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

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

1. Related Work: Zero-shot learning with VLM

CLIP [21] pioneers the research of building a vision-language model that can be used for zero-shot image classification, followed by other works with a different training scheme [7,22,27] or supervision signals [2,12,18,26]. After pre-training on hundreds of millions of web-crawled image-caption pairs, these models obtain the transfer ability to directly conduct inference on a wide range of downstream datasets. Recent research trends also push the boundaries of VLM applications into more challenging tasks such as zero-shot object detection [11,31] and instance segmentation [30]. In this paper, we further extend these efforts by exploring a new task - zero-shot HOI detection, with knowledge distillation from VLM for relationship understanding.

2. Spatial Feature Generation

For each pair of human and object proposals $\langle x_h, x_o \rangle$, we follow the similar pipeline as [28] to compute their spatial feature $v_{sp} \in \mathbb{R}^D$. This involves encoding the spatial information of their bounding boxes, including center coordinates, heights, widths, aspect ratios, and area sizes. All these values are normalized by the corresponding dimensions of the image. Additionally, we incorporate the intersection over union (IoU) to represent the pairwise relationships and characterize their distance. All these spatial cues are encoded with a multi-layer perceptron to obtain the spatial feature v_{sp}.

3. Ablation Studies of Multi-level Knowledge Distillation without CLIP Components

To isolate and analyze the impact of CLIP-oriented representation learning and CLIP supervision, we perform additional ablation studies. We begin with a simpler experimental setup, where our multi-branch network does not use CLIP visual and textual encoders. In this setup, both the visual encoder and HOI embedding are randomly initialized.

As shown in Table 1, all the results exhibit a substantial decrease compared to the results in the main paper (we replicated the results from Table 3 for a clear comparison). Besides, the mAP in non-rare classes is higher than in rare classes, even with the same supervision signals. This phenomenon demonstrates the integration of CLIP components into our model design facilitates the transfer of its generalization capability to HOI representation.

Table 1. **Ablation study of multi-level incorporation on HICO-DET dataset.** *Rand* means our model is randomly initialized, and *CLIP* indicates that the visual encoder and HOI embedding are provided by CLIP. The union branch is added with a late fusion strategy.

	Branch			mAP (%)		
	h-o	union	global	Full	Rare	Non-Rare
Rand	✓	-	-	6.31	5.21	6.63
	✓	-	✓	8.78	5.52	9.76
	✓	✓	✓	8.88	5.68	9.83
CLIP	✓	-	-	10.48	9.45	10.78
	✓	-	✓	15.84	17.91	15.21
	✓	✓	✓	**17.12**	**20.26**	**16.18**

4. Results comparison with existing works

In Table 2, we present a comprehensive set of results that encompasses a broader array of existing works in the realm of HOI detection.

5. More Visualizations

In Figure 1, we present additional HOI predictions following the same visualization process as Figure 4 in our main paper. We observe the same phenomena in both success and failure cases. Notably, our model excels in recognizing challenging HOIs, particularly when the human/object regions are small or occluded. This success can be attributed to the integration of CLIP, which enables our model to leverage contextual information and gain a better understanding of the surrounding environment.

Table 2. **Results comparison of different methods on HICO-DET test set**. †means re-implementation in [24]. Here FS, WS, and ZS indicate fully-supervised, weakly-supervised, and zero-shot HOI detection methods, respectively. The notation (D) means the visual encoder or the detector is pre-trained on dataset D, D∈ {COCO, HICO-DET, YFCC-15M}.

S	Methods	Visual Encoder	Detector	HICO-DET (%)		
				Full	Rare	Non-Rare
FS	InteractNet [5]	RN50-FPN (COCO)	FRCNN (COCO)	9.94	7.16	10.77
	iCAN [4]	RN50 (COCO)	FRCNN (COCO)	14.84	10.45	16.15
	TIN [15]	RN50-FPN (COCO)	FRCNN (COCO)	17.22	13.51	18.32
	PMFNet [25]	RN50-FPN (COCO)	FRCNN (COCO)	17.46	15.56	18.00
	DJ-RN [13]	RN50 (IN-1K&COCO)	FRCNN (COCO)	21.34	18.53	21.18
	IDN [14]	RN50 (IN-1K&COCO)	FRCNN (HICO-DET)	26.29	22.61	27.39
	SCG [28]	RN50-FPN (IN-1K&HICO-DET)	FRCNN (HICO-DET)	31.33	24.72	33.31
	HOTR [8]	RN50+Transformer (COCO)	DETR (HICO-DET)	25.10	17.34	27.42
	QPIC [23]	RN101+Transformer (COCO)	DETR (COCO)	29.90	23.92	31.69
	CATN [3]	RN50+Transformer (IN-1K&HICO-DET&COCO)	DETR (HICO-DET)	31.86	25.15	33.84
	MSTR [9]	RN50+Transformer (COCO)	DETR(HICO-DET)	31.17	25.31	33.92
	DisTr [32]	RN50+Transformer (IN-1K&COCO)	DETR (HICO-DET)	31.75	27.45	33.03
	IF [17]	RN50+Transformer	DETR (HICO-DET)	33.51	30.30	34.46
	CPC [20]	RN50+Transformer	DETR (COCO)	29.63	23.14	31.57
	SSRT [6]	R101+Transformer (COCO)	DETR (COCO)	31.34	24.31	33.32
	GEN-VLKT [16]	RN50+Transformer (HICO-DET)	DETR (HICO-DET)	33.75	29.25	35.10
	HOICLIP [19]	RN50+Transformer (HICO-DET)	DETR (HICO-DET)	34.69	31.12	35.74
WS	Explanation-HOI† [1]	ResNeXt101 (COCO)	FRCNN (COCO)	10.63	8.71	11.20
	MX-HOI [10]	RN101 (COCO)	FRCNN (COCO)	16.14	12.06	17.50
	PPR-FCN† [29]	RN50 (YFCC-15M)	FRCNN (COCO)	17.55	15.69	18.41
	PGBL [24]	RN50 (YFCC-15M)	FRCNN (COCO)	22.89	22.41	23.03
ZS	*baseline*	RN50 (YFCC-15M)	FRCNN (COCO)	10.48	9.45	10.78
	ours	RN50 (YFCC-15M)	FRCNN (COCO)	17.12	20.26	16.18

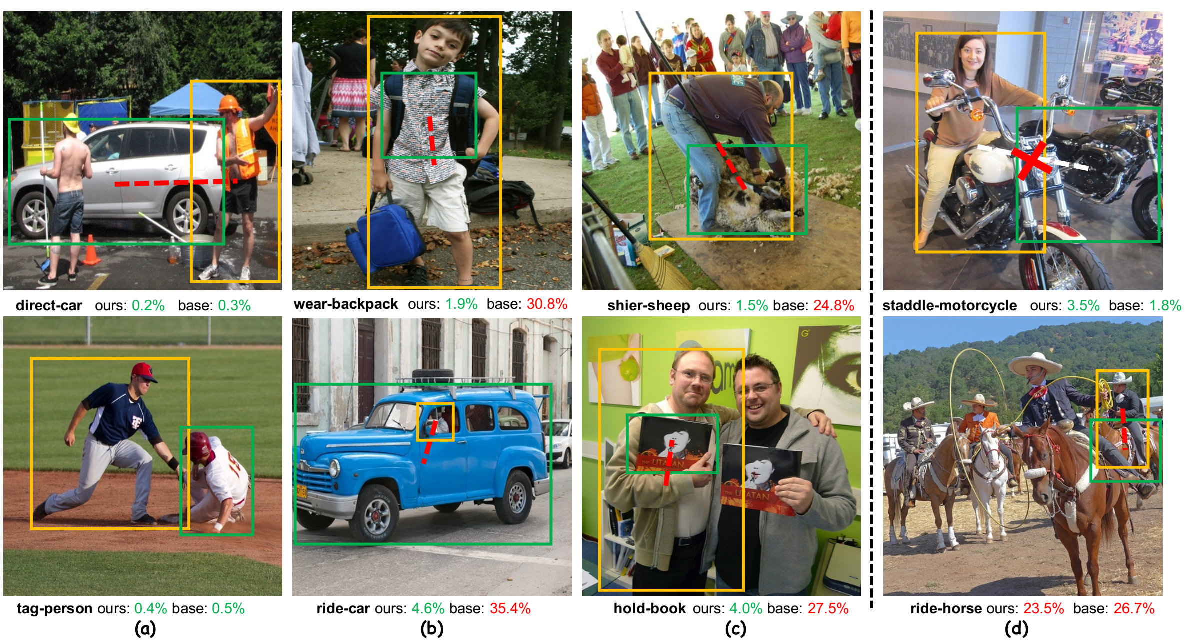

Figure 1. **More visualization of the HOI detection results.** Green percentiles signify the model's confident HOI predictions, and red percentiles denote the negative HOI predictions that the model treats as background.

References

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

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

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

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

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

[6] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. *arXiv preprint arXiv:2204.00746*, 2022.

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

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

[9] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. *arXiv preprint arXiv:2203.14709*, 2022.

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

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

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

[13] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020.

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

[15] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019.

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

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

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

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

[20] Jihwan Park, SeungJun Lee, Hwan Heo, Hyeong Kyu Choi, and Hyunwoo J Kim. Consistency learning via decoding path augmentation for transformers in human object interaction detection. *arXiv preprint arXiv:2204.04836*, 2022.

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

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

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

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

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

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

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

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

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

[30] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022.

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

[32] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. *arXiv preprint arXiv:2204.09290*, 2022.