Exploring Predicate Visual Context in Detecting Human–Object Interactions

Frederic Z. Zhang[1†] Yuhui Yuan[2] Dylan Campbell[1] Zhuoyao Zhong[2] Stephen Gould[1]
[1]The Australian National University [2]Microsoft Research Asia
frederic.zhang@anu.edu.au yuhui.yuan@microsoft.com

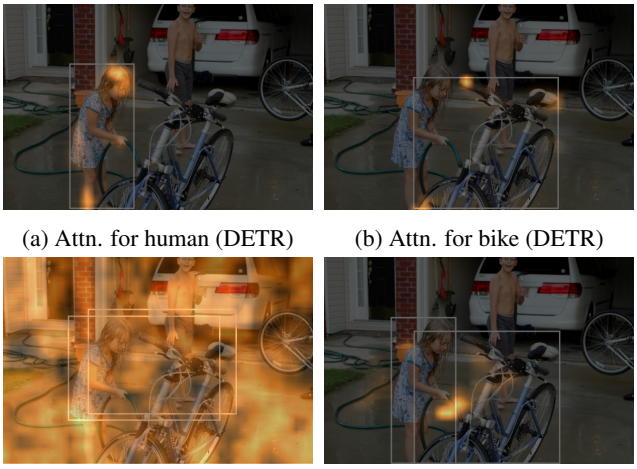
https://github.com/fredzzhang/pvic

Abstract

*Recently, the DETR framework has emerged as the dominant approach for human–object interaction (HOI) research. In particular, two-stage transformer-based HOI detectors are amongst the most performant and training-efficient approaches. However, these often condition HOI classification on object features that lack fine-grained contextual information, eschewing pose and orientation information in favour of visual cues about object identity and box extremities. This naturally hinders the recognition of complex or ambiguous interactions. In this work, we study these issues through visualisations and carefully designed experiments. Accordingly, we investigate how best to re-introduce image features via cross-attention. With an improved query design, extensive exploration of keys and values, and box pair positional embeddings as spatial guidance, our model with enhanced **p**redicate **vi**sual **c**ontext (PViC) outperforms state-of-the-art methods on the HICO-DET and V-COCO benchmarks, while maintaining low training cost.*

1. Introduction

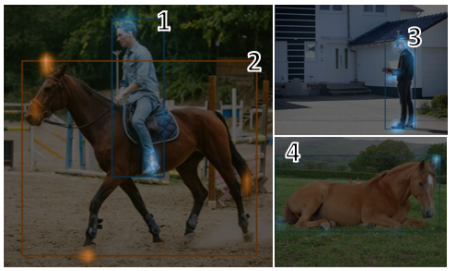
Detecting human–object interactions (HOI) is the task of localising and recognising interactive human–object pairs. It extends the detection of objects to include their relationships and facilitates a deeper understanding of visual scenes. Recent developments in the detection of human–object interactions have largely adhered to the encoder–decoder style introduced by the detection transformers (DETR) [**?**], where learnable queries are randomly initialised with Gaussian noise, and progressively decoded into the desired *human–predicate–object* triplets. Such one-stage detectors [**?**, **?**, **?**, **?**, **?**, **?**] require pre-trained DETR weights for initialisation to facilitate stable convergence. As we will demonstrate empirically, the pre-trained encoder features have overfitted to object cues and lack the necessary information for recognising human–object interactions. This means that the transformer encoder weights need to change significantly to produce discriminative features for such tasks. Together with the need to repurpose the decoder to detect HOI triplets rather than unary objects, this results in long training schedules that often amount to hundreds of GPU hours. On the other hand, two-stage detectors adopt a different methodology, wherein an object detector is fine-tuned and then frozen. These approaches focus on the extraction and exploitation of the rich information residing in the frozen detector. Naturally, two-stage detectors require significantly less time and resources to train, facilitating more model analysis and experimentation.



(a) Attn. for human (DETR) (b) Attn. for bike (DETR)
(c) Attn. for triplet (QPIC) (d) Attn. for predicate (Ours)

Figure 1. Visual context for the (a, b) two-stage HOI detector UPT [**?**], (c) one-stage HOI detector QPIC [**?**] and (d) our method. Cross-attention weights from the last decoder layer are used for visualization. UPT uses coarse object features that favor visual cues about object identity and box extremities. QPIC fails to detect the triplet *person-washing-bike* as it struggles to locate the relevant visual context (person, bike, and the water hose). The box pair with the highest IoUs against the ground truth is selected for display. Our two-stage method with pre-detected objects successfully recognises the predicate *washing* as it pinpoints the location of the image region containing the water hose.

[†]Work done at Microsoft Research Asia.



(a) Object feature replacement for predicate *riding*.

(b) Object feature replacement for predicate *throwing*.

Figure 2. Object features from a frozen object detector (DETR [**?**]) are extracted from image regions that are indicative of the object identity and are near the bounding box extremities, as shown by the cross-attention weights overlaid on each detected object. These features often lack the fine-grained information to recognise HOIs. As a result, replacing object features with those from an object in a different (a, b) pose, orientation or even (a) identity does not impact the classification score significantly. Experiments are conducted on UPT [**?**], where the spatial configuration for pair (1, 2) is used for all pairs in each set of images while the object features are replaced.

Current state-of-the-art two-stage detector UPT [**?**] employs a fine-tuned DETR detector, and performs self-attention on unary (object) and pairwise (human–object) tokens. Despite its overall high performance and low cost, it only utilises object features from the frozen detector, complemented by hand-crafted spatial features, to construct the final representations. As we show in Figures 1a and 1b, these frozen features are obtained by attending to image regions indicative of object identity and box extremities, thus lacking the necessary information for recognising HOIs.

In Figure 2, we study the impacts of this lack of information by replacing the object features with those of a different object. For predicates with a distinct spatial pattern, such as *riding*, we observe that the predicted scores do not change significantly when object features are swapped, suggesting that the spatial information dominates the visual information, as shown in Figure 2a. Even when the replacement features come from objects of a different identity, UPT still gives confident predictions for the same predicate, such as (4, 2) *horse–riding–horse*. Yet, the majority of predicates do not exhibit a prominent spatial pattern, e.g. *throwing*. In such cases, visual context plays a crucial role. Naturally, replacing object features amounts to a more tangible impact (Figure 2b). However, the score drop when replacing the human features with those of a non-interactive person is still not significant, indicating that such features do not contain enough visual cues to differentiate interactiveness. As we will show in Figure 3, failure cases of UPT often require much richer visual context. In particular, we identify two types of context that coarse object features lack: fine-grained information about the subject or object, such as human pose, and information about other relevant context in the scene, such as another object involved in the interaction.

To address the aforementioned issues, we investigate how to enrich the contextual cues for human–object pair representations. Our contribution is twofold. We conduct thorough analysis with abundant visualisations to characterise the two types of visual contexts lacking in current two-stage models and the damage this causes. Accordingly, we develop a superior two-stage detector with a lightweight decoder, where we improve the query design with a more streamlined architecture, explore various choices and compositions of keys/values and introduce positional embeddings tailored for bounding box pairs. In particular, we demonstrate that the positional embeddings function as spatial guidance in cross-attention, and shed light on this mechanism with rich visualisations.

2. Related Works

There is a large body of works [**?**, **?**, **?**, **?**] centred around adapting the detection transformer [**?**] to one-stage HOI detectors. Since Tamura et al. [**?**] established a strong baseline, the focus has shifted to improving the architecture design. Zhang et al. [**?**] proposed to partially decouple the feature representation of humans and objects from that of the predicates. Qu et al. [**?**] investigated ways to better utilise the ground truth with data distillation. Tu et al. [**?**] and Kim et al. [**?**] explored multiscale backbone features by either exploiting irregular window attention or extending Deformable DETR [**?**] to HOI detection. Last, Wu et al. [**?**] demonstrated the value of human pose by applying body-part masks in transformer cross-attention.

Two-stage detection has received much less attention in comparison. Graph-based methods [**?**, **?**, **?**] were the state of the art for an extended period of time. Since the advent of transformer-based approaches, much of the focus has been shifted to one-stage detection. Recently, Zhang et al. [**?**] demonstrated that self-attention can be repeatedly applied to unary objects and human–object pairs, achieving complementary effects. However, the lack of contextual information is its major weakness. A concurrent work [**?**] addressed this by integrating hand-coded HOI structures into transformer cross-attention. Nonetheless, this introduces even

(a) Feeding. UPT: 0.07 (b) Tying. UPT: 0.12 (c) Washing. UPT: 0.12 (d) Cutting. UPT: 0.10

(e) Feeding. Ours: **0.75** (f) Tying. Ours: **0.52** (g) Washing. Ours: **0.66** (h) Cutting. Ours: **0.61**

(i) Feeding. Ours (masked): 0.09 (j) Tying. Ours (masked): 0.14 (k) Washing. Ours (masked): 0.21 (l) Cutting. Ours (masked): 0.08

Figure 3. Existing two-stage HOI detectors (e.g., UPT [**?**]) lack relevant visual context, including (a, b) fine-grained information about the subject or object, such as human pose, and (c, d) other relevant contextual information in the scene, such as another object involved in the interaction. The predicted score for each example is listed in the caption. UPT (first row) uses frozen object features which often pool information from the box boundary since this aids localisation. Consequently, such features do not cover other aspects of the object and are not discriminative enough to recognise complex human–object interactions. Our method (second row) solves such failure cases with spatially guided cross-attention, pinpointing the image regions corresponding to the relevant body parts or the additional object besides the human–object pair. To demonstrate that these regions are indeed highly relevant to the prediction score, we mask out those image regions with the highest attention weights (third row), and observe a significant drop in prediction scores.

more hand-crafted elements into two-stage detection. Seeking a more streamlined model design, we show that a dedicated query positional embedding yields better performance and more interpretable visualisations.

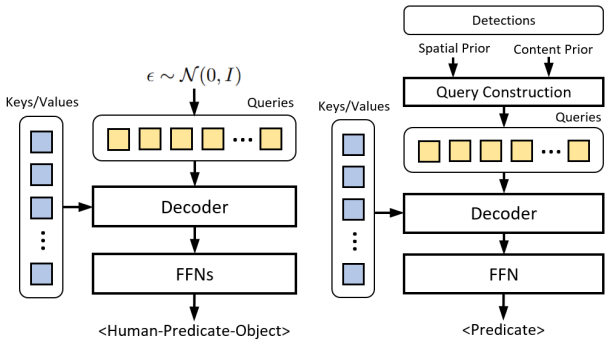
In addition, there have been some works focusing on other aspects of HOI research. Specifically, Wang et al. [**?**] studied the object bias and explored ways to mitigate it. Yuan et al. [**?**] conducted contrastive language–image pre-training for HOI representations and demonstrated its effectiveness. Liao et al. [**?**] explored data distillation from CLIP [**?**] features and showed competitive performance.

3. Spatially Guided Cross-Attention

The underpinning of query-based detection systems is the transformer cross-attention mechanism [**?**], which acts as a form of soft RoI pooling where the weights are computed dynamically from the data. Stacking cross-attention layers allows the queries to aggregate useful information from the keys/values (image features) gradually. In the detection transformer, queries are randomly initialised with Gaussian noise and learn to represent spatial priors (box centre positions, widths and heights) [**?**] as training progresses. We refer to such queries as *implicit queries* (Figure 4a), commonly used in one-stage HOI detectors. For their two-stage counterparts [**?**, **?**], thanks to the rich information in the detections, there is no need for such learned queries. Instead, the queries are explicit human–object pair representations, injected with spatial and content priors. We refer to them as *explicit queries* (Figure 4b).

3.1. Explicit Queries

Prior to the query construction, we filter the detected objects by their scores and perform self-attention to refine the object features. As Zhang et al. [**?**] pointed out, such self-attention promotes information flow between the interactive objects and helps increase scores of positive examples. Based on the observation that interactive instances tend to appear close together in an image, we apply positional embeddings for bounding boxes, which encourages attention



(a) Implicit Queries (b) Explicit Queries

Figure 4. Implicit queries (a) for one-stage detectors and explicit queries (b) for two-stage detectors.

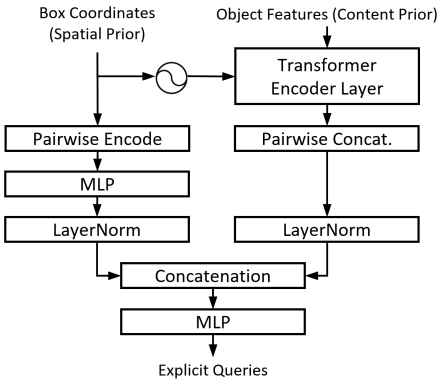


Figure 5. An illustration of the construction of explicit queries.

between near objects. Insights on this design will be detailed in the next section. Formally, denote the mapping from a scalar to a sinusoidal embedding by $\phi : \mathbb{R} \rightarrow \mathbb{R}^d$,

$$\phi(x)_{2i} = \sin\left(\frac{x}{\tau^{2i/d}}\right), \ \phi(x)_{2i-1} = \cos\left(\frac{x}{\tau^{2i/d}}\right), \quad (1)$$

where $i = 1, \ldots, d/2$ and τ is a temperature parameter. A bounding box can be encoded by concatenating the sinusoidal embeddings of the centre coordinates, width and height, which are then applied via element-wise sum as a convention [**?**, **?**]. We show in the experiment section that with the positional embeddings, vanilla self-attention achieves better performance than the custom layer in UPT.

To construct explicit queries, we enumerate all human–object pairs. For each pair, the representation is obtained by fusing the concatenated object features and their spatial representations, following previous practice [**?**, **?**]. In addition, we apply LayerNorm [**?**] to both modalities before fusion. This greatly stabilises the training process and prevents numeric overflow, which was previously resolved by using large batch sizes. The full process of query construction is illustrated in Figure 5.

3.2. Positional Embeddings as Guidance

Cross-Attention. Even though the explicit query representation of the human–object pair already contains spatial priors, positional embeddings are still critical since they function as spatial biases on the attention weights. This is particularly important in the case of cross-attention. To shed light on its impact, let us denote the keys and queries as $\mathbf{k}_c$ and $\mathbf{q}_c$, and their respective positional embeddings as $\mathbf{k}_p$ and $\mathbf{q}_p$. For simplicity, let us omit the linear transformations and the normalisation. Dot-product attention is computed as

$$(\mathbf{k}_c + \mathbf{k}_p)^\mathsf{T}(\mathbf{q}_c + \mathbf{q}_p) = {\mathbf{k}_c}^\mathsf{T}\mathbf{q}_c + \ldots + {\mathbf{k}_p}^\mathsf{T}\mathbf{q}_p. \quad (2)$$

Intuitively, first term on the RHS measures the similarity between the content features of the keys (image features) and queries, while the last term measures that of the positional embeddings. More specifically, for an image token with normalised spatial indices (i, j) and a 2D point with normalised coordinate (x, y), the last term can be expanded as a simple sum of similarity between coordinates,

$${\mathbf{k}_p}^\mathsf{T}\mathbf{q}_p = \phi(i)^\mathsf{T}\phi(x) + \phi(j)^\mathsf{T}\phi(y). \quad (3)$$

As an advantage of explicit queries, the availability of box coordinates allows us to use box centres to construct positional embeddings, directly adding a bias to the attention map in the corresponding position. A weakness of the aforementioned positional embeddings is the lack of information on box dimensions. Although the subsequent linear transformations have the potential to shift and deform the dot-product attention, Liu et al. [**?**] showed that the positional embeddings can be modulated with box widths and heights, saving the network from learning the relevant transforms. For a bounding box $\mathbf{b} = [x, y, w, h]$, we follow their practice by using the normalised widths and heights as different temperature parameters in horizontal and vertical directions for the subsequent softmax normalisation, leading to a bias term on attention weights as below

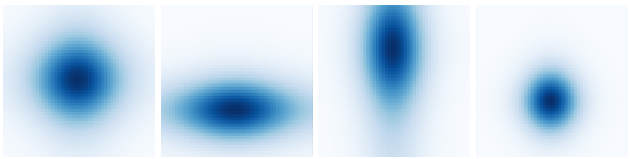
$${\mathbf{k}_p}^\mathsf{T}\mathbf{q}_p = \phi(i)^\mathsf{T}\phi(x)\frac{w_\text{ref}}{w} + \phi(j)^\mathsf{T}\phi(y)\frac{h_\text{ref}}{h}, \quad (4)$$

where w_ref and h_ref are reference values learned from the box features using a two-layer MLP, as follows

$$w_\text{ref}, h_\text{ref} = \sigma(\text{MLP}(\mathbf{f})), \quad (5)$$

with σ being the sigmoid function and $\mathbf{f}$ being the box appearance features obtained from the object detector.

To extend the usage to bounding box pairs, we concatenate the positional embeddings of the two box centres. The concatenation of two positional embeddings is equivalent to spatially summing the attention weights (Eq. 4) of two boxes prior to the softmax normalisation. We defer the mathematical details to the supplementary materials, but



(a) (0.5, 0.5) (b) (0.5, 0.7) (c) (0.5, 0.3) (d) (b, c)

Figure 6. Visualisation of the dot-product attention between the position embeddings of 2D points and those of image patches, without height and width modulation (a), with modulation (b, c). Point coordinates are listed in the captions. The spatially summed attention weights *after* softmax normalisation is shown in (d).

show visualisations in Figure 6. Furthermore, we show in the experiment section that for cross-attention, the two omitted cross-terms in Eq. 2 produce mostly noise, hampering the effectiveness of positional embeddings. Thus, we follow Meng et al. [**?**] to concatenate the keys and queries with their respective positional embeddings, essentially removing said terms. In addition, separate linear layers are employed for the content features (keys and queries) and positional embeddings to avoid undesired information flow, in the same spirit as removing the two cross-terms. We show visualisations on the impacts of the positional embeddings in the experiment section (Figure 8).

Self-Attention. For explicit queries, self-attention acts mainly as a form of suppression [**?**]. Interactive human–object pairs, which are often the most salient ones, suppress the non-interactive pairs via the attention mechanism. The positional embeddings, on the other hand, add an inductive bias such that box pairs in close proximity attend to each other more. While such inductive bias is intuitive in cross-attention, it does not reflect the way human–object pairs interact with each other. As we did not observe any improvement, we do not use positional embeddings in self-attention between human–object pairs. In contrast, self-attention amongst the unary objects does benefit from positional embeddings, because interactive objects tend to appear together, and often share an intersected area. Thus, such an inductive bias promotes attention between near instances and aids the training process.

3.3. Keys/Values

In one-stage methods, the encoder features serve as dedicated keys/values and are end-to-end trained. Two-stage methods, on the other hand, employ pre-trained object detectors, mostly with frozen weights to ensure the performance of the detector. Although an additional feature head can be used for refinement, the source of the keys/values is of utmost importance. We empirically found (see Section 4.2) that backbone ResNet [**?**] C5 features are the most informative, and a very lightweight feature head with window attention [**?**] improves the performance further, while higher feature resolutions and multiscale features do not introduce additional benefits.

3.4. Training and Inference

During training, we use the focal loss [**?**] on the predicted action logits following previous practice [**?**, **?**], where invalid actions for each object are masked out. During inference, we combine the object detection scores (s_h, s_o) and action prediction scores $(\mathbf{s}_a)$ using the geometric mean with hyperparameter $\lambda \in [0, 1]$ as follows

$$\mathbf{s} = (s_h s_o)^{1-\lambda} \mathbf{s}_a^{\lambda}. \tag{6}$$

4. Experiments

In this section, we first present a thorough ablation study by progressively building up the proposed model, demonstrating the impact of each design choice. We then compare our method against state-of-the-art models and show its superior performance, even against methods that perform data distillation on large pre-trained vision and language models. Last, to shed light on how spatial priors are used to guide cross-attention, we show visualisations of the attention weights for different terms in Eq. 2 and demonstrate why concatenated positional embeddings are superior.

Datasets: The primary dataset used for model design and validation is HICO-DET [**?**], which contains 37 633 training images and 9 546 test images. The dataset includes the same 80 object classes as in MS COCO [**?**], 117 action classes and 600 interaction classes. For legacy reasons we also report on V-COCO [**?**], a much smaller dataset with 2 533 training images, 2 867 validation images and 4 946 test images. The dataset has 24 action classes.

4.1. Implementation Details

We use fine-tuned DETR provided by Zhang et al. [**?**] and freeze the weights. In addition, we fine-tuned deformable DETR [**?**] with iterative box refinement and the two-stage options. During training, we adopt the same sampling scheme in UPT, by filtering detections with a threshold of 0.2 and sampling a minimum of 3 and a maximum of 15 human and object instances each. For focal loss, we use $\alpha = 0.5$ and $\gamma = 0.1$. The hyper-parameter λ in the geometric mean is set to be 0.26, which is simply a normalised value and has an equivalent effect as the setup in UPT. For the feature head, we use one encoder layer with window attention and a window size of 8×8. For the decoder, we use two layers. We apply the same data augmentation in previous works [**?**, **?**, **?**], including multiscale resizing, random cropping and random colour jittering. AdamW [**?**] is used as the optimiser, with both the learning rate and weight decay being 10^{-4}. Unless otherwise specified, all models are trained for 30 epochs, with a learning rate drop by a factor

Table 1. The mAP (×100) of model variants with different components of the decoder on the HICO-DET test set. Variant C is equivalent to UPT [**?**]. Results are averaged across three runs.

#	Decoder			C.A. src.	Default Setting (mAP)		
	Self	Cross	FFN		Full	Rare	N-rare
A				None	30.71	25.16	32.37
B			✓	None	30.98	25.36	32.62
C	✓		✓	None	31.47	25.98	33.11
D	✓	✓	✓	Encoder	31.51	26.12	33.13
E	✓	✓	✓	Backbone	**32.89**	**27.91**	**34.38**

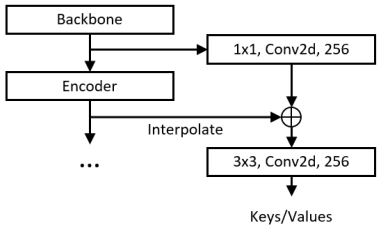


Figure 7. Illustration of the composition of backbone and encoder features. Encoder features are upsampled with bilinear interpolation when there is a resolution discrepancy.

Table 2. The mAP (×100) of model variants with different choices and compositions of features as cross-attention keys/values on the HICO-DET test set. Results are averaged across three runs.

#	Keys/Values	Feature head	Default Setting (mAP)		
			Full	Rare	N-rare
E1	Backbone C3	C4, C5	29.59	22.41	31.73
E2	Backbone C4	C5	30.80	25.34	32.43
E3	Backbone C5	None	**32.89**	**27.91**	**34.38**
F1	C3 + encoder	None	30.21	25.08	31.74
F2	C4 + encoder	None	31.27	25.09	33.12
F3	C5 + encoder	None	32.69	27.38	34.27
G1	FPN P3	None	32.44	28.19	33.71
G2	FPN P4	None	32.40	28.34	33.61
H1	Backbone C5	1× Self Attn.	33.50	29.80	**34.60**
H2	Backbone C5	2× Self Attn.	33.45	29.73	34.56
H3	Backbone C5	1× Win. Attn.	33.54	30.17	34.55
H4	Backbone C5	2× Win. Attn.	**33.57**	**30.32**	34.54

of 5 at the 20$^{\text{th}}$ epoch. Training is conducted on 8 Nvidia Tesla V100 GPUs, with a batch size of 16.

4.2. Ablation Study

We present the ablation study as a progressive build-up from the baseline model (variant A in Table 1), which directly feeds the explicit queries into a classifier. All subsequent model variants in this section along with the baseline use a fine-tuned DETR [**?**] with ResNet50 backbone [**?**]. In Table 1, we show that introducing cross-attention with encoder features as the keys/values leads to minimal improvement. This indicates that the frozen encoder features may have overfitted to object cues and therefore do not contain orthogonal information beneficial to the understanding of HOIs. The backbone features, on the other hand, contain more general contextual features and result in substantial performance improvement. Nevertheless, there is likely to be a certain degree of overfitting in the backbone, thus warranting investigation into earlier convolutional stages.

Table 3. The mAP (×100) of model variants with different query components and designs on HICO-DET test set. Results are averaged across three runs.

#	Self-Attn.	Modality Fusion		Default Setting (mAP)		
		Spatial	Content	Full	Rare	N-rare
H3	Modified	✓	✓	33.54	**30.17**	34.55
I1	Modified		✓	33.04	28.31	34.46
I2	None		✓	32.60	26.79	34.34
I3	None	✓		31.30	26.19	32.82
I4	None	✓	✓	32.87	29.20	34.27
J1	Vanilla	✓	✓	33.26	29.01	34.53
J2	Vanilla + pe	✓	✓	**33.59**	29.65	**34.76**

Table 4. The mAP (×100) of model variants with different positional embeddings and number of decoder layers on HICO-DET test set. Results are averaged across three runs.

#	Positional Embed.	#Dec.	Default Setting (mAP)		
			Full	Rare	N-rare
J2	None	1	33.59	29.65	34.76
K1	Standard, additive	1	33.43	**29.83**	34.50
K2	Standard, concat.	1	33.72	29.14	35.09
K3	Modulated, concat.	1	**33.91**	29.28	**35.29**
L1	Modulated, concat.	2	**34.18**	**31.09**	35.10
L2	Modulated, concat.	3	34.03	30.18	**35.18**
L3	Modulated, concat.	4	34.05	30.44	35.12

We present the relevant findings in Table 2. For fairness, when using C3 and C4 features (variants E1, E2), we added a feature head equivalent to the missing convolutional stages, and observed that C5 features still yield the best performance. We also explored the composition of the backbone and encoder features illustrated in Figure 7. The results (F variants) show that although the addition of encoder features benefits the lower-level backbone features, it does not introduce orthogonal information to C5 features. In addition, with the G variants, we train a feature pyramid network [**?**] to propagate the semantics in C5 features to lower levels. The results show that higher-resolution features, albeit helpful for object detection and segmentation [**?**, **?**], do not benefit the recognition of HOIs. We further investigated adding attention layers to refine the backbone features (H variants) and observed similar performance with self-attention [**?**] and window attention [**?**]. Due to the lower complexity, we use window attention in subsequent model variants. We do not observe significant performance increases with additional attention layers (H2, H4).

Table 5. Comparison of detection performance (mAP×100) on the HICO-DET [?] and V-COCO [?] test sets. We report results with the common DETR [?] detector and ResNet50 backbone, while showing the scalability of our method using the more advanced $\mathcal{H}$-DETR with Swin-L backbone. Best performance in each section is highlighted in bold.

		HICO-DET						V-COCO	
		Default Setting			Known Objects Setting				
Method	**Backbone**	Full	Rare	Non-rare	Full	Rare	Non-rare	AP^{S1}_{role}	AP^{S2}_{role}
InteractNet [?]	ResNet-50-FPN	9.94	7.16	10.77	-	-	-	40.0	-
iCAN [?]	ResNet-50	14.84	10.45	16.15	16.26	11.33	17.73	45.3	52.4
TIN [?]	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26	47.8	54.2
VSGNet [?]	ResNet-152	19.80	16.05	20.91	-	-	-	51.8	57.0
PPDM [?]	Hourglass-104	21.94	13.97	24.32	24.81	17.09	27.12	-	-
VCL [?]	ResNet-50	23.63	17.21	25.55	25.98	19.12	28.03	48.3	-
DRG [?]	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43	51.0	-
IDN [?]	ResNet-50	24.58	20.33	25.86	27.89	23.64	29.16	53.3	60.3
HOTR [?]	ResNet-50	25.10	17.34	27.42	-	-	-	55.2	64.4
FCL [?]	ResNet-50	25.27	20.57	26.67	27.71	22.34	28.93	52.4	-
HOI-Trans [?]	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57	52.9	-
AS-Net [?]	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14	53.9	-
SCG [?]	ResNet-50-FPN	29.26	24.61	30.65	32.87	27.89	34.35	54.2	60.9
QPIC [?]	ResNet-101	29.90	23.92	31.69	32.38	26.06	34.27	58.8	61.0
MSTR [?]	ResNet-50	31.17	25.31	32.92	34.02	28.82	35.57	62.0	65.2
CDN [?]	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38	**63.9**	65.9
UPT [?]	ResNet-101-DC5	32.62	28.62	33.81	36.08	31.41	37.47	61.3	**67.1**
RLIP [?]	ResNet-50	32.84	26.85	34.63	-	-	-	61.9	64.2
GEN-VLKT [?]	ResNet-50	**33.75**	**29.25**	**35.10**	**36.78**	**32.75**	**37.99**	62.4	64.5
PViC w/ DETR	ResNet-50	34.69	32.14	35.45	38.14	35.38	38.97	62.8	67.8
PViC w/ $\mathcal{H}$-DETR	Swin-L	**44.32**	**44.61**	**44.24**	**47.81**	**48.38**	**47.64**	**64.1**	**70.2**

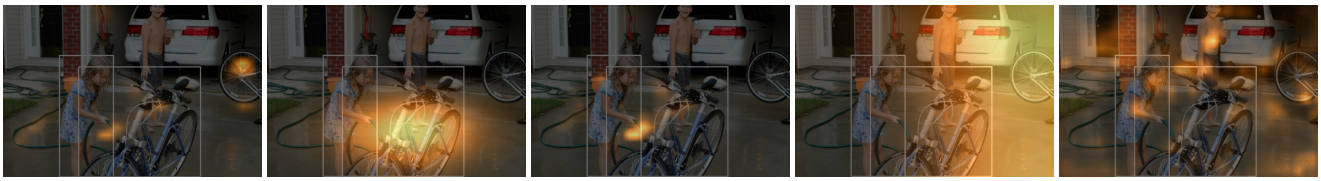
Next, we ablate the query components in Table 3. Using variant H3 as the reference, we show the impacts of the object self-attention and modality fusion of object features and the spatial features. Importantly, we show that a vanilla transformer encoder with box positional embeddings achieves comparable performance to the modified encoder in UPT, validating our removal of this custom layer.

Last, we demonstrate the effectiveness of modulated positional embeddings in cross-attention as well as the scaling of decoders in Table 4. Notably, using the additive positional embeddings (variant K1) does not help the model. This is due to the noise introduced by the two cross-terms in Eq. 2. Removing the cross-terms by concatenating the positional embeddings (K2) results in a slight improvement, while the modulated positional embeddings yields additional improvement. Similar to the feature head for keys and values, improvements brought by the decoder saturate after two layers, likely due to the use of frozen features. In summary, we observe a very significant improvement in the rare classes (5 mAP) between our model (variant L1) and the previous state-of-the-art UPT (variant C). This is in line with our understanding that contextual cues introduced with cross-attention greatly benefit the more ambiguous interactions, often the rare ones in the HICO-DET dataset.

4.3. Comparison with State-of-the-Art Methods

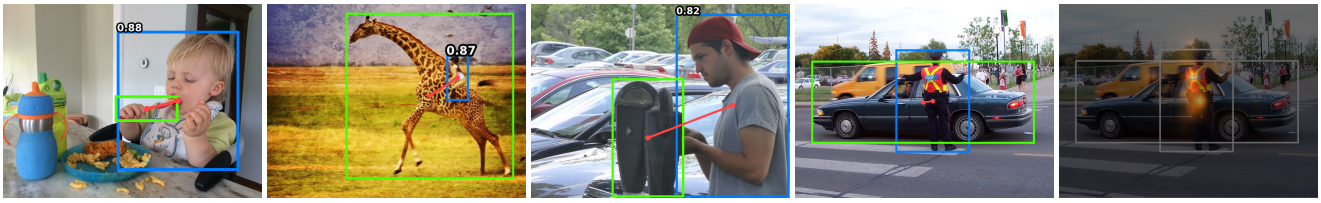
We report the performance of our method on HICO-DET [?] and V-COCO [?] datasets. For HICO-DET, evaluation is conducted under two different settings. The *default setting* is the primary setting under which different methods are being compared. The criteria for a successful detection extends that of the Pascal VOC challenge [?] to bounding box pairs. Specifically, both the human and object boxes need to have an intersection over union (IoU) larger than 0.5 with ground truth for the detected pair to be identified as positive. The *known objects setting* considers the sets of object types of the ground truth pairs in an image to be known, thus automatically removing detections where the object class is outside the set. For V-COCO, there are also two evaluation scenarios, differentiated by the protocol when handling occluded objects. Scenario 1 (S1) requires an empty box prediction for the detection to be considered a match, while scenario 2 (S2) neglects the occluded object and assumes it is always matched.

We report the performance of our model with two backbones to demonstrate its scalability. For the object detector, we use DETR [?] and the most recent $\mathcal{H}$-DETR [?], showing the detector-agnostic nature of our approach. As shown in Table 5, our method with the ResNet50 already outper-



(a) $\sigma(\mathbf{k}_c^\intercal \mathbf{q}_c)$ (b) $\sigma(\mathbf{k}_p^\intercal \mathbf{q}_p)$ (c) $\sigma(\mathbf{k}_c^\intercal \mathbf{q}_c + \mathbf{k}_p^\intercal \mathbf{q}_p)$ (d) $\sigma(\mathbf{k}_c^\intercal \mathbf{q}_p)$ (e) $\sigma(\mathbf{k}_p^\intercal \mathbf{q}_c)$

Figure 8. For the image in Figure 3c, we show visualisations of attention weights computed from the content features (a), positional embeddings (b) and the concatenated formulation (c). We also show the noisy attention weights computed from the two omitted terms from Eq. 2 in (d) and (e). Here we use σ to denote the softmax function and omit the scalar normalisation for brevity of exposition.



(a) *licking a fork* (b) *riding a giraffe* (c) *checking a parking meter* (d) *directing a car* (e) Visualised attention for (d)

Figure 9. Qualitative results (a, b, c) and failure case (d) on HICO-DET test set with fine-tuned DETR-R50 as the object detector.

forms the previous state-of-the-art two-stage detector UPT by 2.5 mAP, despite it using the heavier ResNet101 and a feature dilation. Furthermore, compared against GEN-VLKT, which distils features from a vision and language model trained on millions of images (CLIP [**?**]), our method achieves higher performance with the same ResNet50 backbone. With a stronger detector and backbone, i.e., $\mathcal{H}$-DETR, our method receives significant performance boost. This highlights one of the great advantages of two-stage detectors, that they can directly benefit from independent advances in object detection.

4.4. Positional Embeddings in Cross-Attention

To elucidate the mechanism of positional embeddings, we separate the attention weights for each term in Eq. 2 and visualise them in Figure 8. Starting with the content features of the keys and queries ($\mathbf{k}_c^\intercal \mathbf{q}_c$), we show in Figure 8a that this term only accounts for the visual similarity. Consequently, a distant object of a relevant class, i.e. the bike in the background, receives a substantial amount of attention. This can be corrected by the similarity term between positional embeddings, which places a spatial bias on locations of the human object pair as depicted in Figure 8b. When the positional embeddings are concatenated to the content features, the resultant attention weights become the sum of the two terms. As shown in Figure 8c, combination of these two terms results in high attention weights on the water hose, which is the key to recognising the interaction *washing a bike*. In addition, we show in Figures 8d and 8e that the two omitted cross-terms from Eq. 2 mostly generate noise, as the content features and positional embeddings are from very different feature spaces, justifying their removal.

4.5. Qualitative Results and Limitations

We show additional qualitative results in this section. In particular, our model performs well on several interactions with little training data, such as *licking a fork* (six training examples, 9a), *riding a giraffe* (two training examples, 9b) and *checking a parking meter* (36 training examples, 9c). We also show an example of missed detections in Figure 9d, due to a severe lack of training examples, one in this case. In addition, as there are other interaction classes involving the same predicate *directing* but with different objects and backgrounds, the model tends to fit towards other classes. Consequently, the model cannot locate the relevant visual context (Figure 9e), hand gesture in this case.

5. Conclusion

In this paper we analysed the visual features used in existing two-stage HOI detectors and concluded that their major weakness was a lack of relevant contextual information, since they were specialised to the localisation task. As such, we proposed an improved design by re-introducing image features into the human–object pair representation via cross-attention. To this end, we performed extensive experiments on the choices of keys/values and introduced box pair positional embeddings as spatial guidance, and visualised the impacts of the attention mechanism. Compared to previous two-stage approaches, we streamlined and simplified the architecture, reducing the need for custom components. Our method achieves state-of-the-art performance on the relevant benchmarks, with particular improvements where fine-grained visual features, like human pose, and additional context, like another object involved in the interaction, are relevant to the classification.

A. Bounding Box Pair Positional Embeddings

We provide more mathematical details on the positional embeddings for bounding box pairs used in cross-attention. Let us first revisit the notations from the main paper. We define sinusoidal embedding of a scalar as $\boldsymbol{\phi} : \mathbb{R} \rightarrow \mathbb{R}^d$,

$$\boldsymbol{\phi}(x)_{2i} = \sin\left(\frac{x}{\tau^{2i/d}}\right), \ \boldsymbol{\phi}(x)_{2i-1} = \cos\left(\frac{x}{\tau^{2i/d}}\right), \quad (7)$$

where $i = 1, \ldots, d/2$ and τ is a temperature parameter we set as 20 following previous practice [**?**]. With box width and height as modulation, the positional embeddings for one bounding box are as follows

$$\text{PE}(x, y, w, h) = \left[\boldsymbol{\phi}(y)\frac{h_{ref}}{h}, \boldsymbol{\phi}(x)\frac{w_{ref}}{w}\right] \in \mathbb{R}^{2d}, \quad (8)$$

where w_{ref}, h_{ref} are reference values learned from box appearance features $\mathbf{f}$ as follows

$$w_{\text{ref}}, h_{\text{ref}} = \sigma(\text{MLP}(\mathbf{f})), \quad (9)$$

where σ is the sigmoid function. As such, the positional embeddings for a human–object pair ($\mathbf{b}_h, \mathbf{b}_o \in \mathbb{R}^4$) are defined by concatenating the positional embeddings of the two boxes,

$$\mathbf{q}_p = \left[\mathbf{q}_p^h, \mathbf{q}_p^o\right] = [\text{PE}(\mathbf{b}_h), \text{PE}(\mathbf{b}_o)] \in \mathbb{R}^{4d}. \quad (10)$$

Denote the positional embeddings of an image patch with normalised spatial indices (i, j) by

$$\mathbf{k}_p = [\boldsymbol{\phi}(j), \boldsymbol{\phi}(i)] \in \mathbb{R}^{2d}. \quad (11)$$

Assuming the number of heads is one, the dot-product attention weights between positional embeddings are computed as

$$(W_k\mathbf{k}_p)^T(W_p\mathbf{q}_p) = \mathbf{k}_p^T W_k^T W_p \mathbf{q}_p, \quad (12)$$

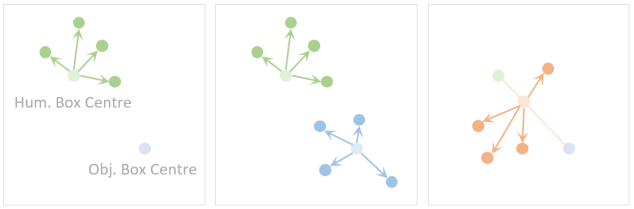
where $W_k \in \mathbb{R}^{2d\times 2d}, W_p \in \mathbb{R}^{2d\times 4d}$ are weight matrices associated with the linear transformations applied to the positional embeddings. In particular, matrix W_p can be partitioned into $\left[W_p^h, W_p^o\right]$, therefore decomposing the linear transformation on query (human–object pair) positional embeddings as follows

$$W_p\mathbf{q}_p = W_p^h\mathbf{q}_p^h + W_p^o\mathbf{q}_p^o. \quad (13)$$

For brevity of exposition, let us now assume that weight matrices $W_k, W_p^h, W_p^o \in \mathbb{R}^{2d\times 2d}$ are identity matrices. This simplifies the dot-product attention weights between positional embeddings in Eq. 12 as follows

$$\mathbf{k}_p^T\mathbf{q}_p^h + \mathbf{k}_p^T\mathbf{q}_p^o, \quad (14)$$

demonstrating that the concatenation of two modulated box positional embeddings results in a weighted sum of the pre-normalised attention weights.



(a) Single ref. (b) Dual ref. (c) Single ref.

Figure 10. Reference point designs for human–object pairs. (a) Human box centre as the single reference point. (b) Human and object box centres as dual reference points. (c) Interpolated point as the single reference point.

B. Multiscale Features as Keys/Values

Deformable DETR [**?**] introduced a simple way to exploit the multiscale structure of convolutional features via deformable attention. Specifically, for each query, a set of offsets with respect to a reference point are predicted to obtain a small number of keys and values. The attention operation is thus reduced to a sparse variant between a query and its corresponding subset of keys and values. As a result of the sparsity, it becomes affordable to extend the cross-attention source from a single feature level to a feature pyramid, where keys/values across different levels are concatenated.

In the case of two-stage HOI detection, we follow the practice in Deformable DETR to construct multiscale features denoted by $\{C_3, C_4, C_5, C_6\}$, where C_3, C_4, C_5 are extracted directly from the backbone and C_6 is obtained from C_5 by applying a 3×3 convolution with stride 2. Four sets of keys/values are sampled for each feature level per query. For the reference points, Deformable DETR uses bounding box centres predicted from the query representations. In our two-stage method, due to the availability of bounding boxes, there is no need to predict the reference points. As such, we focus on designing reference points for human–object pairs and explore three variants depicted in Figure 10.

As humans play a centric role in HOIs, we start with a simple variant with human box centres as the reference point for each query (Figure 10a). Naturally, this can be extended to dual reference points to also include the object box centre (Figure 10b). Last, we experiment with a variant where the reference point is a convex combination (linear interpolation) of the human and object box centres (Figure 10c). Formally, denote the box centres by $x, y \in \mathbb{R}^2$. The reference point is computed as $\beta x + (1 - \beta)y$, where the scaling factor $\beta \in [0, 1]$ is predicted from the query representation and normalised with sigmoid.

We compare the performance against the L1 variant, which uses the C5 features as keys/values and employs a vanilla decoder with box pair positional embeddings intro-

Table 6. The mAP (×100) of model variants with different reference point designs for deformable attention. Results are averaged across three runs.

#	Ref. points	#Keys	Default Setting (mAP)		
			Full	Rare	N-rare
L1	N/A	N/A	**34.18**	**31.09**	**35.10**
M1	single (hum.)	4	33.55	29.70	34.70
M2	dual	8	33.59	30.31	34.57
M3	dual	4	33.55	29.88	34.65
M4	single (interp.)	4	33.40	29.58	34.54

(a) *dribbling a sports ball* (b) *catching a frisbee*

(c) *washing a car*

Figure 11. Predicted human–object interactions and visualised attention weights on data in the wild.

duced in the main paper. For fair comparison, we use one deformable encoder layer to refine the multiscale features and two deformable decoder layers, a similar setup to the L1 variant. As shown in Table 6, we observed insignificant performance differences amongst the M variants with multiscale features, while their performance in general lacks behind the single-scale variant L1.

We believe the inferior performance of multiscale deformable attention is likely due to the visual complexity of human–object interactions. In particular, the role of reference points is somewhat similar to the box pair positional embeddings. Although the offsets with respect to the reference points are dynamically predicted from the query representation, they tend not to have large values. Therefore, they act as a form of inductive bias to encourage high attention weights on keys/values (image patches) closer to the

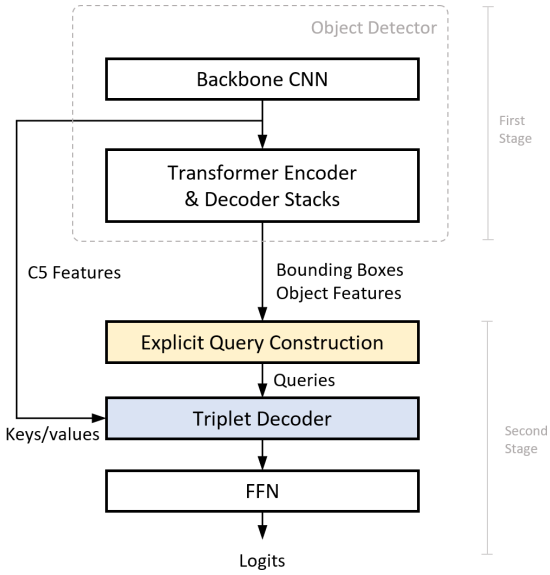


Figure 12. Illustration of the overall pipeline.

reference point, analogous to the $\mathbf{k}_p^T \mathbf{q}_p$ term in the vanilla decoder. On the other hand, a weakness of deformable attention is that, it does not have a mechanism to encourage attention based on visual similarity, i.e. the $\mathbf{k}_c^T \mathbf{q}_c$ term in the vanilla decoder. As we have demonstrated in the main paper, the recognition of HOIs requires more complex visual context, compared to object detection. Therefore, the offsets predicted from the query features are not sufficient to locate the relevant context.

C. Demonstration on Data in The Wild

For more evidence on the two types of visual context exploited by our model, we provide additional qualitative results on data in the wild, with cross-attention weights overlaid. As shown in Figures 11a and 11b, the model extracts contextual features from image regions containing relevant human body parts, and successfully predicts the correct interactions with high scores. In Figure 11c, we highlight the other type of context, i.e. another involved object. Notably, three out of the four human–object pairs place high attention weights on the water buckets, which are indicators of the corresponding interaction.

D. Pipeline

For better clarity, we attach an illustration of the entire pipeline, as shown in Figure 12. Due to the popularity of the DETR framework, the first stage is depicted as a transformer-based object detector. But the method itself is detector-agnostic.

E. Advanced Variants of DETR

To demonstrate the detector-agnostic nature of our method, we fine-tuned the recent state-of-the-art object detector $\mathcal{H}$-DETR [**?**] and reported the performance of our method with it, as shown in Table 5. However, we would like to point out that the performance of our method with $\mathcal{H}$-DETR-R50 was surprisingly lower than that with DETR-R50, although $\mathcal{H}$-DETR-R50 outperforms DETR-R50 significantly in terms of object detection mAP on HICO-DET [**?**]. To investigate this issue, we first made the observation that $\mathcal{H}$-DETR was trained using a multi-label classification objective, that is, the scores are individual normalised using the Sigmoid function as opposed to Softmax. As a result, the predicted object detection scores tend to be lower, thus less over-confident. To this end, we increased the value of hyper-parameter λ from Eq. 6 to 0.37. Nevertheless, this only resulted in marginal improvement.

In addition, $\mathcal{H}$-DETR employs deformable attention [**?**] and utilises the multi-scale feature. Our method, on the other hand, uses a single-scale feature in the spatially-guided cross-attention. Therefore, it is likely that the different levels of the feature maps learned to extract different information, Unfortunately, our attempts in exploiting such multi-scale features did not lead to concrete improvements. As such, we leave this problem to potential future work.