2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00379

HOD: Human-Object Decoupling Network for HOI Detection

1st Hantao Zhang
Key Laboratory of Electromagnetic Space Information
University of Science and Technology of China
Hefei, China
zhanghantao@mail.ustc.edu.cn

2nd Shouhong Wan*
Key Laboratory of Electromagnetic Space Information
University of Science and Technology of China
Hefei, China
wansh@ustc.edu.cn

3rd Weidong Guo
Key Laboratory of Electromagnetic Space Information
University of Science and Technology of China
Hefei, China
gwd200@mail.ustc.edu.cn

4th Peiquan Jin
Key Laboratory of Electromagnetic Space Information
University of Science and Technology of China
Hefei, China
jpq@ustc.edu.cn

5th Mingguang Zheng
Key Laboratory of Electromagnetic Space Information
University of Science and Technology of China
Hefei, China
zhmg@mail.ustc.edu.cn

***Abstract*—Single-stage Human-Object Interaction (HOI) detection methods have attracted considerable attention due to their high efficiency. Existing methods tend to concentrate the detection of humans and objects in one decoder without considering the differences between them, which causes tremendous pressure on a single decoder and affects the detection effect. This paper aims to decouple the detection decoder of humans and objects. In particular, we advocate and propose a novel human-object decoupling network (HOD) that divides the decoder into three tasks: human detection, object detection, and action classification. The network uses the random erasure training strategy to improve the model's generalization ability and introduces pose features to handle the long-tailed problem. In addition, we design a pose fusion branch to alleviate the semantic gap between pose and HOI datasets. The experimental results suggest that our method achieves consistent improvements over the state-of-the-art across different datasets, utilizing only image information. Specifically for HICO-Det, our method outperforms existing methods by a large margin, with a significant relative mAP gain of 7.8%. Our source code will be publicly available upon acceptance.**

***Index Terms*—HOI, Task decoupling, Random erasure, Pose fusion**

I. Introduction

In these years, the relationship between humans and objects has attracted enormous attention due to its vital role in understanding static image high-level semantics. Human-Object Interaction (HOI) detection mainly aims to identify a set of HOI$< human, object, action >$triplets for a given image. Based on the process, the methods can be roughly viewed as single-stage and two-stage. Single-stage methods [1]–[13] aim to detect the HOI triplets directly, which improves the efficiency. In contrast, two-stage methods [14]–[22] try to obtain the human and object information at first and then obtain the interaction action classification, which brings additional complexity and slows detection speed. Thanks to the long-distance modeling capability of transformer in object detection [23], enormous single-stage work such as [5]–[12] have been developed to improve the HOI detection performance. However, due to the limitation of architecture, the single-stage HOI problem still faces the following challenges:

Challenge 1: *how to better obtain the location information of humans and objects.* Conventional HOI detection methods [1]–[4], [14]–[22] are mostly based on convolution operation to detect the tiny feature information in the interactions. However, the interactive parts in the images (such as people's hands and feet) often distribute discretely, and a considerable distance usually exists between them, which may enlarge the semantic gap and make the regional feature information hard to extract. To overcome this problem, basic transformer-based HOI methods [5]–[12] utilize the transformer [23] to capture correlation among objects. However, these methods concentrate the tasks of human-object pair detection and classification in one decoding module, which ignores the task characteristics of action classification and detection of human and object, causing heavy pressure on the decoder and affecting the detection effect.

Challenge 2: *how to balance the status between humans and objects.* In HOI detection, the importance of objects is usually lower than humans. In the experiments, we find that

This work is supported by Natural Science Foundation of Anhui Province (Grant No. 2208085MF157).

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00379

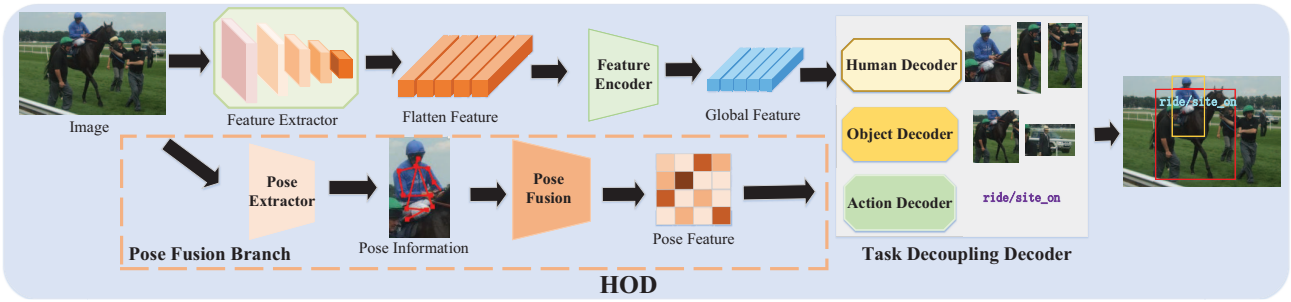

Fig. 1. The framework of our proposed human-object decoupling network (HOD)

excessive attention paid to the object will reduce the model's generalization and lead to poor performance. However, the current mainstream methods do not distinguish the importance between humans and objects.

Challenge 3: *how to effectively introduce human pose information into single-stage HOI detection architecture.* Pose information is essential to ease HOI detection's long-tailed problem. It is usually introduced to improve the accuracy of action classification for the two-stage methods [14], [18], [19], [21]. However, these methods usually extract people's location information and then obtain a single human's pose through estimation methods. This implementation form can not directly apply to a single-stage method and ignores the domain gaps between different datasets.

To overcome the above challenges, we propose a novel network named HOD, shown in Fig.1. **Our contributions are as follows: 1)** To help the model better locate the humans and objects, we decouple the decoder into the human decoder and object decoder, which can help different module focus on their tasks. Thus the effectiveness and accuracy can be improved. **2)** To improve the model's generalization and eliminate excessive dependence on objects, we propose a novel method to erase objects' information randomly. **3)** To better utilize pose information in single-stage architecture, we design the pose branch to integrate pose information in the network.

II. Related Work

Single-stage HOI detector: Mainstream single-stage HOI detectors can be viewed as an implicit two-stage process. The first stage aims to locate human-object pairs, while the second stage tries to integrate the correlation information to infer the HOI results. These methods are usually based on DETR [23] and detect the interactive human-object pair without further matching process. QPIC [8] is the first work that uses a query-based pairwise approach to detect the interactive human-object pairs as a whole. Based on QPIC, zhang et al. [9] uses the different decoders to decode human-object pairs and interaction classification. QAHOI [24] utilizes the deformable DETR [25] as the backbone to reduce the computation of the model and support larger-size feature maps. However, these detectors [1]–[12] always integrate the tasks related to human and objects in one module, which increases the task complexity and pressure of a single module on trade-off of multi-task learning, since the human detection and object detection are so various that different features are needed.

Pose information fusion: HOI detection is facing a long-tailed distribution problem due to the difficulty of collecting labeled training data for all HOI categories. Many works [11], [26]–[29] have been done to tackle this problem. Introducing pose information is an effective way to mitigate the long-tailed problem of HOI datasets and improve the identification performance of interaction for HOI detection. Many two-stage HOI detection methods [14], [18], [19], [21] have tried to utilize specific modules to obtain pose estimation at the first stage and then introduce it into HOI detection methods to improve the accuracy of action classification tasks at the second stage. However, these methods are usually limited by shortcomings: 1. Bounding box labels are usually unavailable for the single-stage model. 2. The mainstream pose estimation methods usually need a series of non-end-to-end post-processing before obtaining the last results, which destroys the architecture of single-stage HOI detectors.

III. Method

In this work, we mainly aim to further decouple the detection decoder of humans and objects and try to introduce pose information under the single-stage framework. The basic architecture of our proposed HOD is illustrated in Fig.1. Our framework divides the detection decoder into three tasks: human detection, object detection and action classification. Through these tasks decoupling, each module can focus on its own job. To improve the model's generalization, we randomly erase the object information in the action classification module to reduce the excessive dependence. Finally, to further improve HOI detection, we introduce human pose information into the framework utilizing the attention mechanism.

A. Human-Object Decoupling

The primary pipeline of task decoupling branch is shown in Fig.2. Our encoder-decoder architecture framework mainly consists of feature encoder and task decoupling decoder. Feature encoder aims to extract richer global contextual information from the original picture. The task decoupling module utilizes our proposed decoder to refine the classification of people, objects, and actions and then gives the final HOI

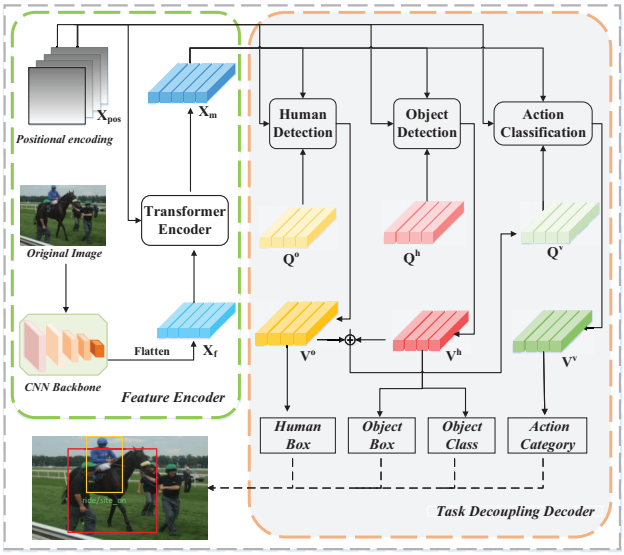

Fig. 2. The pipeline of human-object decoupling branch.

results. These essential parts are detailedly introduced as follows.

Feature Encoder: The feature encoder mainly consists of a CNN feature extraction network and a transformer encoder. For an input image I with shape (H, W, C), it is firstly fed into a CNN backbone network to get the feature $F \in R^{H' \times W' \times C'}$ of shape (H', W', C'), then a projection convolution layer with the kernel size of 1×1 is used to generate the flatten feature $X_f \in R^{(H' \times W') \times C'}$. After that, the flatten feature and positional encoding $X_{pos} \in R^{(H' \times W') \times C'}$ are fed into a transformer encoder where we use its multi-head self-attention mechanism to obtain richer global contextual information feature $X_m \in R^{(H' \times W') \times C'}$.

Task Decoupling Decoder: We divide the basic interaction detection task into three sub-tasks: human detection, object detection, and action classification. Moreover, we design an independent decoder module for each task to reduce the whole task's burden. Specifically, we propose a novel human detection decoder and object detection decoder based on the framework of CDN [9]. Each task decoder owns its unique self-attention module and multi-head co-attention module. The calculation process of task decoupling is as follows Eq.(1), where V^h, V^o, and V^v are the sequenced visual features obtained through the decoder of human detection f_h, object detect detection f_o and action classification f_v, respectively. Then V^h, V^o, and V^v are passed through their own full connection layer FC_h, FC_h, and FC_v to obtain the final HOI$< human, object, action >$triplets.

$$\begin{aligned} V^h &= f_h(X_m, X_{pos}, Q^h) \\ V^o &= f_o(X_m, X_{pos}, Q^o) \\ V^v &= f_v(X_m, X_{pos}, Q^v) \\ HOI &=< FC_h(V^h), FC_o(V^o), FC_v(V^v) > \end{aligned} \tag{1}$$

Specifically, each task utilizes its series of query vectors $Q^h, Q^o, Q^v \in R^{N_d \times C}$ where N_d indicates the number of vectors. For Q^h and Q^o, a series of randomly trainable vectors are used as initialization. The initialization for Q^h and Q^o are entirely independent, which enables the decoupling decoder to focus on extracting the features suitable for their respective tasks according to different query vectors. In this way, the performance decay caused by various distributions of people and objects can be alleviated to a certain extent. The contextual information features X_m and positional encoding X_{pos} are both fed into each decoder as input. It should be noted that Q^v for action classification is obtained through the fusion between V^h and V^o, which calculates as follows Eq.(2). In this way, the obtained fusion query vector Q^v can fully use the information of people and objects to guide the action classification module to extract the interactive features.

$$Q^v = V^h + V^o \tag{2}$$

After these three task modules decoding, the sequenced visual features $V^h, V^o, V^v \in R^{N_d \times C}$ are finally obtained. Then, the output value vectors are passed through their full connection layers to obtain the human bounding box, object bounding box, and action category. The complete output is obtained by combining the above results.

B. Random Erasure Strategy:

One thing to note is that multiple information about people, such as hand or foot movements, posture, and other information, always play a vital role in judging the action category. Excessive attention paid to the object will reduce the model's generalization due to weakening the influence of human information. Hence, we propose a random erasure strategy to erase part information of objects in the training process randomly. Specifically, for the action classification module in the task decoupling decoder, we use the probability of α making $Q^v = V^h$ to erase its object information and the probability of $1 - \alpha$ making $Q^v = V^h + V^o$ to use the complete human and object information. By randomly erasing the object information in Q^v, our proposed model is forced to judge the action category more accurately when the object information is missing, thus reducing the dependence of the model on object information and improving its generalization ability.

C. Pose Fusion Branch

In this work, we propose a novel pose fusion subnetwork utilizing heatmap to obtain pose estimation from bottom to up, which breakthroughs the limitation of necessity for human bounding box. The pose fusion subnetwork mainly consists of the feature interaction module and pose attention module, which is shown in Fig.3. In our design, we use HrHRNet [30] as the backbone of our pose extraction network. The HrHRNet finally outputs the pose key points heatmap, label prediction map, and scale prediction map. In our work, we only retain the pose key points heatmap and focus on the fusion of pose information. It should be noticed that the existing HOI

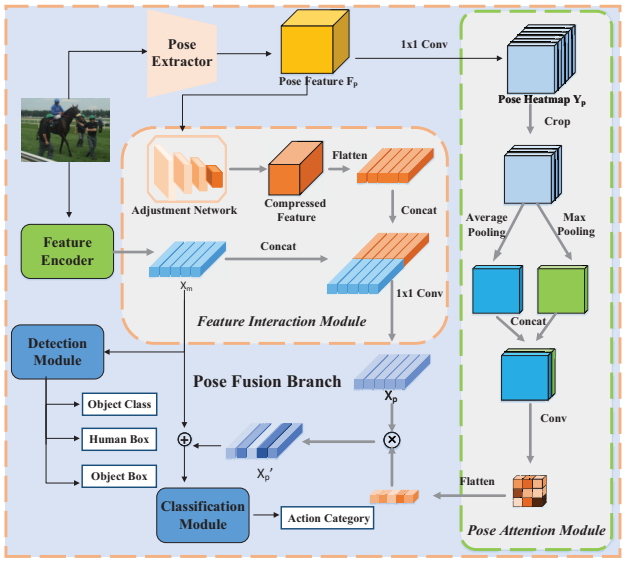

Fig. 3. The framework of pose fusion branch.

detection datasets do not contain pose information labels which means the pose estimate network HrHRNet can not be trained in the target source. Therefore, to avoid contaminating the pose estimation network, all parameters of HrHRNet are frozen. Moreover, we design the feature interaction and pose attention modules, mainly aiming to alleviate the semantic gap between the pose datasets and HOI datasets. Next, we will detailedly introduce these two modules.

Feature Interaction Module: The pose information extraction network is pre-trained on the pose key point datasets. Thus, the pose feature F_P obtained through the pose extractor contains more pose semantic information. To fuse pose information and alleviate the semantic gap between F_P and X_m, we first use an adjustment network to pull closer semantic information and adjust the size of F_P to be the same as the mainstream feature X_m. The adjustment network consists of one layer of max pooling and three layers of convolution operations. Then, the two features are concatenated together, and one 1×1 convolution is used to fuse the feature and obtain the pose heatmap Y_p.

Pose Attention Module: The most critical features of HOI discrimination are usually located in the posture key points of human hands, feet, knees, and other places. To capture the significant feature in HOI, we design a pose attention module shown in Fig.3. It is noticed that redundant key points may also interfere with the detection, so in practice, we only retain the key points heatmap feature of eyes, shoulders, wrists, hips, and feet which is cropped from Y_p. After obtaining the pose attention map, the final feature X_p' is obtained through the dot product between pose attention and X_p. Finally, to keep the training process stable, the final feature fed in classification is through the residual processing with the origin global feature X_m.

IV. EXPERIMENTS

Methods	Backbone	T	F	R	NR
Two-stage methods					
TIN [14]	R50		17.2	13.5	18.3
PMFNet [15]	R50-FPN	✓	17.4	15.6	18.0
CHG [16]	R-0		17.5	16.8	17.7
VSGNet [17]	R152		19.8	16.0	20.9
FCMNet [18]	R50	✓	20.4	17.3	21.5
ACP [19]	R152	✓	20.5	15.9	21.9
PD-Net [20]	R152	✓	20.8	15.9	22.2
DJ-RN [21]	R50		21.3	18.5	22.1
IDN [22]	R50		23.3	22.4	23.6
One-stage methods					
IPNet [1]	HG		19.5	12.7	21.5
DIRV [2]	EDet-d3		21.7	16.3	23.3
PPDM [3]	HG		21.7	13.7	24.1
GG-Net [4]	HG		23.4	16.4	25.6
HOI-Trans [5]	R50		23.4	16.9	25.4
HOTR [6]	R50		25.1	17.3	27.4
AS-Net [7]	R50		28.8	24.2	30.2
QPIC [8]	R101		29.9	23.9	31.6
CDN-S [9]	R50		31.4	27.3	32.6
CDN-L [9]	R101		32.0	27.1	33.5
Iwin-S(2022) [10]	R50		24.3	18.5	26.0
Iwin-L(2022) [10]	R101		32.0	27.6	34.1
PHOI-S(2022) [11]	R50	✓	29.2	22.0	31.4
PHOI-L(2022) [11]	R101	✓	30.0	23.4	31.9
IPAD-B(2023) [12]	R50		30.5	22.7	33.8
HOD-S(Ours)	R50		**33.2**	**28.4**	**34.6**
HOD-L(Ours)	R101		**34.5**	**30.4**	**35.7**

TABLE I
PERFORMANCE COMPARISON ON THE HICO-DET TEST SET. THE 'TEXT' REPRESENTS WHETHER TO UTILIZE EXTRA LANGUAGE INFORMATION. 'T', 'F', AND 'NR' REFER TO FULL, RARE, AND NO-RARE SET OF HOI CLASSES RESPECTIVELY.

A. Datasets and Evaluation Metrics

Comprehensive experiments are conducted on two widely-used HOI detection benchmarks: HICO-DET [31] and V-COCO [32] to demonstrate the superiority of our designed HOD. For both datasets, one person can interact with multiple objects at the same time. For HICO-DET, we follow the standard evaluation as [9], [31] and use the mean average precision (mAP) as the evaluation metric. Both full set and rare set of HOI classes' results are in detailly reported. For V-COCO, we report two scenarios mAP as [9]: scenario 1 includes all cases even without any objects, and scenario 2 ignores these cases. Due to the lack of pose key points label in HICO-DET and V-COCO, pose extraction network HrHRNet [30] is pre-trained on COCO [33].

B. Implementation Details

We implemented our HOI detection framework HOD with Python 3.7.11 and Pytorch 1.7.0 on 8 NVIDIA RTX 3090 GPU cards (8×24 GB memory) for all experiments with a batch size of 24. Two variant architectures of HOD: HOD-S and HOD-L, are implemented, where 'S' and 'L' denote small and large, respectively. We adopt the same configuration as CDN [9] for a fair comparison. Specifically, HOD-S utilizes ResNet-50 with a 3-layer transformer for both the encoder and decoder. HOD-L is equipped with ResNet-101 and 6-layer transformers for both the encoder and decoder. Other hyperparameters are the same as CDN. More implementations and results are put in the Appendix.

Methods	Backbone	Text	S1	S2
Two-stage methods				
TIN [14]	R50		54.2	-
PMFNet [15]	R50-FPN	✓	52.0	-
CHG [16]	R50		52.7	-
VSGNet [17]	R152		57.0	-
FCMNet [18]	R50	✓	53.1	-
ACP [19]	R152	✓	53.23	-
PD-Net [20]	R152	✓	52.6	-
IDN [22]	R50		53.3	60.3
One-stage methods				
UnionDet [34]	R50-FPN		47.5	56.2
IPNet [1]	HG		51.0	-
HOI-Trans [5]	R50		52.9	-
GG-Net [4]	HG		54.7	-
DIRV [2]	EDet-d3		56.1	-
HOTR [6]	R50		55.2	64.4
AS-Net [7]	R50		53.9	-
QPIC [8]	R101		58.3	60.7
CDN-S [9]	R50		61.6	63.7
CDN-L [9]	R101		63.9	65.8
IPAD-B(2023) [12]	R50		62.5	-
Iwin-S(2022) [10]	R50		51.8	-
Iwin-L(2022) [10]	R101		63.9	-
PHOI(2022) [11]	R50	✓	57.4	-
HOD-S(Ours)	R50		**61.6**	**63.8**
HOD-L(Ours)	R101		**64.0**	65.7

TABLE II
PERFORMANCE COMPARISON ON THE V-COCO TEST SET. THE 'TEXT' REPRESENTS WHETHER TO UTILIZE EXTRA LANGUAGE INFORMATION. 'S1' AND 'S2' INDICATE SCENARIO 1 AND SCENARIO 2, RESPECTIVELY.

C. Comparison to State-of-the-Art

In this work, we use the official evaluation code to verify the effectiveness of our method. The mAPs are computed for both HICO-Det and V-COCO. The comparisons of our proposed HOD with the recent HOI detection methods are shown in Table I and Table II where the 'Text' indicates whether extra language information is utilized. For backbone, 'R', 'HG', and 'EDet' refer to ResNet, Hourglass, and EfficientDet, respectively. For model scale, 'B' denotes that some methods [12] utilize ResNet-50 with a 6-layer transformers (double blocks than 'S').

For the HICO-Det dataset, as shown in Table I, our HOD-S outperforms all existing single-stage and two-stage HOI detection methods that only utilize the image without additional information. It should be noted that even just utilizing the small-scale configurations, our proposed HOD-S achieves a 1.2 mAP gain than CDN-L with large-scale configurations. Moreover, HOD-L, which utilizes the same configurations as CDN-L, achieves a relative 7.8% mAP gain with a margin of mAP 2.5. Especially for the rare categories, HOD-L achieves 30.4 mAP, which achieves a relative 12.2% mAP growth with a margin of mAP 3.3. This indicates that our design of human-object decoupling and pose fusion effectively alleviates the long-tailed problem. For V-COCO, as shown in Table II, our proposed HOD also achieves the state-of-the-art with the same configurations. However, due to the limitation of insufficient samples of V-COCO compared to HICO-Det to train such a large number of 263 categories classification, our decoupling decoder and attention module can be trained effectively. Hence, the promotion is not as significant as that on HICO-Det.

Setting	Decoupling	Erase	Full	Rare	NonRare
H-O-A			28.92	20.86	31.33
H-O, A	✓		31.25	26.04	32.81
H-A, O	✓		30.94	26.76	32.19
O-A, H	✓		29.04	23.05	30.83
H, O, A	✓		31.79	27.29	33.13
H, O, A	✓	✓	**33.19**	**28.37**	**34.63**

TABLE III
ABLATION STUDIES OF DECOUPLING AND TRAINING STRATEGY ON THE HICO-DET TEST SET WITH SMALL MODEL (HOD-S).

Pose	Attention	Feature Interaction	Full	Rare	NonRare
			33.28	29.10	34.52
✓	✓		33.63	28.14	35.27
✓		✓	33.89	28.38	35.54
✓	✓	✓	**34.48**	**30.38**	**35.71**

TABLE IV
ABLATION STUDIES OF POSE FUSION ON THE HICO-DET TEST SET WITH LARGE MODEL (HOD-L).

D. Ablation Study

This subsection analyzes the effectiveness of our proposed strategies and components in detail. All experiments are eventuated on the HICO-Det dataset to keep the results' stability and reproducibility. The ablation experiments mainly consist of the following two parts:

Task Decoupling and Random Erasure Strategy: As shown in Table III, we first carry out experiments based on the HOD-S with ResNet-50 as the backbone to prove the superiority of our proposed human-object decoupling and randomly erase strategy. In Table III, 'H', 'O', and 'A' respectively indicate human, object, and action. Moreover, '-' refers to the corresponding parts decoded in the same module. The decoupling ablation results show that too many tasks centralized on one module will give much pressure on the decoder and thus hinder the performance. Through the experiments, we find that human, object, and action all decoupling can achieve the best results, causing a 2.87 mAP improvement. Randomly erasing strategy can force the model to pay more attention to the human, which achieves an additional 1.4 mAP improvement.

Pose Fusion: As shown in Table IV, ablation experiments are carried out on the pose attention module and feature interaction module. In experiments, we find that our HOD-L model equipped with a 6-layer transformers decoder has enough ability to use the pose feature fully. Hence, the ablation experiments in this part mainly carry out based on the HOD-L with ResNet-101. These results indicate that pose information is also vital to HOI detection for query-based transformer architecture. Moreover, the pose feature's power is limited due to the semantic gap between origin pose label datasets and HOI datasets. Our proposed pose attention module and feature interaction module can alleviate this problem to a certain extent and achieve a 1.2 mAP gain.

V. CONCLUSION

In this paper, we explore the different decoupling ways of multi-task learning and propose a novel single-stage framework that decouples the human-object decoder and intro-

duces the pose feature into end-to-end architecture. Our HOD keeps the independent detection decoder process of humans and objects while utilizing a novel random erasure training strategy to improve the model's generalization. The proposed approach significantly improves the state-of-the-art results in HOI datasets and alleviates the long-tailed problem to some extent. In the future, we aim to refine the fusion of pose for HOI detection and explore the multi-information fusion strategy.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of Anhui Province (Grant No. 2208085MF157).

REFERENCES

[1] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian 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.

[2] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu, "Dirv: Dense interaction region voting for end-to-end human-object interaction detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 1291–1299.

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

[4] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng 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. 13234–13243.

[5] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen 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. 11825–11834.

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

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

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

[9] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li, "Mining the benefits of two-stage and one-stage hoi detection," *arXiv preprint arXiv:2108.05077*, 2021.

[10] D. Tu, X. Min, H. Duan, G. Guo, G. Zhai, and W. Shen, "Iwin: Human-object interaction detection via transformer with irregular windows," *European Conference on Computer Vision*, 2022.

[11] Zhimin Li, Cheng Zou, Yu Zhao, Boxun Li, and Sheng Zhong, "Improving human-object interaction detection via phrase learning and label composition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, pp. 1509–1517.

[12] Lin Bai, Fenglian Chen, and Yang Tian, "Automatically detecting human-object interaction by an instance part-level attention deep framework," *Pattern Recognition*, vol. 134, pp. 109110, 2023.

[13] Zeyu Ma, Ping Wei, Huan Li, and Nanning Zheng, "Hoig: end-to-end human-object interactions grounding with transformers," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

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

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

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

[17] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13617–13626.

[18] Yang Liu, Qingchao Chen, and Andrew Zisserman, "Amplifying key cues for human-object-interaction detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 248–265.

[19] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon, "Detecting human-object interactions with action co-occurrence priors," in *European Conference on Computer Vision*. Springer, 2020, pp. 718–736.

[20] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao, "Polysemy deciphering network for human-object interaction detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 69–85.

[21] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10166–10175.

[22] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu, "Hoi analysis: Integrating and decomposing human-object interaction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5011–5022, 2020.

[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

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

[25] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[26] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa, "Detecting human-object interactions via functional generalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 10460–10469.

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

[28] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao, "Visual compositional learning for human-object interaction detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 584–600.

[29] Guangzhi Wang, Yangyang Guo, Yongkang Wong, and Mohan Kankanhalli, "Chairs can be stood on: Overcoming object bias in human-object interaction detection," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer, 2022, pp. 654–672.

[30] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou, "Rethinking the heatmap regression for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13264–13273.

[31] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng, "Learning to detect human-object interactions," in *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2018, pp. 381–389.

[32] Saurabh Gupta and Jitendra Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.

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

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