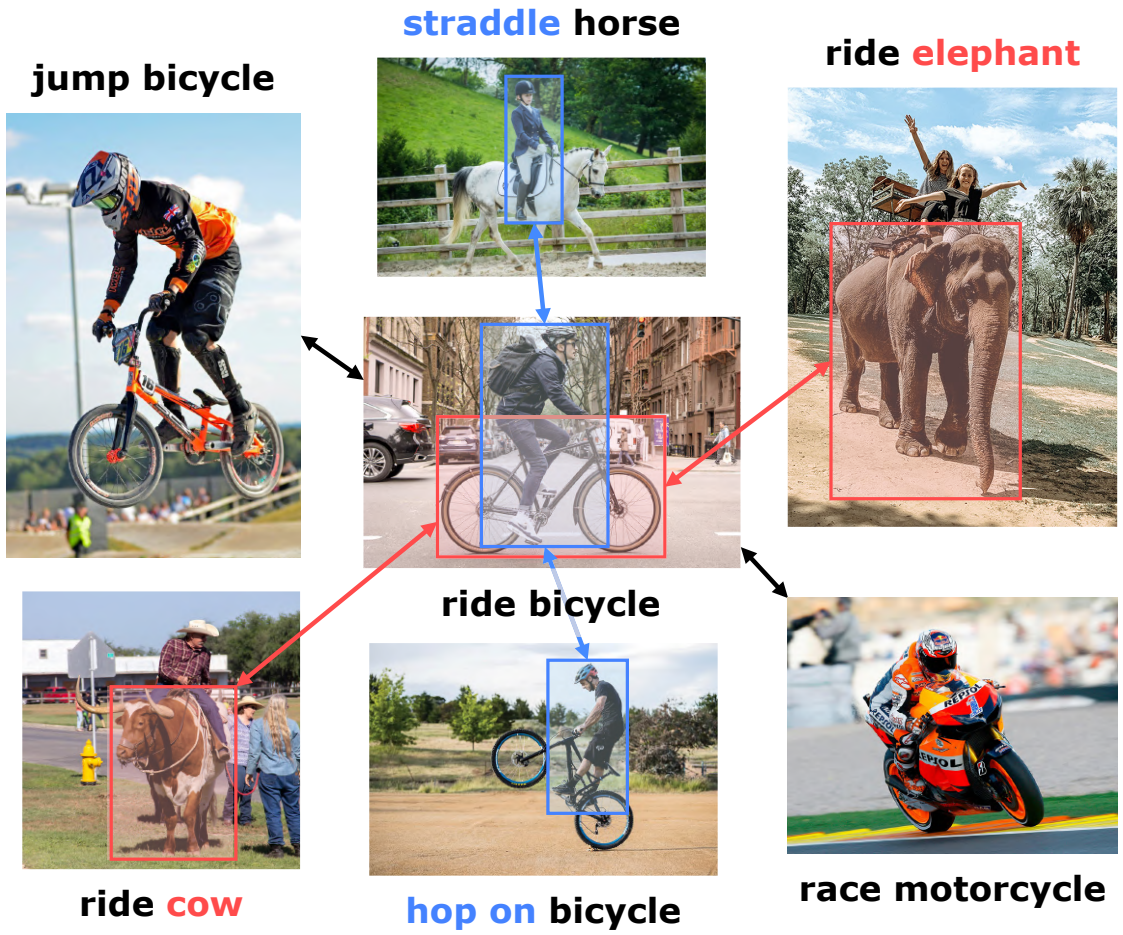
Class-wise long-tail distribution

Challenges

ride horse | ride snowboard | ride skis | ride motorcycle | ride boat | ride bus

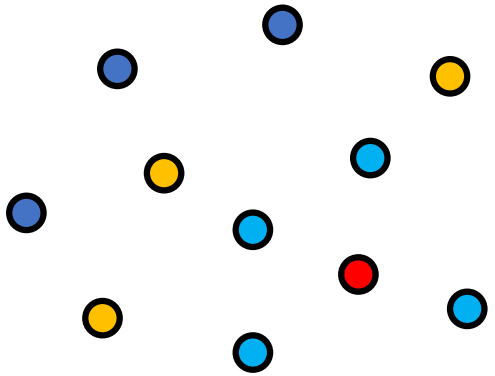
Polysemy of action labels

Multi-level Consistencies

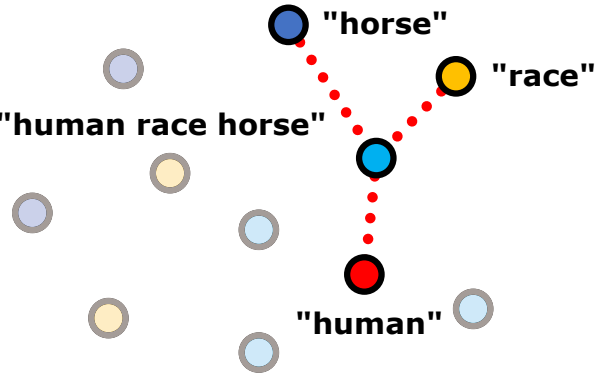


Functional Consistency Behavioral Consistency Interactional Consistency

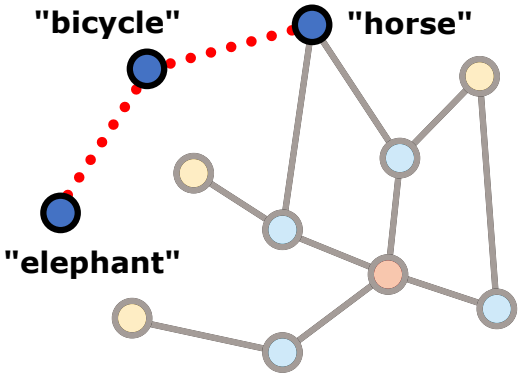
Consistency Graph



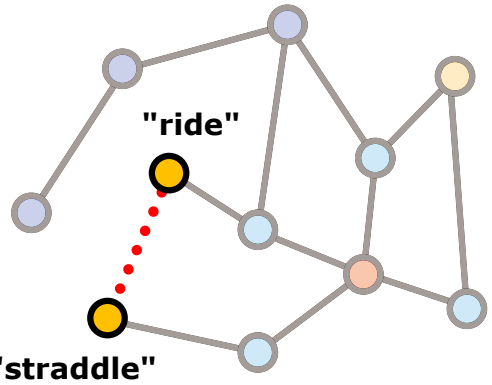
(a) Nodes



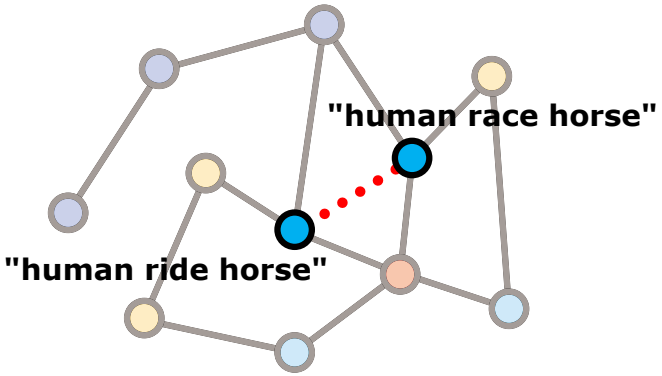
(b) Granularity Bridges



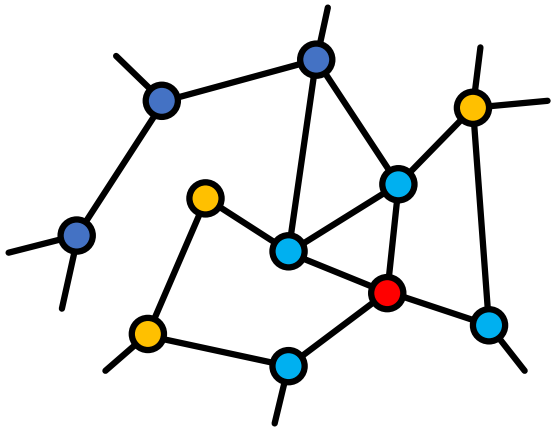
(c) Functional Consistency



(d) Behavioral Consistency

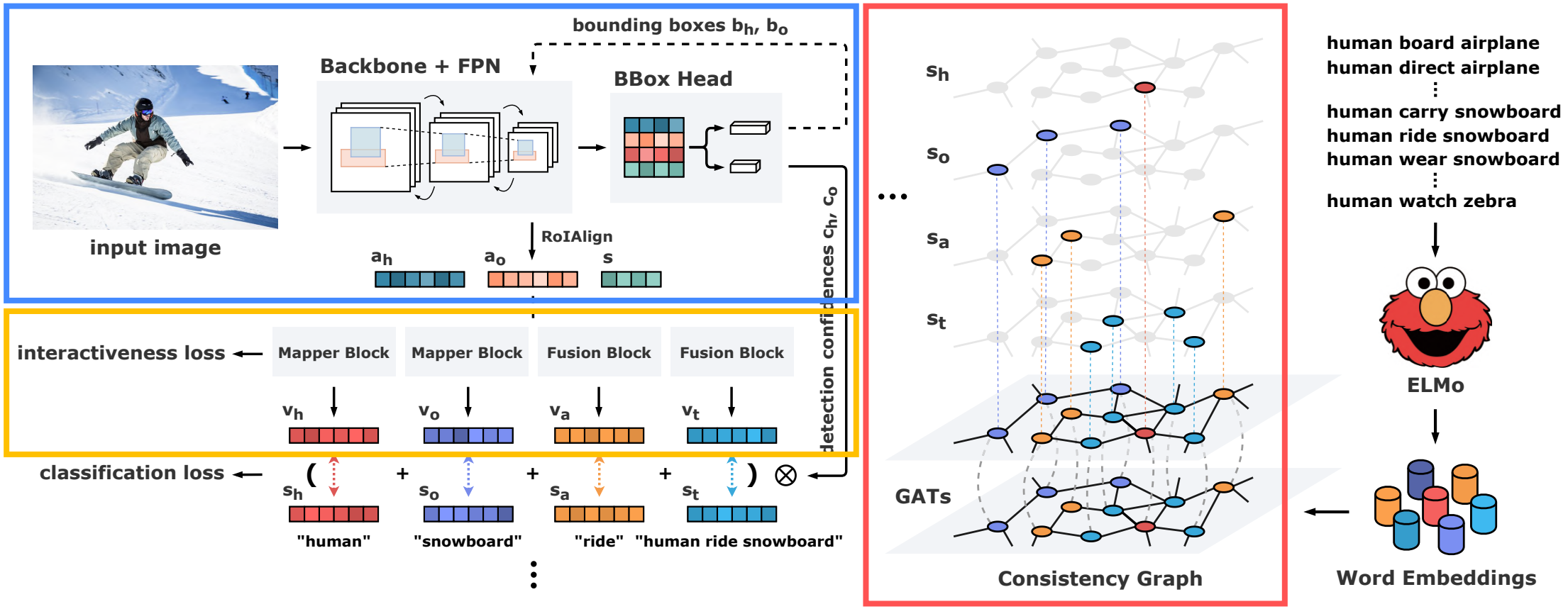


(e) Interactional Consistency



(f) Consistency Graph

Model Architecture



Object Detection Module Visual Embedding Network Semantic Embedding Network

Quantitative Results

Method	Backbone	mAP_{role}
Gupta *et al.* [13]	ResNet-50-FPN	31.8
InteractNet [11]	ResNet-50-FPN	40.0
GPNN [34]	DCN	44.0
iCAN [9]	ResNet-50	45.3
TIN-$RP_{T2}C_D$ [25]	ResNet-50	48.7
BAR-CNN [21]	Inception-ResNet	43.6
Wang *et al.* [41]	ResNet-50	47.3
PMFNet [39]	ResNet-50	52.0
IP-Net [42]	Hourglass-104	51.0
VSGNet [37]	ResNet-152	51.8
ConsNet (ours)	ResNet-50-FPN	**53.2**

Table 1: Role detection results on V-COCO dataset

Method	Backbone	Full	Rare	Non-Rare
Shen *et al.* [36]	VGG-19	6.46	4.24	7.12
HO-RCNN [4]	CaffeNet	7.81	5.37	8.54
InteractNet [11]	R-50-FPN	9.94	7.16	10.77
GPNN [34]	DCN	13.11	9.34	14.23
iCAN [9]	R-50	14.84	10.45	16.15
TIN-$RP_{T2}C_D$ [25]	R-50	17.22	13.51	18.32
HOID [40]	R-50-FPN	17.85	12.85	19.34
Wang *et al.* [41]	R-50-FPN	16.24	11.16	17.75
Gupta *et al.* [14]	R-152	17.18	12.17	18.68
PMFNet [39]	R-50-FPN	17.46	15.65	18.00
Peyre *et al.* [33]	R-50-FPN	19.40	15.40	20.75
IP-Net [42]	H-104	19.56	12.79	21.58
VSGNet [37]	R-152	19.80	16.05	20.91
ConsNet (ours)	R-50-FPN	**22.15**	**17.12**	**23.65**
Bansal *et al.* [1]	R-101	21.96	16.43	23.62
PPDM [26]	H-104	21.73	13.78	24.10
ConsNet-F (ours)	R-50-FPN	**24.39**	**17.10**	**26.56**

Table 2: HOI detection results on HICO-DET dataset

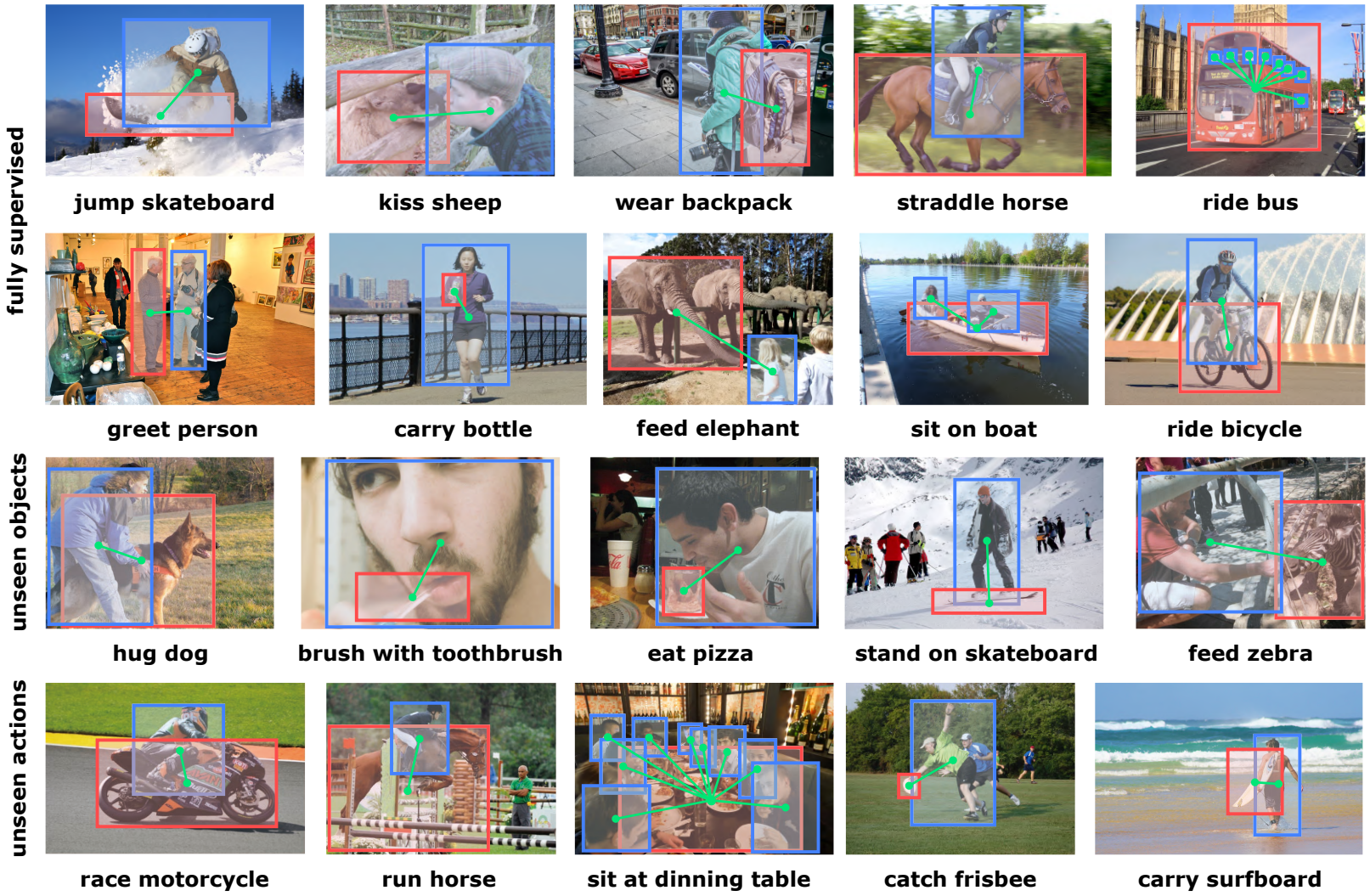
Method	Type	Full	Unseen	Seen
Shen *et al.* [36]	UC	6.26	5.62	-
Bansal *et al.* [1]		12.45±0.16	11.31±1.03	12.74±0.34
ConsNet (ours)		**14.48**±0.26	**13.46**±1.24	**14.74**±0.57
Bansal *et al.* [1]	UO	13.84	11.22	14.36
ConsNet (ours)		**14.48**	**13.51**	**14.67**
ConsNet (ours)	UA	**14.35**	**12.50**	**14.72**

Table 3: Zero-shot HOI detection results on HICO-DET dataset

Type	Embedder	Depth	Full	Rare	Non-Rare
∅	-	-	18.90	10.57	21.40
MLP	ELMo	3	19.01	11.82	21.15
SGC	ELMo	3	19.63	14.85	21.05
GCN	ELMo	3	20.15	15.12	21.66
SAGE	ELMo	3	20.07	15.05	21.58
GAT	ELMo	2	21.16	16.82	22.46
GAT	**ELMo**	**3**	**22.15**	**17.12**	**23.65**
GAT	ELMo	4	21.12	16.35	22.54
GAT	Word2Vec	3	20.59	15.94	21.98
GAT	GloVe	3	20.63	15.66	22.12
GAT	FastText	3	20.58	15.68	22.04

Table 4: Ablation study on HICO-DET dataset

Qualitative Results



ACM multimedia 2020 Seattle

Thank you!

WUHAN UNIVERSITY 武漢大學 University at Buffalo The State University of New York 鹏城实验室 PENG CHENG LABORATORY 香港中文大學（深圳） The Chinese University of Hong Kong, Shenzhen