Relational Context Learning for Human-Object Interaction Detection

Sanghyun Kim Deunsol Jung Minsu Cho
Pohang University of Science and Technology (POSTECH), South Korea
{sanghyun.kim, deunsol.jung, mscho}@postech.ac.kr
http://cvlab.postech.ac.kr/research/MUREN

Abstract

Recent state-of-the-art methods for HOI detection typically build on transformer architectures with two decoder branches, one for human-object pair detection and the other for interaction classification. Such disentangled transformers, however, may suffer from insufficient context exchange between the branches and lead to a lack of context information for relational reasoning, which is critical in discovering HOI instances. In this work, we propose the multiplex relation network (MUREN) that performs rich context exchange between three decoder branches using unary, pairwise, and ternary relations of human, object, and interaction tokens. The proposed method learns comprehensive relational contexts for discovering HOI instances, achieving state-of-the-art performance on two standard benchmarks for HOI detection, HICO-DET and V-COCO.

1. Introduction

The task of Human-Object Interaction (HOI) detection is to discover the instances of ⟨*human, object, interaction*⟩ from a given image, which reveal semantic structures of human activities in the image. The results can be useful for a wide range of computer vision problems such as human action recognition [1, 25, 42], image retrieval [9, 33, 37], and image captioning [12, 34, 36] where a comprehensive visual understanding of the relationships between humans and objects is required for high-level reasoning.

With the recent success of transformer networks [31] in object detection [2, 45], transformer-based HOI detection methods [4, 15, 16, 29, 38, 44, 46] have been actively developed to become a dominant base architecture for the task. Existing transformer-based methods for HOI detection can be roughly divided into two types: single-branch and two-branch. The single-branch methods [16, 29, 46] update a token set through a single transformer decoder and detect HOI instances using the subsequent FFNs directly. As a single transformer decoder is responsible for all sub-tasks (*i.e.,*

Figure 1. The illustration of relation context information in an HOI instance. We define three types of relation context information in an HOI instance: unary, pairwise, and ternary relation contexts. Each relation context provides useful information for detecting an HOI instance. For example, in our method, the unary context about an interaction (green) helps to infer that a human (yellow) and an object (red) are associated with the interaction, and vice versa. Our method utilizes the multiplex relation context consisting of the three relation contexts to perform context exchange for relational reasoning.

human detection, object detection, and interaction classification), they are limited in adapting to the different subtasks with multi-task learning, simultaneously [38]. To resolve the issue, the two-branch methods [4, 15, 38, 40, 44] adopt two separated transformer decoder branches where one detects human-object pairs from a human-object token set while the other classifies interaction classes between human-object pairs from an interaction token set. However, the insufficient context exchange between the branches prevents the two-branch methods [15, 38, 40] from learning relational contexts, which plays a crucial role in identifying HOI instances. Although some methods [4, 44] tackle this issue with additional context exchange, they are limited to propagating human-object context to interaction context.

To address the problem, we introduce the **MU**ltiplex

RElation **N**etwork (MUREN) that performs rich context exchange using unary, pairwise, and ternary relations of human, object, and interaction tokens for relational reasoning. As illustrated in Figure 1, we define three types of relation context information in an HOI instance: unary, pairwise, and ternary, each of which provides useful information to discover HOI instances. The ternary relation context gives holistic information about the HOI instance while the unary and pairwise relation contexts provide more fine-grained information about the HOI instance. For example, as shown in Figure 1, the unary context about an interaction (*e.g.,* 'riding') helps to infer which pair of a human and an object is associated with the interaction in a given image, and the pairwise context between a human and an interaction (*e.g.,* 'human' and 'riding') helps to detect an object (*e.g.,* 'bicycle'). Motivated by this, our multiplex relation embedding module constructs the context information that consists of the three relation contexts, thus effectively exploiting their benefits for relational reasoning. Since each sub-task requires different context information for relational reasoning, our attentive fusion module selects requisite context information for each sub-task from multiplex relation context and propagates the selected context information for context exchange between the branches. Unlike previous methods [4, 15, 38, 44], we adopt three decoder branches which are responsible for human detection, object detection, and interaction classification, respectively. Therefore, the proposed method learns discriminative representation for each sub-task.

We evaluate MUREN on two public benchmarks, HICO-DET [3] and V-COCO [10], showing that MUREN achieves state-of-the-art performance on two benchmarks. The ablation study demonstrates the effectiveness of the multiplex relation embedding module and the attentive fusion module. Our contribution can be summarized as follows:

- We propose multiplex relation embedding module for HOI detection, which generates context information using unary, pairwise, and ternary relations in an HOI instance.
- We propose the attentive fusion module that effectively propagates requisite context information for context exchange.
- We design a three-branch architecture to learn more discriminative features for sub-tasks, *i.e.*, human detection, object detection, and interaction classification.
- Our proposed method, dubbed MUREN, outperforms state-of-the-art methods on HICO-DET and V-COCO benchmarks.

2. Related Work

2.1. CNN-based HOI Methods.

Previous CNN-based HOI methods can be categorized into two groups: two-stage methods and one-stage methods. Two-stage HOI methods [7, 8, 13, 18, 19, 26, 30, 32, 39] first detect the human and the object instances using an off-the-shelf detector (*e.g.,* Faster R-CNN [27]) and predict the interaction between all possible pairs of a human and an object. To create discriminative instance features for HOI detection, they additionally utilize spatial features [8, 19, 35], linguistic features [7, 23], and human pose features [11, 19] with visual features. Some approaches [7, 26, 30, 32, 39] utilize the graph structure and exchange the context information of the instance features for relational reasoning between the nodes. DRG [7] proposes human-centric and object-centric graphs to perform context exchange focused on relevant context information. SCG [39] transforms and propagates the context information to the nodes in a graph conditioned on spatial relation. On the other hand, previous one-stage HOI methods [6, 14, 20] detect human-object pairs and classify the interactions between human-object pairs in an end-to-end manner. These methods utilize the interaction region to match the interaction and a pair of a human box and an object box. UnionDet [14] proposes a union-level detector to find the union box of human and object for matching a human-object pair. PPDM [20] detects interaction centers and points to the center point of the human and object box to predict HOI instances.

2.2. Transformer-based HOI Methods.

Inspired by DETR [2], a number of work [4, 15, 16, 29, 40, 44, 46] have adopted the transformer-based object detector to solve HOI detection. They can be divided into two folds: single-branch and two-branch methods. The single-branch methods [16, 29, 46] predict the HOI instances with a single transformer decoder. MSTR [29] utilizes multi-scale features to extract discriminative features for the HOI instances. In contrast, two-branch methods [4, 15, 38, 40, 44] adopt two transformer decoder branches, one is responsible for human-object pair detection and the other for interaction classification. HOTR [15] detects the instances in an image in detection branch and predicts the interaction with additional offsets to associate humans and objects in interaction branch. Although they extract discriminative features for each sub-task, there is no context exchange for relational reasoning, bringing performance degradation in HOI detection. To alleviate this, AS-NET [4] and DisTR [44] perform the message passing for relational reasoning between two branches. However, they only propagate human-object context information for interaction classification. In this paper, we exchange the context among branches with the multiplex relation context. The multiplex relation context,

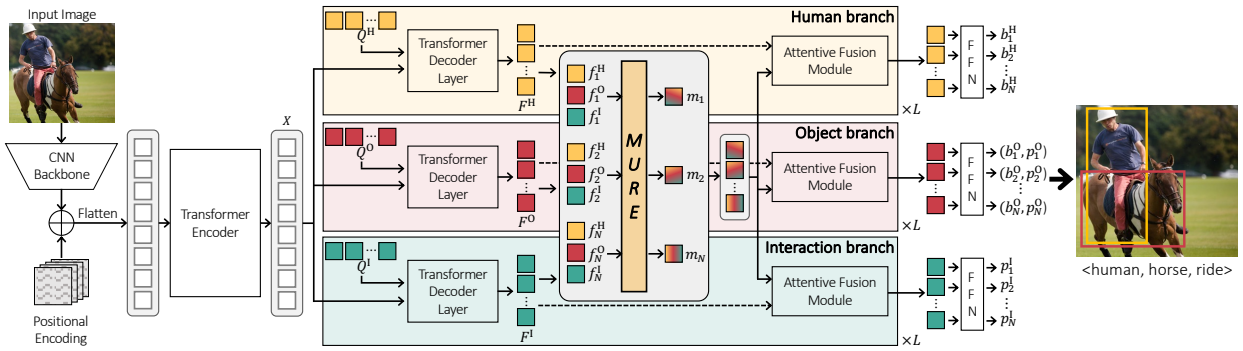

Figure 2. The overall architecture of MUREN. The proposed method adopts three-branch architecture: human branch, object branch, and interaction branch. Each branch is responsible for human detection, object detection, interaction classification. The input image is fed into the CNN backbone followed by the transformer encoder to extract the image tokens. A transformer decoder layer in each branch layer extracts the task-specific tokens for predicting the sub-task. The MURE takes the task-specific tokens as input and generates the multiplex relation context for relational reasoning. The attentive fusion module propagates the multiplex relation context to each sub-task for context exchange. The outputs at the last layer of each branch are fed into to predict the HOI instances.

which considers all relation contexts in an HOI instance, gives relational semantics for relational reasoning. We also extract more discriminative features for each sub-task via three-branch.

3. Problem Definition

Given an input image, the goal of HOI detection is to predict a visually-grounded set of HOI instances for object classes $\mathcal{O}$ and interaction classes $\mathcal{I}$. An HOI instance consists of four components: a bounding box of human $\mathbf{b}_i^{\mathrm{H}} \in \mathbb{R}^4$, a bounding box of object $\mathbf{b}_i^{\mathrm{O}} \in \mathbb{R}^4$, a one-hot vector of object label $\mathbf{c}_i^{\mathrm{O}} \in \{0,1\}^{|\mathcal{O}|}$, and a one-hot vector of interaction label $\mathbf{c}_i^{\mathrm{I}} \in \{0,1\}^{|\mathcal{I}|}$, where $|\cdot|$ denotes the size of a set. The output of HOI detection is thus expressed by a set of HOI instances $\{(\mathbf{b}_i^{\mathrm{H}}, \mathbf{b}_i^{\mathrm{O}}, \mathbf{c}_i^{\mathrm{O}}, \mathbf{c}_i^{\mathrm{I}})\}$.

4. Method

The proposed network, MUREN, is illustrated in Figure 2. Given an input image, it extracts image tokens via a CNN backbone followed by a transformer encoder. The image tokens are fed to three independent branches to perform three sub-task: human detection, object detection, and interaction classification. In each branch, a transformer decoder layer refines N learnable tokens using the image tokens as keys and values to extract task-specific tokens. Using the task-specific tokens of each branch, our multiplex relation embedding module (MURE) generates the context information for relational reasoning. The attentive fusion module then integrates the context information across the task-specific tokens for human, object, and interaction branches, propagating the results to the next layer. After repeating this process for L times, FFNs predict the set of HOI instances. In the remainder of this section, we explain the details of each component in MUREN.

4.1. Image Encoding

Following the previous work [2, 29, 47], we use a transformer encoder with a CNN backbone to extract image tokens. The CNN backbone takes an input image to extract an image feature map. The image feature map is fed into 1×1 convolution layer to reduce the channel dimension to D, and the positional encoding [2] is added to the image feature map to reflect the spatial configuration of the feature map. The feature map is then tokenized by flattening and fed into the transformer encoder to produce image tokens $\mathbf{X} \in \mathbb{R}^{T \times D}$ for the subsequent networks, where T and D are the number of the image tokens and the channel dimension, respectively.

4.2. HOI Token Decoding

Different from previous two-branch methods [4, 15, 44], we design an architecture consisting of three branches which is responsible for human detection, object detection, and interaction classification, respectively. Each branch τ, consisting of L layers, takes the tokens $\mathbf{Q}^\tau = \{\mathbf{q}_i^\tau\}_{i=1}^N$ and the image tokens $\mathbf{X}$ as inputs , where $\tau \in \{\mathrm{H}, \mathrm{O}, \mathrm{I}\}$ indicates human, object, and interaction respectively. At each layer, $\mathbf{Q}^\tau$ is refined through a transformer decoder layer followed by a MURE module and an attentive fusion module. Specifically, the three branches take learnable tokens $\mathbf{Q}^{\mathrm{H}}, \mathbf{Q}^{\mathrm{O}}, \mathbf{Q}^{\mathrm{I}} \in \mathbb{R}^{N \times D}$ for human, object, and interaction branches, respectively. In l-th layer of the branch τ, a transformer decoder layer $\mathrm{Dec}_{(l)}^\tau$ updates $\mathbf{Q}_{(l-1)}^\tau$, the output of previous layer of the branch τ, by attending $\mathbf{X}$ to generate task-specific tokens $\mathbf{F}_{(l)}^\tau = \{\mathbf{f}_{(l),i}^\tau\}_{i=1}^N$ which contain the context information for predicting a sub-task which the branch τ is responsible for:

$$\mathbf{F}_{(l)}^\tau = \mathrm{Dec}_{(l)}^\tau(\mathbf{Q}_{(l-1)}^\tau, \mathbf{X}), \tag{1}$$

where $\mathrm{Dec}(q, kv)$ denotes a transformer decoder layer.

4.3. Relational Contextualization

As mentioned above, relational reasoning is crucial to identify HOI instances. However, since the task-specific tokens are generated from the separated branches, the tokens suffer from a lack of relational context information. To mitigate this issue, we propose multiplex relation embedding module (MURE) which generates multiplex relation context for relational reasoning. The multiplex relation context contains the unary, pairwise, and ternary relation contexts to exploit useful information in each relation context, as shown in Figure 3.

Specifically, the MURE first constructs the ternary relation context $\mathbf{f}_i^{\mathrm{HOI}} \in \mathbb{R}^D$ for i-th HOI instance by concatenating each $\mathbf{f}_i^{\tau}$ followed by an MLP layer.

$$\mathbf{f}_i^{\mathrm{HOI}} = \mathrm{MLP}([\mathbf{f}_i^{\mathrm{H}}; \mathbf{f}_i^{\mathrm{O}}; \mathbf{f}_i^{\mathrm{I}}]), \tag{2}$$

where $[\cdot;\cdot]$ is a concatenation operation. We omit the subscript l for the sake of simplicity. Since the ternary relation takes the overall understanding of each sub-task into account, it gives holistic context information about the HOI instance. On the other hand, since the unary and the pairwise relations take a fine-grained level understanding of each sub-task into account, they give the fine-grained context information about the HOI instance. To exploit both holistic and fine-grained context information, we embed the unary and the pairwise relation contexts within the ternary relation context with a sequential manner.

In detail, we apply a self-attention on a set of i-th task-specific tokens $\{\mathbf{f}_i^{\mathrm{H}}, \mathbf{f}_i^{\mathrm{O}}, \mathbf{f}_i^{\mathrm{I}}\}$ to consider the unary relation for i-th HOI instance as Eq. 3. Then, the unary-relation context U_i is embedded into ternary relation context using a cross-attention as Eq. 4:

$$U_i = \mathrm{SelfAttn}(\{\mathbf{f}_i^{\mathrm{H}}, \mathbf{f}_i^{\mathrm{O}}, \mathbf{f}_i^{\mathrm{I}}\}), \tag{3}$$

$$\tilde{\mathbf{f}}_i^{\mathrm{HOI}} = \mathrm{CrossAttn}(\mathbf{f}_i^{\mathrm{HOI}}, U_i), \tag{4}$$

where we denote $\mathrm{SelfAttn}(\cdot)$ as a self-attention operation and $\mathrm{CrossAttn}(q, kv)$ as a cross-attention operation for simplicity. To embed the pairwise relation context within the ternary relation context, we extract the pairwise features of $\mathbf{f}^{\mathrm{HO}}, \mathbf{f}^{\mathrm{HI}}, \mathbf{f}^{\mathrm{OI}} \in \mathbb{R}^D$ for respective human-object, human-interaction, object-interaction relation as follows:

$$\mathbf{f}_i^{\mathrm{HO}} = \mathrm{MLP}([\mathbf{f}_i^{\mathrm{H}}; \mathbf{f}_i^{\mathrm{O}}]), \tag{5}$$

$$\mathbf{f}_i^{\mathrm{HI}} = \mathrm{MLP}([\mathbf{f}_i^{\mathrm{H}}; \mathbf{f}_i^{\mathrm{I}}]), \tag{6}$$

$$\mathbf{f}_i^{\mathrm{OI}} = \mathrm{MLP}([\mathbf{f}_i^{\mathrm{O}}; \mathbf{f}_i^{\mathrm{I}}]). \tag{7}$$

Similar to the above, we apply the self attention on a set of pairwise features to consider the pairwise relation for i-th HOI instance, and the cross attention to embed the pairwise relation contexts within ternary relation context:

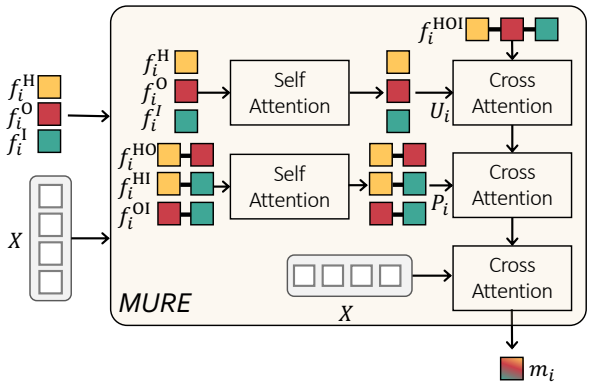

Figure 3. The architecture of the multiplex relation embedding module (MURE). MURE takes i-th task-specific tokens and the image tokens as input, and embed the unary and pairwise relation contexts into the ternary relation context. The multiplex relation context, the output of MURE, is fed into subsequent attentive fusion module for context exchange.

$$P_i = \mathrm{SelfAttn}(\{\mathbf{f}_i^{\mathrm{HO}}, \mathbf{f}_i^{\mathrm{HI}}, \mathbf{f}_i^{\mathrm{OI}}\}), \tag{8}$$

$$\hat{\mathbf{f}}_i^{\mathrm{HOI}} = \mathrm{CrossAttn}(\tilde{\mathbf{f}}_i^{\mathrm{HOI}}, P_i). \tag{9}$$

Finally, the $\hat{\mathbf{f}}_i^{\mathrm{HOI}}$ is transformed to generate the multiplex relation context $\mathbf{m}_i$ as follows by attending the image tokens $\mathbf{X}$:

$$\mathbf{m}_i = \mathrm{CrossAttn}(\hat{\mathbf{f}}_i^{\mathrm{HOI}}, \mathbf{X}). \tag{10}$$

It is noteworthy that our high-order (ternary and pairwise) feature functions have a form of non-linear function, *i.e.*, MLP, on top of a tuple of multiple inputs, which is not reducible to a sum of multiple functions of individual lower-order inputs in general. Such a high-order feature function thus can learn the structural relations of the inputs in the tuple, considering all the inputs jointly. For example, a ternary function of three coordinates $f(a, b, c)$ can compute the angle feature between $\overline{ab}$ and $\overline{ac}$, which cannot be computed by an individual unary function, $g(a)$, $g(b)$, or $g(c)$ as well as their linear combination. In a similar vein, our ternary feature functions, *i.e.*, Eq. 2, can effectively learn to capture structural relations which are not easily composable from unary and pairwise feature functions.

4.4. Attentive Fusion

Our attentive fusion module aims to propagate the multiplex relation context to the task-specific tokens for context exchange. Since each sub-task requires different context information for relational reasoning, the multiplex relation context is transformed using MLP with each task-specific token to propagate the context information conditioned on each sub-task. We further utilize the channel attention to select the requisite context information for each sub-task.

Then, the refined tokens $\mathbf{Q}^{\tau}_{(l)}$, the output of l-th layer of branch τ, is generated by propagating the requisite context information to the task-specific tokens $\mathbf{F}^{\tau}_{(l)}$. Formally, the channel attention α and the refined tokens $\mathbf{Q}^{\tau}_{(l)}$ are formulated as follows:

$$\alpha = \sigma(\text{MLP}([\mathbf{f}^{\tau}_{(l),i}; \mathbf{m}_{(l),i}])) \tag{11}$$

$$\mathbf{q}^{\tau}_{(l),i} = \mathbf{f}^{\tau}_{(l),i} + \alpha \odot \text{MLP}([\mathbf{f}^{\tau}_{(l),i}; \mathbf{m}_{(l),i}]), \tag{12}$$

where we denote $\odot$ and σ as element-wise multiplication, and sigmoid function, respectively. As the refined tokens $\mathbf{Q}^{\tau}_{(l)}$ is generated via context exchange with the multiplex relation context, it deduces the comprehensive relational understanding to discover HOI instances.

The $\mathbf{Q}^{\tau}_{(L)}$, the output of last layer of branch τ, is fed into FFNs to predict a set of the HOI predictions. Formally, given the $\mathbf{Q}^{\tau}_{(L)}$, the MUREN predicts a set of HOI predictions $\{(\mathbf{b}^{\text{H}}_i, \mathbf{b}^{\text{O}}_i, \mathbf{p}^{\text{O}}_i, \mathbf{p}^{\text{I}}_i)\}^{N}_{i=1}$ using FFNs as follows:

$$\mathbf{b}^{\text{H}}_i = \text{FFN}_{\text{hbox}}(\mathbf{q}^{\text{H}}_{(L),i}) \in \mathbb{R}^4, \tag{13}$$

$$\mathbf{b}^{\text{O}}_i = \text{FFN}_{\text{obox}}(\mathbf{q}^{\text{O}}_{(L),i}) \in \mathbb{R}^4, \tag{14}$$

$$\mathbf{p}^{\text{O}}_i = \delta(\text{FFN}_{\text{oc}}(\mathbf{q}^{\text{O}}_{(L),i})) \in \mathbb{R}^{|\mathcal{O}|}, \tag{15}$$

$$\mathbf{p}^{\text{I}}_i = \sigma(\text{FFN}_{\text{ic}}(\mathbf{q}^{\text{I}}_{(L),i})) \in \mathbb{R}^{|\mathcal{I}|}, \tag{16}$$

where δ is a softmax operation, and $\mathbf{p}^{\text{O}}_i$, $\mathbf{p}^{\text{I}}_i$ are class probability of object and interaction, respectively.

4.5. Training Objective

For training our proposed method, we follow previous transformer-based methods [29, 38, 44]. We adopt the Hungarian Matching [17] to assign the ground-truth HOI instances to the predictions. MUREN is trained with multi-task loss composed of four losses: L1 loss [27] $\mathcal{L}_{\text{L1}}$ and GIoU loss [28] $\mathcal{L}_{\text{GIoU}}$ for the bounding box regression, cross-entropy loss $\mathcal{L}_{oc}$ for the object classification, and focal loss [21] $\mathcal{L}_{ic}$ for the interaction classification. The total loss $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \lambda_{\text{L1}}\mathcal{L}_{\text{L1}} + \lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}} + \lambda_{\text{oc}}\mathcal{L}_{\text{oc}} + \lambda_{\text{ic}}\mathcal{L}_{\text{ic}}, \tag{17}$$

where λ_{L1}, λ_{GIoU}, λ_{oc}, and λ_{ic} are the hyper-parameters for weighting each loss. Additionally, we apply intermediate supervision for better representation learning. Specifically, we attach the same FFNs to each decoding branch layer to calculate the intermediate loss. This auxiliary loss is computed the same as $\mathcal{L}$.

4.6. Inference

Given the set of HOI predictions, we generate a set of HOI instances $\{(\mathbf{b}^{\text{H}}_i, \mathbf{b}^{\text{O}}_i, \mathbf{c}^{\text{O}}_{i,j'}, \mathbf{c}^{\text{I}}_{i,t}) | \, i \in N, \, k \in \mathbb{R}^{|\mathcal{I}|}, \, j' = \text{argmax}_j \mathbf{p}^{\text{O}}_{i,j}\}$, where $c^{O}_{i,j'} \in \mathbb{R}^{|\mathcal{O}|}$, $c^{I}_{i,t} \in \mathbb{R}^{|\mathcal{I}|}$ are one-hot vectors with the j-th and t-th index set to 1, respectively. Following [38], we then select top-k score HOI instances, where the score is given by $\mathbf{p}^{\text{O}}_{i,j'} \cdot \mathbf{p}^{\text{I}}_{i,t}$.

5. Experiments

5.1. Datasets and Metrics

We evaluate our model on the two public benchmark datasets: HICO-DET [3] and V-COCO [10].

HICO-DET has 38,118 images for training and 9,658 images for testing. It contains 80 object classes, 117 interaction classes and 600 HOI classes, which are a pair of an object class and an interaction class (*e.g.,* 'riding bicycle'). We evaluate the proposed method on Default and Known Object settings. In the Default setting, the AP is calculated across all testing images for each HOI class. The Known Object setting calculates the AP of an HOI class over the images containing the object in the HOI class (*e.g.,* the AP of an HOI class 'riding bicycle' is only calculated on the images which contain the object 'bicycle'). Following the previous work [38], we report the mAP under three splits (Full, Rare, and Non-Rare) for each setting. The Full, Rare, and Non-Rare splits contain all 600 HOI classes, 138 HOI classes, which have less than 10 training samples for each class, and 462 HOI classes, which have more than 10 training samples for each class, respectively.

V-COCO is a subset of the MS-COCO [22] dataset. It consists of 5400 and 4,946 images for training, and testing. It has 80 object classes and 29 action classes. Following the evaluation settings in [15], we evaluate the proposed method on scenario 1 and scenario 2, and report role average precision under two scenarios ($\text{AP}^{\#1}_{\text{role}}$ for scenario 1 and $\text{AP}^{\#2}_{\text{role}}$ for scenario 2). In scenario 1, the model should predict the bounding box of the occluded object as [0,0,0,0]. In contrast, the predicted bounding box of the occluded object is ignored on calculating the AP_{role} in scenario 2.

5.2. Implementation Details

The encoder in MUREN adopts ResNet-50 as a CNN backbone followed by a 6-layer transformer encoder. We set the number of branch layers L to 6. For the training, we set the number of queries N to 64 for HICO-DET and 100 for V-COCO following [38]. The weight of loss λ_{L1}, λ_{GIoU}, λ_{oc}, λ_{ic} is set to 2.5, 1, 1, 1, respectively. The network is initialized with the parameters of DETR [2] pretrained on MS-COCO [22]. We optimize our network by AdamW [24] with the weight decay $1e{-}4$. We set the initial learning rate of the CNN backbone to $1e{-}5$ and the other component to $1e{-}4$. The model is trained with 100 epoch. For the V-COCO, we freeze the CNN backbone to prevent overfitting, and set the learning rate to $4e{-}5$. All experiments are conducted with a batch size of 16 on 4 RTX 3090 GPUs.

5.3. Comparison with State-of-the-Art

Table 1 and Table 2 show the performance comparison of the proposed method with the previous HOI methods. As shown in Table 1, on the HICO-DET dataset, the pro-

Method	Backbone	Feature	Default			Known Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
CNN-based methods								
iCAN [8]	R50	A+S	14.84	10.45	16.15	16.26	11.33	17.73
TIN [19]	R50	A+S+P	22.90	14.97	25.26	-	-	-
GPNN [26]	R101	A	13.11	9.34	14.23	-	-	-
DRG [7]	R50-FPN	A+S+L+M	24.53	19.47	26.04	27.98	23.11	29.43
VSGNet [30]	R152	A+S	19.80	16.05	20.91	-	-	-
wang *et al*. [32]	R50-FPN	A+S+M	17.57	16.85	17.78	21.00	20.74	21.08
IDN [18]	R50	A+S	26.29	22.61	27.39	28.24	24.47	29.37
VCL [13]	R50	A	23.63	17.21	25.55	25.98	19.12	28.03
UnionDet [14]	R50	A	17.58	11.72	19.33	19.76	14.68	21.27
GGNet [43]	HG104	A	28.83	22.13	30.84	27.36	20.23	29.48
SCG [39]	R50-FPN	A+S+M	31.33	24.72	33.31	34.37	27.18	36.52
Transformer-based methods								
PST [5]	R50	A	23.93	14.98	26.60	26.42	17.61	29.05
HoiTrans [46]	R101	A	26.61	19.15	28.84	29.13	20.98	31.57
HOTR [15]	R50	A	25.10	17.34	27.42	-	-	-
AS-Net [4]	R50	A	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [29]	R101	A	29.90	23.92	31.69	32.38	26.06	34.27
MSTR [16]	R50	A+M	31.17	25.31	32.92	34.02	28.83	35.57
CDN [38]	R101	A	32.07	27.19	<u>33.53</u>	34.79	29.48	36.38
UPT [40]	R50	A+S	31.66	25.94	33.36	35.05	29.27	<u>36.77</u>
DisTR [44]	R50	A	31.75	27.45	33.03	34.50	30.13	35.81
STIP [41]	R50	A+S+L	<u>32.22</u>	<u>28.15</u>	33.43	<u>35.29</u>	**31.43**	36.45
Ours	R50	A	**32.87**	**28.67**	**34.12**	**35.52**	<u>30.88</u>	**36.91**

Table 1. Performance comparison on the HICO-DET [3] dataset. The letters in Feature column stand for A: Appearance/Visual features, S: Spatial features, L: Linguistic features, P: Human pose features, M: Multi-scale features. The best score is highlighted in bold, and the second-best score is underscored.

posed method achieves state-of-the-art performance on Default and Known Object settings against existing CNN- and transformer-based methods. Compared with the previous CNN-based methods [7, 26, 30, 32, 39], which utilize the graph structure for context exchange, MUREN shows significant improvements. We also surpass the previous single-branch methods [16, 29, 46]. It illustrates that it is crucial extracting the task-specific tokens for each sub-task with different branches. In particular, we outperform the previous two-branch methods [4, 15, 38, 40, 44]. DisTR [44] and AS-NET [4] perform context exchange for relational reasoning, but they only propagate the context information of the human and the object to the interaction branch for interaction classification. Instead, we exchange the context information among the three branches, selecting requisite context information from the multiplex relation context for each sub-task. These results illustrate the advantage of context exchange between each branch using the multiplex relation context for relational reasoning. Moreover, MUREN shows better performance without using any additional information (*e.g.,* spatial and linguistic information) compared with [16, 39–41]. We also outperform [29, 38, 46] which utilize a deeper backbone to extract discriminative features for each sub-task. These results illustrate that three-branch architecture and context exchange with multiplex relation context for relational reasoning provide more discriminative features to predict each sub-task. We further evaluate MUREN on the V-COCO dataset and observe similar results as in the HICO-DET dataset. As shown in Table 2, MUREN achieves state-of-the-art performances across all the metrics compared with existing methods.

5.4. Ablation Study

We conduct various ablation studies on the V-COCO dataset to validate the effectiveness of MUREN.

Impact of each relation context information on relational reasoning. We utilize the multiplex relation context, which contains the unary, pairwise, and ternary relation context, for relational reasoning. To investigate the impact of each relation context information on relational reasoning, we gradually add each relation context information to the baseline, which predicts the HOI instances without context exchange among each branch for relational reasoning. As shown in Table 3, we observe that context exchange using the ternary relation context gives 4.55%p, 4.22%p improvement with a large margin in $\mathrm{AP}^{\#1}_{\mathrm{role}}$ and $\mathrm{AP}^{\#2}_{\mathrm{role}}$, respectively. This result indicates that context exchange for relational reasoning is essential for discovering the HOI instance and ternary relation context promotes relational reasoning providing holistic information about the HOI instances. Besides, when the model exploits ternary and unary relation contexts, the model shows an additional

Method	Backbone	Feature	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
CNN-based methods				
GPNN [26]	R101	A	44.0	-
iCAN [8]	R50	A+S	45.3	52.4
TIN [19]	R50	A+S+P	47.8	54.2
VSGNet [30]	R152	A+S	51.8	57.0
DRG [7]	R50-FPN	A+S+L+M	51.0	-
VCL [13]	R101	A	48.3	-
UnionDet [14]	R50	A	47.5	56.2
GGNet [43]	HG104	A	54.7	-
IDN [18]	R50	A+S	53.3	60.3
SCG [39]	R50-FPN	A+S+M	54.2	60.9
Transformer-based methods				
QPIC [29]	R50	A	58.8	61.0
MSTR [16]	R50	A+M	62.0	65.2
HOTR [15]	R50	A	55.2	61.0
AS-NET [4]	R50	A	53.9	-
CDN [38]	R101	A	63.9	65.9
UPT [40]	R50	A	59.0	64.5
STIP [41]	R50	A+S+L	66.0	<u>70.7</u>
DisTR [44]	R50	A	<u>66.2</u>	68.5
Ours	R50	A	**68.8**	**71.0**

Table 2. Performance comparison on V-COCO [10] dataset. The letters in Feature column stand for A: Appearance/Visual features, S: Spatial features, L: Linguistic features, P: Human pose features, M: Multi-scale features. The best score is highlighted in bold, and the second-best score is underscored.

performance improvement. We observe similar results on the model which utilizes both ternary and pairwise relation contexts. It indicates that the fine-grained relation contexts provide useful information for relational reasoning to predict HOI instances. When we use all the relation context information in HOI instance, the model shows a significant performance increase of 6.23%p and 5.86%p in $AP^{\#1}_{role}$ and $AP^{\#2}_{role}$, compared with the baseline. It demonstrates that each relation context information complements the others, and thus the multiplex relation context provides rich information for relational reasoning and brings performance gain in HOI detection.

Impact of the multiplex relation context on each sub-task. For investigating the propagation impact of the multiplex relation context on the sub-tasks, we gradually add the propagation the multiplex relation context to each branch. When we propagate the multiplex relation context to one of the detection branches (*i.e.,* human branch and object branch), we observe that the model consistently shows performance improvement compared with the baseline, as shown in Table 4. We also observe the performance gains when the model propagates the multiplex relation context to both human and object branch. It indicates that relational context information is required to detect the human and the object in the HOI detection. In particular, when the model propagates the multiplex relation context to the interaction branch, MUREN shows the notable performance gains of 3.19%p and 2.77%p on scenario 1 and scenario 2. It indicates that the multiplex relation context is essential to interaction classification which requires a comprehensive relational understanding between the human and the object. The entire model of MUREN, which propagates the relation context information to all sub-tasks, achieves the highest performance with a significant margin compared with the other model variants. The results demonstrate that context exchange among the three branches is essential to identify HOI instances and plays a crucial role in the comprehensive relational understanding.

ternary	unary	pairwise	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
-	-	-	62.52	65.14
✓	-	-	67.07	69.36
✓	✓	-	68.12	70.31
✓	-	✓	67.67	70.02
✓	✓	✓	**68.75**	**71.00**

Table 3. The impact of each relation context information on relational reasoning. The 'ternary', 'unary', and 'pairwise' columns indicate the ternary, unary and pairwise relation context.

human	object	interaction	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
-	-	-	62.52	65.14
✓	-	-	64.44	66.62
-	✓	-	63.66	66.00
✓	✓	-	65.29	67.5
-	-	✓	65.71	67.91
✓	✓	✓	**68.75**	**71.00**

Table 4. The impact of the multiplex relation context on each sub-task. The 'human', 'object', and 'interaction' columns indicate the propagation of the multiplex relation context to human, object, and interaction branch, respectively.

conditioning	channel	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
-	-	66.50	68.96
✓	-	66.95	69.23
-	✓	67.10	69.49
✓	✓	**68.75**	**71.00**

Table 5. Ablations studies on each component in the attentive fusion module. 'conditioning' and 'channel' indicate transforming multiplex relation context conditioned on a task-specific token and channel attention mechanism.

Impact of attentive fusion module on context exchange. MUREN exchanges relational context information between each branch via the attentive fusion module. To investigate the impact of the attentive fusion module, we remove the attentive fusion module and fuse both the task-specific tokens and the multiplex relation context with an element-wise addition operation for the baseline. As shown in Table 5, the performance drops by 2.25%p and 2.04%p in the two scenarios. It shows the effectiveness of our attentive fusion module for context exchange between the branches.

Method	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$	Params (M)
MUREN-(0)	68.8	71.0	69.3
MUREN-(3)	67.1	69.3	64.3
MUREN-(6)	66.6	69.1	59.6
MUREN†	68.3	70.6	59.6

Table 6. The Impact of disentangling human and object branches. MUREN-(k) denotes the sharing of parameters between the human and object branches across k layers. The parameters are shared only between corresponding layers. MUREN† is variant of MUREN by adjusting the number of layer L.

Impact of the context information selection for each sub-task. In the attentive fusion module, we select requisite context information for each sub-task from the multiplex relation context. We further analyze the impact of the context information selection as shown in Table 5. To select the requisite context information for each sub-task, we utilize 1) transforming multiplex relation context conditioned on a task-specific token ('conditioning' in Table 5) and 2) channel attention mechanism ('channel' in Table 5). We observe that the model, which utilizes one of 'conditioning' and 'channel', gains performance improvement. We also observe that the model with both 'conditioning' and 'channel' shows better performance than the other model variants. The results demonstrate that each sub-task requires different context information for relational reasoning, and thus it is important to propagate the requisite context for each sub-task. Our attentive fusion module effectively selects requisite context information for each sub-task.

Impact of disentangling human and object branches. Human plays a central and an active role for HOI, which is distinctive from a relatively passive role of object, and thus requires a dedicated module to capture relevant attributes and semantics such as pose and clothing. We evaluated in Table 6 the effect of sharing parameters between human and object branches; we gradually increased the number of layers that share parameters between the two branches. The results show that increasing the number of shared layers drops the performance and the full-sharing model, MUREN-(6), results in 2.2%p and 1.9%p decrease in performance at two scenarios, respectively, compared with non-sharing model, MUREN-(0). This is a significant drop also compared to MUREN†, which has a similar number of parameters with MUREN-(6) by adjusting the number of layer L of MUREN, indicating that separating human and object branches is important indeed for HOI detection.

5.5. Qualitative Results

We visualize HOI detection results and the cross attention map of each branch and the multiplex relation embedding module (MURE) in Fig. 4. As shown in Fig. 4b, c, the human and the object branches focus on the instance extremities to detect the human and the object. In the Fig. 4d, we observe that the interaction branch attends to the regions where the interaction exists between the human and the object. These results indicate that the task-specific tokens contain context information for predicting each sub-task. We also observe that the cross-attention map in MURE highlights the overall region that contains the relational semantics about the HOI instance as shown in Fig. 4e. It demonstrates that MURE captures the context information about HOI instance for relational reasoning.

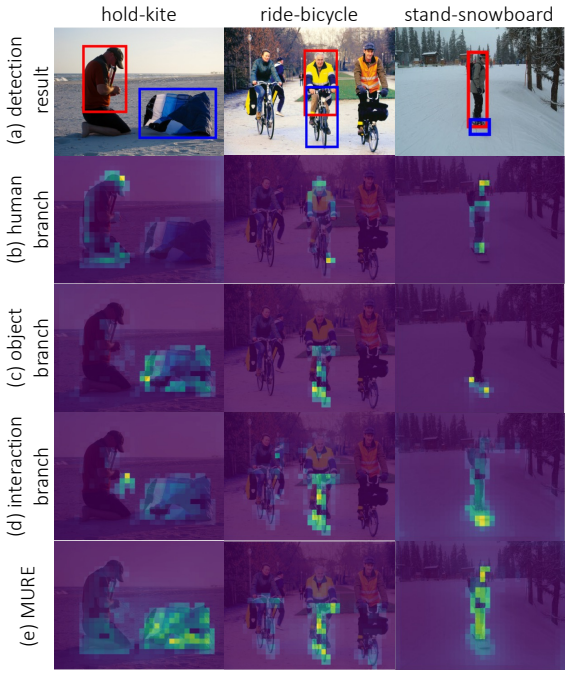

Figure 4. The visualization of the HOI detection results and the cross-attention map in each branch and the multiplex relation embedding module (MURE). Best viewed in color.

6. Conclusion

We have proposed MUREN, a one-stage method that effectively performs the context exchange between the three branches for HOI detection. By leveraging relation contexts for relational reasoning in MURE and using the attention fusion module to select requisite context information for each sub-task, MUREN can learn discriminative features to predict each sub-task. Our extensive experiments demonstrate the importance of context exchange between the branches and the effectiveness of MUREN, which achieves state-of-the-art performance on both HICO-DET and V-COCO benchmarks and its components.

Acknowledgements. This work was supported by the IITP grants (2021-0-00537: Visual common sense through self-supervised learning for restoration of invisible parts in images (50%), 2021-0-02068: AI Innovation Hub (40%), and 2019-0-01906: AI graduate school program at POSTECH (10%)) funded by the Korea government (MSIT).

References

[1] Carlo Bretti and Pascal Mettes. Zero-shot action recognition from diverse object-scene compositions. *arXiv preprint arXiv:2110.13479*, 2021. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 3, 5

[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. 2, 5, 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, 3, 6, 7

[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. 6

[6] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1291–1299, 2021. 2

[7] 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. 2, 6, 7

[8] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. 2, 6, 7

[9] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6589–6598, 2017. 1

[10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 5, 7

[11] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9677–9685, 2019. 2

[12] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[13] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020. 2, 6, 7

[14] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020. 2, 6, 7

[15] 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, 3, 5, 6, 7

[16] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022. 1, 2, 6, 7

[17] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[18] 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*, 33:5011–5022, 2020. 2, 6, 7

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 2, 6, 7

[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. 2

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[23] 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. 2

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[25] Gyeongsik Moon, Heeseung Kwon, Kyoung Mu Lee, and Minsu Cho. Integralaction: Pose-driven feature integration for robust human action recognition in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2021. 1

[26] 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. 2, 6, 7

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 5

[28] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 5

[29] 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, 3, 5, 6, 7

[30] 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. 2, 6, 7

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[32] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *European Conference on Computer Vision*, pages 248–264. Springer, 2020. 2, 6

[33] Hui Wu, Min Wang, Wengang Zhou, Houqiang Li, and Qi Tian. Contextual similarity distillation for asymmetric image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9489–9498, June 2022. 1

[34] Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. Difnet: Boosting visual information flow for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18020–18029, June 2022. 1

[35] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[36] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 1

[37] Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10718–10726, 2021. 1

[38] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34:17209–17220, 2021. 1, 2, 5, 6, 7

[39] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 2, 6, 7

[40] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 1, 2, 6, 7

[41] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19548–19557, 2022. 6, 7

[42] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9975–9984, 2019. 1

[43] 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. 6, 7

[44] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022. 1, 2, 3, 5, 6, 7

[45] 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. 1

[46] 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

[47] 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. 3