Exploiting CLIP for Zero-shot HOI Detection Requires Knowledge Distillation at Multiple Levels

Bo Wan Tinne Tuytelaars
ESAT, KU Leuven
{bwan,tinne.tuytelaars}@esat.kuleuven.be

Abstract

In this paper, we investigate the task of zero-shot human-object interaction (HOI) detection, a novel paradigm for identifying HOIs without the need for task-specific annotations. To address this challenging task, we employ CLIP, a large-scale pre-trained vision-language model (VLM), for knowledge distillation on multiple levels. Specifically, we design a multi-branch neural network that leverages CLIP for learning HOI representations at various levels, including global images, local union regions encompassing human-object pairs, and individual instances of humans or objects. To train our model, CLIP is utilized to generate HOI scores for both global images and local union regions that serve as supervision signals. The extensive experiments demonstrate the effectiveness of our novel multi-level CLIP knowledge integration strategy. Notably, the model achieves strong performance, which is even comparable with some fully-supervised and weakly-supervised methods on the public HICO-DET benchmark. Code is available at https://github.com/bobwan1995/Zeroshot-HOI-with-CLIP.

1. Introduction

HOI detection aims to identify triplets of ⟨*human, object, interaction*⟩ within the context of a given image, which requires localization of human and object regions and recognition of their interactive behavior, e.g., play-basketball. It enables the intelligent system to understand and interpret human behavior in real-world scenarios, thus playing an instrumental role in anomalous behavior detection [34, 40], motion tracking [39, 52] and visual scene understanding [27, 41].

HOI detection has typically been investigated in a fully-supervised learning paradigm [5, 10, 13, 54, 63], where aligned HOI annotations (i.e. human-object locations and interaction types) are provided during the training stage, as illustrated in Fig. 1(a). Despite the high performance due to such comprehensive annotations, it suffers from the labor-intensive process of labeling HOI instances. This limitation has led recent studies to transition towards a weakly-supervised setup [1, 22, 25, 53, 64], where only image-level

(a) play-basketball (b) play-basketball (c) `Null`

Figure 1. **Comparison on training annotations for different setups:** (a) Fully-supervised; (b) Weakly-supervised; (c) Zero-shot.

HOI categories (without corresponding bounding boxes) are provided for model learning, as demonstrated in Fig. 1(b). While this setup significantly reduces the labeling costs and enhances robustness to label errors, it nonetheless necessitates image-level annotations on HOI datasets. These are still costly to obtain, often noisy and incomplete. Moreover, annotators may inadvertently introduce bias. Drawing inspiration from the great success of zero-shot learning [21, 26, 43, 60, 65, 66], we introduce the new challenging problem of zero-shot HOI detection, where *NO* HOI annotations are required for model learning, as shown in Fig.1(c). Importantly, note how our setup diverges from previous zero-shot HOI detection frameworks [2, 14, 36, 38, 45], which predominantly focus on knowledge transfer from observed HOI concepts to unseen categories (c.f. Sec. 2 for more details). In contrast, this work takes it a step further by tackling the extreme scenario where none of the HOI categories have been annotated during training, although we assume the category space is known.

Large-scale pre-trained VLM, such as CLIP [43], have shown substantial promise in various domains of zero-shot learning, including image classification [7, 21, 28, 37, 46, 57, 60], object detection [26, 66], and instance segmentation [65]. However, extending these models to zero-shot HOI detection poses a unique challenge, primarily due to the high-level relational understanding required in this context. To the best of our knowledge, this paper makes the first effort to propose

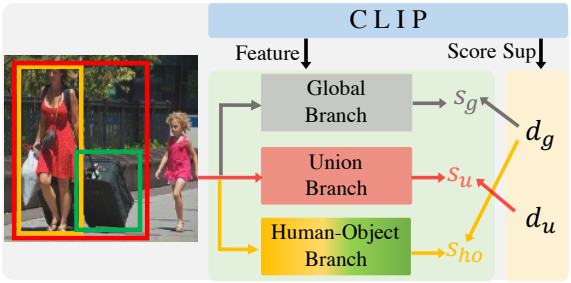

Figure 2. **Overview of our multi-level knowledge distillation.**

and tackle this challenging task. A naive solution might be directly employing CLIP on union regions (the joint areas of human-object proposals detected by an external object detector) to produce corresponding HOI scores. However, this approach proves to be suboptimal in terms of both inference speed and performance, as illustrated in Sec. 4.4. Alternatively, we leverage the power of CLIP in two ways: i) In terms of model design, we build a multi-branch network that extracts CLIP-oriented features on multiple levels, thus incorporating its robust generalization capabilities into HOI representations. ii) For model learning, we use the CLIP scores generated from global images and union regions as supervisory signals to train our multi-branch network.

Prior works [36, 54, 56] have revealed that HOI relations manifest across various levels, including global images, relational union regions, and individual human-object instances. Inspired by this insight, our work employs a multi-branch neural network to leverage the capabilities of CLIP for multi-level HOI representation learning. As sketched in Fig. 2, the network architecture comprises a global branch, a union branch, and a human-object branch. The process begins with computing an HOI embedding by using the CLIP text encoder to encode HOI label prompts. This embedding is shared across all branches. Concurrently, we detect human-object proposals with an off-the-shelf object detector and generate an image feature map with the CLIP visual encoder. Each branch within our network adopts a similar design, where the HOI features of global image, union regions, and human-object pairs are extracted from the image feature map, and then serve to compute scores corresponding to the HOI embedding. Finally, we bring these branches together using a late fusion strategy: rather than focusing on merging semantic features, our approach concentrates on fusing the multi-level HOI scores, offering a comprehensive view of HOIs on different scales.

To generate supervision for model training, we leverage CLIP to produce meaningful HOI scores for both global images and local union regions. For global images, the entire image is re-scaled and fed into the CLIP model to derive a global HOI score that captures the entire context of the image. Given that CLIP is pre-trained to understand a wide range of image-text pairs, it can provide a holistic understanding of visual relationships. Similarly, for union regions, we extract regions of interest that include both the human and the object. These regions, capturing more focused and localized interactions, are separately fed into CLIP to generate local and region-specific HOI scores. The local HOI scores can be noisy due to the coexistence of multiple HOIs in the same region or the presence of HOI irrelevant distractions, but they complement the global supervision signal. We conduct ablative studies on the supervision strategy in Tab. 4, which demonstrate the application of global supervision on the global & human-object branch, and local supervision on the union branch, yields the most effective results.

By incorporating CLIP knowledge on multiple levels for both model design and learning, our approach substantially enhances the zero-shot detection capability in a variety of HOI scenarios. In summary, our main contributions are three-fold:

- We pioneer the challenging task of zero-shot HOI detection, a new learning setup where no HOI annotations are used during training. This is a significant leap forward in the field of HOI detection.
- We propose a multi-level knowledge distillation strategy from CLIP for this task, where we seamlessly incorporate CLIP into the model design for detecting HOIs on different scales, and capture both global and local contexts to provide rich supervision for model training.
- Extensive experiments are conducted to verify the effectiveness of our CLIP integration strategies. Impressively, our method achieves a strong performance even on par with some fully-supervised and weakly-supervised methods on HICO-DET benchmarks.

2. Related Works

HOI detection ***Fully-supervised HOI detection*** has been the most common setup due to its superior performance. Research in this area generally falls into two categories: two-stage and one-stage frameworks. Two-stage methods [10, 11, 15, 29, 31, 49, 54, 63, 67, 68] adopt a *hypothesize-and-classify* strategy, which first generates a set of human-object proposals with the off-the-shelf object detector, and then enumerates all possible HOI pairs to classify their interactions. One-stage methods predict human & object locations and their interaction types simultaneously in an end-to-end manner, which are currently dominated by transformer-based architectures [4, 8, 24, 61, 62].

To decrease the reliance on HOI annotations, ***weakly-supervised HOI detection*** is proposed to learn HOIs with only image-level annotations. Due to the lack of location annotations, current works in this domain adopts the two-stage framework, and they focus on recognizing HOIs by developing advanced network structures to encode context [1,

25] and integrating external knowledge for representation learning [50, 53]. In this work, we propose a novel zero-shot setup without the need for any manual annotations in HOI detection.

Zero-shot HOI detection As most HOI classes are distributed in a long-tail manner [14, 45, 53] due to the inherent compositionality of HOIs [38], previous works on zero-shot HOI detection [3, 14, 18–20, 32, 36, 38, 42, 45, 56] aim to distill knowledge from observed HOI concept to unseen classes. They can be categorized into three scenarios: *unseen object*, *unseen action*, and *unseen combination*. There are mainly two streams of research for solving this problem. One stream [3, 14, 18, 20, 45] focuses on factorizing the human and object features by performing disentangled reasoning on verbs and objects, which allows the composition of novel HOI triplets for training and inference. Another stream [32, 36, 38, 56] transfers knowledge from knowledge graphs or pre-trained VLM to recognize unseen HOI concepts. Despite the substantial success of these approaches in knowledge transfer, they still rely heavily on the base knowledge provided by the seen HOI categories. In contrast, our setup does not require any HOI annotations for learning.

3. Method

3.1. Problem Setup

Formally, zero-shot HOI detection aims to learn an HOI detector that takes an image I as input and generates a collection of tuples $\mathcal{O} = \{(\mathbf{b}_h, \mathbf{b}_o, r_{h,o}, s^r_{h,o})\}$. Each tuple corresponds to a HOI instance, where $\mathbf{b}_h, \mathbf{b}_o \in \mathbb{R}^4$ indicate human and object bounding boxes, $r_{h,o} \in \{1, ..., N\}$ represents the interaction type between $\mathbf{b}_h$ and $\mathbf{b}_o$, and $s^r_{h,o} \in \mathbb{R}$ is the confidence score of the detected interaction.

3.2. Method Overview

To address the challenging zero-shot HOI detection task, we leverage CLIP for multi-level knowledge integration. To this end, we exploit the visual and textual encoders of CLIP to construct a multi-branch network for HOI representation learning, and use CLIP to generate global and local supervision for model training.

Model Design Due to the lack of HOI location annotations, we adopt a typical *two-stage* formulation [1, 25, 64] for HOI detection: in the first stage, we generate a group of human proposals $\{(\mathbf{b}_h, s_h)\}$ and object proposals $\{(\mathbf{b}_o, c_o, s_o)\}$ with an off-the-shelf object detector [44], where $s_h, s_o \in \mathbb{R}$ are detection scores and $c_o \in \{1, ..., C\}$ is the object class. In the second stage, we pair up all human and object proposals and predict the interaction class for each combination.

In order to infuse the generalization capability of CLIP into the HOI representation, we design a multi-branch deep network by incorporating CLIP's visual and textual encoders, as sketched in Fig. 3. Specifically, the global branch performs image-level HOI recognition, utilizing the HOI embedding produced by CLIP textual encoder as a classifier. In parallel, for each detected human-object pair $(\mathbf{b}_h, \mathbf{b}_o)$, a union branch extracts the contextual cues in their shared region of interest, providing a comprehensive view of the surrounding environment and potential interactions. On top of that, a human-object branch focuses on fine-grained HOI features and encodes the specific relational attributes of the interactive pairs, which are used to predict their interaction types. All the branches are integrated with a late fusion strategy, where the HOI scores from different levels are combined to obtain the final predictions.

Model Learning To train our model, we first employ CLIP on global image and local union regions to compute the corresponding HOI scores as supervision. Then we apply global image supervision on the global branch and human-object branch, and local union supervision on the union branch. Our training procedure can be viewed as a multi-level knowledge distillation approach from the pre-trained CLIP model. The primary objective of this strategy is to ensure that the HOI scores derived from distinct branches align with the CLIP scores.

3.3. Model Design

3.3.1 CLIP Backbone

CLIP builds a powerful vision-language model by pretraining on large-scale image-text pairs. It consists of a visual encoder $\mathcal{F}_V$ (e.g., a ResNet [17] or Vision Transformer [9]), a self-attention module $\mathcal{F}_{ATT}$ and a textual encoder $\mathcal{F}_T$ (e.g., a Transformer [51]), to map the visual and textual inputs to a shared latent space.

Specifically, for an input image I, the visual encoder $\mathcal{F}_V$ produces a feature map $\mathbf{\Gamma} \in \mathbb{R}^{H \cdot W \cdot D}$, where H, W, D denote the height, width, and depth of $\mathbf{\Gamma}$, respectively. Then a self-attention module $\mathcal{F}_{ATT}$ is adopted to encode $\mathbf{\Gamma}$ to a feature vector $v \in \mathbb{R}^D$: it takes a linear projection of the average pooling on the spatial dimensions of $\mathbf{\Gamma}$ as query $Q \in \mathbb{R}^D$, and a linear projection of the reshaped feature columns as key and value $K, V \in \mathbb{R}^{(HW) \cdot D}$:

$$\begin{aligned} Q = \mathcal{F}_Q(AvgPool(\mathbf{\Gamma})); \quad K, V = \mathcal{F}_K(\mathbf{\Gamma}), \mathcal{F}_V(\mathbf{\Gamma}) \\ v = \mathcal{F}_{MHA}(Q, K, V) \end{aligned} \tag{1}$$

where $\mathcal{F}_{Q,K,V}$ are linear projection layers and $\mathcal{F}_{MHA}$ is a standard multi-head attention module [51], all incorporated in $\mathcal{F}_{ATT}$.

To leverage CLIP for HOI detection, we utilize CLIP textual encoder $\mathcal{F}_T$ to generate HOI embedding $\mathcal{W}_T \in \mathbb{R}^{N \cdot D}$. In a prompting strategy akin to CLIP, we adopt a common template 'a person is $\{verb\}$-ing $\{object\}$' to convert HOI labels into text prompts. For instance, 'play basketball' would be converted to 'a person is playing basketball'. These sentences are then processed by $\mathcal{F}_T$ to create the HOI embedding $\mathcal{W}_T$, which is used to classify different levels of visual features into corresponding HOI scores.

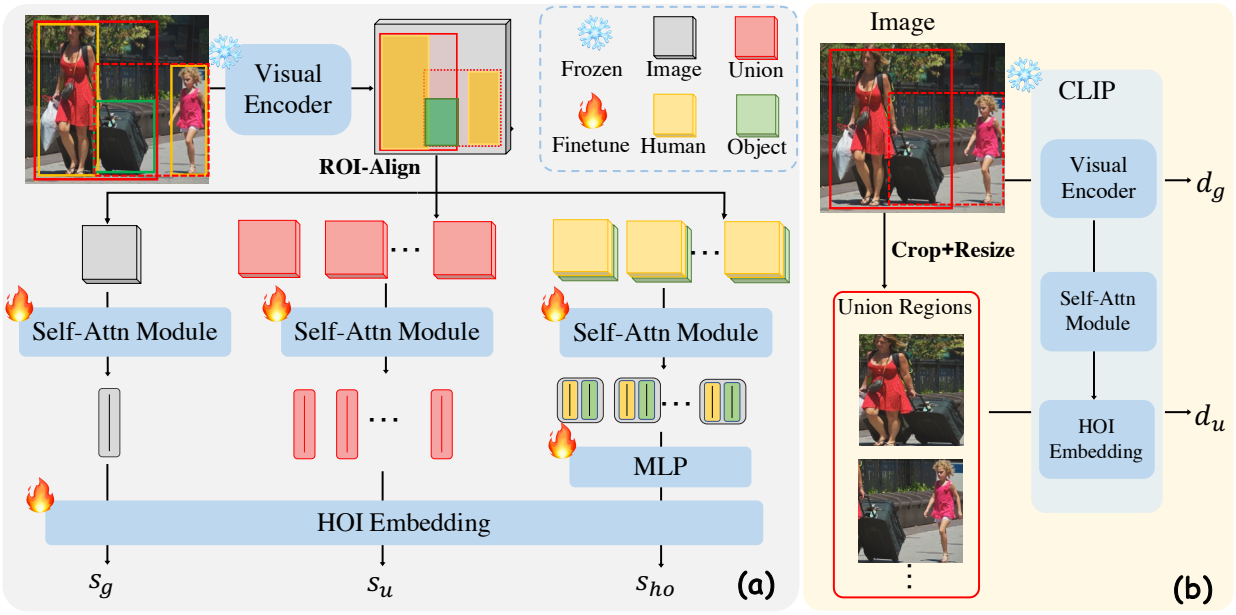

Figure 3. **A detailed introduction of our method:** We design a multi-branch neural network that incorporates CLIP components to extract multi-level information for HOI detection, which is supervised by the CLIP scores derived from global images and local union regions.

Modified Visual Encoder By default, a CLIP model with a ResNet visual encoder downscales the feature map by a factor of 32. Consequently, the spatial dimension of the feature map is insufficient for detailed region-specific feature extraction. To adapt this visual encoder for HOI detection with a larger feature map, we tweak the original ResNet structure. This is done by discarding the last average pooling module and adding an upsampling layer, thereby reducing the feature map size to only 8 times smaller than the image resolution. To maintain compatibility with the CLIP's self-attention module for feature aggregation, we implement an ROI-Align module [16] to resize the cropped feature map to a 7×7 grid for global images, union regions, and human-object proposals.

3.3.2 Global Branch

The global branch adapts the visual encoder and HOI embedding to perform an image-wise HOI recognition task, thereby capitalizing on the knowledge embedded within the CLIP model. Firstly, we adopt an ROI-Align on the image feature map followed by the self-attention module to compute a global feature vector $v_g \in \mathbb{R}^D$. Then the global HOI scores $s_g \in \mathbb{R}^N$ are predicted by conducting the inner product between the global vector v_g and the HOI embedding $\mathcal{W}_T$: $s_g = \text{Softmax}(\mathcal{W}_T \times v_g)$, where $\times$ is matrix multiplication.

3.3.3 Union Branch

The union region encapsulates the contextual relationship between humans and objects, which is crucial in comprehending their context. To exploit these context cues, we compute the union region $\mathbf{b}_u \in \mathbb{R}^4$ for each human proposal $\mathbf{b}_h$ and object proposal $\mathbf{b}_o$, and extract the corresponding appearance feature $v_u \in \mathbb{R}^D$ via RoI-align over the feature map $\mathbf{\Gamma}$ and self-attention aggregation. Similar union scores $s_u \in \mathbb{R}^N$ are computed with HOI embedding: $s_u = \text{Softmax}(\mathcal{W}_T \times v_u)$.

3.3.4 Human-object Branch

The human-object branch performs a fine-level classification for interaction pairs. For each human proposal $\mathbf{b}_h$ and object proposal $\mathbf{b}_o$, we crop the feature maps from $\mathbf{\Gamma}$ using RoI-Align, followed by a self-attention operation to generate their appearance features $v_h, v_o \in \mathbb{R}^D$. We also compute a spatial feature v_{sp} by encoding the relative positions of their bounding boxes $(\mathbf{b}_h, \mathbf{b}_o)$[1]. The holistic HOI representation $v_{ho} \in \mathbb{R}^D$ is an embedding of the human and object appearance features and their spatial feature: $v_{ho} = \mathcal{F}_{ho}([v_h; v_o; v_{sp}])$, where $[;]$ is the concatenation operation and $\mathcal{F}_{ho}$ is a multi-layer perceptron (MLP). Finally, we use the shared HOI embedding to predict the pairwise interaction scores $s_{ho} \in \mathbb{R}^N$ for each human-object combination: $s_{ho} = \mathcal{W}_T \times v_{ho}$

3.4. Model Learning with CLIP Supervision

In this section, we first utilize CLIP to generate two types of HOI supervision that are based on global images and local union regions, and then design a multi-task loss on various levels to train our multi-branch network. The overall

[1]For details c.f. Appendix

loss function $\mathcal{L}$ consists of three terms: i) an image-wise recognition loss $\mathcal{L}_g$ to detect global HOIs; ii) a union loss $\mathcal{L}_u$ for identifying contextual regional HOIs; and iii) a pairwise interaction classification loss $\mathcal{L}_{ho}$ to guide the learning of instance-specific HOIs. Formally, the overall loss is written as: $\mathcal{L} = \mathcal{L}_g + \mathcal{L}_u + \mathcal{L}_{ho}$.

CLIP Supervision Generation To obtain supervision for model learning, we directly employ CLIP on the image and union regions to generate the corresponding scores on the training set. As shown in Fig. 3(b), we crop the image to a square region at its center, with the side length equal to its shortest edge. This region is subsequently resized to 224×224 pixels and fed into the pre-trained CLIP model to generate the global supervision $d_g \in \mathbb{R}^N$. Similarly, for each union box $\boldsymbol{b}_u$, we crop the corresponding region in the raw image and resize it to 224×224 pixels, and then apply CLIP to generate local union supervision $d_u \in \mathbb{R}^N$.

Image-wise loss $\mathcal{L}_g$: Given the global image HOI scores s_g and CLIP supervision d_g, $\mathcal{L}_g$ is a standard Kullback-Leibler (KL) divergence defined as: $\mathcal{L}_g = \mathcal{D}_{KL}(s_g||d_g)$. Notably, d_g and s_g are independent predictions from the original CLIP and the global branch, respectively. The aim of $\mathcal{L}_g$ is to align the up-scaled image feature map with the original CLIP representation, which plays a crucial role in extracting regional features for the union and human-object branches.

Union loss $\mathcal{L}_u$: Similarly, we take a KL divergence on union HOI scores s_u and the local union supervision d_u to formulate the union loss: $\mathcal{L}_u = \frac{1}{M}\sum_{m=1}^{M} \mathcal{D}_{KL}(s_u^m||d_u^m)$. Here M is the total number of human-object combinations for a given image.

Human-object pairwise loss $\mathcal{L}_{ho}$: Inspired by [53, 55], we adopt a Multiple Instance Learning (MIL) strategy to train the human-object branch. In detail, we first string together all the interaction scores to form a bag, denoted as $S_{ho} = [s_{ho}^1; ...; s_{ho}^M] \in \mathbb{R}^{M \cdot N}$, where s_{ho}^m stands for the score of the m-th pair. Then we take maximization over all pairs to obtain the image-wise interaction scores: $\hat{s}_{ho} = \underset{m}{max}\, S_{ho}$. This step essentially distills the most representative HOIs from the pairwise predictions, providing a consolidated representation of image-wise interaction scores that can be guided by the global CLIP supervision d_g. Formally, the human-object loss $\mathcal{L}_{ho}$ is a KL divergence defined on $\hat{s}_{ho}$ and d_g: $\mathcal{L}_{ho} = \mathcal{D}_{KL}(\hat{s}_{ho}, d_g)$.

3.5. Inference

During the inference stage, we combine multiple scores to obtain the final interaction score $s_{h,o}^r$ for each human-object pair $(\mathbf{b}_h, \mathbf{b}_o)$. It includes the global HOI scores s_g, the union score s_u, the normalized pairwise interaction scores p_{ho} (rather than s_{ho}), and the object detection scores $\langle s_h, s_o \rangle$ as follows:

$$s_{h,o}^r = s_g \cdot s_u \cdot p_{ho} \cdot (s_h \cdot s_o)^\gamma \tag{2}$$

where γ is a hyper-parameter to balance the HOI scores and the object detection scores.

It is noteworthy that we avoid using the original pairwise interaction score s_{ho} as it fails to measure the contribution of each pair when multiple pairs in an image share the same interaction. Instead, we institute a competitive environment among the pairs by applying a Softmax operation on S_{ho}: $\bar{S}_{ho} = \underset{m}{\text{Softmax}}(S_{ho})$. Following this, we derive the normalized pairwise interaction scores $p_{ho} = \sigma(\hat{s}_{ho}) \cdot \bar{s}_{ho}$, where $\bar{s}_{ho}$ is a row from $\bar{S}_{ho}$ and σ is Sigmoid function.

4. Experiments

4.1. Datasets and Metrics

We use the public HOI detection dataset HICO-DET to benchmark our model. The dataset contains 37,633 training images and 9,546 test images. It includes $C = 80$ common objects (the same as MSCOCO [33]) and 117 unique action categories, together forming $N = 600$ HOI categories.

We adopt the mean average precision (mAP) metric [6] for evaluating HOI detection results. A human-object pair is deemed positive when the predicted human and object boxes have an IoU of at least 0.5 with their ground truth boxes, and the HOI class is correctly classified.

4.2. Implementation Details

We use an off-the-shelf Faster R-CNN [44] pre-trained on MSCOCO to generate up to 100 object candidates for each image. It's crucial to note that we only keep the detection results and do not re-use the feature maps. We rather employ a CLIP with a ResNet-50 visual encoder, which is pre-trained on the YFCC-15M dataset [48], with an image resolution of 224×224. The CLIP model we used was implemented by OpenCLIP [2], which achieved an mAP of 35.5 on zero-shot image-wise HOI recognition on the test set, suggesting that it is capable of providing a comprehensive holistic understanding of HOIs.

During training, we use a larger image resolution with a minimum edge length of 384, while maintaining the original aspect ratio of the input images in our multi-branch network. We freeze the weights of the pre-trained visual encoder and optimize the remaining modules by AdamW, with a learning rate of 1e-4 and batch size of 16. The model is trained for 30K iterations on two NVIDIA V100 GPUs. After 15K iterations, the learning rate is decayed by a factor of 10. Following previous works [30, 63], we set feature dimension D as 1024 and the detection score weight γ as 2.8.

[2] https://github.com/mlfoundations/open_clip

Table 1. **Results comparison of different methods on HICO-DET test set** (a full table of results comparison c.f. Appendix). †means re-implementation in [53]. Here FS, WS, and ZS indicate fully-supervised, weakly-supervised, and zero-shot HOI detection methods, respectively. The notation (D) means the visual encoder or the detector is pre-trained on dataset D, D∈ {COCO, HICO-DET, YFCC-15M}.

S	Methods	Visual Encoder	Detector	HICO-DET (%)		
				Full	Rare	Non-Rare
FS	InteractNet [12]	RN50-FPN (COCO)	FRCNN (COCO)	9.94	7.16	10.77
	iCAN [11]	RN50 (COCO)	FRCNN (COCO)	14.84	10.45	16.15
	TIN [30]	RN50-FPN (COCO)	FRCNN (COCO)	17.22	13.51	18.32
	PMFNet [54]	RN50-FPN (COCO)	FRCNN (COCO)	17.46	15.56	18.00
	HOTR [23]	RN50+Transformer (COCO)	DETR (HICO-DET)	25.10	17.34	27.42
	QPIC [47]	RN101+Transformer (COCO)	DETR (COCO)	29.90	23.92	31.69
	GEN-VLKT [32]	RN50+Transformer (HICO-DET)	DETR (HICO-DET)	33.75	29.25	35.10
	HOICLIP [38]	RN50+Transformer (HICO-DET)	DETR (HICO-DET)	34.69	31.12	35.74
WS	Explanation-HOI† [1]	ResNeXt101 (COCO)	FRCNN (COCO)	10.63	8.71	11.20
	MX-HOI [25]	RN101 (COCO)	FRCNN (COCO)	16.14	12.06	17.50
	PPR-FCN† [64]	RN50 (YFCC-15M)	FRCNN (COCO)	17.55	15.69	18.41
	PGBL [53]	RN50 (YFCC-15M)	FRCNN (COCO)	22.89	22.41	23.03
ZS	*baseline*	RN50 (YFCC-15M)	FRCNN (COCO)	10.48	9.45	10.78
	ours	RN50 (YFCC-15M)	FRCNN (COCO)	17.12	20.26	16.18

4.3. Quantitative Results

As shown in Tab. 1, with our multi-level knowledge integration strategy from CLIP, our approach achieves 17.12 mAP with ResNet-50, which is on par or even surpasses some fully-supervised (FS) [11, 12, 30, 54] and weaky-supervised (WS) [1, 25, 64] methods. For the FS and WS methods, we observe the mAP on Non-rare classes (i.e., those with more than 10 HOI instance annotations in the training set) is always higher than on Rare classes. This skew is to be expected given that HICO-DET is inherently an imbalanced dataset [14, 45]. Models tend to learn frequently occurring patterns for which they have training supervision. Despite certain HOIs being simpler to learn, their performance may lag due to the relative lack of supervision, compared to the more challenging yet annotated HOIs.

Remarkably, we obtain a higher mAP on Rare classes compared to Non-Rare classes in our results. This consequence stems from the integration of CLIP for HOI representation learning and model supervision. Firstly, CLIP is pre-trained on large-scale image-text pairs and has potentially encountered every imaginable HOI scenario during its pre-training phase. We build our model on top of CLIP components, which allows us to exploit its strong generalization capability for learning a better HOI representation. Secondly, when comparing our results with PGBL [53], which also exploits CLIP for HOI representation learning and achieves the best performance in a weakly-supervised setting (i.e., image-level HOI annotations are available). We experience a small drop on Rare classes (from 22.41 → 20.26), but a significant drop on Non-Rare classes (from 23.03 → 16.18). This disparity suggests that the learning of Non-Rare HOIs is more reliant on strong annotations, whereas Rare HOIs can be effectively learned by distilling the 'dark knowledge' from CLIP scores. Consequently, we sidestep issues associated with the long-tailed distribution.

Table 2. **Comparison of inference speed and performance** between baseline, training-free (TF) approach, and our method. TF* means d_g is added to predict s^r_{ho} on top of TF.

Exp	Speed (fps)	mAP (%)		
		Full	Rare	Non-Rare
base	**56.19**	10.48	9.45	10.78
TF	6.52	11.19	13.98	10.37
*TF**	6.47	12.24	15.75	11.19
ours	35.64	**17.12**	**20.26**	**16.18**

4.4. Ablation Studies

In this section, we mainly assess the effectiveness of each component with detailed ablation studies on HICO-DET dataset. We first introduce our baseline, based on which we will answer some interesting questions regarding the model design and learning.

Baseline: Our baseline model is constructed on top of the human-object branch, where the human-object representation v_{ho} is used to predict their normalized interaction scores s_{ho}, which are supervised by $\mathcal{L}g$. During the inference process, the final interaction scores in 2 are recomputed as $s^r_{h,o} = p_{ho} \cdot (s_h \cdot s_o)^\gamma$.

Why not a training-free (TF) approach? In a TF approach, we directly apply CLIP scores on union regions d_u along with object detection scores $\langle s_h, s_o \rangle$ for inference. This reformulates Eq. 2 as: $s^r_{h,o} = d_u \cdot (s_h \cdot s_o)^\gamma$.

While this straightforward approach only relies on a pre-trained CLIP and avoids the need for model training, it has two notable drawbacks compared to our method: (i) It is

Table 3. **Ablation study of multi-level incorporation on HICO-DET dataset.** The baseline is the human-object (h-o) branch, and we add other branches on top of it. We denote 'early' as the early fusion of union features and 'late' as the late fusion of union scores.

Exp	Branch			mAP (%)		
	h-o	union	global	Full	Rare	Non-Rare
0	✓	-	-	10.48	9.45	10.78
1	✓	✓(late)	-	14.49	16.33	13.95
2	✓	-	✓	15.84	17.91	15.21
3	✓	✓(early)	✓	14.64	16.09	14.21
4	✓	✓(late)	✓	**17.12**	**20.26**	**16.18**

Table 4. **Ablation study of different supervision strategies on HICO-DET dataset. g** means supervision from d_g and **u** means supervision from union region CLIP scores d_u. **g+u** indicates using both supervisions. In this table, the global branch is added for all the experiments and supervised with d_g.

Exp	Branch		mAP (%)		
	union	h-o	Full	Rare	Non-Rare
0	g	g	15.05	16.76	14.55
1	u	u	14.22	15.00	13.99
2	u	g	**17.12**	**20.26**	**16.18**
3	u	g+u	15.83	18.20	15.13
4	g+u	g	16.96	19.54	16.18

time-consuming for inference. For each image, the approach requires forwarding the CLIP model for the image and all union regions, resulting in $M + 1$ forward passes. As indicated in Tab. 2, given the detected bounding boxes, the inference speed for the TF approach is only 6.52 frames per second (fps) on a single Nvidia V100 GPU with a batch size of 1, whereas our method achieves an fps of 35.64. (ii) The performance is low, even enhanced by d_g (i.e., TF*). This issue arises because both d_g and d_u correspond to large regions, and as such, they fail to specify the interacted human and object pair. Consequently, the CLIP scores, when directly used for inference, tend to be noisy. In contrast, our method opts to distill knowledge from these scores at different levels, leading to a significant performance increase, from an mAP of 12.24 to 17.12.

Does multi-level CLIP knowledge distillation strategy work? As demonstrated in Tab. 3, the answer is definitively yes. We added the union branch and the global branch on top of our baseline model (Exp 0). The results indicate that the union branch enhances the mAP from 10.48 to 14.49 (Exp 0 vs. 1), while the global branch raises the mAP from 10.48 to 15.84 (Exp 0 vs. 2). When we combine both branches, the mAP sees a more significant improvement, from 10.48 to 17.12 (Exp 0 vs. 4). These results underscore the effectiveness of the multi-level CLIP knowledge distillation strategy.

How to incorporate contextual cues from the union region? In Tab. 3, we also explore two different designs for integrating the contextual union branch, including an early fusion strategy and a late fusion strategy. For the early fusion strategy, we concatenate the union feature v_u with human-object appearance features $\langle v_h, v_o \rangle$ and their spatial encoding v_{sp} to compose v_{ho}. The experimental results (Exp 3 vs. 4) indicate that the early fusion strategy performs considerably worse than a late fusion strategy (14.64 vs 17.12), and even underperforms Exp 2 where the union branch is not included.

We hypothesize that this is because the union feature, encompassing a large region, may include other HOIs or background distractions, making it noisy. These irrelevant features may not offer valuable contextual information for HOI representation learning, particularly in a zero-shot setup where the training signals are also somewhat noisy. Conversely, with the late fusion strategy, the union branch aims to learn context-aware union HOI scores. Although these union scores cannot precisely describe the specific human-object pair, they provide some contextual HOI cues that are compatible with the HOI predictions from the other branches.

Table 5. **Generalization to Unseen Categories**. We randomly select N' HOI categories for model learning.

N'	mAP (%)		
	Full	Rare	Non-Rare
100	15.25	18.97	14.15
300	16.45	19.48	15.55
600	**17.12**	**20.26**	**16.18**

What kind of supervision works best for different branches? In Tab. 4, we compare different supervision strategies for each branch. By default, we supervise the image branch with global supervision d_g. Then, we apply d_g on the union branch in the same manner as $\mathcal{L}_{ho}$ in Exp 0. This results in an mAP drop from 17.12 to 15.05, indicating that the union features, being noisy and non-discriminative, are not suitable for a MIL strategy. Besides, we apply local supervision d_u on the human-object branch in Exp 1 and observe the drop in mAP to 14.22. This is because of the noisiness of d_u, as it does not always correspond accurately to a specific human-object pair.

Experimental results reveal that the mismatch of the supervision signals can result in a performance drop, and Combining both types of supervision does not necessarily lead to better results (as seen in Exp 3 and 4).

Can our method generalize to unseen categories? To answer this question, we randomly select N' out of $N = 600$ HOI categories on HICO-DET during training. This implies the class dimensions of predicted scores $\langle s_g, s_u, s_{ho} \rangle$ and CLIP supervisions d_g, d_u are set to N'. For inference, we evaluate across all 600 categories. As shown in Tab. 5, even though the training is limited to just 100 categories, we observed only a slight 1.87 mAP drop compared to the

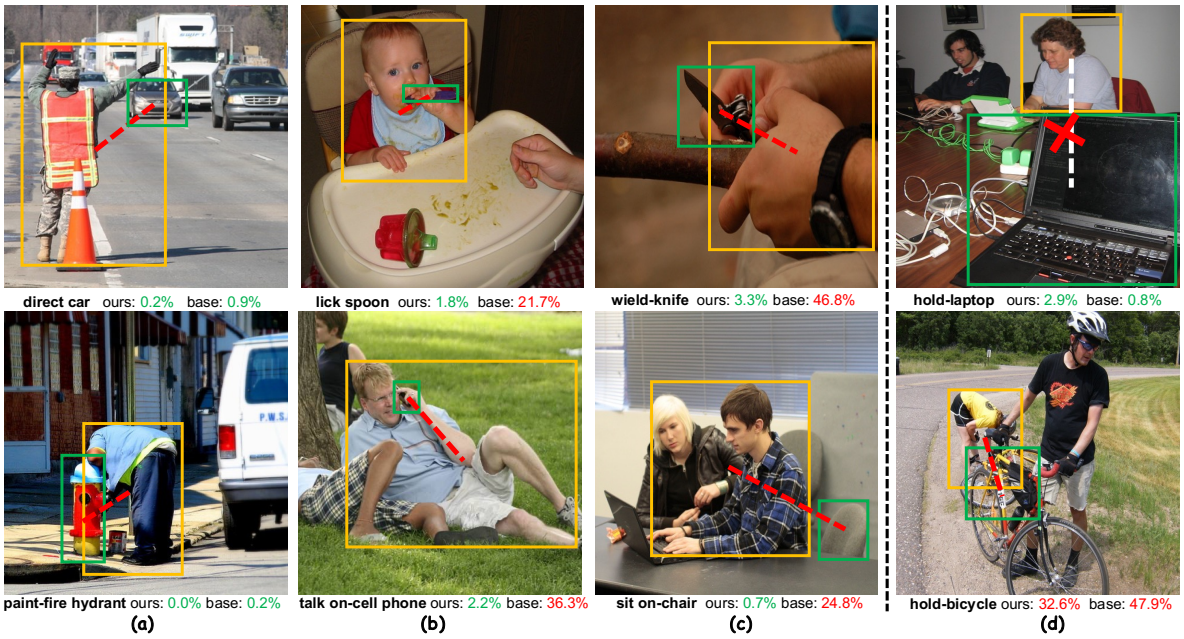

Figure 4. **Visualization of the HOI detection results.** We compare our method with the baseline model based on the relative ranking of the detection scores, which is presented in percentile format. The percentiles highlighted in green signify the model's confident HOI predictions, whereas those in red indicate negative HOI predictions that the model treats as background.

final model. This indicates that our approach effectively captures common knowledge that can be shared across all HOI categories.

4.5. Qualitative Results

Figure 4 offers a qualitative evaluation of our method. Due to the way we incorporate multiple scores into our final prediction s^r_{ho} in our methodology, it's not practical to directly compare the HOI scores with the baseline. Instead, our visualization and comparison with the baseline model are based on the relative ranking of the detection scores rather than their absolute values. Concretely, for each HOI prediction, we present its ranking amongst all predictions belonging to the same category across the entire test set. The ranking is exhibited in the form of a percentile (top $p\%$), wherein a lower percentile value (smaller p) signifies the model's strong confidence in the positive prediction (represented by green colored numbers).

As depicted in Fig.4(a), both our method and the baseline successfully identify some Rare HOI classes. However, when the objects are quite small, as shown in Fig.4(b), our model tends to offer more confident predictions, attributing this advantage to its ability to factor in contextual cues. For example, in the top image, our model infers that a baby is likely licking a spoon due to the context of sitting in a baby chair with residual sauce visible. In a similar vein, Fig.4(c) showcases situations where objects are heavily occluded. Despite this challenge, our model manages to discern some unapparent relationships by considering the overall environment. As an illustration, in the image at the bottom, a man in an office, engaged in computer work on the table, is more likely identified by our model as 'sitting on a chair', which would be a challenging task for the baseline method.

5. Limitations and Future Works

Although inspiring results have been achieved by our method, the zero-shot HOI detection is far from satisfactory. As an example of its limitations, the top image in Fig.4(d) shows a typical failure case where our model incorrectly associates a person with a computer that is considerably distant. This error can be attributed to the absence of adequate supervision for pairwise associations. Furthermore, the bottom image exhibits a scenario where our method struggles to recognize relations when the object is completely obscured.

A potential area for further exploration based on this work involves the detection of ambiguous HOI associations. Previous research has investigated this issue within fully-supervised or weakly-supervised settings [31, 35, 53]. However, transferring these learnings to a zero-shot setup remains a largely unexplored area. Besides, this study employs a classic CLIP structure for zero-shot HOI detection. Nonetheless, it is intriguing to explore various adaptations of CLIP [58–60] for enhancing performance in this task.

Acknowledgement

We acknowledge funding from European Research Council under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 101021347) and Flemish Government under the Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen programme.

References

[1] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-based weakly-supervised learning of visual relations with graph networks. In *ECCV*, 2020. 1, 2, 3, 6

[2] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020. 1

[3] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020. 3

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

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 5

[7] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022. 1

[8] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. *arXiv preprint arXiv:2204.04911*, 2022. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[10] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 1, 2

[11] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 2, 6

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 6

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

[14] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019. 1, 3, 6

[15] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV2017*, 2017. 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[18] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 3

[19] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 3

[20] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 3

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021. 1

[22] Mert Kilickaya and Arnold Smeulders. Human-object interaction detection via weak supervision. *arXiv preprint arXiv:2112.00492*, 2021. 1

[23] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 6

[24] 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. *arXiv preprint arXiv:2203.14709*, 2022. 2

[25] Suresh Kirthi Kumaraswamy, Miaojing Shi, and Ewa Kijak. Detecting human-object interaction with mixed supervision. In *WACV*, 2021. 1, 2, 3, 6

[26] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 1

[27] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*, 2021. 1

[28] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. 1

[29] 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 *CVPR*, 2020. 2

[30] 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 *CVPR*, 2019. 5, 6

[31] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. In *CVPR*, 2019. 2, 8

[32] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. *arXiv preprint arXiv:2203.13954*, 2022. 3, 6

[33] 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 *ECCV*, 2014. 5

[34] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao Gao. Future frame prediction for anomaly detection – a new baseline. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[35] Xinpeng Liu, Yong-Lu Li, Xiaoqian Wu, Yu-Wing Tai, Cewu Lu, and Chi-Keung Tang. Interactiveness field in human-object interactions. *arXiv preprint arXiv:2204.07718*, 2022. 8

[36] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACMMM*, 2020. 1, 2, 3

[37] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*, 2021. 1

[38] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *CVPR*, 2023. 1, 3, 6

[39] Hitoshi Nishimura, Satoshi Komorita, Yasutomo Kawanishi, and Hiroshi Murase. Sdof-tracker: Fast and accurate multiple human tracking by skipped-detection and optical-flow. *arXiv preprint arXiv:2106.14259*, 2021. 1

[40] Guansong Pang, Cheng Yan, Chunhua Shen, van den Hengel Anton, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[41] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019. 1

[42] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019. 3

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015. 3, 5

[45] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 1, 3, 6

[46] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. In *NeurIPS*, 2022. 1

[47] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 6

[48] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015. 5

[49] Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 2

[50] Mesut Erhan Unal and Adriana Kovashka. Weakly-supervised hoi detection from interaction labels only and language/vision-language priors. *arXiv preprint arXiv:2303.05546*, 2023. 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[52] Mrabti Wafae, Baibai Kaoutar, Bellach Benaissa, Oulad Haj Thami Rachid, and Tairi Hamid. Human motion tracking: A comparative study. *Procedia Computer Science*, 148:145–153, 2019. 1

[53] Bo Wan, Yongfei Liu, Desen Zhou, Tinne Tuytelaars, and Xuming He. Weakly-supervised hoi detection via prior-guided bi-level representation learning. In *ICLR*, 2023. 1, 3, 5, 6, 8

[54] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019. 1, 2, 6

[55] Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. MAF: Multimodal alignment framework for weakly-supervised phrase grounding. In *EMNLP*, 2020. 5

[56] Mingrui Wu, Jiaxin Gu Gu, Yunhang Shen, Mingbao Lin, Chao Chen, Xiaoshuai Sun, and Rongrong Ji. End-to-end zero-shot hoi detection via vision and language knowledge distillation. *arXiv preprint arXiv:2204.03541*, 2022. 2, 3

[57] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, 2022. 1

[58] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 8

[59] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023. 8

[60] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 1, 8

[61] 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. *NeuIPS*, 2021. 2

[62] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. *arXiv preprint arXiv:2112.01838*, 2021. 2

[63] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021. 1, 2, 5

[64] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *ICCV*, 2017. 1, 3, 6

[65] 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. 1

[66] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 1

[67] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019. 2

[68] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020. 2