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 \text{(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 \text{(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 \text{(i.e., two-stage HOI detectors [1, 6–8, 10, 16, 18, 19, 21–24, 26, 28–30, 32, 33, 35]) or triplet matching \text{(i.e., one-stage HOI detectors [12, 20, 31])}. The additional complexity caused by this indirect inference structure and post-processing \text{(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.

However, due to the huge computational costs raised when processing multi-scale feature maps \text{(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-
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 TRasformer (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 $f_{q}^{sg} = \text{SAAttn}(z_q, x)$ 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

$$f_{q}^{sg} = \sum_{m=1}^{M} W_{m} \left[ \sum_{k \in \Omega_k} A_{mqk} \cdot W'_{mqk} \right],$$

where $A_{mqk}$ indicates an attention weight calculated with learnable weights $U_{m}, V_{m} \in \mathbb{R}^{C \times C}$ as $\text{exp}(\frac{z_{q}^{T}U_{m}^{T}V_{m}z_{k}}{\sqrt{C}})$. 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 $\Omega_q$ and $\Omega_k$ 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 $O(HWC^2 +$
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 \( \{x^l\}_{l=1}^{L} \) (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 \( \times 20 \) 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 a prediction 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 \( f^{ms}_q = \text{MSDeformAttn}(z_q, p_q, \{x^l\}_{l=1}^{L}) \) for query element \( z_q \) is calculated using a set of predicted sampling locations \( p_q \) as follows:

\[
f^{ms}_q = \sum_{m=1}^{M} W_m \left[ \sum_{l=1}^{L} \sum_{k=1}^{K} A_{mlqk} \cdot W'_m \Phi_{mlqk} \right],
\]

where \( l, k, m \) 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. \( \Phi_{mlqk} \) 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 \( (\sum_{i=1}^{L} H_i W_i) \), 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-
tions easier, given an HOI query feature \( q \), our Dual-Entity attention separately identifies sampling locations for the humans \((p^h_q)\) and objects \((p^o_q)\). First, we project \( q \) with two linear layers to obtain \( z^h_q \) and \( z^o_q \). The \( k^{th} \) sampling location at \( l^{th} \) feature level and \( m^{th} \) attention head for human and object are represented by

\[
p^h_m l q k = \phi_l(h_q) + \Delta h_m l q k, \\
p^o_m l q k = \phi_l(o_q) + \Delta o_m l q k,
\]

(3)

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

\[
 f^h_q = \text{MSDeformAttn}(z^h_q, p^h_q, \{x^l\}_{l=1}^L), \\
 f^o_q = \text{MSDeformAttn}(z^o_q, p^o_q, \{x^l\}_{l=1}^L).
\]

(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 \( \Delta c_q \) from the HOI query feature, obtaining \( p^c_m l q k = \phi_l(c_q) + \Delta c_m l q k \). Finally, the attended feature for contextual information \( f^c_q \) is computed using sampling location \( p^c_q \) as follows:

\[
f^c_q = \text{MSDeformAttn}(z_q, p^c_q, \{x^l\}_{l=1}^L).
\]

(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 Inference. Given the cross attention results of the final decoder layer where \( f^h_q \) and \( f^o_q \) is obtained by Eq. (4) and \( f^c_q \) is obtained by Eq. (5), the final prediction heads in MSTR predict the \( \langle \text{bbox}_h, \text{bbox}_o, \text{cls}_o, \text{act}_q \rangle \) using FFN as follows:

\[
\begin{align*}
(u_{qx}, u_{qy}, u_{qw}, u_{qh}) &= \text{FFN}_{\text{bbox}}(f^h_q), \\
(v_{qx}, v_{qy}, v_{qw}, v_{qh}) &= \text{FFN}_{\text{bbox}}(f^o_q), \\
\text{cls}_q &= \sigma(\text{FFN}_{\text{cls}}(f^o_q)), \\
\text{act}_q &= \sigma(\text{FFN}_{\text{act}}(f^c_q)),
\end{align*}
\]

where \( \text{cls}_q \) and \( \text{act}_q \) each denote predictions for object the class and the action class after sigmoid function, and final \( \text{bbox}_h \) is predicted with the center point \( \sigma(u_{qx} + \sigma^{-1}(h_{qx})), \sigma(u_{qy} + \sigma^{-1}(h_{qy})) \), width \( u_{qw} \), and height \( u_{qh} \). Likewise, the \( \text{bbox}_o \) is predicted with center point as \( \sigma(v_{qx} + \sigma^{-1}(o_{qx})), \sigma(v_{qy} + \sigma^{-1}(o_{qy})) \), width \( v_{qw} \), and height \( v_{qh} \). \( \sigma \) and \( \sigma^{-1} \) 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$_\text{role}$ over 25 interactions in two scenarios. HICO-DET contains 37,536 and 9,515 images for each training and test split with annotations for VERB, OBJECT, and ROLE 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 AP$_\text{role}^{#1}$ and +4.2p in AP$_\text{role}^{#1}$). 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).

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>AP$_\text{role}^{#1}$</th>
<th>AP$_\text{role}^{#2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIN (RP$_\text{C5}$) [19]</td>
<td>R50</td>
<td>47.8</td>
<td>-</td>
</tr>
<tr>
<td>Verb Embedding [32]</td>
<td>R50</td>
<td>45.9</td>
<td>-</td>
</tr>
<tr>
<td>RPNN [35]</td>
<td>R50</td>
<td>-</td>
<td>47.5</td>
</tr>
<tr>
<td>PMFNet [28]</td>
<td>R50-FPN</td>
<td>52.0</td>
<td>-</td>
</tr>
<tr>
<td>PastaNet [18]</td>
<td>R50-FPN</td>
<td>51.0</td>
<td>57.5</td>
</tr>
<tr>
<td>PD-Net [33]</td>
<td>R50</td>
<td>52.0</td>
<td>-</td>
</tr>
<tr>
<td>ACP [14]</td>
<td>R152</td>
<td>53.0</td>
<td>-</td>
</tr>
<tr>
<td>FCNet [21]</td>
<td>R50</td>
<td>53.1</td>
<td>-</td>
</tr>
<tr>
<td>ConsNet [22]</td>
<td>R50-FPN</td>
<td>53.2</td>
<td>-</td>
</tr>
<tr>
<td>Sequential HOI Detectors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VSRL [9]</td>
<td>R50-FPN</td>
<td>31.8</td>
<td>-</td>
</tr>
<tr>
<td>InteractNet [8]</td>
<td>R50-FPN</td>
<td>40.0</td>
<td>48.0</td>
</tr>
<tr>
<td>BAR-CNN [15]</td>
<td>R50-FPN</td>
<td>43.6</td>
<td>-</td>
</tr>
<tr>
<td>GPNN [24]</td>
<td>R152</td>
<td>44.0</td>
<td>-</td>
</tr>
<tr>
<td>iCAN [7]</td>
<td>R50</td>
<td>45.3</td>
<td>52.4</td>
</tr>
<tr>
<td>TIN (RC$_\text{P}$) [19]</td>
<td>R50</td>
<td>43.2</td>
<td>-</td>
</tr>
<tr>
<td>DCA [30]</td>
<td>R50</td>
<td>47.3</td>
<td>-</td>
</tr>
<tr>
<td>VCL [11]</td>
<td>R50-FPN</td>
<td>48.3</td>
<td>-</td>
</tr>
<tr>
<td>DRG [6]</td>
<td>R50-FPN</td>
<td>51.0</td>
<td>-</td>
</tr>
<tr>
<td>VSGNet [26]</td>
<td>R152</td>
<td>51.8</td>
<td>57.0</td>
</tr>
<tr>
<td>IDN [17]</td>
<td>R50</td>
<td>53.3</td>
<td>60.3</td>
</tr>
<tr>
<td>Parallel HOI Detectors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UnionDet [12]</td>
<td>R50-FPN</td>
<td>47.5</td>
<td>56.2</td>
</tr>
<tr>
<td>IPNet [31]</td>
<td>HG104</td>
<td>51.0</td>
<td>-</td>
</tr>
<tr>
<td>HOI Transformer [37]$^*$</td>
<td>R101</td>
<td>52.9</td>
<td>-</td>
</tr>
<tr>
<td>ASNet [4]$^*$</td>
<td>R50</td>
<td>53.9</td>
<td>-</td>
</tr>
<tr>
<td>GGNet [34]</td>
<td>HG104</td>
<td>54.7</td>
<td>-</td>
</tr>
<tr>
<td>HOTR [13]$^*$</td>
<td>R50</td>
<td>55.2</td>
<td>64.4</td>
</tr>
<tr>
<td>QPIC [25]$^*$</td>
<td>R50</td>
<td>58.8</td>
<td>61.0</td>
</tr>
</tbody>
</table>

Table 1. Comparison of performance on V-COCO test set. AP$_\text{role}^{#1}$, AP$_\text{role}^{#2}$ 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

$^1$https://github.com/YueLiao/PPDM
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.

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.
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 transformer-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


