Mining Cross-Person Cues for Body-Part Interactiveness Learning in HOI Detection

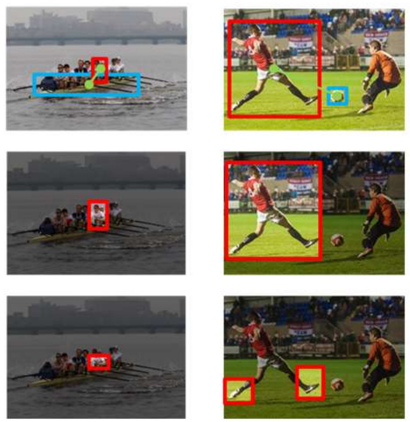
Xiaoqian Wu*, Yong-Lu Li*, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, Cewu Lu

mvig.org

Code

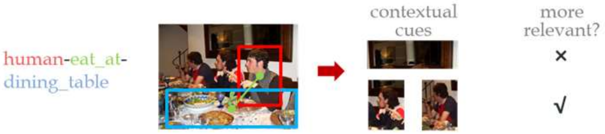
Motivation

■ Local perspective (previous works)
- only focus on the **targeted** person
- overlook the information of the other persons

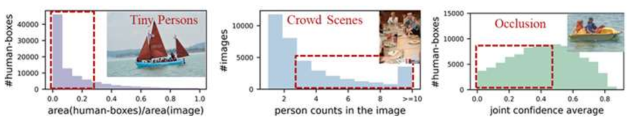


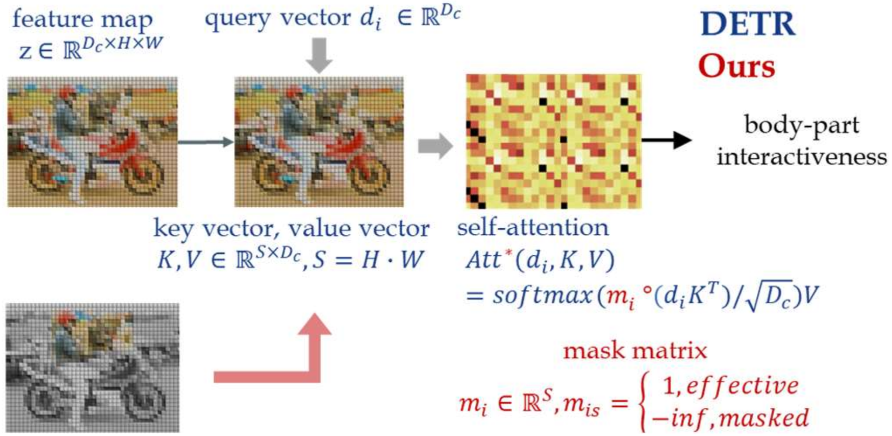
■ Global perspective (ours)
- comparing body-parts of **multi-person** simultaneously
- more **useful** and **supplementary** interactiveness cues

☐ exploit **contextual** cues from the whole image, easier and more stable



☐ alleviate the difficulty of HOI **hard cases**



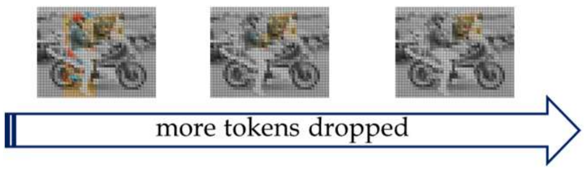


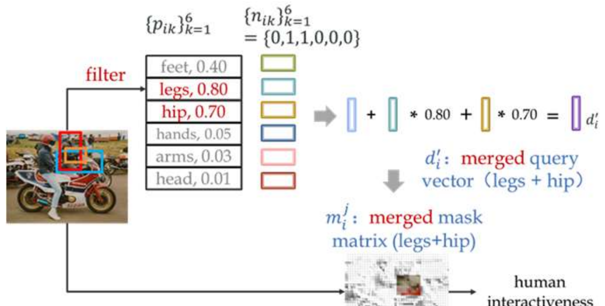
☐ utilizing **self-attention** calculation in transformer

☐ constructing body-part saliency maps via image patches (i.e., transformer tokens) **masking**

☐ progressively body-part masking to encode diverse visual patterns more flexibly

☐ different attention mask is applied in successive transformer layers and more tokens are dropped in the late layers





☐ one-time passing via body-parts filtering and merging to improve computation efficiency

Discussion: Sparse vs. Crowded Scene

■ We focus on crowded scenes, then what about **sparse** scenes?
- Our model is **adapted to both** crowded and sparse scenes.
- Crowded scenes is more important in interactiveness learning.
- Thus, we further propose a novel sparsity adaptive sampling strategy on train set to put more emphasis on crowded scenes.

Experiment & Results

■ With our holistic global-local interactiveness detector, we achieve state-of-the-art for interactiveness detection and HOI detection on HICO-DET & V-COCO.

Method	Full	Sparse/Crowded	Normal/Tiny	Less/More Occ
TIN++	14.35	16.96/9.64	16.11/8.94	16.49/8.06
PPDM	27.34	34.67/26.69	31.79/26.33	29.83/17.25
QPIC	32.96	36.80/27.04	34.02/26.14	32.08/19.75
CDN	33.55	39.92/28.84	36.10/25.11	34.55/21.69
Ours	**38.74**	**43.62/33.10**	**39.85/32.47**	**38.60/22.75**

Method	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
iCAN [7]	14.84	10.45	16.15	16.26	11.33	17.73
TIN [22]	17.03	13.42	18.11	19.17	15.51	20.26
PMFNet [31]	17.46	15.65	18.00	20.34	17.47	21.20
DJ-RN [17]	21.34	18.53	22.18	23.69	20.64	24.60
PPDM [23]	21.73	13.78	24.10	24.58	16.65	26.84
VCL [11]	23.63	17.21	25.55	25.98	19.12	28.03
IDN [19]	26.29	22.61	27.39	28.24	24.47	29.37
Zou *et al.* [37]	26.61	19.15	28.84	29.13	20.98	31.57
ATL [12]	28.53	21.64	30.59	31.18	24.15	33.29
AS-Net [2]	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [29]	29.07	21.85	31.23	31.68	24.14	33.93
FCL [13]	29.12	23.67	30.75	31.31	25.62	33.02
GGNet [34]	29.17	22.13	30.84	33.50	26.67	34.89
SCG [34]	31.33	24.72	33.31	34.37	27.18	36.52
CDN [33]	31.78	27.55	33.05	34.53	29.73	35.96
Ours	**35.15**	**33.71**	**35.58**	**37.56**	**35.87**	**38.06**

Method	HICO-DET [1]	V-COCO [10]
TIN++ [18]	14.35	29.36
PPDM [23]	27.34	-
QPIC [29]	32.96	38.33
CDN [33]	33.55	40.13
ours	**38.74 (+5.19)**	**43.61 (+3.48)**

Method	AP_{role}(S1)	AP_{role}(S2)
iCAN [7]	45.3	52.4
TIN [22]	47.8	54.2
VSGNet [30]	51.8	57.0
PMFNet [31]	52.0	-
IDN [19]	53.3	60.3
AS-Net [2]	53.9	-
SCG [34]	54.2	-
GGNet [34]	54.7	-
HOTR [14]	55.2	64.4
QPIC [29]	58.8	61.0
CDN [33]	62.3	64.4
Ours	**63.0**	**65.1**

■ Some visualization results, where informative cues are extracted from other persons in the image.