Disentangled Pre-training for Human-Object Interaction Detection

Supplementary Material

This supplementary material includes four sections. Section 1 provides more details about pre-training datasets. Section 2 describes more training details of DP-HOI. We provide more experimental results on HOI detection methods in Section 3. Section 4 provides experimental results on various zero-shot settings, i.e., UV, RF-UC, and NF-UC.

1. More Details of Pre-training Datasets

Objects365 [66]. Objects365 is a large-scale object detection dataset, which contains nearly 1,724K images with annotations for object detection only. From the 365 classes in Objects365, we select the classes that are overlapped with the 80 classes in COCO. Subsequently, we randomly sampled 117,266 images in the selected object classes.

Haa500 [8]. Haa500 is a video-based action recognition dataset. For each long video, we conduct sampling uniformly with a time interval of 0.5s. For each video that is shorter than 2s, we uniformly sample 4 frames in the video. Since the action changes in each video are very small, we utilized the sampled 526,44 video frames as an image-based action recognition database for pre-training.

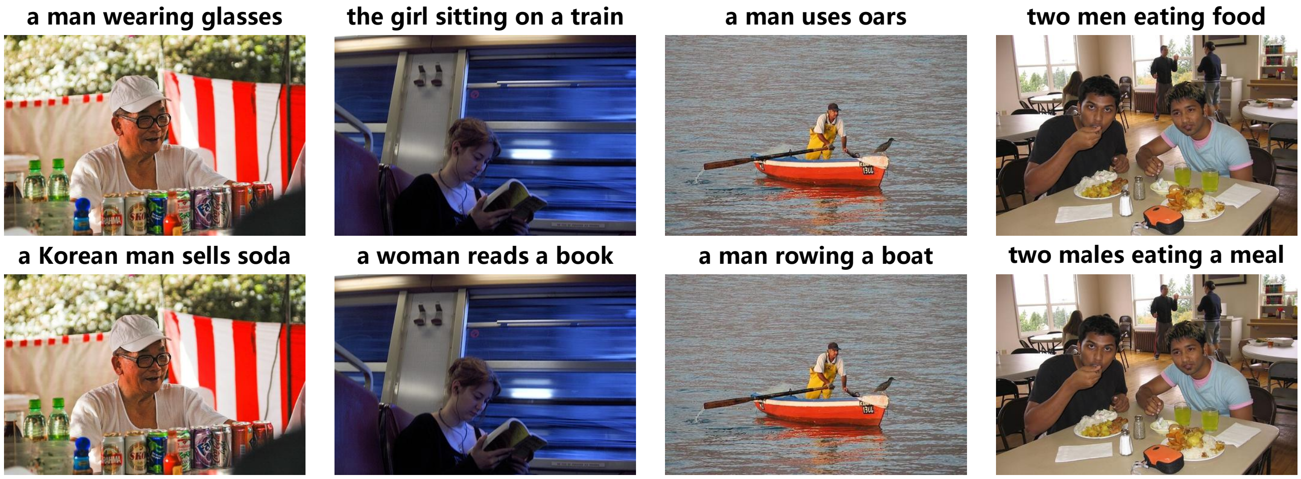
Kinetics-700 [73]. Kinetics-700 is a large-scale video-based action recognition dataset, which contains over 650K videos in 700 classes. For each long video, we randomly select a starting frame and sample 16 frames with a frame interval of 4. We uniformly sample videos in these 700 classes and obtain 117K videos.

Flickr30k [86]. Flickr30k dataset contains nearly 30K images collected from Flickr. Each image owns 5 different captions. We use our rule-based language parser [60] to obtain qualified HOI triplets from captions. For example, given an image with captions {"a man drives a car", "car runs on the road", "a man on the road"}, we remove the triples where the subject is not a person and the relation is not a verb, i.e. {"car runs on the road", "a man on the road"}. We visualize some examples of obtained HOI triplets in Figure 1. In the first three columns, the obtained HOI triplets exhibit diverse actions. As shown in the last column, there are several HOI triplets with similar semantics in our data. Different HOI triplets of similar meaning could enrich the diversity of text embeddings, which helps to increase the robustness of the model and prevent overfitting. Therefore, we do not perform additional processing for these synonyms.

Visual Genome [7]. Visual Genome(VG) consists of 101,174 images sampled from MS-COCO [6], with densely annotated object, attribute and relationship labels. We utilize the captions provided in VG and search for effective HOI triplets according to the same role for Flickr30k. In addition, we do not utilize the VG images that are overlapped with the V-COCO test set to avoid information leakage.

2. More Training Details

We adopt the denoising (DN) strategy [2] to accelerate the pre-training and fine-tuning stages. In the pre-training stage, we first add noise to the ground-truth coordinates of each object bounding box and then use a two-layer FFN with ReLU to encode the coordinates [43]. We also used the label denoising strategy strategy in [10] to speed up pre-training. In the fine-tuning stage, we adopt the ground-truth coordinates of labeled human-object pairs to construct an auxiliary group of queries. Specifically, we add noise to



a man wearing glasses

the girl sitting on a train

a man uses oars

two men eating food

a Korean man sells soda

a woman reads a book

a man rowing a boat

two males eating a meal

Figure 1. Visualization of obtained HOI triplets on Flickr30k. Each column indicates an image and its obtained HOI triplets in Flickr30k.

Table 1. Performance comparisons on HICO-DET. GEN$_s$ denotes that distillation via CLIP is removed from GEN-VLKT$_s$. † means DN is adopted in the fine-tuning stage. * denotes using a data augmentation strategy [43].

Methods	Backbone	DT Mode			Known Object		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
UnionDet [32]	ResNet-50-FPN	17.58	11.72	19.33	19.76	14.68	21.27
DRG [61]	ResNet-50-FPN	19.26	17.74	19.71	23.40	21.75	23.89
PD-Net [24]	ResNet-152	20.81	15.90	22.28	24.78	18.88	26.54
PPDM [30]	Hourglass-104	21.73	13.78	24.10	24.81	17.09	27.12
GGNet [31]	Hourglass-104	23.47	16.48	25.60	27.36	20.23	29.48
HOI-Trans [37]	ResNet-50	23.46	16.91	25.41	26.15	19.24	28.22
AS-Net [32]	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [34]	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
QPIC+Ours	ResNet-50	**30.63**	**25.27**	**32.23**	**32.94**	**27.24**	**34.64**
CDN-S [38]	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42
CDN-S† [38]	ResNet-50	31.98	28.61	32.99	34.77	31.34	35.80
CDN-S+Ours	ResNet-50	34.27	30.02	35.54	37.05	33.09	38.23
CDN-S†+Ours	ResNet-50	35.00	32.38	**35.78**	37.83	35.43	**38.54**
CDN-S†+CCS*+Ours	ResNet-50	**35.38**	**34.61**	35.61	**38.21**	**37.43**	38.44
GEN$_s$+VLKT [49]	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
GEN$_s$+Ours	ResNet-50	**34.40**	**31.17**	**35.36**	**38.25**	**35.64**	**39.03**
HOICLIP [77]	ResNet-50	34.69	31.12	35.74	37.61	34.47	38.54
HOICLIP+Ours	ResNet-50	**36.56**	**34.36**	**37.22**	**39.37**	**36.59**	**40.20**

the ground-truth coordinates of each human-object pair. We then adopt the encoding method proposed in [43] to obtain the auxiliary group of queries.

The obtained auxiliary group of queries and the original group of learnable queries are fed into the decoder for prediction. This enables DETR-based models to converge more quickly [10, 43]. For simplicity, the label denoising strategy in [10] is not used in fine-tuning stages.

Moreover, data augmentation strategies are different for image datasets and video datasets. As to image datasets, it includes random scaling, random horizontal flipping, random color jittering and gaussian blurring. The input images are resized to at least 800 pixels on the short size and at most 1333 pixels on the long side. As to video datasets, it includes random scaling, random cropping and random horizontal flipping. The spatial resolution of the input frames is set to 256 × 256.

Pre-training lasts for 200 epochs according to the MS-COCO dataset. The action datasets, including the action recognition and image caption datasets, are added in the 150th epoch. In each batch, the number of samples from object detection and action datasets is the same. When training with both action recognition and image-caption data, we keep the sampling ratio of object detection, action recognition and image-caption data as 2:1:1.

3. More Experimental Results on HICO-DET

In this section, we demonstrate the effectiveness of DP-HOI in the Known-Object(KO) mode under default setting.

As shown in Table 1, DP-HOI significantly boosts HOI detection performance on both DT and KO modes. When the pre-trained DETR weights by DP-HOI are applied to QPIC [34], CDN-S† [38], GEN$_s$+VLKT [49] and HOICLIP [77], we observe consistent performance gains by 1.26 3.06%, 1.47% and 1.76% mAP in KO mode for the full categories. Moreover, the performance of QPIC, CDN-S†, GEN$_s$+VLKT and HOICLIP on the rare HOI category is promoted by 3.42%, 4.09%, 2.89% and 2.12%, respectively.

4. Zero-shot HOI Detection

In this section, we conduct experiments on three zero-shot settings, i.e. Unseen Verb (UV), Rare First Unseen Combination (RF-UC), and Non-rare First Unseen Combination (NF-UC), following previous work [49, 77].

We adopt GEN$_s$ [49] and HOICLIP [77] as our baseline to verify the performance of DP-HOI on zero-shot settings. For clean comparison, we follow the data split protocol on

Table 2. Application to zero-shot HOI detection on HICO-DET. GEN$_s$ denotes distillation via CLIP is removed from GEN-VLKT$_s$[49].

Methods	UV			RF-UC			NF-UC		
	Unseen	Seen	Full	Unseen	Seen	Full	Unseen	Seen	Full
GEN$_s$+VLKT [49]	20.96	30.23	28.74	21.36	32.91	30.56	25.05	23.38	23.71
GEN$_s$+Ours	**23.01**	**31.29**	**30.13**	**23.73**	**33.59**	**31.61**	**25.78**	**25.05**	**25.20**
Improvement	+2.05	+1.06	+1.39	+2.37	+0.68	+1.05	+0.73	+1.67	+1.49
HOICLIP [77]	24.30	32.19	31.09	25.53	34.85	32.99	26.39	28.10	27.75
HOICLIP+Ours	**26.30**	**34.49**	**33.34**	**30.49**	**36.17**	**35.03**	**28.87**	**29.98**	**29.76**
Improvement	+2.00	+2.30	+2.25	+4.96	+1.32	+2.04	+2.48	+1.88	+2.01

each zero-shot setting in the original papers [49, 77].

As illustrated in Table 2, our DP-HOI outperforms the baseline on most zero-shot settings. Compared with GEN$_s$+VLKT [49], we achieve an impressive 2.05%, 2.37% and 0.73% mAP gain for unseen categories under UV, RF-UC and NF-UC settings. Compared with HOICLIP [77], we observe consistent performance gains by 2.00%, 4.96% and 2.48% mAP for unseen categories under UV, RF-UC and NF-UC settings. These experimental results further demonstrate the effectiveness of our pre-trained weights.