SOON CHUN HYANG UNIVERSITY

SAIL Seminar 2023

QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information

Masato Tamura[1], Hiroki Ohashi[2], Tomoaki Yoshinaga[1]
Lumada Data Science Lab[1], Center for Technology Innovation[2]

IEEE/CVF Conference on CVPR. 2021

2023. 10. 30

순천향대학교 미래융합기술학과

석사과정 김병훈

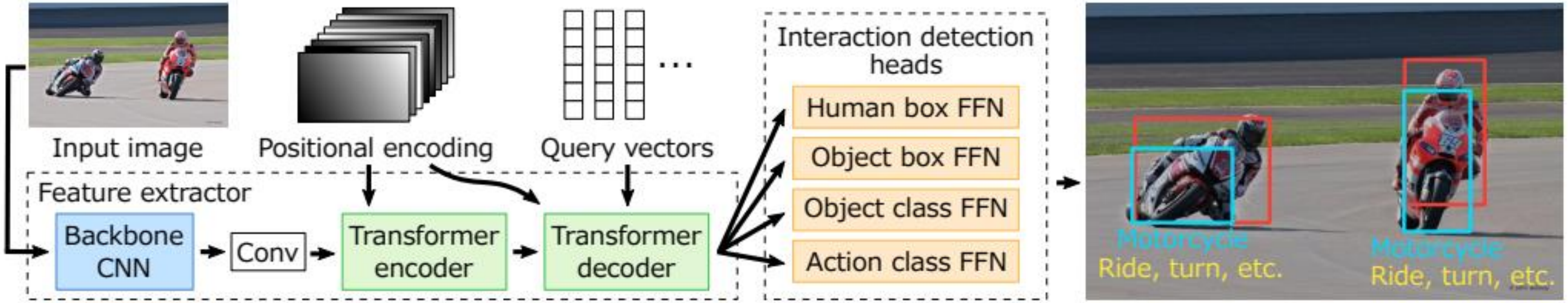
Index
Contents

1 Overview

1. Overview

QPIC

- ✓ This is the first work to use Attention- and Query-based methods in the HOI(Human-Object Interaction)
- ✓ Used DETR(End-to-End Object Detection with Transformers) as a base detector and extend it for HOI detector
- ✓ The feature extractor consists of an off-the-shelf CNN backbone network and a transformer base.



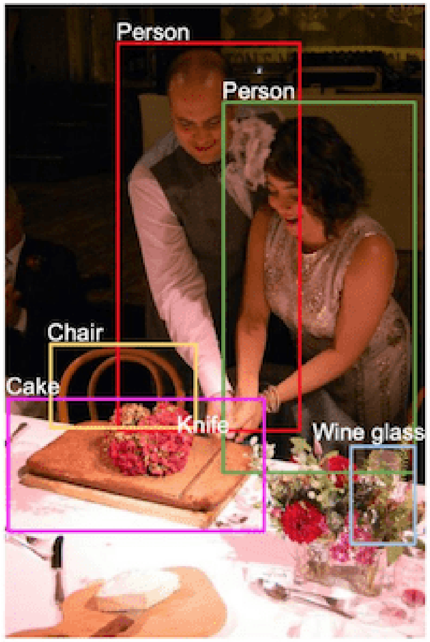
2 Background

2. Background

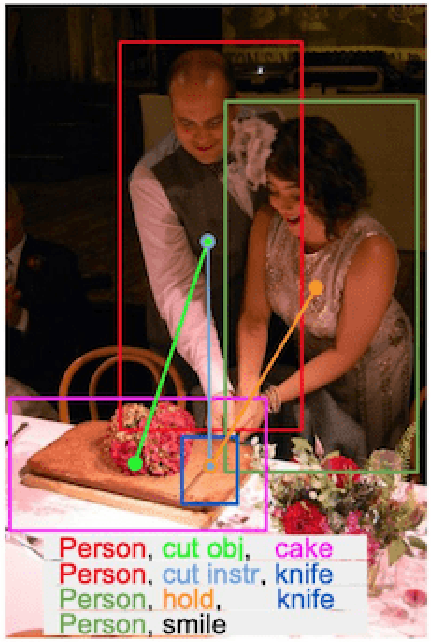
What is HOI(Human-Object Interaction)

- ✓ The task of detecting interactions between objects
- ✓ Further to object detection, add the process of finding interaction associations

Object detection
Classification
Localization



Object detection



HOI detection

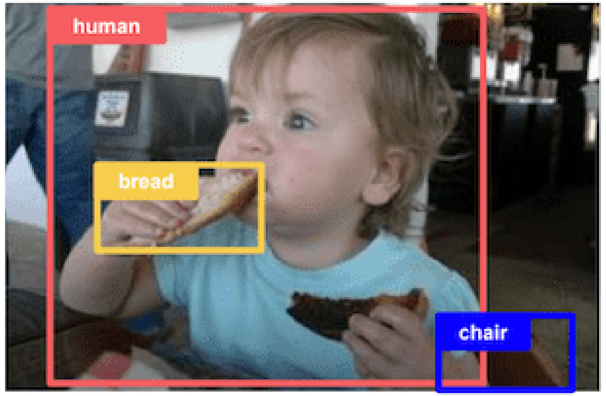
HOI detection
Classification
Localization
Interaction Association

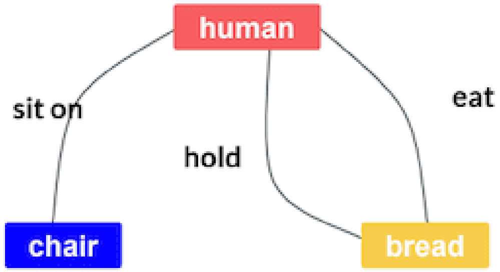
Find triplet

<human, object, interaction>

2. Background

What is HOI(Human-Object Interaction)





Set{ $(bbox_1^{human}, bbox_1^{obj}, Interaction_1), (bbox_2^{human}, bbox_2^{obj}, Interaction_2), (bbox_3^{human}, bbox_3^{obj}, Interaction_3)$}

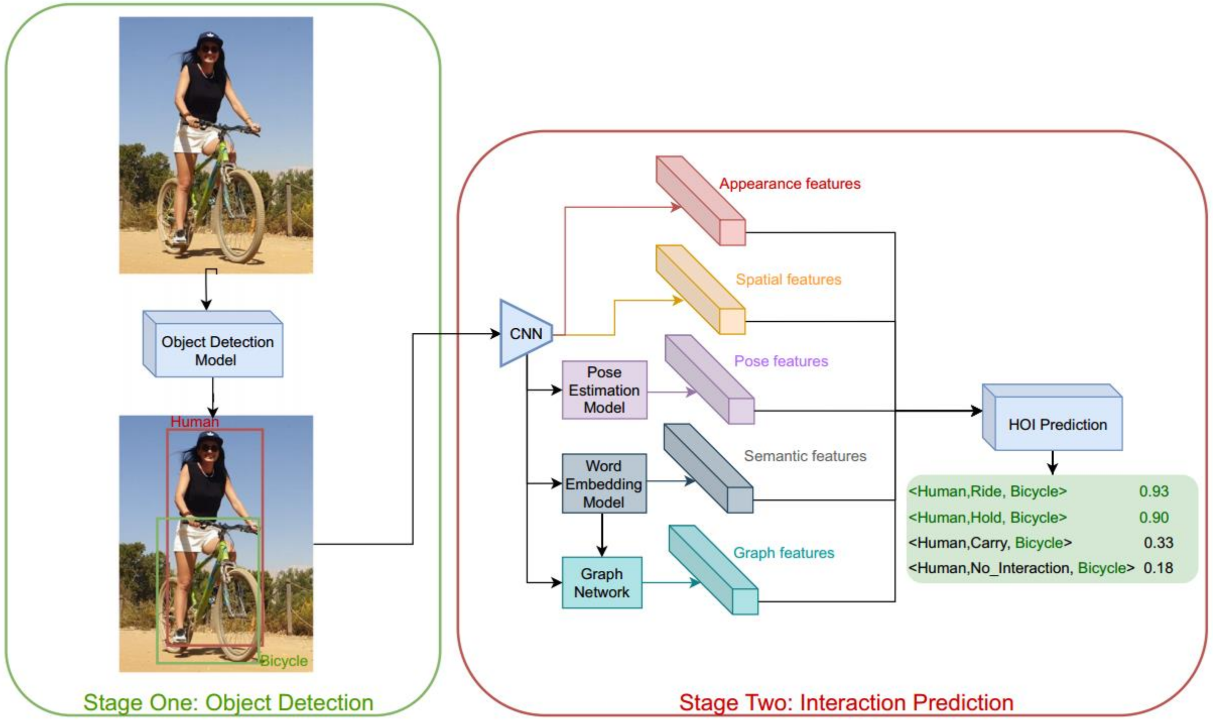
$bboxes^{human}, bboxes^{obj}$

{Human, Chair, Sit on}
{Human, Bread, Hold}
{Human, Bread, Eat}

2. Background

Previous HOI detector(two-stage methods)

- ✓ Consists of Stage One (Object Detection) and Stage Two (Interaction Prediction)
- ✓ The process is to detect all the objects in the image and then use a neural network to find all the parallel interaction scores.



[Representative models]

iCAN, InteractNet

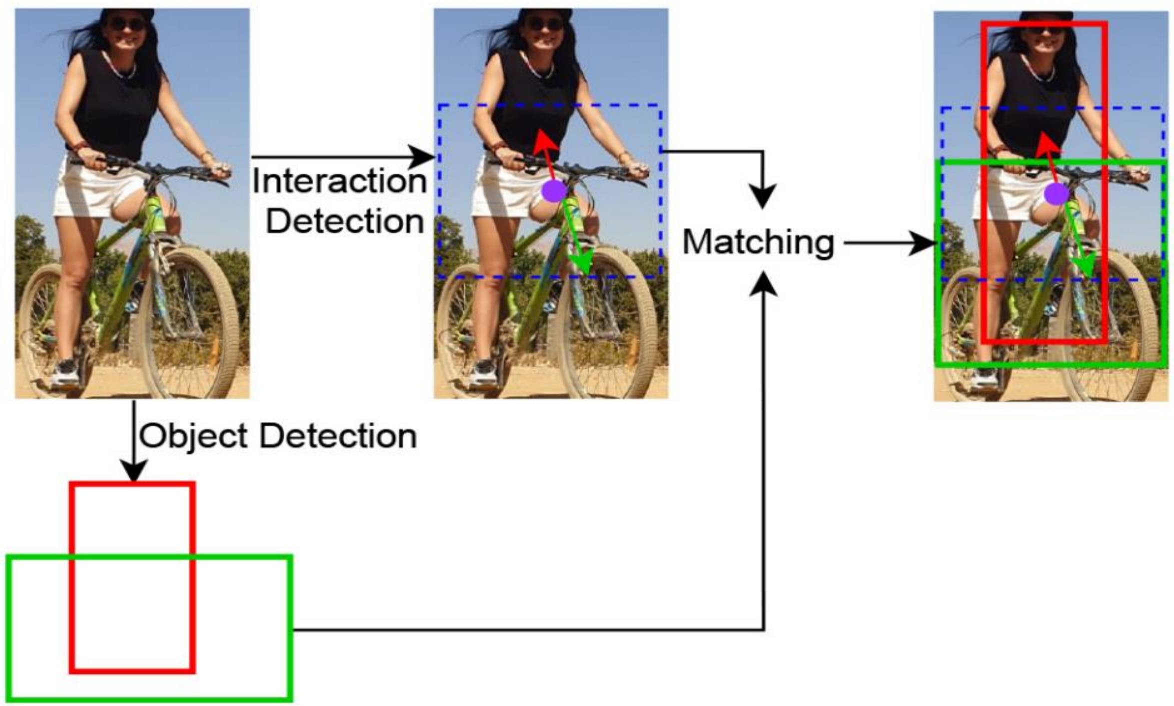
[Limitation]

1. Images are missing contextual features
2. It uses a pairwise neural network, which has the limitation of being inefficient in terms of time and computational cost.

2. Background

Previous HOI detector(single-stage methods)

- ✓ It uses a matching method that performs object detection and interaction detection in parallel.
- ✓ Use interaction boxes or union boxes to reduce inference time while maintaining performance.



[Representative models]

PPDM, CenterNet

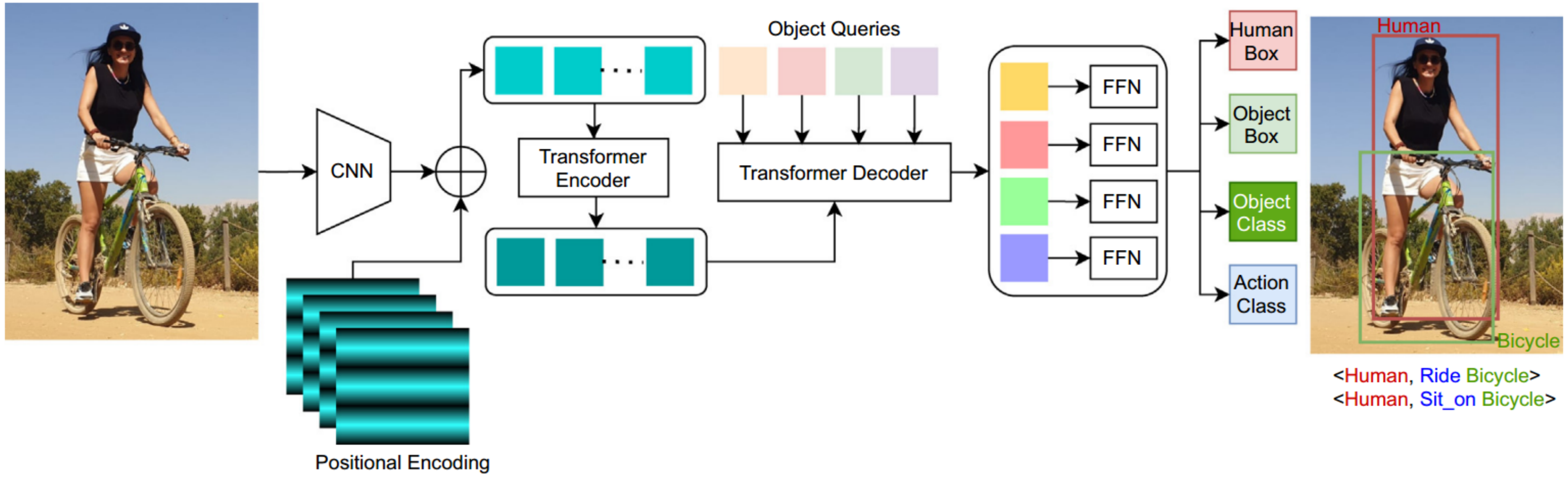
[Limitation]

1. Images are missing contextual features
2. Requires additional post-processing steps or heuristic thresholding

2. Background

Transformer based method

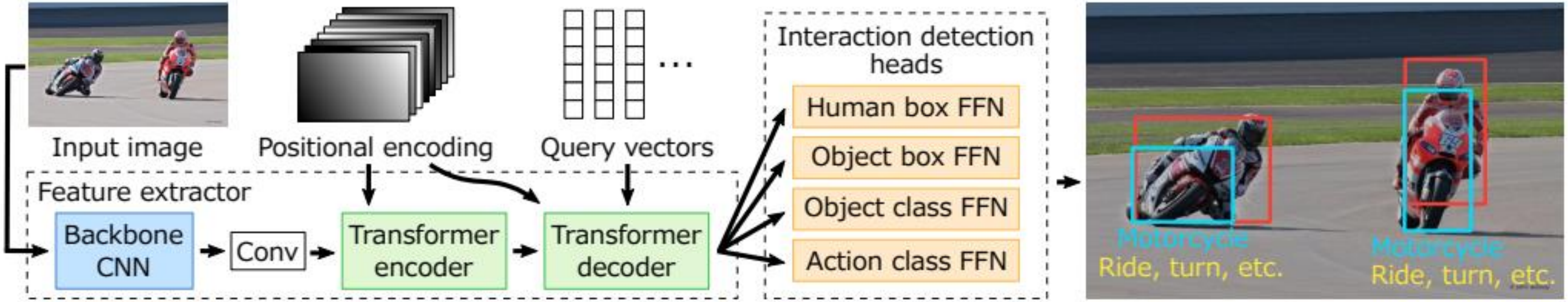
- ✓ Transformers have had success with Natural Language Processing (NLP) and recently applied to images with image transformers.
- ✓ Attention mechanisms can be used to extract overall features of an image.
- ✓ It consists of an encoder and a decoder to predict the hoi triplet at once.



3 Proposed Methods

3. Proposed Method

Overall Architecture



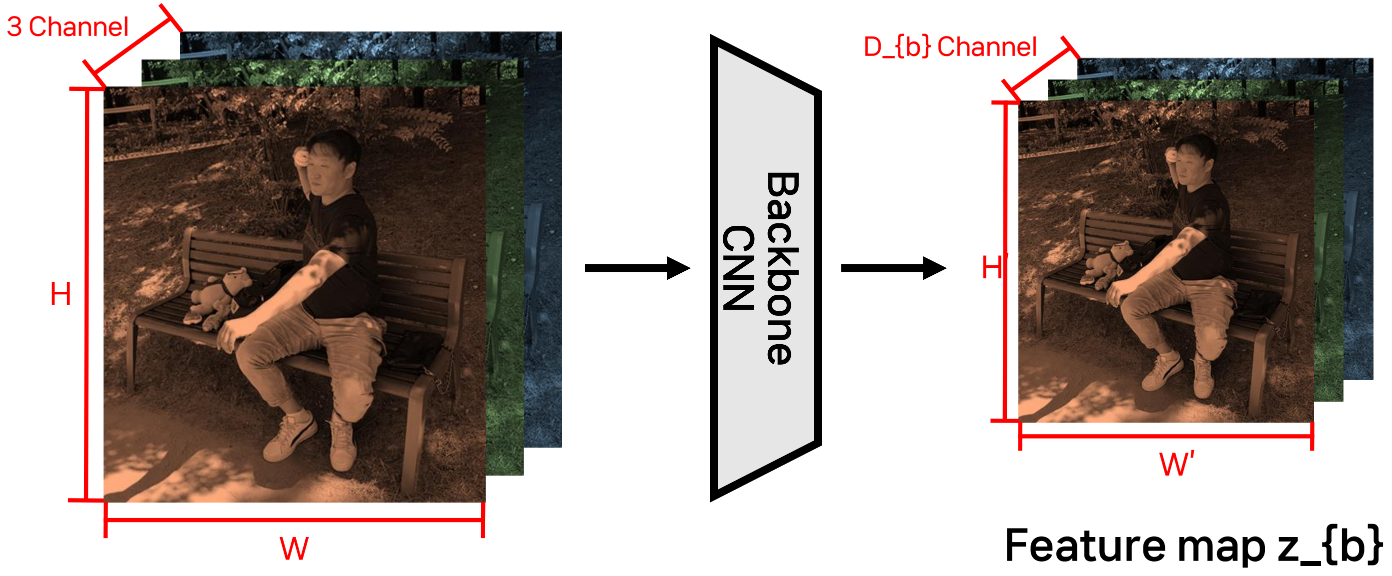
The sections are presented in two parts:

1. Feature Extractor
2. interaction detection head.

3. Proposed Method

Feature Extraction (backbone network)

Given an input image $x \in \mathrm{R}^{3\times\mathrm{H}\times\mathrm{W}}$, it is calculated as a feature map $z_b \in \mathbb{R}^{D_b\times H'\times W'}$ using off-the-shelf backbone network, where H and W are the height and width of the input image, H' and W' those the output feature map, and D_b is the number of channels. Typically $H' < \mathrm{H}$, $W' < \mathrm{W}$. z_b is then input to a projection convolution layer with a kernel size of 1 × 1 to reduce the dimension from D_b to D_c.

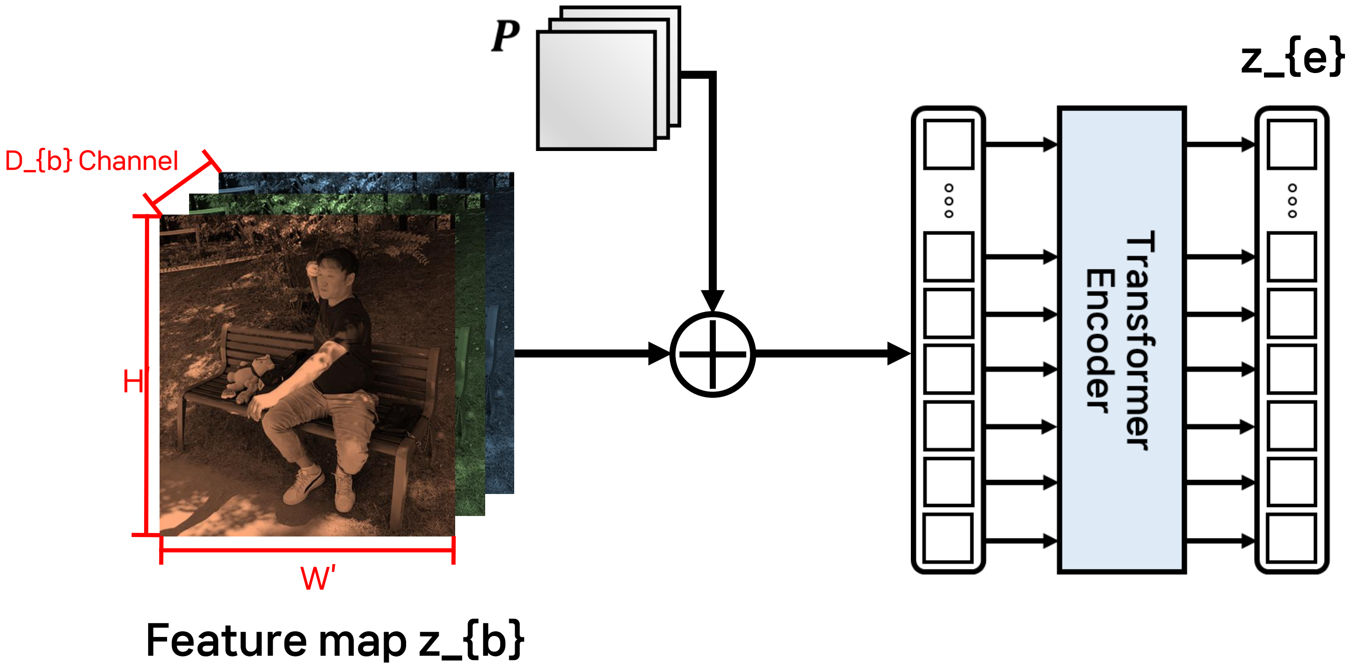


Input image X

3. Proposed Method

Feature Extraction (Transformer Encoder)

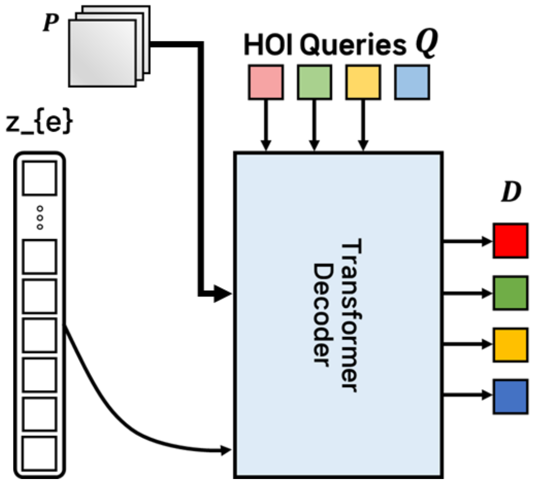
The transformer encoder takes as input a feature map $z_b \in \mathbb{R}^{D_c \times H' \times W'}$ and a fixed positional encoding $p \in \mathbb{R}^{D_c \times H' \times W'}$ that contains positional information. Then, it extracts a feature map that is rich in contextual information using a self-attention mechanism. The encoded feature map is $z_e \in \mathbb{R}^{D_c \times H' \times W'}$, which can be obtained via $z_e = f_{enc}(z_c, p)$. where $f_{enc}(\cdot,\cdot)$ is a set of stacked transformer encoder layers.



3. Proposed Method

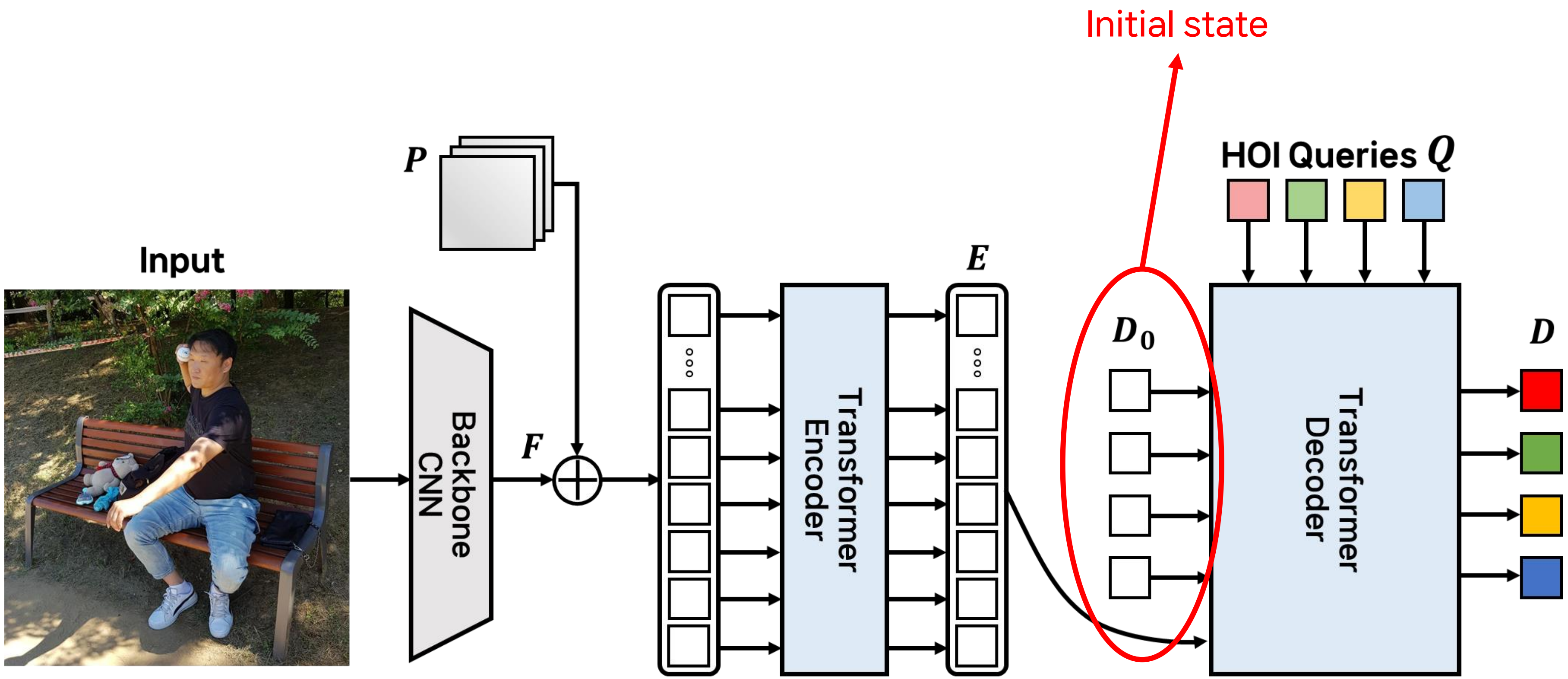
Feature Extraction (Transformer Decoder)

The transform decoder takes as input the encoder output $z_e \in \mathbb{R}^{D_c \times H' \times W'}$, a learnable query vector $Q = \{q_i | q_i \in \mathbb{R}^{D_c}\}_{\mathrm{i}=1}^{N_q}$, and a positional encoding p containing spatial information. Then, it outputs an embedding vector $\mathrm{D} = \{d_i | d_i \in \mathbb{R}^{D_c}\}_{\mathrm{i}=1}^{N_q}$ containing image-wide contextual information for HOI detection. One query vector q_i is designed to contain at most one human-object pair and an interaction between them, which means that the number of queries N_q is always larger than the number of human-object pairs in the image. The decoded embeddings are then obtained as $\mathrm{D} = f_{dec}(z_e, p, Q)$, where $f_{dec}(\cdot,\cdot,\cdot)$ is a set of stacked transformer decoder layers.



3. Proposed Method

Feature Extraction (Architecture)



3. Proposed Method

Interaction Detection Head

The interaction detection head defines the embedding result D as follows:

1. human-bbox vector : $b^{(h)} \in [0.1]^4$
2. object-bbox vector : $b^{(o)} \in [0.1]^4$
3. object class(one-hot vector) : $\mathrm{c} \in [0.1]^{N_{obj}}$, N_{obj} is the number of object class
4. action class : $\mathrm{a} \in [0.1]^{N_{act}}$, N_{act} is the number of action class

Action class is not necessarily a one-hot vector because there may be multiple actions. The vectors listed are input to the 4 heads(f_h, f_o, f_c, f_a), respectively

3. Proposed Method

Interaction Detection Head

The prediction of normalized follows:

1. human-bbox : $\left\{\hat{b}_i^{(h)} \middle| \hat{b}_i^{(h)} \in [0,1]^4\right\}_{\mathrm{i=1}}^{N_q}$
2. object-bbox : $\left\{\hat{b}_i^{(o)} \middle| \hat{b}_i^{(o)} \in [0,1]^4\right\}_{\mathrm{i=1}}^{N_q}$
3. probability of object classes : $\left\{\hat{c}_i | \hat{c}_i \in [0,1]^{N_{obj}+1}, \sum_{j=1}^{N_{obj}+1} \hat{c}_i(j) = 1\right\}_{\mathrm{i=1}}^{N_q}$, where $v(j)$ denotes the j-th element of v
4. probability of action classes : $\{\hat{a}_i | \hat{a}_i \in [0,1]^{N_{act}}\}_{\mathrm{i=1}}^{N_q}$

This predictions are calculated as (σ, ς / sigmoid, softmax):

1. $\hat{b}_i^{(h)} = \sigma\left(f_h(d_i)\right)$

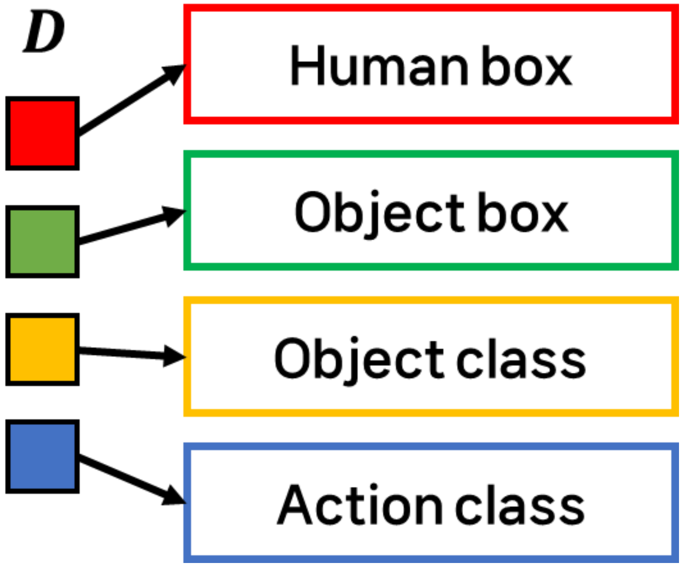
2. $\hat{b}_i^{(o)} = \sigma(f_o(d_i))$

3. $\hat{c}_i = \varsigma(f_c(d_i))$

4. $\hat{a}_i = \sigma(f_a(d_i))$

3. Proposed Method

Interaction Detection Head



$$\hat{b}_i^{(h)} = \sigma\left(f_h(d_i)\right)$$

$$\hat{b}_i^{(o)} = \sigma(f_o(d_i))$$

$$\hat{c}_i = \varsigma(f_c(d_i))$$

$$\hat{a}_i = \sigma(f_a(d_i))$$

4 Experiments

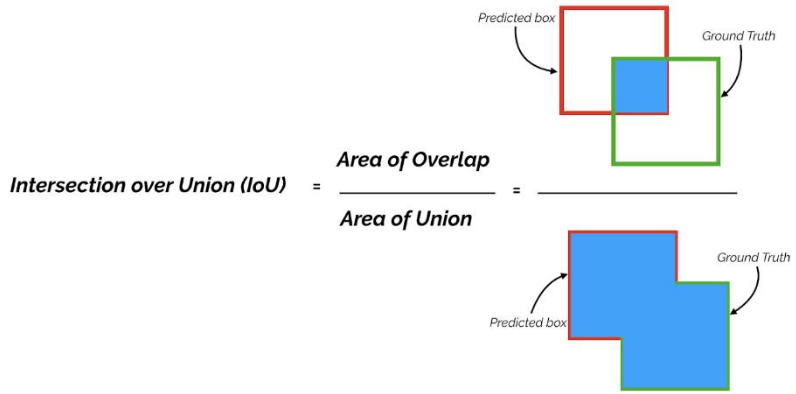
4. Experiments

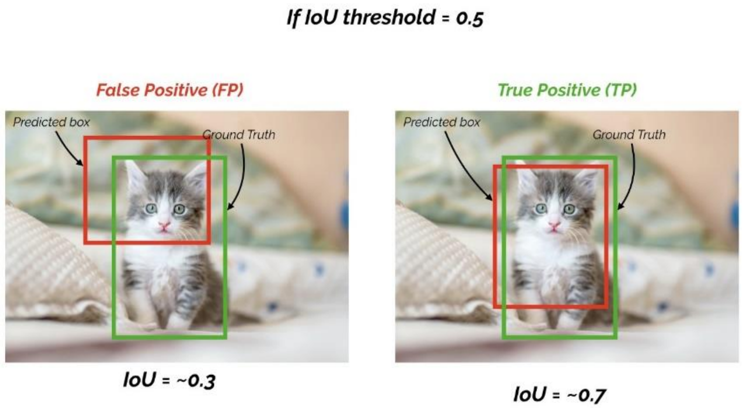
Datasets and Evaluation Metrics

- Datasets

Datasets	Train	Validation	Test	Object class	Action class
V-COCO	2,533	2,867	4,946	80	29
HICO-DET	38,118	-	9,658	80	117

- Evaluation Metrics : mean average precision(mAP)





4. Experiments

Comparison to State-of-the-Art

- HICO-DET
  - Default : APs are calculated on the basis of all the test images
  - Known object : each AP is calculated only on the basis of images that contain the object class corresponding to each AP
  - full, rare, non-rare : 600(entire), 138(less than 10), 462(more than 10)
- V-COCO
  - Scenario 1 : It should correctly detect the 'no-object' class.
  - Scenario 2 : Ignore the 'no-object' class.

	Default			Known object		
Method	full	rare	non-rare	full	rare	non-rare
FCMNet [20]	20.41	17.34	21.56	22.04	18.97	23.13
ACP [13]	20.59	15.92	21.98	–	–	–
VCL [11]	23.63	17.21	25.55	25.98	19.12	28.03
DRG [4]	24.53	19.47	26.04	27.98	23.11	29.43
UnionDet [12]	17.58	11.72	19.33	19.76	14.68	21.27
Wang *et al.* [32]	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [17]	21.73	13.78	24.10	24.58	16.65	26.84
Ours (ResNet-50)	29.07	21.85	31.23	31.68	24.14	33.93
Ours (ResNet-101)	**29.90**	**23.92**	**31.69**	**32.38**	**26.06**	**34.27**

HICO-DET

Method	Scenario 1	Scenario 2
VCL [11]	48.3	–
DRG [4]	51.0	–
ACP [13]	53.0	–
FCMNet [20]	53.1	–
UnionDet [12]	47.5	56.2
Wang *et al.* [32]	51.0	–
Ours (ResNet-50)	**58.8**	**61.0**
Ours (ResNet-101)	58.3	60.7

V-COCO

5 Conclusion

5. Conclusion

Result

- This paper introduces a QPIC that performs the task of predicting HOI using transformer-based DETR.

- It overcomes the limitations of existing single-stage and two-stage methods by using the attention mechanism.

- High performance on HICO-DET and V-COCO, benchmark datasets for HOI task

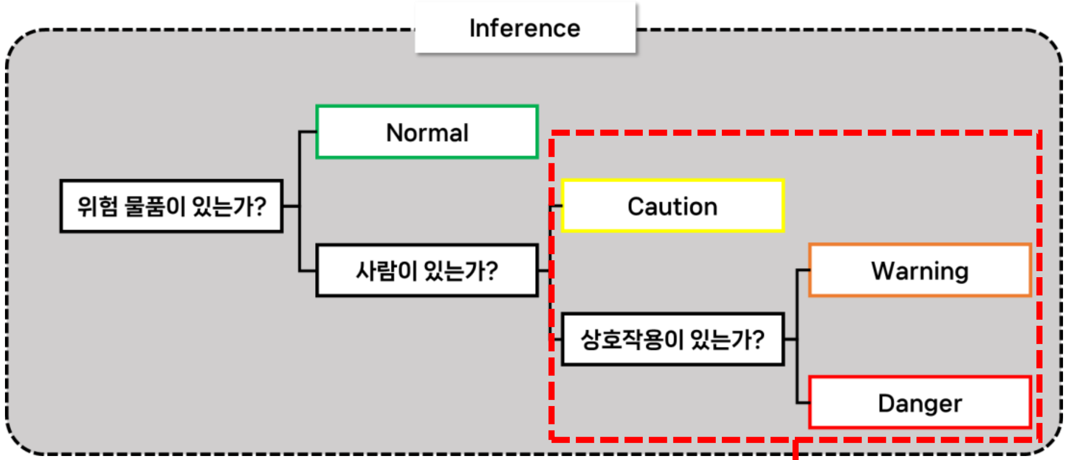
- Provides simple and intuitive detection heads

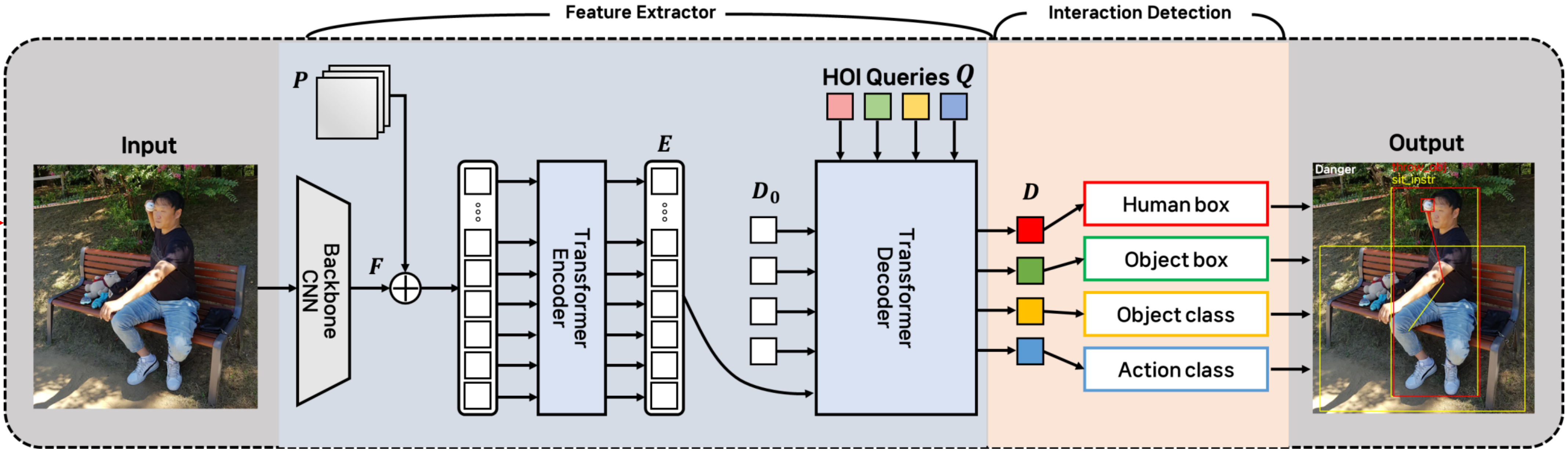
5. Conclusion

How to apply?

[기체 내 위험 상황 정의]

1. Normal : 위험 물품 x
2. Caution : 위험 물품o, 사람 x
3. Warning : 위험 물품o, 사람o, 상호작용 x
4. Danger : 위험 물품o, 사람o, 상호작용 o





Thanks