GHOI: A Green Human-Object-Interaction Detector

1st Tsung-Shan Yang
Department of Electrical Engineering
University of Southern California
Los Angeles, USA
tsungsha@usc.edu

2nd Yun-Cheng Wang
Department of Electrical Engineering
University of Southern California
Los Angeles, USA
yunchenw@usc.edu

3rd Chengwei Wei
Department of Electrical Engineering
University of Southern California
Los Angeles, USA
chengwei@usc.edu

4th C.-C. Jay Kuo
Department of Electrical Engineering
University of Southern California
Los Angeles, USA
jckuo@usc.edu

***Abstract*—Human-Object Interaction (HOI) detection is a fundamental task in image understanding. All recent high-performance HOI methods are based on deep learning (DL) models, which are computationally expensive with an opaque inference process. A green HOI (GHOI) detector is proposed in this work to strike a good balance between detection performance, inference complexity (i.e., low carbon footprints), and mathematical transparency. GHOI is a two-stage method. In the first stage, it conducts object detection and extracts various features from the input images as intermediate outputs. In the second stage, it uses the first-stage outputs to predict the interaction type using the XGBoost classifier. One novel contribution is the application of error correction codes (ECCs) to encode rare interaction cases. This reduces the model size and the complexity of the XGBoost classifier in the second stage. Experimental results demonstrate the advantages of ECC-coded interaction labels and the nice balance of detection performance and complexity of the proposed GHOI method.**

***Index Terms*—Human-Object Interaction (HOI) Detection, Error Correction Code, Green Learning, Image Understanding.**

I. Introduction

Human-Object Interaction (HOI) detection is an important task for image understanding [1], [2]. The labels in HOI datasets are triplets in the form of <Human-Interaction-Object>. HOI detection focuses on human-centric relations and can be applied to human-related applications such as action detection. HOI detection can be challenging, as illustrated in Fig.1. The examples are taken from a popular HOI detection dataset, HICO-DET [1]. Images may contain the label *'no_interaction'*, but not every no_interaction human-object pair is labeled. Furthermore, some images may share the same verb in different scenarios, known as verb polysemy [3].

HOI methods can be categorized into two main categories: one-stage and two-stage models. One-stage models are obtained via end-to-end optimization of certain neural network architectures, where all ground truth bounding boxes and labels are used to define a loss function in the training stage. Although they can achieve better prediction performance, they are difficult to interpret. Two-stage methods decompose the processing procedure into two decoupled modules: 1) identifying where humans and objects are located and 2) determining the type of interaction between them. In the first stage, it conducts object detection using a pre-trained object detector and extracts various features from input images. In the second stage, it leverages the first-stage outputs, such as object classes, bounding boxes, spatial relationships, etc., to predict the interaction type. Deep learning (DL) models are often employed in both stages, leading to a high model complexity. Typically, one-stage models outperform two-stage models in detection accuracy at the expense of larger model sizes and higher training/inference complexities. On the other hand, two-stage models are easier to understand due to their modular design.

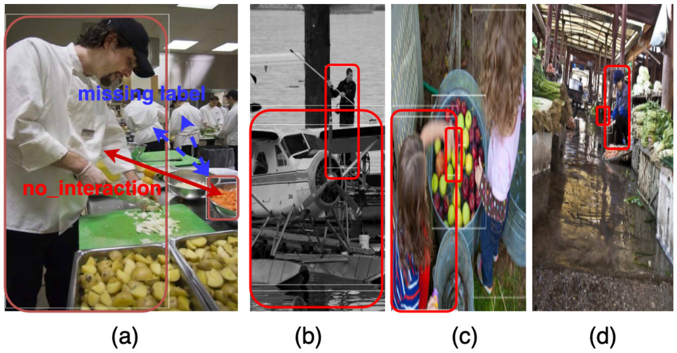

Fig. 1. Illustration of challenges in the HOI problem with images from the HICO-DET dataset: (a) images labeled as 'no_interaction,' (b)-(d) three images with the same verb, 'wash,' but humans behave differently.

Fig. 2 shows the comparison between different HOI detectors in terms of model performance, model sizes, and carbon footprint. Specifically, the model performance is expressed in mAP (%) in the y-axis, model sizes are expressed as the number of model parameters in the x-axis, and carbon footprint is reflected by inference floating-point operation (FLOP) numbers, respectively. This figure includes one-stage transformer-based models (QAHOI [4] and ERNet [5]) and one-stage models based on interaction point prediction (IP-Net [6] and PPDM [7]). We also list two SOTA two-stage methods, namely, UPT [8] and SCG [9]. Aiming at inter-

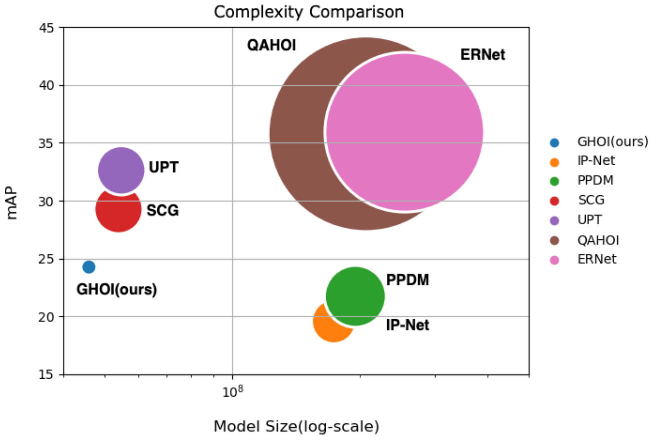

Fig. 2. Complexity comparison between the proposed GHOI and several other state-of-the-art (SOTA) detectors for the HICO-DET dataset, where the x-axis is the model size in the log scale, the y-axis is mAP (%), and the bubble size is proportional to the inference FLOP numbers.

pretability and lower carbon footprints, we propose a new two-stage method called Green HOI (GHOI) in this paper. As shown in Fig. 2, its mAP performance is worse than other two-stage models, UPT and SCG, yet its FLOP number is significantly lower. As reported in Sec. IV, the FLOP number of GHOI is 4,500 times smaller than SCG and 15,800 times smaller than UPT per query, respectively. GHOI outperforms two one-stage models (IP-Net and PPDM) and underperforms another two one-stage models (QAHOI and ERNet) in mAP. However, it has tremendous advantages in the model size and FLOP numbers. In terms of carbon footprint (FLOP number) and memory (model size), GHOI offers an attractive AI/ML solution for mobile and edge devices.

One main challenge in HOI detection is to handle the imbalanced distribution of interaction pairs in the training samples. We propose a hybrid coding scheme to address this problem. That is, we partition interaction pairs into rare and non-rare cases. For non-rare cases, we adopt the traditional one-hot coding. For rare cases, we group them into one superclass, and then adopt binary error correction codes (ECCs) to encode these rare cases. This is one of the major contributions of this work.

II. Related Work

A. One-stage HOI Detection

Wang et al. [6] and Liao et al. [7] proposed the use of interaction point prediction to improve the Average Precision (AP) performance. Besides object detection, they exploited the overlapping features between human and object bounding boxes. The supervised loss from interaction points helps the model obtain better features and remove unlikely interactions. With the thriving visual transformers (ViTs), transformer-based models have been investigated. Kim et al. [10] developed a DETR [11] backbone and combined it with pairwise human/object queries for the interaction decoder. The interaction Feed-Forward Networks (FFN) were trained by the corresponding relation labels. Chen et al. [12] introduced auxiliary interaction vector prediction for interaction FFNs optimization. Tamura et al. [13] used the Hungarian algorithm [14] to calculate the loss of matched human/object pairs in the loss function. Liao et al. [15] combined the SOTA language and visual transformer, CLIP [16], whose comprehensible embeddings improve the relation decision significantly. Chen et al. [4] replaced the ResNet backbone with deformable transformers. Lim et al. [5] used the EfficientNet as the backbone to extract multi-scale features.

Although one-stage models offer SOTA HOI detection performance, they do have some shortcomings. First, the training of one-stage models is highly dependent on the dataset. Zhu et al. [17] pointed out that one-stage models could be biased in detection results under a skewed data distribution. Second, it is challenging to interpret one-stage methods since the semantic information in images is hidden in numerous cascading latent spaces. Researchers analyzed the models using convolutional filter responses and attention matrices in Visual Transformers (ViTs) indirectly. Third, they suffer from a large model size and extremely high computational complexity.

B. Two-stage HOI Detection

HOI is a human-centric classification task. The linkage between human and object representations is crucial. In two-stage models, the first stage extracts various human and object representations, while the second stage is a multi-pair (i.e., human-object pairs) and multi-label (i.e., interaction labels) classification problem. Gupta et al. [2] used pose estimation models to obtain semantic information. By exploiting the human-object correlation, Hou et al. [18] constructed a model to yield human and object streams and handled the relation between the two streams based on the co-occurrence of <human-relation-object> triplets. The human-object relationship can also be formulated as a graph, where human and object features can be viewed as the vertices in the graph. Gao et al. [19] proposed a dual structure to model the relation. It combined the human-centric and object-centric graphs to predict the relation. Zhang et al. [9] exploited the spatial information between objects in the graph convolution structure. With the development of transformers, Zhang et al. [8] adopted the FFN decoder structure for the pairwise relation classification.

To address the increasing computational burdens of DL networks, Kuo et al. [20] proposed a statistical-based learning framework called Green Learning (GL). The GL paradigm does not have neurons, neural networks, and end-to-end optimization via backpropagation. Instead, it adopts a feedforward and modular design in both training and inference based on data statistics. The whole processing pipeline is purely data-driven and transparent. The GL solution focuses on reducing FLOPs to relieve power consumption and carbon footprint, addressing environmental concerns. Our work follows this principle, as detailed in the next section.

III. GHOI Method

A. System Overview

The system diagram of the proposed GHOI method is shown in Fig. 3, which is a two-stage method. The first stage is a pre-trained object detector, where we select DETR [11] as the object detector. It uses ResNet50 as the backbone and achieves good object detection performance trained by suitable object detection datasets. Since our main contributions lie in the second stage, we will emphasize the data processing pipeline of the second stage in this section. It consists of the following four tasks in cascade.

- Module A: Visual Features Construction
  We utilize the Region of Interest (RoI) alignment and pooling [21] to generate human and object representations. It yields an input query pair that contains human and object features and their associated spatial information.
- Module B: Hybrid Interaction Coding
  To address the imbalanced distribution of interaction pairs in the training samples, we propose a hybrid coding scheme. That is, we partition interaction pairs into rare and non-rare cases. For non-rare cases, we adopt the traditional one-hot coding. For rare cases, we group them into one super-class and then adopt binary codes with error correction codes (ECCs) to encode rare cases within the super-class.
- Module C: Discriminant Features Selection
  The discriminant feature selection process is conducted based on the interaction codes. That is, we identify a subset of discriminant features against every bit assignment of the interaction type.
- Module D: Conditional Decision on the Interaction Type
  The final prediction is the aggregation of the probabilities outputs from each interaction bit.

The model in the second stage is efficient in terms of the number of model parameters and Floating-point Operations (FLOP) numbers. Modules B-C are statistics-based, allowing interpretability. Furthermore, the application of error correction codes (ECCs) to encode interaction labels is a novel contribution in the AI/ML literature. Its advantages are demonstrated in the experiments section.

B. Processing Modules in the Second Stage

1) Module A: Constructing a rich feature set for human and object representations is critical. Intuitively, the relative distances and other scenarios between the human and object locations in images are useful for HOI detection. Utilizing the first-stage model, we can capture the features of corresponding regions by RoI pooling and alignment [21]. The human and object features can be obtained from the aggregation of different layer feature maps in the detector. The relative spatial information includes the interaction vector and the relative sizes of human and object bounding boxes. The interaction vector is defined to be the difference between the centers of human and object bounding boxes. It can be represented in Euclidean or polar coordinate systems. Furthermore, the background information can be extracted from the whole image features obtained from the backbone network. To summarize, the features under consideration comprise human RoI features, object RoI features, relative spatial features, and whole image features. We also split human and object representations as individual queries and determine their spatial features accordingly. Some detailed descriptions are illustrated in Fig. 3. Yet, we should point out that these features are far from perfect since they lack precise semantic information, and discriminant features may be concealed in noisy training samples.

To deal with imbalanced labels, our model fits subsets of the interaction samples instead of the whole dataset. For human queries, the classifiers are trained by subsets containing a common object. That is, the desired outputs of a classifier can be denoted as

$$P(relation \mid human) = \sum_{c \in \{Object\}} P(relation, object = c \mid human),$$

where c denotes an object type and $\{Object\}$ denotes the whole object set. Under the constraint, $object = c$, we can reduce the overall long-tail distribution to a few short-tail conditional distributions. Similarly, the performance of the classifier for object query can also be improved by conditional probabilities. We employ clustering algorithms such as KMeans to create subsets and use them to train classifiers in each subset. We can assign a pseudo-label for each subset, and then the classifier can be formulated in a similar manner. The ultimate classifier for object queries can be obtained by combining multiple subset classifiers in a weighted manner. To make relation decisions in the inference stage, we can aggregate the results of both human and object queries.

2) Module B: The foundation of modern machine learning models is to capture the distribution in the training dataset and generalize it to unseen samples. It is essential to find a representation space that generalizes well between training and testing samples. Representations could be highly diversified, and the labeled data may possess a long-tail distribution in real-world applications. It is typical to adopt the one-hot vector to represent the classes of interest in the context of AI/ML. There are two problems with the one-hot representation. First, if the class number is large, the dimension of these one-hot vectors can be high. Second, the labeled data possess a long-tail distribution, as mentioned above. Take the HOI benchmark, HICO-DET, as an example. It has 600 interaction triplets. However, 138 of them have less than ten samples and are called rare cases. If we adopt the one-hot encoding scheme for all, we need 600-dimensional vectors to represent them and have to train 600 one-versue-the-rest binary classifiers. The classification performance of each rare case is expected to be poor due to high data imbalance since it is challenging for a classifier to learn from less than ten samples among more than 10^6 queries.

To handle this challenge, we merge all rare cases into a super-class and adopt the traditional one-hot coding to encode non-rare cases plus this super-class. Then, to differentiate

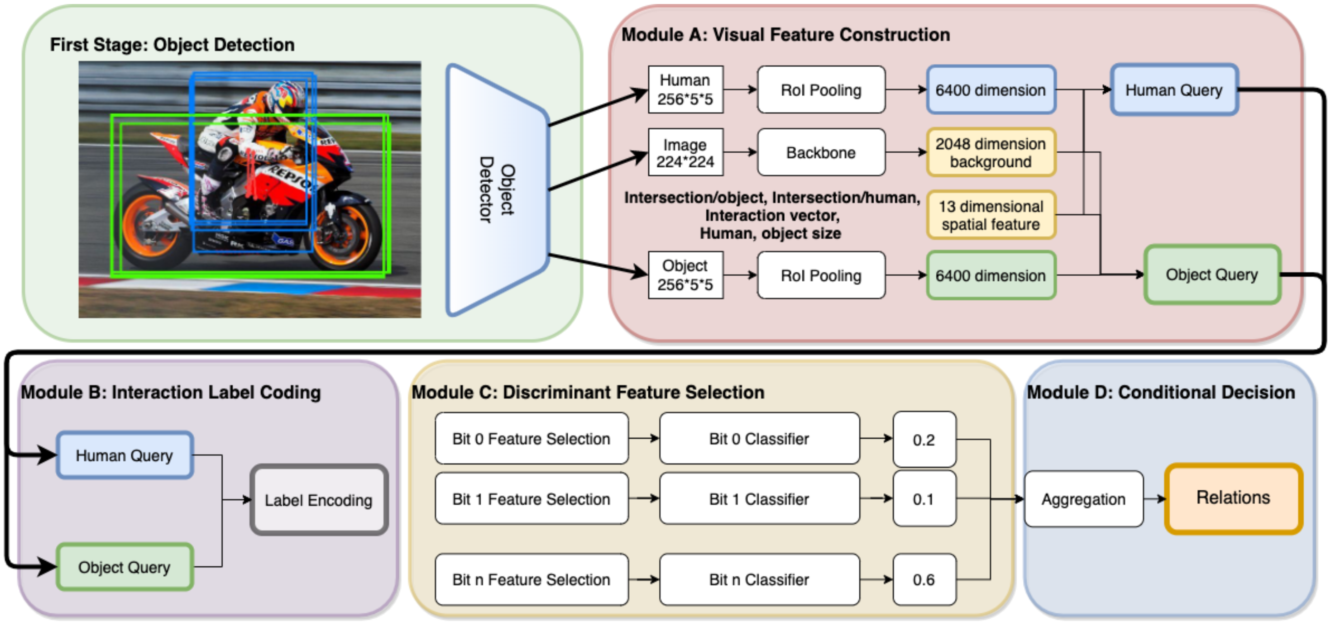

Fig. 3. The overall system diagram of the proposed GHOI. Its first stage is a pre-trained object detector. The main contributions of GHOI lie in the data processing pipeline in the second stage. It consists of four modules: A) visual features construction, B) interaction label coding, C) discriminant features selection, and D) conditional decision on the interaction type.

rare cases inside the super-class, we adopt the binary coding scheme. To compare the difference between the one-hot coding and the binary coding, we take a 4-class classification problem as an example. The four classes are represented as $\{1000, 0100, 0010, 0001\}$ in the one-hot encoding and as $\{00, 01, 10, 11\}$ in the binary coding. Each bit represents a binary split. For the one-hot coding, we can train four binary classifiers that handle the one-versus-the-rest classification problem. For the binary coding, we only train two binary classifiers. The first one separates $\{00, 01\}$ from $\{10, 11\}$ based on the first bit while the second one splits $\{00, 10\}$ from $\{01, 11\}$ based on the second bit. Each binary classifier can be viewed as a decoder. Nevertheless, each classifier may have mistakes, leading to wrong aggregated results. We use error correction codes (ECC) to enhance the robustness to have a remedy. To follow the above example, we can assign three-bit codewords to them, i.e., $\{000, 011, 101, 110\}$. Every codeword pair has a Hamming distance of 2 (i.e., have two different bits) after adding the error correction bit. Here, we use Hamming codes [22] to improve the performance of straightforward binary codes and ensure that each representation differs from others with a Hamming distance no less than 3 in GHOI.

The performance of four coding schemes is compared in Table I for the HICO-DET dataset. They are one-hot codes, binary codes, Hamming codes, and a hybrid coding scheme. The hybrid coding scheme adopts the one-hot coding for non-rare cases plus the super-class of all rare cases and the Hamming codes for rare cases. We see from the table that the hybrid coding scheme achieves the best results, whose mAP value is substantially higher than that of one-hot codes. It is also worthwhile to point out that Hamming codes have a smaller model size, which helps reduce the model size of the hybrid scheme.

TABLE I

PERFORMANCE COMPARISON BETWEEN FOUR CODING SCHEMES FOR INTERACTION LABELS IN MAP (%). UNDER THE SAME ARCHITECTURE, THE HAMMING CODES GIVE THE BEST PERFORMANCE FOR RARE CASES, WHILE THE ONE-HOT CODES OFFER THE BEST PERFORMANCE FOR NON-RARE CASES. THE HYBRID CODING SCHEMES YIELD THE BEST OVERALL PERFORMANCE.

Methods	Default			Model Size
	Full	Rare	Non-Rare	
GHOI (one-hot codes)	20.55	13.47	22.66	56.4M
GHOI (binary codes)	16.35	7.97	18.86	15.4M
GHOI (Hamming codes)	19.19	14.50	20.59	27.7M
GHOI (hybrid)	**24.53**	**19.26**	**26.09**	45.9M

3) Module C: Each bit representation in Hamming codes corresponds to a partition of labeled interactions into two sets. In other words, we relabel interactions of rare cases into two types denoted by 0 and 1, respectively. For a given binary label, we need to select discriminant features to facilitate the classifier in the next module. This can be achieved by applying the Discriminant Feature Test (DFT) [23] to all input features one by one. For a given 1D input feature, we place the feature value of each labeled training sample in a line segment bounded by the range of the maximum and minimum values, as shown in Fig. 4. Then, we search for the optimal partition point on this line segment to minimize the loss function, which is defined as the weighted sum of binary cross-entropies of the left and right partitions. A feature is more discriminant if it has a lower loss value. Then, we can plot the loss value curve from the lowest to the highest and use the elbow point to select a set of discriminant features from the whole feature set.

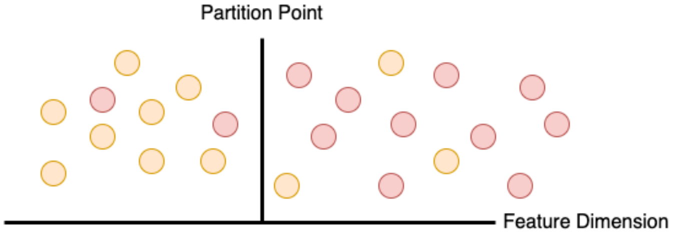

Fig. 4. Visualization of DFT, where pink and orange dots represent the “0” and “1” binary labels, and the loss function is the weighted cross-entropy sum of samples in the left and right parts of the partition line.

4) Module D: We divide the desired decisions into sequential subproblems instead of training a complex classifier for the skewed data distribution. Each subproblem can be expressed clearly, step by step. First, we attempt to maximize the conditional probability of an interaction (or relation) conditioned on the human and object representations, which can be written as

$$\begin{aligned}
&P(relation \mid human, object)\\
&= P(relation \mid human) * \frac{P(object \mid relation, human)}{P(object \mid human)}\\
&= P(relation \mid object) * \frac{P(human \mid relation, object)}{P(human \mid object)}\\
&\sim \alpha P(relation \mid human) + \beta P(relation \mid object),
\end{aligned}$$

where α, β are learnable parameters. Then, the two conditional probabilities in the last equation can be further expressed as

$$\begin{aligned}
P(relation|human) &= \sum_{c\in\{Object\}} P(relation, object = c|human)\\
P(relation|object) &= \sum_{d\in\{Human\}} P(relation, human = d|object),
\end{aligned}$$

where c and d are class labels and $\{Object\}, \{Human\}$ are the object sets and clustered human representations. Suppose we use a bit stream, $B = (b_0, b_1, \cdots, b_{n-1})$, to represent a relation. Then, we would like to maximize $P(relation, object = c|human)$ and $P(relation, human = d|object)$, called the human query and the object query, respectively. The conditional probability of human queries can be written as

$$\begin{aligned}
&P(relation, object = c|human)\\
&= P(B, object = c|human)\\
&= \bigcap_{0\le i<n} P(b_i, object = c|human).
\end{aligned}$$

The conditional probability of object queries can be found in the same manner.

All the probability estimators in GHOI are XGBoost [24]. For each bit classifier, we set the number of estimators and the depth of the tree to 300 and 3, respectively. The aggregation of the bit stream prediction is conducted by Linear Discriminant Analysis (LDA).

IV. Experiments

A. Datasets

V-COCO [30] and HICO-DET [1] are two commonly used HOI detection datasets. V-COCO is a subset of the MS-COCO dataset. It contains 2,533 training images, 2,867 validation images, 4,946 test images, and 24 actions. HICO-DET is larger than V-COCO. It comprises 37,633 training images, 9,546 test images, 117 actions, and 600 interactions for various action-object pairs. Its training set has 117,871 human–object pairs with annotated bounding boxes, while its testing set contains 33,405 such pairs. HICO-DET is a challenging dataset. The 600 labeled interactions can be divided into 138 rare cases and 462 non-rare cases. Rare cases have less than ten samples in the training set. We use the mean Average Precision (mAP) as the evaluation metric, which is the mean of the average precision of all classes.

B. Experimental Results

1) Performance Benchmarking against Other Two-stage Models: We compare the performance of GHOI against other SOTA two-stage models, including both multi-stream and graph-based models, in Table II, where the top and the second performers are in bold and underlined, respectively. GHOI achieves the second-best mAP values in most categories for HICO-DET. Our GHOI model also has the smallest number of learnable parameters. It’s important to note that GHOI relies solely on visual features and doesn’t incorporate external word embeddings or pose estimation information during training. DRG [19] uses extra language models in the training and prediction. Graph-based HOI models need iterations of operations in graph convolutional networks, leading to a higher computation complexity, to be discussed at the end of this subsection.

2) Performance Benchmarking against One-stage Models: Transformer models achieve impressive performance in various computer vision tasks at the expense of very high computational complexities in both training and inference. In the encoder-decoder-based detector, the model requires auxiliary queries for detection. There is no rule of thumb to determine the hyperparameters of the queries and save on computation requirements. Besides model efficiency, the training process is nontrivial for transformer-based one-stage HOI models. HICO-DET is a dataset with a long-tailed distribution. The number of rare cases with less than ten samples has more than 10^6 labeled pairs. The performance of one-stage models would drop dramatically if no finetuning were conducted on object detection and action detection individually. We compare the performance of our GHOI method and those without delicate finetuning in Table III. It was observed in [17] that imbalanced triples tend to decrease the object detection performance. In contrast, our two-stage GHOI method is more robust with respect to imbalanced triples. It can retain the performance of object detectors and plug-on relation detection features.

3) Comparison of Computational Complexity and Carbon Footprint: We compare the computational cost of GHOI,

TABLE II
DETECTION PERFORMANCE COMPARISON OF SOTA TWO-STAGE MODELS IN MAP (%) FOR THE HICO-DET DATASET UNDER THE DEFAULT AND KNOWN OBJECT SETTINGS AND FOR THE V-COCO DATASET. THE MODEL SIZES IN THE PARAMETER NUMBER (M) ARE ALSO COMPARED, WHERE THE NUMBERS ARE TAKEN FROM LIM ET AL. [5].

Architecture	Method	Param(M)(↓)	Backbone	Default(↑)			Known Object(↑)			V-COCO(↑)	
				Full	Rare	Non-Rare	Full	Rare	Non-Rare	AP^{S1}_{role}	AP^{S2}_{role}
Two-Stage Methods											
Multi-Stream	No-Frill [2]	72.3	ResNet152	17.18	12.17	18.08	-	-	-	-	-
	PMFNet [25]	49.3	ResNet50	17.46	15.65	18.00	20.34	17.47	21.20	-	-
	ACP [26]	-	ResNet101	21.96	16.43	23.62	-	-	-	53.2	-
	PD-Net [3]	-	ResNet152	22.37	17.61	23.79	26.86	21.70	28.44	52.0	-
	VCL [18]	-	ResNet50	23.63	17.21	25.55	25.98	19.12	28.03	48.3	-
Graph-Based	RPNN [27]	-	ResNet-50	17.35	12.78	18.71	-	-	-		
	VSGNet [28]	84.9	ResNet-152	19.80	16.05	20.91	-	-	-	*51.8*	*57.0*
	DRG [19]	46.1	ResNet50-FPN	21.66	19.66	22.25	-	-	-	51.0	-
	SCG [9]	53.9	ResNet50-FPN	**29.26**	**24.61**	**30.65**	**32.87**	**27.89**	**34.35**	**54.2**	**60.9**
Green Learning	GHOI (Ours)	**45.9**	ResNet50-FPN	24.53	19.26	26.09	27.64	22.70	29.12	50.8	56.3

TABLE III
DETECTION PERFORMANCE COMPARISON IN MAP (%) BETWEEN GHOI AND THREE END-TO-END TRAINED ONE-STAGE METHODS (WITHOUT FINETUNING OBJECT AND ACTION DETECTORS INDIVIDUALLY) FOR THE HICO-DET DATASET, WHERE THE MAP RESULTS OF HOTR, AS-NET, AND QPIC ARE TAKEN FROM MA ET AL. [29].

Methods	Default		
	Full	Rare	Non-Rare
HOTR [10]	23.46	16.21	25.65
AS-Net [12]	24.40	**22.39**	25.01
QPIC [13]	24.21	17.51	**26.21**
GHOI (ours)	**24.53**	19.26	26.09

TABLE IV
THE COMPARISON OF FLOP NUMBERS PER QUERY BETWEEN GHOI AND TWO SOTA TWO-STAGE MODELS.

Two-Stage Models	Default	Parameters	FLOPs
	Full		(per query)
SCG (Graph)	29.26	53.9M (1.2x)	54M (4,500x)
UPT (Transformer)	32.62	54.7M (1.2x)	190M (15,800x)
GHOI (Ours)	24.53	**45.9M (1x)**	**12K (1x)**

UPT [8] and SCG [9] in Table IV. UPT is a transformer-based method, while SCG is a graph-based method. They are SOTA two-stage HOI methods (see Fig. 2 and Table II). Both of them demand a large number of iterated computations. The table shows that in the inference stage, the FLOP number per query of GHOI is 1/4,500 and 1/15,800 of that of SCG and UPT. Clearly, GHOI is the most eco-friendly in the carbon footprint measure.

V. CONCLUSION AND FUTURE WORK

A green HOI detector, called GHOI, was proposed in this work. It is both mathematically transparent and computationally efficient while offering competitive detection performance. The use of ECC for the coding of rare interaction types helps improve the robustness of GHOI. It appears that the same idea can be generalized to other detection problems. It is worthwhile for further investigations in the future.

ACKNOWLEDGEMENT

Computation for the work was supported by the University of Southern California's Center for High Performance Computing (hpc.usc.edu).

References

[1] 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.

[2] Tanmay Gupta, Alexander Schwing, and Derek Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9677–9685.

[3] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao, "Polysemy deciphering network for robust human–object interaction detection," *International Journal of Computer Vision*, vol. 129, pp. 1910–1929, 2021.

[4] Junwen Chen and Keiji Yanai, "Qahoi: Query-based anchors for human-object interaction detection," in *2023 18th International Conference on Machine Vision and Applications (MVA)*. IEEE, 2023, pp. 1–5.

[5] JunYi Lim, Vishnu Monn Baskaran, Joanne Mun-Yee Lim, KokSheik Wong, John See, and Massimo Tistarelli, "Ernet: An efficient and reliable human-object interaction detection network," *IEEE Transactions on Image Processing*, vol. 32, pp. 964–979, 2023.

[6] 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.

[7] 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.

[8] Frederic Z Zhang, Dylan Campbell, and Stephen 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. 20104–20112.

[9] Frederic Z Zhang, Dylan Campbell, and Stephen Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13319–13327.

[10] 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.

[11] 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.

[12] 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.

[13] Masato Tamura, Hiroki Ohashi, and Tomoaki 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. 10410–10419.

[14] Harold W Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[15] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si 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. 20123–20132.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[17] Fangrui Zhu, Yiming Xie, Weidi Xie, and Huaizu Jiang, "Diagnosing human-object interaction detectors," *arXiv preprint arXiv:2308.08529*, 2023.

[18] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao, "Visual compositional learning for human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 584–600.

[19] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang, "Drg: Dual relation graph for human-object interaction detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 696–712.

[20] C-C Jay Kuo and Azad M Madni, "Green learning: Introduction, examples and outlook," *Journal of Visual Communication and Image Representation*, vol. 90, pp. 103685, 2023.

[21] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[22] Richard W Hamming, "Error detecting and error correcting codes," *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.

[23] Yijing Yang, Wei Wang, Hongyu Fu, C-C Jay Kuo, et al., "On supervised feature selection from high dimensional feature spaces," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[24] Tianqi Chen and Carlos Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[25] 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.

[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] Penghao Zhou and Mingmin Chi, "Relation parsing neural network for human-object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 843–851.

[28] 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.

[29] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei, "Fgahoi: Fine-grained anchors for human-object interaction detection," *arXiv preprint arXiv:2301.04019*, 2023.

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