Learning Transferable Human-Object Interaction Detector with Natural Language Supervision

Suchen Wang[1], Yueqi Duan[2], Henghui Ding[1], Yap-Peng Tan[1], Kim-Hui Yap[1], Junsong Yuan[3]
[1]Nanyang Technological University [2]Tsinghua University [3]State University of New York at Buffalo
{wang.sc,ding0093,ekhyap,eyptan}@ntu.edu.sg, duanyueqi@tsinghua.edu.cn, jsyuan@buffalo.edu

Abstract

It is difficult to construct a data collection including all possible combinations of human actions and interacting objects due to the combinatorial nature of human-object interactions (HOI). In this work, we aim to develop a transferable HOI detector for unseen interactions. Existing HOI detectors often treat interactions as discrete labels and learn a classifier according to a predetermined category space. This is inherently inapt for detecting unseen interactions which are out of the predefined categories. Conversely, we treat independent HOI labels as the natural language supervision of interactions and embed them into a joint visual-and-text space to capture their correlations. More specifically, we propose a new HOI visual encoder to detect the interacting humans and objects, and map them to a joint feature space to perform interaction recognition. Our visual encoder is instantiated as a Vision Transformer with new learnable HOI tokens and a sequence parser to generate unique HOI predictions. It distills and leverages the transferable knowledge from the pretrained CLIP model to perform the zero-shot interaction detection. Experiments on two datasets, SWIG-HOI and HICO-DET, validate that our proposed method can achieve a notable mAP improvement on detecting both seen and unseen HOIs. Our code is available at https://github.com/scwangdyd/promting_hoi.

1. Introduction

Human-object interaction (HOI) detection plays a vital role in human-centric visual analysis tasks and provides deeper understanding of human intentions and behaviors [6, 10, 24, 31, 37, 41, 42]. It aims to localize the interacting humans and objects and then recognize their interactions. The interaction can be treated as a pair of human actions and objects, *e.g.*, riding bicycle. Given its combinatorial nature, it is impractical to create a data collection to include all possible HOIs, especially when the action and object category space becomes large (*e.g.*, 400 actions and 1000 objects in SWIG-HOI [47]). This motivates us to study a transferable HOI detector which can be readily extended to numerous potential unseen interactions.

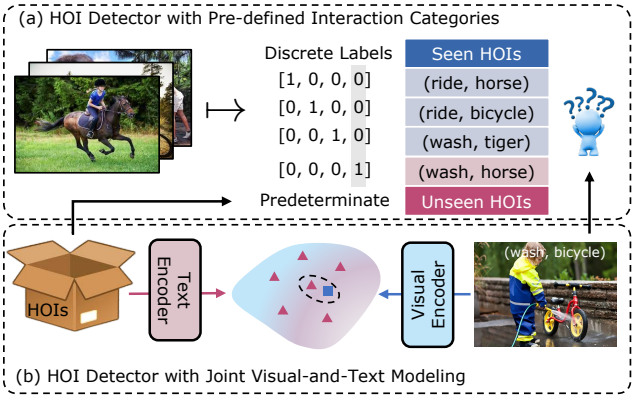

Figure 1. (a) Most existing HOI detectors classify the novel interactions in a predetermined manner. They are difficult to handle novel interactions that are out of the predefined list. (b) We aim to develop a transferable HOI detector via joint visual-and-text modeling which is more suitable to handle unseen interactions. Given an image, the visual encoder detects the interacting humans and objects and maps them into the joint space (*e.g.*, the blue square). We then conduct the nearest search among text features (triangles) for HOI recognition.

Recent works [2, 15, 17, 21] have used compositional learning to enhance the generalization ability of HOI detectors on unseen interactions. Their core idea is to decompose the interaction into an action and an object, and conduct data augmentation to generate new combinations of actions and objects [15–17]. However, there is a basic assumption - the list of unseen interactions is available, so that specific samples of interactions can be generated from existing ones accordingly. However, it is still an open problem how to automatically determine the validity of the generated interactions without any given priors. In this sense, existing methods are only suitable for predetermined cases but not transferable to other unseen interactions.

In this paper, we aim to train a transferable HOI detector without assuming any priors for the unseen interactions.

Figure 1 shows the main difference of our method with existing solutions. Notably, most previous works use discrete labels to learn a classifier with a fixed size of weights. This predetermined setting limits their generalizability and efficacy since it cannot handle any novel HOIs out of the predefined list. Motivated by the recent success of CLIP [35], we transform independent one-hot HOI labels into natural language supervision by joint visual-and-text modeling. In this way, we can reformulate HOI detection as a visual-to-text matching problem and enable the recognition for the unseen interactions.

Specifically, we propose a new one-stage HOI detector comprising one visual encoder and one text encoder. For the visual encoder, we present (1) a novel HOI Vision Transformer by designing additional `[HOI]` tokens and (2) a sequence parser module to encourage the unique HOI detections in the image. We take the output from the final layer with respect to `[HOI]` tokens as the representation of interactions. Then, we feed them into two heads for the bounding box regression and interaction recognition, respectively. We use a regressor to predict the bounding boxes of interacting humans and objects, and estimate a confidence score for the prediction. Furthermore, we project the visual feature to the joint visual-and-text feature space to search for the nearest interaction labels embedded by the text encoder.

The interaction category is typically defined as a pair of actions and objects. Instead of treating them as discrete labels, we aim to build natural language supervision using the action and object names and encode them to the joint visual-and-text space to explore their semantic correlations. Recent works [35, 54] have shown that the context words surrounding the class name can significantly influence the recognition accuracy. In our case, it becomes more complex and may require different sentence formats from case to case. For example, given an action "riding" and an object "bicycle", we can form a sentence like "a photo of a person riding bicycle". However, given an action "fishing" and an object "fishing pole", it does not make sense for the sentence, "a photo of a person fishing fishing pole". To facilitate the text generation, we propose an automated way to form the sentence using learnable tokens to replace the manually determined words. This brings in more generalized results for both seen and unseen interactions.

Our contributions are summarized below. (1) We reformulate the HOI detection as a visual-to-text matching problem and enable the detection of unseen interactions. (2) We propose a new one-stage HOI detector with the Vision Transformer by designing additional HOI tokens and a HOI sequence parser to jointly detect humans and objects and recognize their interaction. (3) Experiments on two datasets, HICO-DET and SWIG-HOI, validate that the proposed method can achieve state-of-the-art results on HOI detection, especially for unseen interactions.

2. Related Work

Generic HOI detection The standard HOI detection [8, 11, 12, 25, 34, 49, 51] mainly focuses on the known interactions. Existing methods can be roughly divided into two groups, one-stage methods [9, 22, 28, 48, 50] and two-stage methods [14, 27, 43, 45, 52, 55, 56]. Two-stage methods usually apply an offline object detector to first detect humans and objects, and then feed detected boxes into an interaction model for classification. Without the need for end-to-end training with object detector, the second stage usually uses a more sophisticated architecture to analyze the relations among the detected bounding boxes, *e.g.*, multi-steams [12, 14, 26] or graphs [11, 34, 52, 53]. Besides, it is advantageous to consider other information to assist the interaction recognition, *e.g.*, human pose [8, 27, 45], spatial distribution [43, 53], *etc*. Compared with two-stage methods, one-stage methods aim to use one model to jointly detect the bounding boxes and recognize the interactions. Early one-stage detectors [9, 28, 50] often apply a parallel structure to simultaneously generate the bounding box candidates and predict the interacting points or pairs, followed by a matching step to form the final HOI predictions. Recent works [5, 23, 40, 57] formulate the HOI detection as a set prediction problem and have proposed various Transformer-based detectors [3].

Novel HOI detection Given the large combinatorial category space of interactions, it is challenging to build a data collection including all possible categories. Several recent works [18, 29, 32, 48] have studied how to handle the novel or zero-shot HOIs. Based on whether the target object is known, novel interactions can be roughly divided into two types. The first type is the novel combinations between the known actions and known objects [15, 17, 29]. Contrast this, it is more challenging to detect unseen interactions with novel objects [16, 18, 48]. Some works have proposed to use semantic word embeddings [2, 29, 48] to assist the zero-shot recognition. But different from our method, they still rely on a predetermined classifier that is difficult to handle other unseen interactions out of the predefined list.

Natural language supervision Vision-language pre-training has recently emerged as a promising approach for image [19, 35] and video understanding [46]. Different from the traditional way of using discrete labels, it provides a new paradigm to perform recognition based on visual and text feature alignment. It naturally suits the use of zero-shot transfer for various downstream tasks [39]. Recent works have also explored how to use the transferable knowledge of pretrained models to address visual question answering (VQA) [20], zero-shot object detection [13], and image captioning [39], *etc*. Motivated by their works, we aim to explore how to learn a transferable HOI detector based on natural language supervision.

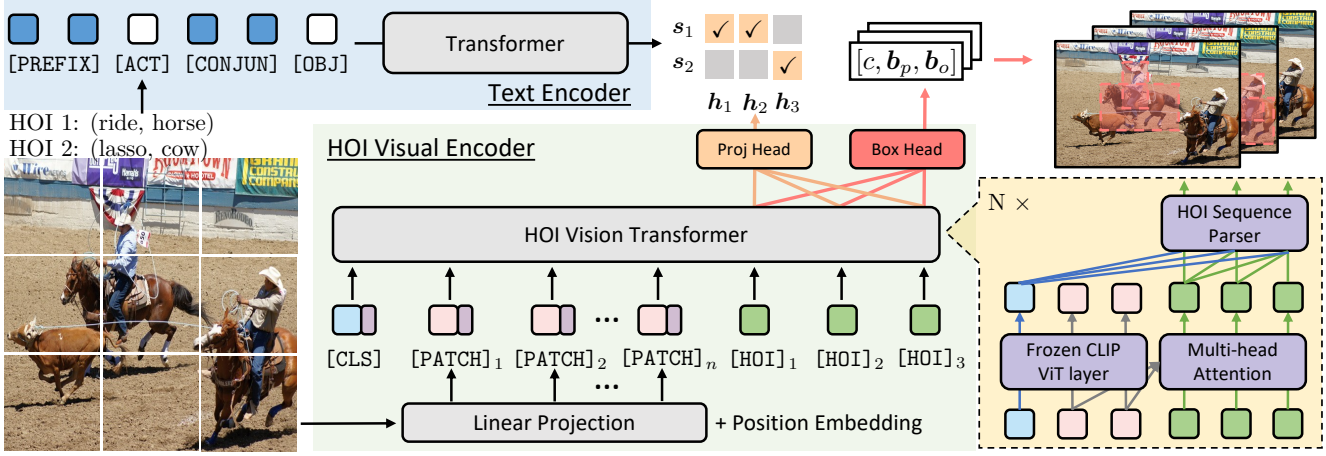

Figure 2. The architecture of our proposed one-stage transferable HOI detector. It consists of one HOI visual encoder and a text encoder. To ease the generation of raw text description of interactions, we use a set of learnable tokens `[PREFIX]` and `[CONJUN]` to automatically learn the sentence format rather than using manually defined context words. The visual encoder is instantiated as a Vision Transformer. We append new `[HOI]` tokens in the input sequence and propose an HOI sequence parser to detect multiple interactions in the image. The visual encoder detects the bounding boxes of interacting humans and objects and maps them to the joint visual-and-text feature space, where we can apply the nearest search with text features to recognize the interaction category.

3. Methodology

We address the problem of human-object interaction (HOI) detection and aim to develop a transferable detector that can handle unseen interactions. Formally, we define the interaction as a tuple $\{(\boldsymbol{b}_p, \boldsymbol{b}_o, a, o)\}$, where $\boldsymbol{b}_p, \boldsymbol{b}_o \in \mathbb{R}^4$ indicates the bounding box of interacting humans and objects, $a \in \mathcal{A} = \{1, \dots, A\}$ represents the human action, and $o \in \mathcal{O} = \{1, \dots, C\}$ denotes the object category. Due to the combinatorial nature of interactions, it is impractical to form a data collection including all possible combinations of actions and objects, especially when $\mathcal{A}$ and $\mathcal{O}$ cover a large space (*e.g.*, around 400 actions and 1000 objects in SWiG-HOI [47]). Similar to previous works [2, 15, 16], we focus on the generalized zero-shot interaction detection setup. Specifically, we assume that both $\mathcal{A}$ and $\mathcal{O}$ are known but only partial combinations between them can be observed during the training. At inference, we would like to detect both the seen and unseen human-object interactions.

3.1. Preliminary: Visual-and-Text Modeling

The classical approach for recognition tasks is to learn a visual classifier that maps the image input to a fixed set of discrete labels. Such a treatment is inherently difficult to handle the unseen categories which are out of the original label set. To address this issue for HOI detection, existing works [2, 15, 16, 38] often rely on compositional learning. The core idea is to decompose HOIs into separate action and object components. For the classifier level decomposition, it assumes that a and o are independent and approximates $p(a, o)$ as $p(a) \cdot p(o)$, where $p(\cdot)$ denotes the confidence score. However, this method fails to model the action-object correlations. Besides, it is error-prone when the human subject performs multiple interactions with different objects. To better model the joint distribution, recent works [2, 15, 16] propose to fill in the missing interactions via data generation. However, these methods need to predetermine the categories of unseen interactions so that specific samples can be generated during the training. This makes it still difficult to handle other unseen interactions that are out of the predefined list.

In this work, we explore recognizing interactions based on their text description. Motivated by CLIP [35], we aim to learn a joint visual-and-text feature space such that the HOIs in image and associated text description can be well aligned. There are two advantages. First, different from seeing interactions as discrete labels, we can leverage the semantic similarity among interactions to eliminate the challenges of unseen interaction recognition. Second, compared with learning a classifier with a fixed label set, this way is more flexible to perform the zero-shot tasks, since we can conduct the nearest search in the joint feature space to link the unseen visual concept with its raw text description. To examine whether this idea works for the HOI detection task, we start with CLIP [35], which has learned a good visual-and-text reference from a large collection of image-text pairs and shown great success on various zero-shot recognition tasks [13, 39].

It should be noted that CLIP is originally designed for image-level recognition, while we aim to detect the bounding boxes of interacting humans and objects and perform instance-level interaction recognition. As a preliminary

Table 1. Preliminary studies on HICO-DET dataset. We implement a simple baseline using a pretrained object detector and CLIP to conduct the HOI detection task. It shows that the baseline can achieve a promising result on unseen interactions, even without any tuning on HICO-DET.

	No tuning	Unseen	Seen	Full
Shen, WACV18 [38]		5.62	-	6.26
FG, AAAI20 [2]		10.93	12.60	12.26
VCL, ECCV20 [15]		10.06	24.28	21.43
ATL, CVPR21 [16]		9.18	**24.67**	**21.57**
FCL, CVPR21 [17]		13.16	24.23	22.01
Pretrained detector + CLIP	✓	**13.42**	15.10	14.76

study, we implement a simple baseline by combining CLIP with an off-the-shelf object detector. Specifically, we use the pretrained Faster RCNN [36] to produce the bounding boxes. Then, we pair every human with an object box and crop their union region as the input of the CLIP visual encoder. To construct the raw text description of interactions, we follow [35] to generate the sentence as "a photo of a person `[ACT] [OBJ]`", where `[ACT]` and `[OBJ]` can be replaced by the action and object category name respectively, *e.g.*, riding horse, lassoing cow, *etc*. Table 1 compares its performance with state-of-the-art methods on HICO-DET dataset [4] using the standard mAP evaluation metric. We observe that such a simple baseline can achieve a promising result on unseen interactions, even without any tuning on the target dataset. Motivated by this result, we further explore if we can reformulate the HOI detection problem in the manner of visual-and-language modeling like CLIP and learn a one-stage HOI detector that is transferable to unseen interactions.

3.2. The Proposed Method

Figure 2 shows the overall architecture of our proposed HOI detector. It mainly consists of an HOI visual encoder and a text encoder. The visual encoder takes as input an RGB image and outputs a set of predictions, $\{(\boldsymbol{h}, c, \boldsymbol{b}_p, \boldsymbol{b}_o)\}$, where $\boldsymbol{h} \in \mathbb{R}^D$ denotes the feature representation of interactions, $\boldsymbol{b}_p, \boldsymbol{b}_o \in \mathbb{R}^4$ denotes the bounding boxes of interacting humans and objects, and $c \in [0, 1]$ represents the confidence score for the bounding box prediction. For the text encoder, it takes as input the raw text of interactions (*e.g.*, riding horse, lassoing cow) and encodes them into a set of semantic features $\{\boldsymbol{s}\}$, where $\boldsymbol{s} \in \mathbb{R}^D$ shares the same feature space as $\boldsymbol{h}$. Then, we perform the interaction recognition based on the similarity $\boldsymbol{h}^\top \boldsymbol{s}$. In the following, we provide the details of our visual and text encoder as well as the loss functions to learn the HOI detector.

3.2.1 HOI Visual Encoder

In our preliminary study, we find that CLIP [35] has obtained a good transferable visual-and-text reference by learning from a large collection of image-text pairs. Our method treats CLIP as an external knowledge base and aim to distill its knowledge for HOI detection. It is a non-trivial task since CLIP is originally designed for image-level recognition, while we want to detect the bounding boxes and perform an instance-level interaction recognition. To mend this gap, we propose a new ViT-based [7] visual encoder.

ViT-based Visual Encoder The visual encoder takes an image $\boldsymbol{I} \in \mathbb{R}^{H \times W \times 3}$ with a fixed resolution (*e.g.*, 224×224). It divides the image into small patches with size $P \times P$ (*e.g.*, 16×16) and projects them as a sequence $[\boldsymbol{x}_1^0; \boldsymbol{x}_2^0; \ldots; \boldsymbol{x}_N^0]$, where $\boldsymbol{x}_i^\ell \in \mathbb{R}^D$ denotes the embedding of i-th image patch and $\ell \in \{0, 1, \ldots, L\}$ denotes the index of Transformer layers. For the ease of presentation, here we consider that $\boldsymbol{x}_i^0$ has already included the positional embedding. In ViT, a learnable embedding $\boldsymbol{z}_0^0 \in \mathbb{R}^D$ (also known as `[CLS]` token) is prepended to the sequence of patch embeddings. Its output $\boldsymbol{z}_0^L$ from the last layer will serve as the representation of the image. The `[CLS]` token aggregates the useful information from the sequence of patches using the softmax attention mechanism [44]. Denote the input as $\boldsymbol{X}_0 = [\boldsymbol{z}_0^0; \boldsymbol{x}_1^0; \boldsymbol{x}_2^0; \ldots; \boldsymbol{x}_N^0] \in \mathbb{R}^{(N+1) \times D}$. The ViT performs the following steps at each layer $\ell = 1, \ldots, L$,

$$\begin{aligned} \boldsymbol{X}'_\ell &= \mathrm{MHA}\big(\boldsymbol{X}_{\ell-1}\big) + \boldsymbol{X}_{\ell-1} \\ \boldsymbol{X}_\ell &= \mathrm{MLP}\big(\mathrm{LN}(\boldsymbol{X}'_\ell)\big) + \boldsymbol{X}'_\ell \end{aligned} \tag{1}$$

where MHA refers to the Multi-Head Attention module [44], LN represents the LayerNorm [1] and MLP is a two-layer perceptron. We assume that MHA has an LN layer inside to normalize the input and it is omitted in the above equation for a concise presentation.

New Tokens for HOI Detection Different from image-level recognition, we aim to perform instance-level HOI detection. In this way, similar to `[CLS]`, we introduce a sequence of new HOI tokens `[HOI]`$_1$, `[HOI]`$_2$, ..., `[HOI]`$_M$. We expect that they can act like `[CLS]` to aggregate the useful information with respect to different interactions in the image. Let $\boldsymbol{H}_0 = [\boldsymbol{h}_1^0; \boldsymbol{h}_2^0; \ldots; \boldsymbol{h}_M^0]$ be the initial state of `[HOI]` tokens, where $\boldsymbol{h}_i^\ell \in \mathbb{R}^D$. We perform the following steps alternatively to aggregate the information from the image patches,

$$\begin{aligned} \boldsymbol{H}'_\ell &= \mathrm{MHA}\big(\boldsymbol{H}_{\ell-1}, \boldsymbol{X}_{\ell-1}^{[1:]}, \boldsymbol{X}_{\ell-1}^{[1:]}\big) + \boldsymbol{H}_{\ell-1} \\ \boldsymbol{H}_\ell &= \mathrm{MLP}\big(\mathrm{LN}(\boldsymbol{H}'_\ell)\big) + \boldsymbol{H}'_\ell \end{aligned} \tag{2}$$

Here we treat $\boldsymbol{H}_{\ell-1}$ as the query and $\boldsymbol{X}_{\ell-1}^{[1:]}$ as the key and value for MHA. We assume that all inputs will first pass a LN layer to be normalized. For a concise presentation, we omit the LN in the above MHA equation. It is worth noting that we mask out the `[CLS]` token $\boldsymbol{z}_0^{\ell-1}$ and only feed the image patch embeddings $\boldsymbol{X}_{\ell-1}^{[1:]} = [\boldsymbol{x}_1^{\ell-1}; \boldsymbol{x}_2^{\ell-1}; \ldots; \boldsymbol{x}_N^{\ell-1}]$.

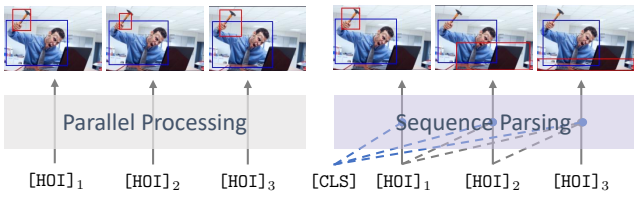

Figure 3. Illustration of the HOI sequence parsing. The main goal is to enable each [HOI] token to be aware of its predecessors to encourage a unique detection.

The main reason is that [HOI] tokens tend to directly copy the feature of [CLS] and fail to identify the true meaningful image patches.

HOI Sequence Parser As shown in Figure 2, there could be multiple interactions between different humans and objects in a single image. In this case, we expect that [HOI] tokens can differentiate them and respond to different interactions. However, we find that it is difficult to achieve so solely based on the initial state, $[\boldsymbol{h}_1^0; \boldsymbol{h}_2^0; \ldots; \boldsymbol{h}_M^0]$. We suspect that this is mainly because [HOI] tokens are processed in parallel. As a result, they are not aware of what interactions have been detected by others. To alleviate this issue, we introduce a module to parse the tokens in a sequence manner.

At each layer ℓ, we obtain the output from the MHA, *i.e.*, $[\boldsymbol{h}_1^\ell; \boldsymbol{h}_2^\ell; \ldots; \boldsymbol{h}_M^\ell]$. The main goal of the parser module is to enable tokens to differentiate themselves and find different interactions. Motivated by this, we propose to parse each [HOI] token based on its predecessor in the sequence. Specifically, we aim to learn a mapping function $\mathcal{F}(\cdot|\cdot)$ to update the feature of $\texttt{[HOI]}_i$, $\boldsymbol{h}_i^\ell$, depending on [CLS] and its predecessors, $\texttt{[HOI]}_1$ to $\texttt{[HOI]}_{i-1}$. This can be written as

$$\boldsymbol{h}_i^\ell := \mathcal{F}(\boldsymbol{h}_i^\ell | \boldsymbol{z}_0^\ell, \boldsymbol{h}_1^\ell, \ldots, \boldsymbol{h}_{i-1}^\ell) \tag{3}$$

The intuition is that we expect the first token $\texttt{[HOI]}_1$ can find one interaction (*e.g.*, maybe the salient one) based on the [CLS] token (*i.e.*, $\boldsymbol{z}_0^\ell$). Then, the second token $\texttt{[HOI]}_2$ can build on what $\texttt{[HOI]}_1$ aims to detect and find a different one, and so on for the other tokens. The parser is instantiated as an MHA module and we use a mask to limit the attention region so that only the preceding ones can be considered for each token. This module is added to each layer after the steps in Eq 2. Figure 3 shows an example of how the parser works compared with the original parallel processing.

Projection Head and Bounding Box Regressor The output from the final layer will serve as the representation of interactions, *i.e.*, $[\boldsymbol{h}_1^L; \boldsymbol{h}_2^L; \ldots; \boldsymbol{h}_M^L]$. Then, we feed them to two different head networks. The first one is a linear projection, $\mathcal{F}_{proj}(\boldsymbol{h}) : \mathbb{R}^D \mapsto \mathbb{R}^D$. It maps the feature to the joint visual-and-text space. Similar to [35], we compute its similarity with the features from the text encoder for interaction recognition. The second one is a bounding box regressor $\mathcal{F}_{bbox}(\boldsymbol{h}) : \mathbb{R}^D \mapsto \mathbb{R}^9$. It predicts a confidence score and the bounding box of the interacting human and object, $[c, \boldsymbol{b}_p, \boldsymbol{b}_o]$, where $\boldsymbol{b}_t \in \mathbb{R}^4, t \in \{p, s\}$ denotes the normalized bounding box coordinate as [3]. It is worth noting that this is a class-agnostic regressor, since the object category is also recognized in the joint visual-and-text space. Furthermore, we use a confidence score $c \in [0, 1]$ to estimate whether the bounding box prediction captures a true interaction. In our method, we introduce a fixed size of M tokens, while the number of interactions in the image can vary from case to case. Suppose there are $m < M$ interactions in the image, we expect that the remaining $M - m$ tokens will have a low score such that their predictions can be filtered out.

3.2.2 Text Encoding

The main goal of the text encoder is to map the raw text description of interactions to the feature space and then compare them with the output from the above HOI visual encoder. The raw text will be tokenized and encoded as a sequence of word embeddings. A learnable [EOS] token will be appended to the end and its output from the last layer will be treated as the representation of the whole sentence. Recent studies [35, 54] have shown that the choices of contexts words surrounding the class name can significantly influence the recognition accuracy. In this work, we use the pretrained CLIP text encoder and freeze it during training. Our main focus is to explore the most suitable way to build the raw text description for human interactions.

The interaction category is typically defined as an actions-objects pair, *e.g.*, (ride, horse), (lasso, cow), *etc*. Follow [35], one naive approach is to use a predefined format to fill in the blank, *e.g.*, "a photo of a person [ACT] [OBJ]". Here [ACT] and [OBJ] can be replaced by the true category names of the action and object respectively. However, we observe that it is difficult to find a universal format that can suit all interactions. For example, given an action "fishing" and an object "fishing pole", it does not make sense for the sentence, "a photo of a person fishing fishing pole". Inspired by [54], we instead form the sentence using learnable context tokens. Figure 2 shows how we build the text input. Specifically, we introduce a few [PREFIX] tokens at the beginning of the sentence to replace the manually defined words, like "a photo of a person". Besides, we use a few learnable [CONJUN] tokens to automatically determine how to connect the category names of the action and object. In Table 4, we investigate other ways to build the text description. We observe that compared with a predefined format, such an automatic way can obtain a notable improvement for the HOI detection.

3.2.3 Training

In this subsection, we present the loss functions used to train the HOI detector. The loss functions can be roughly divided into two parts: (1) losses for the visual-and-text alignment and (2) losses for box regression.

We first perform a bipartite matching [3] between the predictions and the ground truth per image to encourage unique detection. In this way, each ground truth can only associate with one `[HOI]` token. Let $\{(\boldsymbol{h}_i, \hat{c}_i, \hat{\boldsymbol{b}}_p^i, \hat{\boldsymbol{b}}_o^i\}_{i=1}^M$ be the predictions per image and $\{(\boldsymbol{s}_i, \boldsymbol{b}_p^i, \boldsymbol{b}_o^i)\}_{i=1}^K$ be the annotated targets. The bipartite matching aims to find an association, $\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_M]$. Here $\phi_i \in \{0, 1, 2, \ldots, K\}$ denotes the index of the associated target for $\texttt{[HOI]}_i$ and $\phi_i = 0$ means that no target is assigned. Let $\mathcal{L}_m(i, \phi_i)$ be the matching cost between the i-th token and ϕ_i-th annotated target. It can be computed by

$$\mathcal{L}_m(i, \phi_i) = \mathcal{L}_b(\hat{\boldsymbol{b}}_p^i, \boldsymbol{b}_p^{\phi_i}) + \mathcal{L}_b(\hat{\boldsymbol{b}}_o^i, \boldsymbol{b}_o^{\phi_i}) + \mathcal{L}_h(\hat{\boldsymbol{h}}_i, \boldsymbol{s}_{\phi_i}) \quad (4)$$

Here $\mathcal{L}_b$ represents the bounding box regression loss, including both ℓ_1 loss and generalized IoU loss as used in [3]. The last item $\mathcal{L}_h$ is the loss for the interaction recognition. We follow CLIP [35] to compute it as the sum of visual-to-text and text-to-visual cross entropy loss,

$$\mathcal{L}_h(\hat{\boldsymbol{h}}_i, \boldsymbol{s}_{\phi_i}) = \mathcal{L}_{v2t} + \mathcal{L}_{t2v} \quad (5)$$

Different from image-level recognition, we need to compute the loss in the instance-level per image. As shown in Figure 2, one text label may correspond to multiple `[HOI]` tokens. To handle this, we rewrite the text-to-visual classification loss as

$$\mathcal{L}_{t2v} = -\log \frac{\exp(\boldsymbol{h}_i^\top \boldsymbol{s}_{\phi_i}/\tau)}{\sum_{j:\phi_j \neq \phi_i} \exp(\boldsymbol{h}_j^\top \boldsymbol{s}_{\phi_i}/\tau)} \quad (6)$$

where τ is a temperature parameter. The main idea is to omit other `[HOI]` tokens with the same text label. As each `[HOI]` token can ensure that there is only one text label to be assigned, the visual-to-text can be written as

$$\mathcal{L}_{v2t} = -\log \frac{\exp(\boldsymbol{h}_i^\top \boldsymbol{s}_{\phi_i}/\tau)}{\sum_j \exp(\boldsymbol{h}_i^\top \boldsymbol{s}_j/\tau)} \quad (7)$$

Before the backpropagation, we first find the best bipartite matching $\boldsymbol{\phi}^*$ by solving the objective function $\min_{\boldsymbol{\phi}} \sum_{i=1}^M \mathcal{L}_m(i, \phi_i) - \hat{c}_i$. Given that M is usually small, the bipartite matching will not incur much computation. Once we find $\boldsymbol{\phi}^*$, we can determine the label c_i for the confidence score, where $c_i = 0$ if $\texttt{[HOI]}_i$ does not match any target; otherwise, $c_i = 1$. Subsequently, we update the model parameters by minimizing the following loss based on the matching result $\boldsymbol{\phi}^*$,

$$\min \sum_i \mathcal{L}_m(i, \phi_i^*) + \mathcal{L}_c(\hat{c}_i, c_i) \quad (8)$$

Table 2. Comparison of our proposed method with two baselines using CLIP for interaction recognition.

	HICO-DET		SWIG-HOI	
	Seen	Unseen	Seen	Unseen
Disjoint detection + CLIP	15.10	13.42	11.98	6.96
One-stage detector + Proj	**25.41**	8.94	16.95	6.21
THID (our method)	24.32	**17.53**	**17.29**	**9.17**

where $\mathcal{L}_c$ is the standard cross-entropy loss. We initialize the MHA in Eq. 2 using the pretrained weights of CLIP. To eliminate the knowledge forgetting during the fine-tuning on target set, we frozen all parameters of CLIP (*i.e.*, parameters in Eq. 1) and only update the newly added modules.

4. Experiments

The main purpose of this work is to develop a Transferable Human-object Interaction Detector (namely, THID) that can also work on the unseen combinations between human actions and objects. In this section, we present our experiment details and report the results on the generalized zero-shot detection setup.

Datasets Our experiments are mainly conducted on two datasets, **HICO-DET** [4] and **SWIG-HOI** [47]. HICO-DET dataset provides 600 combinations between 117 human actions and 80 objects. We follow previous works [15, 16] to simulate a zero-shot detection setting by holding out 120 rare interactions. That is, we only use 480 seen interactions to train the HOI detector, while the detectors need to detect all 600 interactions. SWIG-HOI dataset provides diverse human interactions with large-vocabulary objects. It includes 400 human actions and 1000 object categories. Instead of simulating the zero-shot scenario via hold-out like HICO-DET, SWIG-HOI has unseen combinations naturally in the test set given the large category space of actions and objects. The test set has ∼14k images and ∼5.5k interactions, where ∼1.8k interactions are not included in the training set.

Implementation Details Our model is built upon the pretrained CLIP and all its parameters are frozen during our training. For the visual encoder, we use the ViT-B/16 version, where it takes as input a $224 \times 224 \times 3$ image and divides it into $16 \times 16 \times 3$ patches, leading to $196 (= 14 \times 14)$ image `[patch]` tokens. In our experiments, we introduce 10 learnable `[HOI]` tokens to detect human-object interactions. The visual encoder has a total of 12 layers. For the text encoder, we introduce 8 `[prefix]` tokens at the beginning and 4 `[conjun]` tokens to connect the words of human actions and objects. These hyperparameters were set based on grid searches. We train our model for 100 epochs on 4 GPUs with a batch size of 128. We set the learning rate as 0.0001 and use the Adam optimizer with decoupled weight decay regularization [30].

Table 3. Experimental results of various parsing modules among [HOI] tokens on SWIG-HOI. † The [CLS] token is ablated from the module.

	Parser	Seq.	[CLS]	Non-rare	Rare	Unseen
ViT-B/16				12.26	7.90	6.27
+ SetParser†	✓			15.14	9.18	6.02
+ SetParser	✓		✓	16.93	10.09	7.07
+ SeqParser†	✓	✓		15.78	11.23	8.43
+ SeqParser	✓	✓	✓	**17.67**	**12.82**	**10.04**

Evaluation Metrics We report the standard mean Average Precision (mAP) for HOI detection. We first compute the AP per interaction and then take the mean. A prediction, to be considered as true positive, requires (1) both of the predicted human and object bounding box have an Intersection-over-Union (IoU) of 0.5 or higher with the ground truth and (2) interaction prediction matches the correct category. Following [4], we further divide the interactions into non-rare, rare, and unseen cases based on their occurrences in the training set.

4.1. Ablation Studies

4.1.1 Vision Transformer with HOI tokens

We first validate our proposed method by comparing it against two straightforward baselines. (1) Disjoint detection + CLIP: We decompose the HOI detection into two disjoint stages. The first stage detects the bounding boxes of humans and objects. We fine-tune the Faster RCNN [36] on the target dataset to produce boxes. Then, their union region is cropped and fed to the pretrained CLIP model for interaction recognition. (2) One-stage detector + Projection: We modify the state-of-the-art Transformer-based one-stage HOI detector QPIC [40] to work with CLIP. Instead of directly predicting the interaction categories, we now use an MLP to project the visual feature to align with the text feature from pretrained CLIP. This is trained in an end-to-end manner. More details of the baseline methods are in the supplementary material.

Table 2 reports the performance of the baselines and our method. For the baseline, disjoint detection + CLIP, the result shows that it generally performs well on unseen interactions. But we observe that it usually fails in cases when the union regions of two different interactions are roughly the same, since it always tends to find the most salient one. For the second baseline, one-stage detector + Projection, we observe that it is good for detecting seen interactions and achieve an mAP of 25.41 and 16.95 on two datasets. But its performance on unseen interactions is generally worse than the others. This suggests that this baseline tends to overfit the seen interactions and fails to transfer the knowledge from CLIP. In comparison, our proposed method can obtain a more balanced result between the seen and unseen interactions.

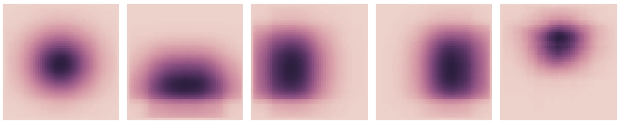

Figure 4. Distribution of detected object w.r.t. the top 5 frequently used [HOI] tokens. We draw the center of object bounding boxes in the normalized image plane.

Table 4. Experimental results of various sentence patterns on SWIG-HOI dataset. [OBJ_DEF] and [ACT_DEF] represent the detailed definition of object and action categories respectively.

Sentence pattern	N-Rare	Rare	Unseen
A photo of a person [ACT] [OBJ]	11.98	7.11	6.96
A photo of a person [ACT] [OBJ]. [OBJ_DEF]	12.22	8.66	7.70
A photo of a person [ACT] [OBJ]. [ACT_DEF]	14.97	9.62	8.22
[Prefix] [ACT] [Conjun] [OBJ]	17.67	12.82	10.04

4.1.2 Sequential HOI Parser

Here we evaluate the necessity of our proposed sequence parser. We introduce a sequence parser to encourage unique detections among [HOI] tokens. In Table 3, we show the performance of different approaches. First, we ablate the sequence parser from the model and it reduces to a vanilla ViT-B/16 [7] with only a few newly appended [HOI] tokens. It shows a huge performance drop, from 17.67 to 12.26 on non-rare interactions and 10.04 to 6.27 on unseen interactions. Second, we remove the attention constraints such that each token can consider all others instead of seeing only the tokens in front. This is named as SetParser in Table 3. We observe that this way slightly drops the mAP on non-rare interactions but significantly affects the performance on the rare and unseen ones. Third, we remove the [CLS] tokens from the sequence parser. The result suggests that including [CLS] token estimated by CLIP can bring extra benefits for the parser. We believe that it is because [CLS] token has a global understanding of the image and thus it can provide some priors for our instance-level HOI detection.

To better understand the functions of sequence parser, Figure 4 visualizes the distributions of object bounding boxes in the normalized image plane with respect to the top 5 frequently used [HOI] tokens. We observe that the first [HOI] token usually responses to the salient object in the image center. And the other tokens focus on different nearby regions to detect other interactions.

4.1.3 Text Description of Interactions

Table 4 reports the experimental results of using different sentence formats. Interestingly, we observe that appending the definition of actions and objects (provided by SWIG [33]) can boost the result. For example, the definition of "washing" is to "remove (a stain or dirt) in a washing manner." And a "car" means "a motor vehicle with four wheels". This observation suggests that a detailed defini-

Table 5. Ablation studies of unsymmetrical classification loss.

	Non-rare	Rare	Unseen
Ablation of text-to-visual loss	15.81	11.83	9.30
Multi-label soft margin loss	13.98	8.01	6.02
Hold-out cross entropy loss	17.67	12.82	10.04

Table 6. Comparison of our proposed THID with state-of-the-art methods on HICO-DET under the simulated zero-shot setting.

	One-stage	Unseen	Seen	Full
Shen, WACV18 [38]	✗	5.62	-	6.26
FG, AAAI20 [2]	✗	10.93	12.60	12.26
VCL, ECCV20 [15]	✗	10.06	24.28	21.43
ATL, CVPR21 [16]	✗	9.18	**24.67**	21.57
FCL, CVPR21 [17]	✗	13.16	24.23	22.01
THID, Ours	✓	**15.53**	24.32	**22.96**

tion can help to reduce the ambiguity to some extent and increase the accuracy. Due to the CLIP text encoder limiting the max sequence length to 76, it does not allow us to attain the performance of appending both the action and object definition. To ease the generation of raw text, we propose to use learnable tokens to replace the specific words, like "a photo of a person". In this way, we can achieve a notable improvement on all non-rare, rare, and unseen cases. In our main experiments, we focus on the above approach without appending the definition, as not all HOI datasets provide detailed explanations of interactions.

4.1.4 Unsymmetrical classification loss

Different from the original CLIP [35], the HOI detection task cannot apply the symmetric image-to-text and text-to-image classification loss, since multiple interactions in the image may share the same text description. Table 5 lists several possible ways to address this issue. A naive solution is to ablate the text-to-visual classification loss. Another possible solution is to treat it as a multi-label classification problem. Then, we can formulate the text-to-visual loss in Eq. 6 as the multi-label soft margin loss. However, we observe that such a hybrid approach significantly compromises the performance. We suspect that the hybrid losses significantly change how the visual-and-text alignment is learned by the CLIP. It learns to build a new joint space instead of fine-tuning, as a result, leads to a huge loss on the transferable knowledge of CLIP. For this reason, we propose to modify the cross-entropy loss by holding out the tokens with the same text label.

4.2. Comparison on HICO-DET dataset

In this section, we compare our method with state-of-the-art methods on the HICO-DET dataset. We mainly compare against the existing methods which are designed to address both the seen and unseen interactions, including FG [2], VCL [15], ATL [16] and FCL [17]. Table 6 reports the experimental results of our detector and existing methods. Our method can obtain a gain of 2.37 mAP on unseen interactions and achieve a comparable result on seen interactions. It is worth noting that compared with other methods, like FCL [17] and ATL [16], we do not predetermine the unseen interactions during the training stage.

Table 7. Comparison of our proposed THID with state-of-the-art methods on SWIG-HOI dataset.

	Non-rare	Rare	Unseen	Full
JSR, ECCV20 [33]	10.01	6.10	2.34	6.08
CHOID, ICCV21 [47]	10.93	6.63	2.64	6.64
QPIC†, CVPR21 [40]	16.95	10.84	6.21	11.12
THID, Ours	**17.67**	**12.82**	**10.04**	**13.26**

4.3. Comparison on SWIG-HOI dataset

Here we compare our method with state-of-the-art methods on the SWIG-HOI dataset. We compare against the recent works JSL [33] and CHOID [47]. We also report one baseline using one-stage detector and projection as introduced in Sec 4.1.1. We train the one-stage QPIC to detect interacting humans and objects and then map them into the joint feature space of CLIP. Table 7 presents the experimental results. As shown, our proposed method can achieve a 2.14 mAP improvement on all interactions and a notable 3.83 mAP gain on unseen interactions.

4.4. Limitations

The current implementation of our visual encoder requires a fixed input resolution (*i.e.*, 224×224). This limits its ability on bounding box detection, especially for small objects. In our experiment, we observe that our detector is good at recognition, while it performs slightly worse on box localization than the other detectors which can use multi-scale training to enhance the detection quality.

5. Conclusion

In this work, we develop a transferable HOI detector via joint visual-and-text modeling. We propose a new Transformer-based visual encoder with new HOI tokens and a sequence parser module to detect multiple human interactions in the image. An automatic text formatting method is presented to ease the generation of raw text descriptions for interaction categories. Our proposed detector is capable of handling a large variety of unseen interactions. Experiments on two datasets show that it can achieve state-of-the-art performance on both seen and unseen HOI detection.

Acknowledgement This research is supported by the National Research Foundation, Singapore, under the NRF Medium Sized Centre Scheme (CARTIN). Any opinions, findings and conclusions expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] Jimmy Ba, J. Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020. 1, 2, 3, 4, 8

[3] 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, 6

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 4, 6, 7

[5] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 2

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 1

[7] 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. 4, 7

[8] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018. 2

[9] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *AAAI*, 2021. 2

[10] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018. 1

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

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 2

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv: 2104.13921*, 2021. 2, 3

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

[15] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 1, 2, 3, 4, 6, 8

[16] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 1, 2, 3, 4, 6, 8

[17] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 1, 2, 4, 8

[18] Dat Huynh and Ehsan Elhamifar. Interaction compass: Multi-label zero-shot learning of human-object interactions via spatial relations. In *ICCV*, 2021. 2

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv: 2102.05918*, 2021. 2

[20] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv: 2110.08484*, 2021. 2

[21] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, 2018. 1

[22] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 2

[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. 2

[24] Shuang Li, Yilun Du, Antonio Torralba, Josef Sivic, and Bryan Russell. Weakly supervised human-object interaction detection in video via contrastive spatiotemporal regions. In *ICCV*, 2021. 1

[25] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *NeurIPS*, 2020. 2

[26] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 2

[27] 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. 2

[28] 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 *CVPR*, 2020. 2

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

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[31] Romero Morais, Vuong Le, Svetha Venkatesh, and Truyen Tran. Learning asynchronous and sparse human-object interaction in videos. In *CVPR*, 2021. 1

[32] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019. 2

[33] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020. 7, 8

[34] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 2

[35] 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. *arXiv preprint arXiv: 2103.00020*, 2021. 2, 3, 4, 5, 6, 8

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4, 7

[37] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 1

[38] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Fei Fei Li. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 3, 4, 8

[39] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv: 2107.06383*, 2021. 2, 3

[40] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 2, 7, 8

[41] Zhigang Tu, Hongyan Li, Dejun Zhang, Justin Dauwels, Baoxin Li, and Junsong Yuan. Action-stage emphasized spatiotemporal vlad for video action recognition. *IEEE Transactions on Image Processing*, 28(6):2799–2812, 2019. 1

[42] Zhigang Tu, Wei Xie, Justin Dauwels, Baoxin Li, and Junsong Yuan. Semantic cues enhanced multimodality multistream cnn for action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1423–1437, 2019. 1

[43] 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

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[45] 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. 2

[46] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition, 2021. 2

[47] Suchen Wang, Kim-Hui Yap, Henghui Ding, Jiyan Wu, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In *ICCV*, 2021. 1, 3, 6, 8

[48] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with novel objects via zero-shot learning. In *CVPR*, 2020. 2

[49] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019. 2

[50] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 2

[51] Tete Xiao, Quanfu Fan, Dan Gutfreund, Mathew Monfort, Aude Oliva, and Bolei Zhou. Reasoning about human-object interactions through dual attention networks. In *ICCV*, 2019. 2

[52] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 2

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

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv: 2109.01134*, 2021. 2, 5

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

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

[57] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 2