MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection

Bumsoo Kim*
LG AI Research
bumsoo.kim@lgresearch.ai

Jonghwan Mun
Kakao Brain

Kyoung-Woon On
Kakao Brain

Minchul Shin
Kakao Brain

Junhyun Lee
Korea University

Eun-Sol Kim
Department of Computer Science
Hanyang University

Abstract

Human-Object Interaction (HOI) detection is the task of identifying a set of ⟨human, object, interaction⟩ triplets from an image. Recent work proposed transformer encoder-decoder architectures that successfully eliminated the need for many hand-designed components in HOI detection through end-to-end training. However, they are limited to single-scale feature resolution, providing suboptimal performance in scenes containing humans, objects, and their interactions with vastly different scales and distances. To tackle this problem, we propose a Multi-Scale TRansformer (MSTR) for HOI detection powered by two novel HOI-aware deformable attention modules called Dual-Entity attention and Entity-conditioned Context attention. While existing deformable attention comes at a huge cost in HOI detection performance, our proposed attention modules of MSTR learn to effectively attend to sampling points that are essential to identify interactions. In experiments, we achieve the new state-of-the-art performance on two HOI detection benchmarks.

1. Introduction

Human-Object Interaction (HOI) detection is a task to predict a set of ⟨*human, object, interaction*⟩ triplets in an image [9]. Previous methods have indirectly addressed this task by detecting human and object instances and individually inferring interaction labels for every pair of the detected instances with either neural networks (*i.e.*, two-stage HOI detectors [1, 6–8, 10, 16, 18, 19, 21–24, 26, 28–30, 32, 33, 35]) or triplet matching (*i.e.*, one-stage HOI detectors [12, 20, 31]). The additional complexity caused by this indirect inference structure and post-processing (*e.g.*, NMS) stage behaved as a major bottleneck in inference time in HOI detection. To deal with this bottleneck, transformer-based HOI detectors [4, 13, 25, 37] have been proposed to achieve end-to-end HOI detection without the need for the post-processing stage mentioned above. These works have shown competitive performance in both accuracy and inference time with direct set-level prediction and transformer attentions that can exploit the contextual information between humans, objects, and their interactions.

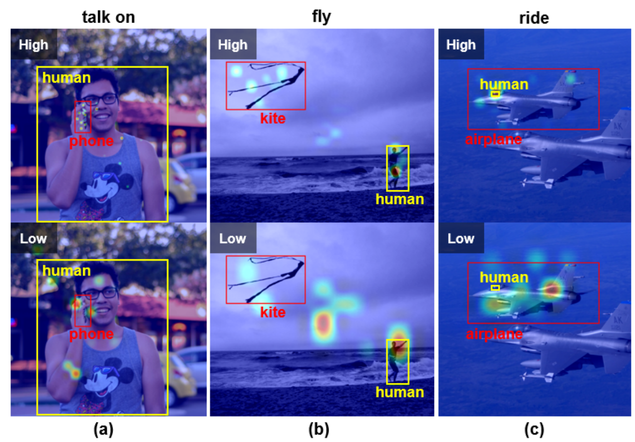

Figure 1. Multi-scale attention of MSTR on interactions including: (a) large human with small object, (b) distant human and object, and (c) small human and a large object. The top row (high resolution) and the bottom row (low resolution) captures the context of the interaction in various scales. Best viewed in color.

However, due to the huge computational costs raised when processing multi-scale feature maps (with about 20× more image tokens) with transformer attention, current transformer-based HOI detectors are limited to using only single-scale feature maps. Due to this limitation, previous transformer-based approaches demonstrate suboptimal per-

*this work was done in Kakao Brain

formance, especially for scenes where humans, objects, and the contextual information for their interactions exist at various scales.

In this paper, we propose Multi-Scale TRasnformer (MSTR), a transformer-based HOI detector that can exploit multi-scale feature maps for HOI detection. Inspired by previously proposed deformable attention for standard object detection [36], we aim to efficiently explore multi-scale feature maps by attending to only a small number of sampling points generated from the query element instead of calculating the attention values for the entire spatial dimension. Yet, we found out in our preliminary experiments that directly applying naïve deformable attention in HOI detection leads to a serious performance drop.

To overcome this deterioration, we equipped MSTR with two novel *HOI-aware* deformable attentions, referred by Dual-Entity Attention and Entity-conditioned Context Attention, which are designed to capture the complicated semantics of Human-Object Interaction throughout multi-resolution feature maps (see Figure 1). Specifically, precise entity-level semantics for humans and objects are captured by *Dual-Entity attention*, while the contextual information for the interaction is conditionally reimbursed by *Entity-conditioned Context attention*. To further improve performance, we delve into decoder architectures that can effectively handle the multiple semantics obtained from the two HOI-aware attentions above.

The main contributions of our work are threefold:

- We propose MSTR, the first transformer-based HOI detector that exploits multi-scale visual feature maps.
- We propose new deformable attention modules, called Dual-Entity attention and Entity-conditioned Context attention, which effectively and efficiently capture human, object, and context information associated with HOI queries.
- We explore decoder architectures to handle the multiple semantics captured by our proposed deformable attentions and further improve HOI detection performance.

2. Preliminary

In this section, we start with a basic pipeline of a transformer-based end-to-end HOI detector [25]. Then, we explain the deformable attention module [36] that reduces computational cost in attention, thus enabling the transformer to take multi-scale feature maps as an input. Afterward, we discuss why the direct application of multi-scale deformable attentions is not suitable for HOI detection.

2.1. End-to-End HOI Detection with Transformers

Out of the multiple candidates [4, 13, 25, 37] using transformers for HOI detection, we adopt QPIC [25] as our baseline due to its simple structure and good performance.

Set Prediction. Transformer-based HOI detectors formulate the task as a set-level prediction problem. It is achieved by exploiting a fixed number of HOI queries, each of which generates four types of predictions: 1) the coordinate of the human bounding box (*i.e*., subject of the interaction), 2) the coordinate of the object bounding box (*i.e*., target of the interaction), 3) the object class and 4) the interaction type. Note that the set-level predictions are learned using losses based on Hungarian Matching with ground-truths.

Transformer Encoder-Decoder Architecture. The architecture of QPIC [25] consists of a backbone CNN, a transformer encoder, and a transformer decoder. Given an image, a single-scale visual feature map is extracted by the backbone CNN, and then positional information is added to the feature map. The transformer encoder takes the visual features and returns contextualized visual features with self-attention layers. In the transformer decoder, HOI queries are first processed by the self-attention layer, and then the cross-attention layer associates the HOI queries with the contextualized visual features (given by the encoder) to capture relevant HOI representations. Finally, predictions for HOI are computed from individual contextualized HOI query embeddings as mentioned above. Note that both self-attention and cross-attention adopt multi-head attention.

To be specific, given a single-scale input feature map $x \in \mathbb{R}^{C\times H\times W}$ where C is the feature dimension, the single-scale multi-head attention $f_q^{sg} = \text{SSAttn}(z_q, x)$ for the q^{th} query feature z_q (either an image token for the encoder or an HOI query for the decoder) is calculated by

$$f_q^{sg} = \sum_{m=1}^{M} W_m \big[\sum_{k\in\Omega_k} A_{mqk} \cdot W'_m x_k \big], \tag{1}$$

where A_{mqk} indicates an attention weight calculated with learnable weights $U_m, V_m \in \mathbb{R}^{C_v\times C}$ as $\exp\big(\frac{z_q^T U_m^T V_m x_k}{\sqrt{C_v}}\big)$. Throughout this paper, for the attention module, we let m index the attention head ($1 \leq m \leq M$), $q \in \Omega_q$ indexes a query element with feature $z_q \in \mathbb{R}^C$, $k \in \Omega_k$ indexes a key element with feature $z_k \in \mathbb{R}^C$, while Ω_q and Ω_k specify the set of query and key elements, respectively. W_m and W'_m are learnable embedding parameters for m^{th} attention head, and A_{mqk} is normalized as $\sum_{k\in\Omega_k} A_{mqk} = 1$.

Complexity. Given an input feature map $x \in \mathbb{R}^{C\times H\times W}$ and N HOI queries, the complexity of transformer encoder and decoder are $O(H^2W^2C)$ and $O(HWC^2 +$

$NHWC + 2NC^2 + N^2C$), respectively. Since the complexity grows in quadratic scale as the spatial resolution (H,W) increases, it raises significant complexity when exploiting multi-resolution feature maps where there are about 20× more features to process.

Towards Multi-Scale HOI detection. In HOI detection, not only do humans and objects exist at various scales, but they also interact at various distances in images. Therefore, it is essential to exploit multi-scale feature maps $\{x\}_{l=1}^{L}$ (where $x^l \in \mathbb{R}^{C\times H_l\times W_l}$, l indexes the feature level) to deal with the various scales of objects and contexts to capture interactions precisely. However, as multi-scale feature maps have almost ×20 more elements to process than a single-scale feature map, it provokes a serious complexity issue in calculating Eq. (1).

2.2. Revisiting Deformable Transformers

The deformable attention module is proposed to deal with the problem of high complexity in the transformer attention. The core idea is to reduce the number of *key* elements in the attention module by sampling the small number of spatial locations related to regions of interest for each *query* element.

Sampling Locations for Deformable Attention. Given a multi-scale input feature map $\{x^l\}_{l=1}^{L}$ where $x^l \in \mathbb{R}^{C\times H_l\times W_l}$, the K sampling locations of interest for each attention head and each feature level are generated from each *query* element $z_q \in \mathbb{R}^C$. Because direct prediction of coordinates of sampling location is difficult to learn, it is formulated as prediction of a reference point $r_q \in [0,1]^2$ and K sampling offsets $\Delta r_q \in \mathbb{R}^{M\times L\times K\times 2}$. Then, the k^{th} sampling location at l^{th} feature level and m^{th} attention head for query element z_q is defined by $p_{mlqk} = \phi_l(r_q)+\Delta r_{mlqk}$ where $\phi_l(\cdot)$ is a function to re-scale the coordinate of reference point to the input feature map of the l^{th} level.

Deformable Attention Module. Given a multi-scale input feature map $\{x^l\}_{l=1}^{L}$, the multi-scale deformable attention $f_q^{ms} = \text{MSDeformAttn}(z_q, p_q, \{x^l\}_{l=1}^{L})$ for query element z_q is calculated using a set of predicted sampling locations p_q as follows:

$$f_q^{ms} = \sum_{m=1}^{M} W_m \big[\sum_{l=1}^{L}\sum_{k=1}^{K} A_{mlqk} \cdot W'_m \Phi_{mlqk}\big], \quad (2)$$

where l, k and m index the input feature level, the sampling location and the attention head, respectively, while A_{mlqk} indicates an attention weight for the k^{th} sampling location at the l^{th} feature level and the m^{th} attention head. Φ_{mlqk} means the sampled k^{th} key element at l^{th} feature level and m^{th} attention head using the sampling location, which is obtained by bilinear interpolation as $\Phi_{mlqk} = x^l(p_{mlqk}) = x^l(\phi_l(r_q) + \Delta r_{mlqk})$. Note that for each query element, the attention computation is performed with only sampled regions of interest where the sampled number ($= LMK$) is much smaller than the number of all the key elements ($\sum_{l=1}^{L} H_l W_l$), thus leads to a reduced computational cost.

Problem with Direct Application to HOI Detection. Deformable attention effectively reduces the complexity of exploiting multi-scale features with transformers to an acceptable level. However, while the sampling procedure above does not deteriorates performance in standard object detection, it causes a serious performance drop in HOI detection (29.07 → 25.53) as shown in Table 3. We conjecture that this is partly due to the following reasons. First, unlike the object detection task where an object query is associated with a single object, an HOI query is entangled with multiple semantics (*i.e.*, human, object, and their interaction); thus learning to sample the region of interest for multiple semantics with individual HOI queries (especially with sparse information) is much challenging compared to the counterpart of object detection. Second, deformable attention is learned to attend only to the sampling points near the localized objects; this leads to the loss of contextual information that is an essential clue for precise HOI detection. The following sections describe how we resolve these issues and improve performance.

3. Method

In this section, we introduce MSTR, a novel deformable transformer architecture that is suitable for multi-scale HOI detection. To resolve the problems described in our preliminary, MSTR features new *HOI-aware* deformable attentions designed for HOI detection, referred by Dual-Entity attention and Entity-conditioned Context attention.

3.1. HOI-aware Deformable Attentions

The objective of our HOI-aware deformable attentions (Dual-Entity attention and Entity-conditioned Context attention) is to efficiently and effectively extract information of HOIs from multi-scale feature maps for a given HOI query. Figure 2 shows conceptual illustrations of (a) deformable attention in literature [36], (b) Dual-Entity attentions and (c) Entity-conditioned Context attention.

Dual-Entity attention for Human/Object. In HOI detection, the HOI query includes complex and entangled information of multiple semantics: human, object, and interaction information. Therefore, it is challenging to accurately predict sampling locations appropriate for each semantic from a single HOI query. To make sampling loca-

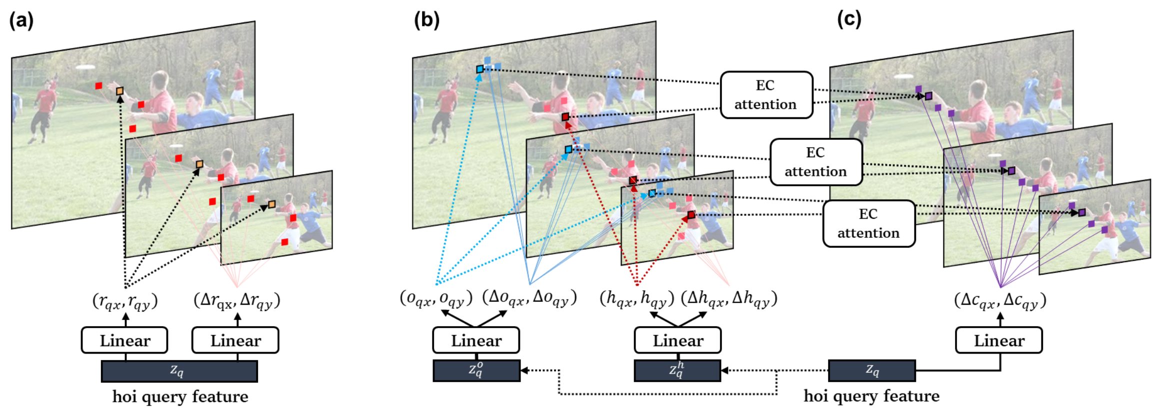

Figure 2. Illustration of (a) Deformable Attention, (b) Dual-Entity Attention, (c) Entity-conditioned Context Attention (abbreviated as EC). The sampling point for deformable attention is obtained by combining the reference points with sampling offset. In (a), both reference points $r_q = (r_{qx}, r_{qy})$ and sampling offsets $\Delta r_q = (\Delta r_{qx}, \Delta r_{qy})$ are obtained from a single hoi query feature z_q. In (b), the reference points and sampling offsets for the humans $h_q = (h_{qx}, h_{qy})$, $\Delta h_q = (\Delta h_{qx}, \Delta h_{qy})$ and objects $o_q = (o_{qx}, o_{qy})$, $\Delta o_q = (\Delta o_{qx}, \Delta o_{qy})$ are obtained from z_q^h and z_q^o, respectively, which is obtained by a linear projection of z_q (dotted line). In (c), the sampling offsets $\Delta c_q = (\Delta c_{qx}, \Delta c_{qy})$ are obtained from z_q while the reference points are obtained in conditional to entities in (b).

tions easier, given an HOI query feature z_q, our Dual-Entity attention separately identifies sampling locations for the humans (p_q^h) and objects (p_q^o). First, we project z_q with two linear layers to obtain z_q^h and z_q^o. The k^{th} sampling location at l^{th} feature level and m^{th} attention head for human and object are represented by

$$\begin{aligned} p_{mlqk}^h &= \phi_l(h_q) + \Delta h_{mlqk}, \\ p_{mlqk}^o &= \phi_l(o_q) + \Delta o_{mlqk}, \end{aligned} \tag{3}$$

where h_q, Δh is the reference point and sampling offsets for humans, and o_q, Δo is the reference point and sampling offsets for objects, each obtained by a linear projection of z_q^h and z_q^o, respectively. Then, based on the sampled locations, attended features for human (f_q^h) and object (f_q^o) are computed by

$$\begin{aligned} f_q^h &= \text{MSDeformAttn}(z_q^h, p_q^h, \{x^l\}_{l=1}^L), \\ f_q^o &= \text{MSDeformAttn}(z_q^o, p_q^o, \{x^l\}_{l=1}^L). \end{aligned} \tag{4}$$

Entity-conditioned Context attention. In HOI detection, contextual information often gives an important clue in identifying interactions. From this point of view, utilizing the local features obtained from near the human and object regions through the Dual-Entity attention is not sufficient to capture contextual information (see our experimental result in Table 3). To compensate for this, we define an attention with an additional set of sampling points, namely Entity-conditioned Context attention, that is designed to capture the contextual information in specific.

Given the 2D reference points for the human $h_q = (h_{qx}, h_{qy})$ and the object $o_q = (o_{qx}, o_{qy})$, the reference point for Entity-conditioned Context attention is conditionally computed with the two references. Motivated by existing works [20, 31, 34], we define the reference points for interaction as the center of human and object, *i.e.*, $c_q = \left(\frac{h_{qx}+o_{qx}}{2}, \frac{h_{qy}+o_{qy}}{2}\right)$. Note that we empirically observe that such simple reference points offer competitive performance compared to ones predicted using an additional network, while being much faster. Then, we predict the sampling offsets Δc_q from the HOI query feature, obtaining $p_{mlqk}^c = \phi_l(c_q) + \Delta c_{mlqk}$. Finally, the attended feature for contextual information f_q^c is computed using sampling location p_q^c as follows:

$$f_q^c = \text{MSDeformAttn}(z_q, p_q^c, \{x^l\}_{l=1}^L). \tag{5}$$

3.2. MSTR Architecture

In this section, the overall architecture of MSTR with our suggested two deformable attentions will be described (see Figure 3). MSTR follows the previous transformer encoder-decoder architecture, where the encoder performs self-attention given the image features while the decoder performs self-attention for HOI queries followed by cross-attention between updated HOI queries and the encoded image features.

Encoder. The encoder of MSTR takes multi-scale input feature maps given by a backbone CNN, performs a series of deformable attention modules in Eq.(2), and finally generates encoded feature maps. Positional encoding [2] is added to preserve spatial information while level embed-

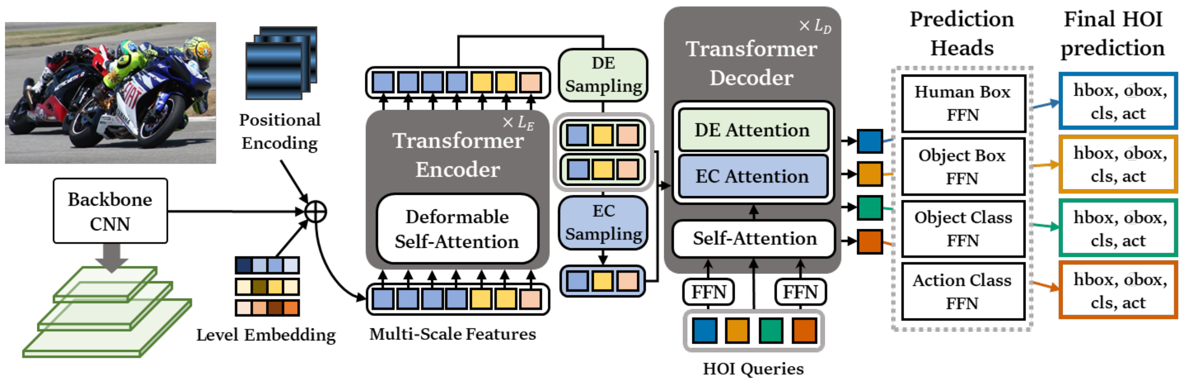

Figure 3. Overall pipeline of MSTR. On top of the standard transformer encoder-decoder architecture for HOI detection (*i.e.*, QPIC), we leverage deformable samplings for the encoder self-attention and the decoder cross-attention modules to deal with the huge complexity caused by using multi-scale feature maps. For the decoder cross-attention, we leverage three sets of key elements sampled for our Dual-Entity attention (denoted as DE sampling, DE attention) and Entity-conditioned Context attention (denoted as EC sampling, EC attention).

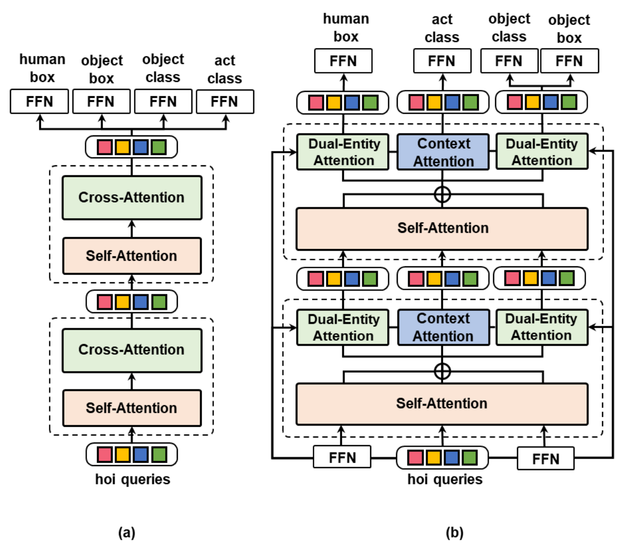

Figure 4. Comparison of a simple 2-layer Decoder architecture for Transformer-based HOI detectors: (a) conventional one introduced in QPIC, and (b) HOI-aware one in MSTR. Entity-conditioned Context attention is abbreviated as Context Attention. MSTR stacks decoder layers by merging the self attention outputs, which further improves performance (see Table 3).

ding [36] is added to denote which resolution did the image feature comes from.

Decoder. By leveraging our HOI-aware deformable attentions, the cross-attention layer in MSTR decoder extracts three different semantics (human, object, and contextual information) for each HOI query from the encoded image features. For each decoder layer, we discovered that compositing the multiple semantics obtained from the previous cross-attention layer [5] by summing the semantics after applying individual self-attention demonstrates the best performance (see Table 3 and Appendix). The input for the $(k+1)$-th layer of our HOI-aware deformable attention $\bar{z}_q^{k+1}$ is written as:

$$\bar{z}_q^{k+1} = \mathrm{SA}(f_q^h(k)) + \mathrm{SA}(f_q^o(k)) + \mathrm{SA}(f_q^c(k)), \tag{6}$$

where $f_q^h(k)$, $f_q^o(k)$, $f_q^c(k)$ denotes the multiple semantic outputs of the previous (k-th) decoder obtained by Eq.(4) and Eq.(5), respectively. SA denotes Multi-Head Self-Attention operation with Eq.(1) [27] and $\bar{z}_q^1 = \mathrm{SA}(z_q) + \mathrm{SA}(z_q^h) + \mathrm{SA}(z_q^o)$.

MSTR Inference. Given the cross attention results of the final decoder layer where f_q^h and f_q^o is obtained by Eq. (4) and f_q^c is obtained by Eq. (5), the final prediction heads in MSTR predict the $\langle \mathrm{bbox}_q^h, \mathrm{bbox}_q^o, \mathrm{cls}_q^o, \mathrm{act}_q \rangle$ using FFN as follows:

$$(u_{qx}, u_{qy}, u_{qw}, u_{qh}) = \mathrm{FFN}_{\mathrm{hbox}}(f_q^h), \tag{7}$$

$$(v_{qx}, v_{qy}, v_{qw}, v_{qh}) = \mathrm{FFN}_{\mathrm{obox}}(f_q^o), \tag{8}$$

$$\mathrm{cls}_q = \sigma(\mathrm{FFN}_{\mathrm{cls}}(f_q^o)), \tag{9}$$

$$\mathrm{act}_q = \sigma(\mathrm{FFN}_{\mathrm{act}}(f_q^c)), \tag{10}$$

where cls_q and act_q each denote predictions for object the class and the action class after sigmoid function, and final bbox_q^h is predicted with the center point $\big(\sigma(u_{qx} + \sigma^{-1}(h_{qx})), \sigma(u_{qy} + \sigma^{-1}(h_{qy}))\big)$, width u_{qw}, and height u_{qh}. Likewise, the bbox_q^o is predicted with center point as $\big(\sigma(v_{qx} + \sigma^{-1}(o_{qx})), \sigma(v_{qy} + \sigma^{-1}(o_{qy}))\big)$, width v_{qw}, and height v_{qh}. σ and σ^{-1} denote the sigmoid and inverse sigmoid function, respectively, and is used to normalize the reference points h_q, o_q and the predicted coordinates of human boxes and object boxes $u_{q\{x,y,w,h\}}, v_{q\{x,y,w,h\}} \in \mathbb{R}$.

4. Experiment

In this section, we show the experimental results of our model in HOI detection. We first describe the experimental settings such as datasets and evaluation metrics. Next, we compare MSTR with state-of-the-art works on two different benchmarks (V-COCO and HICO-DET) and provide detailed ablation study for each component. Through the experiments, we demonstrate that MSTR successfully extends conventional transformer-based HOI detectors to utilize multi-scale feature maps, and emphasize that each component of MSTR contributes to the final HOI detection performance. Lastly, we provide extensive qualitative results of MSTR.

4.1. Datasets and Metrics

We evaluate our model on two widely-used public benchmarks: the V-COCO (*Verbs in COCO*) [9] and HICO-DET [3] datasets. V-COCO is a subset of COCO composed of 5,400 trainval images and 4,946 test images. For V-COCO dataset, we report the AP_{role} over 25 interactions in two scenarios. HICO-DET contains 37,536 and 9,515 images for each training and test split with annotations for 600 $\langle verb, object\rangle$ interaction types. We follow the previous settings and report the mAP over two evaluation settings (Default and Known Object), each with three different category sets: (1) all 600 HOI categories in HICO (Full), (2) 138 HOI categories with less than 10 training instances (Rare), and (3) 462 HOI categories with 10 or more training instances (Non-Rare). See Appendix for details of the evaluation settings.

4.2. Quantitative Results

We use the standard evaluation code [1] following the previous works [4,13,25,37] to calculate metric scores for both V-COCO and HICO-DET.

Comparison to State-of-The-Art. We compare MSTR with state-of-the-art methods in Table 1 and Table 2. In Table 1, MSTR outperforms the previous state-of-the-art method in V-COCO dataset by a large margin (+3.2p in $\text{AP}^{\#1}_{\text{role}}$ and +4.2p in $\text{AP}^{\#1}_{\text{role}}$). Similar to this, in Table 2, MSTR achieves the highest mAP on HICO-DET dataset in all Full, Rare, and Non-Rare classes obtaining +2.1p, +3.46p, and +1.69p gain for each compared to the previous state-of-the-art. We use the same scoring function as QPIC without any modification for a fair comparison. Note that MSTR benefits from the advantages of using deformable attention: the fast convergence for training [36] (see more details and the convergence graph in our Appendix).

[1] https://github.com/YueLiao/PPDM

Method	Backbone	$AP^{\#1}_{\text{role}}$	$AP^{\#2}_{\text{role}}$
*Models with **external features***			
TIN ($\text{RP}_\text{D}\text{C}_\text{D}$) [19]	R50	47.8	-
Verb Embedding [32]	R50	45.9	-
RPNN [35]	R50	-	47.5
PMFNet [28]	R50-FPN	52.0	-
PastaNet [18]	R50-FPN	51.0	57.5
PD-Net [33]	R50	52.0	-
ACP [14]	R152	53.0	-
FCMNet [21]	R50	53.1	-
ConsNet [22]	R50-FPN	53.2	-
Sequential HOI Detectors			
VSRL [9]	R50-FPN	31.8	-
InteractNet [8]	R50-FPN	40.0	48.0
BAR-CNN [15]	R50-FPN	43.6	-
GPNN [24]	R152	44.0	-
iCAN [7]	R50	45.3	52.4
TIN (RC_D) [19]	R50	43.2	-
DCA [30]	R50	47.3	-
VCL [11]	R50-FPN	48.3	-
DRG [6]	R50-FPN	51.0	-
VSGNet [26]	R152	51.8	57.0
IDN [17]	R50	53.3	60.3
Parallel HOI Detectors			
UnionDet [12]	R50-FPN	47.5	56.2
IPNet [31]	HG104	51.0	-
HOI Transformer [37]†	R101	52.9	-
ASNet [4]†	R50	53.9	-
GGNet [34]	HG104	54.7	-
HOTR [13]†	R50	55.2	64.4
QPIC [25]†	R50	58.8	61.0
MSTR (Ours)	R50	**62.0**	**65.2**

Table 1. Comparison of performance on V-COCO test set. $AP^{\#1}_{\text{role}}$, $AP^{\#2}_{\text{role}}$ denotes the performance under Scenario 1 and Scenario 2 in V-COCO, respectively. † denotes end-to-end HOI detectors with transformers, which are the main baselines for our work.

4.3. Ablation Study

We perform ablations to check the effects of our proposed Dual-Entity attention, Entity-conditioned Context attention, and our proposed decoder architecture that merges the self-attention of the multiple semantics.

Baselines. On basis of QPIC [25] structure, we define several variants for baselines by applying different combinations of sub-components from MSTR: *multi-scale feature maps (MS)*, *Deformable Attention (DA)*, *Dual-Entity attention (DE)*, and *Entity-conditioned Context attention (EC)*. Specifically, since deformable attention can be also applied to a single-scale feature map, *SS-Baseline* denotes QPIC where the attention in the transformer is replaced by DA. Our work can be seen as a process of improving the score

Method	Detector	Backbone	Feature	Default			Known Object		
				Full	Rare	Non Rare	Full	Rare	Non Rare
Sequential HOI Detectors									
Functional Gen. [1]	HICO-DET	R101	A+S+L	21.96	16.43	23.62	-	-	-
TIN [19]	HICO-DET	R50	A+S+P	22.90	14.97	25.26	-	-	-
VCL [11]	HICO-DET	R50	A+S	23.63	17.21	25.55	25.98	19.12	28.03
ConsNet [22]	HICO-DET	R50-FPN	A+S+L	24.39	17.10	26.56	30.34	23.40	32.41
DRG [6]	HICO-DET	R50-FPN	A+S	24.53	19.47	26.04	27.98	23.11	29.43
IDN [17]	HICO-DET	R50	A+S	24.58	20.33	25.86	27.89	23.64	29.16
Parallel HOI Detectors									
UnionDet [12]	HICO-DET	R50-FPN	A	17.58	11.72	19.33	19.76	14.68	21.27
PPDM [20]	HICO-DET	HG104	A	21.10	14.46	23.09	24.81	17.09	27.12
HOI Transformer [37]†	HICO-DET	R50	A	23.46	16.91	25.41	26.15	19.24	28.22
HOTR [13]†	HICO-DET	R50	A	25.10	17.34	27.42	-	-	-
GGNet [34]	HICO-DET	HG104	A	28.83	22.13	30.84	27.36	20.23	29.48
AS-Net [4]†	HICO-DET	R50	A	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [25]†	HICO-DET	R50	A	29.07	21.85	31.23	31.68	24.14	33.93
MSTR (Ours)	HICO-DET	R50	A	**31.17**	**25.31**	**32.92**	**34.02**	**28.83**	**35.57**

Table 2. Performance comparison in HICO-DET. The Detector column is denoted as 'HICO-DET' to show that the object detector is fine-tuned on the HICO-DET training set. Each letter in Feature column stands for A: Appearance (Visual Features), S: Interaction Patterns (Spatial Correlations), P: Pose Estimation, L: Linguistic Priors, V: Volume. † denotes end-to-end HOI detectors with transformers. Note that all the baseline models without † are already based on multi-scale feature maps.

Method	MS	DA	DE	EC	mAP
(a) QPIC					29.07
(b) SS-Baseline		✓			25.53
(c) SS-Baseline + DE		✓	✓		27.06
(d) SS-Baseline + DE + EC		✓	✓	✓	27.70
(e) MS-Baseline	✓	✓			27.52
(f) MS-Baseline + DE	✓	✓	✓		28.30
(g) MS-Baseline + DE + EC	✓	✓	✓	✓	30.14
(h) MSTR (Ours)	✓	✓	✓	✓	**31.17**

Table 3. Comparison of MSTR with our baseline QPIC and its variants in the HICO-DET test set. SS and MS denote the models using single scale feature map and multi-scale feature maps, respectively. DE and EC indicate our proposed Dual-Entity attention and Entity-conditioned Context attention, respectively.

to the state-of-the-art by adapting *MS, DE, EC* step by step to *SS-Baseline*. *MS-Baseline+DE+EC* represents MSTR without merging with self-attention, instead simply passing the sum of the outputs to the next decoder layer.

HOI-Aware Deformable Attentions. In Table 3, we explore the effect of our proposed HOI-Aware Deformable Attentions: Dual-Entity attention and Entity-conditioned Context attention. As deformable attentions can also be applied in a single-scale feature map, we verify the effectiveness of our proposed deformable attentions on both single-scale and multi-scale baselines. As we described in our preliminary, the naïve implementation of deformable attention on top of QPIC (for single-scale) significantly degrades the score in both single-scale and multi-scale environments (see (a vs. b) and (a vs. e)). The use of Dual-Entity attention (DE) consistently improves the score in both single-scale (+1.53p in (b vs. c)) and multi-scale environments (+0.78p in (e vs. f)). As well, Entity-conditioned Context attention (EC) contributes in the multi-scale environment when jointly used with DE (+0.64p in SS and +1.84p in MS). Therefore, we conclude that disentangling the references (DE) and conditionally reimbursing context information (EC) each gradually contributes to the final performance of HOI detection in both single-scale and multi-scale environments, enabling MSTR to effectively explore multi-scale feature maps to achieve state-of-the-art performance.

Single-scale vs. Multi-scale. In Table 1 and Table 2, we demonstrate that our method using the multi-scale feature maps outperform all previous methods, including transformer-based methods [4, 13, 25, 37] and the ones that already use multi-scale feature maps heavily [6, 11, 12, 17, 22, 34]. To analyze further, Table 3 compares single-scale version and the multi-scale version of our baselines (see (b-e) and (e-h)). In all cases of converting the single-scale feature map to the multi-scale one, we observe consistent performance gains (see (b vs. e), (c vs. f), and (d vs. g,h)). The gain is maximized when *DE* and *EC* are used together. We further provide a detailed analysis of the effectiveness of MSTR in multi-scale environments in our Appendix.

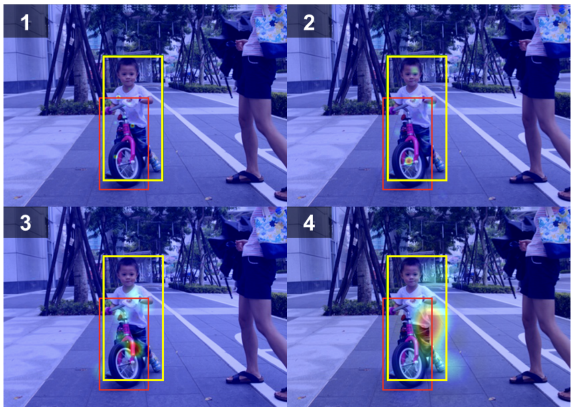

Figure 5. Visualization of our Entity-conditioned Context attentions on different levels of feature map (1 being the highest and 4 being the lowest resolution). Best viewed in color and scale.

Figure 6. Visualization of the HOI-aware attention of MSTR on different scales of humans and objects.

Decoder Architecture. We verify the effectiveness of Figure 4 (b) architecture in Table 3 (g vs. h). As MSTR considers multiple semantics with two suggested deformable modules, it is important to find suitable decoder architecture which can effectively merge the semantics [5]. According to the possible combination ways when merging three kinds of semantics, various types of decoder architecture can be candidates for the decoder architectures (described in Appendix). In our Appendix, we empirically verify that Figure 4 (b) architecture shows the most powerful and robust performance across all datasets.

4.4. Qualitative Results

We conduct qualitative analysis of MSTR to observe how MSTR captures interactions. Figure 1 and Figure 5 show the visualization of the attention map in MSTR in various feature levels. Interestingly, we can observe that in the higher resolution feature maps, the sampling points capture the detail of the interacting human and object while the lower resolution feature maps tend to capture the overall pose or context of the interaction. In Figure 1 and Figure 6, we can observe how MSTR attends to test images that include various scales of humans, target objects, and distances. More details along with quantitative results will be provided in our Appendix.

5. Related Work

Transformer Based HOI Detectors. Human-Object Interaction detection has been initially proposed in [9], and has been developed in two main streams: sequential methods [1, 6–8, 10, 16, 18, 19, 21–24, 26, 28–30, 32, 33, 35] and parallel methods [12, 20, 31]. However, since these works required hand-crafted post-processing, HOI detectors with transformers have been proposed to eliminate the post-processing step through an end-to-end fashioned set prediction approach [4, 13, 25, 37]. Yet, all these methods are limited to a single-scale feature map due to the complexity caused when processing multi-scale feature maps with transformer attention.

Deformable Transformers for Object Detection. DETR has been recently proposed to eliminate the need for many hand-designed components in object detection [2]. Deformable DETR [36] mitigates the slow convergence and high complexity issues of DETR and successfully exploits multi-resolution feature maps. The deformable attention modules in [36] attend to a small set of sampling locations as a pre-filter for prominent key elements out of all the feature map pixels. However, unlike object detection, we observed that this pre-filter seriously deteriorates performance when applied to HOI detection. Therefore, in this paper, we focus on finding a proper way to incorporate deformable attention into HOI detection for exploiting multi-scale feature maps.

6. Conclusion

In this paper, we present MSTR, the first multi-scale approach in transformer-based HOI detectors. MSTR overcomes the issues of extending transfomer-based HOI detectors to multi-scale feature maps with novel HOI-Aware Deformable attentions named as Dual-Entity attention and Entity-conditioned Context attention. In virtue of the two attention modules and our decoder architecture that effectively collects the multiple semantics from each of the attentions, MSTR achieves the state-of-the-art performance in two benchmark datasets in HOI detection.

Acknowledgements. This work was partly supported by the IITP grants (2020-0-01373) and Hanyang University (HY-202100000003160).

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, pages 10460–10469, 2020. 1, 7, 8

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 4, 8

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 6

[4] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 1, 2, 6, 7, 8

[5] Qi Dong, Zhuowen Tu, Haofu Liao, Yuting Zhang, Vijay Mahadevan, and Stefano Soatto. Visual relationship detection using part-and-sum transformers with composite queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3550–3559, 2021. 5, 8

[6] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. 1, 6, 7, 8

[7] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 1, 6, 8

[8] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. 1, 6, 8

[9] Jitendra Gupta, Saurabh Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 6, 8

[10] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9677–9685, 2019. 1, 8

[11] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. *arXiv preprint arXiv:2007.12407*, 2020. 6, 7

[12] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo Kim. Uniondet: Union-level detection towards real-time human-object interaction detection. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 1, 6, 7, 8

[13] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 1, 2, 6, 7, 8

[14] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. *arXiv preprint arXiv:2007.08728*, 2020. 6

[15] Alexander Kolesnikov, Alina Kuznetsova, Christoph Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[16] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10166–10175, 2020. 1, 8

[17] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33, 2020. 6, 7

[18] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 1, 6, 8

[19] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 1, 6, 7, 8

[20] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 1, 4, 7, 8

[21] Y Liu, Q Chen, and A Zisserman. Amplifying key cues for human-object-interaction detection. *Lecture Notes in Computer Science*, 2020. 1, 6, 8

[22] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020. 1, 6, 7, 8

[23] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1981–1990, 2019. 1, 8

[24] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018. 1, 6, 8

[25] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 1, 2, 6, 7, 8

[26] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13617–13626, 2020. 1, 6, 8

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[28] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9469–9478, 2019. 1, 6, 8

[29] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. *arXiv preprint arXiv:2010.10001*, 2020. 1, 8

[30] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. *arXiv preprint arXiv:1910.07721*, 2019. 1, 6, 8

[31] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 1, 4, 6, 8

[32] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 6, 8

[33] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for human-object interaction detection. In *Proc. Eur. Conf. Comput. Vis*, 2020. 1, 6, 8

[34] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021. 4, 6, 7

[35] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–851, 2019. 1, 6, 8

[36] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3, 5, 6, 8

[37] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 1, 2, 6, 7, 8