Weakly-HOI: Multi-Modal Knowledge Transfer and Instance Relation Mining in Weakly Supervised HOI Detection

Yuxiao Wang, Xinyu Jiang*, Zhenao Wei, Yu Lei, Weiying Xue, Yanwu Xu, *Senior Member, IEEE*, Qi Liu†, *Senior Member, IEEE*

***Abstract*—Human-object interaction (HOI) detection refers to the task of extracting interactive relationships between humans and objects in a given image, serving as a crucial step toward a more comprehensive understanding of scene semantics. Although existing fully supervised methods can achieve excellent performance, they require huge manpower to annotate complex label information. In this work, we propose a novel weakly supervised detection method, denoted as Weakly-HOI, which exclusively leverages image-level annotations for efficient HOI detection, significantly reducing annotation costs. Specifically, Weakly-HOI assembles an instance localization and pairing module to construct candidate images representing all possible interactions. Besides, the cross-modal similarity matching module utilizes a text-image model to establish correspondences between actions and images effectively. Lastly, the image annotation transformation module systematically converts image-level labels into instance-level labels. Extensive experimental results validate the efficacy of our method, achieving xx mAP on the HICO dataset. This marks an improvement of xx mAP over the current state-of-the-art. Similarly, on the VCOCO dataset, our method also achieves the best performance, reaching xx mAP. Code will be available at https://drliuqi.github.io/.**

***Index Terms*—Human-object interaction, weakly supervised, object detection, CLIP.**

I. Introduction

HUMAN-OBJECT interaction (HOI) detection aims to precisely locate humans and objects within images of real-world scenes while categorizing the interactive relationships between humans and objects [1]. This task, centered around human-centric detection, plays a pivotal role in enhancing computer understanding of human behavior. Its real-world applications span various domains, including action recognition [2], [3], scene graph generation [4], and autonomous driving [5].

Currently, there are two types of HOI methods: two-stage [1], [6]–[14] and one-stage [15], [15]–[24]. Two-stage methods use object detection [25], [26] models to detect humans and objects before performing action classification.

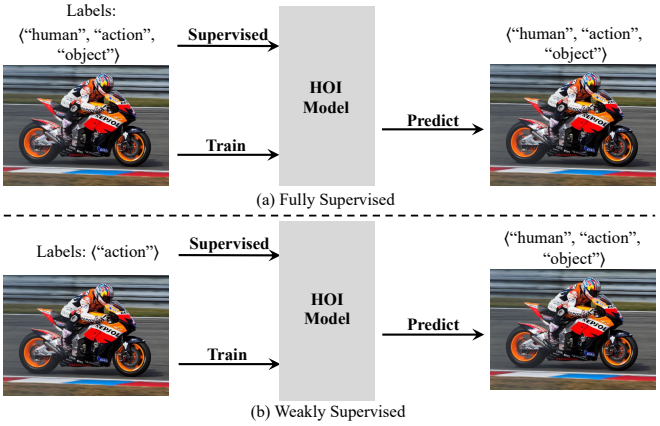

Fig. 1. Fully supervised HOI vs weakly supervised HOI. Fully supervised HOI involves training with ⟨“human”, “action”, “object”⟩ labels, requiring annotations for the positions of humans and objects, the category of objects, and the category of action. This process demands significant resources. Weakly supervised HOI trains using only ⟨“action”⟩ labels to predict complex HOI relationships.

However, these methods result in significant time consumption due to the separation of detection and classification. Moreover, the lack of coordination between detection and classification tasks can directly impact the method’s performance [22]. To track this issue, one-stage methods have been proposed. For instance, HOITrans [27] uses a transformer model with fully connected layers to directly output detection and classification results, thus improving model efficiency. Similar methods to HOITrans include QPIC [28] and Iwin [29]. Recently, some one-stage HOI methods have begun to explore the integration of textual information to enhance performance [30], [31]. For example, RPL [30] improves relation expression through phrase learning and label composition. Based on textual information, most methods have demonstrated that combining the Contrastive Language Image Pretraining (CLIP) [32] model can also assist HOI tasks [15], [22], [24], [33]. CLIP is a model capable of computing text-image similarity. For instance, GEN-VLKT [22] enhances performance by initializing its action classification network with weights obtained from CLIP, leading to a significant improvement in method performance. The HOI detection models mentioned above have predominantly been supervised, utilizing complex labels in the form of ⟨“human”, “action”, “object”⟩ triplets [15], as shown in Figure 1(a). However, the intricate nature of these labels

* Equal contribution as first authors
† Corresponding author: Qi Liu (drliuqi@scut.edu.cn)
Y. Wang (ftwangyuxiao@mail.scut.edu.cn), X. Jiang (202164020201@mail.scut.edu.cn), Z. Wei, W. Xue, Y. Xu (ywxu@ieee.org), and Q. Liu are with the School of Future Technology, South China University of Technology, China 511400.
Y. Lei is with the School of Information Science & Technology, Southwest Jiaotong University, China 611730 (e-mail:leiyu1117@my.swjtu.edu.cn).

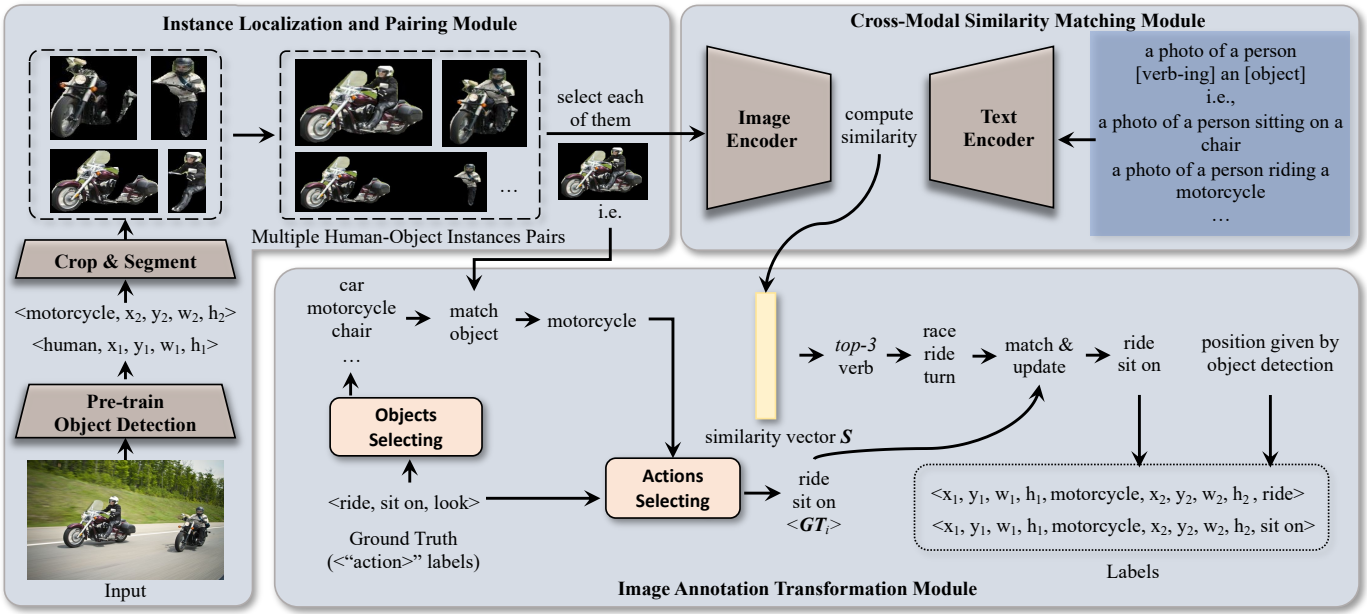

Fig. 2. Method overview. Given an image, the instance localization and pairing module is responsible for detecting and segmenting all instances, pairing humans and objects to create new candidate images. The cross-modal similarity matching module sequentially matches candidate images with a pre-defined HOI text template. The image annotation transformation module combines and transforms the action labels from ground truth with the information from the first two modules, ultimately outputting the ⟨"human", "action", "object"⟩ labels used for training. The "objects selecting" function is primarily responsible for identifying all objects related to the GT action labels. For example, in the case of "sit on", relevant objects could include "car", "motorcycle", and so on. The "actions selecting" function is tasked with filtering out action labels that correspond to the object category specified in the GT.

necessitates laborious annotation procedures, demanding considerable time and resources for handling large datasets [34].

To alleviate the annotation burden, methods of weakly-supervised learning have been proposed, leveraging only a small fraction of labels for training [34]–[36]. For example, AlignFormer [36] leverages image-level HOIs and box annotations, while others like MX-HOI [35] utilize image-level object and action information. Although these methods reduce the need for extensive data annotation, they still require manual annotation of object information. Later, VLHOI [34] solely relies on action labels to accomplish HOI tasks, as shown in Figure 1(b). It utilizes a vision-language model (VLM) and a large language model (LLM) to query possible interactions, thereby finishing HOI detection. However, VLHOI does not fully consider the action relationships between instances, resulting in lower accuracy.

Unlike other weak supervision methods, we seek to explore the latent relationships between humans and objects and the relationships between actions to remedy the negative impact caused by insufficiently considered implicit relationships. To this end, we first design an instance localization and pairing module to compose candidate images, which encompass all human-object action relationships. Then we employ the CLIP model for cross-modal text and image encoding pairs, facilitating the transfer of multi-modal knowledge to the HOI task and assisting in the exploration of human-object action relationships. In addition, we propose an image annotation transformation module to convert potential actions into ⟨"human", "action", "object"⟩ for training. Extensive experiments demonstrate that our method surpasses existing state-of-the-art methods on the HICO-Det and V-COCO datasets.

II. METHOD

Our method, termed Weakly-HOI, consists of three key modules: instance localization and pairing (ILP) module, cross-modal similarity matching (C-MSM) module, and image annotation transformation (IAT) module. The architecture of the proposed framework is shown in Figure 2. The Weakly-HOI aims to transform image-level annotation (⟨"action"⟩ labels) into instance-level annotation (⟨"human", "action", "object"⟩) for training. In brief, the instance localization and pairing module partitions the input image into multiple instance images based on the detection positions of humans and objects. Subsequently, different instance images are fed into the cross-modal similarity matching module as inputs to the image encoder. This process involves similarity matching with the text features fed into the text encoder, resulting in a similarity matrix. Finally, the image annotation transformation module extracts similarity-inferred action sets from the similarity matrix and then matches them with instance-level action sets to obtain the final annotated information for training.

A. Instance Localization and Pairing Module

In pursuit of automated label generation, our method commences with the utilization of the Yolov8[1] model for instances localization within the input image $\boldsymbol{I}$. This yields a set of instances bounding boxes $\boldsymbol{B} = \{\boldsymbol{b}_i | i = 1, 2, \ldots, N\}$, where each bounding box $\boldsymbol{b_i}$ encompasses the center coordinates (x_i, y_i), width W_i, height H_i, and the category c_i of the instance. Subsequently, we perform image cropping on the original image based on $\boldsymbol{b_i}$, generating a series of cropped images. Additionally, to mitigate the impact of redundant

[1]https://github.com/ultralytics/ultralytics

instances and background information on the model, instance segmentation techniques are used to process the cropped images. Through this step, we ensure that the model concentrates more effectively on the intricate features of humans and objects, providing clearer inputs for subsequent processes.

$$\boldsymbol{B} = \text{Yolov8}(\boldsymbol{I}). \tag{1}$$

Subsequently, the detected human and object instances are systematically paired, forming a set of candidate images $\boldsymbol{I}_C = \{\boldsymbol{I}_C^1, \boldsymbol{I}_C^2, \ldots, \boldsymbol{I}_C^{N_I}\}$. N_I denotes the number of candidate images, determined by the product of N_h and N_o, where N_h represents the number of humans and N_o represents the number of objects. N represents the total number of instances, i.e., $N_h + N_o = N$. This combinatorial method ensures that each candidate image encapsulates potential human-object interaction scenarios comprehensively.

B. Cross-Modal Similarity Matching Module

To explore potential relationships between humans and objects, we initialize the image encoder and text encoder in Figure 2 using the weights of the CLIP model. By computing the cross-modal similarity, we match the HOI relationship description texts with the encoded representations of the candidate images. Specifically, N_I candidate images are feed to the image encoder $\boldsymbol{F}_{IE}$ for processing, obtaining the image encoding $\boldsymbol{I}_E \in \mathbb{R}^{N_I \times C_{IE}}$, where C_{IE} represents the output dimensionality of the image encoder. The mathematical expression is given by:

$$\boldsymbol{I}_E = \boldsymbol{F}_{IE}(\boldsymbol{I}_C). \tag{2}$$

Next, text encoder $\boldsymbol{F}_{TE}$ is used for extracting the text features of the text template TT_1, resulting in the text information matrix $\boldsymbol{T}_{E1} \in \mathbb{R}^{N_T \times C_{TE}}$. Here, N_T represents the number of texts, corresponding to the number of HOI action categories, and C_{TE} is the output dimensionality of the encoded text. The text template TT_1 follows the format "a photo of a person verb-ing an object". For instance, the triplet ⟨"human", "ride", "motorcycle"⟩ is transformed into "a photo of a person riding a motorcycle".

Additionally, to emphasize the importance of verbs in HOI relationships, we construct another text template TT_2 with the format "a photo of a person verb-ing". The information matrices $\boldsymbol{T}_{E1}$ and $\boldsymbol{T}_{E2}$ corresponding to TT_1 and TT_2 are given by:

$$\boldsymbol{T}_{Ei} = \boldsymbol{F}_{TE}(\text{TT}_i), \quad i = 1, 2. \tag{3}$$

Finally, we compute the cosine similarity between the image encoding $\boldsymbol{I}_E$ and the text information $\boldsymbol{T}_{Ei}$ as follows:

$$\boldsymbol{simE}_i(\boldsymbol{I}_E, \boldsymbol{T}_{Ei}) = \frac{\boldsymbol{I}_E \cdot \boldsymbol{T}_{Ei}}{\|\boldsymbol{I}_E\| \cdot \|\boldsymbol{T}_{Ei}\|}, \quad i = 1, 2, \tag{4}$$

$$\boldsymbol{S} = \boldsymbol{simE}_1 + \boldsymbol{simE}_2, \tag{5}$$

where $\boldsymbol{I}_E \in \mathbb{R}^{N_I \times C_{IE}}$, $\boldsymbol{T}_{Ei} \in \mathbb{R}^{N_T \times C_{TE}}$, $C_{IE} = C_{TE}$, $\boldsymbol{simE}_i \in \mathbb{R}^{N_I \times N_T}$, and $\boldsymbol{S} \in \mathbb{R}^{N_I \times N_T}$. Through this process, we obtain the comprehensive similarity matrix $\boldsymbol{S}$, reflecting the degree of association between images and text.

C. Image Annotation Transformation Module

The core of the image annotation transformation module is to convert image-level annotations to instance-level annotations. Rather than directly considering the results of image-text similarity as the annotated actions for the current instance image, we incorporate an additional step of judgment and filtering through image encoder and text encoder. Based on the aforementioned reasoning and the similarity matrix $\boldsymbol{S} \in \mathbb{R}^{N_I \times N_T}$, we extract the top-3 most probable action for each candidate image, forming the candidate action set, also referred to as the similarity inference action set. Each instance candidate action follows the process outlined below:

$$\boldsymbol{K}_i = \texttt{sort\_indices}(\boldsymbol{S}_i), \tag{6}$$

$$\boldsymbol{A}_i = \{\boldsymbol{K}_i^j \mid j \in \{1, 2, 3\}\}, \tag{7}$$

where $\boldsymbol{S}_i \in \mathbb{R}^{N_T}$ represents the likelihoods of all possible actions occurring between human and object in the ith candidate image. The function `sort_indices` returns the indices that sort $\boldsymbol{S}_i$ in ascending order by value, denoted as $\boldsymbol{K}_i \in \mathbb{R}^{N_T}$. $\boldsymbol{A}_i \in \mathbb{R}^3$ represents the indices of the top-3 most probable actions for the ith candidate image.

In weakly supervised HOI, the true labels only include actions without specifying the particular object involved. This poses a challenge as the HOI model cannot learn the target task. Our objective is to match the detected instances with action labels, thereby completing HOI labels for training. Specifically, based on the object information from the current candidate image, we extract the ground truth action set $\boldsymbol{GT}_i$ from the image-level annotated data. Subsequently, we compute the intersection between $\boldsymbol{A}_i$ and $\boldsymbol{GT}_i$, denoted as:

$$\boldsymbol{M}_i = \boldsymbol{GT}_i \cap \boldsymbol{A}_i \tag{8}$$

Based on $\boldsymbol{M}_i$, we complete the HOI label for the candidate image as follows:

$$\boldsymbol{G}_i = \begin{cases} \boldsymbol{GT}_i, & \text{if } \boldsymbol{M}_i \neq \emptyset, \\ \emptyset, & \text{if } \boldsymbol{M}_i = \emptyset. \end{cases} \tag{9}$$

In other words, if there is an overlap between the similarity-inferred action set and the instance-level action set, we consider the actions from the instance-level set for the current candidate image. Otherwise, there is no human-object interaction in this candidate image.

Next, combining the object information $\boldsymbol{B}_o$ and human information $\boldsymbol{B}_h$ previously annotated by the detection model, we complete the annotation data for the current candidate image:

$$\boldsymbol{\Upsilon}_i = (\boldsymbol{B}_i^h, \boldsymbol{B}_i^o, C_i^o, \boldsymbol{G}_i). \tag{10}$$

In the instance localization and pairing module, we cropped the input image into multiple candidate images (pairing each detected human with each detected object). Therefore, it is necessary to integrate the labels of all candidate images to form instance-level annotation labels for the input image. This process is denoted as:

$$\boldsymbol{\Upsilon} = \{\boldsymbol{\Upsilon}_i \mid i \in [1, N_i], i \in \mathbb{Z}\}. \tag{11}$$

Using the above-mentioned method, we successfully transformed image-level labels into trainable instance-level labels.

III. EXPERIMENTS

A. Setup

Datasets and metrics. We conduct experiments on HICO-Det [1] and V-COCO [37] datasets to evaluate the effectiveness of our method. HICO-Det consists of 47,776 images, encompassing 117 action categories and 80 object categories, forming a total of 600 different HOI types. Evaluation of HICO-Det includes three scenarios: Full (600 HOI categories), Rare (HOI categories appearing less than 10 times), and Non-Rare (the remaining HOI categories). V-COCO comprises 10326 images, including 29 action categories and 80 object categories, forming a total of 263 HOI categories. The very common metrics mean Average Precision (mAP) is used to evaluate the model performance.

Parameter Details. We use GEN-VLKT as our base model, and the learning rate is set to 1e-4. Additionally, the AdamW optimizer is used to optimize the model, training for a total of 90 epochs. The experiments were conducted on the Ubuntu 20.04 system, using PyTorch version 1.7.1, and the model was trained on 8 A6000 (48G) GPUs with a batch size of 16.

B. Effectiveness for Regular HOI Detection

The experiments have been conducted on the HICO-Det, comparing them to the 10 recently published methods, and the results are tabulated in Table I. Our method achieves state-of-the-art performance in weakly supervised learning, surpassing VLHOI with a performance improvement of xx mAP in the Full scenario. Additionally, Weakly-HOI achieves xx mAP and xx mAP in the Rare and Non-Rare scenarios, respectively. Furthermore, even when using only ⟨"action"⟩ labels, our method outperforms other methods that use ⟨"action", "object"⟩ labels, demonstrating its robust capability.

Table II presents the experimental results of various methods on the V-COCO. Our method also attains state-of-the-art performance in weakly supervised learning, surpassing the second-best method with a xx mAP. Similarly, in the case of using only ⟨"action"⟩ labels, Weakly-HOI has also surpassed other methods that use ⟨"action", "object"⟩ labels.

C. Ablation Studies

To demonstrate the effectiveness of each component of Weakly-HOI, a series of ablation studies are performed on the HICO dataset. The results are listed in Table III. We first investigate the impact of preserving the background. The experiments demonstrate a significant improvement in mAP when deleting the background (Row 1 and Row 2 in Table III). The reason is that in most images, there are multiple human-object pairs. When annotating labels for a specific human-object pair, other human-object pairs can influence the calculation of action similarity. Additionally, we conduct experiments on the utilized text templates (TT_1 and TT_2). The experimental results indicate that using only TT_1 yields an average increase of xx mAP compared to using only TT_2 (Row 2 and Row 3 in table III). The optimal performance is achieved when both TT_1 and TT_2 are employed, with respective mAP values of xx, xx, and xx (Row 4 in Table III).

TABLE I
PERFORMANCE COMPARISONS ON HICO-DET DATASETS.

Method	Backbone	mAP↑		
		Full	Rare	Non-Rare
Fully supervised (using ⟨"human", "action", "object"⟩ labels)				
iCAN	RN50	14.84	10.45	16.15
VSGNet	RN152	19.80	16.05	20.91
SCG	RN50 FPN	21.85	18.11	22.97
IDN	RN50	23.36	22.47	23.63
HOTR	RN50	25.10	17.34	27.42
MSTR	RN50	31.17	25.31	33.92
Weakly+ supervised (using <"action", "object">labels)				
Explanation-HOI	RNXt101	10.63	8.71	11.20
MX-HOI	RN101	16.14	12.06	17.50
PPR-FCN	RN50	17.55	15.69	18.41
AlignFormer	RN50	19.26	14.00	20.83
Weakly supervised (using <"action">labels)				
SCG	RN50 FPN	7.05	-	-
VLHOI	RN50 FPN	8.35	-	-
Weakly-HOI	RN50	xx.xx	xx.xx	xx.xx

* Weakly+ supervised indicates training using ⟨"action", "object"⟩ labels, while weakly supervised implies training solely with ⟨"action"⟩ labels. Weakly supervised is more challenging, as it involves searching a broader interaction space due to the uncertainty of object categories. RN and RNXt represent ResNet and ResNetXt, respectively.

TABLE II
PERFORMANCE COMPARISONS ON V-COCO DATASETS.

Method	Backbone	mAP↑
Weakly+ supervised		
MX-HOI	RN101	-
AlignFormer	RN50	14.15
Weakly supervised		
SCG	RN50 FPN	20.05
VLHOI	RN50	17.71
VLHOI	RN50 FPN	29.59
Weakly-HOI	RN50	**39.52**

Similarly, experiments on the V-COCO dataset show that deleting the background and simultaneously using both TT_1 and TT_2 templates achieves optimal performance, reaching xx mAP in Table IV

The number of selected actions for each candidate image (denoted as $\boldsymbol{K}_i$ in Eq. 6) also affects the performance of the method. As shown in Eq. 6-9, the more actions select (i.e., larger $\boldsymbol{K}_i$), the higher the probability of having intersections with the ground truth ⟨"action"⟩ labels. This makes it easier to assign actions not belonging to the specific human-object pair, leading to a decrease in performance. Conversely, selecting fewer actions (smaller $\boldsymbol{K}_i$) may result in missing actions that actually belong to the pair, also impacting performance. Therefore, Table V presents the impact of the number of candidate actions on performance. The experiments show that the best performance is achieved when the candidate number is 3, with performance reaching xx mAP, xx mAP, and xx mAP in various scenarios of HICO-Det, and xx mAP on V-COCO.

D. Visualization

We visualized some results of our model on HICO-Det and V-COCO, shown in Figure 3 and Figure 4, respectively. Taking HICO-Det as an example, the first row visualizes different actions and objects. The second and third rows

TABLE III
EXPERIMENTAL RESULTS OF DIFFERENT COMPONENTS ON THE HICO-DET.

Row	Retain BG	Delete BG	TT_1	TT_2	mAP↑ Full	mAP↑ Rare	mAP↑ Non-Rare
1	✓		✓		xx.xx	xx.xx	xx.xx
2		✓	✓				
3		✓		✓			
4		✓	✓	✓			

* BG represents background.

TABLE IV
EXPERIMENTAL RESULTS OF DIFFERENT COMPONENTS ON THE V-COCO.

Row	Retain BG	Delete BG	TT_1	TT_2	mAP↑
1	✓		✓		xx.xx
2		✓	✓		
3		✓		✓	
4		✓	✓	✓	

depict scenarios with the same action and different objects, as well as different actions and the same object. The last row demonstrates that our method can also recognize images without human-object interaction actions. It can be observed that our method accurately identifies the interactive actions between humans and objects, as well as the positions of humans and objects in the images.

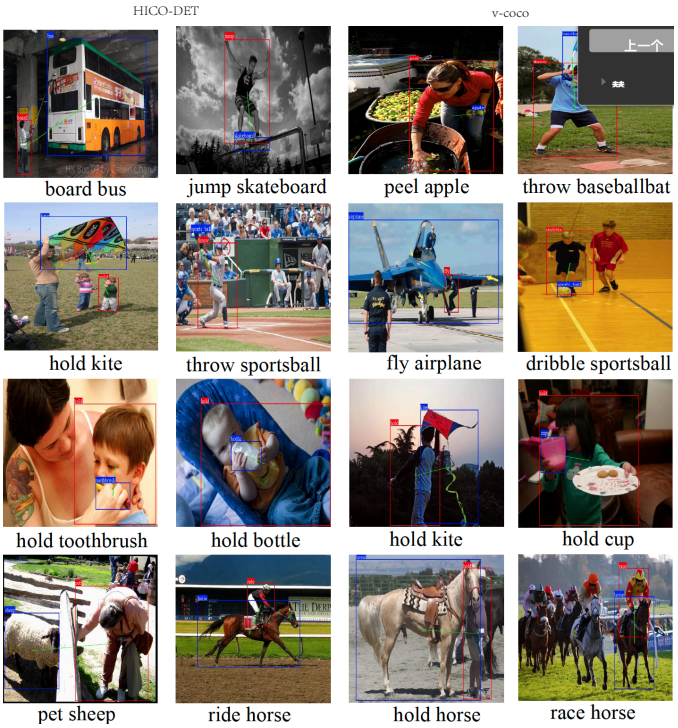

Fig. 3. HOI detection samples in HICO-Det.

IV. CONCLUSION

In this work, we introduce a novel weakly supervised method for HOI detection, referred to as Weakly-HOI. Weakly-HOI constructs multi-modal data to explore the latent relationships between humans and objects, as well as the relationships between actions. It completes the weakly supervised task by utilizing the proposed instance localization and pairing module, cross-modal similarity matching module, and image annotation transformation module to supplement the ⟨"human", "action", "object"⟩ labels for train, achieving the HOI task. In contrast to both fully supervised and other weakly supervised methods, our proposed method alleviates the need for intricate annotation labels. Furthermore, when compared to state-of-the-art methods, Weakly-HOI demonstrates substantial advantages in terms of performance.

TABLE V
EXPERIMENTAL RESULTS OF SELECTING DIFFERENT NUMBERS OF CANDIDATE ACTIONS.

Row	Top-num	mAP↑ (HICO-Det) Full	mAP↑ (HICO-Det) Rare	mAP↑ (HICO-Det) Non-Rare	mAP↑ (V-COCO)
1	Top-1	xx.xx	xx.xx	xx.xx	xx.xx
2	Top-2				
3	Top-3				
4	Top-4				

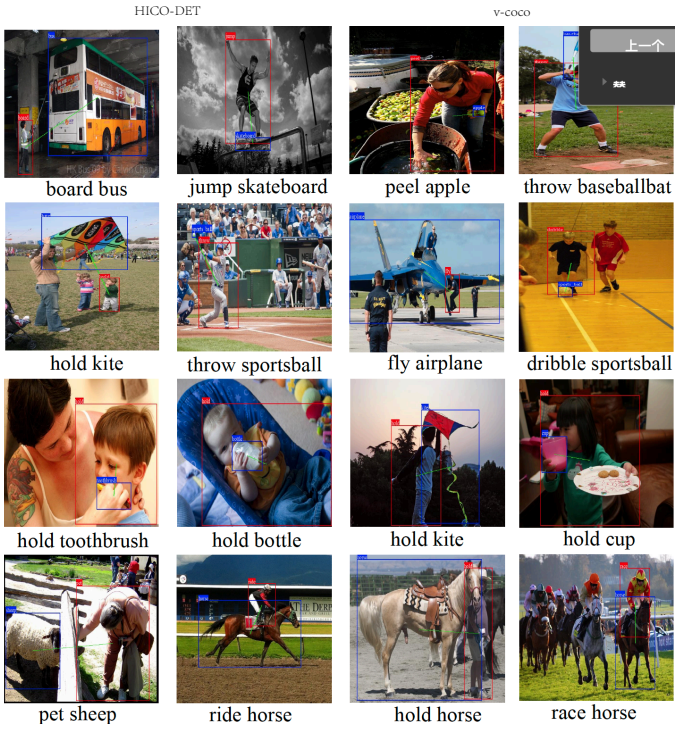

Fig. 4. HOI detection samples in V-COCO.

REFERENCES

[1] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 381–389.
[2] N. Wang, G. Zhu, H. Li, M. Feng, X. Zhao, L. Ni, P. Shen, L. Mei, and L. Zhang, "Exploring spatio–temporal graph convolution for video-based human–object interaction recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5814–5827, 2023.
[3] H. Fan, T. Zhuo, X. Yu, Y. Yang, and M. Kankanhalli, "Understanding atomic hand-object interaction with human intention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 275–285, 2022.
[4] Z. Fu, C. Zheng, J. Feng, Y. Cai, X.-Y. Wei, Y. Wang, and Q. Li, "DRAKE: Deep pair-wise relation alignment for knowledge-enhanced

multimodal scene graph generation in social media posts," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 7, pp. 3199–3213, 2023.

[5] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[6] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 104–20 112.

[7] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8359–8367.

[8] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 248–265.

[9] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 10 460–10 469.

[10] A. Iftekhar, S. Kumar, R. A. McEver, S. You, and B. Manjunath, "GTNet: Guided transformer network for detecting human-object interactions," *arXiv preprint arXiv:2108.00596*, 2021.

[11] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 69–85.

[12] Y. Liu, J. Yuan, and C. W. Chen, "ConsNet: Learning consistency graph for zero-shot human-object interaction detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4235–4243.

[13] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–417.

[14] H. Wang, L. Jiao, F. Liu, L. Li, X. Liu, D. Ji, and W. Gan, "Ipgn: Interactiveness proposal graph network for human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 6583–6593, 2021.

[15] Y. Wang, Q. Liu, and Y. Lei, "Ted-net: Dispersal attention for perceiving interaction region in indirectly-contact hoi detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[16] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4116–4125.

[17] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 234–13 243.

[18] B. Kim, T. Choi, J. Kang, and H. J. Kim, "UnionDet: Union-level detector towards real-time human-object interaction detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 498–514.

[19] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "DIRV: Dense interaction region voting for end-to-end human-object interaction detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1291–1299.

[20] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating HOI detection as adaptive set prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9004–9013.

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 825–11 834.

[22] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu, "GEN-VLKT: Simplify association and enhance interaction understanding for hoi detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 123–20 132.

[23] W.-K. Lin, H.-B. Zhang, Z. Fan, J.-H. Liu, L.-J. Yang, Q. Lei, and J. Du, "Point-based learnable query generator for human–object interaction detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 6469–6484, 2023.

[24] T. He, L. Gao, J. Song, and Y.-F. Li, "Toward a unified transformer-based framework for scene graph generation and human-object interaction detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 6274–6288, 2023.

[25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[26] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[27] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. 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*, 2021, pp. 74–83.

[28] M. Tamura, H. Ohashi, and T. 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*, 2021, pp. 10 410–10 419.

[29] D. Tu, X. Min, H. Duan, G. Guo, G. Zhai, and W. Shen, "Iwin: Human-object interaction detection via transformer with irregular windows," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 87–103.

[30] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1509–1517.

[31] A. Iftekhar, H. Chen, K. Kundu, X. Li, J. Tighe, and D. Modolo, "What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5353–5363.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[33] S. Ning, L. Qiu, Y. Liu, and X. He, "Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 507–23 517.

[34] M. E. Unal and A. Kovashka, "Vlms and llms can help detect human-object interactions with weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, url: https://asu-apg.github.io/odrum/posters_2023/poster_6.pdf*, 2023.

[35] S. K. Kumaraswamy, M. Shi, and E. Kijak, "Detecting human-object interaction with mixed supervision," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1228–1237.

[36] M. Kilickaya and A. Smeulders, "Human-object interaction detection via weak supervision," *arXiv preprint arXiv:2112.00492*, 2021.

[37] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.