# 인공지능은 물체를 어떻게 이해할까?

## 소주제 3: 물체간 관계 추론

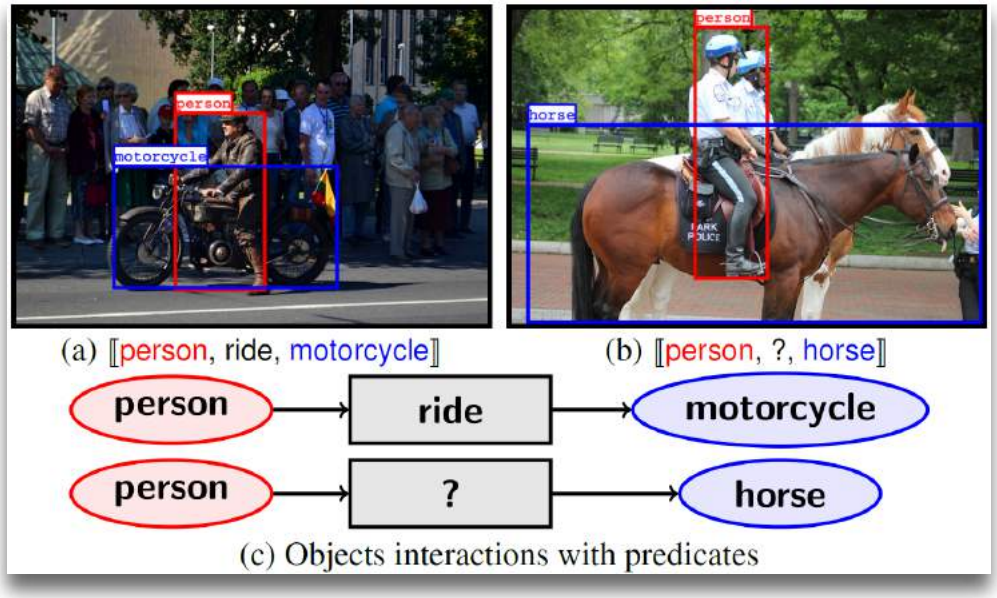
김현우 (hyunwoojkim@korea.ac.kr)

(정보대학 컴퓨터학과 기계학습 및 비전 연구실)

# 인공지능은 물체를 어떻게 이해할까?

- ~~소주제 1: 물체 탐지 원리~~
- ~~소주제 2: 물체 탐지 기법~~
- 소주제 3: 물체간 관계 추론

# 물체간 관계 추론

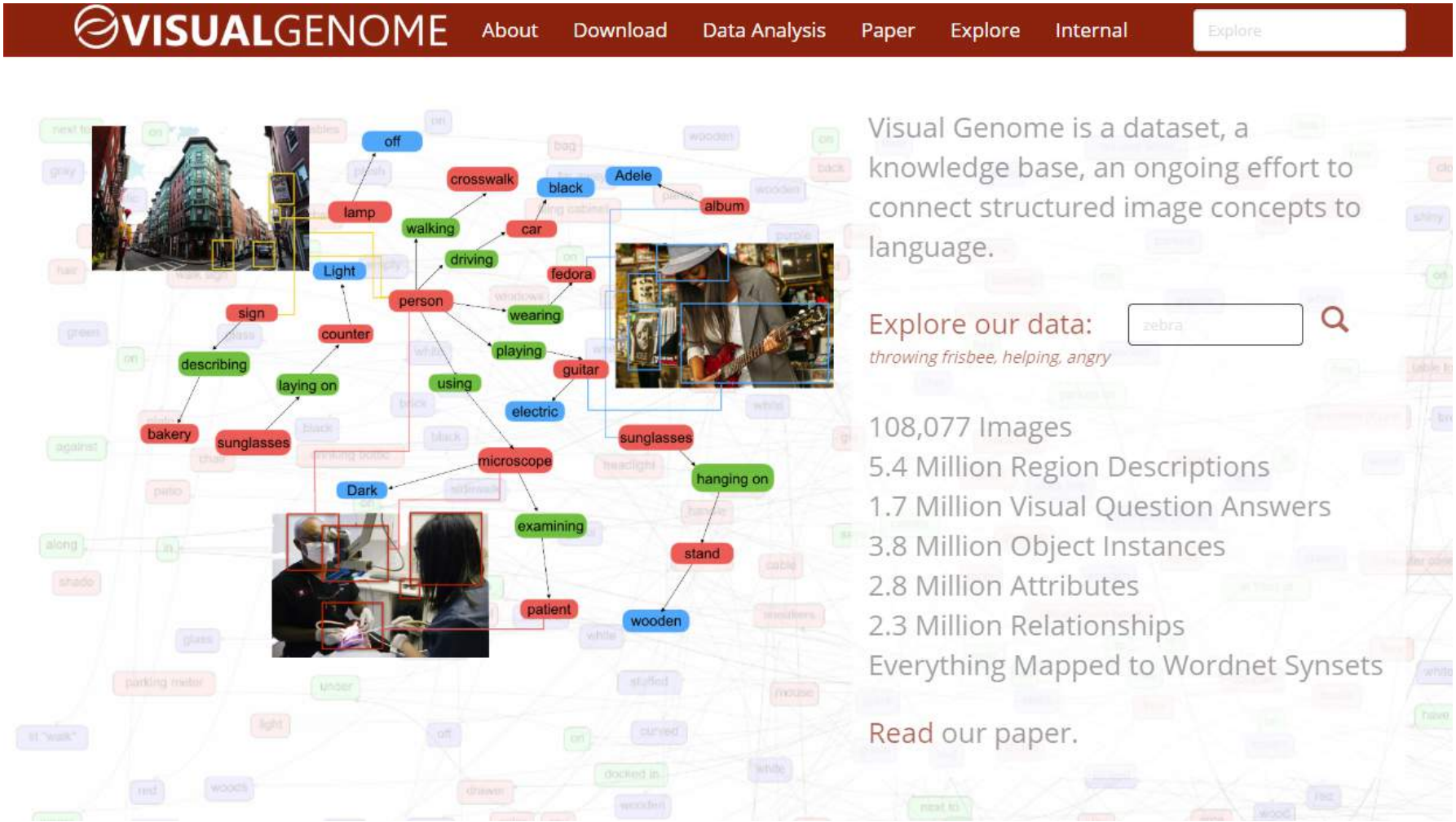



- **태스크 정의:**

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

- Goal:
- **관계 추론** < **물체1**, **관계** , **물체2** >

# 비주얼 게놈 (Visual Genome)



VISUALGENOME About Download Data Analysis Paper Explore Internal

Explore our data: tennis

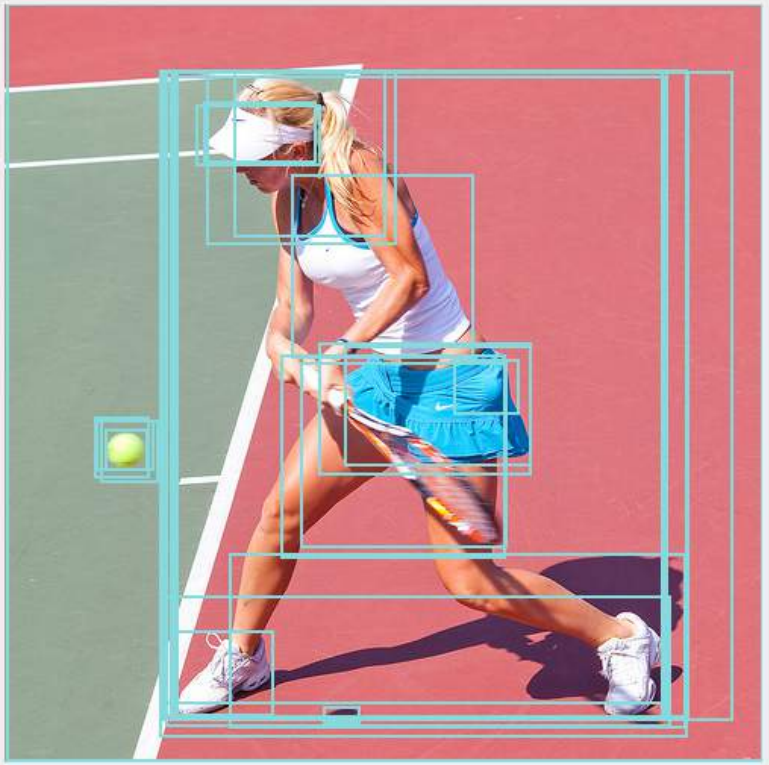
Try one of these: throwing Frisbee, helping or angry.

tennis ball is round
tennis ball is yellow
tennis ball is bright green

next Last

Tennis ball in pocket.
tennis ball is in mid air

| Regions | Attributes | Relationships |
|---|---|---|
| Woman is playing tennis. | woman is playing tennis | skort |
| Woman is holding tennis racquet. | skort is blue | woman WEARING tennis shoe |
| Woman is wearing blue skort. | tennis ball is in mid air | woman WEARING visor |
| Tennis ball in pocket. | tennis ball is yellow | woman swinging tennis racquet |
| Tennis ball in mid air. | tennis ball is round | woman WEARING shirt |
| Tennis ball is yellow. | visor is white | woman has hair |
| Tennis ball is round. | hair is woman's | shadow ON tennis court |
| | hair is in ponytail | shadow OF tennis ball |
| | woman is swinging | |



Question Answers

| | |
|---|---|
| How many people on the court? | One. |
| What color is the woman's shorts? | Blue. |
| What color are the woman's shoes? | White. |
| Who is on the court? | A woman. |
| Why is the woman on the court? | Playing tennis. |

https://visualgenome.org/VGViz/explore?query=tennis

Regions

- white letter on television
- closed caption on television
- cat with a brown coat
- fox symbol on TV program
- cord draped down the wall
- part of wood trim around window
- controls for the television

Attributes

- subtitles is spanish
- flash is camera
- television is old
- tv set is black
- cat is gray
- shelf is small
- shelf is black
- window trim is white

Relationships

- woman ON screen
- cat ON tv
- cat lying on tv


Question Answers

| | |
|---|---|
| What animal is in the picture? | A feline. |
| What is the cat doing? | Laying down. |
| What is under the cat? | A tv. |
| What color is the cloth behind the cat? | Black. |
| Who is on top the tv? | A cat. |

Regions

- white letter on television
- closed caption on television
- cat with a brown coat
- fox symbol on TV program
- cord draped down the wall
- part of wood trim around window
- controls for the television

Attributes

- subtitles is spanish
- flash is camera
- television is old
- tv set is black
- cat is gray
- shelf is small
- shelf is black
- window trim is white

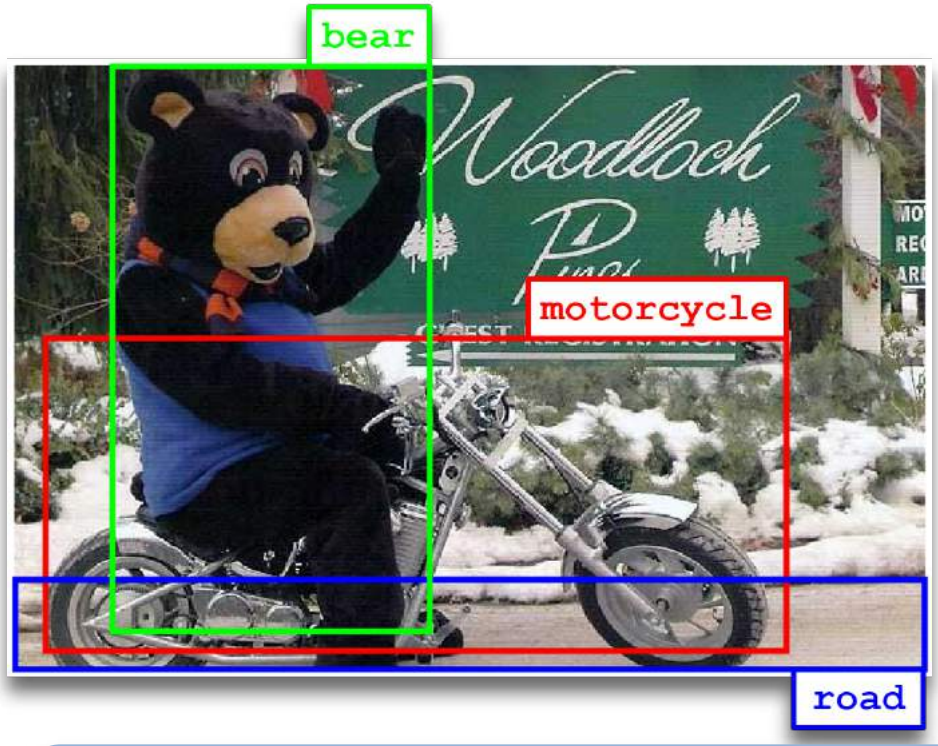
Relationships

- woman ON screen
- cat ON tv
- cat lying on tv


Question Answers

| | |
|---|---|
| What animal is in the picture? | A feline. |
| What is the cat doing? | Laying down. |
| What is under the cat? | A tv. |
| What color is the cloth behind the cat? | Black. |
| Who is on top the tv? | A cat. |

# 장면 그래프 생성 도전성



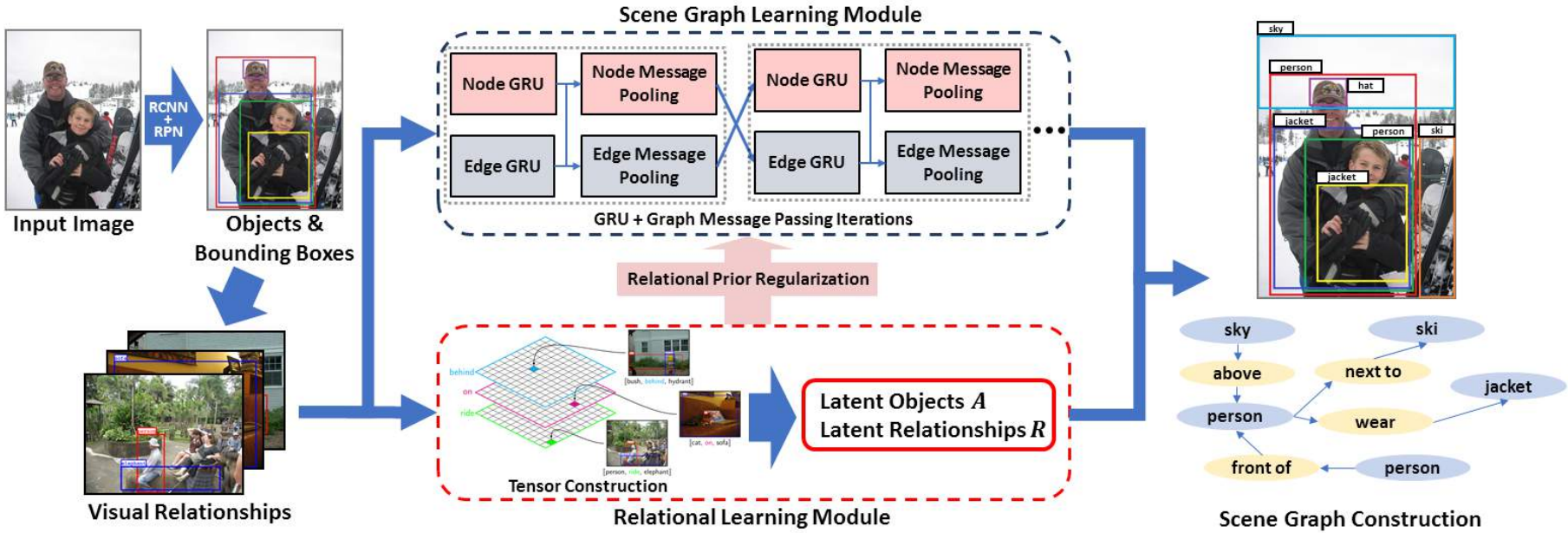

- **의미 추론** 의 어려움
- 데이터 의존성
- 물체관계의 다양성 및 큰 관계공간

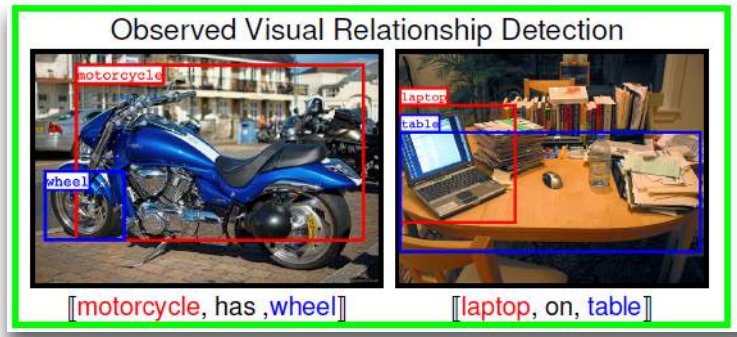
  **< 물체1, 관계 , 물체2 >**

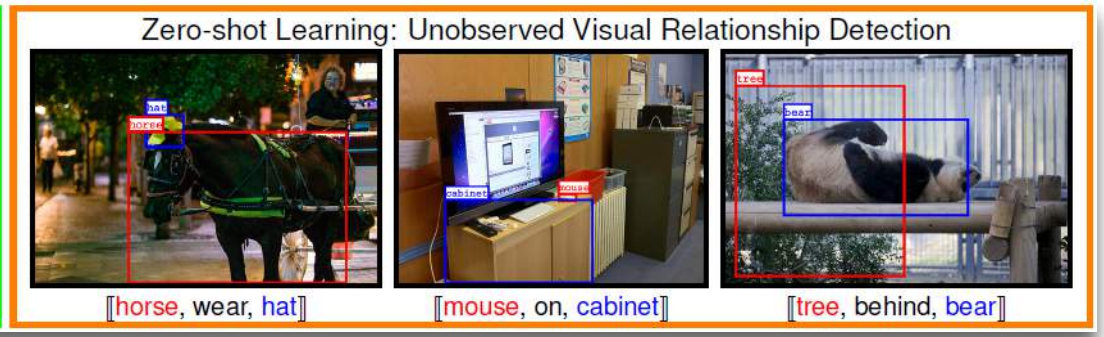
  100 x 100 x 100 ~ 1M
- 제한된 데이터 (Visual Genome)
  - >1M 이미지
  - **가능한 조합의 2% 만 커버**
- **제로샷 학습 (Zero-shot Learning):**
  한번도 본적 없는 관계 예측 필요

**해결 방안: 다중관계 텐서 분해. 관계에 대한 사전 지식 생성/습득**

# 장면 그래프 생성 과정

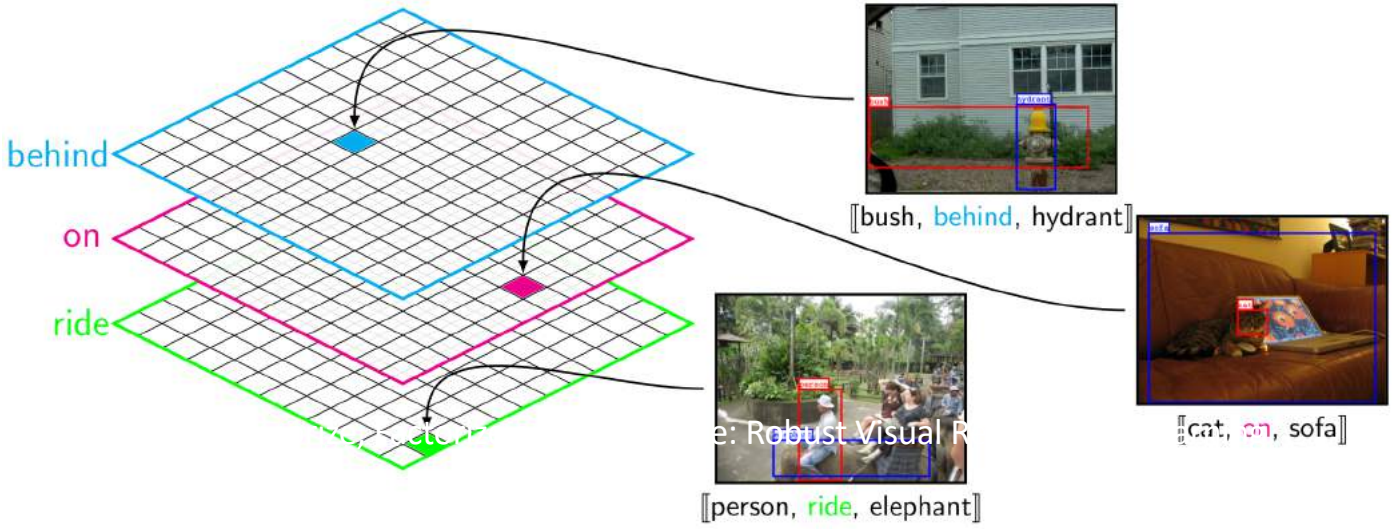

Input Image
RCNN + RPN
Objects & Bounding Boxes
Visual Relationships
Scene Graph Learning Module
Node GRU
Node Message Pooling
Edge GRU
Edge Message Pooling
GRU + Graph Message Passing Iterations
Relational Prior Regularization
Tensor Construction
Latent Objects $A$
Latent Relationships $R$
Relational Learning Module
sky
person
hat
jacket
ski
Scene Graph Construction


Observed Visual Relationship Detection
motorcycle
wheel
laptop
table
[motorcycle, has ,wheel]
[laptop, on, table]


Zero-shot Learning: Unobserved Visual Relationship Detection
hat
horse
cabinet
mouse
tree
bear
[horse, wear, hat]
[mouse, on, cabinet]
[tree, behind, bear]

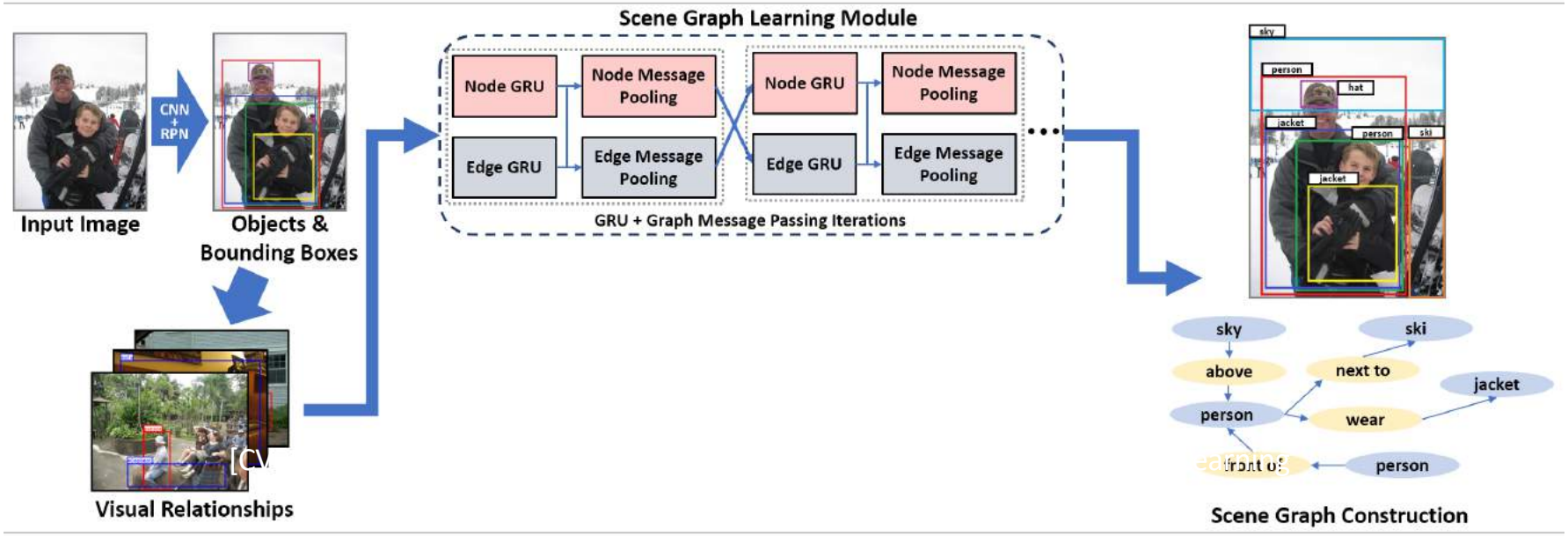
# 물체 관계 텐서

- 물체 관계 텐서: < 물체1, 관계 , 물체2 >
- 확률적 높은 관계 추정
- 텐서는 매우 희소함
- 텐서 분해 (tensor factorization)을 통해 저랭크 추정(low-rank estimation)을 수행



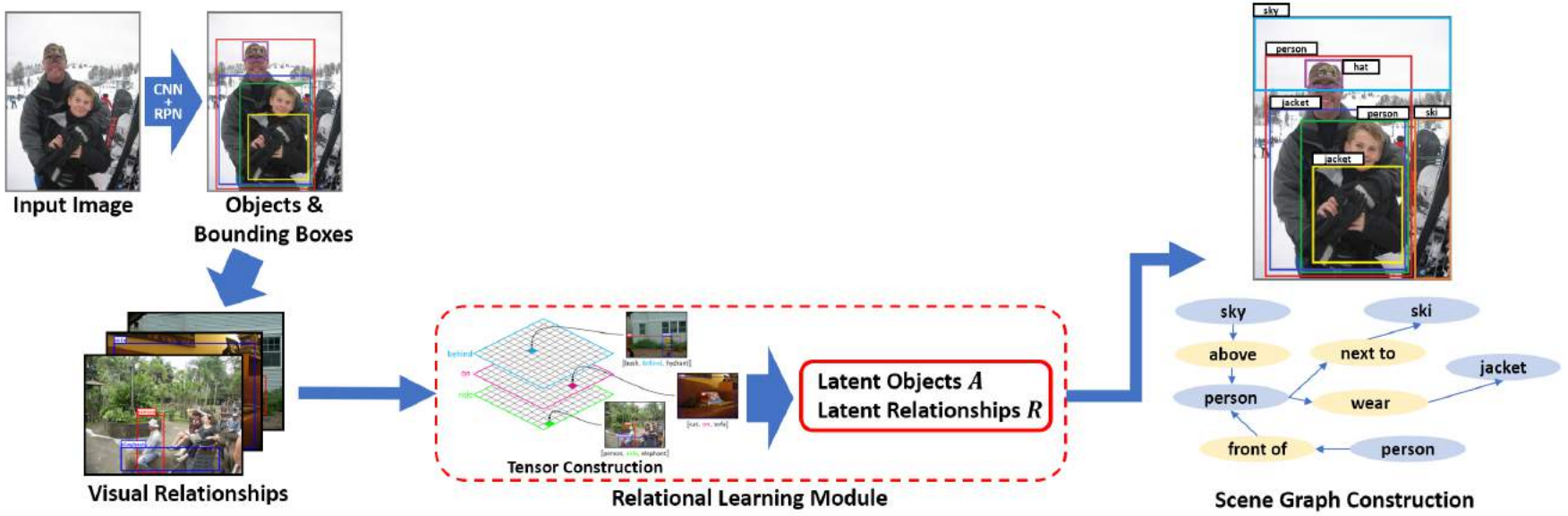
# 모듈 1: 장면 그래프 생성 (Xu et al.)

1. Faster R-CNN: 물체를 탐지하고 탐지된 물체 영역에서 특성값을 추출
2. 초안 그래프 생성: 탐지된 물체를 노드 물체 사이의 관계를 에지로 생성함
3. 메세지 패싱 네트워크 (GNNs): 그래프 인공신경망을 이용한 추론



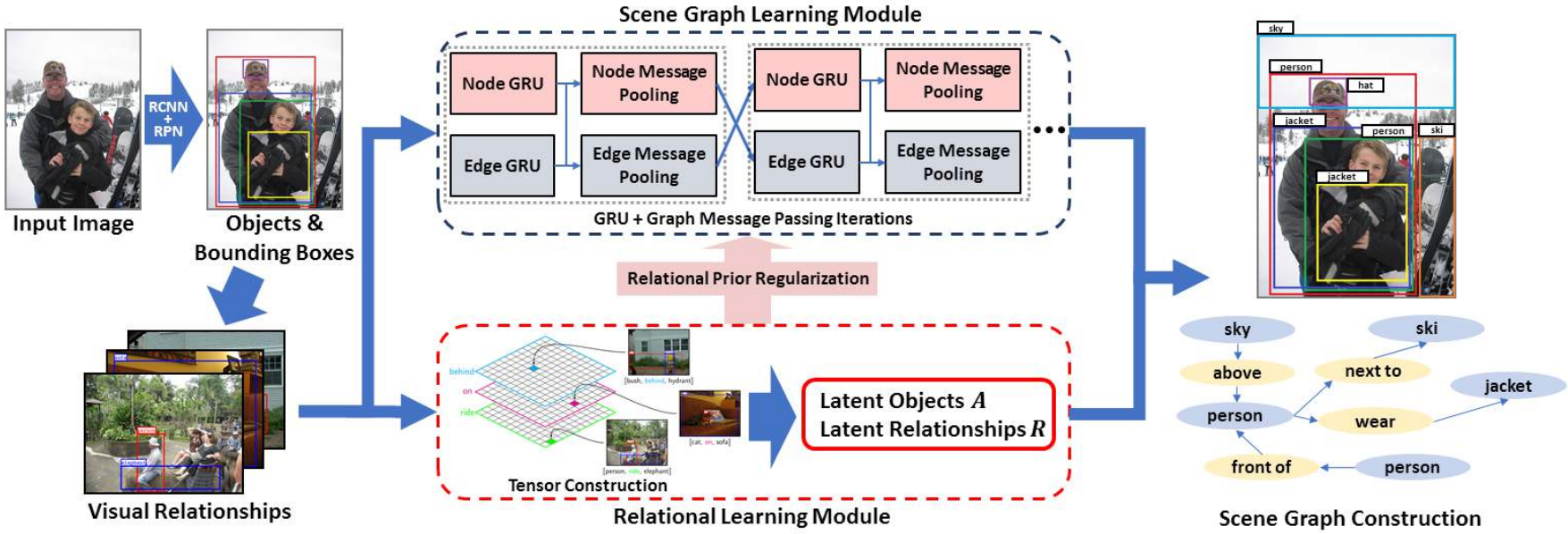
# 모듈 2: 물체 관계 업데이트 (Ours)

1. **Faster R-CNN**: features of object bounding boxes (**nodes**) and intersecting bounding boxes (**edges**).
2. **Tensor factorization:** Using the detected labels, predict predicates



# 최종 관계 예측

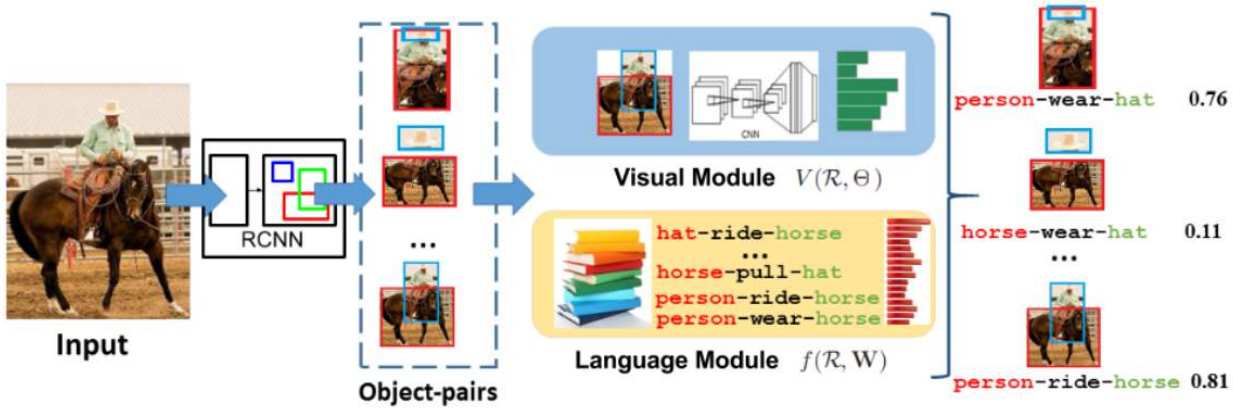
1. Faster R-CNN: 물체 탐지, 노드 및 에지 생성
2. Tensor factorization: 물체 관계 사전 지식 활용

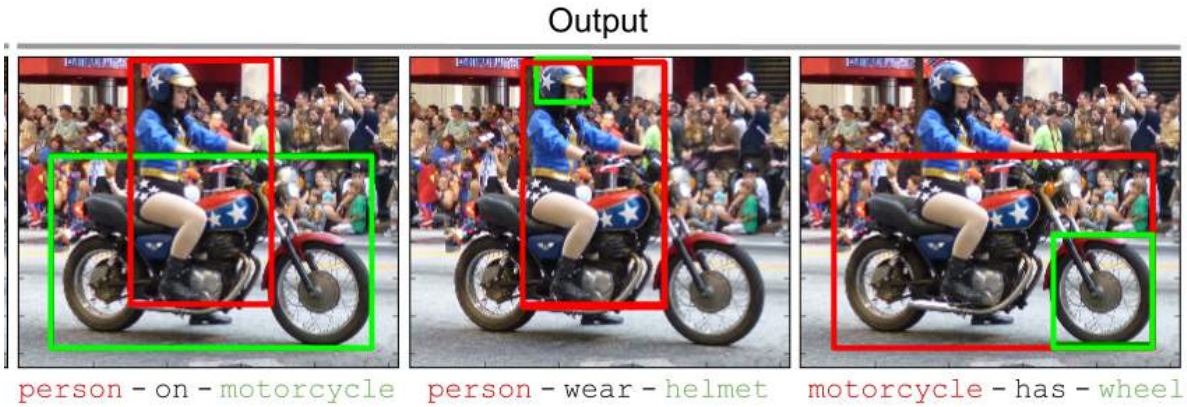


## Visual Relationship Detection with Language Priors


Abstract

Visual relationships capture a wide variety of interactions between pairs of objects in images (e.g. "man riding bicycle" and "man pushing bicycle"). Consequently, the set of possible relationships is extremely large and it is difficult to obtain sufficient training examples for all possible relationships. Because of this limitation, previous work on visual relationship detection has concentrated on predicting only a handful of relationships. Though most relationships are infrequent, their objects (e.g. "man" and "bicycle") and predicates (e.g. "riding" and "pushing") independently occur more frequently. We propose a model that uses this insight to train visual models for objects and predicates individually and later combines them together to predict multiple relationships per image. We improve on prior work by leveraging language priors from semantic word embeddings to finetune the likelihood of a predicted relationship. Our model can scale to predict thousands of types of relationships from a few examples. Additionally, we localize the objects in the predicted relationships as bounding boxes in the image. We further demonstrate that understanding relationships can improve content based image retrieval.

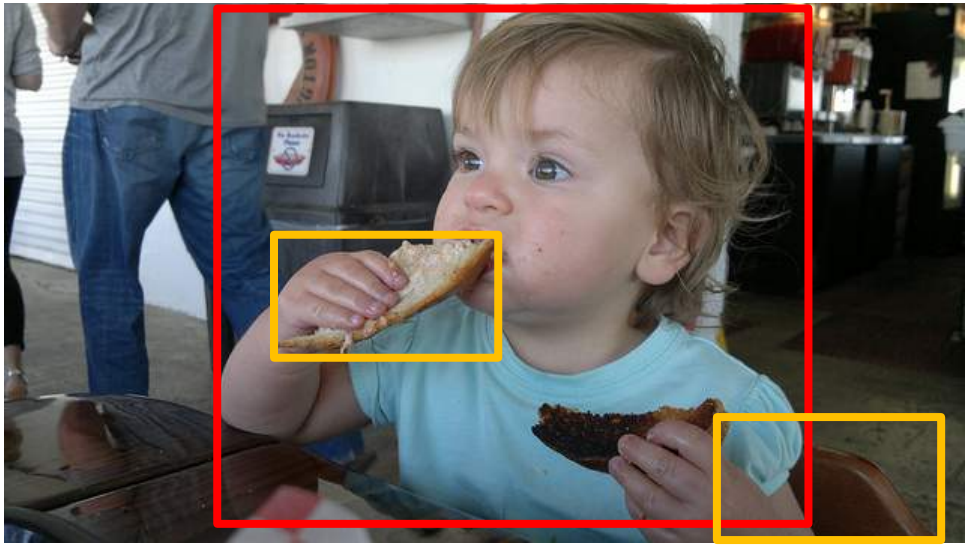




https://cs.stanford.edu/people/ranjaykrishna/vrd/

**Dataset:**

**5000** images
**37,993** relationships.
**100 object** classes
**70 predicate (relationship)** classes

# 사람-사물- 관계 탐지 (HOI Detection)



There is a human

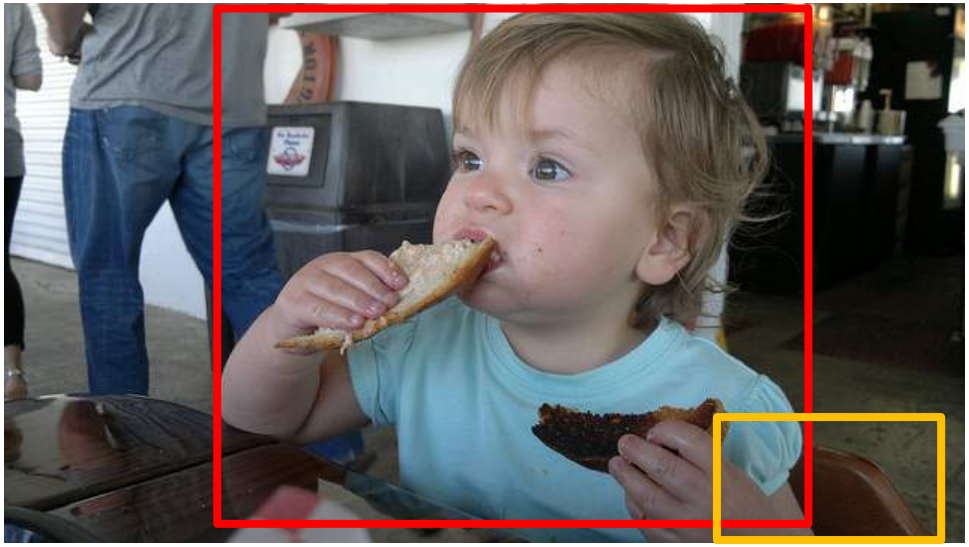
There is a bread

There is a chair

# 사람-사물- 관계 탐지 (HOI Detection)

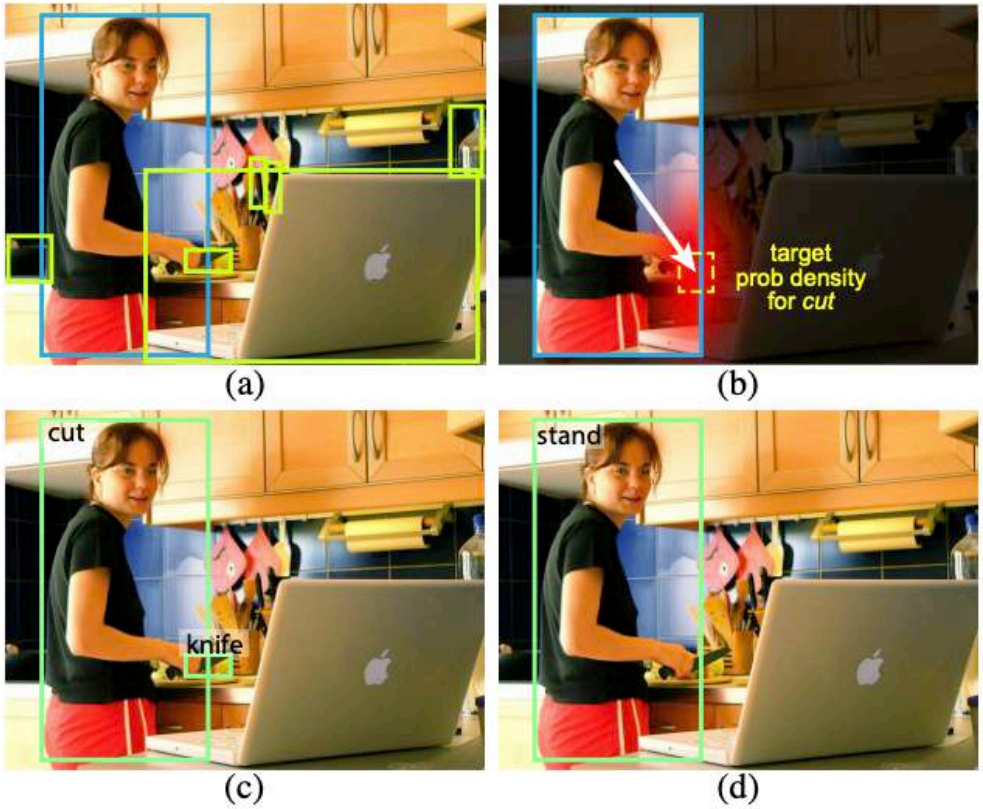
- Human is eating the bread

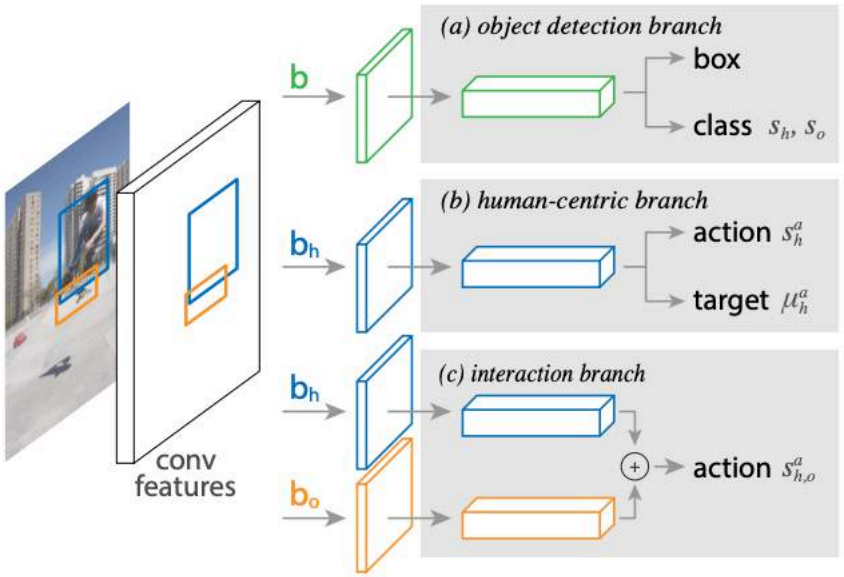
# 사람-사물- 관계 탐지 (HOI Detection)



- Human is eating the bread
- Human is sitting on the chair

# InteractNet

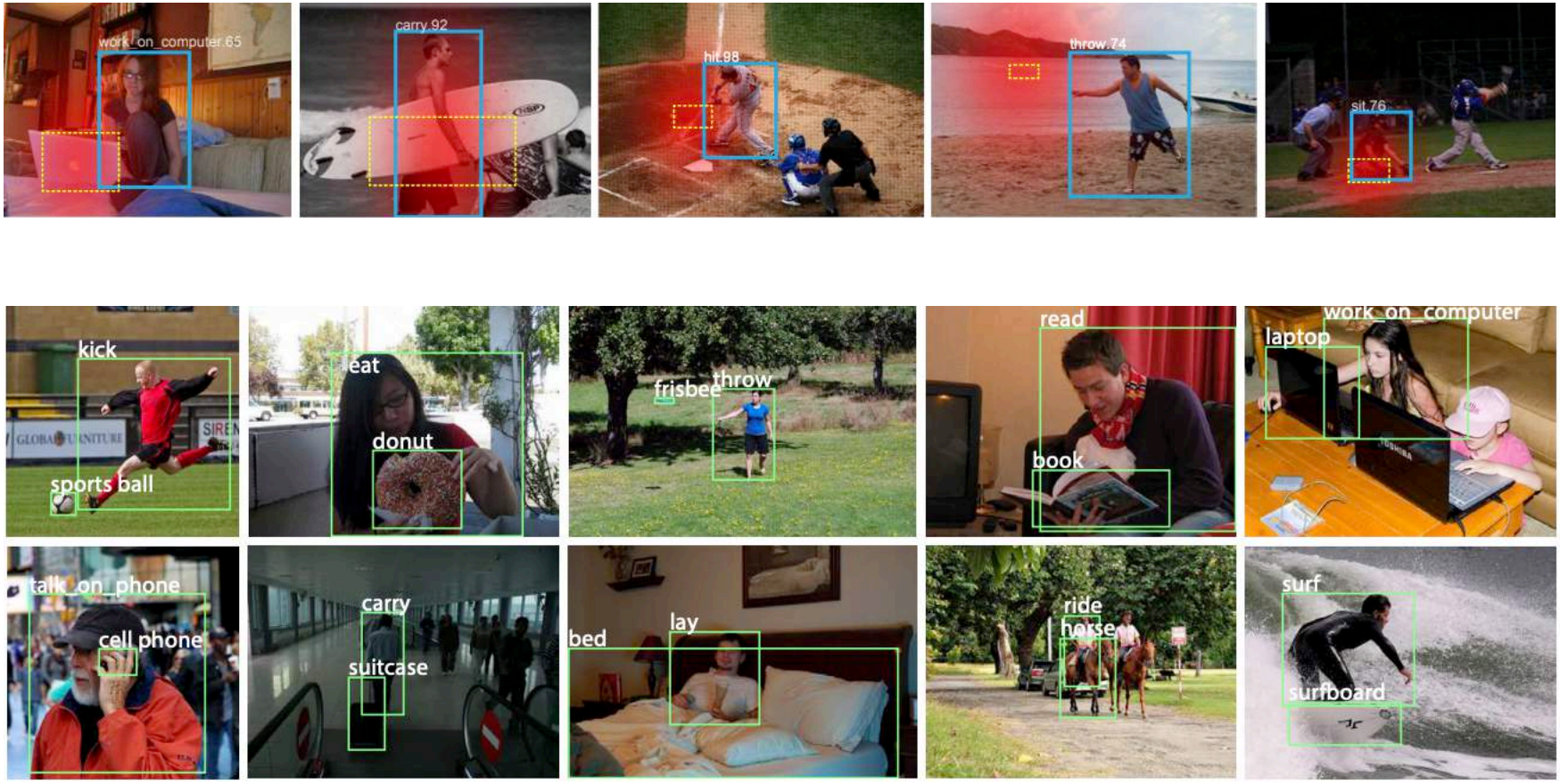



Facebook AI Research (FAIR)

Gkioxari, Georgia, et al. "Detecting and recognizing human-object interactions." *CVPR* 2018.

# InteractNet




< Human, Verb, Object >

Facebook AI Research (FAIR)

ECCV'20 ONLINE
23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON
COMPUTER VISION

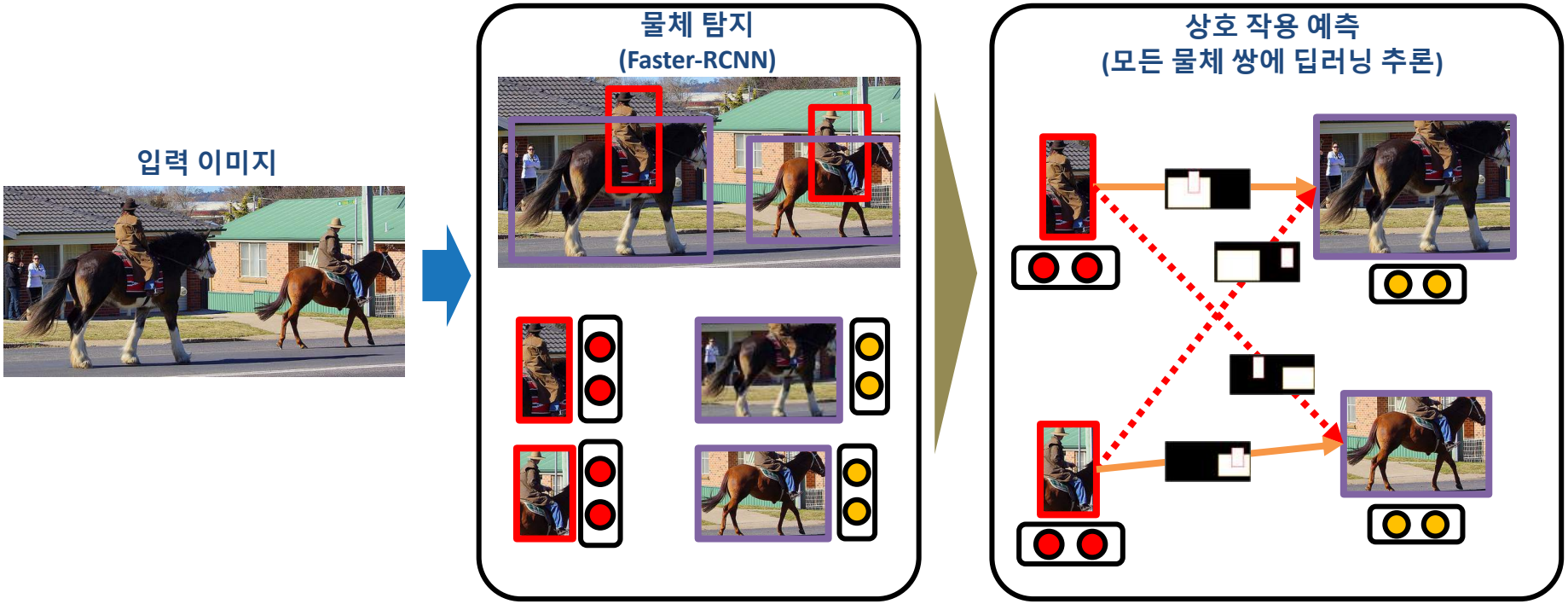
WWW.ECCV2020.EU

# UnionDet: Union-level Detector Towards Real-Time HOI Detection

Bumsoo Kim, Taeho Choi, Jaewoo Kang, Hyunwoo J. Kim

# 사람-사물- 관계 탐지: 기존 연구

- 다단계 기법
- 순차적 기법

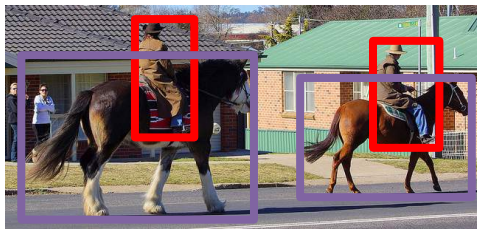


# 상호작용과 물체 탐지 동시에?

- UnionDet

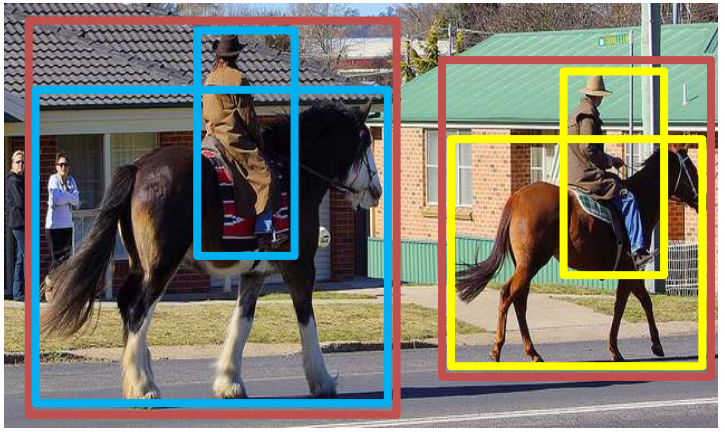
입력 이미지

물체 탐지
(Faster-RCNN)



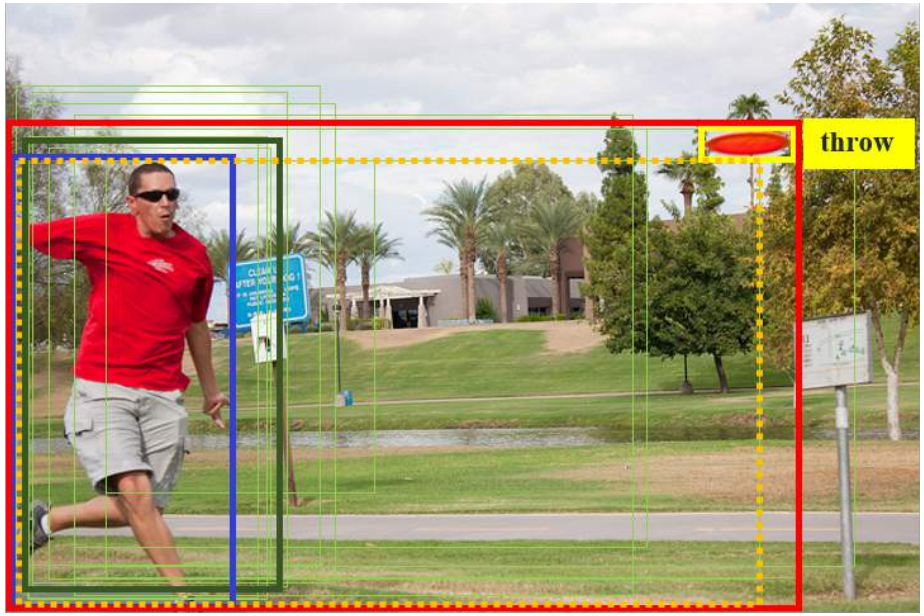
상호작용 영역 예측

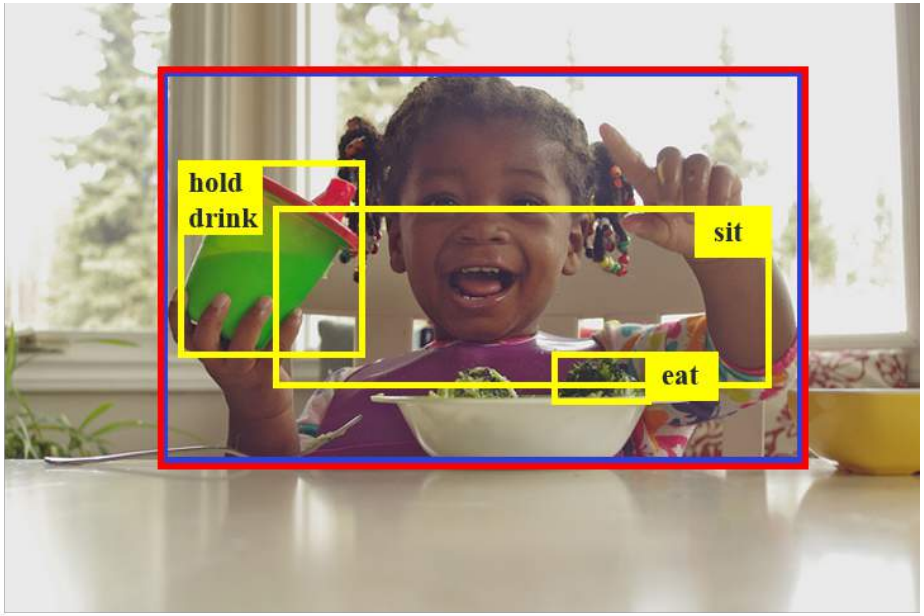
상호 작용 예측
(모든 물체 쌍에 딥러닝 추론)



- 탐지를 동시에 수행
- 물체 쌍 마다의 딥러닝 추론 제거

# 상호작용 영역 탐지의 문제

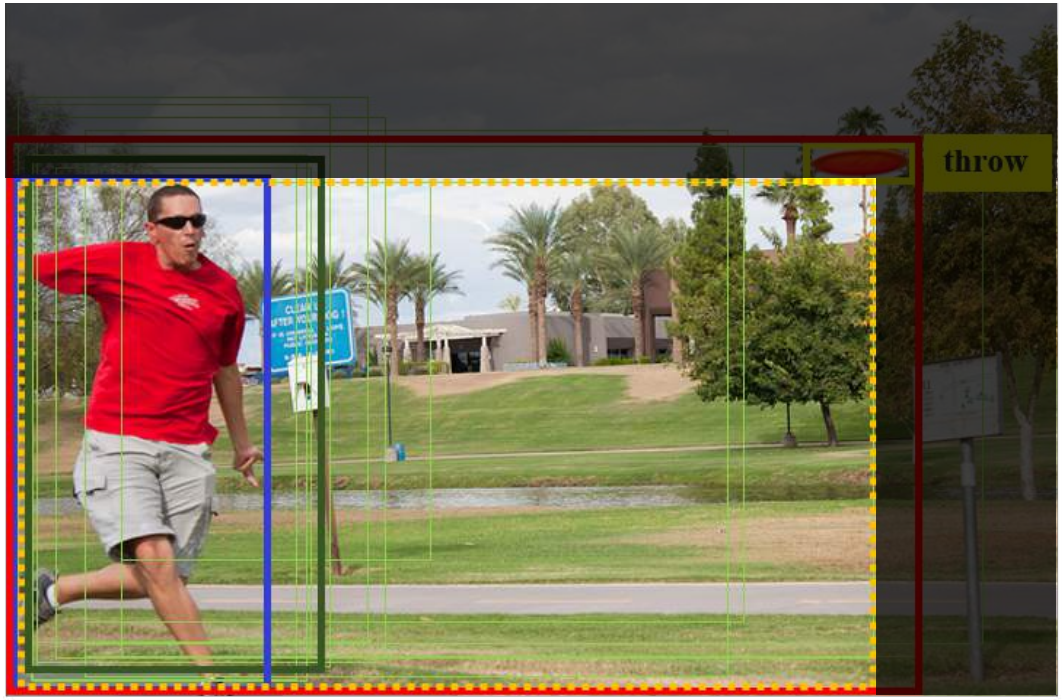


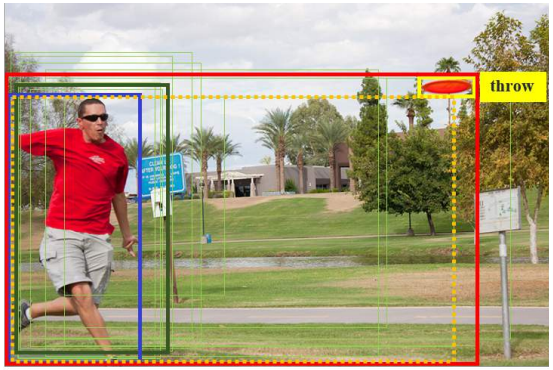

Prediction (union) Prediction (highest IoU) GT (human) GT (object) GT (union)

- IoU 상호작용을 탐지 영역 탐지에 부적합.
- 사람 영역에 치우침
- 동일 상호작용 영역에 다수의 상호작용이 존재

# 상호작용 영역 탐지의 문제



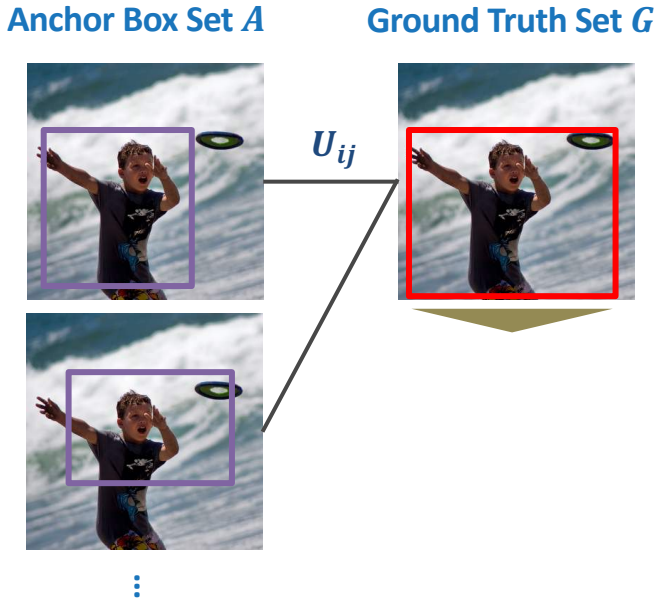

Prediction (union) Prediction (highest IoU)

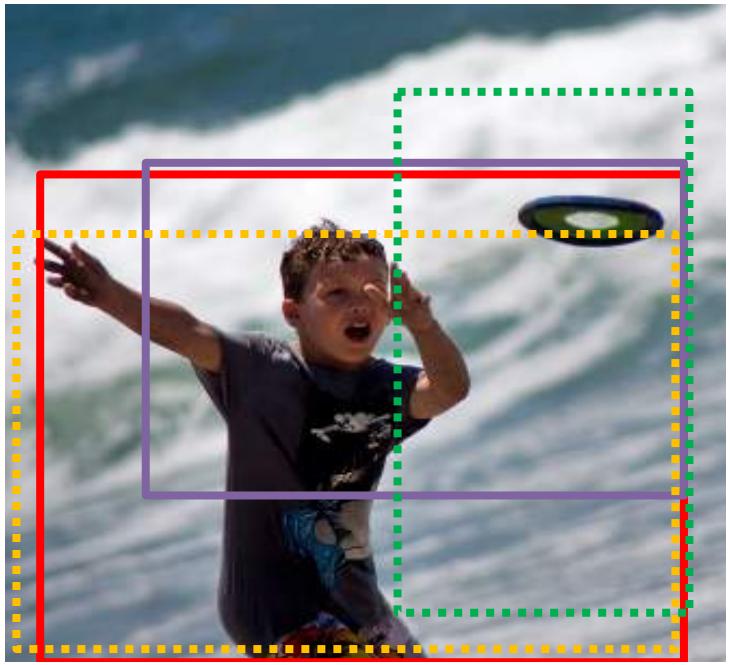



- IoU 상호작용을 탐지 영역 탐지에 부적합.
- 사람 영역에 치우침
- 동일 상호작용 영역에 다수의 상호작용이 존재

# 상호작용 영역 위치 정확도 함수

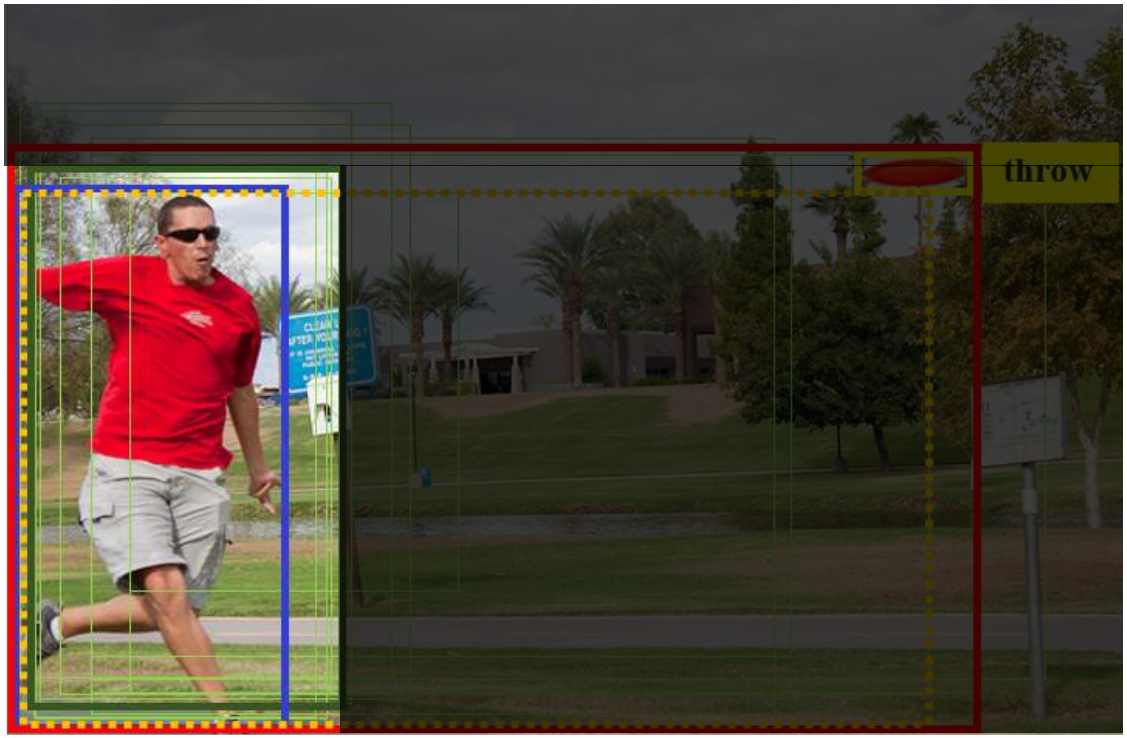
상호작용 영역 정답 매칭 및 위치 정확도 함수



✓ $$U_{ij} = \mathbb{1}(\text{IoU}(a_j, \breve{g}_i^{loc}) > t_u) \cdot \mathbb{1}\left(\frac{a_j \cap \breve{h}_i^{loc}}{\breve{h}_i^{loc}} > t_h\right) \cdot \mathbb{1}\left(\frac{a_j \cap \breve{o}_i^{loc}}{\breve{o}_i^{loc}} > t_o\right)$$

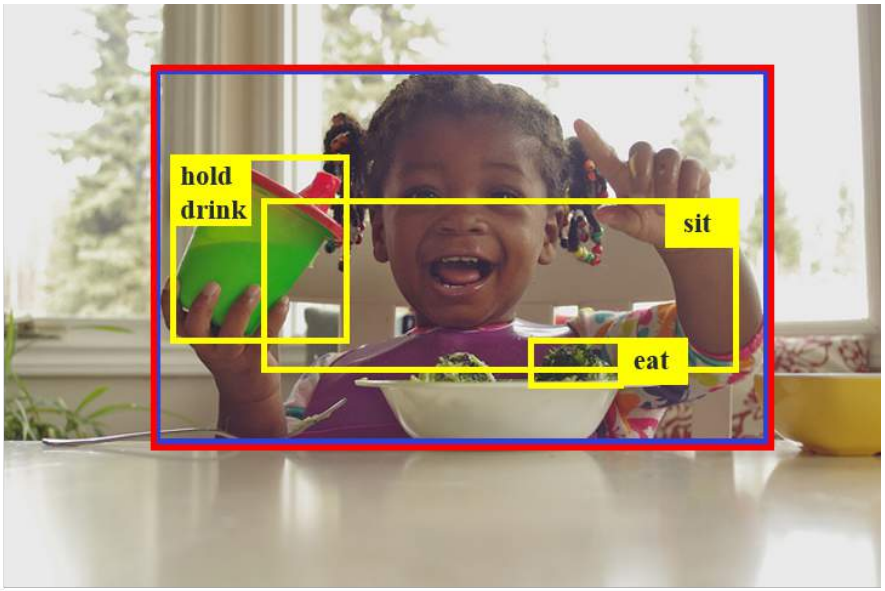
# 사람 영역 선호 문제




**기본 손실 함수+ 목적 물체 클래스 예측**

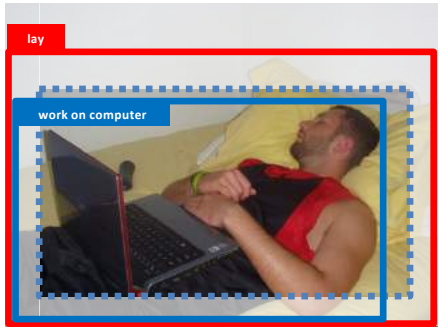
$$\mathcal{L}_u(\breve{\theta}) = \sum_{a_j \in A_+} \sum_{\breve{g}_i \in \breve{\mathcal{G}}} U_{ij} \left[ \mathcal{L}_{ij}^{act}(\breve{\theta}) + \mathcal{L}_{ij}^{loc}(\breve{\theta}) + \mathcal{L}_{ij}^{cls}(\breve{\theta}) \right] + \sum_{a_j \in A_-} \mathcal{L}_j^{bg}(\breve{\theta})$$

# 동일 지역 복수 상호작용

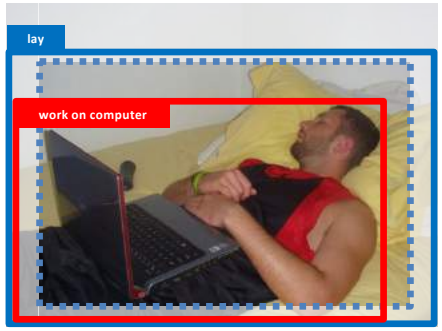



GT (human) GT (object) GT (union)

동일 상호작용 영역에
다수의 상호작용이 존재



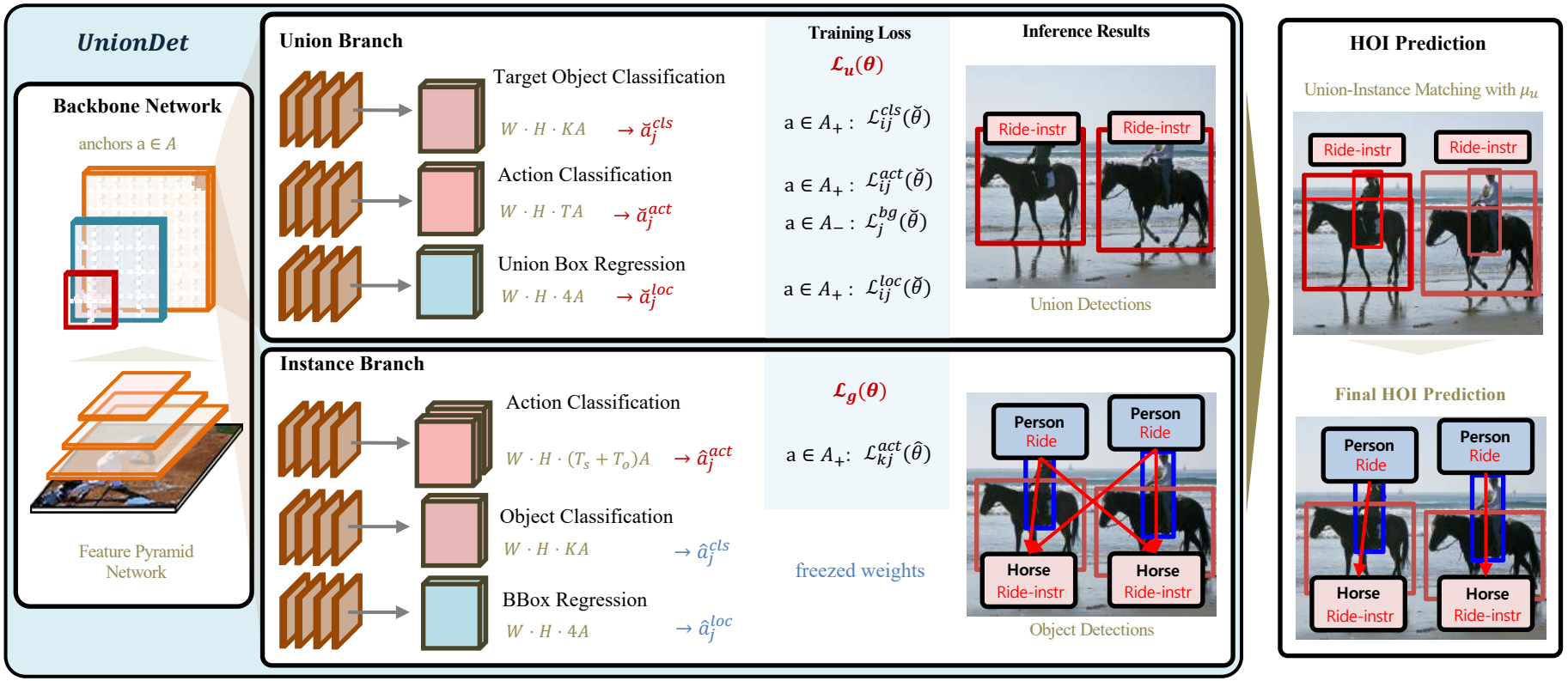

lay : ***Negative***
work on computer : ***Positive***



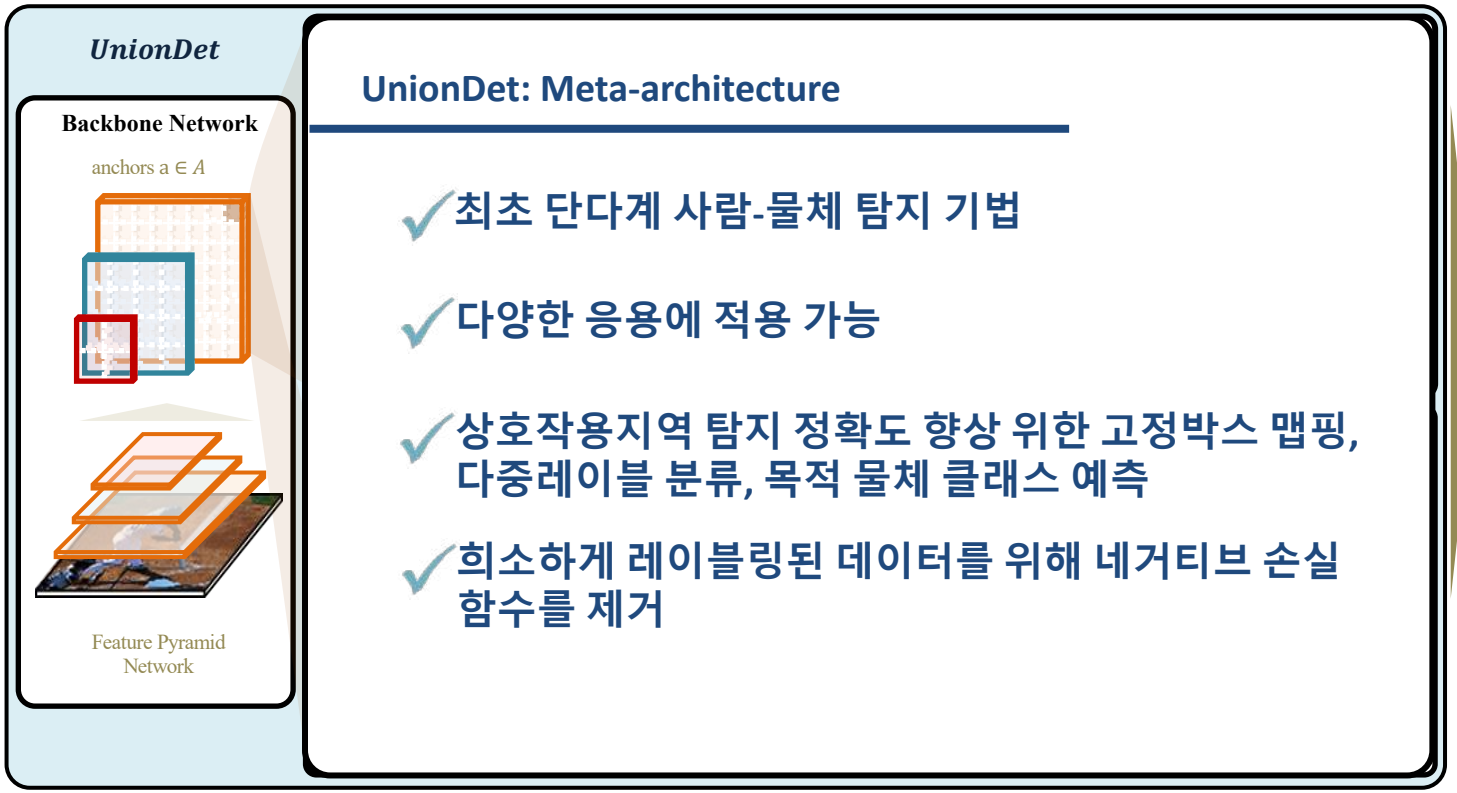

lay : ***Positive***
work on computer : ***Negative***

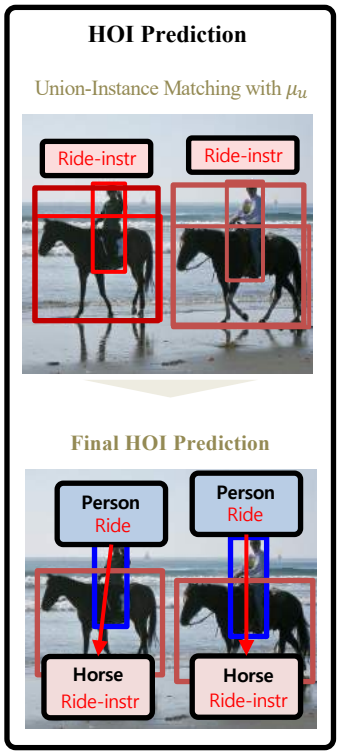
네거티브 패널티 제외

# UnionDet

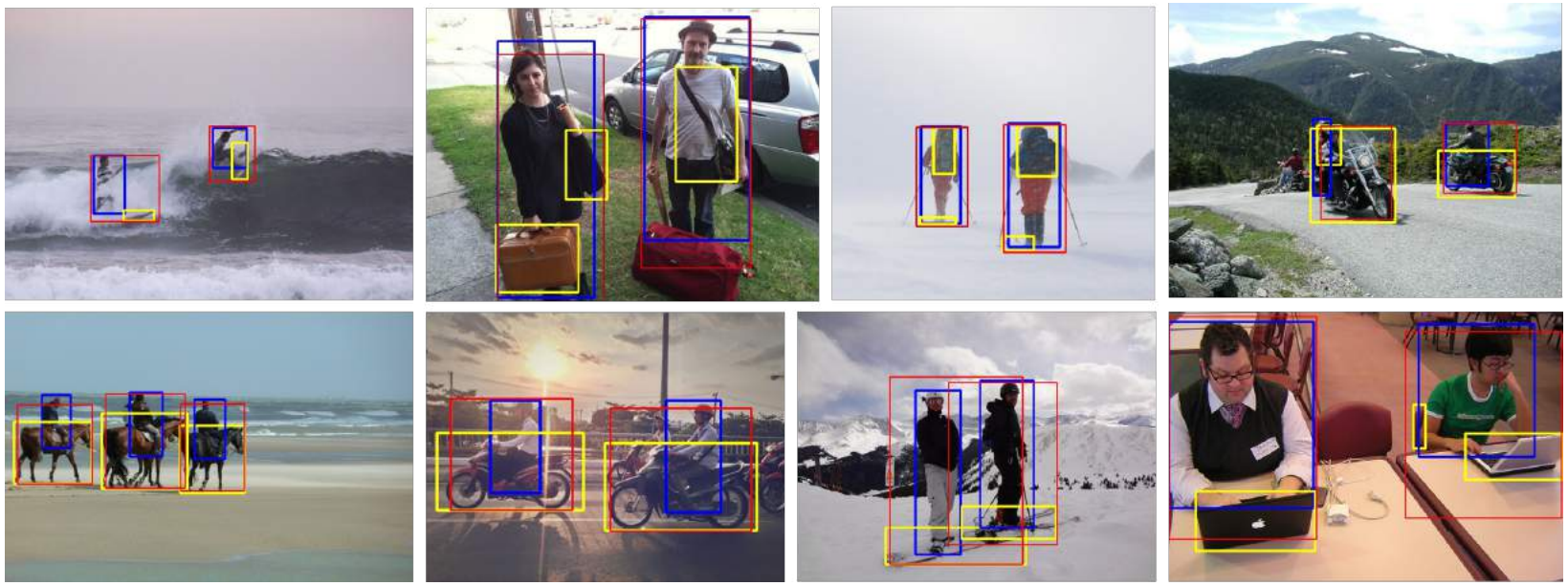

UnionDet
Backbone Network
anchors a ∈ A
Feature Pyramid Network
Union Branch
Target Object Classification
W · H · KA → ă_j^cls
Action Classification
W · H · TA → ă_j^act
Union Box Regression
W · H · 4A → ă_j^loc
Training Loss
L_u(θ)
a ∈ A+ : L_ij^cls(ă)
a ∈ A+ : L_ij^act(ă)
a ∈ A− : L_j^bg(ă)
a ∈ A+ : L_ij^loc(ă)
Inference Results
Ride-instr
Ride-instr
Union Detections
Instance Branch
Action Classification
W · H · (T_s + T_o)A → â_j^act
Object Classification
W · H · KA → â_j^cls
BBox Regression
W · H · 4A → â_j^loc
L_g(θ)
a ∈ A+: L_kj^act(â)
freezed weights
Person Ride
Person Ride
Horse Ride-instr
Horse Ride-instr
Object Detections
HOI Prediction
Union-Instance Matching with μ_u
Ride-instr
Ride-instr
Final HOI Prediction
Person Ride
Person Ride
Horse Ride-instr
Horse Ride-instr

# UnionDet

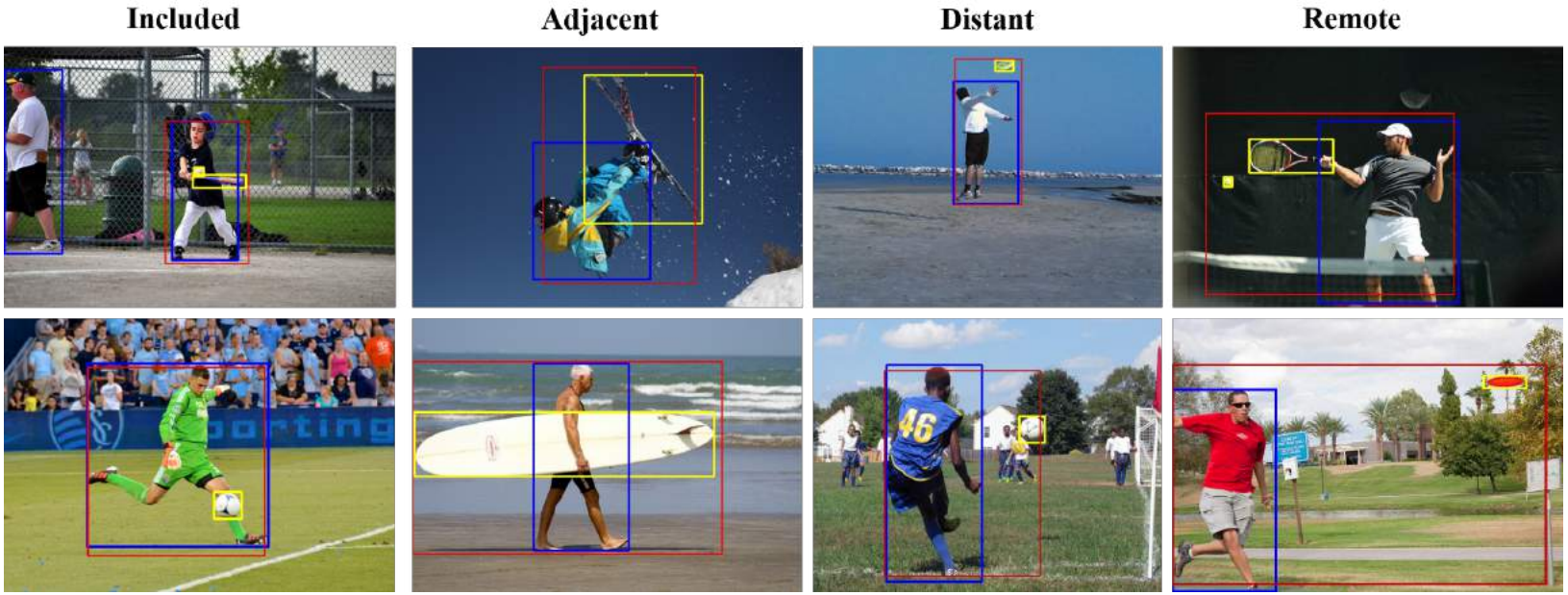

UnionDet
Backbone Network
anchors a ∈ A
Feature Pyramid Network
UnionDet: Meta-architecture
✓최초 단다계 사람-물체 탐지 기법
✓다양한 응용에 적용 가능
✓상호작용지역 탐지 정확도 향상 위한 고정박스 맵핑, 다중레이블 분류, 목적 물체 클래스 예측
✓희소하게 레이블링된 데이터를 위해 네거티브 손실 함수를 제거


HOI Prediction
Union-Instance Matching with $\mu_u$
Ride-instr
Ride-instr
Final HOI Prediction
Person Ride
Person Ride
Horse Ride-instr
Horse Ride-instr

# 동일 관계 독립적 탐지 가능

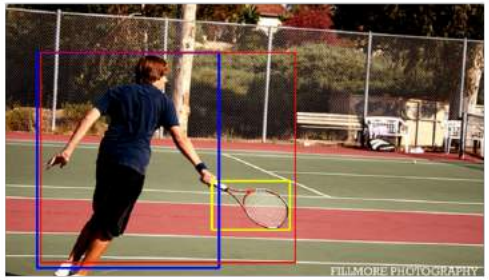


# 다양한 거리의 상호작용 탐지



# 다대다 상호작용 탐지

**One vs One**

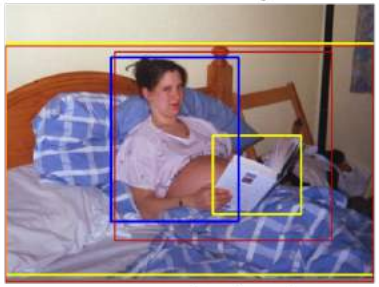


person – **hit instr** – tennis racket

**Many vs One**

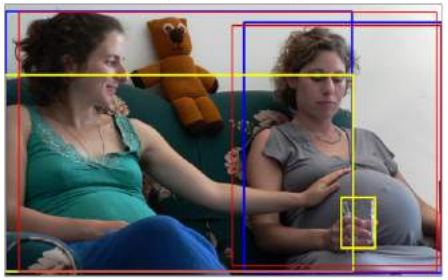
person (1) – **ride** – motorcycle
person (2) – **ride** – motorcycle

**One vs Many**

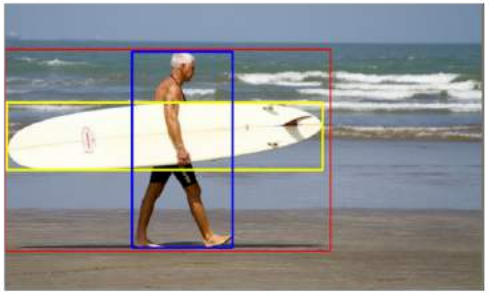


person – **read** – book
person – **lay** – bed

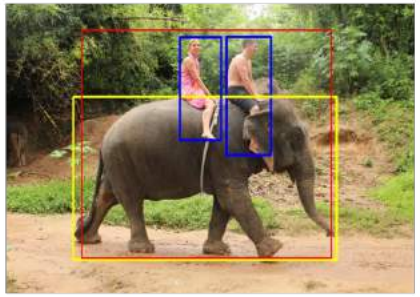
**Many vs Many**



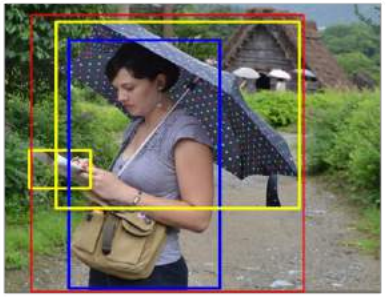
person(1) – **sit** – couch
person(2) – **sit** – couch
person(2) – **hold** – cup



person – **carry** – surfboard



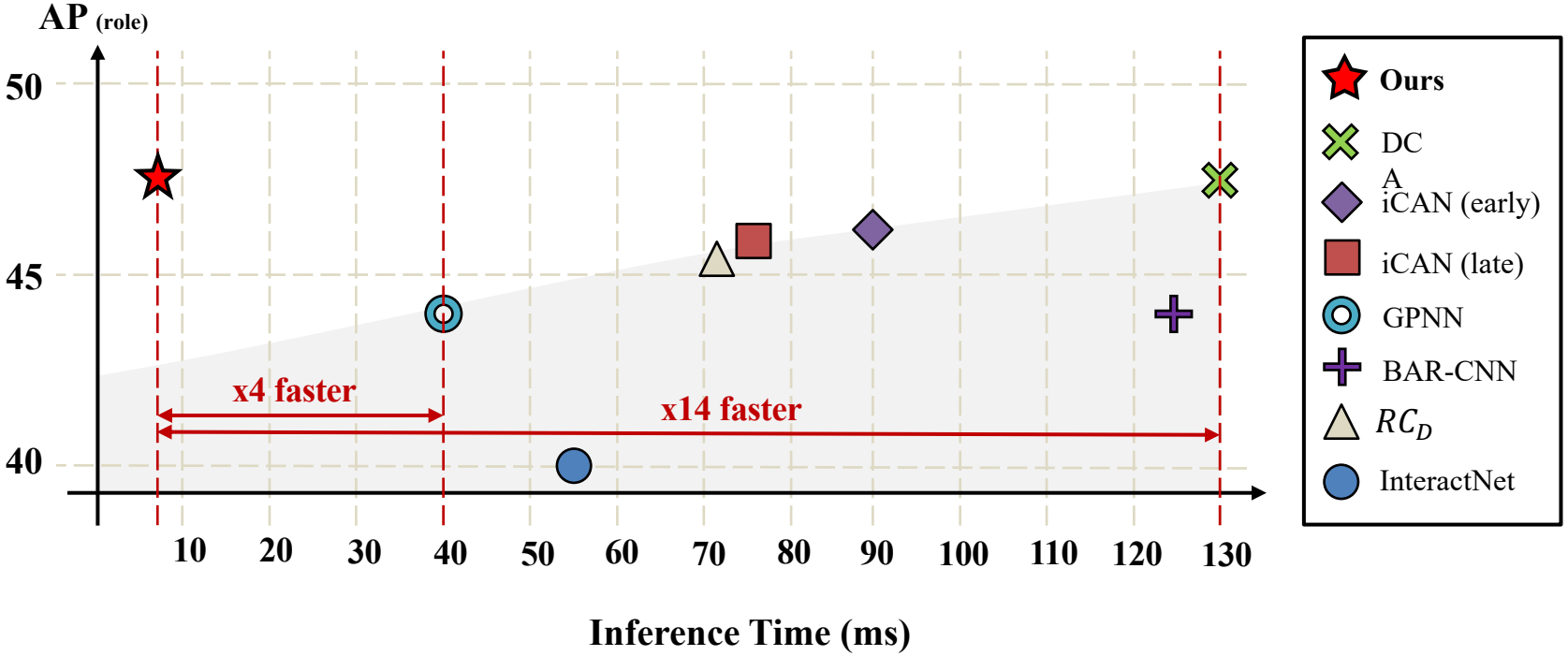
person (1) – **ride** – elephant
person (2) – **ride** – elephant



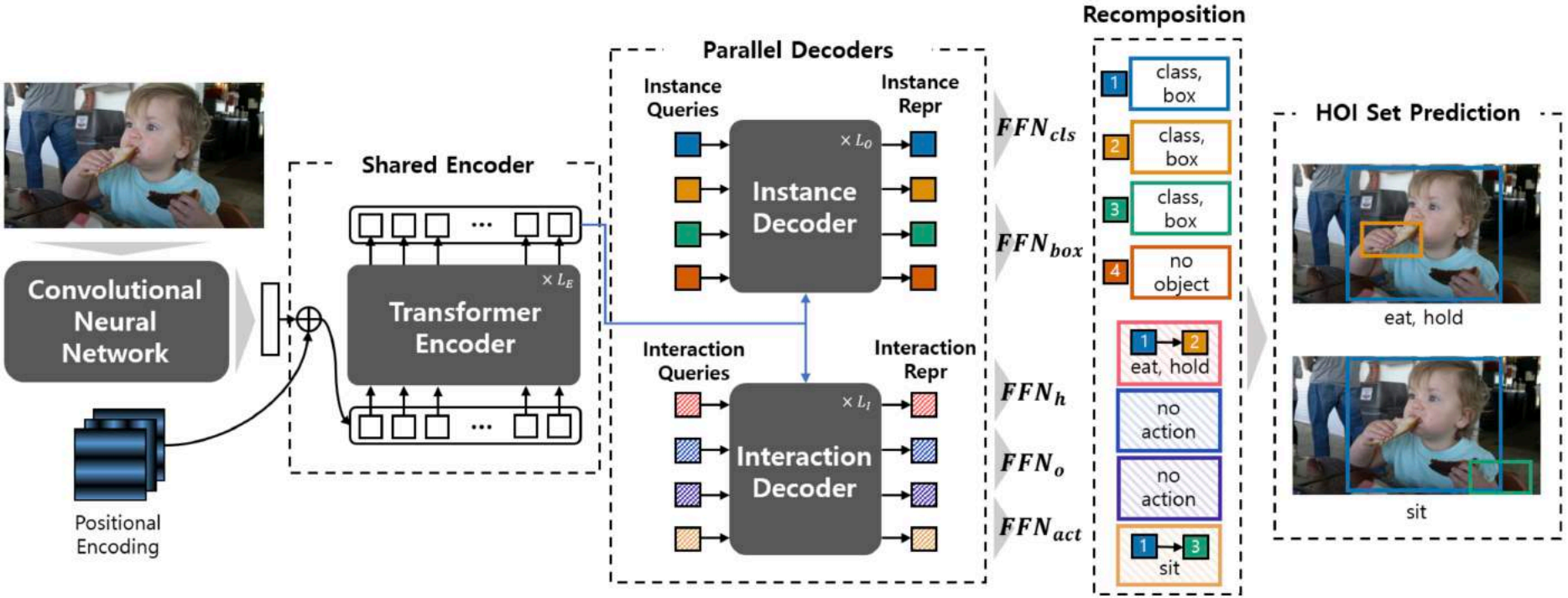
person – **hold** – umbrella
person – **read** – book

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

# 속도 개선 및 향상된 정확도

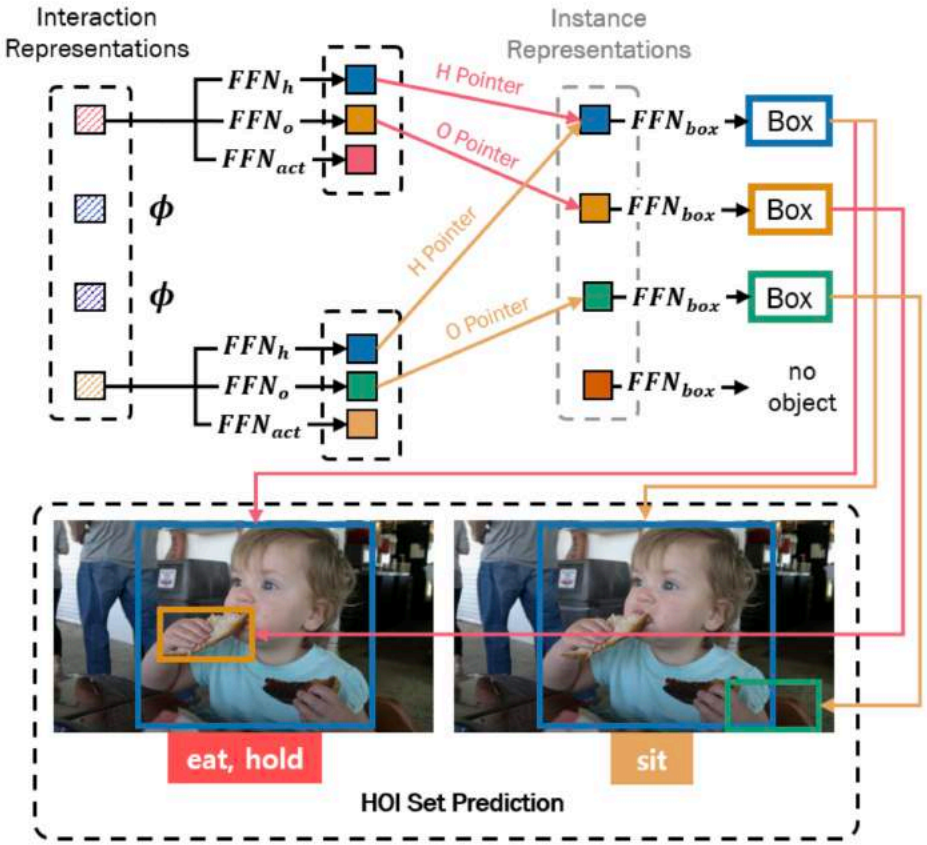


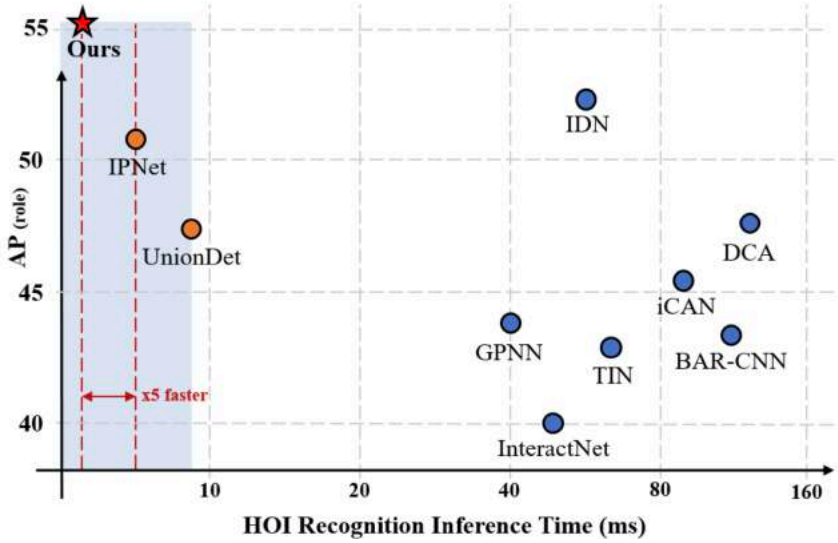
# 트랜스포머 기반 사람-사물 상호작용 탐지



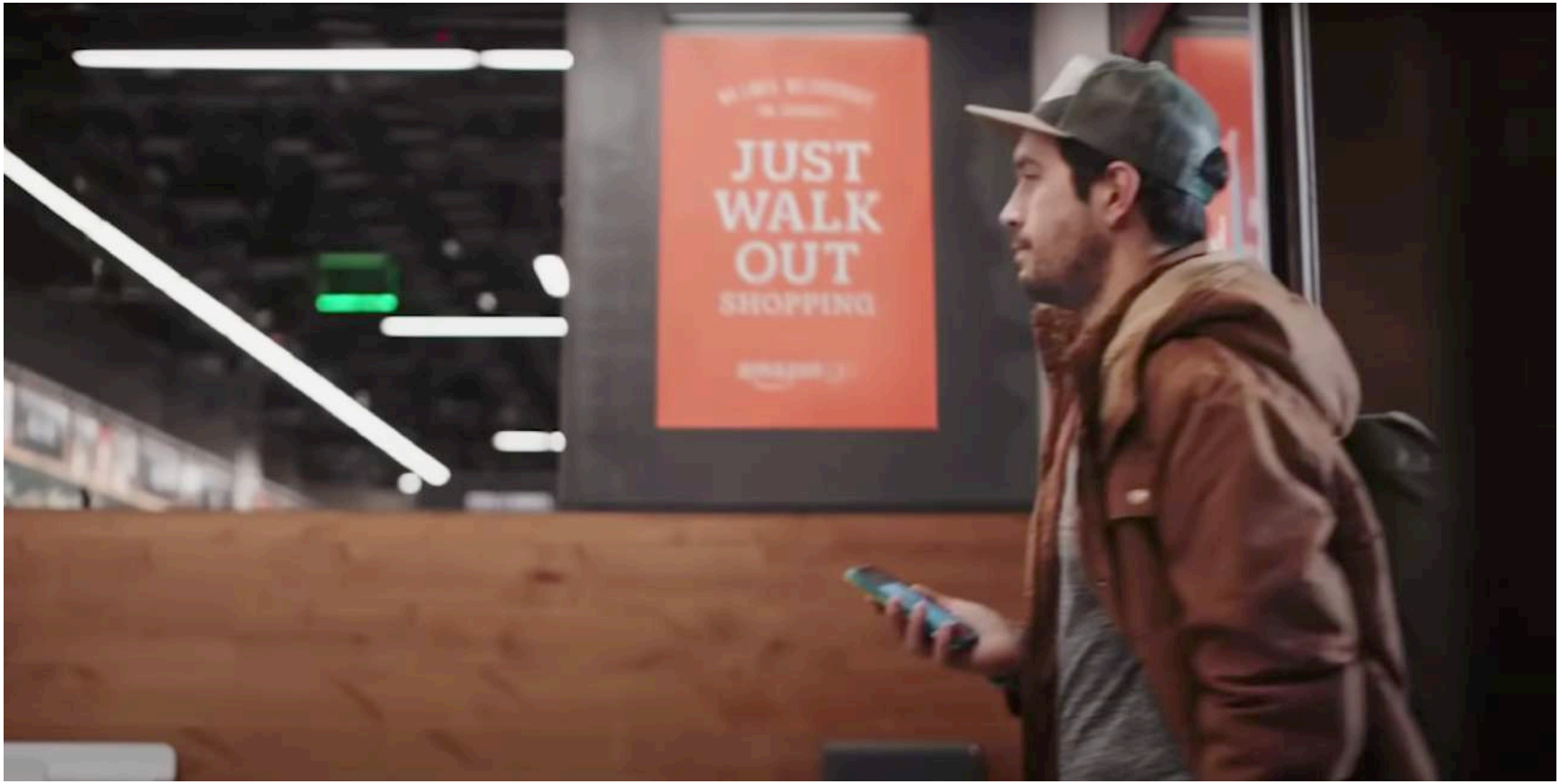

CVPR 2021, 구두 발표 논문. 김범수, 이준현, 강재우, 김은솔 (카카오 브레인), 김현우

# 트랜스포머 기반 사람-사물 상호작용 탐지



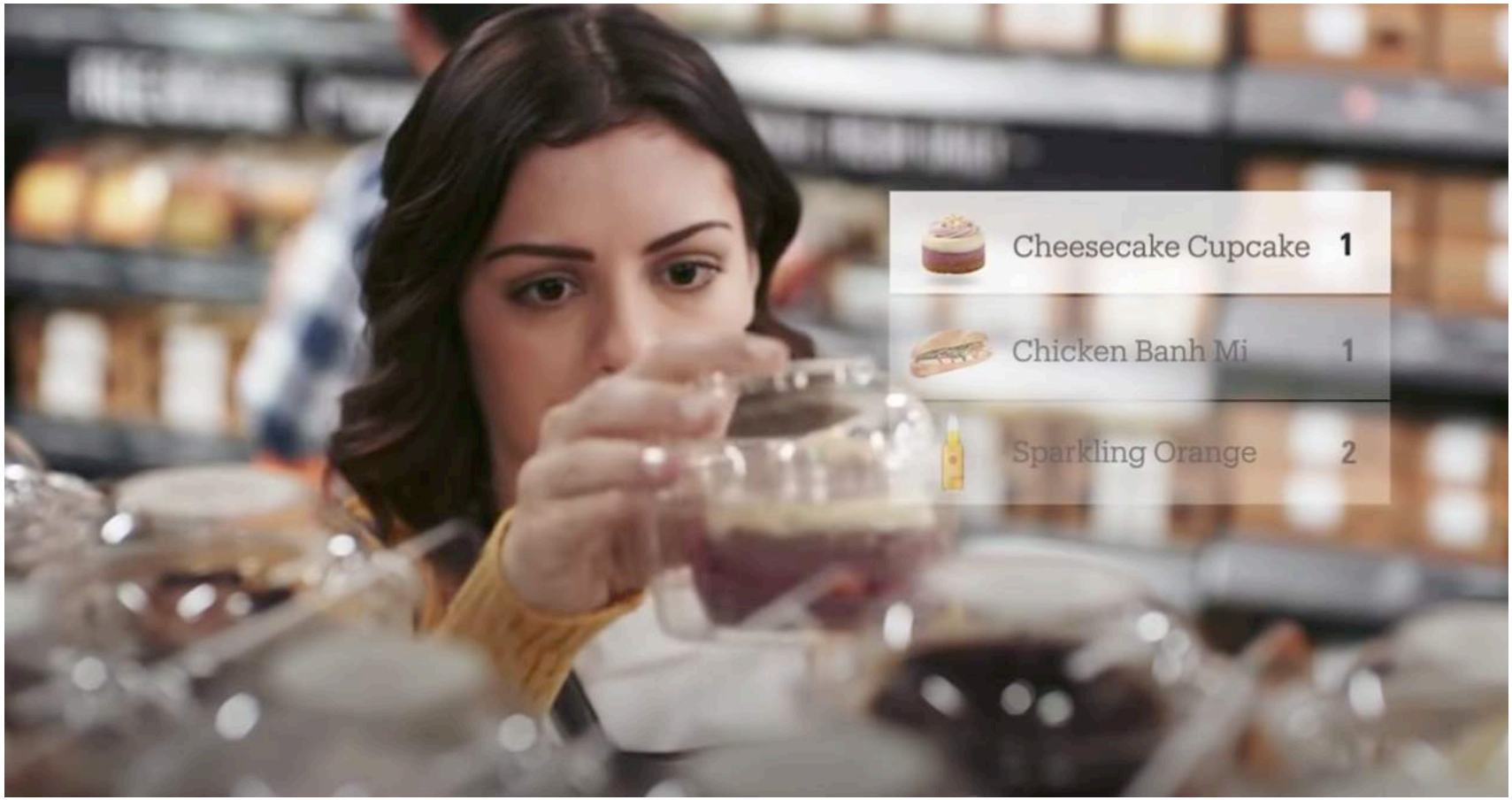



# Smart Store (AmazonGo)

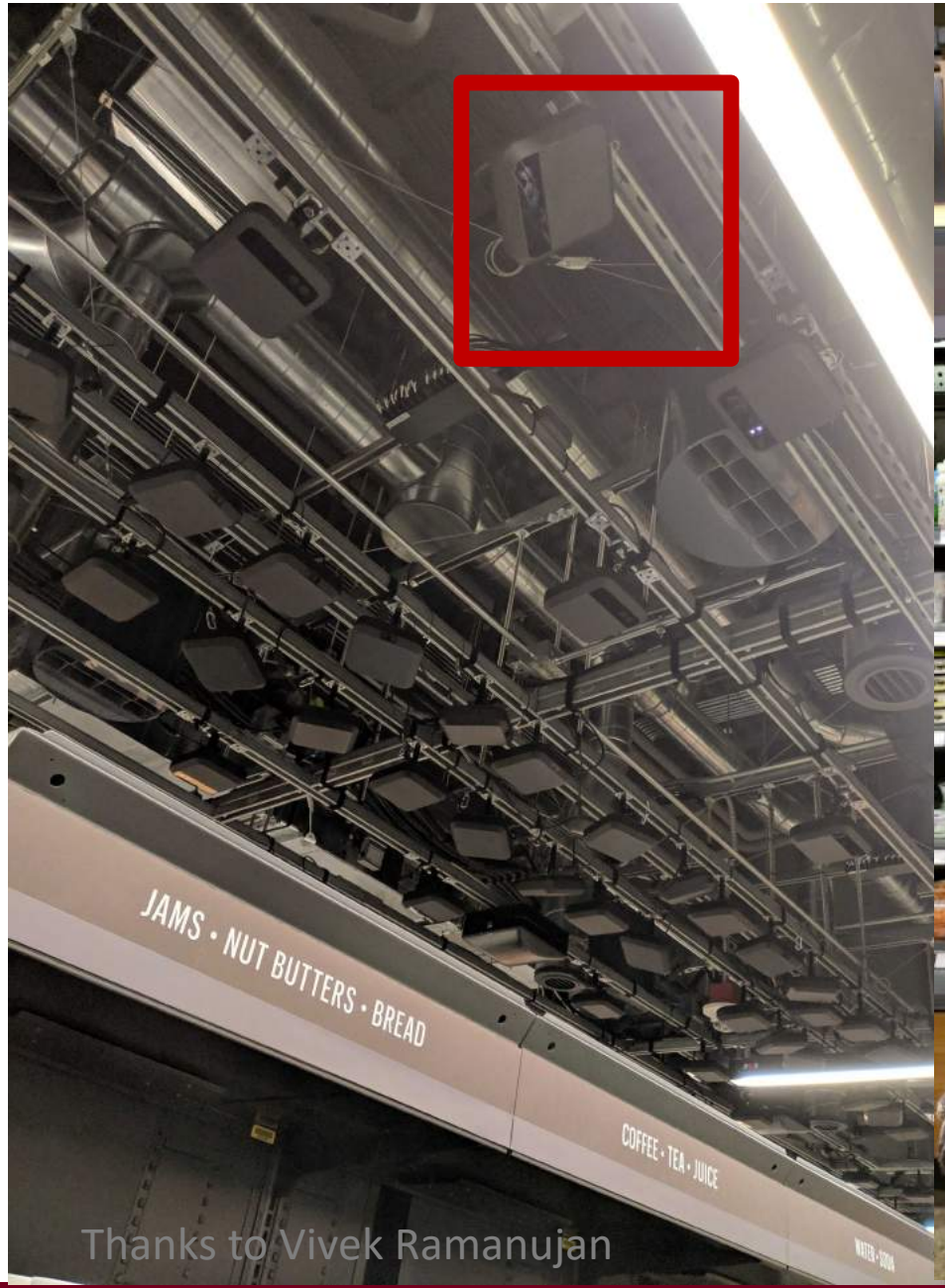



https://www.youtube.com/watch?v=NrmMk1Myrxc

# Amazon Go




https://www.youtube.com/watch?v=NrmMk1Myrxc




Thanks to Vivek Ramanujan

# 소주제 3: 물체간 탐지 기법 요약

- 물체간 상호작용 탐지
- 다단계 탐지 기법: InteractNet, Message-passing Network …
- 단단계 탐지 기법: UnionDet, HOTR, IPNet
- 사전 지식을 활용한 상호작용 탐지 성능 향상
  - 상호작용 사전 확율 학습, 언어 사전 지식 활용 등
- 희소 레이블링, 고정 박스 매핑 문제
- 고정 박스(Anchor box) 없는 최신 기법: HOTR (CVPR '21)

# 참고 문헌


- Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, **김현우**, HOTR: End-to-End Human-Object Interaction Detection with Transformers, *CVPR*, 2021. (구두발표).
- Bumsoo Kim, Taeho Choi, Jaewoo Kang, **김현우**, UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection, *ECCV, 2020.*
- Seong Jae Hwang, Sathya N. Ravi, Zirui Tao, **김현우**, Maxwell D. Collins, Vikas Singh, Tensorize, Factorize and Regularize: Robust Visual Relationship Learning, *CVPR*, 2018.
- Lu, Cewu, et al., Visual relationship detection with language priors, *ECCV*, 2016.
- Xu, Danfei, et al., Scene graph generation by iterative message passing, *CVPR*, 2017.
- Gkioxari, Georgia, et al., Detecting and recognizing human-object interactions, *CVPR*, 2018.
- Krishna, Ranjay, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *IJCV*, 2017.