南京航空航天大學
Nanjing University of Aeronautics and Astronautics

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

Bumsoo Kim[1,2] Junhyun Lee[2] Jaewoo Kang[2] Eun-Sol Kim[1,†] Hyunwoo J. Kim[2,†]
[1]Kakao Brain [2]Korea University

CVPR 2021

Human-Object Interaction（HOI）

(a) (b) (c)

(a) There can be many possible objects (green boxes) interacting with a detected person (blue box)

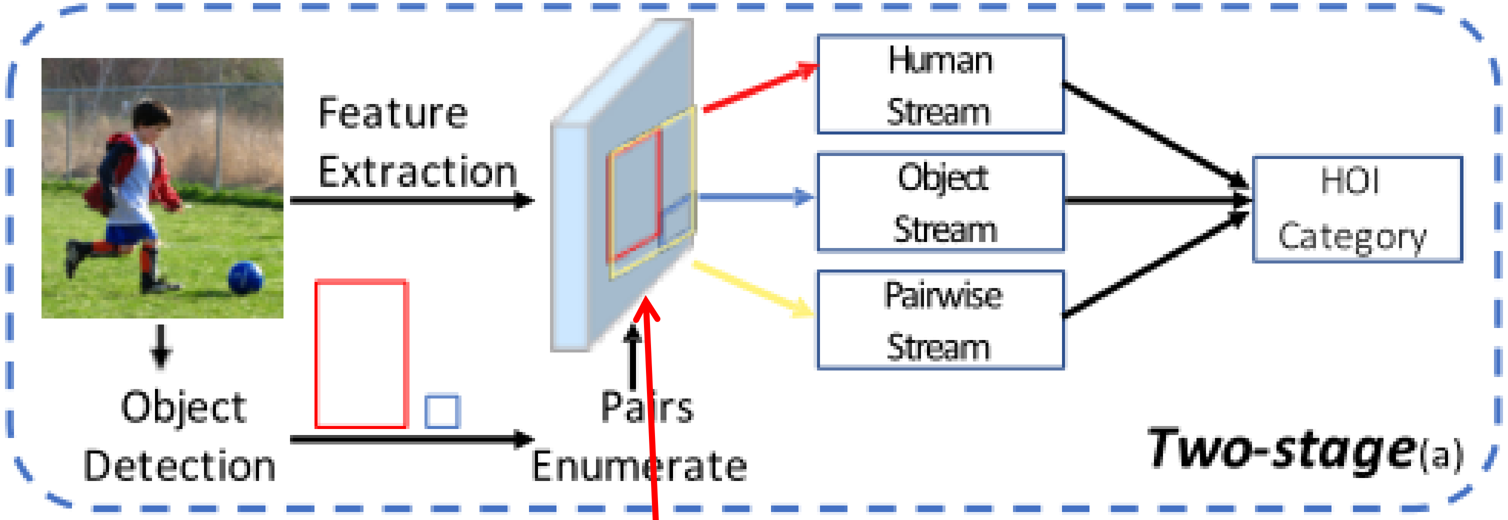
(b) A <human, object, action> triplet was detected. <human, knife, cut>

(c) Another predicted action (stand) <human, , stand>

Existing methods

1. sequential HOI detectors(two stage)

2. parallel HOI detectors(one stage)

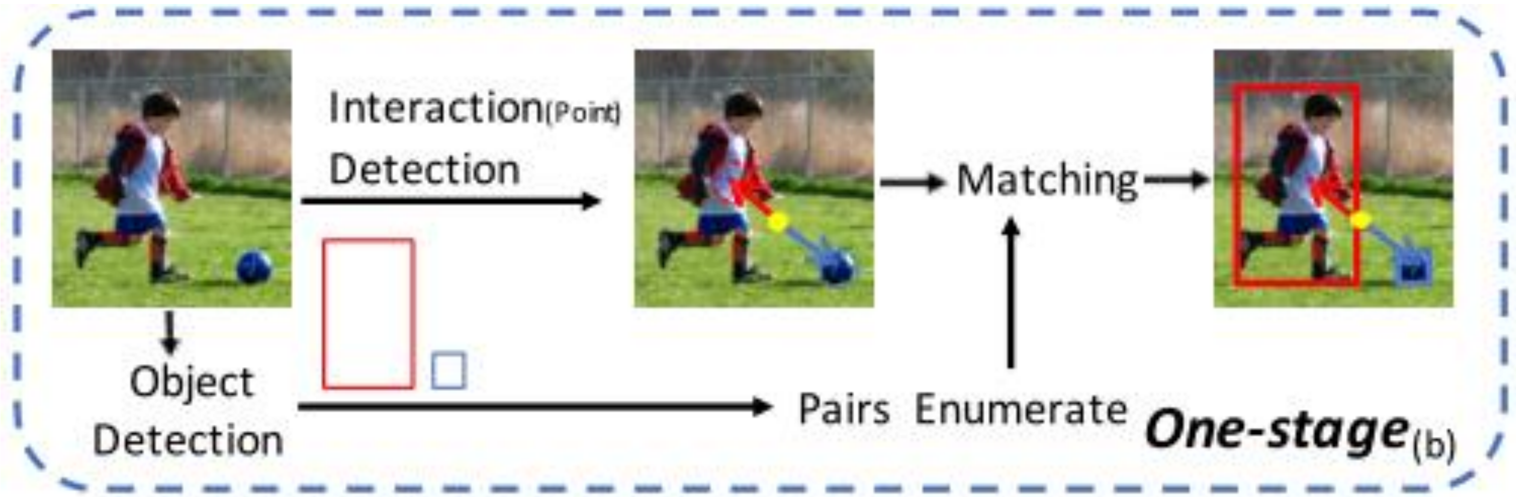


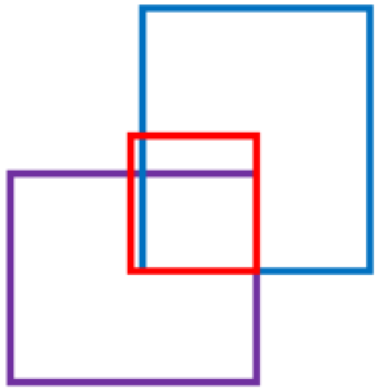
separate neural network

time-consuming、 computationally expensive

Existing methods

interaction point





interaction box

Since they can be parallelized with existing object detectors, they feature fast inference time.
However, these works are limited in that they require a hand-crafted postprocessing stage to associate the localized interactions with object detection results.

Abstract of HOTR

<human, object, interaction>

considering the inherent semantic relationships between the triplets in an end-to-end manner

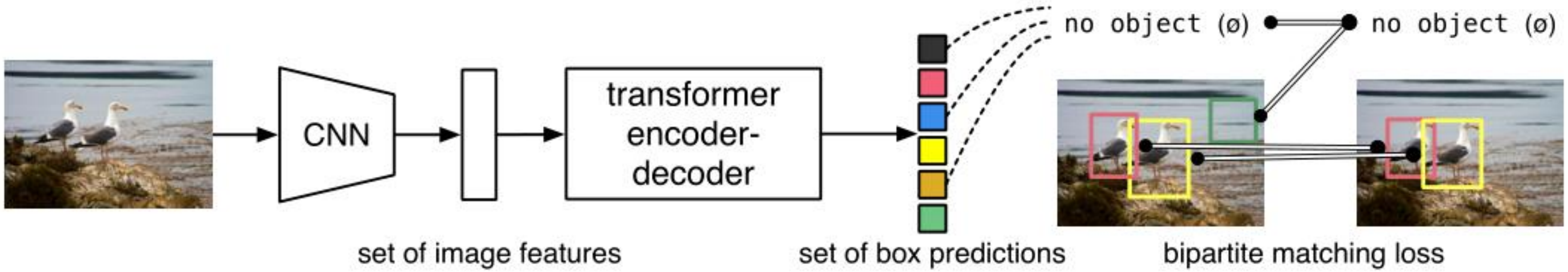
formulate HOI detection as set prediction

parallelly predicts a set of object detection and associates the human and object of the interaction, while the self-attention in transformers models the relationships between the interactions.

Object Detection with Transformers--DETR

Object Detection as Set Prediction: the transformer encoder-decoder structure in DETR transforms N positional embeddings to a set of N predictions for the object class and bounding box.

N queries → N predictions N predictions → N ground truth

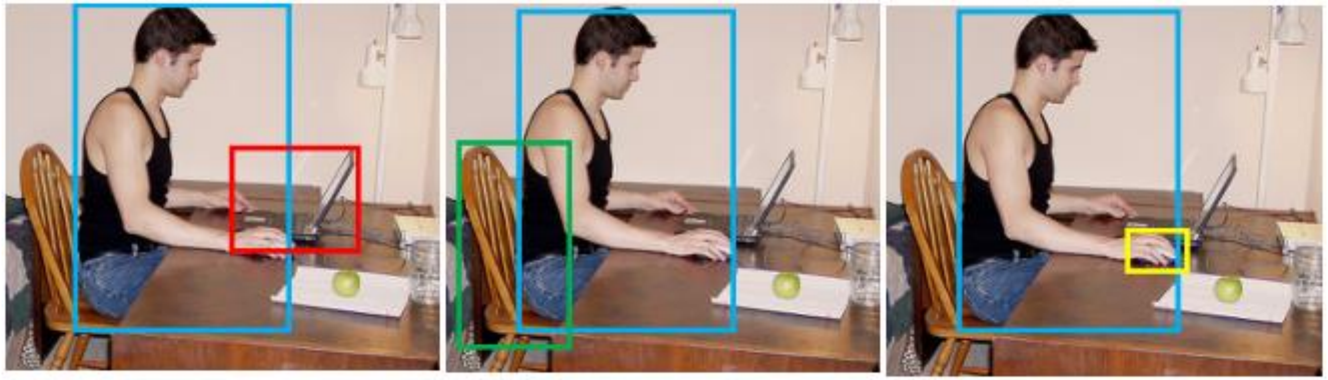


Similar to object detection, **HOI detection** can be defined as a **set prediction problem** where each prediction includes the localization of a human region (i.e., subject of the interaction), an object region (i.e., target of the interaction) and multi-label classification of the interaction types.

How to modify?

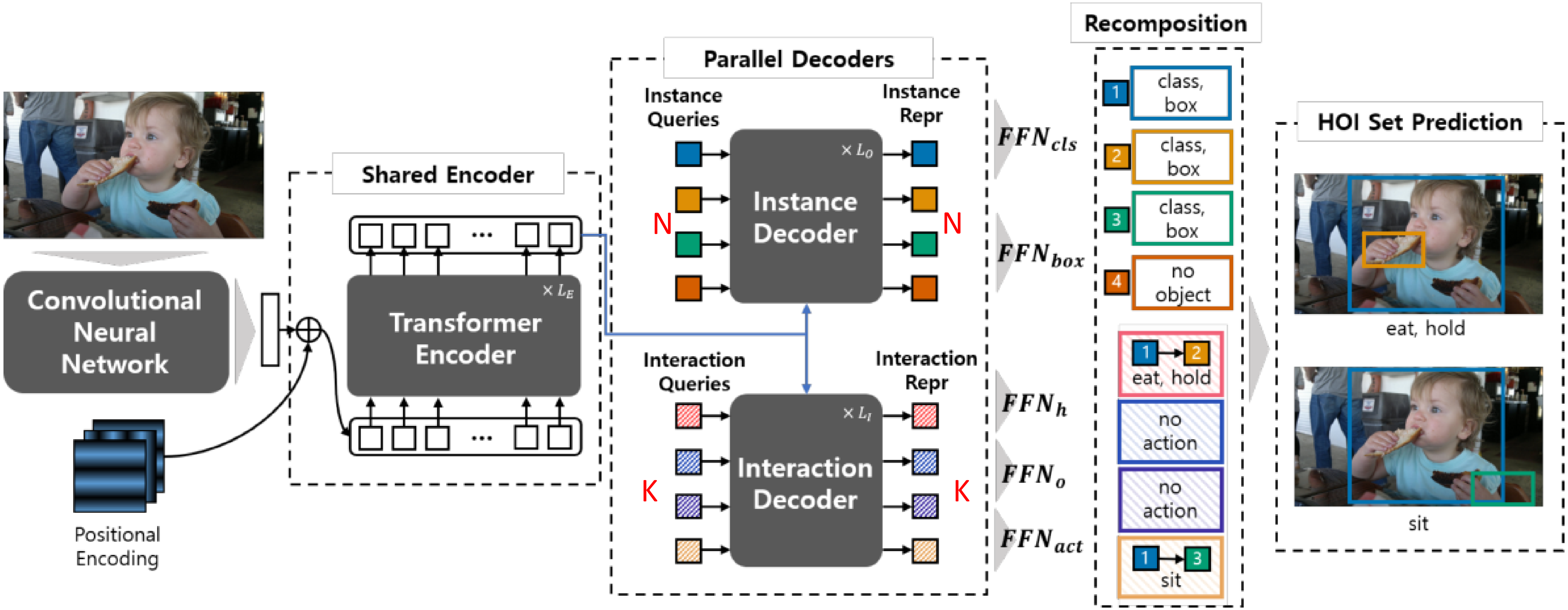
One straightforward extension is to modify the **MLP heads** of DETR to transform each **positional embedding** to predict a **human box**, **object box**, and **action classification**.

However, this architecture poses a problem where the localization for the same object needs to be redundantly predicted with multiple positional embeddings.



- HO pointers

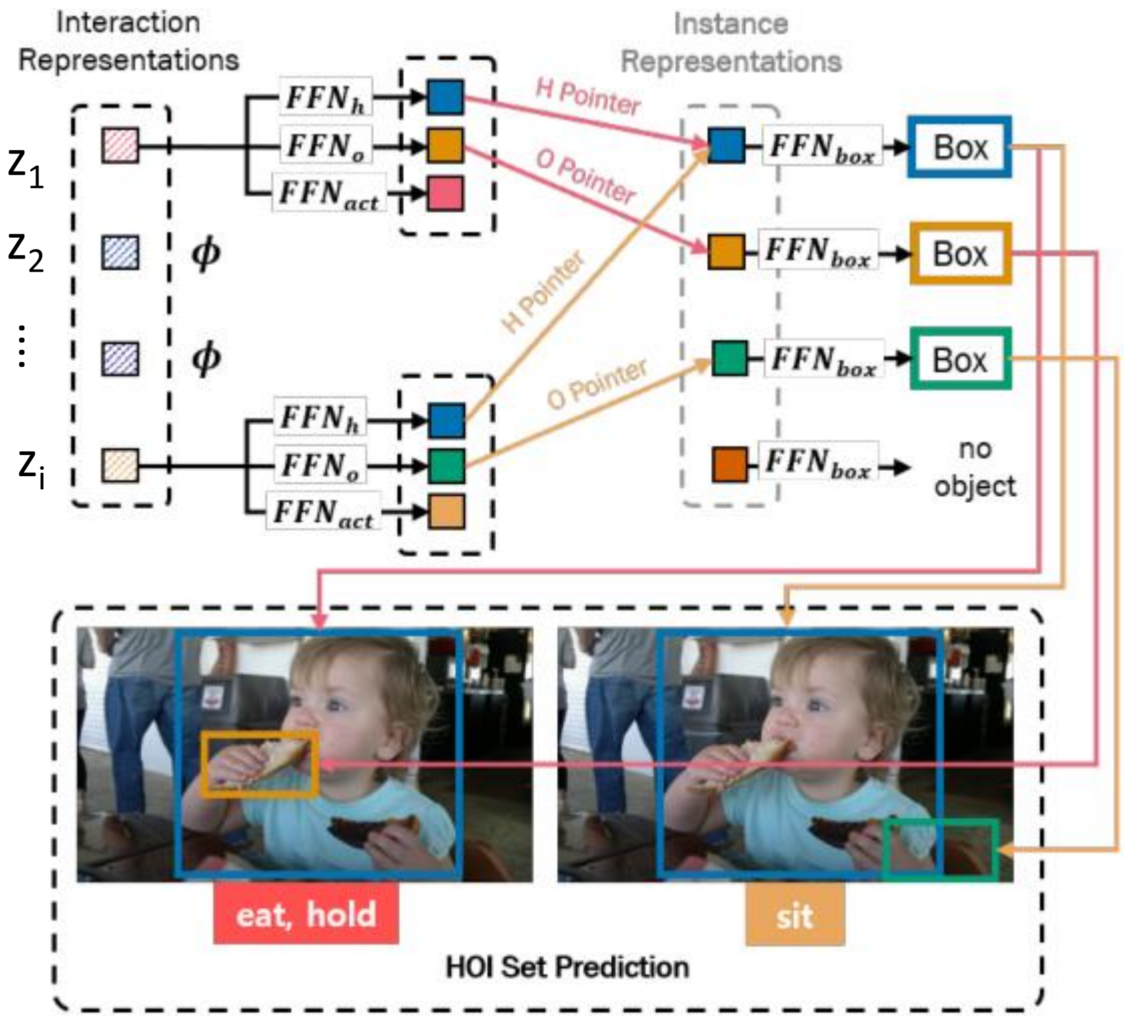
Framework



The architecture features a transformer encoder-decoder structure with a shared encoder and two parallel decoders(i.e., instance decoder and interaction decoder).

The results of the two decoders are associated with using HO Pointers to generate final HOI triplets.

HO Pointers



$$v_i^h = \mathrm{FFN}_h(z_i)$$

$$v_i^o = \mathrm{FFN}_o(z_i)$$

- indices of the instance representations with the highest similarity scores：

$$\hat{c}_i^h = \underset{j}{\operatorname{argmax}} \left(\mathrm{sim}(v_i^h, \mu_j)\right)$$

$$\hat{c}_i^o = \underset{j}{\operatorname{argmax}} \left(\mathrm{sim}(v_i^o, \mu_j)\right)$$

μ_j is the j-th instance representation

$$\mathrm{sim}(u, v) = u^\top v / \|u\| \|v\|$$

Recomposition for HOI Set Prediction

μ: N instance representations
z: K interaction representations
$\hat{c}^h$ and $\hat{c}^o$: their HO Pointers

$$\hat{b}_i^h = \mathrm{FFN}_{\text{box}}(\mu_{\hat{c}_i^h}) \in \mathbb{R}^4$$
$$\hat{b}_i^o = \mathrm{FFN}_{\text{box}}(\mu_{\hat{c}_i^o}) \in \mathbb{R}^4$$
$$\hat{a}_i = \mathrm{FFN}_{\text{act}}(z_i) \in \mathbb{R}^\gamma.$$

The final HOI prediction by the HOTR is the set of K triplets, $\{\langle \hat{b}_i^h, \hat{b}_i^o, \hat{a}_i \rangle\}_{i=1}^K$.

Training HOTR

1. Introduce the cost matrix of Hungarian Matching for unique matching between the ground-truth HOI triplets and HOI set predictions obtained by recomposition.

2. Using the matching result, defining the loss for HO Pointers and the final training loss.

Hungarian Matching for HOI Detection

Let **Y** denote the set of ground truth HOI triplets and $\hat{\mathcal{Y}} = \{\hat{y}_i\}_{i=1}^{K}$ as the set of K predictions.

a permutation of K elements with the lowest cost:

$$\hat{\sigma} = \underset{\sigma \in \mathfrak{S}_K}{\operatorname{argmin}} \sum_{i}^{K} \mathcal{C}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

$$\begin{aligned}\mathcal{C}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = &-\alpha \cdot \mathbb{1}_{\{a_i \neq \varnothing\}} \hat{P}^h[\sigma(i), c_i^h] \\ &-\beta \cdot \mathbb{1}_{\{a_i \neq \varnothing\}} \hat{P}^o[\sigma(i), c_i^o] \\ &+\mathbb{1}_{\{a_i \neq \varnothing\}} \mathcal{L}_{\text{act}}(a_i, \hat{a}_{\sigma(i)})\end{aligned}$$

$$\mathcal{L}_{\text{act}}(a_i, \hat{a}_{\sigma(i)}) = \text{BCELoss}(a_i, \hat{a}_{\sigma(i)})$$

$$\boxed{\begin{aligned}\mu' &= \mu/\|\mu\| \\ M &= [\mu'_1 \dots \mu'_N] \\ \hat{P}^h &= \|_{i=1}^{K} \text{softmax}((\bar{v}_i^h)^T M)\end{aligned}}$$

$\hat{P}[i, j]$ denotes the element at i-th row and j-th column.

Final Set Prediction Loss for HOTR

compute the Hungarian loss for all pairs matched:

$$\mathcal{L}_{\mathrm{H}} = \sum_{i=1}^{K} \left[\mathcal{L}_{\mathrm{loc}}(\mathrm{c}_i^h, \mathrm{c}_i^o, z_{\sigma(i)}) + \mathcal{L}_{\mathrm{act}}(a_i, \hat{a}_{\sigma(i)}) \right]$$

$$\mathcal{L}_{\mathrm{loc}} = -\log \frac{\exp(\mathrm{sim}(\mathrm{FFN}_h(z_{\sigma(i)}), \mu_{c_i^h})/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(\mathrm{FFN}_h(z_{\sigma(i)}), \mu_k)/\tau)}$$

$$-\log \frac{\exp(\mathrm{sim}(\mathrm{FFN}_o(z_{\sigma(i)}), \mu_{c_i^o}/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(\mathrm{FFN}_o(z_{\sigma(i)}), \mu_k)/\tau)}$$

where τ is the temperature that controls the smoothness of the loss function.

Experiment

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

Dataset: VCOCO

Method	Backbone	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Models with external features			
TIN (RP_DC_D) [18]	R50	47.8	
Verb Embedding [31]	R50	45.9	
RPNN [33]	R50	-	47.5
PMFNet [27]	R50-FPN	52.0	
PastaNet [17]	R50-FPN	51.0	57.5
PD-Net [32]	R50	52.0	-
ACP [13]	R152	53.0	
FCMNet [20]	R50	53.1	-
ConsNet [21]	R50-FPN	53.2	-
Sequential HOI Detectors			
VSRL [8]	R50-FPN	31.8	-
InteractNet [6]	R50-FPN	40.0	48.0
BAR-CNN [14]	R50-FPN	43.6	-
GPNN [24]	R152	44.0	-
iCAN [5]	R50	45.3	52.4
TIN (RC_D) [18]	R50	43.2	-
DCA [29]	R50	47.3	-
VSGNet [26]	R152	51.8	57.0
VCL [10]	R50-FPN	48.3	
DRG [4]	R50-FPN	51.0	
IDN [16]	R50	53.3	60.3
Parallel HOI Detectors			
IPNet [30]	HG104	51.0	-
UnionDet [12]	R50-FPN	47.5	56.2
Ours	R50	**55.2**	**64.4**

Table 1. Comparison of performance on V-COCO test set. $AP^{\#1}_{role}$, $AP^{\#2}_{role}$ denotes the performance under Scenario1 and Scenario2 in V-COCO, respectively.

Dataset: HICO-DET

Method	Detector	Backbone	Feature	Default Full	Default Rare	Default Non Rare
Sequential HOI Detectors						
InteractNet [6]	COCO	R50-FPN	A	9.94	7.16	10.77
GPNN [24]	COCO	R101	A	13.11	9.41	14.23
iCAN [5]	COCO	R50	A+S	14.84	10.45	16.15
DCA [29]	COCO	R50	A+S	16.24	11.16	17.75
TIN [18]	COCO	R50	A+S+P	17.03	13.42	18.11
RPNN [33]	COCO	R50	A+P	17.35	12.78	18.71
PMFNet [27]	COCO	R50-FPN	A+S+P	17.46	15.65	18.00
No-Frills HOI [9]	COCO	R152	A+S+P	17.18	12.17	18.68
DRG [4]	COCO	R50-FPN	A+S+L	19.26	17.74	19.71
VCL [10]	COCO	R50	A+S	19.43	16.55	20.29
VSGNet [26]	COCO	R152	A+S	19.80	16.05	20.91
FCMNet [20]	COCO	R50	A+S+P	20.41	17.34	21.56
ACP [13]	COCO	R152	A+S+P	20.59	15.92	21.98
PD-Net [32]	COCO	R50	A+S+P+L	20.81	15.90	22.28
DJ-RN [15]	COCO	R50	A+S+V	21.34	18.53	22.18
ConsNet [21]	COCO	R50-FPN	A+S+L	22.15	17.12	23.65
PastaNet [17]	COCO	R50	A+S+P+L	22.65	21.17	23.09
IDN [16]	COCO	R50	A+S	23.36	**22.47**	23.63
Functional Gen. [1]	HICO-DET	R101	A+S+L	21.96	16.43	23.62
TIN [18]	HICO-DET	R50	A+S+P	22.90	14.97	25.26
VCL [10]	HICO-DET	R50	A+S	23.63	17.21	25.55
ConsNet [21]	HICO-DET	R50-FPN	A+S+L	24.39	17.10	26.56
DRG [4]	HICO-DET	R50-FPN	A+S	24.53	19.47	26.04
IDN [16]	HICO-DET	R50	A+S	24.58	**20.33**	25.86
Parallel HOI Detectors						
UnionDet [12]	COCO	R50-FPN	A	14.25	10.23	15.46
IPNet [30]	COCO	R50-FPN	A	19.56	12.79	21.58
Ours	COCO	R50	A	**23.46**	16.21	**25.62**
UnionDet [12]	HICO-DET	R50-FPN	A	17.58	11.72	19.33
PPDM [19]	HICO-DET	HG104	A	21.10	14.46	23.09
Ours	HICO-DET	R50	A	**25.10**	17.34	**27.42**

Table 2. Performance comparison in HICO-DET. The Detector column is denoted as 'COCO' for the models that freeze the object detectors with the weights pre-trained in MS-COCO and 'HICO-DET' if the object detector is fine-tuned with the HICO-DET train set. The each letter in Feature column stands for A: Appearance (Visual features), S: Interaction Patterns (Spatial Correlations [5]), P: Pose Estimation, L: Linguistic Priors, V: Volume [15].

Experiment

Method	$AP^{\#1}_{\text{role}}$	Default(Full)
HOTR	**55.2**	**23.5**
w/o HO Pointers	39.3	17.2
w/o Shared Encoders	33.9	14.5
w/o Interactiveness Suppression	52.2	22.0

Table 3. Ablation Study on both V-COCO test set (scenario 1, $AP^{\#1}_{\text{role}}$) and HICO-DET test set (Default, Full setting without fine-tuning the object detector)

Thanks