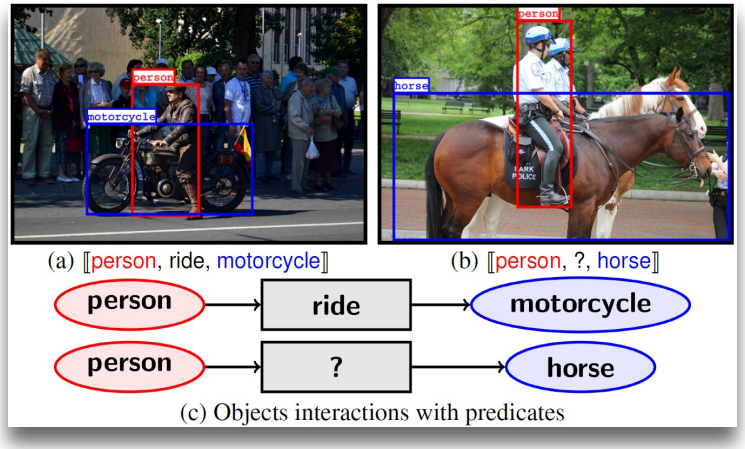
Human-Object Interaction Detection

KCCV 2022 Tutorial
2022.02.24

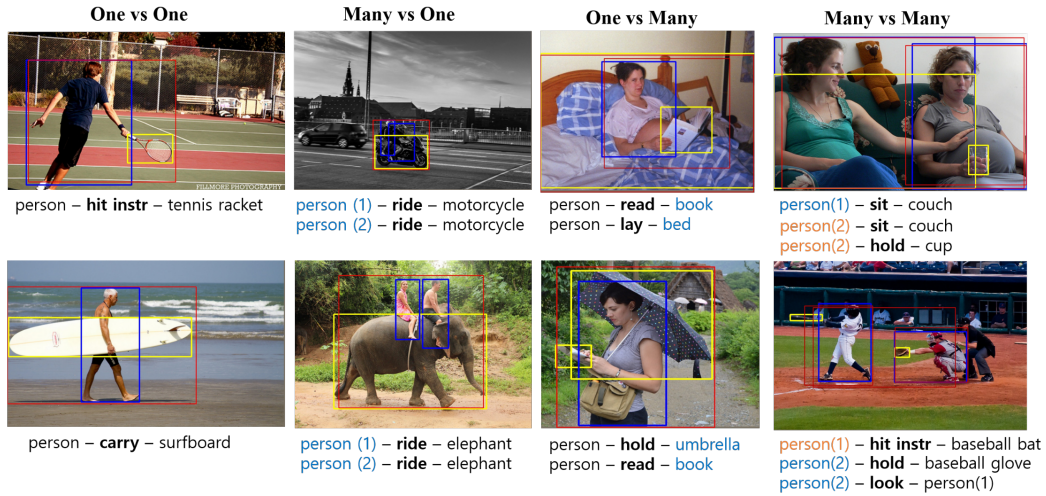
김현우 (MLV Lab)
고려대학교 기계학습 및 비전 연구실
(hyunwoojkim@korea.ac.kr)

목차

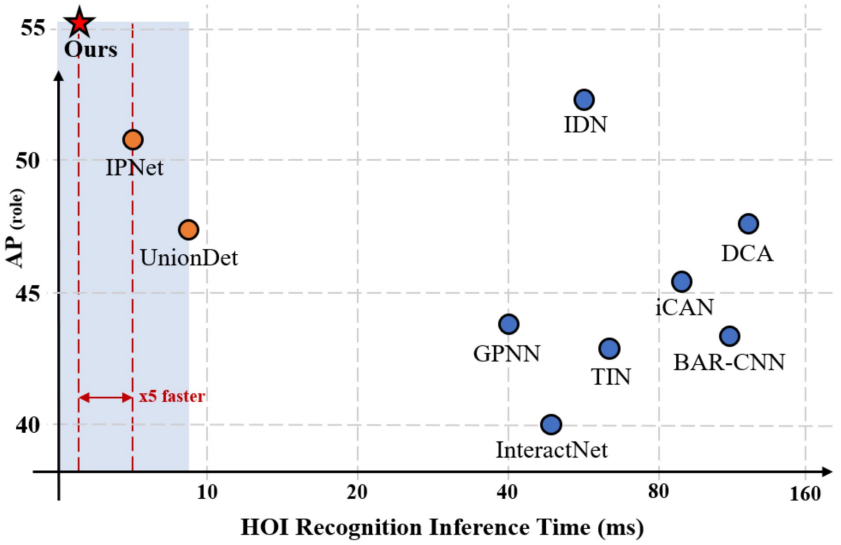
고차원 장면 이해 및 물체 탐지

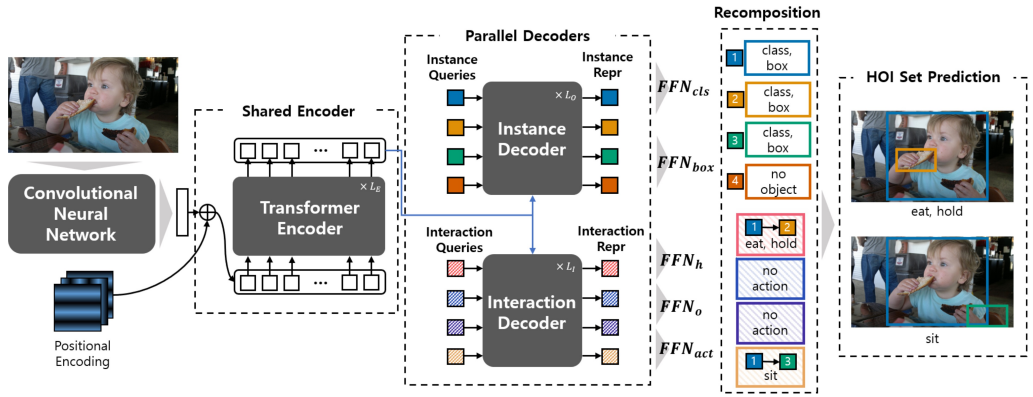


[CVPR18] Tensorize



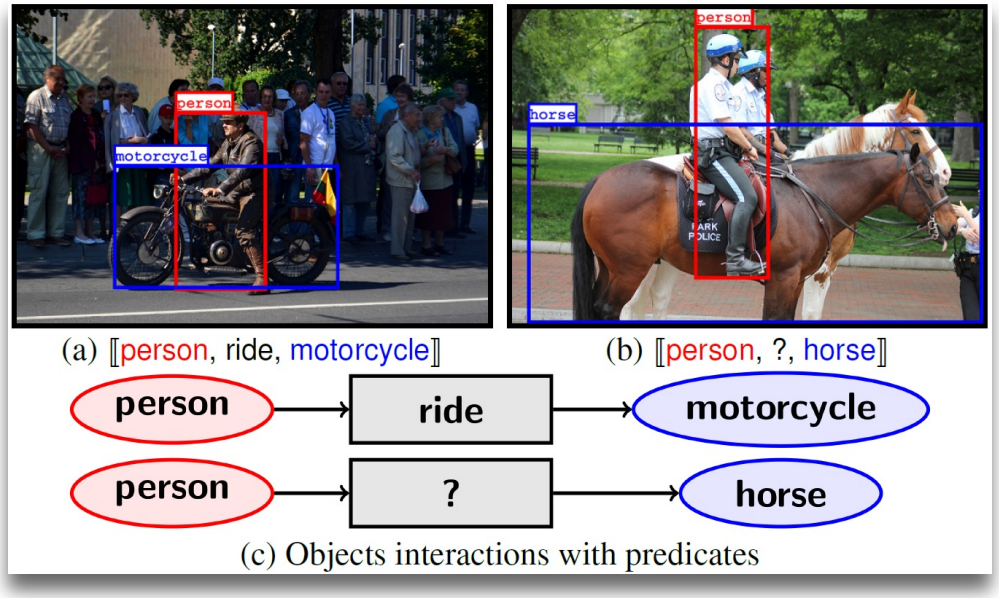
[ECCV20] UnionDet





[CVPR '21 (ORAL)] HOTR

물체간 관계 추론 (사물-사물 상호작용 탐지)

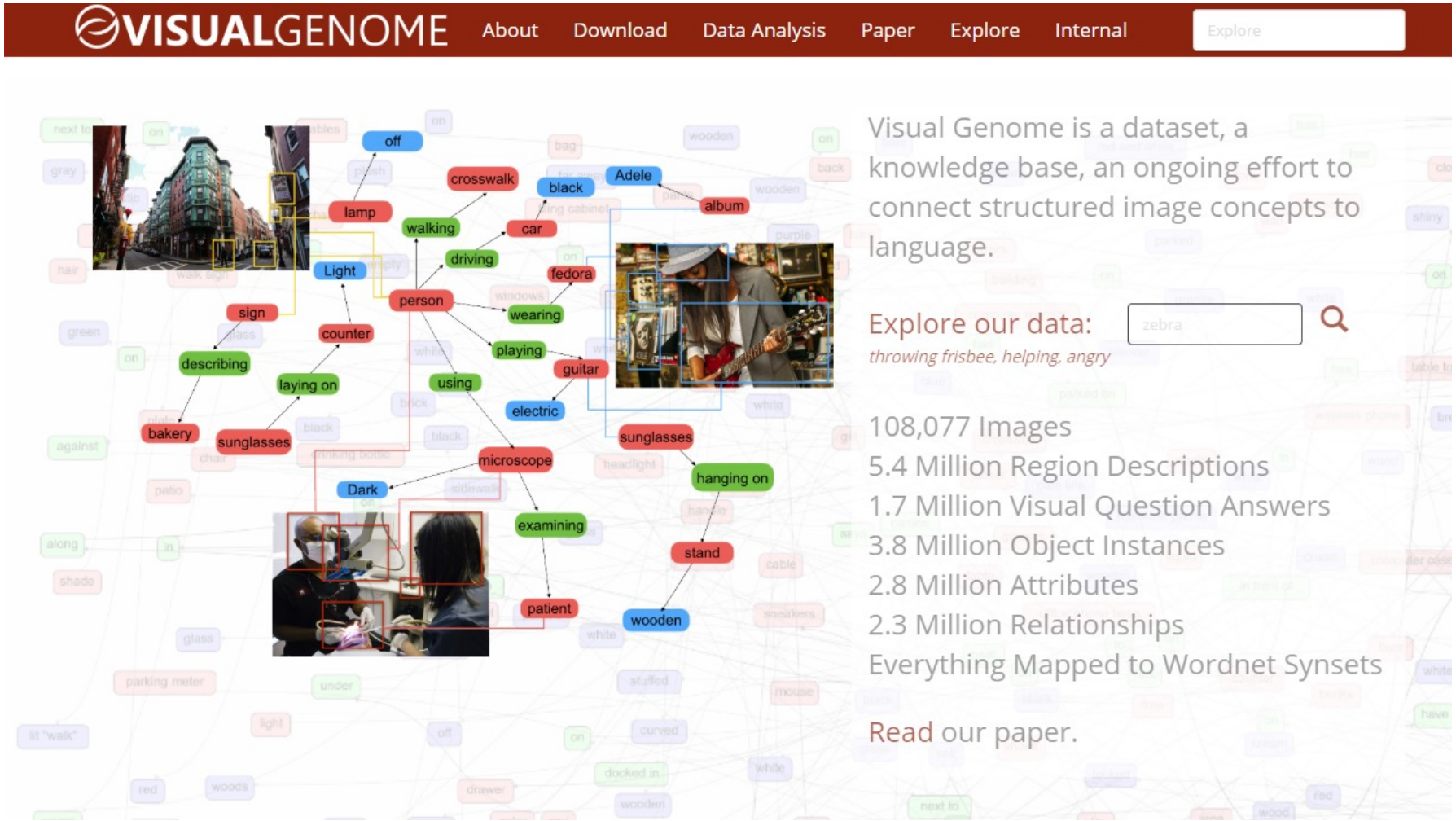


- **태스크 정의:**

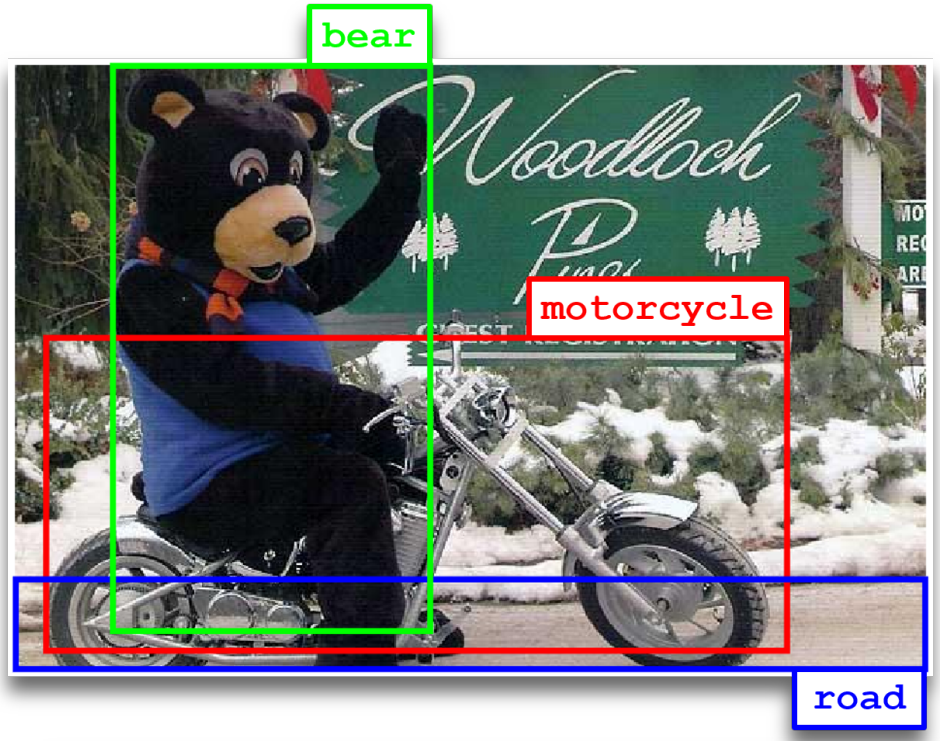
1. 주어진 이미지에서 물체를 탐지
2. 탐지된 물체들의 관계를 추론

- **Goal:**
- **관계 추론 〈물체1, 관계 , 물체2 〉**

비주얼 게놈 (Visual Genome)



장면 그래프 생성 도전성



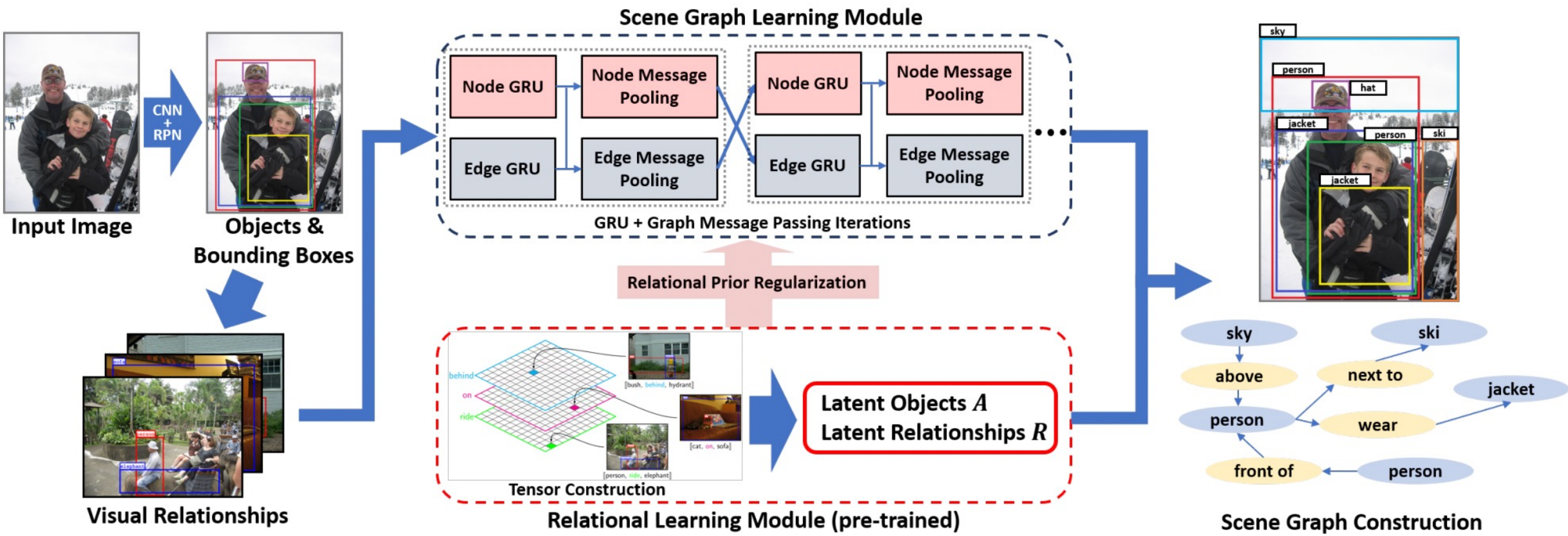
- **의미 추론** 의 어려움
- 데이터 의존성
- 물체관계의 다양성 및 큰 관계공간

  < 물체1, 관계 , 물체2 >

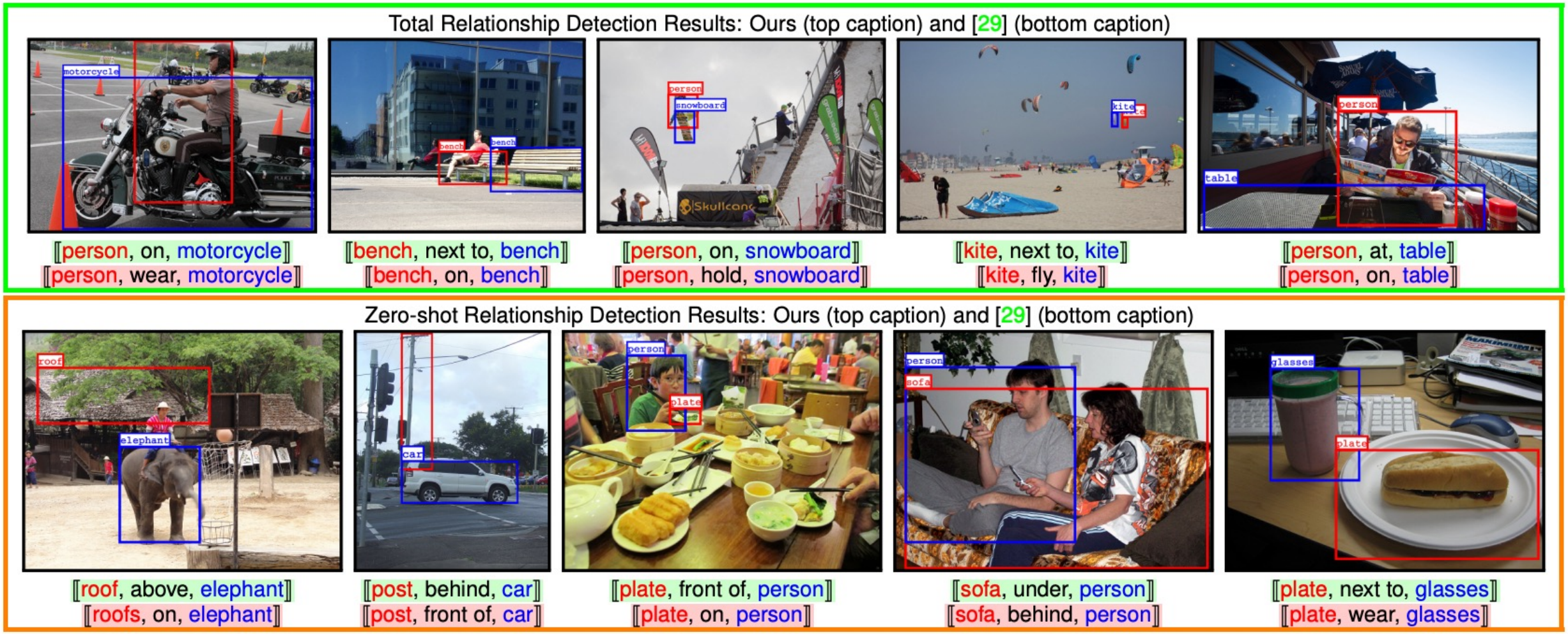
  100 x 100 x 100 ~ 1M
- 제한된 데이터 (Visual Genome)
  - >1M 이미지
  - **가능한 조합의 2% 만 커버**
- **제로샷 학습 (Zero-shot Learning):**
  한번도 본적 없는 관계 예측 필요

해결 방안: 다중관계 텐서 분해. 관계에 대한 사전 지식 생성/습득

장면 그래프 생성 과정


Input Image
CNN + RPN
Objects & Bounding Boxes
Visual Relationships
Scene Graph Learning Module
Node GRU
Node Message Pooling
Edge GRU
Edge Message Pooling
Node GRU
Node Message Pooling
Edge GRU
Edge Message Pooling
GRU + Graph Message Passing Iterations
Relational Prior Regularization
behind
on
ride
[bush, behind, hydrant]
[cat, on, sofa]
[person, ride, elephant]
Tensor Construction
Latent Objects A
Latent Relationships R
Relational Learning Module (pre-trained)
sky
person
hat
jacket
person
ski
jacket
sky
above
person
front of
person
next to
ski
wear
jacket
Scene Graph Construction

실험 결과

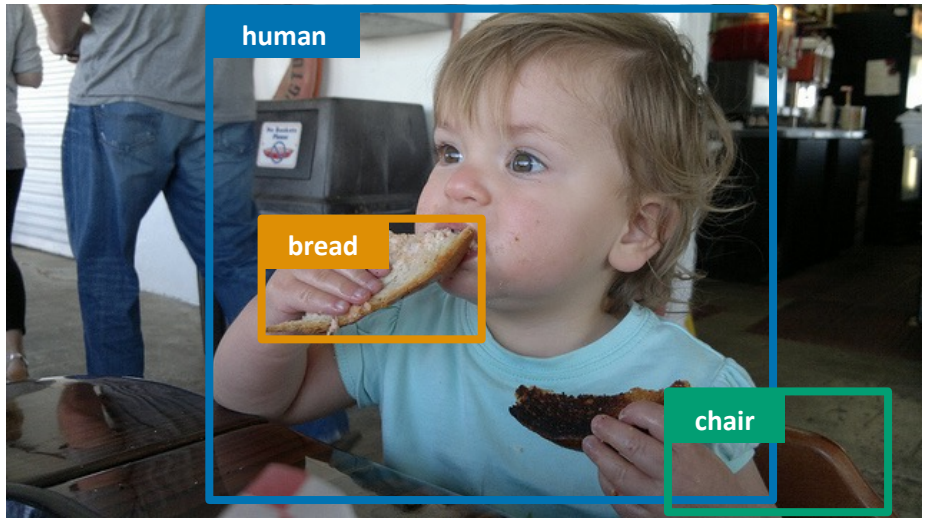

Total Relationship Detection Results: Ours (top caption) and [29] (bottom caption)
motorcycle
[person, on, motorcycle]
[person, wear, motorcycle]
bench
bench
[bench, next to, bench]
[bench, on, bench]
person
snowboard
[person, on, snowboard]
[person, hold, snowboard]
kite
[kite, next to, kite]
[kite, fly, kite]
person
table
[person, at, table]
[person, on, table]
Zero-shot Relationship Detection Results: Ours (top caption) and [29] (bottom caption)
roof
elephant
[roof, above, elephant]
[roofs, on, elephant]
car
[post, behind, car]
[post, front of, car]
person
plate
[plate, front of, person]
[plate, on, person]
person
sofa
[sofa, under, person]
[sofa, behind, person]
glasses
plate
[plate, next to, glasses]
[plate, wear, glasses]

사람-사물 상호작용 탐지

[CVPR ‘18] Gkioxari, Georgia, et al. "Detecting and recognizing human-object interactions."

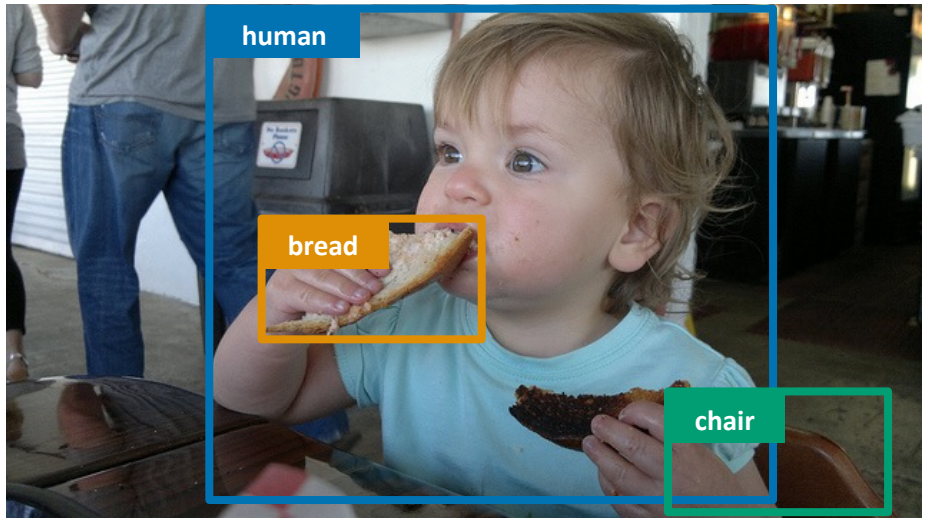
Human-Object Interaction (HOI) Detection

◦ Set {(bbox_1^h, bbox_1^o,[eat, hold]), (bbox_2^h, bbox_2^o,[sit])}



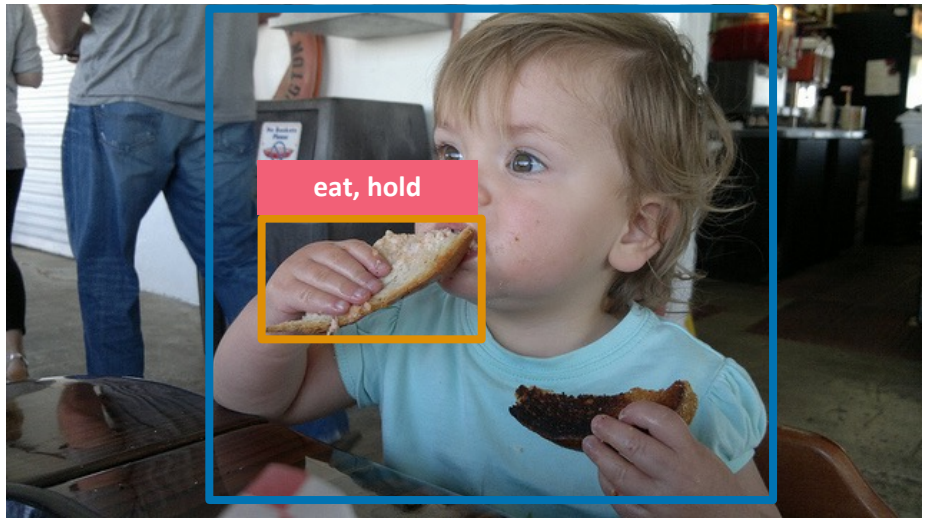
Human-Object Interaction (HOI) Detection

◦ Set {(bbox_1^h, bbox_1^o), (bbox_2^h, bbox_2^o)}



Human-Object Interaction (HOI) Detection

◦ Set {(bbox_1^h, bbox_1^o,[eat, hold]), (bbox_2^h, bbox_2^o)}



Human-Object Interaction (HOI) Detection

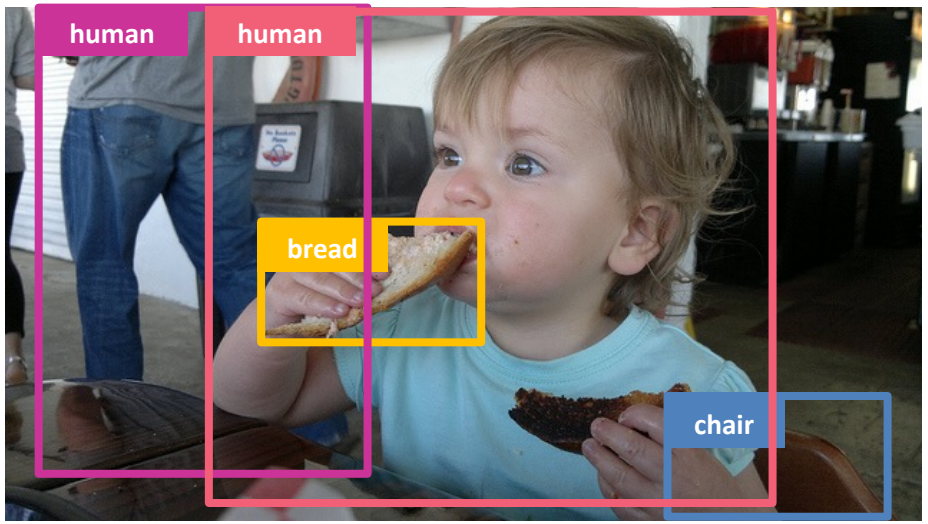
- Set {(bbox_1^h, bbox_1^o,[eat, hold]), (bbox_2^h, bbox_2^o,[sit])}

Previous approach in HOI detection

- Sequential HOI Detectors

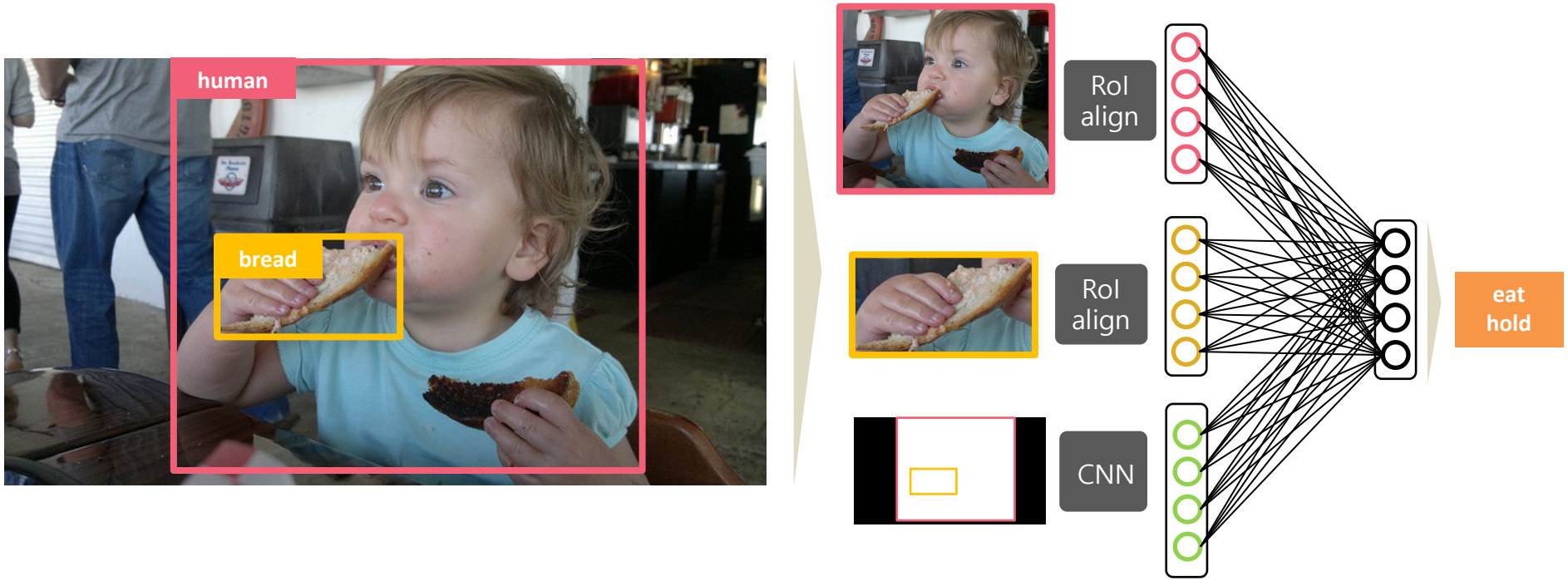
Previous approach in HOI detection

- Sequential HOI Detectors



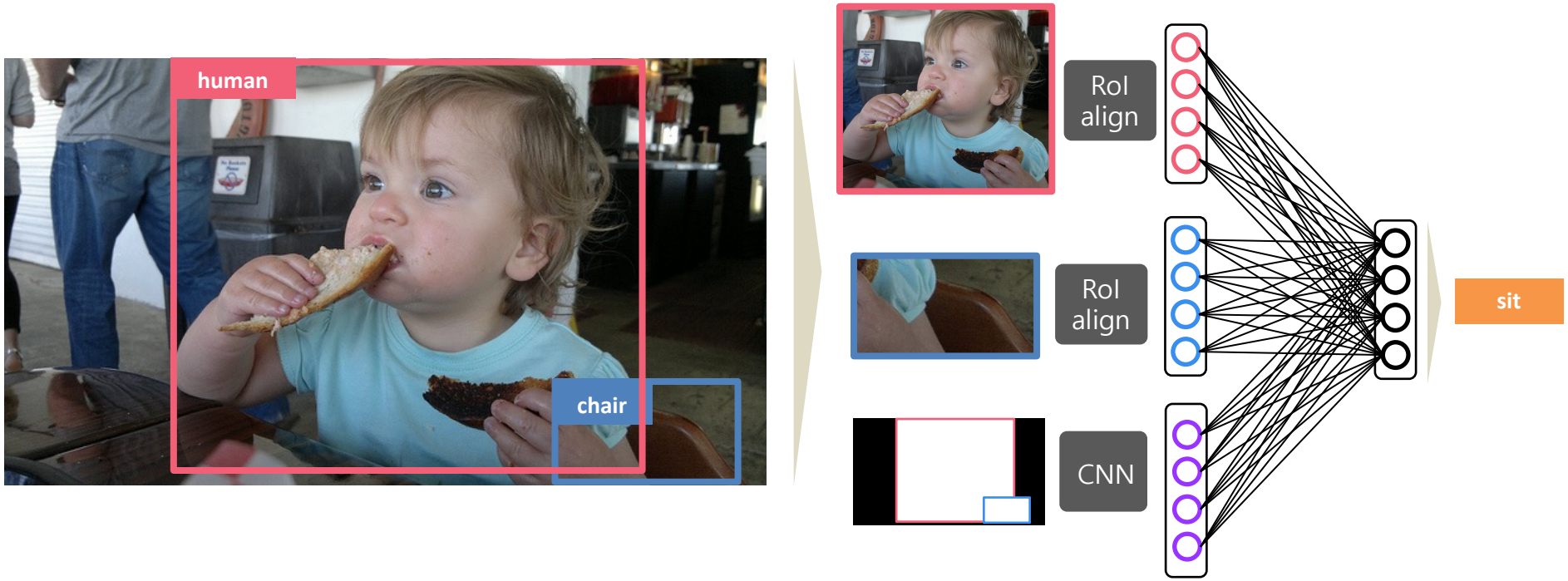
Previous approach in HOI detection

- Sequential HOI Detectors



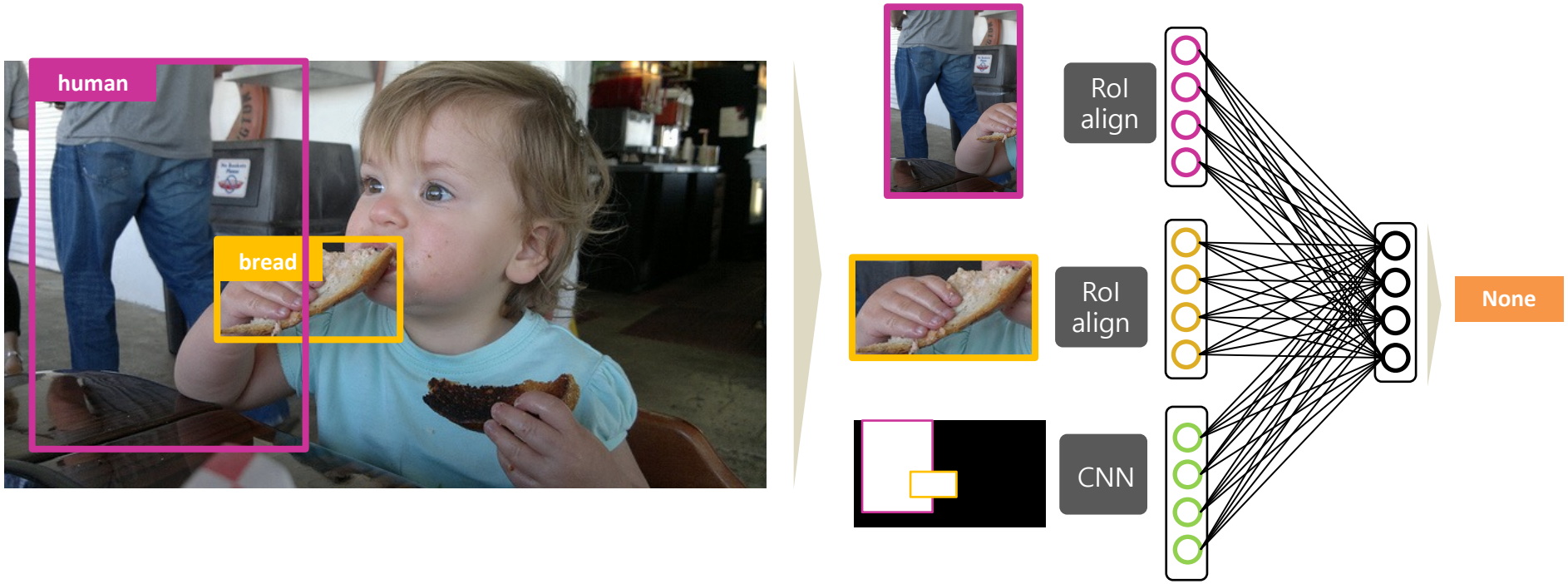
Previous approach in HOI detection

- Sequential HOI Detectors



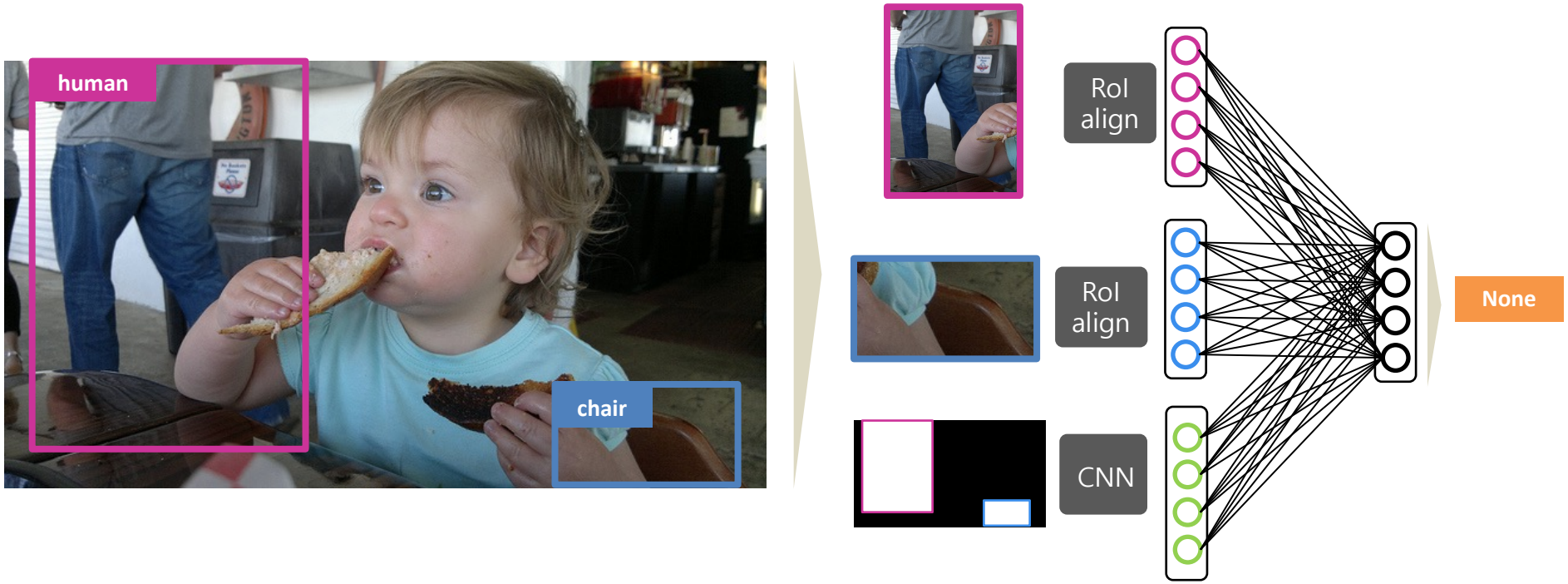
Previous approach in HOI detection

- Sequential HOI Detectors



Previous approach in HOI detection

- Sequential HOI Detectors



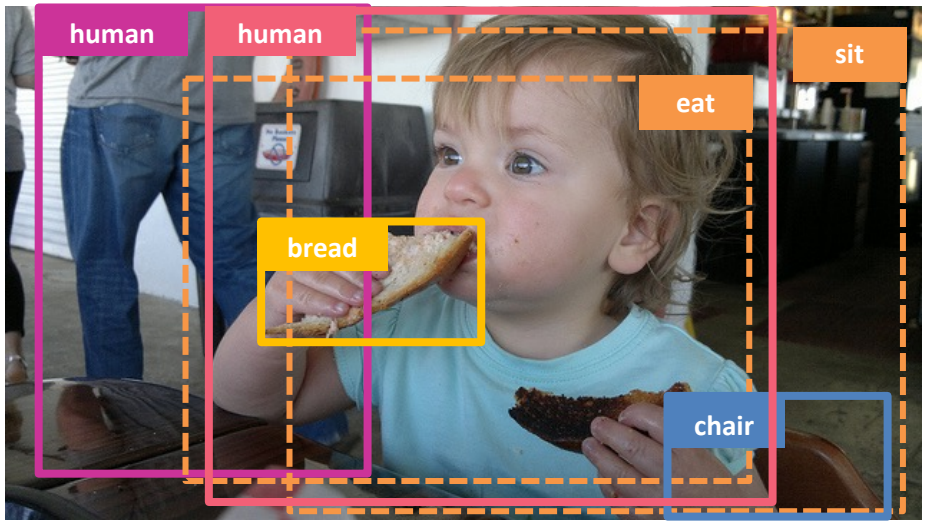
Previous approach in HOI detection

◦ Sequential HOI Detectors

- Intuitive Pipeline
- Pairwise Neural Network Inference : Slow

Previous approach in HOI detection

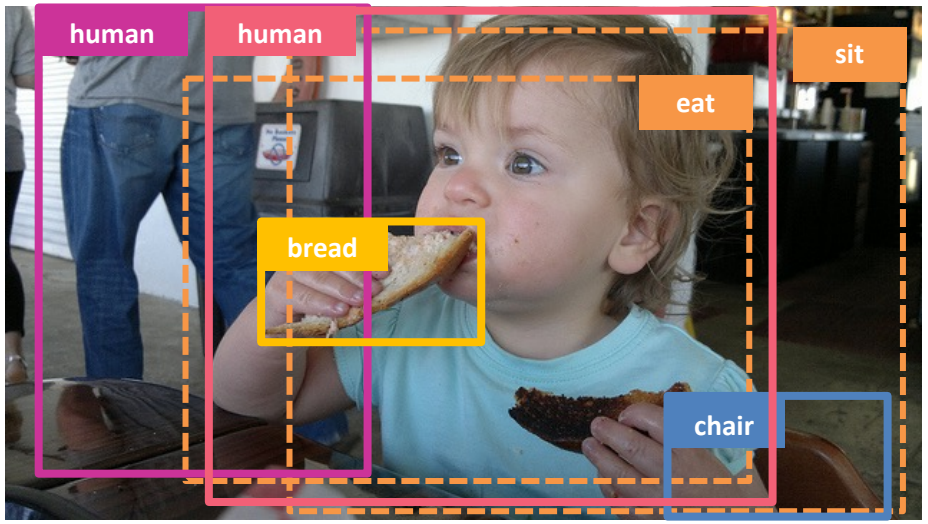
- Parallel HOI Detectors



[ECCV20] UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Previous approach in HOI detection

- Parallel HOI Detectors

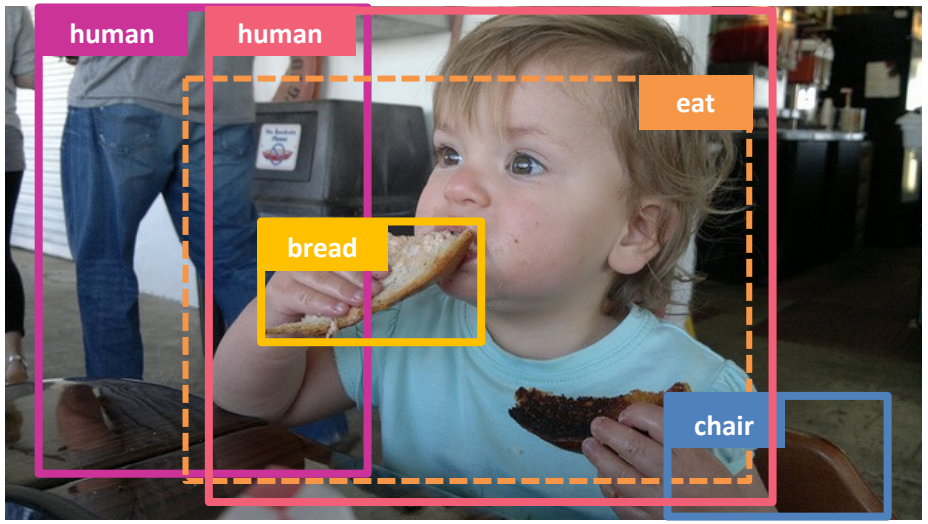


Region of Interaction

[ECCV20] UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Previous approach in HOI detection

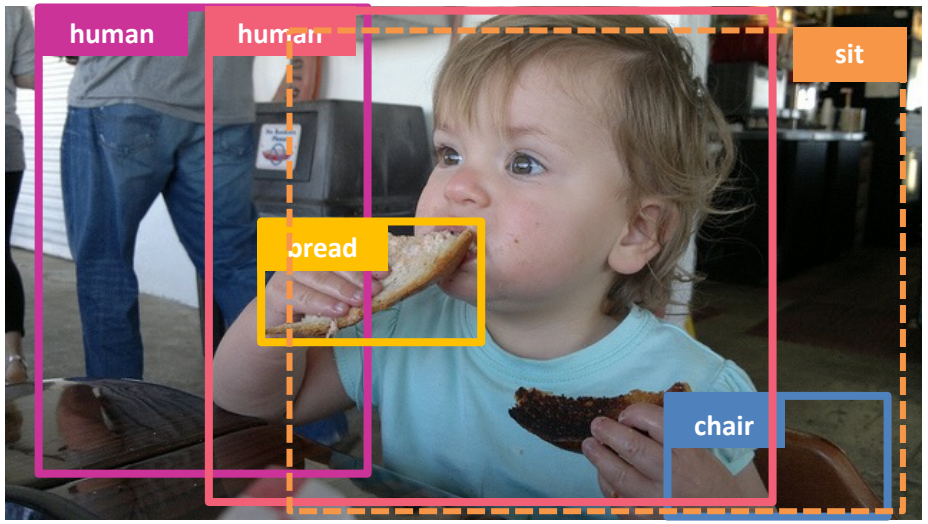
- Parallel HOI Detectors



[ECCV20] UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Previous approach in HOI detection

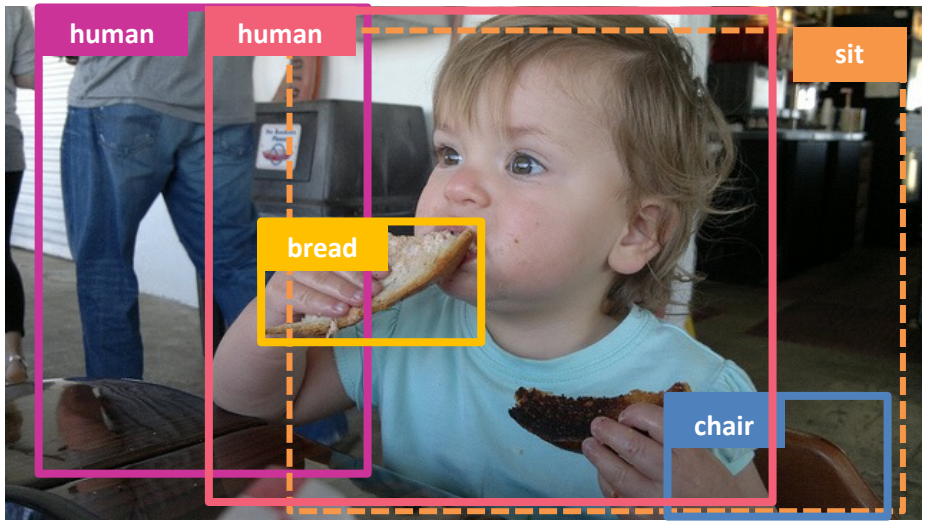
- Parallel HOI Detectors



[ECCV20] UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Previous approach in HOI detection

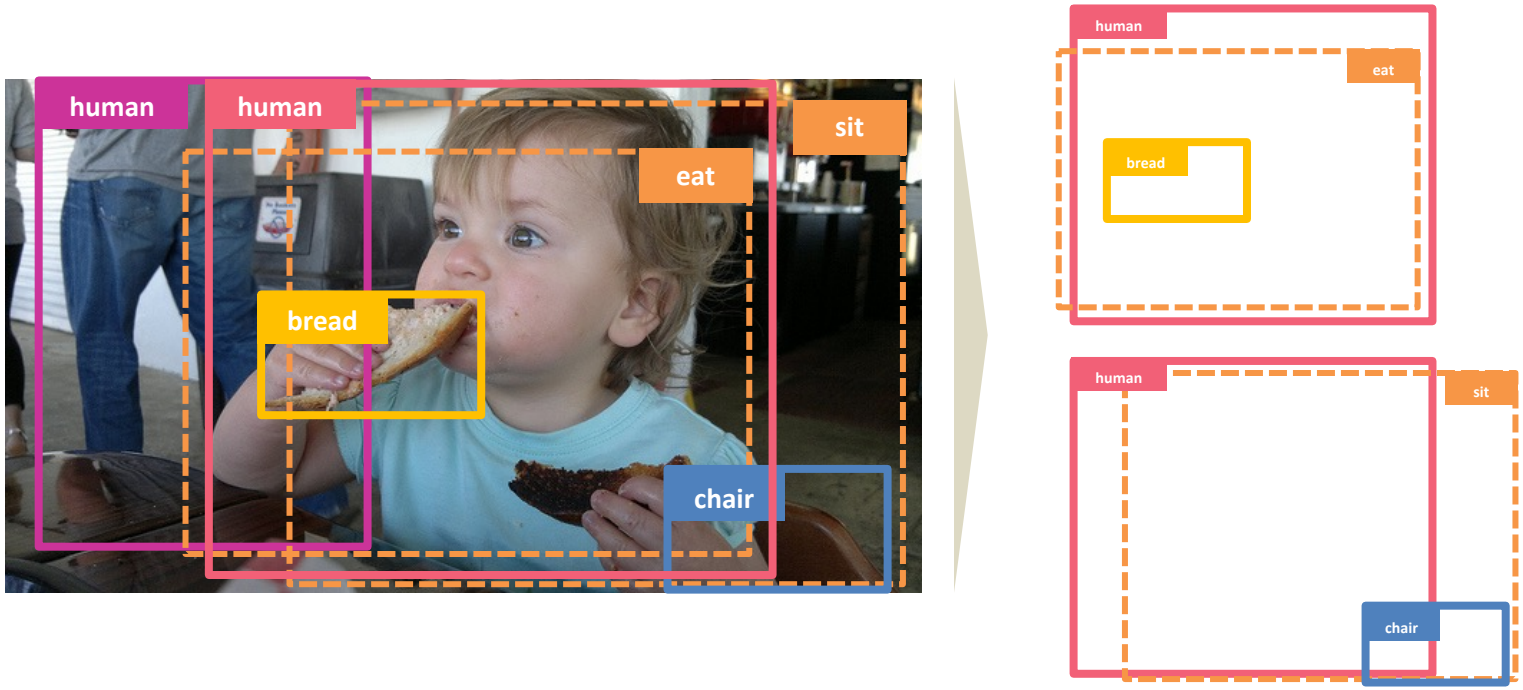
- Parallel HOI Detectors



[ECCV20] UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Previous approach in HOI detection

- Parallel HOI Detectors



[ECCV20] UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Previous approach in HOI detection

◦ Sequential HOI Detectors

- Intuitive Pipeline
- Pairwise Neural Network Inference : Slow

◦ Parallel HOI Detectors: UnionDet [ECCV20]

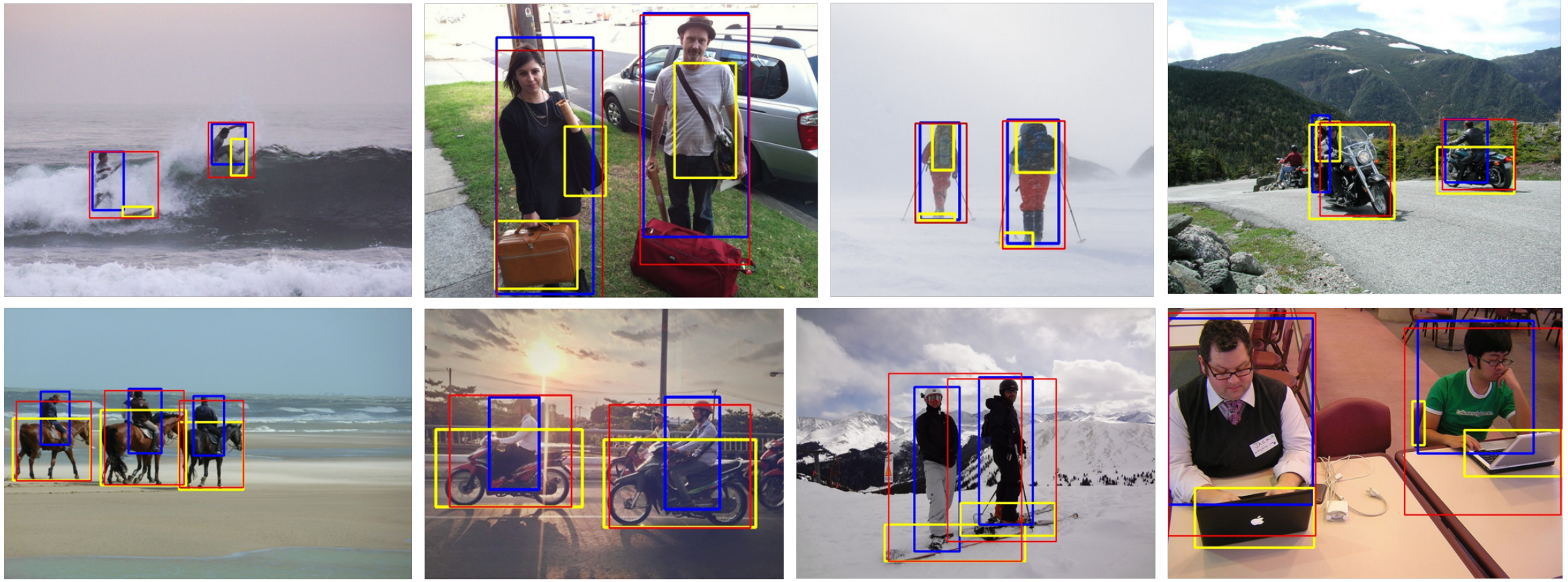
- Define “region of interaction” : Union / Interaction
- Speed-up in HOI inference time
- However, the triplet search is still a bottleneck and has room for improvement

UnionDet

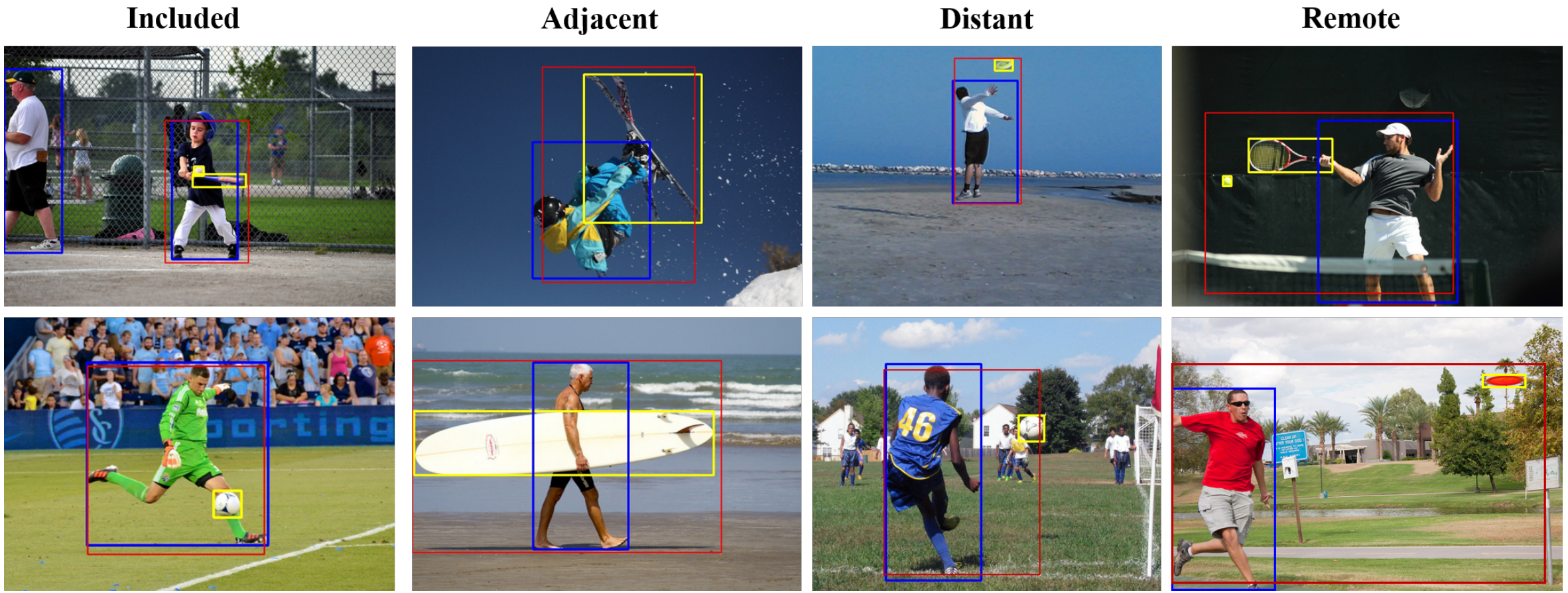
- **UnionDet**

- 탐지를 동시에 수행
- 물체 쌍 마다의 딥러닝 추론 제거

동일 관계 독립적 탐지 가능

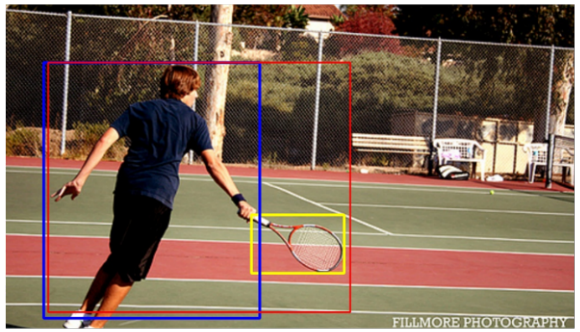


다양한 거리의 상호작용 탐지



다대다 상호작용 탐지

One vs One

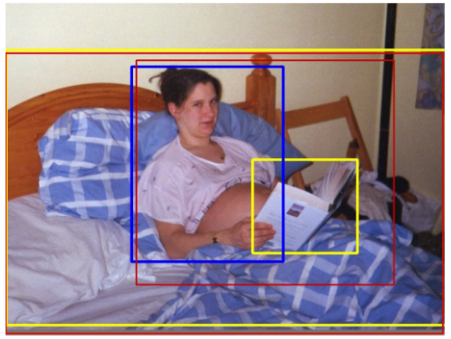


person – **hit instr** – tennis racket

Many vs One

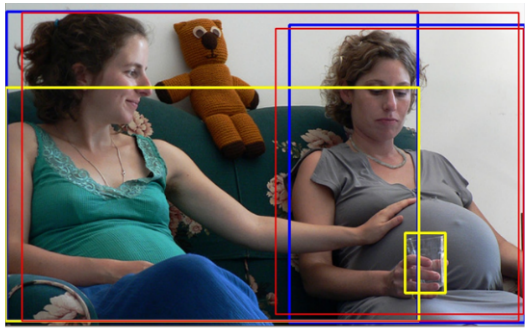
person (1) – **ride** – motorcycle
person (2) – **ride** – motorcycle

One vs Many

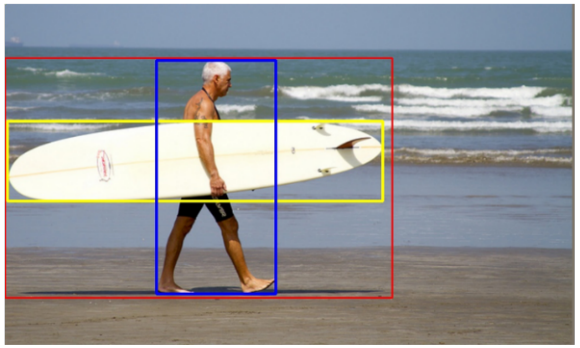


person – **read** – book
person – **lay** – bed

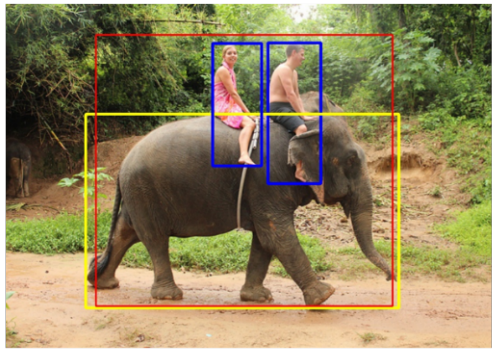
Many vs Many



person(1) – **sit** – couch
person(2) – **sit** – couch
person(2) – **hold** – cup

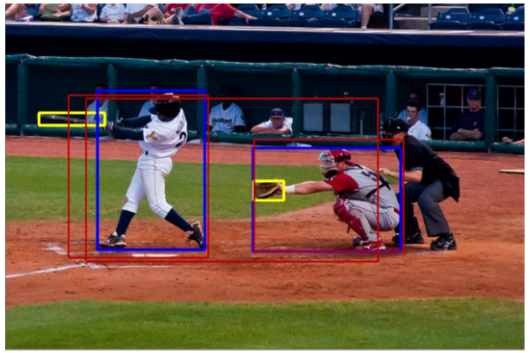


person – **carry** – surfboard



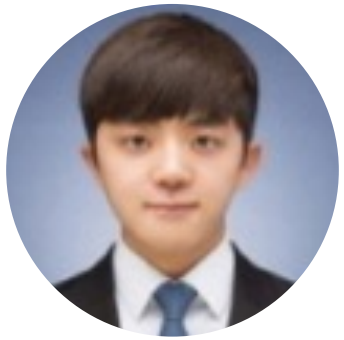
person (1) – **ride** – elephant
person (2) – **ride** – elephant

person – **hold** – umbrella
person – **read** – book



person(1) – **hit instr** – baseball bat
person(2) – **hold** – baseball glove
person(2) – **look** – person(1)

HOTR: End-to-End Human-Object Interaction Detection with Transformers



Bumsoo Kim
Kakao Brain
Korea Univ.

Junhyun Lee
Korea Univ.

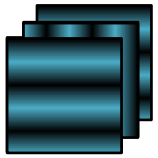
Jaewoo Kang
Korea Univ.

Eun-Sol Kim
Kakao Brain

Hyunwoo J. Kim
Korea Univ.

Architecture Overview

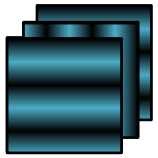
Convolutional Neural Network



Positional Encoding

Architecture Overview

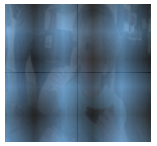
Convolutional Neural Network



Positional Encoding

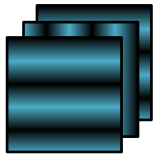
Architecture Overview

Convolutional Neural Network



Transformer Encoder

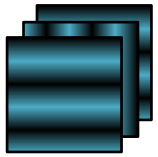
$\times L_E$



Positional Encoding

Architecture Overview

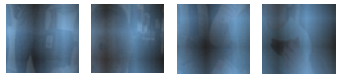
Convolutional Neural Network



Positional Encoding

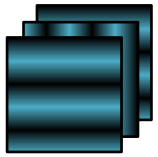
$\times L_E$

Transformer Encoder



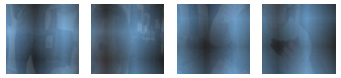
Architecture Overview

Convolutional Neural Network



Positional Encoding

Instance Queries



Transformer Encoder $\times L_E$

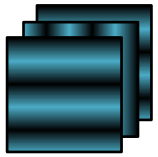
Instance Decoder $\times L_O$

Interaction Decoder $\times L_I$

Interaction Queries

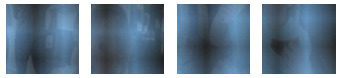
Architecture Overview

Convolutional Neural Network



Positional Encoding

Instance Queries



Transformer Encoder $\times L_E$

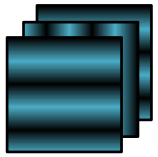
Instance Decoder $\times L_O$

Interaction Decoder $\times L_I$

Interaction Queries

Architecture Overview

Convolutional Neural Network

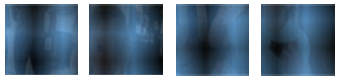


Positional Encoding

Instance Queries

$\times L_O$

Instance Decoder



$\times L_E$

Transformer Encoder

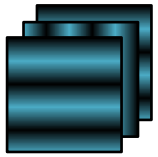
$\times L_I$

Interaction Decoder

Interaction Queries

Architecture Overview

Convolutional Neural Network



Positional Encoding

Transformer Encoder $\times L_E$

Instance Decoder $\times L_O$

FFN

1 class, box

2 class, box

3 class, box

4 class, box

Interaction Decoder $\times L_I$

FFN

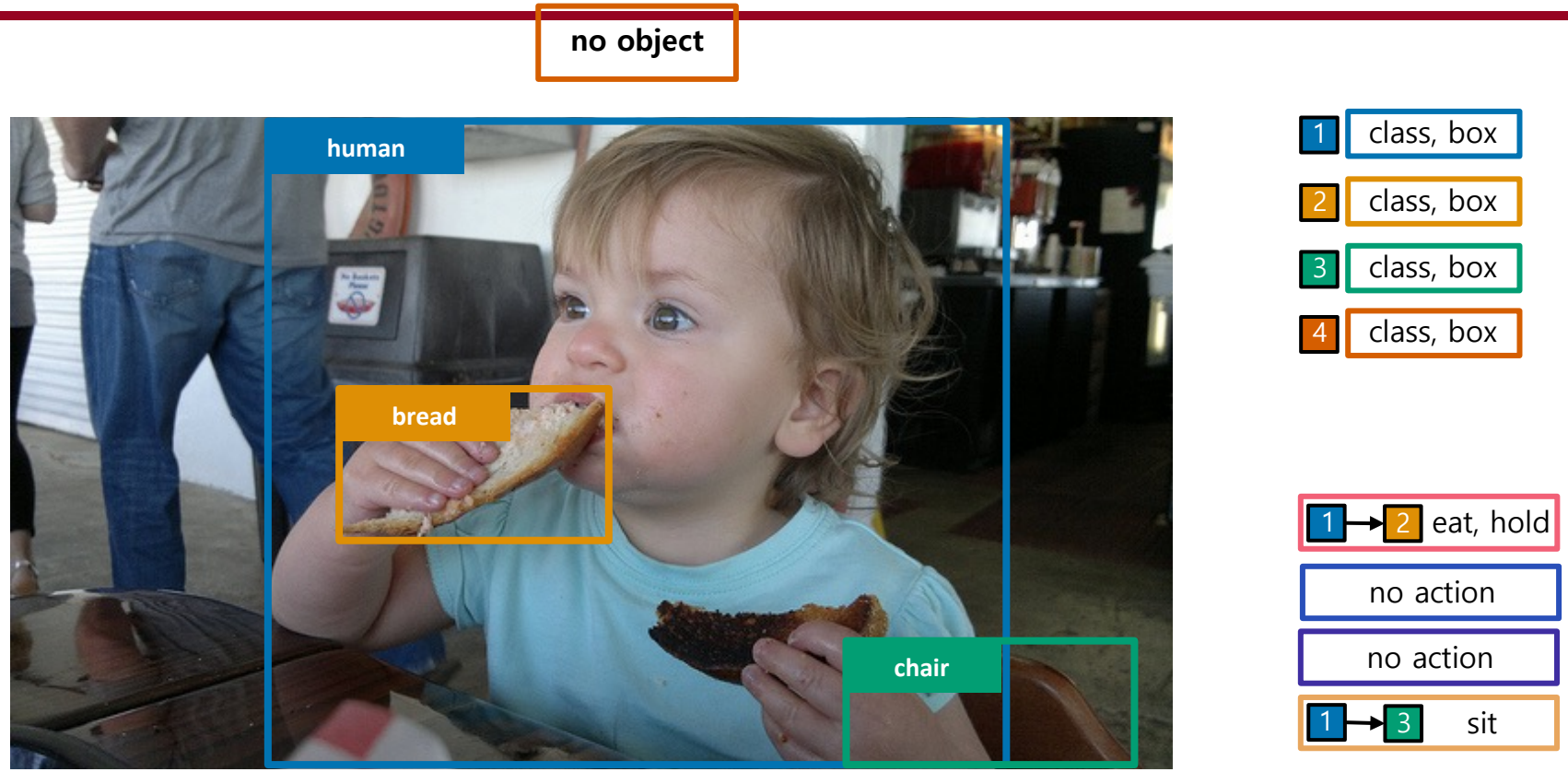
1 → 2 eat, hold

no action

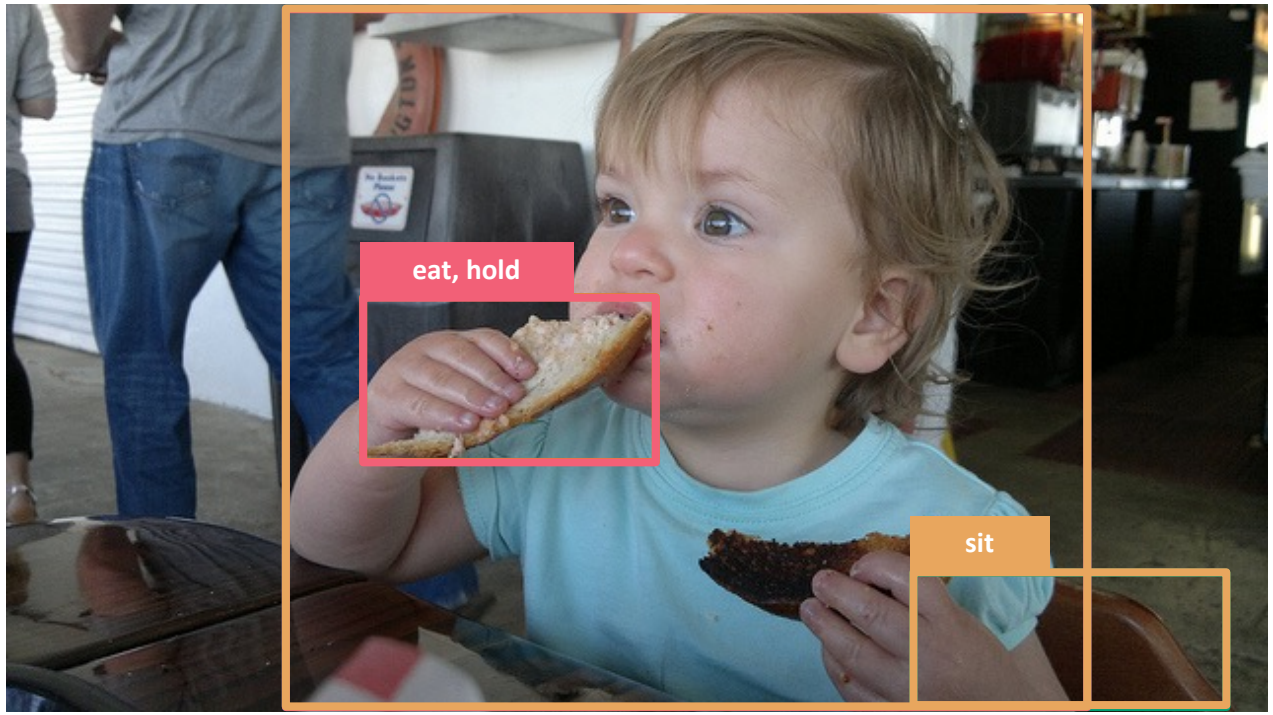
no action

1 → 3 sit

Architecture Overview



Architecture Overview



1 class, box

2 class, box

3 class, box

4 class, box

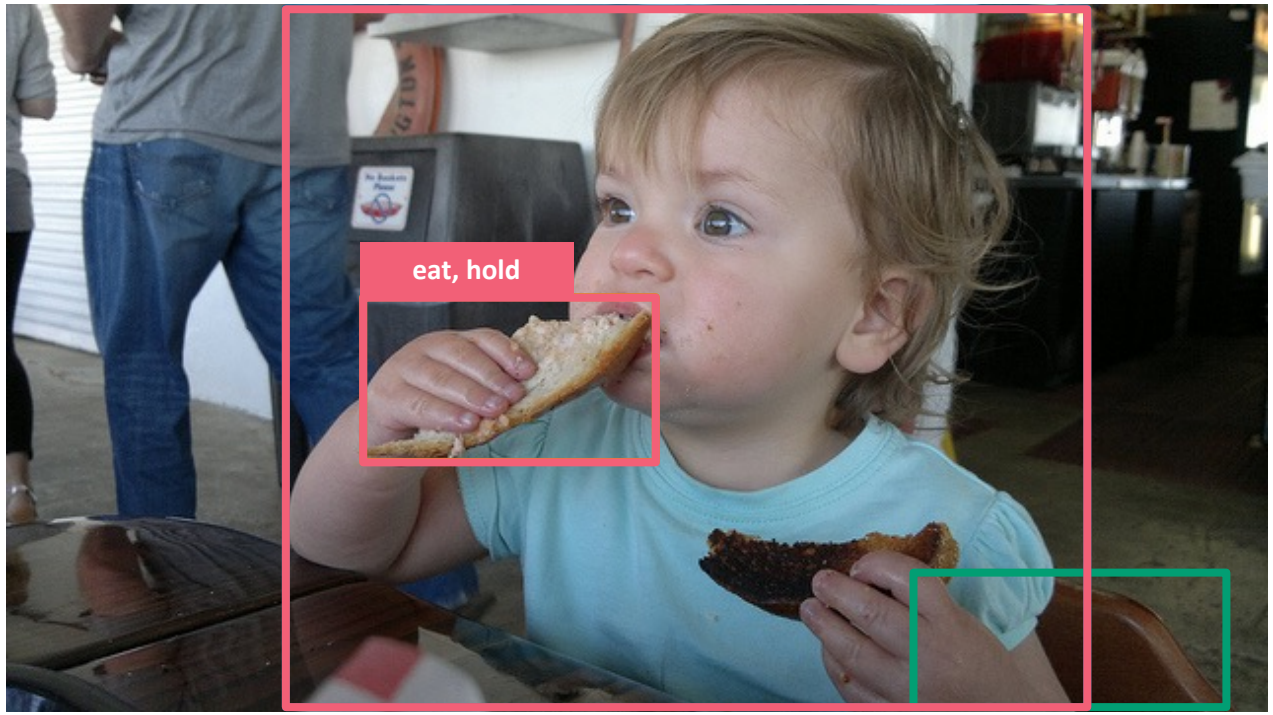
1 → 2 eat, hold

no action

no action

1 → 3 sit

Architecture Overview



1 class, box

2 class, box

3 class, box

4 class, box

1 → 2 eat, hold

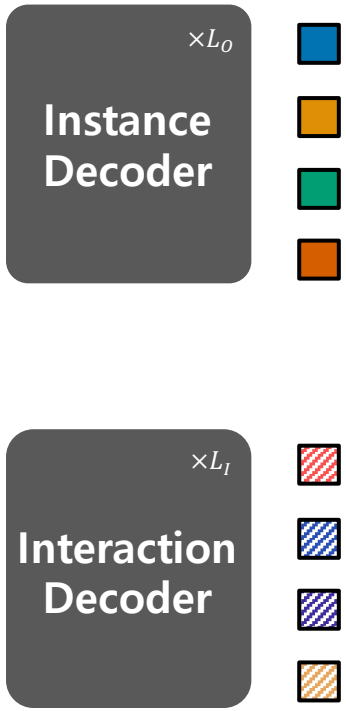
no action

no action

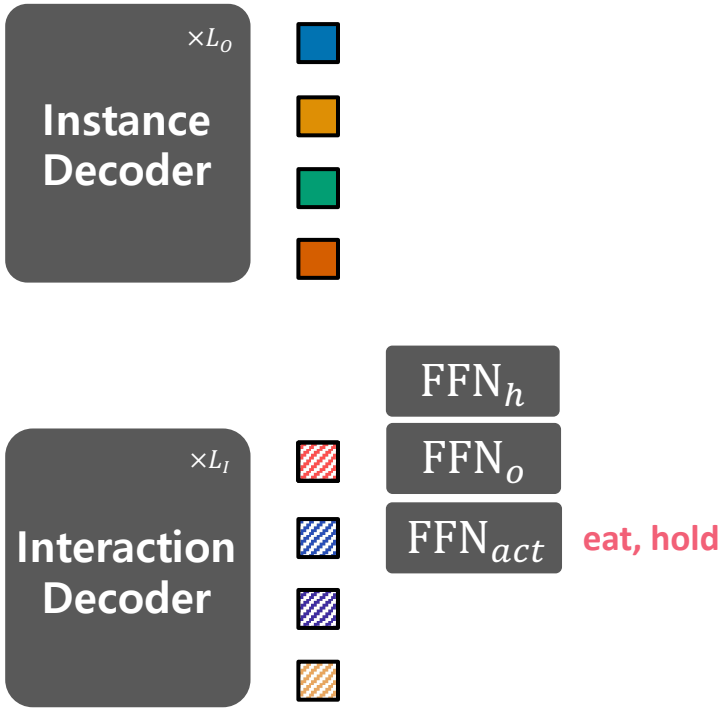
1 → 3 sit

HO Pointers

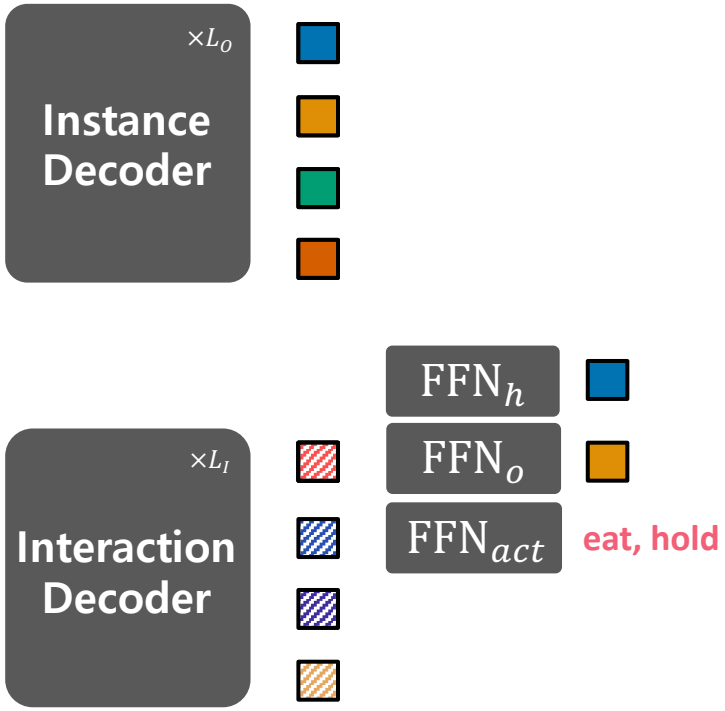
HO Pointers



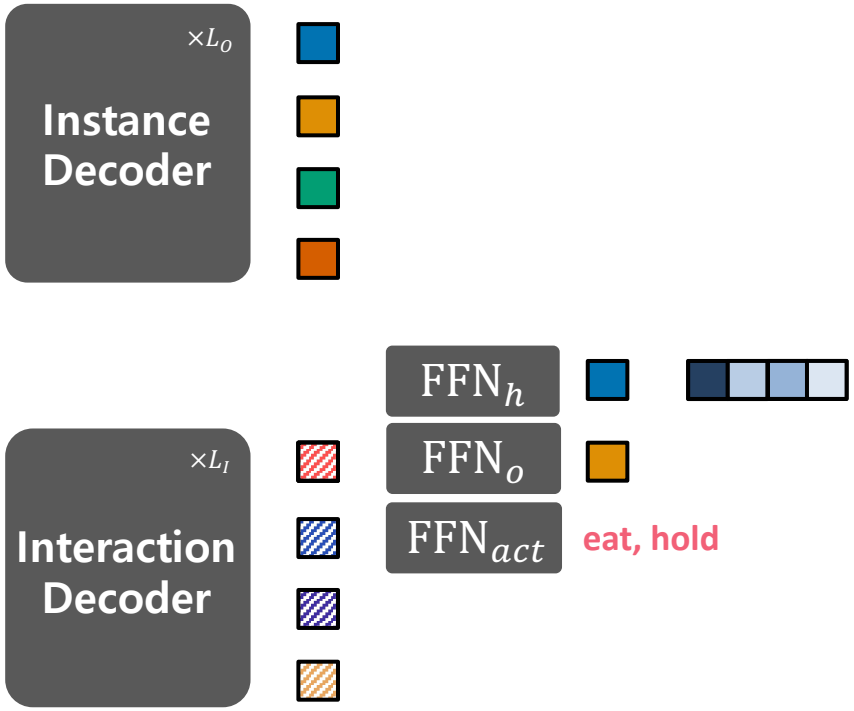
HO Pointers


$\times L_O$
Instance Decoder
$\times L_I$
Interaction Decoder
FFN_h
FFN_o
FFN_{act}
eat, hold

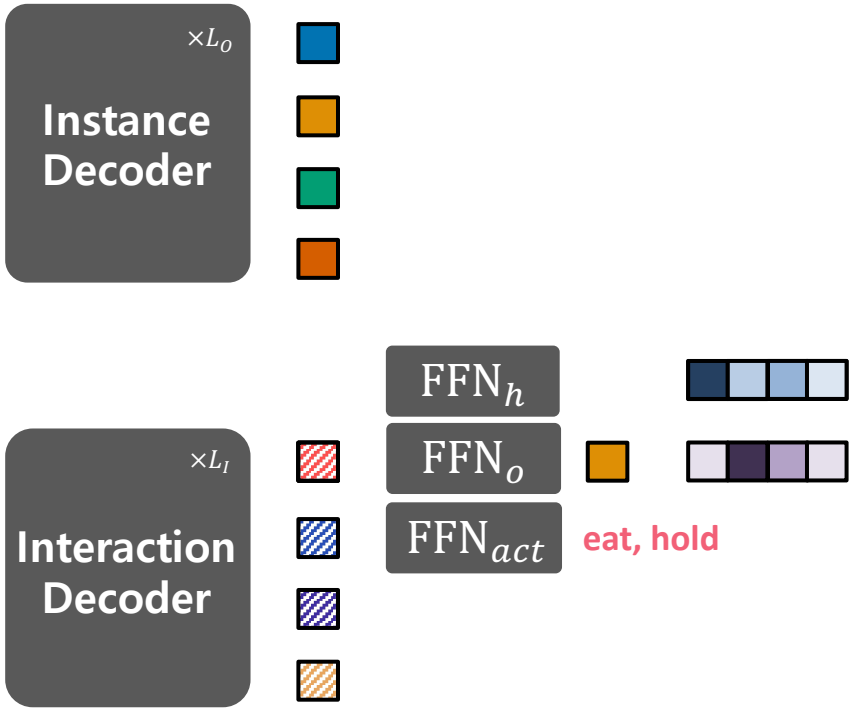
HO Pointers


$\times L_O$
Instance Decoder
$\times L_I$
Interaction Decoder
FFN_h
FFN_o
FFN_{act}
eat, hold

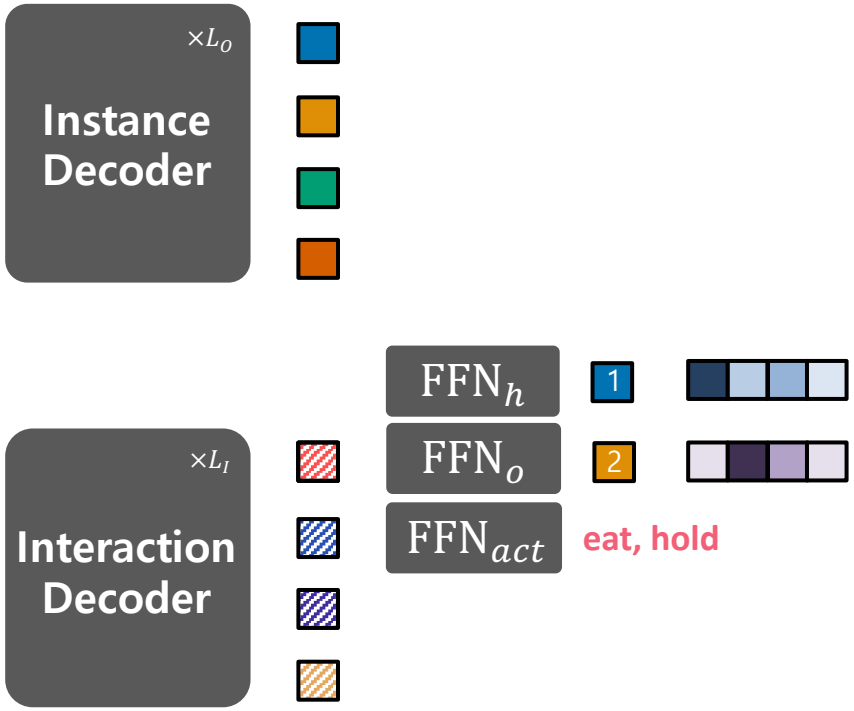
HO Pointers


$\times L_O$
Instance Decoder
$\times L_I$
Interaction Decoder
FFN_h
FFN_o
FFN_{act}
eat, hold

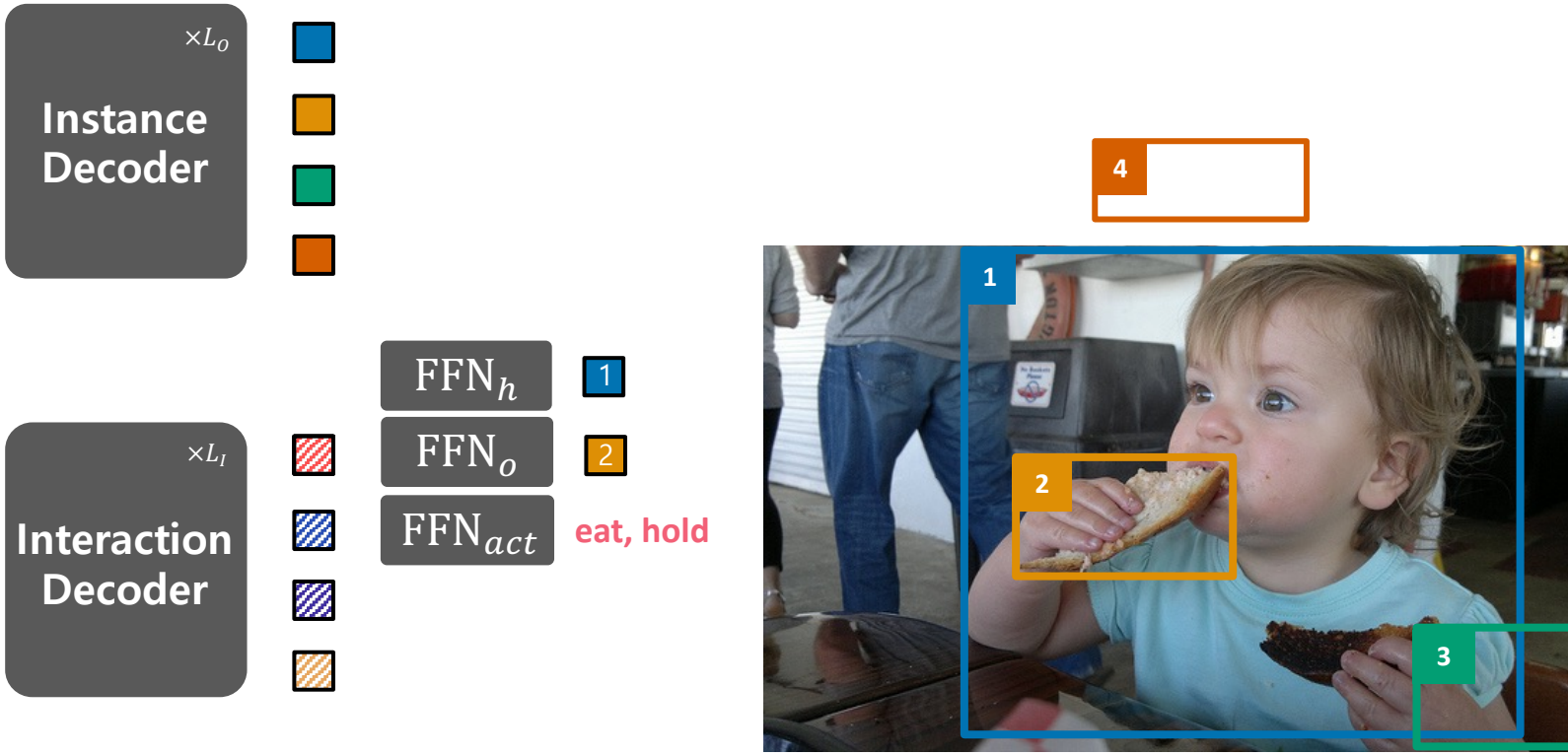
HO Pointers


$\times L_O$
Instance Decoder
$\times L_I$
Interaction Decoder
FFN_h
FFN_o
FFN_{act}
eat, hold

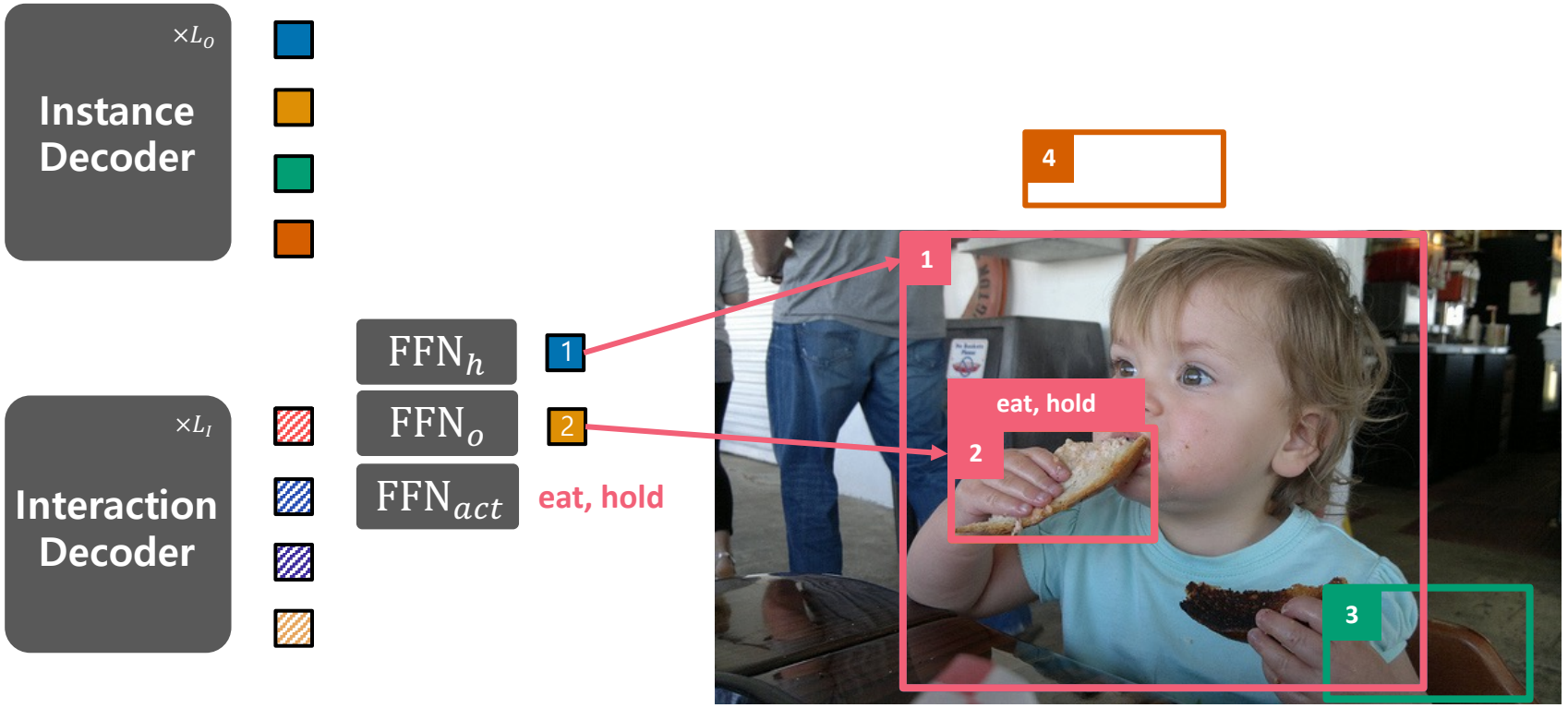
HO Pointers


$\times L_O$
Instance Decoder
$\times L_I$
Interaction Decoder
FFN_h
FFN_o
FFN_{act}
1
2
eat, hold

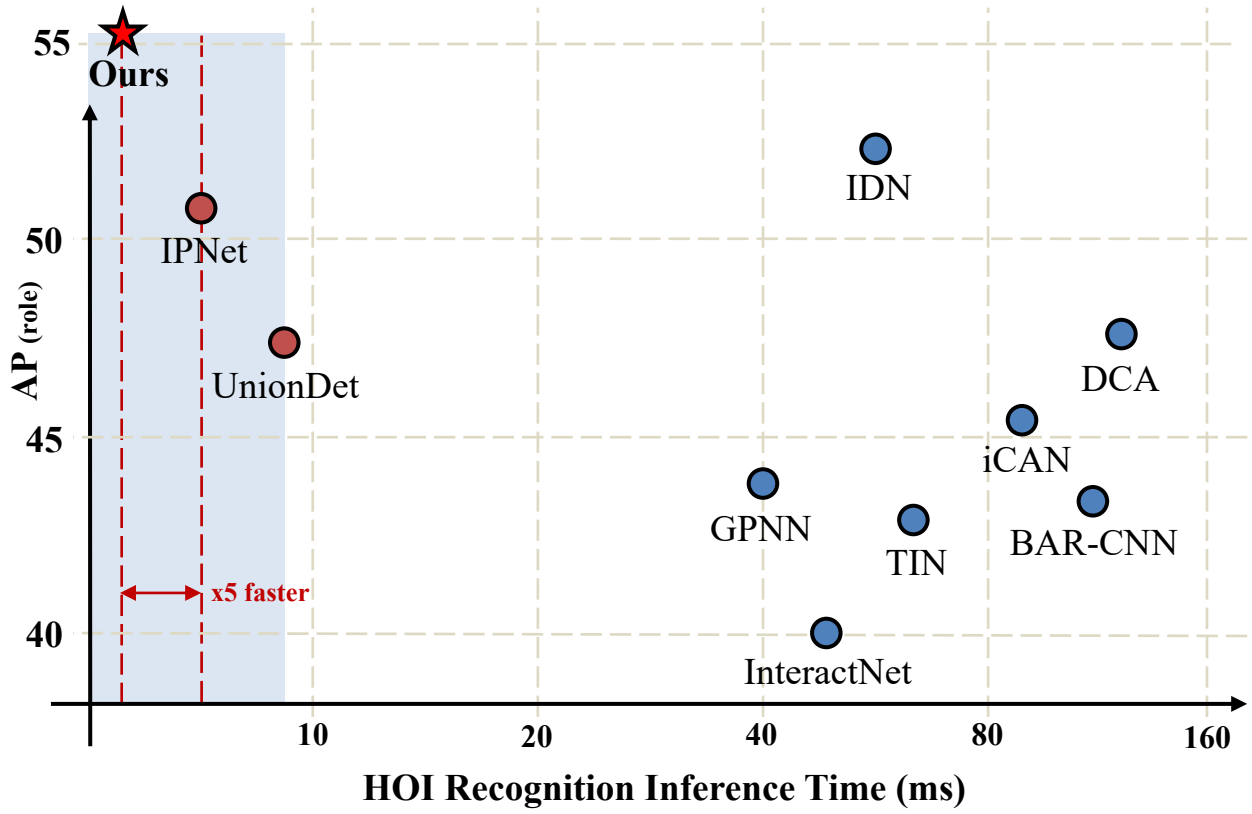
HO Pointers

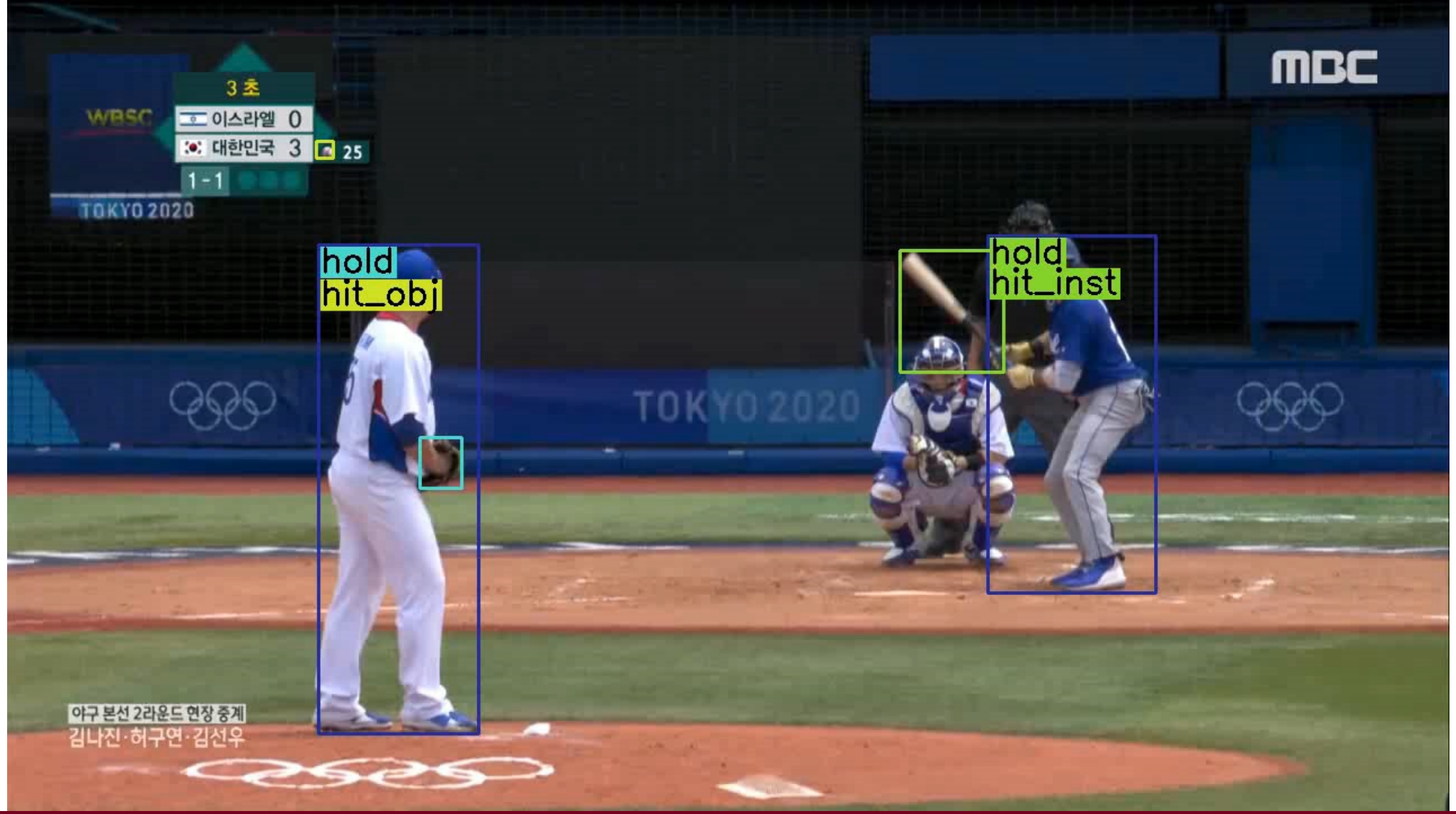

×L_O
Instance Decoder
×L_I
Interaction Decoder
FFN_h
FFN_o
FFN_{act}
1
2
eat, hold
4
1
2
3

HO Pointers


×L_O
Instance Decoder
×L_I
Interaction Decoder
FFN_h
FFN_o
FFN_{act}
eat, hold
1
2
4
1
eat, hold
2
3

Experimental Results (HOI Inference Time)




3 초
WBSC
이스라엘 0
대한민국 3
25
1 - 1
TOKYO 2020
MBC
hold
hit_obj
hold
hit_inst
TOKYO 2020
야구 본선 2라운드 현장 중계
김나진·허구연·김선우

- **Ride Giraffe**

BCE

Focal+CB

참고 문헌

- [CVPR 21 (ORAL)] HOTR: End-to-End Human-Object Interaction Detection with Transformers
- [CVPRW 19] ANTNets: Mobile Convolutional Neural Networks for Resource Efficient Image Classification
- [CVPR 18] Tensorize, Factorize and Regularize: Robust Visual Relationship Learning
- [ECCV 20] UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Questions?