https://arxiv.org/abs/1912.12898

PPDM: Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection

Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, Jiashi Feng

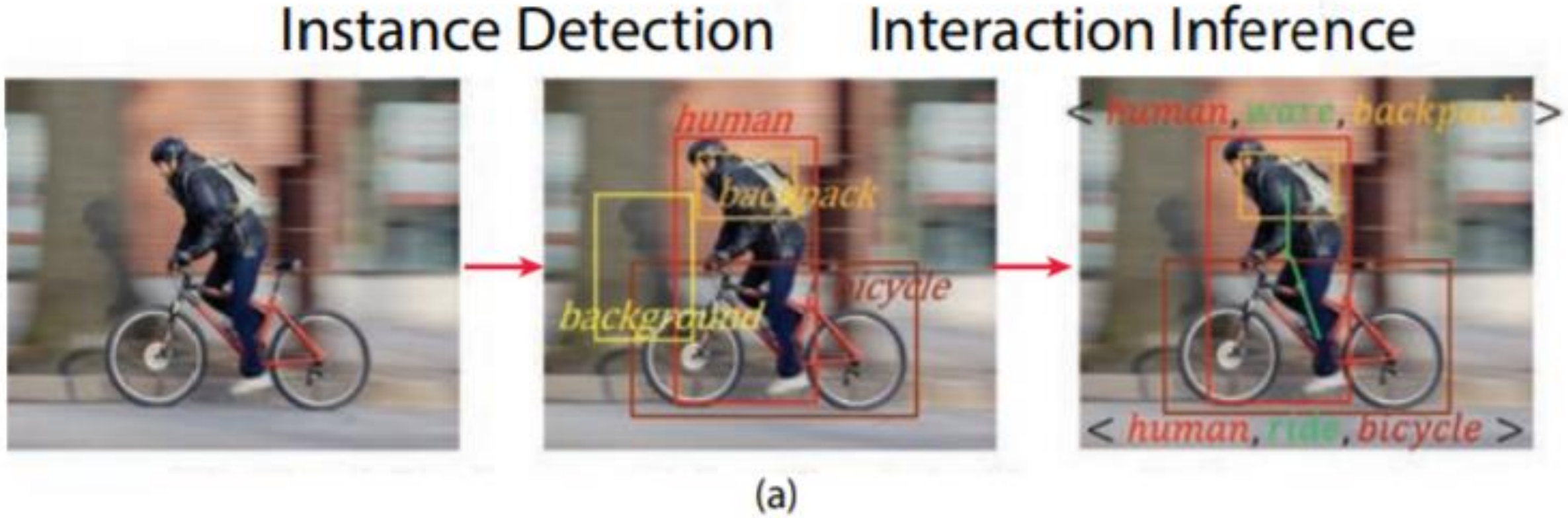
School of Computer Science and Engineering, Beihang University,
SenseTime Research, National University of Singapore

2020 CVPR

인공지능 연구실
석사과정 구자봉

문제 정의 :
HOI(Human Object Interaction)

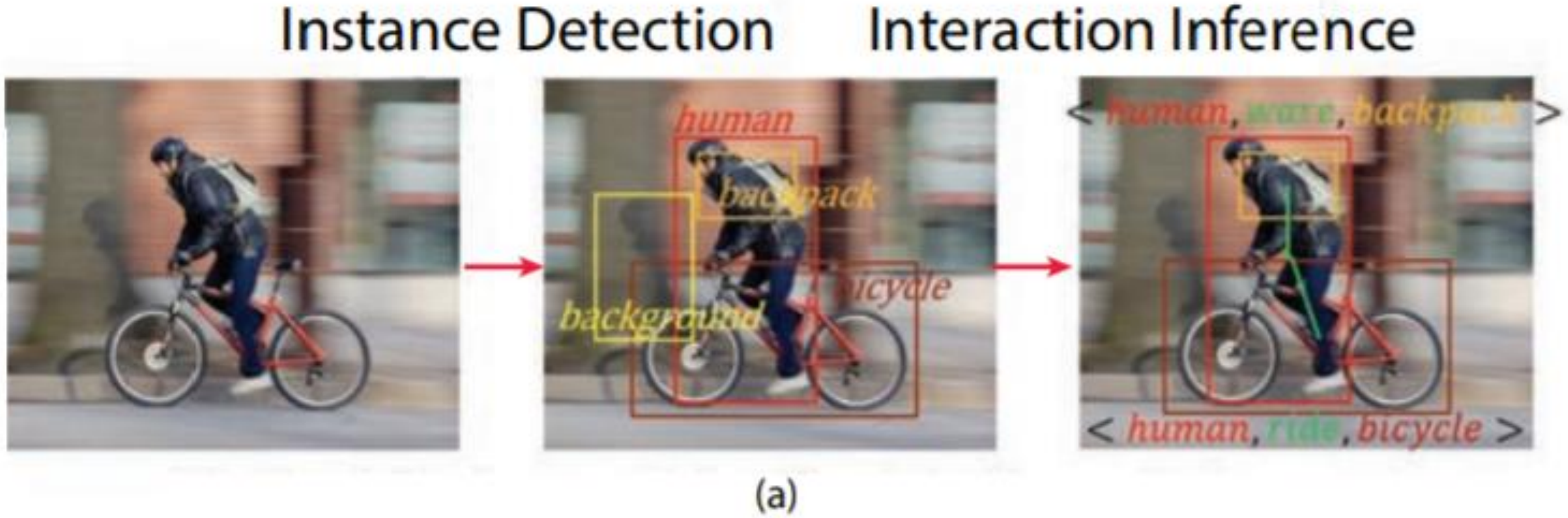
이미지에서 오브젝트 디텍션, 인간과 상호작용이 큰 객체쌍을 선택, 술어(상관관계)를 찾는 것이 목적



(a)

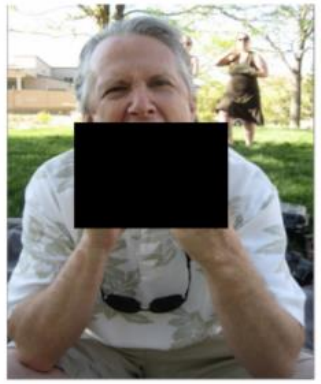
문제 제기 :

대부분의 HOI 모듈은
1)인간-대상 제안서 생성, 2)제안서 분류
2단계로 나뉘어 구성된다. 순차적이고 개별적인 아키텍처를 사용하여 성능이 제한된다.

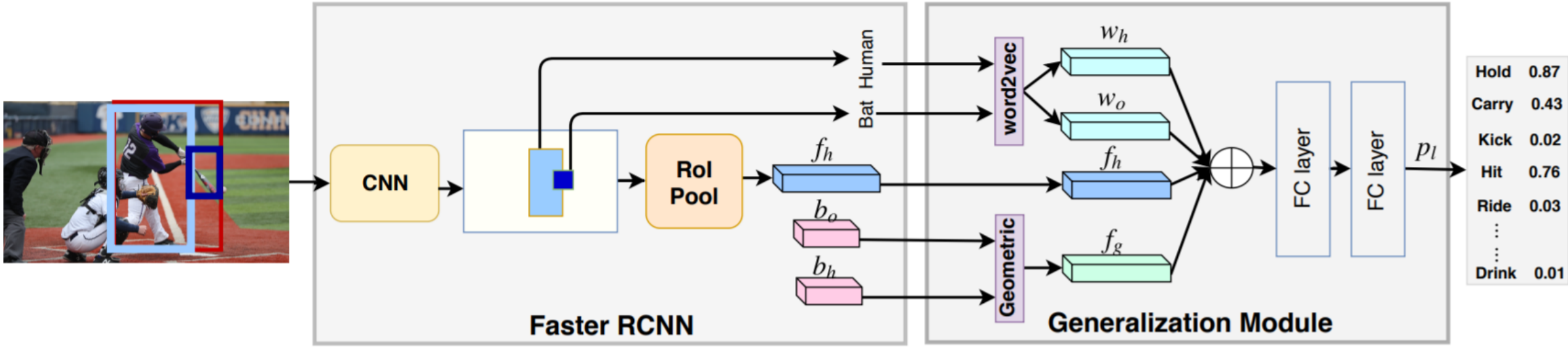


(a)

Detecting Human-Object Interactions via Functional Generalization

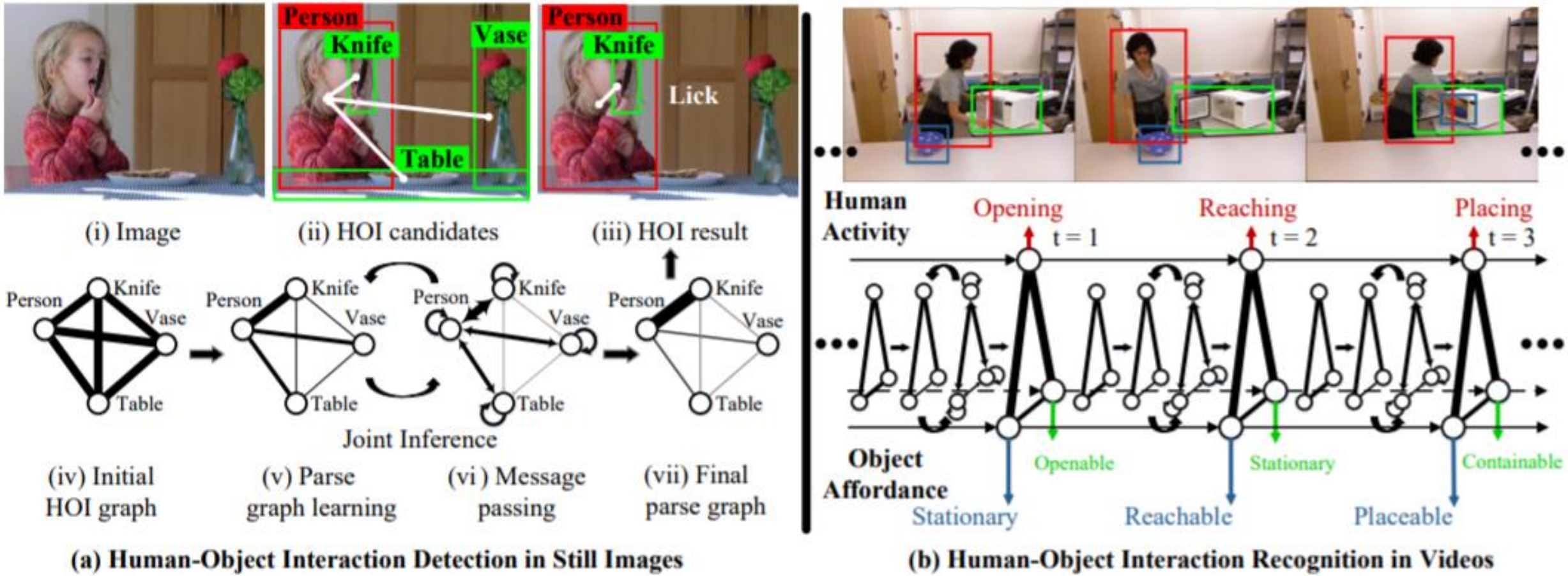


⟨human, eat, __



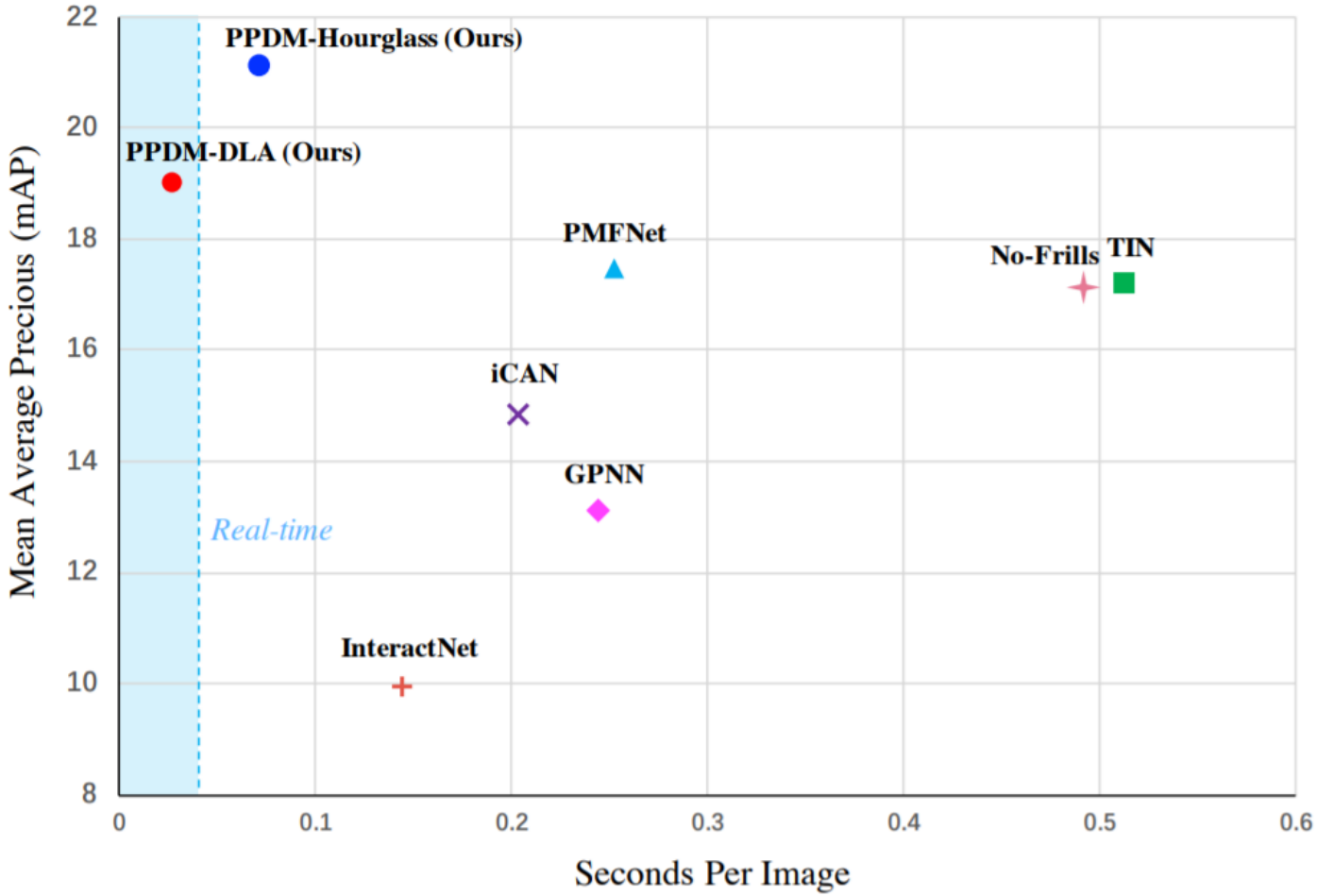
https://arxiv.org/pdf/1710.10903.pdf

Learning Human-Object Interactions by Graph Parsing Neural Networks (GPNN):

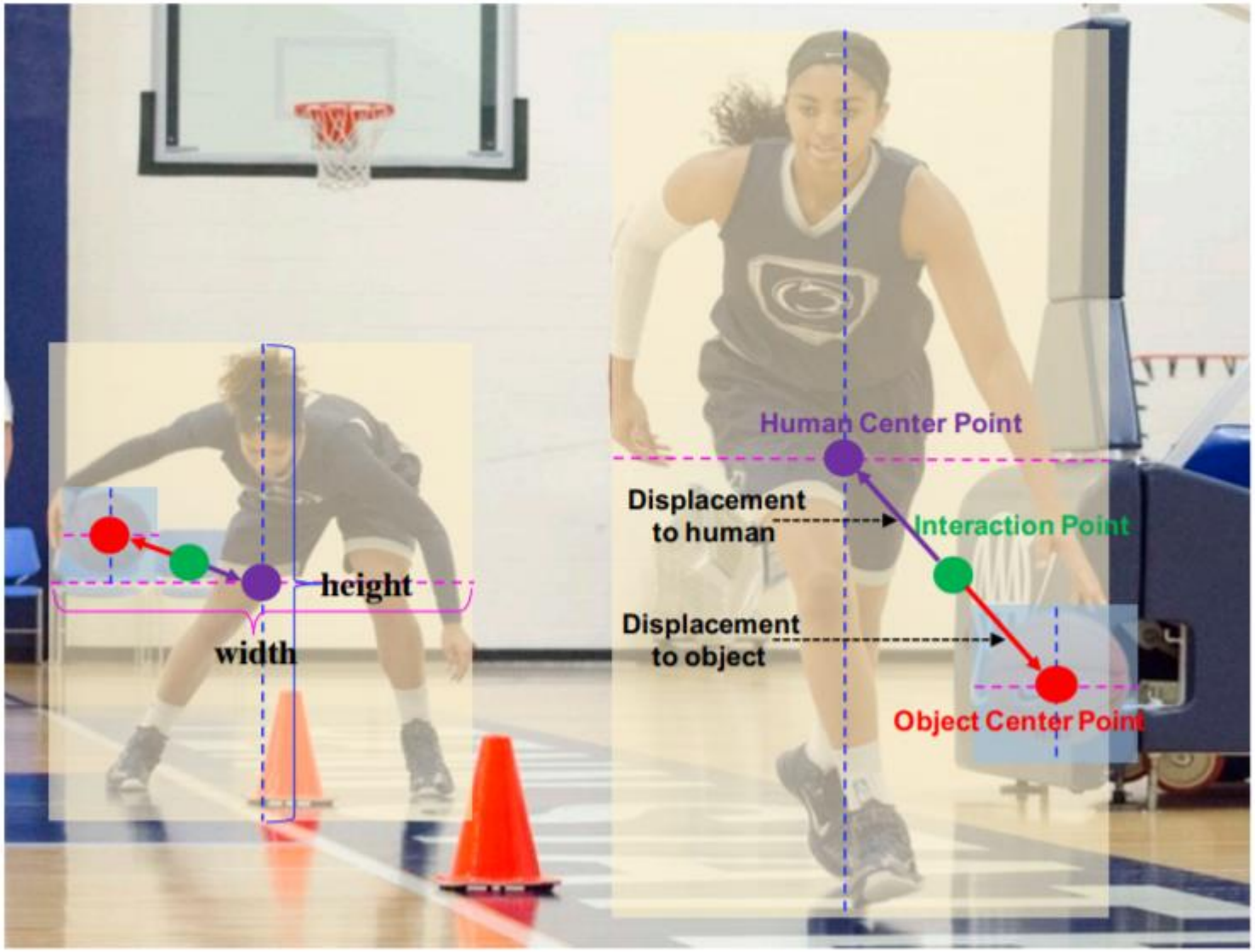


목표:

HOI 문제를 점 삼중항으로 정의하여 포인트 감지분기와 포인트 일치 분기를 병렬 아키텍처를 사용하여 쓸모없는 삼중항을 억제함으로써 정확도를 높임과 동시에 계산 비용을 절감한다.



메인 아이디어

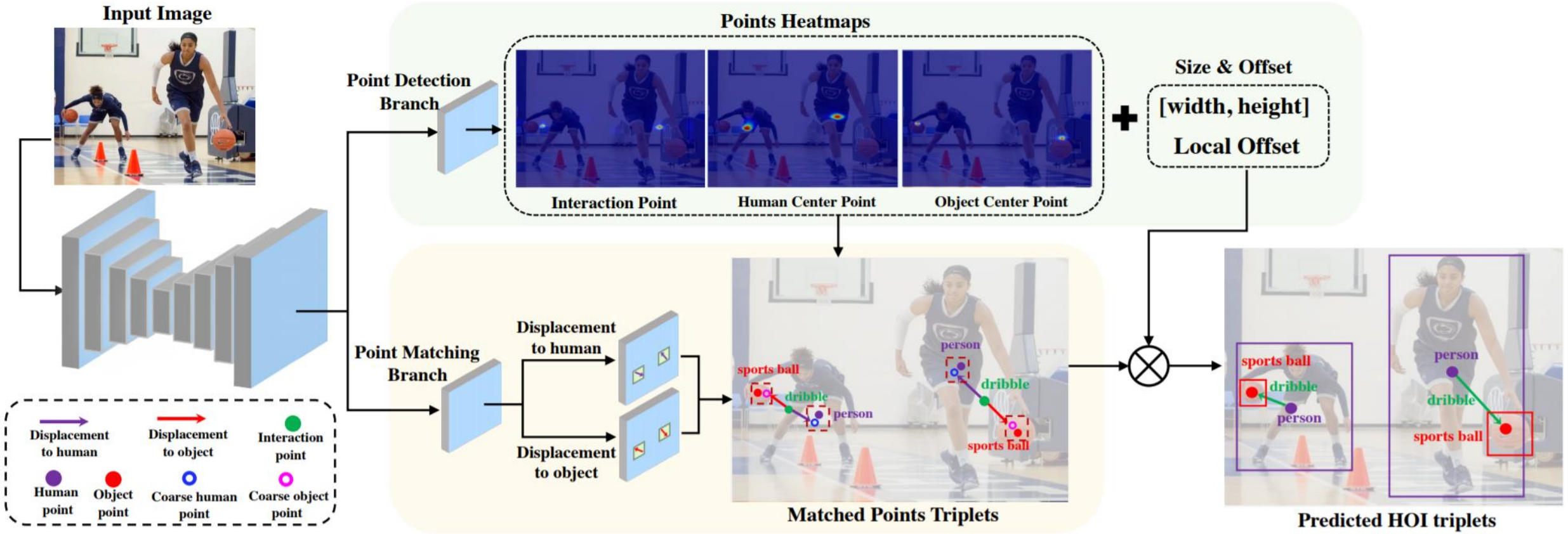


두개의 병렬 분기

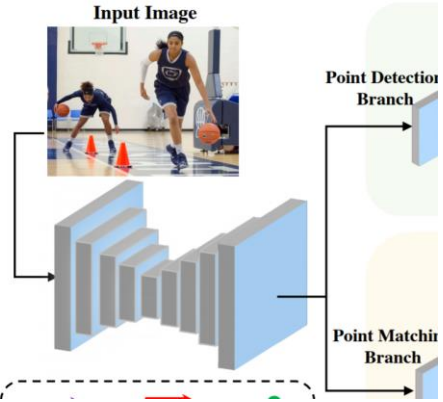
1) 포인트 검출
   - 사람 중심점, 객체 중심점, 상호작용 중심점 찾기
2) 포인트 매칭
   - 상호작용 중심점에서 인간과 물체로의 변위 추정

➢ 동일한 상호작용 중심점에서의 변위가 일치하면 상호작용 쌍으로 간주함

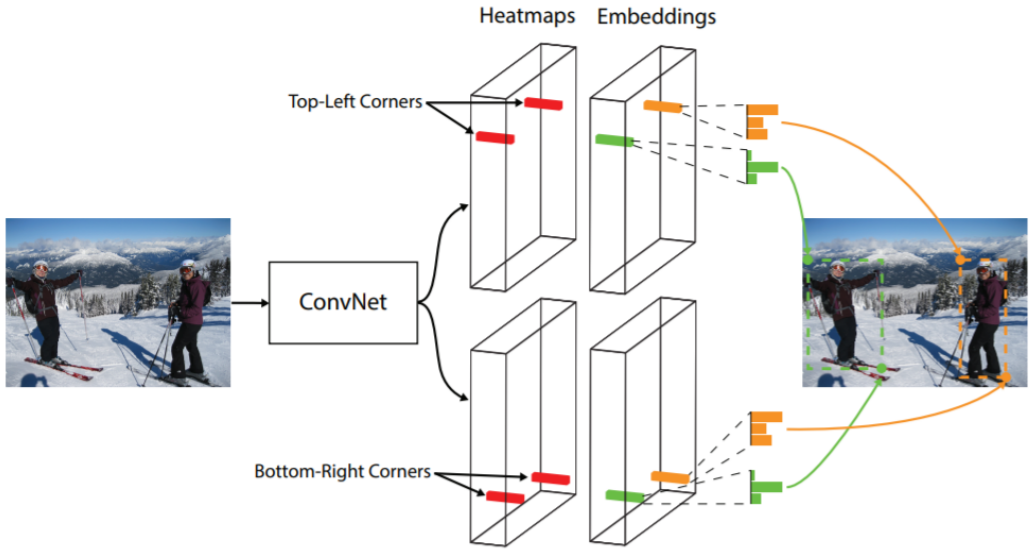
PPDM : Parallel Point Detection and Matching for Real-time Human-Object Interaction Detection



피쳐 추출기(heatmap prediction networks)



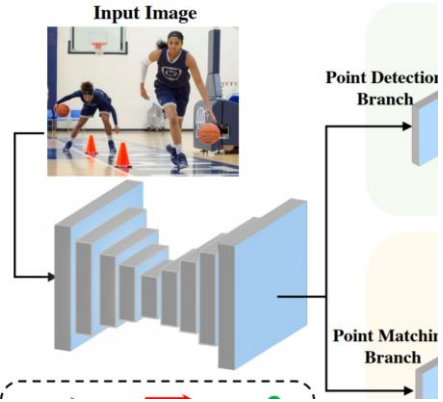
Hourglass-104(CornetNet)



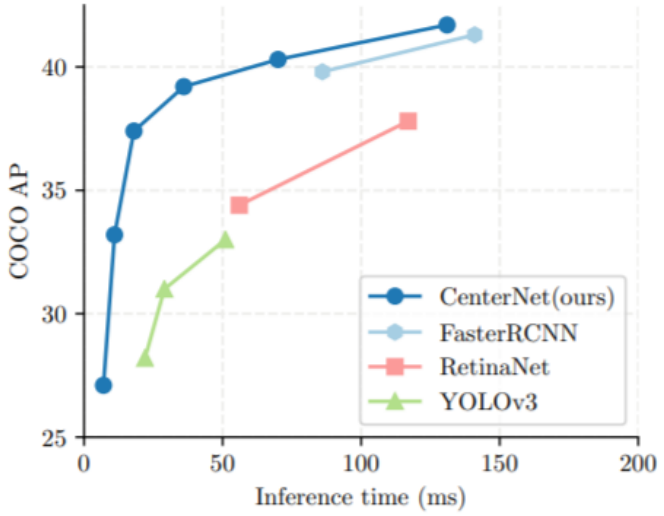
Method	Backbone	AP
Two-stage detectors		
DeNet (Tychsen-Smith and Petersson, 2017a)	ResNet-101	33.8
CoupleNet (Zhu et al., 2017)	ResNet-101	34.4
Faster R-CNN by G-RMI (Huang et al., 2017)	Inception-ResNet-v2 (Szegedy et al., 2017)	34.7
Faster R-CNN+++ (He et al., 2016)	ResNet-101	34.9
Faster R-CNN w/ FPN (Lin et al., 2016)	ResNet-101	36.2
Faster R-CNN w/ TDM (Shrivastava et al., 2016)	Inception-ResNet-v2	36.8
D-FCN (Dai et al., 2017)	Aligned-Inception-ResNet	37.5
Regionlets (Xu et al., 2017)	ResNet-101	39.3
Mask R-CNN (He et al., 2017)	ResNeXt-101	39.8
Soft-NMS (Bodla et al., 2017)	Aligned-Inception-ResNet	40.9
LH R-CNN (Li et al., 2017)	ResNet-101	41.5
Fitness-NMS (Tychsen-Smith and Petersson, 2017b)	ResNet-101	41.8
Cascade R-CNN (Cai and Vasconcelos, 2017)	ResNet-101	42.8
D-RFCN + SNIP (Singh and Davis, 2017)	DPN-98 (Chen et al., 2017)	45.7
One-stage detectors		
YOLOv2 (Redmon and Farhadi, 2016)	DarkNet-19	21.6
DSOD300 (Shen et al., 2017a)	DS/64-192-48-1	29.3
GRP-DSOD320 (Shen et al., 2017b)	DS/64-192-48-1	30.0
SSD513 (Liu et al., 2016)	ResNet-101	31.2
DSSD513 (Fu et al., 2017)	ResNet-101	33.2
RefineDet512 (single scale) (Zhang et al., 2017)	ResNet-101	36.4
RetinaNet800 (Lin et al., 2017)	ResNet-101	39.1
RefineDet512 (multi scale) (Zhang et al., 2017)	ResNet-101	41.8
CornerNet511 (single scale)	Hourglass-104	40.6
CornerNet511 (multi scale)	Hourglass-104	42.2

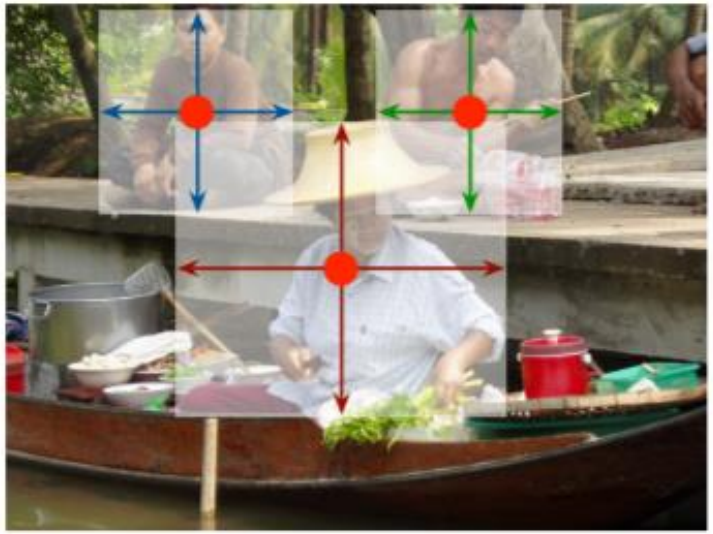
https://arxiv.org/abs/1808.01244

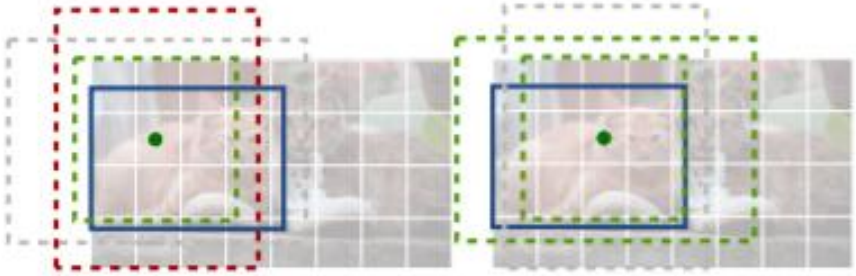
피쳐 추출기(heatmap prediction networks)

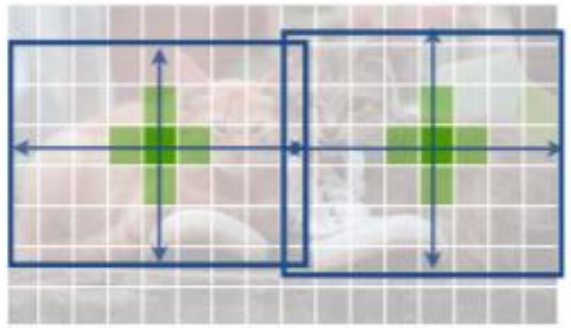


DLA-34







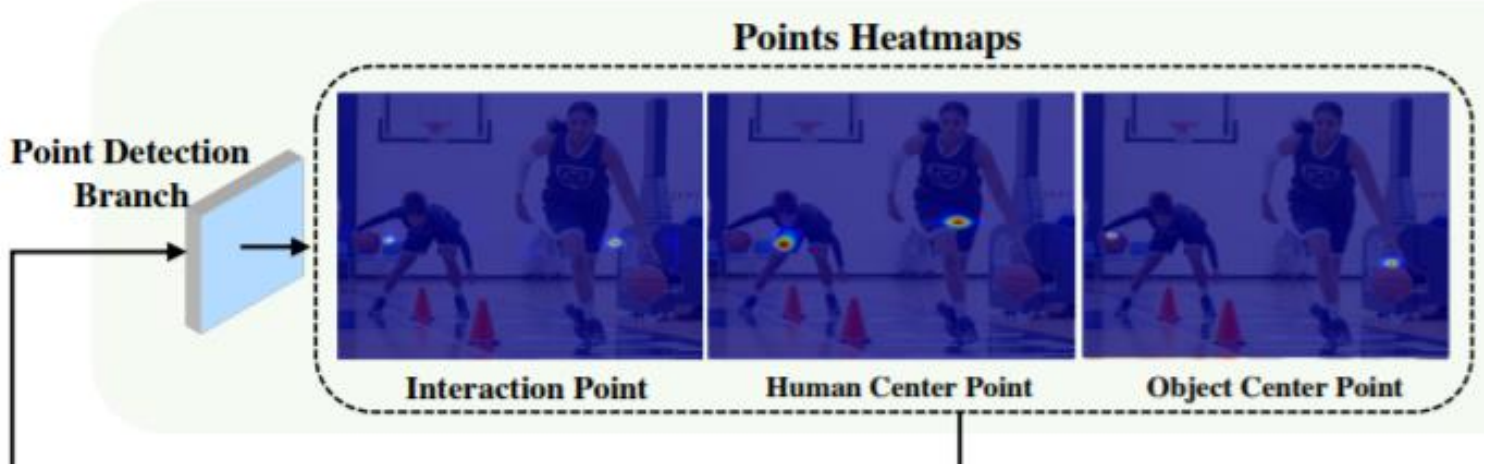


	AP			Time (ms)			FPS		
	N.A.	F	MS	N.A.	F	MS	N.A.	F	MS
Hourglass-104	**40.3**	**42.2**	**45.1**	71	129	672	14	7.8	1.4
DLA-34	37.4	39.2	41.7	19	36	248	52	28	4
ResNet-101	34.6	36.2	39.3	22	40	259	45	25	4
ResNet-18	28.1	30.0	33.2	**7**	**14**	**81**	**142**	**71**	**12**

https://arxiv.org/abs/1904.07850

포인트 디텍션

사람 상자, 객체 상자, 상호작용 지점을 추정(가우시안 커널을 사용하여 히트맵으로 표시)



$(x^h, y^h) \in \mathbb{R}^2$

$(w^h, h^h) \in \mathbb{R}^2$

$\boldsymbol{\delta c^h} \in \mathbb{R}^2$

(x^h, y^h)

$\tilde{\boldsymbol{C}}^{\boldsymbol{h}} \in [0,1]^{\frac{H}{d} \times \frac{W}{d}}$

$(x^a, y^a) \in \mathbb{R}^2$

(x^o, y^o)

$\tilde{\boldsymbol{C}}^{o} \in [0,1]^{T \times \frac{H}{d} \times \frac{W}{d}}$

$(\tilde{x}^a, \tilde{y}^a) = (\lfloor \frac{\tilde{x}^h + \tilde{x}^o}{2} \rfloor, \lfloor \frac{\tilde{y}^h + \tilde{y}^o}{2} \rfloor).$

(x^a, y^a)

$\tilde{\boldsymbol{C}}^{a} \in [0,1]^{K \times \frac{H}{d} \times \frac{W}{d}}$

포인트 디텍션

로스, 알파2, 배타4, 사람과 오브젝트도 동일한 식 사용하여 로스 적용

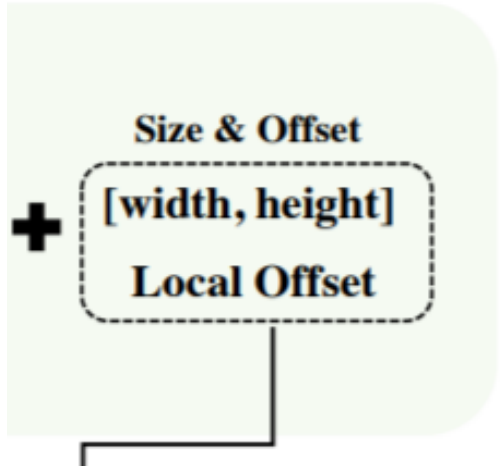
$$L_a = -\frac{1}{N}\sum_{kxy}\begin{cases}(1-\hat{C}^a_{kxy})^{\alpha}\log(\hat{C}^a_{kxy}) & \text{if } \tilde{C}^a_{kxy}=1\\ (1-\tilde{C}^a_{kxy})^{\beta}(\hat{C}^a_{kxy})^{\alpha} & \text{otherwise}\\ \log(1-\hat{C}^a_{kxy}), & \end{cases} \tag{1}$$

크기 및 오프셋 손실

$$(\tilde{\delta}^x_{(\tilde{x}^h,\tilde{y}^h)}, \tilde{\delta}^y_{(\tilde{x}^h,\tilde{y}^h)}) = (\frac{x^h}{d} - \tilde{x}^h, \frac{y^h}{d} - \tilde{y}^h).$$

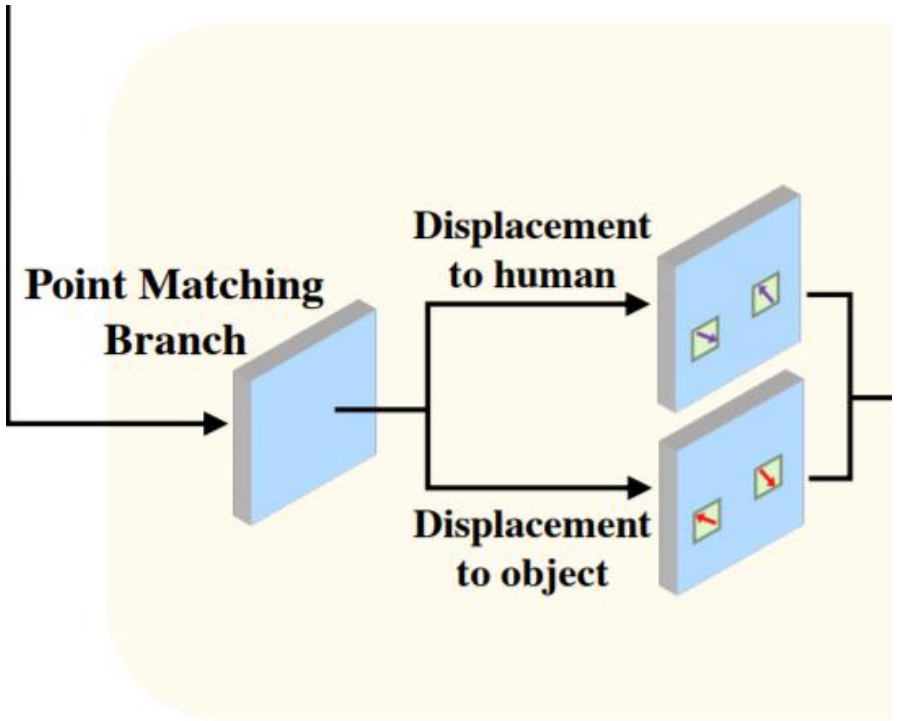
$$L^h_{off} = \sum_{(\tilde{x}^h,\tilde{y}^h)\in\tilde{S}^{\bar{h}}} (|\tilde{\delta}^x_{(\tilde{x}^h,\tilde{y}^h)} - \hat{\delta}^x_{(\tilde{x}^h,\tilde{y}^h)}| + |\tilde{\delta}^y_{(\tilde{x}^h,\tilde{y}^h)} - \hat{\delta}^y_{(\tilde{x}^h,\tilde{y}^h)}|, \tag{3}$$

$$L_{off} = \frac{1}{M+D}(L^h_{off} + L^o_{off}) \qquad M = |\tilde{S}^h| \text{ and } D = |\tilde{S}^o|$$



포인트 매칭

상호작용 점을 앵커로 하여 사람 쪽과 객체 쪽으로 변위를 추정



로스 ~(그라운드트루), ^(예측된), L_{ao}

$$(\tilde{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)}, \tilde{d}^{hy}_{(\tilde{x}^a,\tilde{y}^a)}) = (\tilde{x}^a - \tilde{x}^h, \tilde{y}^a - \tilde{y}^h)$$

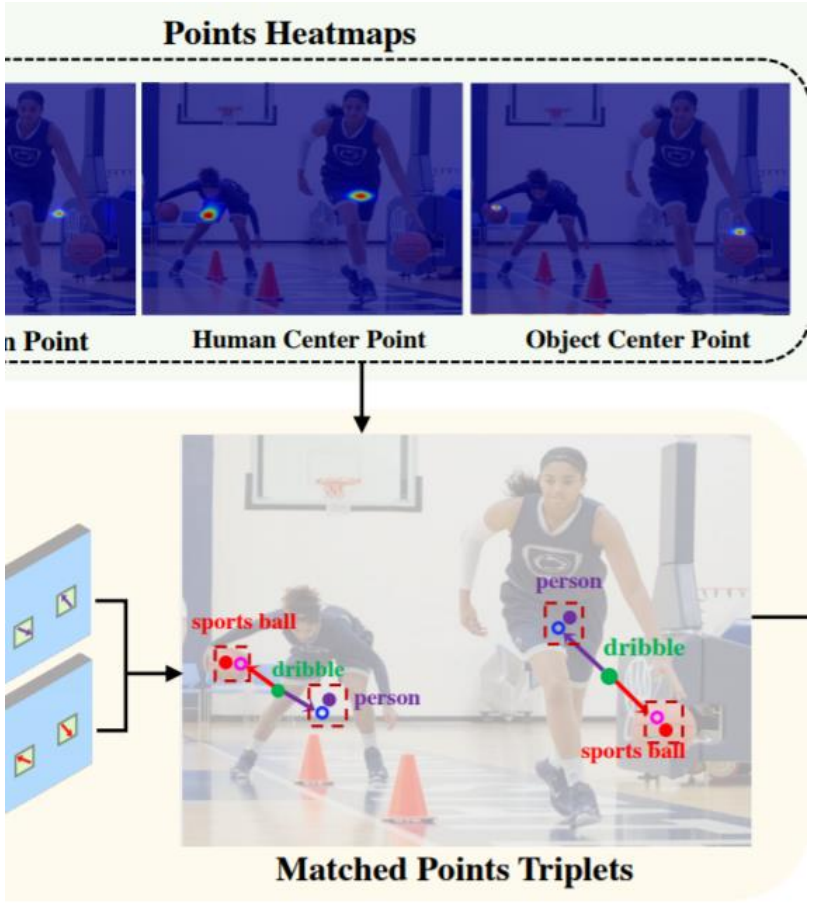
$$L_{ah} = \frac{1}{N} \sum_{(\tilde{x}^a,\tilde{y}^a)\in\tilde{S}^a} |\hat{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)} - \tilde{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)}| + |\hat{d}^{hx}_{(\tilde{x}^a,\tilde{y}^a)} - \tilde{d}^{hy}_{(\tilde{x}^a,\tilde{y}^a)}|, \quad (4)$$

$$\boldsymbol{d^{ah}} = (d^{ah}_x, d^{ah}_y)$$

$$\boldsymbol{d^{ao}} = (d^{ao}_x, d^{ao}_y)$$

포인트 매칭

트리플넷 매칭
가우시안 분포인 C점수가 크고, 예측된 값의 차이가 가까운 것들중 최적의 것을 선택
객체상자도 같은 식으로 선택

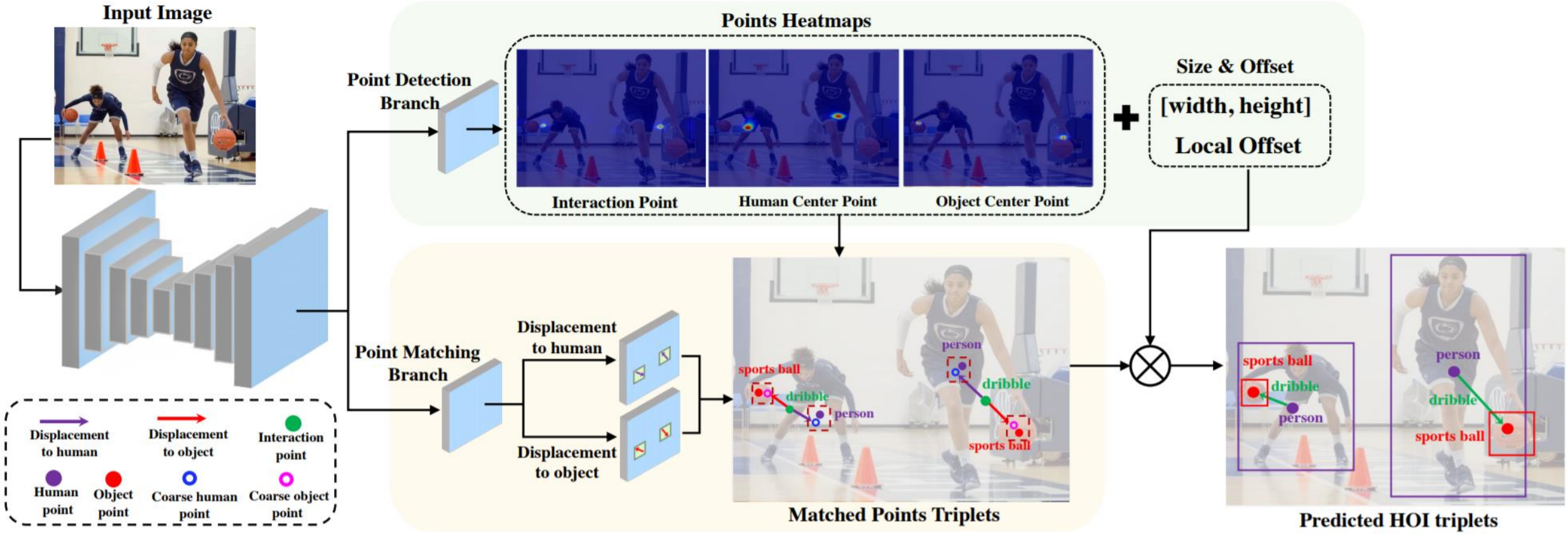


$$(\hat{x}^h_{opt}, \hat{y}^h_{opt}) = \underset{(\hat{x}^h,\hat{y}^h)\in\hat{S}^h}{\arg\min} \frac{1}{C^h_{(\hat{x}^h,\hat{y}^h)}} \cdot (|(\hat{x}^a,\hat{y}^a) - (\hat{d}^{hx}_{(\hat{x}^a,\hat{y}^a)}, \hat{d}^{hy}_{(\hat{x}^a,\hat{y}^a)}) - (\hat{x}^h,\hat{y}^h)|) \quad (5)$$

최종 손실

$$\hat{x}^h_{ref} = \hat{x}^h_{opt} + \hat{\delta}^x_{(\hat{x}^h_{opt},\hat{y}^h_{opt})} \quad (\hat{x}^h_{ref} - \frac{\hat{w}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}, \hat{y}^h_{ref} - \frac{\hat{h}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2},$$

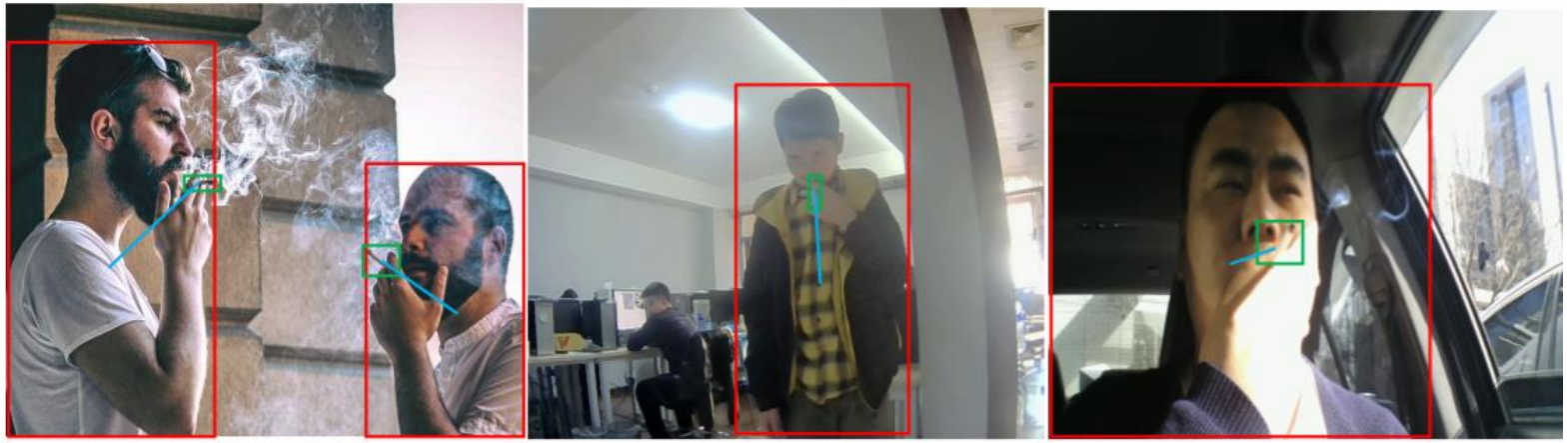
$$\hat{y}^h_{ref} = \hat{y}^h_{opt} + \hat{\delta}^y_{(\hat{x}^h_{opt},\hat{y}^h_{opt})} \quad \hat{x}^h_{ref} + \frac{\hat{w}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}, \hat{y}^h_{ref} + \frac{\hat{h}_{(\hat{x}^h_{opt},\hat{y}^h_{opt})}}{2}. \quad (7)$$



$$L = L_a + L_h + L_o + \lambda(L_{ah} + L_{ao} + L_{wh}) + L_{off} \quad (6)$$

$$\hat{C}^p_{\hat{x}^h_{ref}\hat{y}^h_{ref}} \hat{C}^o_{\hat{x}^o_{ref}\hat{y}^o_{ref}} \hat{C}^a_{\hat{x}^a_{ref}\hat{y}^a_{ref}}.$$

HOI-A Dataset(Human-Object Interaction for Application)



a. <human, smoke, cigarette>
outdoor

b. <human, smoke, cigarette>
indoor

c. <human, smoke, cigarette>
in car & intense illumination

d. <human, smoke, cigarette>
in dark scene

e. Attacking smoke: no cigarette
negative sample

f. no predefined interaction
negative sample

HOI를 위한 긍정적인 이미지와 부정적인 이미지를 모아 훈련할때 더 효과적으로 할 수 있도록 하는 데이터셋
38,668 이미지, 11개 객체, 10개 액션

43,820 휴먼 등장, 60,438개 오브젝트 등장, 96,160개 상호작용 등장

Dataset

HICO-DET
Images 47,776 (38,118, 9,658)
Objects 80 (airplane, apple…)
Verbs 117 (carry, catch…)
HOI 600 (airplane – board, direct, exit, fly…)

HOI Remark >= 150k

HOI-A

평가지표

mAP

실험 (HICO-DET)

Method	Feature	Full(mAP %) ↑	Rare(mAP %) ↑	Non-Rare(mAP %) ↑	Inference Time (ms) ↓	FPS ↑
Shen *et. al* [23]	A + P	6.46	4.24	7.12	-	-
HO-RCNN [2]	A + S	7.81	5.37	8.54	-	-
VSRL [10]	A	9.09	7.02	9.71	-	-
InteractNet [8]	A	9.94	7.16	10.77	145	6.90
GPNN [21]	A	13.11	9.34	14.23	197 + 48 = 245	4.08
Xu *et. al* [27]	A + L	14.70	13.26	15.13	-	-
iCAN [6]	A + S	14.84	10.45	16.15	92 + 112 = 204	4.90
PMFNet-Base [25]	A + S	14.92	11.42	15.96	-	-
Wang *et. al* [26]	A	16.24	11.16	17.75	-	-
No-Frills [11]	A + S + P	17.18	12.17	18.68	197 + 230 + 67 = 494	2.02
TIN [15]	A + S + P	17.22	13.51	18.32	92 + 98 + 323 = 513	1.95
RPNN [32]	A + P	17.35	12.78	18.71	-	-
PMFNet [25]	A + S + P	17.46	**15.65**	18.00	92 + 98 + 63 = 253	3.95
PPDM-DLA	A	19.02	12.65	20.92	**27**	**37.03**
PPDM-Hourglass	A	**21.10**	14.46	**23.09**	71	14.08

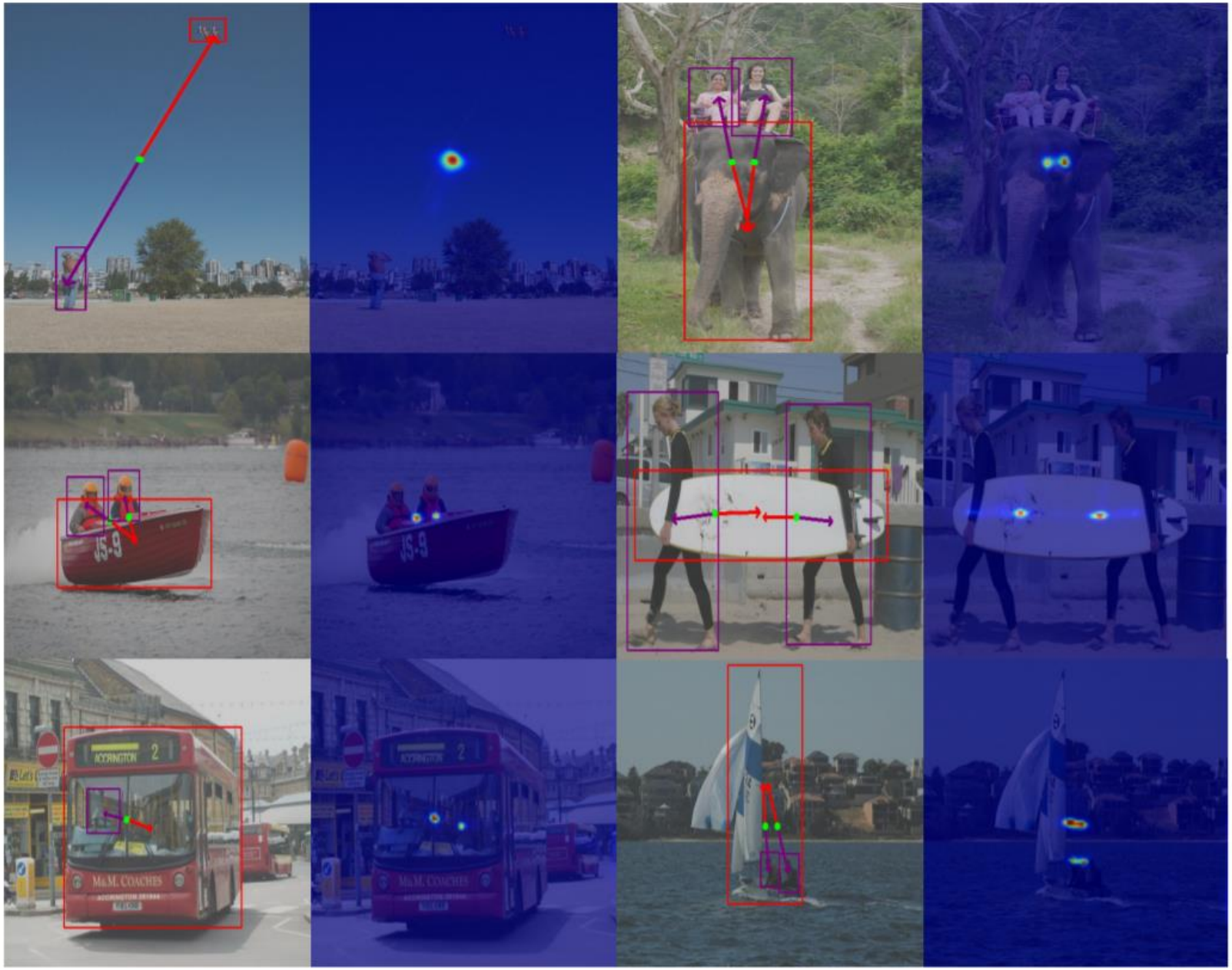
	Method	Full	Rare	Non-Rare	Time
1	Basic Model	18.46	11.97	20.40	24
2	+ Feature Fusion	18.66	11.86	20.69	26
3	+ Global Reasoning	18.63	12.61	20.42	26
4	Union Center	18.07	11.53	20.02	27
5	PPDM-DLA	19.02	12.65	20.92	27

실험 (HOI-A)

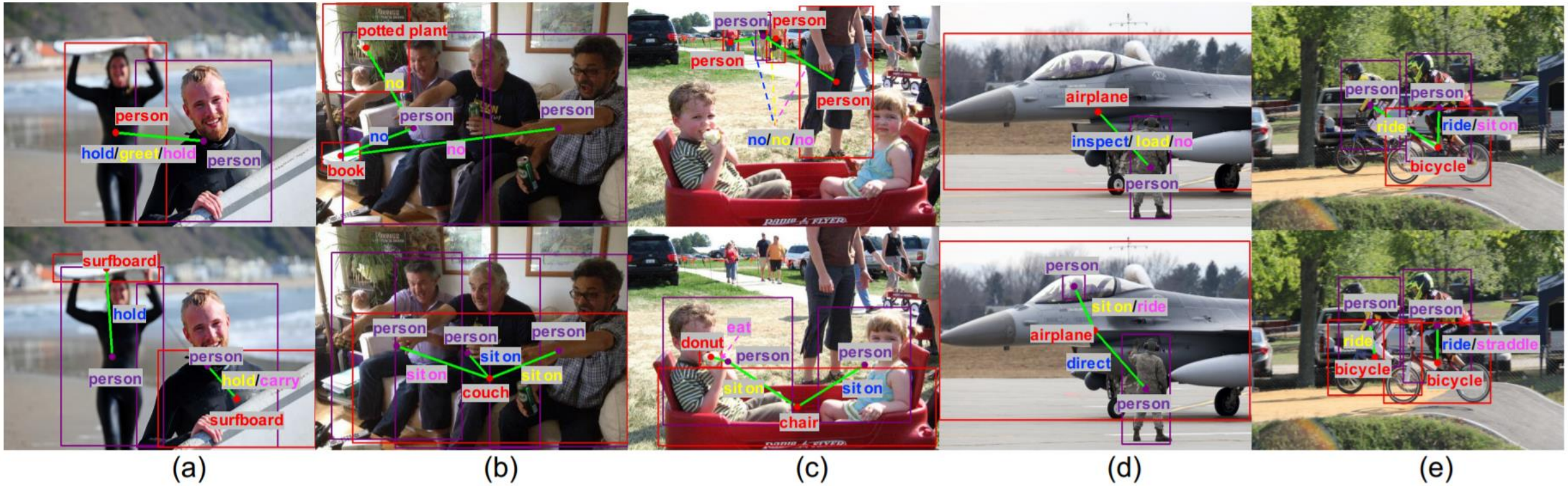
Method	mAP (%)	Time (ms)
Faster Interaction Net [1]	56.03	-
GMVM [1]	60.26	-
URNet [1]	66.04	-
iCAN [6]	44.23	194
TIN [15]	48.64	501
PPDM-DLA	67.03	**27**
PPDM-Hourglass	**71.45**	71

Table 3. Performance comparison on HOI-A test set.

실험 (인터렉션 포인트 히트맵)



실험 (iCAN 과 결과 비교)



Q & A