Disentangled Pre-training for Human-Object Interaction Detection

Zhuolong Li[1*]   Xingao Li[1*]   Changxing Ding[1,2,3†]   Xiangmin Xu[2]

[1]School of Electronic and Information Engineering, South China University of Technology
[2]School of Future Technology, South China University of Technology   [3]Pazhou Lab, Guangzhou

{eezhuolong, eexingao}@mail.scut.edu.cn, {chxding, xmxu}@scut.edu.cn

Abstract

Detecting human-object interaction (HOI) has long been limited by the amount of supervised data available. Recent approaches address this issue by pre-training according to pseudo-labels, which align object regions with HOI triplets parsed from image captions. However, pseudo-labeling is tricky and noisy, making HOI pre-training a complex process. Therefore, we propose an efficient disentangled pre-training method for HOI detection (DP-HOI) to address this problem. First, DP-HOI utilizes object detection and action recognition datasets to pre-train the detection and interaction decoder layers, respectively. Then, we arrange these decoder layers so that the pre-training architecture is consistent with the downstream HOI detection task. This facilitates efficient knowledge transfer. Specifically, the detection decoder identifies reliable human instances in each action recognition dataset image, generates one corresponding query, and feeds it into the interaction decoder for verb classification. Next, we combine the human instance verb predictions in the same image and impose image-level supervision. The DP-HOI structure can be easily adapted to the HOI detection task, enabling effective model parameter initialization. Therefore, it significantly enhances the performance of existing HOI detection models on a broad range of rare categories. The code and pre-trained weight are available at https://github.com/xingaoli/DP-HOI.

1. Introduction

Human-Object Interaction (HOI) detection involves simultaneous object detection and verb classification for every interactive human-object pair in an image. It is a fundamental scene and action understanding task with various potential applications in robotics [1], image captioning [2, 3], image retrieval [4, 8], and visual question answering [5]. The labeling costs for HOI detection datasets are higher than those for image classification and object detection due to the in-

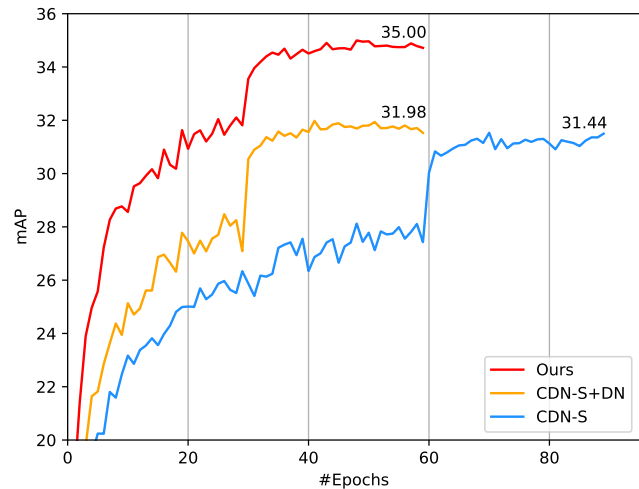*The first two authors contribute equally.
†Corresponding author.

Figure 1. CDN-S [44] mAP and convergence curves with the pretrained DETR weights [10] on MS-COCO [9] and our DP-HOI, respectively. DN denotes the denoising strategy [11] is adopted to speed up convergence. Experiments are conducted on the HICO-DET dataset [6].

clusion of all meaningful $\langle human, verb, object \rangle$ triplets in an image. Therefore, existing HOI detection datasets are usually small, considerably affecting HOI detection performance.

Current HOI detection models are usually based on detection transformer (DETR) [10]. Since the DETR training is data hungry, most existing works [40, 41, 47, 52, 57, 60, 63] initialize their model according to pre-trained DETR weights on object detection datasets (e.g., MS-COCO [9]). This strategy is sub-optimal for HOI detection because the pre-trained DETR model does not contain any action knowledge. As a result, recent studies [71, 72] have adopted large-scale pseudo-labeled scene graph data for HOI model pre-training, which has shown significant potential.

However, the scene graph data pseudo-labeling process is complex and the obtained pseudo-labels are error-prone. In this paper, we observed that HOI detection can be decomposed into two sub-tasks: interactive human-object pair

detection and interaction classification. These sub-tasks are closely related to the object detection and action recognition tasks, respectively. The labeling of both object detection and action recognition tasks is easier; therefore, they both own large-scale labeled datasets (e.g., Objects365 [83] and Kinetics-700 [84]). Based on these observations, we propose utilizing these labeled datasets for pre-training HOI detection models. However, these datasets are partially labeled (i.e., only objects or actions are labeled). Thus, they cannot be directly utilized for training according to standard HOI detection architecture. Therefore, a tailored pre-training architecture that is as close as possible to that of the downstream HOI detection task is required for efficient knowledge transfer.

In this study, we propose the disentangled pre-training method for human-object interaction (DP-HOI). DP-HOI conducts object detection and verb classification using two parallel branches. The first branch contains a detection decoder trained with object detection datasets, according to the standard DETR structure and training strategy [10, 11]. The second branch is trained using readily available action recognition datasets.

Moreover, we design the verb classification branch to mimic popular HOI detection structures [44, 47, 64]. This branch contains a detection and interaction decoder. The detection decoder shares parameters with the object detection branch and identifies all human instances in each training image from the action recognition datasets. Then, we adopt each human instance's output decoder embedding as a reliable person query (RPQ) for the interaction decoder. Each RPQ is responsible to search for the action cues of the specified person and predict the person's action. Since there are only image-level action labels and there might be several RPQs for each image, we introduce a verb-wise prediction fusion (VPF) strategy to merge the RPQ prediction results and impose supervision. In addition, we extend our approach to video and image captioning data, which contain significant action categories vital for pre-training purposes.

Furthermore, we demonstrate DP-HOI's effectiveness through comprehensive experiments on two popular benchmarks (i.e., HICO-DET and V-COCO), observing that DP-HOI consistently boosts the performance of state-of-the-art HOI detection models. For example, as illustrated in Figure 1, DP-HOI promotes the performance of CDN-S with denoising (DN) [11] by 3.02% mAP.

2. Related Work

2.1. Human-Object Interaction Detection

Existing HOI detection models can be divided into one- and two-stage methods. Previous two-stage methods [14–25] employ an off-the-shelf detector to execute object detection before predicting interactions. These methods intro-

duce additional features [14–18, 26–30, 37, 38] or external knowledge [31, 32, 39] to promote the interaction classification accuracy. For example, Park et al. [37] propose a pose-conditioned self-loop graph neural network to enhance interaction features, while Cao et al. [39] incorporates structured text knowledge to promote HOI detection performance. Due to their multi-stage nature, the two-stage methods generally have a slow inference process. To overcome this problem, one-stage methods [33–36] were proposed and they typically perform object detection and interaction classification in parallel.

Based on DETR's success [10], recent studies have focused on developing DETR-based HOI detection models, achieving significant performance improvement [40–52]. This is mainly because the cross-attention operation in transformer decoder layers flexibly extracts image-wide context information for interaction classification. DETR-based methods can be divided into two groups. The first group directly utilizes the conventional DETR structure [40–44, 48, 52]. The second group increases the power of DETR models and can be further divided into two sub-categories: query-enhanced methods [53, 58, 59] and structure-enhanced methods [47, 54–56, 61–65]. The query-enhanced methods enhance HOI detection performance with semantically clear queries. In contrast, the structure-enhanced methods aim to develop customized model architectures for HOI detection. Some studies [47, 64, 67–69] recently proposed improving HOI detection performance by transferring knowledge from visual-linguistic pre-trained models (e.g., CLIP [70]).

The above methods have achieved impressive performances. However, they still initialize model parameters according to a pre-trained DETR model using object detection datasets. Therefore, their performance in interaction classification may still be sub-optimal.

2.2. Pre-training Methods for Detection Tasks

Pre-training and fine-tuning have become popular pipelines for object detection. Due to the DETR architecture's increase in popularity for object detection, researchers have started studying DETR-specialized pre-training methods. For example, UP-DETR [73] utilizes a proxy task that uses a randomly cropped image patch as the query and forces the DETR model to predict the patch location in the image. Moreover, DETReg [74] uses an unsupervised region proposal generator to produce potential object-bounding boxes. These boxes are used to pre-train the DETR model via the bounding-box regression task.

Since the transformer training is data hungry [75] and existing HOI detection datasets are usually small [6, 7], DETR-based HOI detection models usually adopt pre-trained DETR weights on object detection datasets. However, this strategy may be unsuitable, since HOI detection

includes object detection and interaction classification. To solve this problem, Yuan et al. [66] utilized manually labeled scene graph data for HOI detection pre-training. In comparison, subsequent studies [71, 72] proposed various pseudo-labeling approaches that associate image-level HOI labels with object-bounding boxes, significantly expanding the pre-training data scale.

In this paper, we separate the pre-training of the two sub-tasks in HOI detection to bypass the tricky and noisy pseudo-labeling process. In this way, both sub-tasks benefit from clean labels. The experimental results indicate that DP-HOI significantly improves HOI detection performance.

3. Methods

In this section, we first briefly describe the research motivation and overall DP-HOI framework. Then, we introduce its detection and verb classification branch structures in Section 3.2 and Section 3.3, respectively. Moreover, in Section 3.4, we broaden our approach to video-based action and image-caption data. Finally, additional details are provided in Section 3.5.

3.1. Overview

HOI detection can be divided into two sub-tasks: interactive human-object pair detection and interaction classification. These sub-tasks are closely related to the object detection and action recognition tasks. These tasks have large-scale labeled datasets because their annotation costs are cheaper. Moreover, action recognition data can be supplemented with image-caption data. Based on this observation, we proposed using the existing datasets for object detection, verb classification and image captioning to pre-train HOI detection models.

Our proposed DP-HOI framework is illustrated in Figure 2. During pre-training, each batch contains a set of images from the object detection $D_d=\{\boldsymbol{X}_i^d, \mathbf{y}_i^d\}_{i=1}^{N_d}$ and another set of images from action recognition datasets $D_a=\{\boldsymbol{X}_i^a, \mathbf{y}_i^a\}_{i=1}^{N_a}$. N_d and N_a denote the number of images. While the annotation $\mathbf{y}_i^d$ contains object bounding boxes and object categories, $\mathbf{y}_i^a$ only contains the verb categories. First, a given image $\boldsymbol{X}_i^k$ ($k \in \{d, a\}$) was fed into the CNN backbone in Figure 2. Then, the output feature maps were flattened and injected with fixed sine positional encoding. Finally, the feature maps were enhanced by the self-attention operations in the transformer encoder.

Since $\mathbf{y}_i^d$ and $\mathbf{y}_i^a$ contain object and action labels only, we utilized the enhanced $\boldsymbol{X}_i^d$ and $\boldsymbol{X}_i^a$ features in a disentangled manner. In summary, there are two branches after the transformer encoder, i.e., one detection branch and one verb classification branch. The enhanced $\boldsymbol{X}_i^d$ and $\boldsymbol{X}_i^a$ features pass through the detection and verb classification branches, respectively.

3.2. The Object Detection Branch

This branch contains a detection decoder. We denote the features enhanced by the transformer encoder as $\mathbf{V}_e$, the learnable object queries as $\mathbf{Q}_o = \{Q_0, Q_1, ..., Q_{N-1}\}$, and the initial decoder embeddings as $\mathbf{o_0}$. The output decoder embeddings $\mathbf{o_d}$ can be represented as follows:

$$\mathbf{o_d} = D_d(\mathbf{Q_o}, \mathbf{o_0}, \mathbf{V_e}), \tag{1}$$

where $D_d(\cdot)$ represents the detection decoder. Finally, $\mathbf{o_d}$ was employed to predict object bounding boxes and categories using feed-forward networks (FFNs):

$$\hat{\mathbf{y}}_{box} = f_h(\mathbf{o_d}), \tag{2}$$

$$\hat{\mathbf{y}}_o = f_o(\mathbf{o_d}), \tag{3}$$

where f_h and f_o denote two FFNs. Finally, we imposed supervision on $\hat{\mathbf{y}}_{box}$ and $\hat{\mathbf{y}}_o$ via bipartite matching [10].

3.3. The Verb Classification Branch

We designed the DP-HOI verb classification branch to maximize the pre-training efficacy according to the structure of recently popular HOI detection models (e.g., CDN [44]). In the experimentation section, we demonstrate that DP-HOI significantly improves the performance of other HOI detection models.

The verb classification branch contains two sequential decoders (i.e., a detection decoder and interaction decoder). The same as [44, 47], the first decoder's output embeddings were utilized as the queries for the second one. The detection decoder shares parameters with that in the object detection branch. The two decoders are utilized for object detection and verb classification, respectively.

A slight difference exists between our detection decoder and that in the CDN [44] model. Specifically, the CDN model's detection decoder detects interactive human–object pairs, enabling its interaction decoder to recognize the verb categories of a specific human–object pair. In comparison, our adopted pre-training datasets do not contain any interactive human–object pair annotations. Hence, we proposed reducing the action recognition within a human–object pair to identifying all the actions performed by a human instance. Thus, we selected reliable human instances according to the detection decoder's predictions.

Reliable Person Queries. In this study, a human instance was regarded as reliable if the detection decoder's human category prediction score was above the threshold T. RPQs are the decoder embeddings in $\mathbf{o_d}$ that predict these reliable instances for the interaction decoder. The collection of the RPQs for one image is denoted as $\mathbf{Q_p}$. Each RPQ searches for action-relevant cues using cross-attention on the specific person within the interaction decoder:

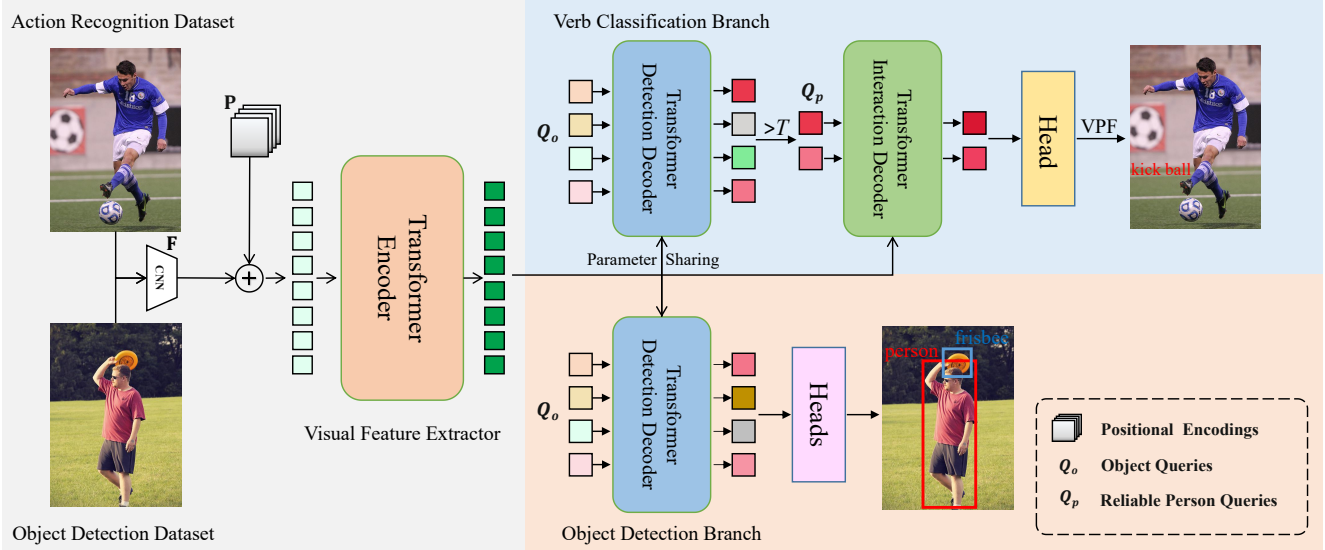$$\mathbf{o_a} = D_a(\mathbf{Q_p}, \mathbf{o_0}, \mathbf{V_e}), \tag{4}$$

Figure 2. Our DP-HOI framework overview. It includes a CNN backbone, a transformer encoder, an object detection branch, and a verb classification branch. The two branches are trained in a disentangled manner, with labeled databases for object detection and action recognition, respectively. Each training image from the action recognition dataset first passes the detection decoder, identifies reliable human instances, and generates reliable person queries (RPQs) for the interaction decoder. Then, each RPQ is responsible for searching for relevant action cues for the specified human instance. Since we only have image-level action labels, we impose supervision on the fused RPQs predictions.

where $D_a(\cdot)$ represents the interaction decoder and $\mathbf{o_a}$ denotes the output decoder embeddings from the interaction decoder. Finally, $\mathbf{o_a}$ is utilized for verb classification:

$$\hat{\mathbf{y}}_a = f_a(\mathbf{o_a}), \qquad (5)$$

where f_a denotes an FFN and $\hat{\mathbf{y}}_a \in \mathbb{R}^{N_p \times C_a}$. N_p and C_a represent the number of RPQs and the number of verb classes, respectively.

Verb-wise Prediction Fusion. Although action recognition datasets generally do not provide instance-level action annotations, a single image may contain multiple human instances. Hence, we propose Verb-wise Prediction Fusion (VPF) to fuse the prediction results in $\hat{\mathbf{y}}_a$ by conducting *max-pooling* along its column dimension. In the experimentation section, we demonstrated that VPF outperforms other fusion strategies and effectively suppresses noisy RPQ predictions.

3.4. Extension to Video and Caption Data

Video Data. Existing action recognition datasets are usually video-based. To utilize these video-based datasets, we randomly sampled N_f frames from each video and fed them into our model to obtain RPQ prediction results, denoted by $\{\hat{y}_a\}_{N_f}$. Then, we utilized the VPF method to fuse these prediction results. Finally, we adopted focal loss [77] to supervise the fused results according to the video label.

Image-Caption Data. Since action recognition datasets are labeled according to fixed action categories, the action semantics they contain are insufficient. As a result, contrastive learning was utilized, enabling the use of image-caption data with robust action semantic information for pre-training. First, we used a rule-based language parser [80] to obtain HOI triplets $\langle human, verb, object \rangle$ for a given image-caption pair. Then, we fed each HOI-triplet prompt (i.e., a photo of {human} {verb} {object}) into the CLIP text encoder to obtain its embedding.

Selecting negative samples during contrastive learning significantly impacts model performance. In this paper, we clustered all the HOI-triplet text embedding categories into 100 offline clusters. Then we sampled 10 HOI categories from each cluster as negative samples for each corresponding RPQ embedding. Next, we calculated the cosine similarity between the RPQs' decoder and text embeddings, selecting the RPQ with the highest similarity score.

Finally, we compute the InfoNCE loss [85] separately in two directions to obtain the image alignment loss $\mathcal{L}_{i2t}$ and the text alignment loss $\mathcal{L}_{t2i}$. The average of these two loss functions is used as the final loss $\mathcal{L}_s$:

$$\mathcal{L}_s = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}). \qquad (6)$$

3.5. Overall Loss Function

We adopted similar loss functions as existing object detection [10] and verb classification [40] studies. The overall

DP-HOI loss function is represented as follows:

$$\mathcal{L} = \mathcal{L}_d + \lambda_v \mathcal{L}_v, \tag{7}$$

$$\mathcal{L}_d = \lambda_b \mathcal{L}_b + \lambda_g \mathcal{L}_g + \lambda_c \mathcal{L}_c, \tag{8}$$

$$\mathcal{L}_v = \lambda_a \mathcal{L}_a + \lambda_s \mathcal{L}_s, \tag{9}$$

where $\mathcal{L}_d$ and $\mathcal{L}_v$ denote the object detection and verb classification branches' loss functions, respectively. $\mathcal{L}_b$, $\mathcal{L}_g$, $\mathcal{L}_c$, $\mathcal{L}_a$ and $\mathcal{L}_s$ represent the losses, including L1 and GIOU [76] for bounding box regression, cross-entropy for object classification, focal [77] and InfoNCE [85] for verb prediction, respectively. In addition, λ_v is a weight that balances the two branches' losses. λ_b, λ_g, λ_c, λ_a and λ_s are set as 5, 2, 1, 1 and 1, respectively.

Moreover, we utilized multiple action recognition and image-caption datasets for pre-training. C_a denotes the total verb category number of all the action recognition datasets. Since semantically overlapping verb categories may exist between different datasets, we only activate the binary classifiers for the verb categories owned by the database that each training sample belongs to. More pre-training details are provided in the supplementary material.

4. Experiments

4.1. The Pre-training Datasets

As illustrated in Table 1, we adopted the MS-COCO [9] and Objects365 [83] datasets for the object detection branch. Then, we employed the action recognition and image-caption datasets in the verb classification branch. First, the action recognition datasets included Haa500 [13] and Kinetics-700 [84]. Haa500 and Kinetics-700 are video-based datasets; therefore, we sampled frames at regular intervals during data processing. Considering the lower quality of video frames in Kinetics-700 compared to Haa500, we treated each sampled frame from Haa500 as an individual supervision sample. Then, we applied video-level supervision to the frame sequences sampled from Kinetics-700. Second, the image-caption datasets included Flickr30k [86] and VG [12]. Aside from captions, Flickr30k and VG include additional annotation information; however, we only used the caption annotations for our pre-training. We also filtered the images from which caption HOI triplets could not be extracted. The datasets and data processing methods are detailed in the supplementary material.

4.2. The HOI Detection Datasets

HICO-DET. HICO-DET [6] is a popular dataset for HOI detection. It consists of 47,776 images with more than 150,000 human-object pairs, 38,118 of which are used for training and 9,658 for testing. This dataset contains the same 80 object classes as MS-COCO [9] and 117 interaction classes. The combination of object and interaction

Table 1. Statistics of the adopted object detection, action recognition and image-caption datasets used for pre-training.

Types	Datasets	#Samples	#Classes
Object Detection	COCO	117266	80
	Objects365	117266	365
Action Recognition	Haa500	52644	500
	Kinetics-700	117266	700
Image-Caption	Flickr30k	25977	-
	VG	54280	-

classes forms 600 HOI categories. Also, there are 138 HOI categories with less than 10 training samples, which are denoted as "rare" categories. We conducted experiments using the default(DT) mode and three zero-shot settings (i.e., UV, RF-UC, and NF-UC). UV and UC represent unseen verb and composition settings, respectively. RF means rare first, and NF is non-rare first.

V-COCO. V-COCO [7] is a relatively small dataset, built on the MS-COCO database [9]. It contains 10,346 images (i.e., 5,400 for training and 4,946 for testing), covering the same 80 object categories as MS-COCO [9] and 26 interaction categories. We use the mean average precision of Scenario 1 ($\mathrm{AP}_{\mathrm{role}}$) [7] for evaluation.

4.3. Implementation Details

We adopted ResNet-50 as our backbone model. We utilized the AdamW [78] optimizer to conduct experiments with a batch size of 64 on 8 A800 GPUs for DP-HOI. In each batch, the number of samples from the object detection and action recognition datasets is equal. The initial learning rate was set to 1e-4 and multiplied by 0.1 after 180 epochs. The pre-training stage lasts for 200 epochs according to the MS-COCO dataset. Regarding the Kinetics-700 samples, we resized the input video frames from their original size to 256×256 pixels. Meanwhile, the other datasets' input samples were resized to a minimum and maximum of 800 pixels and 1,333 pixels on the short and long sides, respectively. λ_v, T, and N were set to 1, 0.9, and 100, respectively. Furthermore, we adopted the DN [11] strategy to accelerate the pre-training stage, and the number of detection and interaction decoder layers was set to 3. Please refer to the supplementary material for more implementation details.

4.4. Comparisons with State-of-the-Art Methods

Comparisons on HICO-DET under default setting. We directly applied the pre-trained DETR weights from DP-HOI to existing popular methods.

As shown in Table 2, DP-HOI significantly enhances the performance of one- and two-stage HOI detection methods. When the pre-trained DP-HOI weights are applied to two-stage methods, UPT [55] and PViC [38], consistent perfor-

Table 2. Performance comparisons for HICO-DET. † means DN was adopted in the fine-tuning stage. * denotes a data augmentation strategy [53] was employed.

Methods	Backbone	DT Mode		
		Full	Rare	Non-Rare
InteractNet [79]	ResNet-50-FPN	9.94	7.16	10.77
GPNN[80]	Res-DCN-152	13.11	9.34	14.23
iCAN [15]	ResNet-50	14.84	10.45	16.15
No-Frills [14]	ResNet-152	17.18	12.17	18.68
UnionDet [35]	ResNet-50-FPN	17.58	11.72	19.33
DRG [81]	ResNet-50-FPN	19.26	17.74	19.71
PD-Net [27]	ResNet-152	20.81	15.90	22.28
PPDM [33]	Hourglass-104	21.73	13.78	24.10
GGNet [34]	Hourglass-104	23.47	16.48	25.60
HOTR [42]	ResNet-50	23.46	16.21	25.62
HOI-Trans [43]	ResNet-50	23.46	16.91	25.41
AS-Net [35]	ResNet-50	28.87	24.25	30.25
QPIC [40]	ResNet-50	29.07	21.85	31.23
CDN-S [44]	ResNet-50	31.44	27.39	32.64
HQM(CDN-S) [59]	ResNet-50	32.47	28.15	33.76
DOQ(CDN-S) [53]	ResNet-50	33.28	29.19	34.50
GEN-VLKT$_s$ [47]	ResNet-50	33.75	29.25	35.10
with our pre-trained model weights				
UPT [55]	ResNet-50	31.66	25.94	33.36
UPT + Ours	ResNet-50	**33.36**	**28.74**	**34.75**
PViC [38]	ResNet-50	34.69	32.14	35.45
PViC + Ours	ResNet-50	**35.77**	**32.26**	**36.81**
CDN-S† [44]	ResNet-50	31.98	28.61	32.99
CDN-S†+Ours	ResNet-50	35.00	32.38	**35.78**
CDN-S†+CCS*+Ours	ResNet-50	**35.38**	**34.61**	35.61
HOICLIP [64]	ResNet-50	34.69	31.12	35.74
HOICLIP+Ours	ResNet-50	**36.56**	**34.36**	**37.22**
comparison with pre-training methods				
OpenCat (754k) [72]	ResNet-101	32.68	28.42	33.75
RLIP (225k) [66]	ResNet-50	32.84	26.85	34.63
RLIPv2 (1,967k) [71]	ResNet-50	35.38	29.61	37.10
CDN-S†+CCS*+Ours (484k)	ResNet-50	35.38	**34.61**	35.61
HOICLIP+Ours (484k)	ResNet-50	**36.56**	34.36	**37.22**

Table 3. Performance comparisons with state-of-the-art methods for zero-shot HOI detection on HICO-DET. UV and UC indicate unseen verb and composition settings, respectively. RF is short for rare first. NF is non-rare first.

Methods	Type	Unseen	Seen	Full
GEN-VLKT$_s$ [47]	UV	20.96	30.23	28.74
EoID [68]	UV	22.71	30.73	29.61
OpenCat [72]	UV	19.48	29.02	27.43
HOICLIP [64]	UV	24.30	32.19	31.09
HOICLIP+Ours	UV	**26.30**	**34.49**	**33.34**
GEN-VLKT$_s$ [47]	RF-UC	21.36	32.91	30.56
EoID [68]	RF-UC	22.04	31.39	29.52
OpenCat [72]	RF-UC	21.46	33.86	31.38
RLIPv2 [71]	RF-UC	21.45	35.85	32.97
HOICLIP [64]	RF-UC	25.53	34.85	32.99
HOICLIP+Ours	RF-UC	**30.49**	**36.17**	**35.03**
GEN-VLKT$_s$ [47]	NF-UC	25.05	23.38	23.71
EoID [68]	NF-UC	26.77	26.66	26.69
OpenCat [72]	NF-UC	23.25	28.04	27.08
RLIPv2 [71]	NF-UC	22.81	29.52	28.18
HOICLIP [64]	NF-UC	26.39	28.10	27.75
HOICLIP+Ours	NF-UC	**28.87**	**29.98**	**29.76**

S†+CCS*. With a stronger baseline HOICLIP [64], DP-HOI outperforms RLIPv2, especially on the rare HOI categories.

Comparisons on HICO-DET under zero-shot settings. To further demonstrate DP-HOI's effectiveness, we conducted experiments using various zero-shot settings, including UV, RF-UC, and NF-UC.

As illustrated in Table 3, DP-HOI achieved competitive performance across all three zero-shot settings. It outperforms state-of-the-art methods under the UV, RF-UC and NF-UC settings, reaching 26.30%, 30.49%, and 28.87% mAP in the unseen categories, respectively. In contrast to HOICLIP [64], we observed consistent performance gains of 2.00%, 4.96% and 2.48% mAP for unseen categories under the UV, RF-UC and NF-UC settings.

Comparisons on V-COCO. Table 4 displays the V-COCO comparisons. We observed that DP-HOI consistently enhances the model's performance on the V-COCO dataset, reaching 63.2%, 64.8%, and 66.6% mAP in the AP_{role} for QPIC [40], CDN-S [44], and GEN$_s$ [47], respectively. GEN$_s$+Ours method outperforms other HOI pre-training approaches with 484k samples. These results demonstrate that DP-HOI provides superior pre-trained weights for HOI detection models.

4.5. The Experiment Using Different Datasets

We conducted experiments using various pre-training data combinations to explore their impact on the datasets, as illustrated in Table 5. COCO indicates pre-training with only the MS-COCO dataset, which is regarded as the baseline. ALL signifies pre-training with the 484k data shown in Table 1.

mance gains of 1.70% and 1.08% mAP were observed in DT mode for the full categories, respectively. Furthermore, applying pre-trained DP-HOI weights to one-stage methods, CDN-S† [44] and HOICLIP [64], yielded consistent performance improvements of 3.02% and 1.87% mAP, respectively. Notably, we observed remarkable performance improvements on the rare HOI categories. For example, compared with DOQ [53], which also adopts CCS as the data augmentation method, the performance of CDN-S† on the rare categories was promoted by 5.42% mAP, reaching 34.61% mAP. These improvements demonstrated DP-HOI's efficiency and universality.

Moreover, compared with other HOI pre-training methods, DP-HOI outperforms OpenCat [72] and RLIP [66]. Its performance was similar to RLIPv2 [71] with less pre-training data. RLIPv2 introduces a complex scheme to obtain pseudo-labeled scene graph data from object detection datasets and adopts 1,967k images for pre-training. In contrast, DP-HOI adopts a concise pre-training strategy that effectively leverages action semantic information from action recognition and image-caption datasets. It also yields similar results with less data for the simple baseline CDN-

Table 4. Performance comparisons for V-COCO. GEN_s indicates that CLIP distillation was removed from GEN-VLKT$_s$.

Methods	Backbone	AP_{role}
InteractNet [79]	ResNet-50-FPN	40.0
DRG [81]	ResNet-50-FPN	51.0
PD-Net [27]	ResNet-152	52.6
IDN [82]	ResNet-50-FPN	53.3
GGNet [34]	Hourglass-104	54.7
HOTR [42]	ResNet-50	55.2
HOI-Trans [43]	ResNet-101	52.9
AS-Net [41]	ResNet-50	53.9
IF [57]	ResNet-50	63.0
PartMap [63]	ResNet-50	63.0
DOQ(QPIC) [53]	ResNet-50	63.5
HQM(QPIC) [59]	ResNet-50	63.6
with our pre-trained model weights		
QPIC [40]	ResNet-50	58.8
QPIC+Ours	ResNet-50	**63.2**
CDN-S [44]	ResNet-50	61.7
CDN-S+Ours	ResNet-50	**64.8**
GEN_s+VLKT [47]	ResNet-50	62.4
GEN_s+Ours	ResNet-50	**66.6**
comparisons with pre-training methods		
OpenCat (754k) [72]	ResNet-50	61.9
RLIP (225k) [66]	ResNet-101	61.9
RLIPv2 (1,967k) [71]	ResNet-50	65.9
GEN_s+Ours (484k)	ResNet-50	**66.6**

Initially, we extended pre-training data from the Haa500, Kinetics-700, and Flickr30k datasets independently, reaching 1.34%, 1.36% and 1.33% improvement on full categories. These results demonstrate that using various action dataset types (e.g., action image, action video and image-caption dataset) could be beneficial for DP-HOI. Furthermore, we observed consistent performance improvements when integrating diverse action datasets. With the 484k pre-training data, the DP-HOI outperforms the baseline by 3.02%, 3.77%, and 2.79% mAP on the full, rare, and non-rare HOI categories, respectively. These experimental results demonstrate DP-HOI's scalability and effectiveness. Moreover, we encountered an intriguing observation: integrating image-caption data led to a remarkable improvement on the non-rare categories. This could be attributed to the diverse range of action classes in the image-caption dataset.

4.6. The Ablation Study

We conducted pre-training using the COCO and Kinetics-700 datasets in the ablation study. Then, we fine-tuned the HICO-DET database with the CDN-S model [44] adopted through DN [11]. The detailed experimental results are summarized in Table 6. COCO indicates pre-training with only the MS-COCO dataset, which is regarded as the baseline. K700 represents the 117k Kinetics-700 data in Table 1. **The Effectiveness of DP-HOI.** As illustrated in Table 6,

Table 5. Performance comparisons using different pre-training datasets.

Datasets	Full	Rare	Non-Rare
COCO	31.98	28.61	32.99
+Haa500	33.32	30.18	34.26
+Kinetics-700	33.34	29.89	34.37
+Flickr30k	33.31	30.76	34.08
+Haa500, Flickr30k	34.06	30.89	35.01
+Flickr30k, VG	33.77	31.54	34.43
ALL	**35.00**	**32.38**	**35.78**

Table 6. Ablation study on each key DP-HOI component.

Methods	Components				mAP		
	COCO	K700	VPF	RPQ	Full	Rare	Non-Rare
Baseline	✓	-	-	-	31.98	28.61	32.99
Incremental	✓	✓	-	-	32.69	28.64	33.90
	✓	✓	✓	-	32.93	28.87	34.14
	✓	✓	-	✓	33.07	28.96	34.30
Ours	✓	✓	✓	✓	**33.34**	**29.89**	**34.37**

DP-HOI significantly outperforms the baseline by 1.36%, 1.28%, and 1.38% mAP in DT mode for the full, rare, and non-rare HOI categories, respectively.

The Effectiveness of VPF. 'max-pooling' was proposed to fuse the RPQ prediction results. As shown in Table 6, when VPF was removed, the HICO-DET full category performance declined by 0.27% mAP. Since the people in the video frames may not have performed any action, directly imposing supervision on each RPQ without fusion is unreasonable.

The Effectiveness of RPQ. RPQ was employed to identify human instances from the detection decoder and generate person-specific queries for the subsequent interaction decoder. This strategy enabled the model to focus on the action cues for each specific person. When RPQ was removed and all detection decoder embeddings were fed into the interaction decoder, the HOI detection performance decreased by 0.41% mAP. The above experimental results verify that query selection for the interaction decoder is essential in DP-HOI.

4.7. Comparisons With DP-HOI Variants

The Comparisons With VPF Variants. We compared the VPF's performance with two of its variants. The experimental results are displayed in Table 7. The first variant is denoted as "w/o fusion", indicating that we imposed supervision on each confident RPQ. As a result, it displayed a performance lower than VPF by 0.27%, 0.93%, and 0.07% mAP in DT mode for the full, rare, and non-rare HOI categories, respectively. We estimate that conducting max-pooling on all RPQ predictions enables VPF to suppress

Table 7. Performance comparisons with the DP-HOI VPF Operation in the HICO-DET DT mode.

Methods	Full	Rare	Non-Rare
w/o fusion	33.07	28.96	34.30
avg-pooling	32.76	28.40	34.06
max-pooling	**33.34**	**29.89**	**34.37**

noisy predictions by non-confident RPQs.

Second, "avg-pooling" means that we average all the RPQ prediction results and imposed supervision. Table 7 shows that this setting in the full HOI categories decreases by 0.58% mAP. This may be because this setting implicitly forces all RPQs to make confident predictions according to the annotations.

Table 8. Performance comparisons with different parameter initializations. † means DN was adopted during the fine-tuning stage.

Methods	w/o decoder	w/o interaction decoder	full
CDN-S†	31.06	31.98	31.39
CDN-S† + Ours	**33.10**	**34.16**	**35.00**

The Comparisons with Various Model Initializations. As illustrated in Table 8, we compared the performance of three model initialization strategies using the CDN-S† model: (a) without the decoder, initializing the backbone and encoder; (b) without the interaction decoder, initializing the backbone, encoder and detection decoder; (c) full, initializing the backbone, encoder, detection and interaction decoder. CDN-S† adopts the model pre-trained using the MS-COCO dataset. We also utilized the pre-trained detection decoder to initialize the CDN-S† detection and interaction decoders in the (c) setting.

As illustrated in Table 8, DP-HOI outperforms the baseline by 2.04%, 2.18% and 3.61% mAP in the (a), (b) and (c) settings, respectively. Moreover, when we only initialized our pre-trained model's backbone and encoder, the performance reached 33.10% mAP, outperforming the original pre-trained model with any of the three initialization strategies. These results demonstrate that DP-HOI incorporates action-related features in the backbone and encoder, enhancing our pre-trained model's applicability across various HOI models.

4.8. The DP-HOI Visualizations

As illustrated in Figure 3, we visualized the attention maps for the last detection (i.e., the first row) and interaction decoder layers (i.e., the second row) of the most confident human query according to the RPQ. We observed that the detection attention maps accurately localize the person's boundaries. Likewise, the interaction attention maps accu-
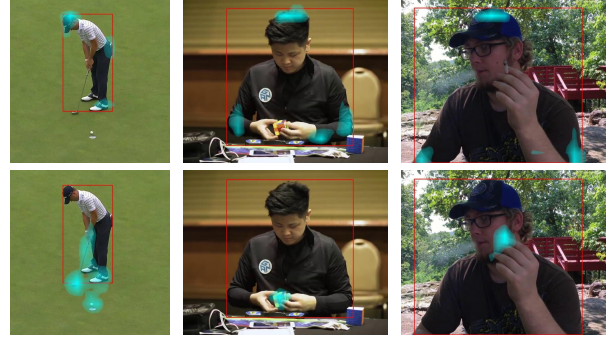

Figure 3. Visualization of the attention maps in the decoder layers. The two rows represent results for the detection and interaction decoders, respectively.

rately localize the interaction regions. Therefore, with the disentangled supervision signals, the two decoders use different features for object detection and interaction classification.

5. Conclusions and Limitations

In this paper, we addressed the pre-training problem for DETR-based HOI detection models. Specifically, we proposed a disentangled pre-training framework that effectively explores readily available and large-scale object detection, action recognition and image-caption datasets. Our pre-training architecture is consistent with the downstream HOI detection task, facilitating efficient knowledge transfer. In addition, we conducted comprehensive experiments on two popular HOI detection benchmarks. The experimental results demonstrated our methods' superiority. A possible limitation of this study is that it requires GPUs with relatively large memories for pre-training. In the future, we will explore more efficient pre-training strategies, that can include more object detection, action recognition and image-caption datasets, further enhancing the pre-training stage.

Broader Impacts. DP-HOI significantly improves the performance of HOI detection models. It could impact human-centric vision applications such as driver monitoring and health care systems. To the best of our knowledge, this study does not have any obvious negative social impacts.

References

[1] E. Mascaro, D. Sliwowski, D. Lee, HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Collaborative roBOTs. In *CoRL*, 2023. 1

[2] T. Yao, Y. Pan, Y. Li, T. Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 1

[3] B. Pan, H. Cai, D. Huang, K. Lee, A. Gaidon, E. Adeli, J. Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *CVPR*, 2020. 1

[4] J. Johnson, R. Krishna, M. Stark, L. Li, D. Shamma, M. Bernstein, L. Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 1

[5] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, Y. Zhuang, Counterfactual samples synthesizing for robust visual question answering. In *CVPR*, 2020. 1

[6] Y. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2, 5

[7] S. Gupta, J. Malik. Visual semantic role labeling. arXiv:1505.04474, 2015. 2, 5

[8] C. Ding, D. Tao. Trunk-Branch Ensemble Convolutional Neural Networks for Video-Based Face Recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1

[9] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 5

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 3, 4

[11] F. Li, H. Zhang, S. Liu, J. Guo, L. Ni, L. Zhang. Dndetr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 1, 2, 5, 7

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. Shamma, Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International journal of computer vision*, 2017. 5

[13] J. Chung, C. Wuu, H. Yang, Y. Tai, C. Tang. Haa500: Human-centric atomic action dataset with curated videos. In *ICCV*, 2021. 5

[14] T. Gupta, A. Schwing, D. Hoiem. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. In *ICCV*, 2019. 2, 6

[15] C. Gao, Y. Zou, J. Huang. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. In *British Machine Vision Conference*, 2018. 6

[16] T. Wang, R. Anwer, M. Khan, F. Khan, Y. Pang, L. Shao, J. Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019.

[17] P. Zhou, M. Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019.

[18] B. Wan, D. Zhou, Y. Liu, R. Li, X. He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019. 2

[19] S. Wang, K. Yap, H. Ding, J. Wu, J. Yuan, Y. Tan. Discovering Human Interactions With Large-Vocabulary Objects via Query and Multi-Scale Detection. In *ICCV*, 2021.

[20] Z. Hou, B. Yu, Y. Qiao, X. Peng, D. Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021.

[21] S. Wang, K. Yap, J. Yuan, Y. Tan. Discovering human interactions with novel objects via zero-shot learning. In *CVPR*, 2020.

[22] Z. Hou, X. Peng, Y. Qiao, D. Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020.

[23] X. Zhong, C. Ding, X. Qu, D. Tao. Polysemy deciphering network for robust human–object interaction detection. In *International Journal of Computer Vision*, 2021.

[24] H. Wang, W. Zheng, L. Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *ECCV*, 2020.

[25] D. Kim, X. Sun, J. Choi, S. Lin, I. Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 2020. 2

[26] Y. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H. Fang, Z. Ma, M. Chen, C. Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 2

[27] X. Zhong, C. Ding, X. Qu, D. Tao. Polysemy deciphering network for human-object interaction detection. In *ECCV*, 2020. 6, 7

[28] O. Ulutan, A. Iftekhar, B. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020.

[29] Y. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, C. Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020.

[30] Y. Liu, J. Yuan, C. Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM MM*, 2020. 2

[31] Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, C. Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 2

[32] T. He, L. Gao, J. Song, Y. Li. Exploiting scene graphs for human-object interaction detection. In *ICCV*, 2021.

[33] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, J. Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 2

[34] X. Zhong, X. Qu, C. Ding, D. Tao. Glance and gaze: Inferring action-aware points for one-stage human-

object interaction detection. In *CVPR*, 2021. 2, 6

[35] B. Kim, T. Choi, J. Kang, H. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 6, 7

[36] T. Wang, T. Yang, M. Danelljan, F. Khan, X. Zhang, J. Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 6

[37] J. Park, J. Park, J. Lee. ViPLO: Vision Transformer based Pose-Conditioned Self-Loop Graph for Human-Object Interaction Detection. In *CVPR*, 2023. 2

[38] F. Zhang, Y. Yuan, D. Campbell, Z. Zhong, S. Gould. Exploring Predicate Visual Context in Detecting of Human-Object Interactions. In *ICCV*, 2023. 2 2, 5, 6

[39] Y. Cao, Q. Tang, F. Yang, X. Su, S. You, X. Lu, C. Xu. Re-mine, Learn and Reason: Exploring the Cross-modal Semantic Correlations for Language-guided HOI detection. In *ICCV*, 2023. 2

[40] M. Tamura, H. Ohashi, T. Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 1, 2, 4, 6, 7

[41] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, C. Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 1, 7

[42] B. Kim, J. Lee, J. Kang, E. Kim, H. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 6, 7

[43] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 6, 7

[44] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, X. Li. Mining the benefits of two-stage and one-stage hoi detection. In *NeurIPS*, 2021. 1, 2, 3, 6, 7

[45] H. Yuan, M. Wang, D. Ni, L. Xu. Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics. In *AAAI*, 2022.

[46] Z. Li, C. Zou, Y. Zhao, B. Li, S. Zhong. Improving human-object interaction detection via phrase learning and label composition. In *AAAI*, 2022.

[47] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, S. Liu. GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection. In *CVPR*, 2022. 1, 2, 3, 6, 7

[48] J. Park, S. Lee, H. Heo, H. Choi, H. Kim. Consistency learning via decoding path augmentation for transformers in human object interaction detection. In *CVPR*, 2022. 2

[49] C. Xie, F. Zeng, Y. Hu, S. Liang, Y. Wei. Category Query Learning for Human-Object Interaction Classification. In *CVPR*, 2023.

[50] X. Zhong, C. Ding, Y. Hu, D. Tao. Disentangled Interaction Representation for One-Stage Human-Object Interaction Detection. arXiv:2312.01713, 2023.

[51] S. Kim, D. Jung, M. Cho. Relational Context Learning for Human-Object Interaction Detection. In *CVPR*, 2023.

[52] D. Zhou, Z. Liu, J. Wang, L. Wang, T. Hu, E. Ding, J. Wang. Human-Object Interaction Detection via Disentangled Transformer. In *CVPR*, 2022. 1, 2

[53] X. Qu, C. Ding, X. Li, X. Zhong, D. Tao. Distillation Using Oracle Queries for Transformer-Based Human-Object Interaction Detection. In *CVPR*, 2022. 2, 6, 7

[54] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, C. Chen. Exploring Structure-Aware Transformer Over Interaction Proposals for Human-Object Interaction Detection. In *CVPR*, 2022. 2

[55] F. Zhang, D. Campbell, S. Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, 2022. 5, 6

[56] A. Iftekhar, H. Chen, K. Kundu, X. Li, J. Tighe, D. Modolo. What to look at and where: Semantic and Spatial Refined Transformer for detecting human-object interactions. In *CVPR*, 2022. 2

[57] X. Liu, Y. Li, X. Wu, Y. Tai, C. Lu, C. Tang. Interactiveness Field in Human-Object Interactions. In *CVPR*, 2022. 1, 7

[58] L. Dong, Z. Li, K. Xu, Z. Zhang, L. Yan, S. Zhong, X. Zou. Category-Aware Transformer Network for Better Human-Object Interaction Detection. In *CVPR*, 2022. 2

[59] X. Zhong, C. Ding, Z. Li, S. Huang. Towards Hard-Positive Query Mining for DETR-based Human-Object Interaction Detection. In *ECCV*, 2022.

[60] G. Wang, Y. Guo, Y. Wong, M. Kankanhalli. Chairs Can be Stood on: Overcoming Object Bias in Human-Object Interaction Detection. In *ECCV*, 2022. 2, 6, 7

[61] B. Kim, J. Mun, K. On, M. Shin, J. Lee, E. Kim. MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection. In *CVPR*, 2022. 1

[62] D. Tu, X. Min, H. Duan, G. Guo, G. Zhai, W. Shen. Iwin: Human-Object Interaction Detection via Transformer with Irregular Windows. In *ECCV*, 2022. 2

[63] X. Wu, Y. Li, X. Liu, J. Zhang, Y. Wu, C. Lu. Mining Cross-Person Cues for Body-Part Interactiveness Learning in HOI Detection. In *ECCV*, 2022. 1, 7

[64] S. Ning, L. Qiu, Y. Liu, X. He. HOICLIP: Efficient Knowledge Transfer for HOI Detection With Vision-Language Models. In *CVPR*, 2023. 2, 6

[65] D. Tu, W. Sun, G. Zhai, W. Shen. Agglomerative Transformer for Human-Object Interaction Detection. In *ICCV*, 2023. 2

[66] H. Yuan, J. Jiang, S. Albanie, T. Feng, Z. Huang, D. Ni, M. Tang. RLIP: Relational Language-Image Pre-

training for Human-Object Interaction Detection. In *NeurIPS*, 2022. 3, 6, 7

[67] T. Lei, F. Caba, Q. Chen, H. Jin, Y. Peng, Y. Liu. Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory. In *ICCV*, 2023. 2

[68] M. Wu, J. Gu, Y. Shen, M. Lin, C. Chen, X. Sun. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *AAAI*, 2023. 6

[69] B. Wan, Y. Liu, D. Zhou, T. Tuytelaars, X. He. Weakly-supervised HOI Detection via Prior-guided Bi-level Representation Learning. arXiv:2303.01313, 2023. 2

[70] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[71] H. Yuan, S. Zhang, X. Wang, S. Albanie, Y. Pan, T. Feng, J. Jiang, D. Ni, Y. Zhang, D. Zhao, RLIPv2: Fast Scaling of Relational Language-Image Pre-training. In *ICCV*, 2023. 1, 3, 6, 7

[72] S. Zheng, B. Xu, Q. Jin Q, Open-Category Human-Object Interaction Pre-Training via Language Modeling Framework. In *CVPR*, 2023. 1, 3, 6, 7

[73] Z. Dai, B. Cai, Y. Lin, J. Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*, 2021. 2

[74] A. Bar, X. Wang, V. Kantorov, C. Reed, R. Herzig, G. Chechik, A. Rohrbach, T. Darrell, A. Globerson. Detreg: Unsupervised pretraining with region priors for object detection. In *CVPR*, 2022. 2

[75] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[76] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 5

[77] T. Lin, P. Goyal, R. Girshick, K. He, Focal loss for dense object detection. In *ICCV*, 2017. 4, 5

[78] I. Loshchilov, F. Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[79] G. Gkioxari, R. Girshick, Detecting and recognizing human-object interactions. In *CVPR*, 2018. 6, 7

[80] S. Qi, W. Wang, B. Jia, J. Shen, S. Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 4, 6

[81] C. Gao, J. Xu, Y. Zou, J. Huang. Drg: Dual relation graph for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 2020. 6, 7

[82] Y. Li, X. Liu, X. Wu, Y. Li, C. Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *NeurIPS*, 2020. 7

[83] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, J. Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 2, 5

[84] J. Carreira, E. Noland, C. Hillier, A. Zisserman, A short note on the kinetics-700 human action dataset. arXiv:1907.06987, 2019. 2, 5

[85] A. Oord, Y. Li, O. Vinyals. Representation learning with contrastive predictive coding. arXiv:1807.03748, 2018. 4, 5

[86] Y. Peter, L. Alice, H. Micah, H. Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, 2014. 5