ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446999

HOICS: ZERO-SHOT HOI DETECTION VIA COMPATIBILITY SELF-LEARNING

Miao Jiang[⋆†] *Min Li*[⋆†] *Junxing Ren*[⋆] *Weiqing Huang*[⋆†]

[⋆] Institute of Information Engineering, Chinese Academy of Sciences
[†]School of Cyberspace Security, University of Chinese Academy of Sciences

ABSTRACT

In recent years, the computer vision community has increasingly focused on detecting and recognizing human-object interactions (HOIs), which are crucial for tasks such as action recognition and scene understanding. However, the imbalanced distribution of interaction categories presents challenges, especially in scenarios with limited data. To address this issue, we propose an innovative end-to-end parallel HOI detection framework that incorporates compatibility self-learning (HOICS) for zero-shot HOI detection. Specifically, our approach introduces a compatibility self-learning strategy that allows the model to extract insights from HOI compatibility data and refine the interaction prediction head. This method involves utilizing combination probabilities and compatibility scores during the self-learning process, which significantly enhances zero-shot detection performance. Additionally, HOICS employs mosaic augmentation to broaden the model's capabilities. Our extensive experiments on benchmark datasets demonstrate HOICS' superiority over state-of-the-art methods, achieving a noteworthy 2.16 percent improvement in zero-shot HOI detection for previously unseen categories.

Index Terms— human-object interaction, zero-shot detection, self-learning

1. INTRODUCTION

Computer vision plays a pivotal role in comprehending and interpreting visual data, and recent years have witnessed remarkable progress fueled by the rapid advancements in artificial intelligence technologies. Among the multifarious tasks encompassed within computer vision, the detection and recognition of human-object interactions (HOIs) have garnered substantial attention. HOI detection seeks to empower computers to locate and analyze interactions between humans and objects, represented as <human, interaction, object> triplets. This capability provides invaluable insights into human behavior and fosters a deeper understanding of visual scenes. Its significance spans across various domains, including action recognition, scene comprehension, and human-computer interaction, to name a few. By precisely detecting and recognizing humans, objects, and their interactions, HOI detection augments the intelligence and practicality of computer vision systems, enabling them to decipher complex visual scenes and derive meaningful insights.

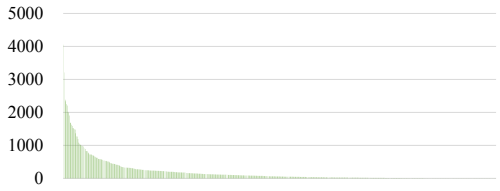

Fig. 1. Illustration of distribution of the number of HOI instances in HICO-DET dataset. The x-axis represents the order of HOI categories after sorting them based on the number of instances.

The rapid evolution of deep learning techniques [1] and the availability of extensive annotated datasets [2, 3] have significantly propelled the advancement of HOI detection [4, 5, 6, 7]. Nevertheless, the categories of interaction behaviors in HOI detection suffers from a pronounced imbalance. The distribution of HOIs is visually depicted in Fig. 1, illustrating the long-tailed number distribution. Conventional methods typically employ multi-label classifiers trained on datasets and grapple with the challenge of a long-tailed distribution. This issue notably hampers the performance of few-shot and zero-shot HOI detection [8]. To address this challenge, significant endeavors have been invested in HOI detection to explore long-tail HOI categories, including few-shot and zero-shot scenarios. Some approaches [9, 10, 11, 12, 13] leverage semantic knowledge to augment unseen interaction or object categories, while others [14, 15, 8, 16] employ compositional techniques to generate novel HOI categories. In alignment with the latter approach, we tackle the long-tailed HOI detection challenge from the perspective of HOI composition.

However, these methods adopt a sequential approach where object detection is initially performed, followed by the association of <human, object> pairs using a subsequent neural network. This sequential process is not only time-consuming but also computationally intensive [4]. Moreover, existing composition methods [8] augment the input feature

This project was funded by the National Key Research and Development Program of China (No. 2021YFB2910109).

space and necessitate breaking the pipeline of the original sequential methods. In this paper, we follow the concept of composition methods but introduce an innovative end-to-end parallel structure to address zero-shot HOI detection.

Our primary contributions can be summarized as follows: (1) We introduce an innovative end-to-end parallel approach named HOICS, marking the first utilization of such a framework to address zero-shot HOI detection. (2) We introduce the concept of a combination probability matrix to represent the likelihood of interaction and object categories combining or interacting to form an HOI category. HOI compatibility scores are proposed to autonomously learn this matrix. (3) We propose a compatibility self-learning strategy that empowers the model to glean insights from HOI compatibility information, facilitating the refinement of the interaction prediction head. (4) We conduct thorough experiments on two benchmark datasets, demonstrating the superiority of our approach over state-of-the-art composition methods.

2. METHOD

In this section, we provide an in-depth discussion of our proposed model. To begin, we outline the comprehensive network architecture. Next, we introduce the concepts of combination probability and compatibility scores. Following that, we delve into the details of our compatibility self-learning strategy tailored for zero-shot HOI detection. Finally, we elucidate the training and inference processes.

2.1. Overview Architecture

We present HOICS, an end-to-end HOI detection framework designed to tackle long-tailed and zero-shot HOI detection challenges. As illustrated in Fig. 2, HOICS employs a transformer encoder-decoder structure with a shared encoder and two parallel decoders. The shared encoder, in conjunction with a backbone CNN, extracts global context from the input image. Two sets of queries are then fed into the parallel decoders. The instance decoder handles instance queries for object detection, while the interaction decoder processes interaction queries for interaction detection. Using feed-forward networks (FFNs), we process interaction representations to obtain HO pointers and identify interaction categories. These outputs are combined with HO pointers to generate final HOI triplets. HOICS incorporates compatibility self-learning after the FFNs, enabling autonomous learning of HOI combination probabilities and refining the interaction prediction head through compatibility information. The subsequent section provides more details on compatibility self-learning.

2.2. Compatibility Self-Learning

We build upon the concept of visual compositional learning. However, in contrast to prior approaches [8, 16], we perform composition without prior knowledge of the labels of the instances being composited. Due to our adoption of an end-to-end parallel structure rather than a sequential one, we lack access to their feature representations. Instead, we introduce combination probabilities and compatibility scores, leveraging the compatibility self-learning strategy to address the zero-shot issue.

2.2.1. Combination Probability

We maintain an HOI combination probability matrix, denoted as $M \in R^{N_v \times N_o}$, during training. Here, N_v and N_o represent the numbers of interaction and object categories, respectively. Each entry in matrix M signifies the probability of combining a specific interaction and object category to form an HOI category. In this context, "combination probability" refers to the likelihood of pairing an interaction category with an object category to create an HOI category, rather than merging interaction and object representations. A simple approach to creating matrix M assigns a probability of 1 to combinations observed in the labeled training data and a probability of 0 to all other entries. However, we aim to enable the network to autonomously learn this matrix. To achieve this, we introduce compatibility scores as part of our approach.

2.2.2. Compatibility Scores

We introduce HOI compatibility scores to quantify the likelihood of interaction representations being associated with object representations. As shown in Fig. 2, we use FFNs to process instance and interaction representations, resulting in object predictions $\hat{Y}_o \in R^{Q_o \times N_o}$ and interaction predictions $\hat{Y}_v \in R^{Q_v \times N_v}$, where Q_o and Q_v represent the numbers of instance and interaction queries, respectively. To encompass all instance and interaction queries in a mini-batch, we compute composite HOI predictions $\hat{Y}_h = \hat{Y}_v \otimes \hat{Y}_o$, where $\hat{Y}_h \in R^{Q_v Q_o \times N_v \times N_o}$. Here, $Q_v Q_o$ represents the total number of potential HOI combinations, including unseen HOI categories. $\hat{Y}_h$ is also denoted as S_h, representing the HOI compatibility scores for that batch. Each $S_h(i,j,k)$ signifies the compatibility score between the j-th interaction category and the k-th object category in the prediction of the i-th composite HOI.

2.2.3. Update Combination Probability

We update M as follows,

$$M \leftarrow \frac{c \cdot M + \sum_i^{Q_v Q_o} S_h(i,:,:) \odot M_f}{c + Q_v Q_o}, \tag{1}$$

$$c \leftarrow c + Q_v Q_o, \tag{2}$$

where $\odot$ indicates the element-wise multiplication and $M_f \in \{0,1\}^{Q_v \times N_v}$. The value of $M_f(i,j)$ is 0 when either the i-th interaction category or the j-th object category is invalid or not involved in the training; otherwise, it is set to 1. Both

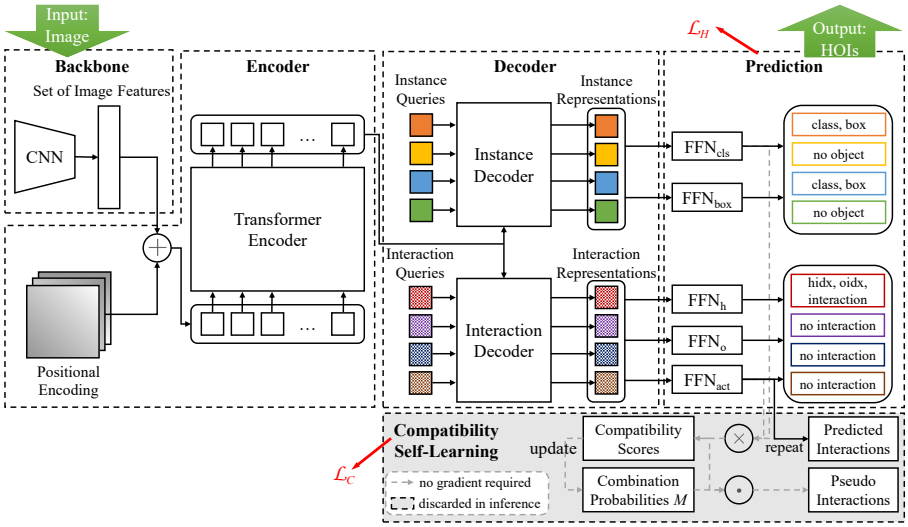

Fig. 2. Overview of the proposed zero-shot HOI detection via compatibility self-learning (HOICS).

matrices M and c start as zero-initialized. Through the optimization of HOI detection, we derive the updated combination probability matrix M.

2.2.4. *Self-Training*

Please note that the calculations for S_h and M described earlier are updated during training but do not directly participate in model training and optimization. To enable the network to learn compatibility information, we introduce compatibility self-learning. This process does not actively contribute to the HOI detection network's training but refines the interaction detection head through back-propagation. Specifically, we repeat $\hat{Y}_v$ for Q_o times to obtain interaction predictions $Z \in R^{Q_v Q_o \times N_v}$ for composite HOIs. We create pseudo interaction labels $\tilde{Y}_v$ using compatibility scores S_h and combination probability M as follows,

$$\tilde{Y}_h = \frac{\sum_i^{N_o} \left(S_h \odot M\right)(:,:,i)}{\max\left(M\right)}, \tag{3}$$

where $\tilde{Y}_v \in R^{Q_v Q_o \times N_v}$ and $Q_v Q_o$ represents the number of composite HOIs in the mini-batch. We use composite HOIs with pseudo interaction labels to self-train, and the interaction classification loss in compatibility self-learning branch is defined as follows,

$$\mathcal{L}_C = \frac{\sum_i^{Q_v Q_o} \sum_j^{N_v} \mathcal{L}_{BCE}\left(Z\left(i,j\right), \tilde{Y}_v\left(i,j\right)\right)}{Q_v Q_o N_v}, \tag{4}$$

where $\mathcal{L}_{BCE}$ indicates the binary cross entropy loss.

2.3. Training and Inference

Training. Our HOICS is trained end-to-end, incorporating both HOI detection loss $\mathcal{L}_H$ and compatibility self-learning loss $\mathcal{L}_C$. $\mathcal{L}_H$ follows the same formulation as in HOTR [6], representing the overall HOI detection loss. The training process and optimization objectives for HOI detection are akin to those employed in HOTR. Importantly, compatibility self-learning does not actively participate in the HOI detection network's training but rather refines the interaction detection head through back-propagation. During training, we continuously update the combination probability matrix, saving it alongside the network parameters.

Inference. The inference process closely resembles that of HOTR, with the exclusion of the compatibility self-learning branch. As the interaction prediction head is refined during self-training, we attain enhanced detection results, particularly in addressing long-tail challenges and zero-shot detection scenarios.

3. EXPERIMENTS

In this section, we commence by introducing the datasets and metrics used in our experiments. Subsequently, we delve into the specifics of our method's implementation. Next, we showcase the effectiveness of our proposed approach by comparing it to state-of-the-art visual compositional methods. Finally, we perform ablation studies to validate the individual components of our method.

3.1. Datasets and Metrics

Datasets. To validate the performance of our model, we evaluate our model on two public benchmark datasets: V-COCO [2] and HICO-DET [3].

Metrics. We adopt the evaluation settings outlined in [3]. For V-COCO, we present the AP_{role} results in two scenarios, denoted as $AP_{role}^{\#1}$ and $AP_{role}^{\#2}$. Regarding HICO-DET, we report our performance in the *Default* setting, where we

Table 1. Comparison of zero-shot detection results with state-of-the-art on HICO-DET.

Method	Default		
	Unseen	Seen	Full
VCL [8] (rare first)	7.55	18.84	16.58
SCL [16] (rare first)	2.26	22.72	18.71
Baseline [6] (rare first)	12.84	23.24	21.16
HOICS (rare first)	**15.00**	**24.29**	**22.43**
VCL (non-rare first)	9.13	13.76	12.76
SCL (non-rare first)	7.05	16.70	14.77
Baseline (non-rare first)	13.35	22.55	20.71
HOICS (non-rare first)	**13.40**	**23.13**	**21.18**

assess detection on the full test set. Our reported metric is the mean average precision (mAP) computed across various category sets.

Zero-Shot Setting. We employ a zero-shot HOI experiment strategy following the approach outlined in [8], consisting of two groups: rare first selection and non-rare first selection. In rare first selection, we prioritize rare labels for unseen classes based on instance counts, while in non-rare first selection, we prioritize non-rare labels. We evaluate the results of unseen HOI detection using mAP metric. Our results are reported in three settings: Unseen (120 HOIs), Seen (480 HOIs), and Full (600 HOIs), using the *Default* mode on HICO-DET.

3.2. Implement Details

We initialize the transformer with a learning rate of 10^{-4} and a weight decay of 10^{-4}, training HOICS using the AdamW optimizer. The transformer weights are initialized using Xavier initialization. In both V-COCO and HICO-DET experiments, we utilize ResNet-50 as the backbone and train the model for 100 epochs, incorporating a learning rate reduction starting from the 80th epoch. We pre-train the backbone, encoder, and instance decoder on MS-COCO and keep them frozen during training. We apply scale augmentation, following the approach in DETR [17], by resizing input images to ensure the shortest side ranges from 480 to 800 pixels, with the longest side not exceeding 1333 pixels.

3.3. Comparison with State-of-the-art

3.3.1. Effectiveness for Zero-Shot HOI Detection

We evaluate our method in HOI zero-shot detection and compare it with state-of-the-art visual compositional methods [8, 16]. Table 1 demonstrates that HOICS outperforms the baseline by 2.16% in Unseen categories. Additionally, both selection strategies consistently yield improvements with HOICS across all categories, highlighting the role of compatibility self-learning in mitigating the scarcity of HOI samples.

Table 2. Effectiveness of HOICS on V-COCO.

Method	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
HOTR [6]	55.2	64.4
QPIC	58.8	61.0
FGAHOI(R-50) [18]	59.0	59.3
Baseline	60.43	65.31
HOICS	**61.64**	**66.53**

Table 3. Ablation studies about compatibility self-learning (denoted as CSL) and data augmentation (denoted as mosaic) in rare first selection on HICO-DET.

Method	Default		
	Unseen	Seen	Full
Baseline	12.84	23.24	21.16
+ CSL	13.62(+0.78)	23.96	21.89
+ mosaic	13.61(+0.77)	24.24	22.11
+ CSL + mosaic	**15.00**(+2.16)	**24.29**	**22.43**

3.3.2. Effectiveness on V-COCO

We also evaluate our HOICS against state-of-the-art HOI detection methods on V-COCO. While V-COCO does not explicitly pose a zero-shot problem, and the dataset itself does not exhibit significant imbalance, our HOICS still shows improvements over the baseline, as demonstrated in Table 2.

3.4. Ablation Study

We employ mosaic image augmentation in the training of our HOI detector. To assess the effectiveness of our HOICS, we conduct ablation studies concerning compatibility self-learning and data augmentation. Our results are reported in various settings: Unseen, Seen, Full, using the Default mode and rare first selection on HICO-DET. Table 3 reveals that both compatibility self-learning and data augmentation contribute to performance improvements on HICO-DET. The compatibility self-learning strategy and mosaic augmentation individually enhance the mAP in the Unseen category by 0.77% and 0.78%, respectively. Notably, their combined impact surpasses the linear sum of their individual improvements, resulting in an overall improvement of 2.16%.

4. CONCLUSION

In this study, we introduced compatibility self-learning for zero-shot HOI detection. This involved the development of a combination probability matrix and HOI compatibility scores, applied to enhance predictions through a self-learning strategy. We also incorporated mosaic augmentation to diversify interaction and object combinations. Our extensive experiments confirm the effectiveness of our HOICS in HOI detection benchmarks, with a significant 2.16% improvement in mAP for the Unseen category over the baseline.

5. REFERENCES

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, Number: 7553 Publisher: Nature Publishing Group.

[2] Saurabh Gupta and Jitendra Malik, "Visual Semantic Role Labeling," *arXiv:1505.04474 [cs]*, May 2015, arXiv: 1505.04474.

[3] 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)*, Mar. 2018, pp. 381–389.

[4] 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 479–487, ISSN: 2575-7075.

[5] 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.

[6] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim, "HOTR: End-to-End Human-Object Interaction Detection with Transformers," *arXiv:2104.13682 [cs]*, Apr. 2021, arXiv: 2104.13682.

[7] Trevor Bergstrom and Humphrey Shi, "Human-Object Interaction Detection: A Quick Survey and Examination of Methods," in *Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis*, New York, NY, USA, Oct. 2020, HuMA'20, pp. 63–71, Association for Computing Machinery.

[8] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao, "Visual Compositional Learning for Human-Object Interaction Detection," in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, Lecture Notes in Computer Science, pp. 584–600, Springer International Publishing.

[9] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel, "HCVRD: a benchmark for large-scale human-centered visual relationship detection," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, Louisiana, USA, Feb. 2018, AAAI'18/IAAI'18/EAAI'18, pp. 7631–7638, AAAI Press.

[10] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa, "Detecting Human-Object Interactions via Functional Generalization," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10460–10469, Apr. 2020, Number: 07.

[11] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan, "Discovering Human Interactions With Novel Objects via Zero-Shot Learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 11649–11658, ISSN: 2575-7075.

[12] 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, Issue: 2.

[13] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu, "GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection," 2022, pp. 20123–20132.

[14] 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)*. 2018, pp. 1568–1576, IEEE.

[15] 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.

[16] Zhi Hou, Baosheng Yu, and Dacheng Tao, "Discovering Human-Object Interaction Concepts via Self-Compositional Learning," *arXiv preprint arXiv:2203.14272*, 2022.

[17] 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*. 2020, pp. 213–229, Springer.

[18] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei, "FGAHOI: Fine-Grained Anchors for Human-Object Interaction Detection," Jan. 2023, arXiv:2301.04019 [cs].