Open-Category Human-Object Interaction Pre-training via Language Modeling Framework

Sipeng Zheng
Renmin University of China
zhengsipeng@ruc.edu.cn

Boshen Xu
UESTC
xuboshen.uestc@gmail.com

Qin Jin*
Renmin University of China
qjin@ruc.edu.cn

Abstract

Human-object interaction (HOI) has long been plagued by the conflict between limited supervised data and a vast number of possible interaction combinations in real life. Current methods trained from closed-set data predict HOIs as fixed-dimension logits, which restricts their scalability to open-set categories. To address this issue, we introduce OpenCat, a language modeling framework that reformulates HOI prediction as sequence generation. By converting HOI triplets into a token sequence through a serialization scheme, our model is able to exploit the open-set vocabulary of the language modeling framework to predict novel interaction classes with a high degree of freedom. In addition, inspired by the great success of vision-language pre-training, we collect a large amount of weakly-supervised data related to HOI from image-caption pairs, and devise several auxiliary proxy tasks, including soft relational matching and human-object relation prediction, to pre-train our model. Extensive experiments show that our OpenCat significantly boosts HOI performance, particularly on a broad range of rare and unseen categories.

1. Introduction

Human-object interaction (HOI) task [5, 6], whose output is usually in the format of a triplet: <human, relation, object>, has drawn increasing attention due to its crucial role in scene understanding. As humans, we have a rich vocabulary to describe one human-object relation in various ways (e.g., near, next to, close to). We can also recognize different combinations of HOI triplets in our real-life scenarios. However, current HOI methods have struggled to achieve such "**open-category**" capability for a long time. We argue that this is primarily due to two deficiencies: *inflexible prediction manner* and *insufficient supervised data*.

Previous works treat HOI learning as a classification problem where the class vocabulary must be pre-defined.

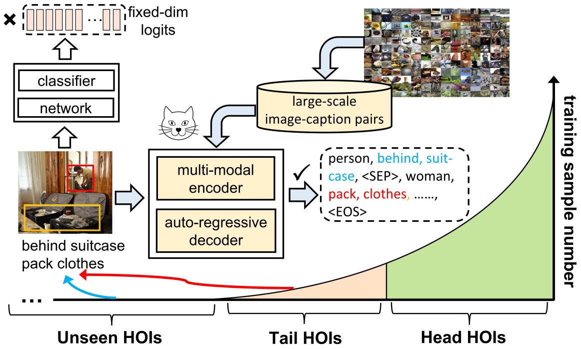

Figure 1. OpenCat reformulates HOI learning as a sequence generation task, rather than a closed-set classification task. Through the aid of task-specific pre-training with weak supervision, our model achieves open-category prediction on a large number of tail and unseen HOI classes.

This approach involves projecting the input image into fixed-dimension logits through a classifier, which restricts the ability to identify new HOI triplets. In contrast, language models [51] are more suited to predict free-form texts, thanks to their extensive token vocabulary. Recently, other works [9, 62] explore to generate visual outputs using a single language modeling objective. Inspired by this line of research, we reformulate HOI learning as a language sequence generation problem as illustrated in Figure 1, which enables our model to leverage an open-set vocabulary, generating HOI triplets with a high degree of freedom.

Moreover, HOI learning requires abundant labels for exhaustive HOI categories. However, due to the high cost of labeling grounded HOIs and the natural long-tailed distribution of HOI categories, it is unrealistic to ensure sufficient instances in each category. In fact, the two most popular benchmarks so far, HICO-DET [5] and V-COCO [21], contain 117 and 50 relation classes respectively, covering just a small portion of the HOI categories in reality. Models trained on such closed-set data fail to handle the large number of possible combinations of human, relation and object.

*Qin Jin is the corresponding author.

Recently, researchers have explored weakly supervised or even self-supervised vision-language (VL) pre-training to address data scarcity. These endeavors have achieved great success, demonstrating their generalization to novel visual or textual concepts [3, 12, 44]. Inspired by these works, one intuitive idea is to leverage pre-training to overcome the problem of insufficient labeled HOI data. However, leveraging weakly-supervised or unsupervised data for HOI pre-training is not trivial. An HOI model must accurately localize the interaction regions in the image and recognize fine-grained differences among massive human activities (e.g., stand on motorcycle vs. sit on motorcycle), which is quite challenging to learn from merely weak supervision (e.g., image-caption pairs). Therefore, the pre-training framework as well as the proxy tasks must be well designed.

In this work, to address the issues of *inflexible prediction manner* and *insufficient supervised data* in human-object interaction tasks, we propose a novel **Open-Cat**egory pre-training framework named OpenCat 🐱. Our framework utilizes a serialization scheme to convert HOI triplets into a sequence of discrete tokens and incorporates several auxiliary proxy tasks to enhance visual representation, including masked language prediction (MLP), human-object relation prediction (HRP) and human-object patch jigsaw (HPJ), all formulated as sequence generation tasks. To enable learning interaction alignment between human and object without the need for grounded HOI annotations, we further devise an additional proxy task named soft relational matching (SRM). The SRM task borrows knowledge from a VL pre-training model [34,50] to create pseudo alignment labels between detected object regions and HOI triplets parsed from the caption. With these proxy tasks, our model improves its generalization to a wide range of novel HOIs.

Our contributions can be outlined as follows:

- We introduce OpenCat, a language modeling framework to effectively model open-category HOIs.
- We collect a large amount of weakly-supervised HOI pre-training data based sorely on textual supervision and devise several proxy tasks to train our model.
- By adapting our model to downstream HOI tasks, we achieve state-of-the-art performance with larger gains observed under zero-shot and few-shot setups.

2. Related Work

Human-Object Interaction Learning. Recognizing human-object interactions (HOI) [18, 20, 45] has been widely studied in recent years. The main challenge of this task is the co-occurrence of multiple human-object pairs in an image where their location is not given. To address it, existing works rely on object [20] or human pose detection [14], or even bodypart-level annotations [36]. In contrast to image-level HOI recognition, instance-level HOI detection [5] aims to accurately localize interactive regions for each human-object pair and predict their interaction class simultaneously. Existing methods can be broadly categorized into two-stage and one-stage models. Two-stage methods [19,49,61] first use an off-the-shelf object detector [53] to ground objects regions offline and then perform HOI prediction. These works primarily focus on second stage to improve human and object embeddings using graph neural networks [49,61] or external information like keypoints [60, 66]. However, two-stage methods do not consider the possibility of combining human and object to form a valid HOI instance, leading to overwhelming negative HOI proposals. One-stage methods [29,38], on the other hand, perform object detection and HOI prediction in parallel to generate HOI proposals with high quality. For example, Liao *et al*. [38] propose a point-based network to heuristically define the position of HOI triplets. Tamura *et al*. [57] go a further step to achieve end-to-end HOI detection based on a DETR-style transformer [4]. Recently, Zhang *et al*. [63] adopt a cascade disentangling decoder to combine the advantages of two-stage and one-stage methods.

Vision-Language Pre-training. Vision-language pre-training (VLP) models [10,43] generally follow two steps: first, using well-designed proxy tasks [13,51] to train models on a vast amounts of data, and then fine-tuning parameters on downstream tasks. These models demonstrate impressive results particularly on zero-shot setups. Previous VLP works haved used task-specific heads [10, 43] with separate parameters for downstream tasks, but recent studies [12, 22, 52] propose a unified framework to perform different downstream task predictions with the same input-output format. Inspired by [9], UniTAB [62] learns to represent both text and box outputs as a discrete token sequence using a language-modeling objective, which enables the model to ground generated texts to object regions and provide a more interpretable description for the image. This language-modeling pre-training paradigm is naturally suitable for open-category HOI learning, as it allows the model to avoid closed-set restriction and leverage the rich semantic knowledge of text supervision to generate diverse HOI combinations. Despite its potential, such paradigm remains unexplored by previous relational pre-training models [11,67]. In this study, we aim to reformulate the HOI model into a language modeling framework that accepts structured inputs (i.e., raw image and language), and auto-regressively outputs HOI triplets in a sequence of tokens.

3. Method

3.1. Overall Architecture

Figure 2 illustrates the overall architecture of our proposed OpenCat model, which follows a visual-language encoder-decoder framwork [59]. We choose ResNet-

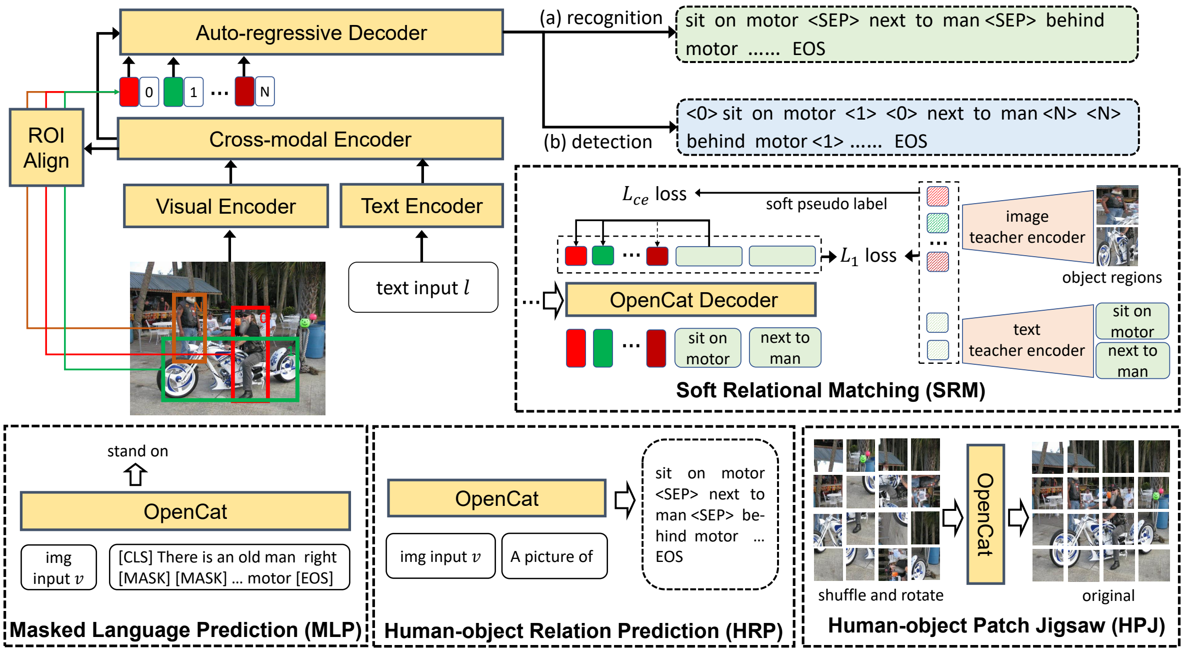
Auto-regressive Decoder
(a) recognition
sit on motor <SEP> next to man <SEP> behind motor ...... EOS
(b) detection
<0> sit on motor <1> <0> next to man <N> <N> behind motor <1> ...... EOS
0
1
N
ROI Align
Cross-modal Encoder
Visual Encoder
Text Encoder
text input l
L_{ce} loss
soft pseudo label
L_1 loss
image teacher encoder
object regions
OpenCat Decoder
sit on motor
next to man
text teacher encoder
sit on motor
next to man
Soft Relational Matching (SRM)
stand on
OpenCat
img input v
[CLS] There is an old man right [MASK] [MASK] ... motor [EOS]
Masked Language Prediction (MLP)
OpenCat
img input v
A picture of
sit on motor <SEP> next to man <SEP> behind motor ... EOS
Human-object Relation Prediction (HRP)
OpenCat
shuffle and rotate
original
Human-object Patch Jigsaw (HPJ)

Figure 2. Our proposed open-category HOI model, OpenCat, auto-regressively generates HOI triplets as a token sequence. To pretrain the model, we utilize large-scale weak text supervision and employ four auxiliary proxy tasks: (1) masked language prediction (MLP); (2) human-object relation prediction (HRP); (3) human-object patch jigsaw (HPJ) and (4) soft relational matching (SRM).

101 [24] and RoBERTa [40] to encode image input v and text input l respectively. A 6-layer transformer encoder is used for cross-modal encoding, followed by another 6-layer transformer decoder for token sequence generation. For each image, an off-the-shelf object detector [53] is used to localize N object regions $\mathcal{B} = \{b_1, b_2, ..., b_N\}$ offline. Region embeddings $\mathcal{R}$ of $\mathcal{B}$ are then cropped and pooled from the image based on ROI align [23]. Similar to language modeling methods [51, 52], our model auto-regressively outputs HOI triplets as a token sequence conditioned on the region embeddings. As in language modeling [51], our model is optimized using a maximum likelihood objective:

$$\mathcal{L}_{LM}(\Theta) = -\sum_{j=1}^{L} \log P_{\Theta}(y_j | v, l, \mathcal{R}, \hat{y}_{1:j-1}) \tag{1}$$

where y and $\hat{y}$ represent the target and input sequence, while Θ denotes the model parameters and L is the length of the target sequence. The region embeddings $\mathcal{R}$ provide box-level prior information during target sequence decoding. By treating HOI learning as a sequence generation task, our model utilizes the semantic knowledge of VL pretraining methods to generate relation phrases between humans and objects (e.g., inferring "read" conditioned on "a man ? book"), Additoinally, the model can predict new HOI categories with a free format.

In the following sections, we explain the learning paradigm of our proposed OpenCat in detail. In Sec 3.2, we outline the formulation of fundamental HOI tasks as sequence generation tasks, encompassing HOI recognition and HOI detection. Sec 3.3 describes how prior knowledge and weakly-supervised data can be utilized to pre-train the model through various proxy tasks.

3.2. HOI Learning via Sequence Generation

In order to perform HOI learning, it is necessary to predict all HOI triplets $\phi = \{(h_1, r_1, o_1), (h_2, r_2, o_2), ...\}$ in an image, where h_i, r_i, o_i refer to the categories of human, relation and object, respectively. Although a category label may contain more than one token, we simplify our notation by using h, r, and o to denote the tokenized category label.

HOI learning comprises two fundamental sub-tasks: image-level HOI recognition and instance-level HOI detection. The former involves predicting all HOI categories in an image, while the latter aims to recognize and localize all HOI triplets, making it more challenging than image-level HOI recognition. OpenCat reformulates all HOI learning tasks as sequence generation tasks. In the following section, we provide a detailed description of these tasks, with a focus on how their training target sequences are constructed.

Image-level HOI Recognition. Our model recognizes all HOI categories in an image and produces a target sequence in the format of $[r_1, o_1, SEP, r_2, o_2, SEP, ..., EOS]$. Each (r, o) pair represents an HOI category, and the special token

$[SEP]$ denotes the separation between different pairs. This is necessary to reduce the difficulty of sequence generation, as the sequence length of (r, o) may vary. The generation process stops after the $[EOS]$ token is predicted. During training, we shuffle the HOI triplets of an image in the target sequence at each step, as they are unordered. During inference, our model predicts all triplets directly. Unlike classifying HOI into fixed-dimension logits, our model is able to predict free-form HOI categories.

Instance-level HOI Detection. The aim of this task is to detect bounding boxes for each HOI instance (h, r, o) in the image while also predicting their categories. However, it is challenging to strike a good balance between human-object detection and interaction prediction in multi-task training, as both tasks are difficult to accomplish [63]. To tackle this issue, we take a different approach from current one-stage HOI methods [30, 57], by decoupling the two tasks from a unified framework through offline object detection. Such decoupling also provides an additional advantage, as we can leverage newly proposed detectors [65] to obtain more accurate object locations and better visual representations.

Conditioned on the detected object regions, our model detects HOI instances as serialized tokens. To be specific, each instance is represented as several discrete tokens such as $[p_h, p_o, r, o]$, where p_h and p_o are two pointer tokens between [0,N−1], indicating the indices of the human and object bounding boxes that have been detected. As a result, the target sequence for HOI detection can be expressed as $[p_{h_1}, p_{o_1}, r_1, o_1, p_{h_2}, p_{o_2}, r_2, o_2, ..., EOS]$. It's worth noting that (p_{h_i}, p_{o_i}) and (p_{h_j}, p_{o_j}), where $i \neq j$, may point to the same (h, o) pair, since relations between human and object can be multi-label. Unlike traditional two-stage HOI methods [19, 49] that suffer from an overwhelming number of negative HOI proposals based only on local region features, our model avoids such interference by setting a maximum target sequence length to limit the proposal number.

Note that our open-category prediction framework may generate triplets that are synonymous with groudtruth HOIs (e.g., the prediction is "stand above bicycle" while the groundtruth is "stand on bicycle"). We thus use WordNet [46] to match the possible synonymous triplets with groundtruth. Details about this process are presented in the supplementary material.

3.3. HOI Pre-training via Proxy Tasks

As described in above section, we unify all HOI learning tasks as sequence generation tasks, which allows for the prediction of new HOI categories in a free-form manner, and also facilitates the more effective utilization of weakly-supervised data for HOI learning. To further enhance model generalization, we design several auxiliary proxy tasks to assist weakly-supervised pre-training :

Masked Language Prediction (MLP). To obtain the HOI triplets for each image, we first employ a rule-based language parser [27] to parse its corresponding image caption. The resulting HOI triplets, denoted as $\phi = \{(h_1, r_1, o_1), (h_2, r_2, o_2), ...\}$, are then used in the MLP task. Specifically, this task randomly selects a subset of HOI triplets from ϕ, and masks either the relation or object tokens of the selected triplets using a special token $[MASK]$. The masked tokens formulate the target sequence y. The aim of this task is to predict the masked text spans related to HOI, based on the visual-textual contexts available.

Human-object Relation Prediction (HRP). In this task, the model is presented with an image and a textual prompt (e.g., "a picture of") and is required to generate all possible HOI categories in the image using an auto-regressive manner. The format of the predicted sequence is $[r_1, o_1, SEP, r_2, o_2, ..., EOS]$. To increase the diversity of the generated sequences, we also augment the target sequence by randomly shuffling the order of HOI classes.

Human-object Patch Jigsaw (HPJ). Drawing inspiration from the jigsaw puzzle solving task [47], which helps the model recognize the key parts of an object, we propose the human-object patch jigsaw (HPJ) task. Given an image-caption pair as input, we slice the image into $H \times W$ patches and randomly select a human-object pair (h, o) from the detected object regions. Assuming the human-object pair contains K image patches, we shuffle the order of these patches and rotate them by an angle $k \in \{0°, 90°, 180°, 270°\}$, which means a rotated patch with k requires 360°-k clockwise rotation to restore. The target sequence of the HPJ task can be denoted as $[y_1^s, y_1^r, y_2^s, y_2^r, ..., y_K^s, y_K^r, EOS]$, where $y_i^s \in$ [0,HW−1] indicates the original location in the image for the i-th region patch, and $y_i^r \in \{0, 1, 2, 3\}$ denotes the restoring angle type. The HPJ task enables our model to explore the relative relationships between distinguishable local information within the human-object pair, leading to a better understanding of potential interactions between them.

Soft Relational Matching (SRM). To learn the alignment between humans and objects in a weakly-supervised manner for HOI detection, we propose the soft relational matching task. In this task, our model outputs a token sequence for each image, similar to the instance-level HOI detection paradigm described in Sec 3.2. However, since discrete box indices for pointer tokens p_h, p_o are not available, we create soft pseudo labels by distilling knowledge from an image-text teacher encoder $\mathcal{V}$ and $\mathcal{T}$ [34, 50] pre-trained on billions of image-caption pairs. Specifically, we use an image teacher encoder $\mathcal{V}$ to extract object region embeddings $\mathcal{V}(\mathcal{B})$ offline and obtain text embeddings $\mathcal{T}(\phi)$ for HOI triplets ϕ. For each $\phi_i = (h_i, r_i, o_i)$ in ϕ, the soft matching label between ϕ_i and object regions $\mathcal{B}$ is computed as the cosine similarity $z(h_i) = cos(\mathcal{V}(\mathcal{B}), \mathcal{T}(h_i))$ and $z(o_i) = cos(\mathcal{V}(\mathcal{B}), \mathcal{T}(o_i))$. Then we apply softmax activation to compute the cross entropy loss:

Table 1. Our pre-training data vs. fully-supervised HOI datasets. $*$ denotes the object boxes are provided by an off-the-shelf object detector. “grounded HOI” denotes exact alignment annotations between human and object in the image.

Dataset	image caption	object bbox	grounded HOI	clean HOI label	#image	#HOI triplet	#bbox	#object	#relation
MPII [1]	×	$*$	×	✓	15,205	-	-	-	393
HICO [6]	×	$*$	×	✓	38,118	-	-	80	117
V-COCO [21]	×	✓	✓	✓	5,400	24,331	50,759	80	26
HICO-DET [5]	×	✓	✓	✓	38,118	70,373	199,733	80	117
ours	✓	$*$	×	×	754,001	1,818,071	7,540,010	9731	2516

$$\mathcal{L}_{\phi_i} = \mathcal{L}_{CE}(z(h_i)/\tau, p_{h_i}) + \mathcal{L}_{CE}(z(h_i)/\tau, p_{o_i}) \tag{2}$$

where τ is a scaling temperature. Moreover, we treat our model Θ as the student model and apply an $\mathcal{L}1$ loss to minimize the distance between the student and teacher output embeddings: $\mathcal{L}1 = ||\Theta(B), \mathcal{V}(B)||_1 + ||\Theta(\phi), \mathcal{T}(\phi)||_1$. The final objective for SMR is calculated as:

$$\mathcal{L}_{srm} = \mathcal{L}_{LM} + \gamma_1 \mathcal{L}_{\phi} + \gamma_2 \mathcal{L}_1 \tag{3}$$

where $\mathcal{L}_{LM}$ denotes the language modeling loss for SRM task, γ_1 and γ_2 are re-weighting hyperparameters.

4. Experiments

To demonstrate the open-category and generalization capabilities of HOI models, we carry out experiments on both HOI recognition and HOI detection tasks using various evaluation settings, including many-shot, few-shot, zero-shot, and weakly-supervised settings.

4.1. Experiment Setup

Pre-training Data. We utilize multiple vision-language datasets to construct our pre-training data, including Flickr-30K [48], MS-COCO [39], Visual Genome [31], Open-Image [33] and ConceptCaption [55]. While these datasets contain diverse structured annotations, we solely employ their text supervision. Similar to [67], we adopt a rule-based language parser [27] proposed by Schuster *et al.* [54] to parse HOI triplets from image captions. After parsing, we obtain triplets with lemmatized words of the subject, relation, and object. We only retain triplets with ”person” as the subject synset while manually removing triplets with obvious typos. In total, we collect over 750K images, including 1.8M HOI triplets that encompass over 2.5K unique relations and 9.7K object categories. We use Faster-RCNN [53] trained on the 600-category Open-Image [33] as the object detector to provide object region candidates for each pre-training image. As shown in Table. 1, our weakly-supervised pre-training data outperforms existing fully-supervised HOI datasets [5,21] in terms of scale, with at least **20**× the size of the image set, **25**× the number of HOI triplets, and **20**× the number of relation categories.

By transferring language knowledge from such diverse text concepts, our model provides prior semantic knowledge for open-set classes (e.g., grab apple, carve sculpture). Moreover, the model can learn new compositions to predict novel HOI triplets based on seen verbs and objects (e.g. leverage “ride horse” & ”feed donkey” for “ride donkey”).

Downstream Datasets. We validate our model on four datasets: **(1) HICO** [6], which comprises 47,776 images with 600 HOI categories, 80 object categories, and 117 unique relations. Each image may contain multiple HOI categories; **(2) MPII** [1], which has 15,205 training images and 5,708 testing images. Unlike HICO, each image contains only one of 393 interaction classes; **(3) HICO-DET** [5], which is an extension of HICO and includes bounding box annotations for HOI detection; **(4) V-COCO** [21], which is created from the MS-COCO dataset and contains 10,346 images with 26 unique interaction categories. We conduct image-level HOI recognition experiments on HICO and MPII, and instance-level HOI detection experiments on HICO-DET and V-COCO.

Implementation Details. Our cross-modal encoder and decoder consist of 6 transformer layers each, with 8 attention heads and a hidden dimension of 256 in every layer. We adopt the same scale and crop augmentation in DETR [4] such that the longest image side is smaller than 1333 pixels while the shortest side falls between 480 and 800. During pre-training, we use the AdamW optimizer [42] and exponential moving average [58] with a decay rate of 0.9998. We pre-train the model for 40 epochs using a batch size of 32. The learning rate is initialized to 1e-4 and 1e-5 for transformer layers and backbones, and is initialized to 3e-5 and 1e-5 when fine-tuning on downstream tasks. For each image, we detect N=100 object regions offline. More details are provided in the supplementary material.

4.2. HOI Recognition

Mean average precision (mAP) is adopted as the evaluation metric for HOI recognition. Figure 3 presents the full-set results on HICO and MPII. We compare our model with different methods, such as Mallya *et al.* [45], Pairwise [14],

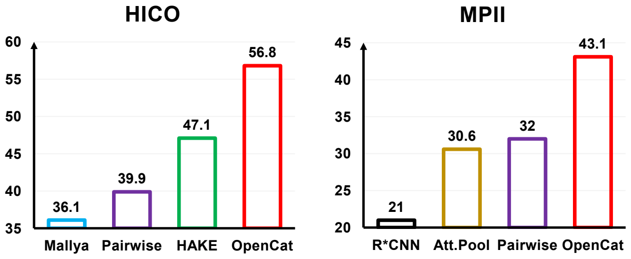

Figure 3. Comparison of many-shot HOI recognition on HICO and MPII with full-set datasets.

Table 2. Comparison of few-shot HOI recognition on HICO, where † denotes using external fine-grained bodypart-level labels.

Method	Few@1	Few@5	Few@10
Pairwise [14]	13.20	19.79	22.78
HAKE† [35]	25.40	32.48	33.71
OpenCat	**37.52**	**42.45**	**44.17**

HAKE† [35] on HICO, and R*CNN [20], Att.Pool [18], Pairwise [14] on MPII. Our model outperforms other methods by at least +9.7% mAP over HAKE† on HICO, and +11.1% mAP over Pairwise on MPII. It's important to note that the improvement isn't solely attributed to the availability of additional data. Despite using only low-cost pretraining data, OpenCat still surpases another pre-trained model HAKE†, which relies on bodypart-level annotations.

To validate our model's effectiveness under the few-shot setting, we present the results on HICO in Table 2. Few@i denotes the mAP metric under few-shot circumstances, with the number of training samples less than i. When i equals to 1, it represents the one-shot problem. Without external fine-grained labels, our model achieves significant improvement with +12.12% mAP on 1-shot, +9.97% mAP on 5-shot and +10.46% mAP on 10-shot compared to HAKE†. These results indicate that our model can effectively handle the long-tailed distribution problem in HOI recognition, and can well adapt to rare interaction classes.

4.3. HOI Detection

For HOI detection, we adopt the same evaluation protocol in [5] and use mAP as the evaluation metric. An HOI prediction is considered as a true positive only if the Interaction of Union (IoU) of human and object bounding boxes is equal to or greater than 0.5, and the interaction label prediction is correct as well. We carry out experiments on two datasets: (1) HICO-DET, which is divided into three HOI category subsets [5]: the *Full* set includes all 600 HOI categories; the *Non-rare* subset consists of 462 categories with 10 or more training samples per category; the *Rare* subset consists of the remaining 138 categories with less than 10 training samples per category. (2) V-COCO, which includes two scenarios. Scenario 1 is required to detect HOI pairs even if there is occlusion between them, while Scenario 2 does not require the detection of occluded HOI pairs.

Table 3. Comparison of HOI detection on HICO-DET and V-COCO. "S1" and "S2" denotes Scenario 1 and Scenario 2.

	Method	Backbone	Full	Rare	Non-rare
HICO-DET	PPDM [38]	Hourglass-104	21.73	13.78	24.10
	DRG [16]	ResNet-50-FPN	24.53	19.47	26.04
	AS-Net [8]	ResNet-50	28.87	24.25	30.25
	QAHOI [7]	Swin-B	29.47	22.24	31.63
	QPIC [57]	ResNet-101	29.90	23.92	31.69
	CDN [63]	ResNet-101	32.07	27.19	33.53
	OpenCat	ResNet-101	**32.68**	**28.42**	**33.75**

	Method	Backbone	S1	S2
V-COCO	TIN [37]	ResNet-50	47.8	-
	AS-Net [8]	ResNet-50	53.9	-
	HOTR [30]	ResNet-50	55.2	64.4
	DIRV [15]	EfficientDet-d3	56.1	
	QPIC [57]	ResNet-101	58.8	61.0
	CDN [63]	ResNet-101	61.7	63.8
	OpenCat	ResNet-101	**61.9**	63.2

Table 4. Comparison of zero-shot HOI detection on HICO-DET. UC, UO, UR, UA denote unseen combination, unseen object, unseen relation and unseen all scenarios, "rare first" and "non rare first" are two HOI class splits provided by [26]

Method	Type	Unseen	Seen	Full
VCL [25]	UC (rare first)	10.06	24.28	21.43
FCL [26]	UC (rare first)	13.16	24.23	22.01
OpenCat	UC (rare first)	**21.46**	**33.86**	**31.38**
VCL [25]	UC (non rare first)	16.22	18.52	18.06
FCL [26]	UC (non rare first)	18.66	19.55	19.37
OpenCat	UC (non rare first)	**23.25**	**28.04**	**27.08**
FG [2]	UO	11.22	14.36	13.84
ConsNet [41]	UO	13.51	14.67	14.48
FCL [26]	UO	15.54	20.74	19.87
OpenCat	UO	**23.84**	**28.49**	**27.72**
ConsNet [41]	UR	12.50	14.72	14.35
OpenCat	UR	**19.48**	**29.02**	**27.43**
OpenCat	UA	15.82	-	-

Table 3 presents the comparison of HOI detection results on HICO-DET and V-COCO. Our model delivers consistent improvement when compared to previous works on the HICO-DET *Full* set. Furthermore, it achieves competitive performance on V-COCO compared with state-of-the-art CDN [63]. Notably, our model achieves 28.42 mAP on the HICO-DET *Rare* set, a larger improvement than that on the *Full* set, which again demonstrates our model's capacity to adapt to rare classes caused by long-tailed distribution.

4.3.1 Zero-shot HOI Detection

The concept of detecting zero-shot interactions, where there is no corresponding image available during training, is first

Table 5. Comparison of weakly-supervised HOI detection on HICO-DET and V-COCO, where “S1” and “S2” denotes Scenario 1 and Scenario 2.

Method	HICO-DET			V-COCO	
	Full	Rare	Non-rare	S1	S2
PPR-FCN [64]	15.14	10.65	16.48	-	-
MX-HOI [32]	16.14	12.06	17.50	-	-
Align-Former [28]	20.85	18.23	21.64	15.8	16.3
OpenCat	**25.82**	**24.35**	**26.19**	**34.4**	**36.1**

introduced by Shen *et al*. [56]. In our experiments, we examine four zero-shot scenarios on HICO-DET [5]:

Unseen Combination (UC). Following [2, 56], we select 120 HOI triplets in HICO-DET as unseen testing set and use the remaining 480 triplets for training. We ensure each relation or object category in unseen set appears at least once in the 480 triplets. Experiments are carried out on two class splits provided by [25]: “rare first”, which prefers selecting the 120 unseen triplets from the tail, and “non rare first”, which prefers selecting the unseen HOIs from the head.

Unseen Object (UO). We choose 12 object categories as unseen objects following [26] and identify 100 HOI triplets containing these objects as unseen HOIs. The remaining 500 HOI triplets are seen during training.

Unseen Relation (UR). Following [41], we select 22 relations out of the 117 relation categories in HICO-DET as unseen. We remove all training samples containing these relations for unseen testing.

Unseen All (UA). No supervision is provided. We directly utilize the pre-trained model to detect HOIs without any fine-tuning on the downstream dataset.

Table 4 showcases the zero-shot HOI detection performance. Our model significantly outperforms previous methods by a large margin in UC, UO, and UR scenarios, demonstrating its ability to detect unseen objects, relations, and their combinations. Our model also performs well in the UA scenario without requiring additional modules, even when all annotations, including bounding box and HOI alignment labels, are previously unseen. To achieve this, our model first generates a sequence of HOI categories similar to the pre-training procedure of SMR. Then, the model matches these predicted HOIs with nearest object regions detected offline based on their embedding similarity. Despite not undergoing fine-tuning, our model still achieves impressive results, reaching 15.82 mAP on HICO-DET.

4.3.2 Weakly-supervised HOI Detection

To further evaluate the robustness of our model, we turn to weakly-supervised HOI detection, which aims to identify HOI triplets without the aid of alignment labels between the human and object in the image. Similar to the UA scenario mentioned above, our model uses a nearest matching mechanism to ground object regions of HOI. The evaluation setting is the same as that for full-set HOI detection.

Table 6. Ablation of different proxy tasks on HICO, where “1, 2, 3, 4” denotes the task ids of MLP, HRP, HPJ and SRM respectively.

Task IDs	HICO	HICO-DET		
	mAP	Full	Rare	Non-rare
w/o pretrain	43.5	21.25	18.47	21.95
1	45.3	22.13	19.57	22.77
1+2	51.7	24.85	22.28	25.49
1+2+3	53.1	26.12	23.84	26.69
1+2+3+4	56.8	32.68	28.42	33.75

Table 7. Ablation of different teacher encoders of SRM on HICO.

Task IDs	HICO	HICO-DET		
	mAP	Full	Rare	Non-rare
w/o SRM	53.1	26.12	23.84	26.69
CLIP w/ Res50 [50]	54.9	29.08	25.26	30.04
CLIP w/ ViT-B/16 [50]	55.7	30.45	26.74	31.38
BLIP w/ ViT-B/16 [34]	56.8	32.68	28.42	33.75

We present the comparison results with previous works in Table 5. Our model outperforms Align-Former [28] with +4.97%, achieving 25.82 mAP on the HICO-DET *Full* set. The +6.12% improvement on the *Rare* set is even more significant, demonstrating OpenCat is highly robust to rare classes with limited samples, even under weakly-supervised training. We achieve larger gains on V-COCO, with +18.6% in Scenario 1 and +19.8% in Scenario 2. These results underscore the advantage of our model as weak supervision is much less costly than HOI alignment labels, enabling us to scale training to a larger number of relations and objects.

4.4. Ablation Study

Ablation of Proxy Tasks. We first validate each proxy task via a thorough ablation study in Table 6. The HRP task brings the largest gain of +6.4% for HOI recognition. This can be attributed to the fact that the HRP task has the most similar target sequence format to the downstream recognition task. In addition, a large number of HOI combinations during pre-training enhance the recognition of HOIs especially those hard cases. In terms of HOI detection, the SRM task makes the largest contribution, with +6.56%, +4.58% and +7.06% mAP improvement on HICO-DET *Full*, *Rare* and *Non-Rare* sets, respectively. It indicates that the SRM task can guide our model to reason about the alignment between human and objects, even without grounded HOI labels. Besides, other proxy tasks also result in stable improvements across different benchmarks.

Ablation of Teacher Encoder. The quality of knowledge we borrow in the SRM task depends largely on the

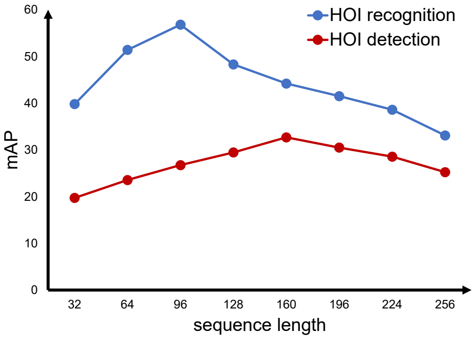

Figure 4. Ablation of sequence length for performance on HICO.

Table 8. Ablation of different sequence design on HICO. $[SEP]$ and $[PTR]$ denote separation and pointer tokens, respectively.

Sequence Design	HICO mAP	HICO-DET Full	 Rare	 Non-rare
OpenCat	56.8	32.68	28.42	33.75
w/o $[SEP]$	49.5	-	-	-
w/o $[PTR]$	-	30.32	25.84	31.44

teacher encoder we choose, making it critical in the SRM task. We compare the impact of different teacher encoders in Table 7. By comparing CLIP [50] with ResNet50 and ViT-B/16, a more powerful backbone performs better on the HICO-DET *Full* set, and BLIP [34] achieves even better performance. We suggest that BLIP, which is based more on in-domain data such as MS-COCO and ConceptCaption, has a smaller gap with our HOI pre-training data compared to other teacher encoders.

Ablation of Sequence Length. Figure 4 shows the impact of different sequence length on performane. On HICO, a sequence length of 96 is found to be sufficient for HOI recognition, while the best performance is achieved with a sequence length of 160 for HOI detection. Note that the optimal maximum sequence length varies depending on the downstream tasks and datasets. For example, since each image in MPII only contains one interaction class, we can use much shorter sequence length for MPII dataset. In contrast, for HOI detection, where there could be over 30 HOI instances in an image, longer sequence lengths might be necessary to capture all the relevant information.

Ablation of Sequence Design. We analyze the impact of sequence design in Table 8. The results show that both separation token $[SEP]$ and pointer token $[PTR]$ are effective in reducing the difficulty of sequence generation. This indicates the necessity of these tokens in our model design.

5. Qualitative Analysis

Figure 5 provides some qualitative examples of rare category prediction. We use iCAN [17] as the baseline. OpenCat is able to predict some challenging HOI cases, such as “milk cow” or “wash dog”, which are often misclassified as “hold cow” or “hold dog” by the baseline due to the dominant number of “hold” samples in the dataset. These examples demonstrate that OpenCat can effectively deal with the long-tailed distribution problem in HOI datasets.

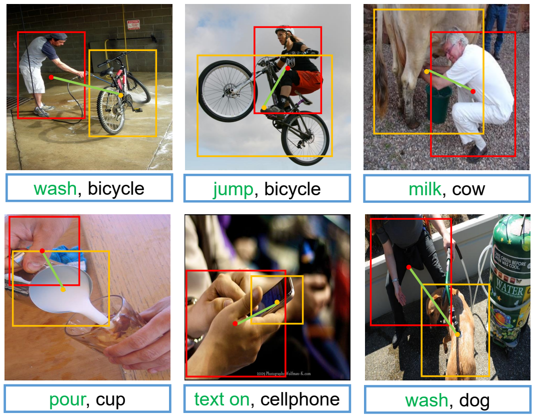

Figure 5. Qualitative examples of rare HOI category prediction, where verbs in green are rare categories that our OpenCat is able to generate while the baseline can not.

6. Conclusion

In this paper, we propose OpenCat, an open-category pre-training model for human-object interaction. OpenCat adopts a language modeling framework, treating HOI learning as a sequence generation task to overcome the constraints of closed-set prediction for novel HOI categories. We leverage massive amounts of weakly-supervised data and propose several proxy tasks for HOI pre-training. As a result, our model achieves state-of-the-art performance on HOI tasks across various benchmarks, with significant improvements observed on rare and novel categories.

Limitations. Our model auto-regressively generates HOI triplets, which may lead to a decrease in inference efficiency due to the long target sequence length. Furthermore, the limitation on sequence length presents challenges when attempting to merge object detection and HOI prediction into a unified framework. We intend to address these challenges in our future work and explore this direction further.

7. Acknowledgments

This work was partially supported by National Natural Science Foundation of China (No. 62072462), National Key R&D Program of China (No. 2020AAA0108600).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 5

[2] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10460–10469, 2020. 6, 7

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

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

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 1, 2, 5, 6, 7

[6] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015. 1, 5

[7] Junwen Chen and Keiji Yanai. Qahoi: Query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021. 6

[8] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. 6

[9] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 1, 2

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[11] Meng-Jiun Chiou, Roger Zimmermann, and Jiashi Feng. Visual relationship detection with visual-linguistic knowledge from multimodal representations. *IEEE Access*, 9:50441–50451, 2021. 2

[12] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 2

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

[14] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 51–67, 2018. 2, 5, 6

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

[16] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. 6

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

[18] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. *Advances in neural information processing systems*, 30, 2017. 2, 6

[19] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 2, 4

[20] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015. 2, 6

[21] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 5

[22] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems. *arXiv preprint arXiv:2104.00743*, 2021. 2

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

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

[26] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 6, 7

[27] Mao Jiayuan, Kasai Seito, and Scene graph parser. Scene graph parser. `https://github.com/vacancy/SceneGraphParser`, 2018. 4, 5

[28] Mert Kilickaya and Arnold Smeulders. Human-object interaction detection via weak supervision. *arXiv preprint arXiv:2112.00492*, 2021. 7

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

[30] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. 4, 6

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 5

[32] Suresh Kirthi Kumaraswamy, Miaojing Shi, and Ewa Kijak. Detecting human-object interaction with mixed supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1228–1237, 2021. 7

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 5

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 2, 4, 7, 8

[35] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019. 6

[36] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. 2

[37] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. 6

[38] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 2, 6

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

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[41] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4235–4243, 2020. 6, 7

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

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[44] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[45] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *European Conference on Computer Vision*, pages 414–428. Springer, 2016. 2, 5

[46] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4

[47] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 4

[48] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 5

[49] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018. 2, 4

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 4, 7, 8

[51] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI*, 2018. 1, 2, 3

[52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 2, 3

[53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 3, 5

[54] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 5

[55] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[56] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1568–1576. IEEE, 2018. 7

[57] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10410–10419, 2021. 2, 4, 6

[58] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 5

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[60] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 2

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

[62] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*. ECCV, 2022. 1, 2

[63] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 4, 6

[64] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE international conference on computer vision*, pages 4233–4241, 2017. 7

[65] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 4

[66] Sipeng Zheng, Shizhe Chen, and Qin Jin. Skeleton-based interactive graph network for human object interaction detection. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 2

[67] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1834, 2021. 2, 5