Re-mine, Learn and Reason: Exploring the Cross-modal Semantic Correlations for Language-guided HOI detection

Yichao Cao
Southeast University
caoyichao@seu.edu.cn

Qingfei Tang
Nanjing Enbo Tech.
qingfeitang@gmail.com

Feng Yang
Southeast University
yangfeng@seu.edu.cn

Xiu Su*
University of Sydney
xisu5992@uni.sydney.edu.au

Shan You
SenseTime
youshan@sensetime.com

Xiaobo Lu
Southeast University
xblu@seu.edu.cn

Chang Xu
University of Sydney
c.xu@sydney.edu.au

Abstract

*Human-Object Interaction (HOI) detection is a challenging computer vision task that requires visual models to address the complex interactive relationship between humans and objects and predict <$human, action, object$> triplets. Despite the challenges posed by the numerous interaction combinations, they also offer opportunities for multi-modal learning of visual texts. In this paper, we present a systematic and unified framework (**RmLR**) that enhances HOI detection by incorporating structured text knowledge. Firstly, we qualitatively and quantitatively analyze the loss of interaction information in the two-stage HOI detector and propose a re-mining strategy to generate more comprehensive visual representation. Secondly, we design more fine-grained sentence- and word-level alignment and knowledge transfer strategies to effectively address the many-to-many matching problem between multiple interactions and multiple texts. These strategies alleviate the matching confusion problem that arises when multiple interactions occur simultaneously, thereby improving the effectiveness of the alignment process. Finally, HOI reasoning by visual features augmented with textual knowledge substantially improves the understanding of interactions. Experimental results illustrate the effectiveness of our approach, where state-of-the-art performance is achieved on public benchmarks.*

*Corresponding author.

1. Introduction

Human-object interaction (HOI) detection [16, 6] is an emerging field of research that builds upon object detection and requires more advanced high-level visual understanding. A high-performing HOI detector should not only accurately localize all interacting Human-Object pairs but also recognize their specific interactions, typically represented as an HOI triplet in the format of <$human, action, object$> [67].

Previous approaches for achieving HOI detection can be divided into two pipelines: those that treat object detection and interaction recognition as separate stages [62, 6, 13, 14, 33, 20], and those that aim to handle both simultaneously [15, 26, 66, 35, 7]. Although both paradigms have made significant progress, the task remains challenging due to the vast variety of human-object interaction combinations in the real world [58, 59]. For example, the HICO-DET dataset [6] contains 600 human-object interaction combinations. A common approach is to optimize the model by mapping these various triplet labels into a discrete one-hot labels. However, this method oversimplifies the intricacy of the HOI task and can be cumbersome for model optimization.

In recent years, multi-modal learning has gained significant attention in the vision-and-language learning domain, where it has achieved state-of-the-art performance on various tasks [25, 3, 4, 30, 1, 23]. By integrating information from multiple modalities, such as images [49, 47, 50, 48] and text [64], multi-modal learning can provide a more comprehensive understanding of entities or events. In the field of HOI, several recent studies [65, 21, 34, 56, 58, 59] have applied image-and-text models to improve interaction detection performance. For example, HOI-VP [65] used a set of

binary classifiers to verify each category and proposed Language Prior-guided Channel Attention (LPCA) to enhance HOI recognition. SSRT [21] pre-selected object-action (OA) prediction candidates and encoded them as text features to refine the queries' representation. PhraseHOI [34] employed a pre-trained word embedding model to generate a phrase embedding that enhances the discriminative ability and capacity of the common knowledge space.

Although the use of vision-and-language pre-training (VLP) or language knowledge injection has motivated the exploration of HOI image-text correspondences through multi-modal learning, their effectiveness in knowledge transfer remains limited. This is due to the heterogeneity gap [3] that exists between different modalities, which requires cross-modal modeling to reduce the inter-modality gap and explore semantic correlations. Additionally, the problem of multi-interaction to multi-text matching in HOI tasks remains unsolved, which may limit the reliability of cross-modal correspondences. Therefore, a systematic and unified solution is needed to better exploit cross-modal HOI detection and improve the generalization ability of HOI detectors.

In this paper, we propose a systematic approach (RmLR) to improve HOI detection in light of the structured text knowledge in cross-modal learning. Concretely, our HOI framework proceeds from three perspectives: *i)* we reveal the problem of interaction information loss in the two-stage HOI detector, and propose the ***Re-mine*** strategy to obtain this crucial visual information; *ii)* more sophisticated cross-modal ***Learning*** method to achieve semantic association from sentence- and word-level; *iii)* ***Reasoning*** using textual knowledge-enhanced representations substantially improves the visual model's understanding of interactions. The main contribution of this paper is summarized as follows:

- We propose a systematic and unified framework so that the inherent challenges of HOI can be elaborated in both visual and cross-modal settings.
- We qualitatively and quantitatively analyze the problem of interaction information loss in two-stage visual HOI detector, and propose a re-mining strategy to capture these crucial interaction-aware representations.
- We formulate the cross-modal learning in HOI domain as a many-to-many matching problem, where multiple interactions need to be matched with their corresponding textual descriptions, and propose appropriate sentence and text alignment strategies to promote learning semantically aligned.
- Extensive experiments show that our RmLR equipped with ResNet-50 outperforms previous SOTA by a large margin and achieves an average mAP increase of about +3.88p and +5.05p on HICO-DET [6] and V-COCO [16], respectively.

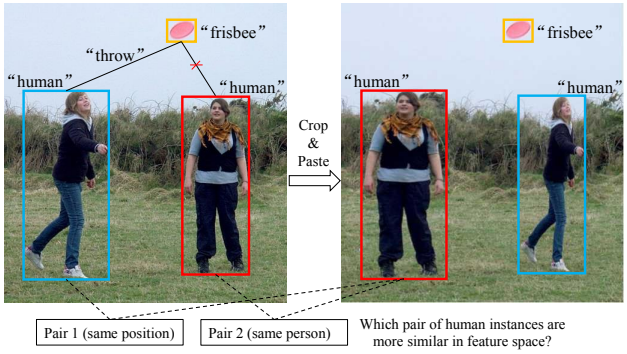

Figure 1. Which pair of human instances is more similar in the HOI detector? According to our analysis using cosine similarity measurement for the human tokens of the DETR-based HOI detector [62], Pair 1 has a similarity score of 0.99, while Pair 2 has a score of only 0.58. These findings are consistent with numerous similar cases observed in our experiments, highlighting the phenomenon of *interaction-related information loss* in which the output tokens of the object detector primarily emphasize spatial position, potentially leading to the loss of crucial information related to the interactions.

2. Related Work

2.1. Generic HOI Detection

According to the network architecture design, current HOI detection approaches can be broadly classified into two categories: two-stage methods [62, 6, 13, 14, 33, 20] and one-stage methods [15, 26, 66, 35, 7]. One-stage methods typically employ multitask learning to jointly perform instance detection and interactive relation modeling [35, 60, 36]. In contrast, two-stage methods first perform object detection, followed by interactive relation modeling for all HO pairs candidates. By leveraging the full potential of each module, two-stage methods have demonstrated improved detection performance [60]. Recent works have also leveraged the power of Transformer [5] in formulating HOI detection as set prediction, resulting in significant performance gains [7].

2.2. Language Semantics for Vision

Motivated by the remarkable success of Large Language Model (LLM) [11] pre-training in NLP, leveraging language semantics to enhance vision models has recently emerged as a promising approach for computer vision tasks [46, 53, 22, 1, 63, 24]. Among them, Vision-and-Language Pre-training (VLP) [42, 30] has become a popular paradigm in many vision-and-language tasks due to its applicability in learning generalizable multi-modal representations from large-scale image-text data [1, 9, 10]. These methods have been recently used in multi-modal retrieval [12], vision-and-language navigation [2], and other fields. Effective inter-modal semantic alignment, especially fine-grained semantic alignment, is a critical component for cross-modal learning [30]. Since different modalities have their own inherent prop-

erties, their semantic organization varies to some extent [8]. Thus, it is crucial to investigate how to efficiently correlate diverse semantic information.

2.3. HOI Vision-and-Language Modeling (HOI-VLM)

Although previous HOI detectors [60, 62, 38] have achieved moderate success, they often treat interactions as discrete labels and ignore the richer semantic text information in triplet labels. More recently, a few researchers [65, 21, 34, 56, 58, 59] have investigated the HOI Vision-and-Language Modeling to further boost the HOI detection performance. Among them, [65], [21], and [59] both tended to aggregate language prior features into the HOI recognition. RLIP [58] and [56] proposed to construct a transferable HOI detector via the VLP approach. As the applications and extensions of Vision-and-Language learning to the HOI domain, these HOI-VLM methods aim to understand the content and relations between visual interaction features and their corresponding triplet texts. However, the natural distribution inconsistency in the two modalities can directly lead to incompatibility of the modal features, as discussed by [8]. The issue of narrowing the heterogeneity gap and effectively ensuring the consistency and correlation of cross-modal features in HOI detection remains unresolved.

3. The Proposed RmLR Framework

3.1. Overview Architecture

We adopt the two-stage HOI detector approach for its superior performance, interpretability, and intuitive intermediate features. Inspired by the DETR family [5], we design the RmLR architecture (see Figure 2). Formally, our RmLR model is trained on an image-text corpus $\mathcal{X} = \left\{\left(\mathcal{I}^i, \mathcal{T}^i\right)\right\}_{i=1}^{|\mathcal{X}|}$, where $\mathcal{I}$ denotes the input image and $\mathcal{T}$ represents all the phrase descriptions (*e.g.* "Human ride bicycle") in $\mathcal{I}$. We can roughly divide RmLR into visual feature learning module $\Phi_{\theta_\mathcal{V}}$, interaction reasoning module $\Phi_{\theta_\mathcal{R}}$ and a pre-trained text encoder $\Phi_{\theta_\mathcal{T}}$, where θ indicates the weights in different modules. The overall training objective is defined as follows,

$$\min \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{X}}\left[\mathcal{L}(\mathcal{GT}, \Phi_{\theta_\mathcal{T}}(\mathcal{T}), \Phi_{\theta_\mathcal{V}} \circ \Phi_{\theta_\mathcal{R}}(\mathcal{I}, \mathcal{Q}_o))\right] \quad (1)$$

where $\mathcal{GT}$ and $\mathcal{L}$ are ground-truth label and overall loss function respectively, $\mathcal{Q}_o$ denotes the set of queries of objects, and $\circ$ is a network compound operator. Details of the module implementation are explained in the subsequent sections.

3.2. Re-mining Visual Features

Visual Entity Detection An input image $\mathcal{I} \in \mathcal{R}^{H\times W\times C}$ is first extracted as low-level visual features $\mathcal{X}^v \in \mathcal{R}^{h\times w\times c}$, and then the features are segmented into patch embeddings $\left\{x_1^v, x_2^v, \ldots, x_{N_v}^v\right\}$, where N_v is the number of patch embeddings. Then the patch embeddings $\left\{x_1^v, x_2^v, \ldots, x_{N_v}^v\right\}$ are flattened and linearly projected through a linear transformation $\mathcal{E}^v \in \mathcal{R}^{c\times D^v}$. Specifically, the input for Transformer-based entity detection are calculated via summing up the patch embeddings and position embeddings $\mathcal{E}_{pos}^v \in \mathcal{R}^{N_v\times D^v}$:

$$\mathcal{Z}^v = \left[x_1^v\mathcal{E}^v; x_2^v\mathcal{E}^v; \ldots; x_{N_v}^v\mathcal{E}^v\right] + \mathcal{E}_{pos}^v \quad (2)$$

Through self-attention, cross-attention, and feed-forward network (FFN) inference in entity detection decoder $\mathcal{F}_{ED}$, we obtain the entity token features $\mathcal{S}^v \in \mathcal{R}^{N\times D^v}$, box locations $\mathcal{B}^v \in \mathcal{R}^{N\times 4}$ and instance classes $\mathcal{C}^v \in \mathcal{R}^{N\times N_c}$:

$$(\mathcal{S}^v, \mathcal{B}^v, \mathcal{C}^v) = \mathcal{F}_{ED}\left(\mathcal{Z}^v, \mathcal{Q}_o\right) \quad (3)$$

where N denotes the number of detected instances, $\mathcal{Q}_o$ denotes the set of queries of objects, and N_c denotes the number of detectable categories. To obtain the pair-wise entity token features and box locations, we construct a set of pair-wise HO indexes $\{(h, o) \mid h \neq o, \mathcal{C}_h^v = \text{"human"}\}$. We form all pairs of detected instances and filter those where the subject is not human, as object–object pairs are beyond the scope of HOI detection. According to the filtered HO indexes, pair-wise entity token features $\tilde{\mathcal{S}}^v \in \mathcal{R}^{N^p\times 2D^v}$ and box locations $\tilde{\mathcal{B}}^v \in \mathcal{R}^{N^p\times 8}$ are able to obtain. This information is used for subsequent interactive relation learning and reasoning.

Interactive Relation Encoder Through a meticulous analysis of numerous cases, we discovered that *the current entity detection models prioritize the object's location information*. As a result, humans performing different actions at the same position are often mapped to similar representations, as illustrated in Figure 1. This phenomenon poses a significant risk to the HOI task, as it may result in the loss of crucial visual information. As two-stage HOI detectors operate independently for entity detection and interaction recognition, the entity token features $\mathcal{S}^v$ obtained from the entity detection model predominantly focus on spatial information and hence may fail to capture enough interaction-relevant cues.

To this end, we design a lightweight Interactive Relation Encoder (IRE) to remine interaction features intuitively and explicitly (see Figure 3). To capture the higher-level relation features from lower-level visual features, we apply a Transformer encoder to process feature map $\mathcal{X}^v$:

$$\mathcal{X}_e^v = \mathcal{F}_{enc}(\mathcal{X}^v) \quad (4)$$

Then, we perform masked RoI operation on the interactive information-rich tensors $\mathcal{X}_e^v$ to compute the direct reflection m^v according to the pair-wise box locations $\tilde{\mathcal{B}}^v$:

$$m^v = FC(GAP(mROI(\mathcal{X}_e^v, \tilde{\mathcal{B}}^v))) \in \mathcal{R}^{D^v} \quad (5)$$

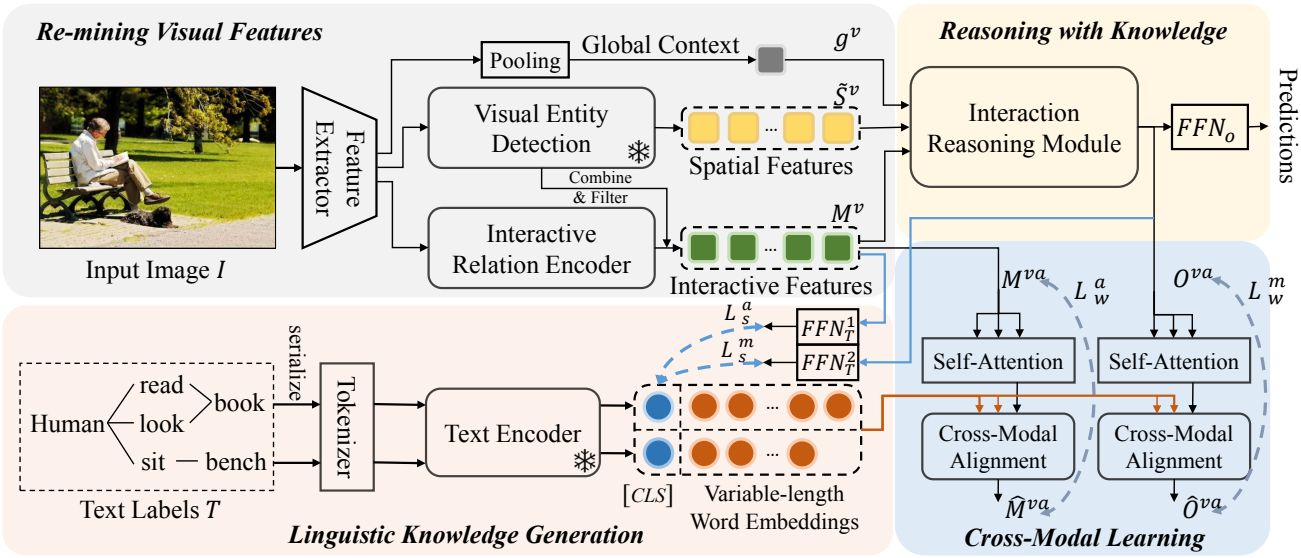

Figure 2. The overall architecture of our proposed RmLR approach, where the Visual Entity Detection module, Interactive Relation Encoder (with the "re-mining visual feature" process), Linguistic Knowledge Generation, Cross-Modal Learning (with the "learning cross-modal content" process), Interaction Reasoning Module (with the "reasoning using knowledge" process) are shown.

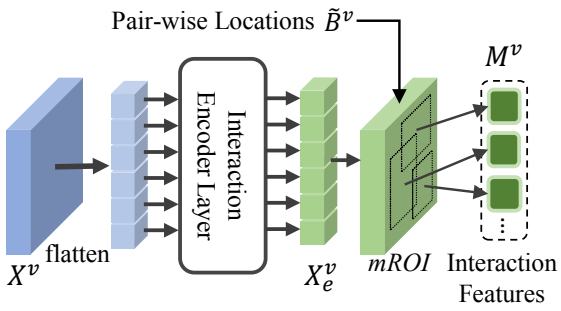

Figure 3. Re-mining the crucial interactive features via an interactive relation encoder.

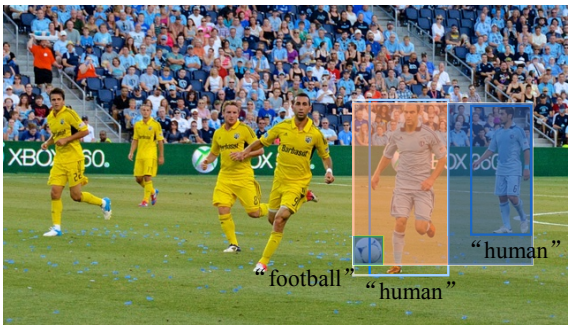

Figure 4. Feature interference problem in the naive union interaction region. According to the rules of the naive union interaction region, the orange and blue part together constitute the interaction area of the rightmost person to the football. It can be seen that the interactive human-object features (orange part) interfere with the non-interactive human-object features (blue part).

Here, we use a fully-connected layer (FC), global average pooling (GAP), and masked region of interest (ROI) operation ($mROI$) to obtain the interaction-aware features. To ensure that the features are only computed within the region of interest, we use a zero mask to cover the regions outside the HO candidate boxes to avoid the feature interference problem, as shown in Figure 4. After that, GAP operation followed by an FC layer are applied on the feature map $\mathcal{X}^v$ to obtain global scene information g^v:

$$g^v = FC(GAP(\mathcal{X}^v)) \in \mathcal{R}^{D^v} \tag{6}$$

So far, we have generated the human and object candidates, global context g^v, pair-wise token $\tilde{\mathcal{S}}^v = \{\tilde{s}_i^v\}_{i=1}^{|\tilde{\mathcal{S}}^v|}$, and interaction cues $\mathcal{M}^v = \{m_j^v\}_{j=1}^{|\mathcal{M}^v|}$, which contain rich visual features for HOI recognition. The detailed ablations for this structure can be founded in Section 4.4 and Table 2.

3.3. Linguistic Knowledge Generation

To integrate linguistic knowledge into the visual HOI framework, we first construct annotation text for every image in HOI datasets. Considering the arrangement of <$person, verb, object$> triplet is very similar to the <$subject, predicate, object$> in language, we directly serialize each triplet annotation $\mathcal{GT}_i$ as a sub-sentence t_i. Then, a special $[SEP]$ token is used to separate multiple sub-sentences. In this way, each input image $\mathcal{I}$ obtains a corresponding variable-length annotation text $\mathcal{T} = \{t_j\}_{j=1}^{|\mathcal{T}|}$, where $|\mathcal{T}|$ denotes the number of ground truth interactions for the input image $\mathcal{I}$.

We utilize a pre-trained language model, such as Mo-

bileBERT [51], to generate semantic representations at the sentence- and word-level. First, the input text $\mathcal{T}$ is tokenized into subword tokens $\{x_1^l, x_2^l, \ldots, x_{N_l}^l\}$ using the WordPiece algorithm [57]. These tokens are then represented as one-hot vectors $z_i^l \in \mathcal{R}^V$, where V is the vocabulary size, and N_l is the number of tokens. The tokens are then linearly transformed into embeddings using a matrix $\mathcal{E}^l \in \mathcal{R}^{V \times D^l}$. Additionally, a special start-of-sequence $[CLS]$ token embedding $z_{cls}^l \in \mathcal{R}^{D^l}$ is added to the beginning of the text. Finally, the input text representations are obtained by summing up the token embeddings and text position embeddings $\mathcal{E}_{pos}^l \in \mathcal{R}^{(N_l+1) \times D^l}$:

$$\mathcal{X}^l = \left[z_{cls}^l; z_1^l \mathcal{E}^l; \ldots; z_{N_l}^l \mathcal{E}^l; z_{end}^l\right] + \mathcal{E}_{pos}^l \tag{7}$$

Using the text encoder $\mathcal{F}_{TE}$, we calculate the $[CLS]$ tokens $\mathcal{E}_{cls}$ and word embeddings $\mathcal{E}^w$ as

$$(\mathcal{E}_{cls}, \mathcal{E}^w) = \mathcal{F}_{TE}\left(\mathcal{X}^l\right) \in \mathcal{R}^{(N_l+1) \times D^l} \tag{8}$$

In this way, the linguistic knowledge corresponding to $|\mathcal{T}|$ ground-truth interactions can be obtained, including sentence-level representation $\mathcal{E}_{cls} = \{e_{cls}^i\}_{i=1}^{|\mathcal{T}|}$ and word-level representation $\mathcal{E}^w = \{e_w^j\}_{j=1}^{|\mathcal{E}^w|}$. We provide a detailed comparison of different text encoders in Section 4.4 and Table 3.

3.4. Cross-Modal Learning

For visual representation, we first concatenate the global context g^v, pair-wise token $\tilde{s}^v = (s_h^v, s_o^v)$ and corresponding interaction cue m^v to generate unified and diverse visual description for HO candidate:

$$\mathcal{H}^v = FC(\text{cat}\left(g^v, \tilde{s}^v, m^v\right)) \in \mathcal{R}^{D^l} \tag{9}$$

Then, we introduce the competitive strategy in UPT [62] to construct a concise Transformer-based interaction reasoning module $\mathcal{F}_{IR}$. After the competitive operation in $\mathcal{F}_{IR}$, the visual features $\mathcal{H}^v$ are converted into $\mathcal{O}^v$. To achieve a more flexible and efficient correlation of variable-length text to the interaction set, we design a dual distillation scheme to guide the training process for Interactive Relation Encoder and Interaction Reasoning Module simultaneously. Among them, the operation for IRE is more focused on pair-wise token $\tilde{s}^v$, and the latter is more focused on $\mathcal{O}^v$. The attention operation in these two mechanisms is defined as follows:

$$\mathcal{ATTN}(q, k, v) = \text{softmax}\left(qk^\top / \sqrt{D_k}\right) \cdot v \tag{10}$$

where q, k, and v are the query, key, value matrices linearly transformed from the corresponding input sequences, respectively, and D_k is the dimension of k. We conduct L self-attention layers to interact representations within the two levels of features:

$$\mathcal{M}^{vs} = \mathcal{ATTN}\left(\mathcal{M}^v, \mathcal{M}^v, \mathcal{M}^v\right) \tag{11}$$

$$\mathcal{O}^{vs} = \mathcal{ATTN}\left(\mathcal{O}^v, \mathcal{O}^v, \mathcal{O}^v\right) \tag{12}$$

where $\mathcal{M}^{vs}$ and $\mathcal{O}^{vs}$ are the self-attention outputs for two representations, respectively. Then, the cross-modal attention are designed to align two modality representations and integrate linguistic information into visual representations in word-level:

$$\widehat{\mathcal{M}}^{va} = \mathcal{ATTN}\left(\mathcal{M}^{va}, \mathcal{E}^w, \mathcal{E}^w\right) \tag{13}$$

$$\widehat{\mathcal{O}}^{va} = \mathcal{ATTN}\left(\mathcal{O}^{va}, \mathcal{E}^w, \mathcal{E}^w\right) \tag{14}$$

where $\mathcal{M}^{va}$ and $\mathcal{O}^{va}$ are the visual representations corresponding to the textual embeddings $\mathcal{E}^w$, $\widehat{\mathcal{M}}^{va}$ and $\widehat{\mathcal{O}}^{va}$ are cross-attention outputs for two visual representations, respectively. In this way, no matter how complex multiple interaction are confronted, it is possible to align their visual features with the fine-grained textual representations. And the number of tokens of $\widehat{\mathcal{M}}^{va}$ and $\widehat{\mathcal{O}}^{va}$ are equal to the number of $\mathcal{M}^{va}$ and $\mathcal{O}^{va}$. In order to transfer linguistic knowledge to a visual model, we adopt the $L1$ distance metric to facilitate the learning between two types of representations:

$$\mathcal{D}_{L1}\left(a_{ho}, b_{ho}\right) = \frac{1}{N} \sum_{i}^{N} |a_{ho} - b_{ho}| \tag{15}$$

where a_{ho} and b_{ho} broadly refer to two types of representations in our RmLR architecture. It is convenient to use word-level semantically enhanced representations to guide the learning of visual models. The two key components are guided as follows:

$$\mathcal{L}_w^m = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{X}}\left[\mathcal{D}_{L1}\left(\mathcal{O}^{va}, \widehat{\mathcal{O}}^{va}\right)\right] \tag{16}$$

$$\mathcal{L}_w^a = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{X}}\left[\mathcal{D}_{L1}\left(\mathcal{M}^{va}, \widehat{\mathcal{M}}^{va}\right)\right] \tag{17}$$

where $\mathcal{L}_w^m$ and $\mathcal{L}_w^a$ denote the word-level cross-modal alignment loss for visual representation and logits, respectively. Even if multiple interactions occur between one HO pair, they can be described by variable-length word embedding sequences. These operations implement a fine-grained alignment and transfer mechanism for variable-length word embedding sequences to visual interaction set in HOI task.

In addition, we also perform sentence-level knowledge transfer for the RmLR. Although the sentence-level text representation is not as detailed as the word-level text representation, it also reflects the interaction information of HO pair to some extent. Thus, we regard sentence-level transfer as an auxiliary objective for our RmLR. Without the cross-modal attention, we directly perform knowledge transfer from $[CLS]$ tokens $\mathcal{E}_{cls}$ to the logits of RmLR:

$$\mathcal{L}_s^m = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{X}}\left[\mathcal{D}_{L1}\left(\mathcal{E}_{cls}, \mathcal{FFN}_T^2\left(\mathcal{O}^{va}\right)\right)\right] \tag{18}$$

Table 1. Experimental results on HICO-DET [6] and V-COCO [16].

Method (Year)	Backbone	HICO-DET						V-COCO	
		Default Setting			Known Objects Setting				
		Full	Rare	Non-rare	Full	Rare	Non-rare	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
One-stage Methods:									
InteractNet (2018) [15]	ResNet-50-FPN	9.94	7.16	10.77	-	-	-	40.0	-
PPDM (2020) [35]	Hourglass-104	21.94	13.97	24.32	24.81	17.09	27.12	-	-
HOTR (2021) [27]	ResNet-50	25.10	17.34	27.42	-	-	-	55.2	64.4
HOI-Trans (2021) [68]	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57	52.9	-
AS-Net (2021) [7]	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14	53.9	-
QPIC (2021) [52]	ResNet-101	29.90	23.92	31.69	32.38	26.06	34.27	58.8	61.0
SSRT (2022) [21]	ResNet-50	30.36	25.42	31.83	-	-	-	63.7	65.9
SSRT (2022) [21]	ResNet-101	31.34	24.31	33.32	-	-	-	65.0	67.1
CDN-S (2022) [60]	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42	61.68	63.77
CDN-B (2022) [60]	ResNet-50	31.78	27.55	33.05	34.53	29.73	35.96	62.29	64.42
CDN-L (2022) [60]	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38	63.91	65.89
DOQ (CDN-S) (2022) [45]	ResNet-50	33.28	29.19	34.50	-	-	-	-	-
Liu et al. (2022) [38]	ResNet-50	33.51	30.30	34.46	36.28	33.16	37.21	63.0	65.2
GEN-VLKT-s (2022) [36]	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99	62.41	64.46
GEN-VLKT-m (2022) [36]	ResNet-101	34.78	31.50	35.77	38.07	34.94	39.01	63.28	65.58
GEN-VLKT-l (2022) [36]	ResNet-101	34.95	31.18	36.08	38.22	34.36	39.37	63.58	65.93
Two-stage Methods:									
HO-RCNN (2018) [6]	CaffeNet	7.81	5.37	8.54	10.41	8.94	10.85	-	-
GPNN (2018) [44]	ResNet-101	13.11	9.34	14.23	-	-	-	44.0	-
TIN (2019) [33]	ResNet-50	17.03	13.42	18.11	19.17	15.51	20.26	47.8	54.2
VCL (2020) [18]	ResNet-50	23.63	17.21	25.55	25.98	19.12	28.03	48.3	-
ATL (2021) [19]	ResNet-50	23.81	17.43	27.42	27.38	22.09	28.96	-	-
VSGNet (2020) [54]	ResNet-152	19.80	16.05	20.91	-	-	-	51.8	57.0
DJ-RN (2020) [31]	ResNet-50	21.34	18.53	22.18	23.69	20.64	24.60	-	-
DRG (2020) [13]	ResNet-50-FPN	24.53	19.47	26.04	27.98	23.11	29.43	51.0	-
IDN (2020) [32]	ResNet-50	24.58	20.33	25.86	27.89	23.64	29.16	53.3	60.3
FCL (2021) [20]	ResNet-50	25.27	20.57	26.67	27.71	22.34	28.93	52.4	-
SCG (2021) [61]	ResNet-50-FPN	29.26	24.61	30.65	32.87	27.89	34.35	54.2	60.9
UPT (2022) [62]	ResNet-50	31.66	25.90	33.36	35.05	29.27	36.77	59.0	64.5
UPT (2022) [62]	ResNet-101	32.31	28.55	33.44	35.65	31.60	36.86	60.7	66.2
RmLR (Ours)	ResNet-50	36.93	29.03	39.29	38.29	31.41	40.34	63.78	69.81
RmLR (Ours)	ResNet-101	37.41	28.81	39.97	38.69	31.27	40.91	64.17	70.23

Similarly, we design an auxiliary objective on IRE, where the task is to guide the output representations of IRE:

$$\mathcal{L}_s^a = \mathbb{E}_{(\mathcal{I},\mathcal{T})\sim\mathcal{X}}\left[\mathcal{D}_{L1}\left(\mathcal{E}_{cls}, \mathcal{FFN}_T^1\left(\mathcal{M}^{va}\right)\right)\right] \quad (19)$$

We also provided detailed ablation experiments and analysis of this structure in Table 2 and Table 7 of Section 4.4.

3.5. Reasoning with Language-enhanced Representations

For the HOI recognition, a concise Transformer-based Interaction Reasoning Module (IRM) $\mathcal{F}_{IR}$ is designed to aggregate representation for each HO candidate. Our work differs from previous work in that these fed-in features are enhanced by textual knowledge, which is richer and more distinct than the unimodal features. After that, we add a classification head $\mathcal{FFN}_o$ to map logits to specific categories:

$$\mathcal{P} = \mathcal{FFN}_o(\mathcal{F}_{IR}(\mathcal{H}^v)) \quad (20)$$

Finally, a Focal loss is adopted as $\mathcal{L}_{\text{hoi}}$ to evaluate the image-level HOI predictions:

$$\mathcal{L}_{\text{hoi}} = Focal(sigmoid(\mathcal{P}), \mathcal{GT}) \quad (21)$$

where $\mathcal{GT}$ are the ground-truth labels corresponding to the predicted interaction set $\mathcal{P}$. Focal loss is defined via $Focal(p) = -(1-p)^\gamma \log(p)$, where γ is set as a hyperparameter. The overall loss is constructed as follows:

$$\mathcal{L} = \lambda_{hoi}\mathcal{L}_{hoi} + \lambda_s^m\mathcal{L}_s^m + \lambda_w^m\mathcal{L}_w^m + \lambda_s^a\mathcal{L}_s^a + \lambda_w^a\mathcal{L}_w^a \quad (22)$$

4. Experiments

4.1. Datasets and Evaluation Metrics

We conducted training and evaluation on the widely used V-COCO [16] and HICO-DET [6], following the established protocols in previous works [33, 62]. Due to the limited space, a detailed description of the datasets and evaluation metrics can be found in Supplementary Material.

4.2. Implementation Details

Following the two-stage HOI detector training paradigm [62], we first pre-train the DETR on a large-scale image dataset and then fine-tune it on the HICO-DET and V-COCO datasets. For HICO-DET, we initialize the network with DETR pre-trained on MS COCO [37]. We adopt the data

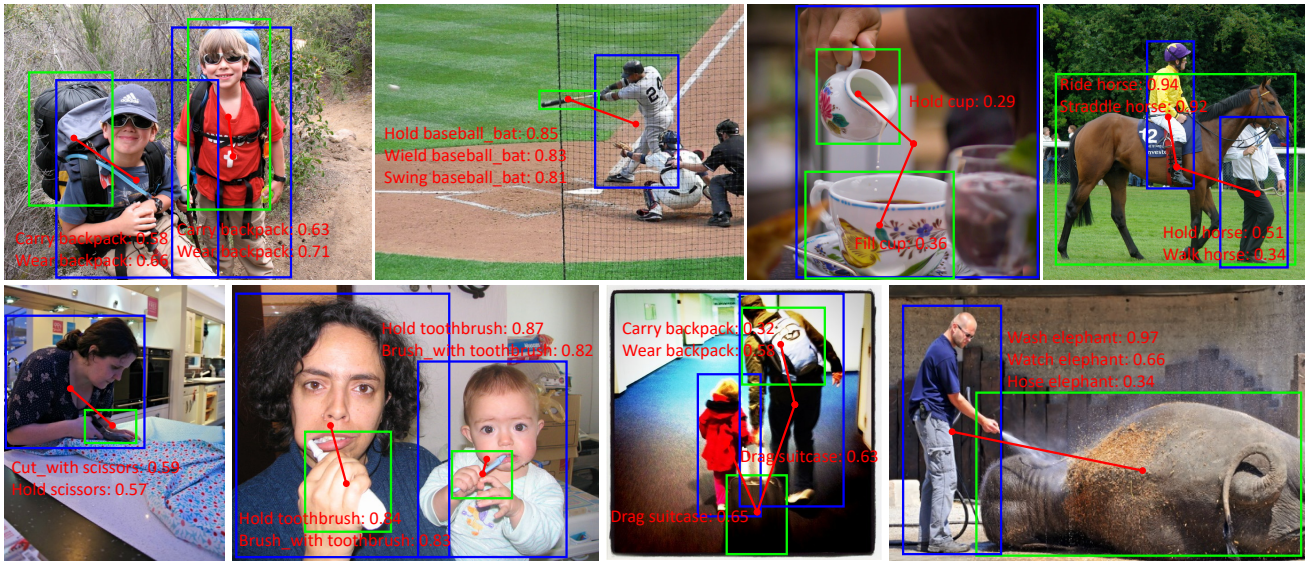

Figure 5. Some representative results of our RmLR method on HICO-DET [6] test set.

augmentation and preprocessing techniques as [62]. For cross-modal learning, the number of self-attention and cross-attention layers is set to 2 and 1, respectively. And the dimension of the hidden state in these two mechanisms is set to 1024. For the Focal loss, we set $\gamma = 0.2$ and $\beta = 0.5$ following [62]. We also provide a detailed description of implementation details in Supplementary Material.

4.3. Main Results

We conducted a comprehensive evaluation of our proposed method in comparison with state-of-the-art HOI methods, such as UPT [62], GEN-VLKT [36], and CDN [60], on the HICO-DET and V-COCO datasets. The results of this comparison are presented in Table 1. Our approach significantly outperforms all previous state-of-the-art methods, and this advantage is maintained across both ResNet-50 and ResNet-101 feature extractors. We also compared our proposed method with some previous methods, such as those relying on extra datasets such as Human Pose [39] and Vision-and-Language [58]), by training on larger and richer datasets, as shown in Tables 5 and 6. These results demonstrate the superiority of our RmLR method.

4.4. Ablation Studies

To illustrate the effectiveness of our proposed approach, we perform ablation studies on each component. Specifically, cross-modal learning contains sentence-level and word-level embedding knowledge distillation for IRE and IRM. The experiments are conducted on the V-COCO [16] dataset with ResNet50 [17] as the CNN backbone, and the results are reported in Table 2. We also provide an analysis of the computational cost of our method in Table 4. The results demonstrate that RmLR achieves a substantial performance improvement while adding only a minor computational cost.

Table 2. Ablations of different modules of our RmLR Framework on V-COCO [16]. "SA" and "WA" indicate sentence- and word-level alignment, respectively. "KT" indicates knowledge transfer.

Variants	IRE	IRE-KT		IRM-KT		V-COCO	
		SA	WA	SA	WA	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Plain model						58.51	63.87
w/o CL	✓					61.13	67.48
w/o Rm				✓	✓	62.89	68.91
w/o WA	✓	✓		✓		62.37	68.29
w/o SA	✓		✓		✓	63.33	69.41
w/o IRM-KT	✓	✓	✓			62.53	68.61
w/o IRE-KT	✓			✓	✓	63.42	69.49
RmLR	✓	✓	✓	✓	✓	63.78	69.81

Table 3. Experimental results of different Text Encoders. The ResNet-50 [17] backbone is adopted as the visual feature extractor.

Text Encoder	V-COCO	
	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
ALBERT-base-v2 [29]	63.45	69.64
RoBERTa [40]	63.49	69.62
MobileBERT [51]	63.78	69.81
BERT-base [11]	63.89	69.98
BERT-large [11]	63.93	70.05

Table 4. FLOPs and Params analysis for HOI detectors on V-COCO [16] dataset with 800×800 resolution.

Method	Backbone	MACs (G)	Parms (M)	FPS
DETR [5]	ResNet-50	57.02	36.59	29.1
	ResNet-101	104.37	55.53	21.3
UPT [62]	ResNet-50	57.11	36.86	27.5
	ResNet-101	104.46	55.80	20.2
RmLR (Ours)	ResNet-50	57.22	36.98	27.2
	ResNet-101	105.57	55.92	19.9

The impact of Interactive Relation Encoder. In "Plain

Table 5. Comparison results with the methods using extra datasets on HICO-DET [6]. For extra datasets, "P" indicates human pose and "L" indicates linguistic knowledge.

Method (Year)	Backbone	Extras	HICO-DET					
			Default Setting			Known Objects Setting		
			Full	Rare	Non-rare	Full	Rare	Non-rare
PMFNet (2019) [55]	ResNet-50	L	17.46	15.65	18.00	20.34	17.47	21.20
TIN (2019) [33]	ResNet-50	P	17.22	13.51	18.32	19.38	15.38	20.57
Peyre et al. (2019) [43]	ResNet-50	P	19.40	14.63	20.87	-	-	-
FCMNet (2020) [39]	ResNet-50	P+L	20.41	17.34	21.56	22.04	18.97	23.12
PD-Net (2021) [65]	ResNet-50-FPN	L	20.76	15.68	22.28	25.59	19.93	27.28
ACP (2020) [28]	ResNet-152	P+L	20.59	15.92	21.98	-	-	-
DRG (2020) [13]	ResNet-50-FPN	P	24.53	19.47	26.04	27.98	23.11	29.43
RLIP-ParSeD (2022) [58]	ResNet-50	L	30.70	24.67	32.50	-	-	-
RLIP-ParSe (2022) [58]	ResNet-50	L	32.84	26.85	34.63	-	-	-
PhraseHOI (2022) [34]	ResNet-50	L	29.29	22.03	31.46	31.97	23.99	34.36
PhraseHOI (2022) [34]	ResNet-101	L	30.03	23.48	31.99	33.74	27.35	35.64
OCN (2022) [59]	ResNet-50	L	30.91	25.56	32.51	-	-	-
OCN (2022) [59]	ResNet-101	L	31.43	25.80	33.11	-	-	-
RmLR (Ours)	ResNet-50	L	36.93	29.03	39.29	38.29	31.41	40.34
RmLR (Ours)	ResNet-101	L	37.41	28.81	39.97	38.69	31.27	40.91

Table 6. Comparison results with the methods using extra datasets on V-COCO [16].

Method	Backbone	Extras	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
TIN [33]	ResNet-50	P	48.7	-
DRG [13]	ResNet-50-FPN	L	51.0	-
FCMNet [39]	ResNet-50	P	53.1	-
ConsNet [41]	ResNet-50-FPN	P	53.2	-
RLIP-ParSeD [58]	ResNet-50	L	61.7	63.8
RLIP-ParSe [58]	ResNet-50	L	61.9	64.2
RmLR (Ours)	ResNet-50	L	63.78	69.81
RmLR (Ours)	ResNet-101	L	64.17	70.23

model", we follow the typical two-stage HOI detector [62] to construct a plain model, which directly adopts the entity token features as visual representations and feds them to HOI classifier. For "w/o CL", we add IRE for the plain model, but not cross-modal learning. In "w/o Rm", we remove the re-mining operation (*i.e.*, IRE) in RmLR to analyze the effect of IRE for the RmLR framework. Since the lack of IRE, we only perform knowledge transfer for IRM in this variant. As shown in Table 2, the introduction of IRE greatly improves the plain model by around 3.1 mAP. And the IRE also shows improvement on RmLR frameworks that are equipped with cross-modal learning.

Effect of sentence- and word-level alignment. For "w/o WA" and "w/o SA", we remove the word- and sentence-level alignment in the cross-modal learning process. In these two variants, IRE and other settings remained the same. Compared to the complete RmLR, these two variants drop in mAP by 1.5 and 0.4 points, respectively. Adding the word- and sentence-level alignment to "w/o CL" variant jointly improves by around 2.5 mAP. Furthermore, the experimental results show that the word-level alignment strategy has a stronger facilitation to cross-modal HOI learning than sentence-level alignment. The possible cause for this phenomenon is that HOI is essentially a variable-size interaction set prediction problem, and a more flexible alignment strategy is beneficial for linguistic knowledge transfer.

The impact of transfer position. In addition, we also verify the necessity of knowledge transfer for IRE and IRM. For "w/o IRM-KT" and "w/o IRE-KT", we remove the linguistic knowledge transfer for IRM and IRE, respectively. The experimental results show that the performance of these two variants decreased by about 1.2 and 0.3 mAP compared to RmLR. These findings suggest that knowledge transfer for IRM in this architecture is a more efficient approach. Moreover, the results also suggest that distillation for IRE can further improve the performance. Therefore, we chose to perform knowledge transfer for both modules simultaneously, with knowledge distillation for IRM as the primary and IRE as the secondary.

Effect of different Text Encoder. We build RmLR variants equipped with other text encoders and conduct comparison experiments on the V-COCO dataset to explore the effect of different text encoders. In Table 3, we show the results of different text encoders. These results indicate that different text encoders impact HOI recognition capability; generally, larger models may perform better. In addition, all these text models promote our RmLR framework to obtain state-of-the-art results on the V-COCO dataset.

The impact of hyperparameters for loss terms. We also present the results for detailed weight settings for loss function to Table 7. The subscript s and w indicates sentence- and word-level alignment loss, respectively. These results demonstrate that our model performance is not very sensitive to the weights of different loss terms.

4.5. Visualization

As depicted in Figure 5, one image may contain multiple individuals and objects, which may or may not interact with each other or engage in several interactions. Hence, we finely aligned and transferred knowledge between visual features

Table 7. Experiments on the V-COCO [16] set *w.r.t* different loss terms. s and w indicates sentence- and word-level alignment loss.

λ_{hoi}	λ_s^m	λ_w^m	λ_s^a	λ_w^a	$AP_{role}^{\#1}$	$AP_{role}^{\#2}$
1.0	1.0	1.0	1.0	1.0	62.98	69.11
1.0	1.0	0.5	1.0	0.5	62.73	68.95
1.0	0.5	1.0	0.5	1.0	63.35	69.52
2.0	0.5	0.5	0.1	0.1	63.05	69.14
2.0	1.0	1.0	0.08	0.08	63.59	69.55
2.0	2.0	2.0	0.1	0.1	63.57	69.62
2.0	1.0	1.0	0.3	0.3	63.55	69.69
2.0	2.0	1.0	0.5	0.1	63.43	69.39
2.0	1.0	2.0	0.1	0.3	63.69	69.77
2.0	1.0	1.0	0.1	0.1	63.78	69.81

and annotation texts and have effectively guided the complex HOI learning process via linguistic prior knowledge. The detection results substantiate the validity of cross-modal alignment and the efficacy of our RmLR approach.

5. Conclusion

In this paper, we introduce a systematic and unified framework called RmLR, which leverages structured text knowledge to enhance HOI detector. To address the issue of interaction information loss in the two-stage HOI detector, we propose a re-mining strategy to generate more comprehensive visual representations. We then develop fine-grained sentence- and word-level alignment and knowledge transfer methods to effectively address the many-to-many matching problem between multiple interactions and multiple texts in HOI-VLM. These strategies alleviate the matching confusion problem caused by simultaneous occurrences of multiple interactions, thus improving the effectiveness of the cross-modal learning process in HOI detection filed. Experimental results on the public datasets demonstrate the effectiveness of our approach, which achieves state-of-the-art performance. We hope the proposed RmLR may serve as an architecture guideline for future research in this area.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China under grant 62271143, and the Big Data Center of Southeast University.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683, 2018. 2

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1, 2

[4] Yichao Cao, Xiu Su, Qingfei Tang, Shan You, Xiaobo Lu, and Chang Xu. Searching for better spatio-temporal alignment in few-shot action recognition. *Advances in Neural Information Processing Systems*, 35:21429–21441, 2022. 1

[5] 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. 2, 3, 7

[6] 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. 1, 2, 6, 7

[7] 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

[8] Wei Chen, Weiping Wang, Li Liu, and Michael S Lew. New ideas and trends in deep multimodal content understanding: A review. *Neurocomputing*, 426:195–215, 2021. 3

[9] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022. 2

[10] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Vista: Vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5184–5193, June 2022. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 7

[12] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021. 2

[13] 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, 2, 6, 8

[14] 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, 2

[15] 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, 2, 6

[16] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 2, 6, 7

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[18] 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. 6

[19] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 6

[20] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 1, 2, 6

[21] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5353–5363, 2022. 1, 2, 3, 6

[22] Yongcheng Jing, Yining Mao, Yiding Yang, Yibing Zhan, Mingli Song, Xinchao Wang, and Dacheng Tao. Learning graph neural networks for image style transfer. In *ECCV*, 2022. 2

[23] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Amalgamating knowledge from heterogeneous graph neural networks. In *CVPR*, 2021. 1

[24] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Meta-aggregator: Learning to aggregate for 1-bit graph neural networks. In *ICCV*, 2021. 2

[25] Yongcheng Jing, Chongbin Yuan, Li Ju, Yiding Yang, Xinchao Wang, and Dacheng Tao. Deep graph reprogramming. In *CVPR*, 2023. 1

[26] 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. 1, 2

[27] 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. 6

[28] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020. 8

[29] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 7

[30] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1, 2

[31] 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. 6

[32] 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. 6

[33] 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. 1, 2, 6, 8

[34] Zhimin Li, Cheng Zou, Yu Zhao, Boxun Li, and Sheng Zhong. Improving human-object interaction detection via phrase learning and label composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1509–1517, 2022. 1, 2, 3, 8

[35] 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, 2, 6

[36] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 2, 6, 7

[37] 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. 6

[38] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20113–20122, 2022. 3, 6

[39] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020. 7, 8

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 7

[41] 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. 8

[42] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 2

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

[44] 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. 6

[45] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19558–19567, 2022. 6

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[47] Xiu Su, Tao Huang, Yanxi Li, Shan You, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Prioritized architecture sampling with monto-carlo tree search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10968–10977, 2021. 1

[48] Xiu Su, Shan You, Tao Huang, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Locally free weight sharing for network width search. *arXiv preprint arXiv:2102.05258*, 2021. 1

[49] Xiu Su, Shan You, Fei Wang, Chen Qian, Changshui Zhang, and Chang Xu. Bcnet: Searching for network width with bilaterally coupled network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2175–2184, 2021. 1

[50] Xiu Su, Shan You, Jiyang Xie, Mingkai Zheng, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Vitas: Vision transformer architecture search. In *European Conference on Computer Vision*, pages 139–157. Springer, 2022. 1

[51] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. *arXiv preprint arXiv:2004.02984*, 2020. 5, 7

[52] 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. 6

[53] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2

[54] 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. 6

[55] 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/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 8

[56] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022. 1, 3

[57] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 5

[58] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. Rlip: Relational language-image pre-training for human-object interaction detection. *arXiv preprint arXiv:2209.01814*, 2022. 1, 3, 7, 8

[59] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. Detecting human-object interactions with object-guided cross-modal calibrated semantics. *arXiv preprint arXiv:2202.00259*, 2022. 1, 3, 8

[60] 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. 2, 3, 6, 7

[61] 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. 6

[62] 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, 3, 5, 6, 7, 8

[63] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 2

[64] Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970*, 2023. 1

[65] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human–object interaction detection. *International Journal of Computer Vision*, 129(6):1910–1929, 2021. 1, 3, 8

[66] 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. 1, 2

[67] Tianfei Zhou, Siyuan Qi, Wenguan Wang, Jianbing Shen, and Song-Chun Zhu. Cascaded parsing of human-object interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2827–2840, 2021. 1

[68] 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. 6