

This item is likely protected under Title 17 of the U.S. Copyright Law. Unless on a Creative Commons license, for uses protected by Copyright Law, contact the copyright holder or the author.

Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

SKGHOI: Spatial-Semantic Knowledge Graph for Human-Object Interaction Detection

Lijing Zhu¹, Qizhen Lan¹, Alvaro Velasquez², Houbing Song³, Acharya Kamal⁴, Qing Tian¹, Shuteng Niu¹

¹ Department of Computer Science, Bowling Green State University

² Department of Computer Science, University of Colorado Boulder

³ Department of Information Systems, University of Maryland, Baltimore County

⁴ Department of Electrical Engineering and Computer Science, Embry-Riddle Aeronautical University

Abstract—Detecting human-object interactions (HOIs) is a challenging problem in computer vision. Existing techniques for HOI detection heavily rely on appearance-based features, which may not capture other essential characteristics for accurate detection. Furthermore, the use of transformer-based models for sentiment representation of human-object pairs can be computationally expensive. To address these challenges, we propose a novel graph-based approach, SKGHOI (Spatial-Semantic Knowledge Graph for Human-Object Interaction Detection), that effectively captures the sentiment representation of HOIs by integrating both spatial and semantic knowledge. In a graph, SKGHOI takes the components of interaction as nodes, and the spatial relationships between them as edges. Our approach employs a spatial encoder and a semantic encoder to extract spatial and semantic information, respectively, and then combines these encodings to create a knowledge graph that captures the sentiment representation of HOIs. Compared to existing techniques, SKGHOI is computationally efficient and allows for the incorporation of prior knowledge, making it practical for use in real-world applications. We demonstrate the effectiveness of our proposed method on the widely-used HICO-DET datasets, where it outperforms existing state-of-the-art graph-based methods by a significant margin. Our results indicate that the SKGHOI approach has the potential to significantly improve the accuracy and efficiency of HOI detection, and we anticipate that it will be of great interest to researchers and practitioners working on this challenging task.

Index Terms—Human-Object Interaction Detection, Knowledge Graph Embedding, Graph Neural Network, Spatial-semantic Representation

I. INTRODUCTION

In recent years, the development of human-object interaction (HOI) detection has been a rapidly growing field in computer vision. The HOI detection technology is used to recognize the interactions between humans and objects in a given image, allowing for the more comprehensive understanding of human behavior and its relation to the surrounding environment. The development of HOI detection has been driven by advances in deep learning and neural networks, which have enabled more accurate and efficient analysis of visual data. The goal of HOI detection is to identify human-object interactions, which are depicted as the anticipated set $\langle \text{person}, \text{verb}, \text{object} \rangle$ triplet, for instance, $\langle \text{person}, \text{read}, \text{book} \rangle$. The detection of HOIs is commonly tackled as a multi-label classification task, whereby an image may entail multiple interactions taking place concurrently. For instance, a single image may exhibit a person engaging in both phone usage and

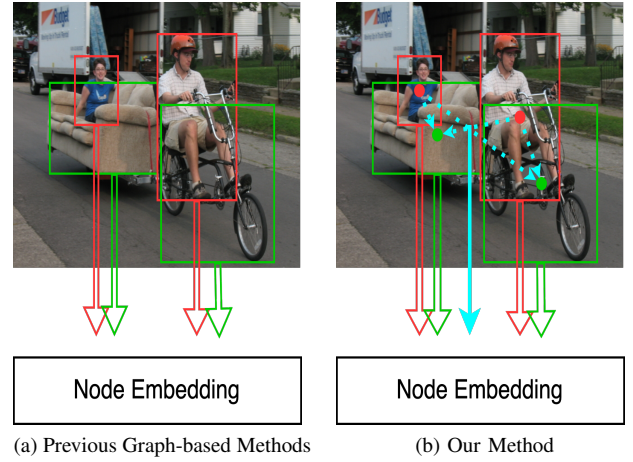


Fig. 1. This is the comparison of the node embedding between previous graph-based approaches and our approach. The detected instances are represented by a colorful rectangular bounding box, with a central dot marking the center of the box. A dotted line connects the center point of the bounding box to indicate the relationship between a pair of human and object in the scene. Figure (a) solely relies on the detected bounding box's appearance feature, whereas figure (b) incorporates both the appearance and the relation features of the human and object pair.

coffee consumption simultaneously, thereby requiring the HOI detector to accurately identify and associate these interactions with the appropriate objects depicted in the image. Detecting the objects and humans present in an image is not the only aspect of the task; recognizing the particular interactions transpiring between them makes it a challenging and intricate undertaking.

As it stands, multiple models and frameworks for human-object interaction (HOI) detection have been proposed. These methods can be broadly classified into one-stage and two-stage approaches. In the case of one-stage approaches [5], [32], [41], [20], [24], the task is accomplished in a single step without the need for detecting the instance first. They can simultaneously detect all pairs of instances on an image and the interaction between these instance pairs. One-stage HOI methods are primarily of two types: interaction-point techniques [19], [39] and transformer-based methods [28], [4]. Interaction-point techniques leverage interaction points to illustrate the region of interaction between humans and objects. This approach efficiently explains the intimate associ-

ation between people and objects. However, its effectiveness diminishes as the distance between them increases. As for the transformer-based method, it attracts enough attention since the DETR [2] developed. The DETR introduced the structure resembling transformers into the computer vision algorithms. Transformer-based models have a more intricate architecture and utilize attention mechanisms to facilitate the learning of the head, although at a higher computational expense due to global feature processing. This methodology accommodates complex interactions and produces superior accuracy compared to interaction-point methods, albeit at a slower convergence rate.

As for the two-stage approaches [26], [29], [36], [40], [34], the detector is responsible for identifying all feasible instances present in the image. Subsequently, another network performs the HOI detection. The output of the object detector is a set of bounding boxes that localize the detected instances, along with corresponding scores that indicate the confidence score of the instance label. The object detector filters out the instances whose scores are below a certain threshold, retaining only the high-confidence instances. The detected instances from the first stage are used as input to a separate network that performs verb classification tasks. The HOI recognition network typically takes as input appearance features, the visual features of the detected instances, and contextual information such as the spatial layout between human and object pairs. In the previous two-stage approach [3], [16], they adopted the multi-stream method, adding a multi-label classifier after the object detector to make action prediction. While the subsequent paper [6], [35] proposes an alternative to the multi-label classifier, such as leveraging pose estimation techniques to incorporate pose information and fuse it with other features to forecast the action category. HOI detection methods often face issues with utilizing a CNN-based backbone feature, which neglects global features. Nonetheless, graph-based methodologies such as those proposed by Li *et al.* [18], Ulutan *et al.* [29] and Zhang *et al.* [36] overcome this limitation by capitalizing on appearance features and effectively utilizing spatial features.

A significant proportion of HOI detection methodologies, whether one-stage or two-stage, depend exclusively on appearance features to predict human-object interactions. Nevertheless, such interactions are inherently intricate and markedly variable, making it difficult for HOI detectors to model them with only instance features of human and object entities. To address this challenge, researchers have developed graph-based techniques, which entail constructing a graph model that captures all human-object pair characteristics. By modeling the relationships between all human-object pairs in the image, the graph-based approach can capture spatial and semantic features between pairs of nodes, leading to more accurate HOI detection. In particular, the graph-based approach is helpful for modeling complex interactions between multiple humans and objects, which can be challenging to capture with traditional two-stage methods that treat each human-object pair in isolation.

The advanced graph-based HOI model adopted the appearance feature as the node representation and the spatial feature as the edge representation, then performed the link prediction,

which is the verb prediction between the human and the object. Despite the promising performance, this approach only partially explores the semantic information between nodes. Specifically, some existing graph-based methods in HOI detection only leverage instance information and ignore other semantic features hidden in relations between nodes. However, there are specific patterns and tendencies in human-object interactions, such as that people are likelier to read books than tables. The trap of implicit ways would enable the provision of additional insights for graph embedding. But the previous graph-based methods did not encode this sentiment information into the node embedding.

In light of the aforementioned limitations, this paper introduces a novel graph-based model for HOI detection named SKGHOI: Spatial-Semantic Knowledge Graph Human-Object Interaction Detection, as illustrated in Figure.2. Our proposed approach employs the translational model to encode relation semantic features and then integrate them into the node embedding, thus strengthening the link between the appearance and relation semantic features. Furthermore, unlike previous graph-based methods that constructed node and edge embedding independently, our approach fortifies the connection between nodes and edges by incorporating relation features into node embedding. For instance, consider the example presented in Figure.1. Previous graph-based techniques only relied on independent instance features to represent nodes. However, our proposed approach encodes relation features (i.e., the relationship between human-object pairs) into node representations. As a result, the node embedding can incorporate relation features and interact more effectively with edge embedding, which also aims to represent relation features. This technique integrates the appearance feature and the novel semantics into the node encoding process, imparting supplementary insights to the graph model. In general, to obtain a more meaningful graph representation, SKGHOI re-encoded both the edge and node embedding, while the node encoder applied the translational model to do semantic feature extraction. To summarize, our contributions are three-fold:

- Our proposed approach involves integrating new semantic representations from the translational model into the node embedding process of the graph-based HOI model, which represents a novel node embedding integration. This is a significant contribution, as it allows the graph-based HOI model to leverage the benefits of the new node and edge embedding representations.
- In our proposed method, we have redesigned the node representation to enhance the coherence between the node and edge of the graph neural network. This is achieved by incorporating a relationship feature into the node embedding, previously absent in the graph network. By doing so, the node and edge can share interactive features, resulting in an improved prediction ability of the graph neural network.
- We modify the knowledge graph embedding model as a translation to fit into the graph-based HOI detection framework and enable domain-transfer learning. By leveraging domain-transfer learning, our model can improve

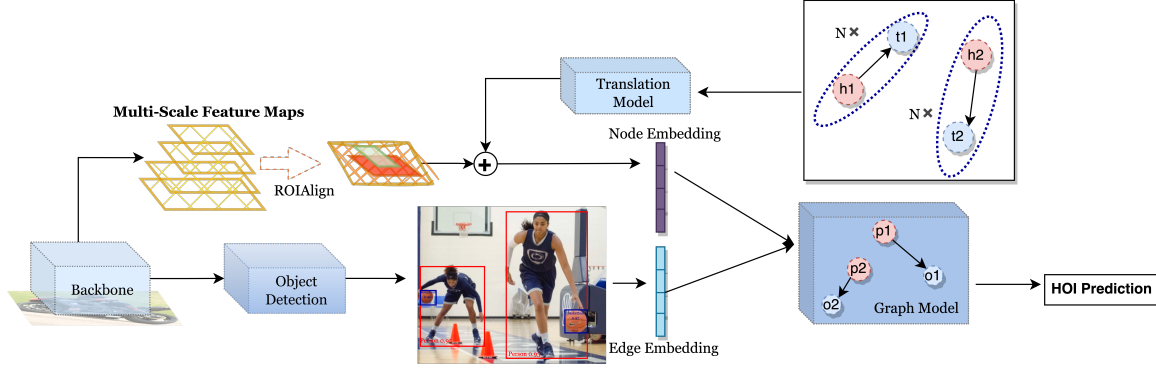


Fig. 2. This is the framework of the proposed method. Given an image, the backbone module extracts its features, which are then passed into object detection. With extracted feature maps, object detection makes predictions for the instance class, score, and box coordinates. Upon detecting each human and object instance pair, a total of $N \times (h, r, t)$ triplets are generated, with the majority of them being negative. These triplets are subsequently fed into the translational model. The node embedding of the HOI graph model is obtained by concatenating the appearance features and transferring embedding from the translational model, the detection information acquired through object detection is utilized as edge embedding, and the HOI prediction is ultimately produced by the graph model.

the accuracy of HOI detection in real-world applications, especially in scenarios where labeled HOI detection data is limited.

Then reminded of the paper is structured as follows: Section II provides a literature review of human-object interaction (HOI) detection. Section III outlines our proposed methodology for addressing the formulated problem. Section IV presents a comprehensive comparison and analysis of our approach with state-of-the-art HOI algorithms. The conclusions of our study and the contributions of our proposed methods are presented in Section V.

II. RELATED WORK

A. Human-Object Interaction Detection

Human-Object Interaction Detection could be a downstream task of object detection. The advancement of HOI detection is intricately linked with the progress of object detection [9]. More recent methods have adopted deep learning architectures, and many HOI detection models now rely on neural network architectures. Existing work widely-used multi-stream framework proposed by Chao *et al.* [3], consists of three streams: a human stream, an object stream, and an interaction stream. Several recent HOI detection algorithms have been proposed that incorporate various improvements.

HOI detection approaches can be broadly categorized into one-stage and two-stage methods based on their model architecture. Until the recent proposal of QPIC [28], two-stage methods have been the predominant approach in HOI detection, with human-pose approaches [30], graph-based approaches [18], [36], and other multi-stream approaches [16] leveraging different frameworks and methods for predicting verbs in the head. In addition to the diverse head designs utilized in two-stage HOI detection, the model's effectiveness is also influenced by the object detector employed. In contrast to two-stage HOI detection, where instances are first detected using an object detector, one-stage approaches directly generate predictions by leveraging the backbone of the object detector without performing explicit instance detection. Most

one-stage HOI detection methods rely on transformer-based architectures, while others employ interaction point-based methods. One such approach, proposed by Liao *et al.* [19], entails a pipeline that does not depend on an object bounding box but directly detects interactions as key points. In recent years, transformer-based HOI detection has gained significant attention and demonstrated promising results, primarily due to QPIC's innovative use of transformer-based techniques in HOI detection. Subsequently, several works have followed in the footsteps of QPIC, such as [4], [20], [34]. However, it is more comprehensive than one-stage approaches adopting transformer-based architectures, as demonstrated by using transformer-based techniques in the two-stage HOI detection method UPT [37]. Despite the effectiveness of transformer-based methods in achieving good results in HOI detection, their application requires substantial computational resources. The resulting predictions are often opaque, posing challenges to subsequent inference work. Furthermore, the relative homogeneity of HOI detection methods in the inference field exacerbates these challenges. Graph-based methods have been proposed as an alternative to address the challenges above faced by previous methods. Despite the fact that accuracy may be compromised to some extent, graph-based methods strike a good balance between computational efficiency and accuracy. Particularly, graph-based methods show great potential in reducing computation costs as they can effectively encode a graph representation with the aid of knowledge graph embedding models.

Graph-based HOI approach The graph-based HOI approach, which was proposed by Qi *et al.* [26], parses the graph neural network [27] and generates features for HOI detection. The box appearance features serve as the initial values for the node features, which are subsequently updated through an iterative message-passing algorithm. In the following work, Wang *et al.* [31] contended that the graph should account for the presence of heterogeneous nodes, human nodes, and object nodes. Gao *et al.* [7] argue that employing homogeneous nodes within separate sub-graphs: human-centric and object-centric graphs, is beneficial because different nodes are responsible for

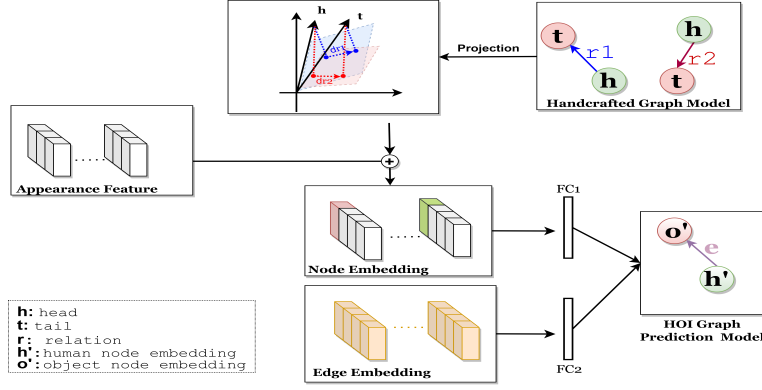


Fig. 3. The design of the novel semantic graph representations. For every human-object pair, we form N triplets $(h, r_1, t), \dots, (h, r_N, t)$. The objective is to project all relationships between the human and object onto the relation-specific hyperplane w_r , with the relation-specific translation vector d_r indicating the distance between the head and tail embedding in w_r . This approach aims to reduce the distance between golden triplets and increase the distance between negative triplets. The node embedding integers are the translation feature and appearance feature, and the edge embedding is the spatial feature, pass these graph representations into the FC layers and the HOI graph prediction model sequentially.

passing and updating distinct information. Then VSGNet [29] enhances the features by considering the spatial relationship between the interacting human-object pair and employs graph convolutions to exploit the inherent structural connections within the pair. A recent graph-based HOI detection model, named SCG [36], innovatively conditions the messages between pairs of nodes on their spatial relationships. To sum up, the graph model builds it up the CNN-based feature extraction, while CNNs are limited to extracting local features, Graph models can leverage spatial information and capture global dependencies, which is a valuable complement to CNN-based backbones.

B. Knowledge Graph Embedding Model

Knowledge graph embedding (KGE) is a method that uses low-dimensional vectors to represent the entities and relations in a knowledge graph. This approach makes a good trade-off between model capacity and efficiency and can be applied to a range of downstream tasks, such as link prediction. By using KGE, it becomes possible to effectively capture the underlying structure of the knowledge graph and use it for various purposes. One of the earliest and most popular knowledge graph embedding methods is TransE [1], which models each relation as a translation operation between the embedding of the head and tail entities. Since then, numerous other embedding methods have been proposed, such as TransH [33], which introduce more complex transformations and modeling techniques to capture different relations of the knowledge graph structure. TransH considers the relation vector as a hyperplane and projects the entity vectors onto this hyperplane to compute the score. This approach enables TransH to model one-to-many and many-to-one relationships more effectively than TransE. We adopted TransH as the translational model for this study since it could handle one-to-many human-object interactions. This study is also the pioneer in using knowledge graph embedding (KGE) methods for HOI detection applications since no such methods have been previously employed in this domain.

III. METHODS

In order to enhance the utilization of the graph-based approach, we will employ AdaMixer [8] as our object detector and develop a more efficient graph head for detecting human-object interactions. The network's general introduction is presented in Section 3.1. While Section 3.2 delves into the design of graph representations. Section 3.3 presents our model training and inference.

A. Algorithm overview

In Figure. 1, we present a high-level overview of our HOI detection network. This model is comprised of two stages. During the first stage, an object detector is utilized to predict all instances within the image. The predicted set should include $(bbox_i, score_i, label_i)$, where $bbox_i \in \mathbb{R}^4$ represents the i -th instance bounding box coordinate which includes the upper left and lower right points $[x_1, y_1, x_2, y_2]$. The i -th confidence score, $score_i$, ranges between 0 and 1, while $label_i \in M$ represents the predicted label for the i -th object. The value of M depends on overall dataset object classes and usually includes the 'person' class as part of the object class. The low-score instances will be filtered out by Non Max Suppression (NMS) [15] to improve the quality of the detected instances. The threshold value is set to 0.5, and low-quality detected objects are not considered for further analysis in the detection process. During the second stage, we extract multi-scale feature maps from the object backbone and use RoIAlign [13] to obtain the appearance features of a $(bbox_i)$. Then pair all human and object instances, and classify them as \mathcal{H} set or \mathcal{O} set with the detected label, where $\mathcal{H} = \{label_i = \text{"person"}\}$ and $\mathcal{O} = \{label_i \neq \text{"person"}\}$. For the pair of the human and object instances, we denote the $label_h$ as the human label and $label_o$ as the object label. Earlier studies [36] employed the appearance feature as the node embedding and the spatial pair information as the edge embedding, which were subsequently fed into a graph model. Building on this research, we also feed the node and edge embedding into an HOI graph prediction model.

However, we use a novel semantic graph representation, and the HOI graph prediction model is capable of generating HOI predictions based on a novel representation. The graph model we employed was a bipartite graph proposed by Zhang *et al.* [36].

B. Graph Representation

Our approach involves adopting an *invisible semantic representation* of the human-object pair. This semantic representation could give the graph model some clues to follow the common pattern among different human-object pairs and improve the accuracy of our HOI predictions. Figure 2 illustrates the design of the translational model, which integrates the transfer relation representation onto the node embedding. We embraced the knowledge graph embedding model as the translational model in this research and denoted it as $\mathcal{G}_T = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E} \in \mathbb{R}^M$ is the set of entities and $\mathcal{R} \in \mathbb{R}^N$ is the set of the relations. Typically, in the knowledge graph embedding model, $h \in \mathcal{E}$ denotes the head entity, $r \in \mathcal{R}$ denotes a relation type, and $t \in \mathcal{E}$ denotes the tail entity. Each triplet is formed as $(h, r, t) \in \mathbb{R}^{\mathcal{E} \times \mathcal{R} \times \mathcal{E}}$. In the latent embedding space, each entity $e \in \mathcal{E}$ and $r \in \mathcal{R}$ are represented as d -dimensional vectors. The corresponding embedding representation of the head, relation, and the tail is denoted as $\mathbf{h}, \mathbf{r}, \mathbf{t}$. The dissimilarity of the triplet (h, r, t) within the embedding space is calculated by the score denoted as $s_r(\mathbf{h}, \mathbf{t})$. The set of triplets is correct in terms of ground truth and is represented by golden triplets Δ .

Followed by the TransH [33], we developed a knowledge graph embedding model that transfers the relationship between humans and objects. The design of the triplet is handcrafted, we would like to map all relations between the human and the object onto the relation-specific space. For each relation $r_i \in \{1, \dots, N\}$, where N denotes the total number of possible relationships. We adopted the $label_h$ as the head since all triplets have the same head, the tail will be the $label_o$, and the relation is the r_i . Given a pair between the human and object, we will have N triplets of this pair in the translational model. This translational model enables the diverse distributed representation of the different relations. As illustrated in Figure 3, for the same head and tail, we project the relation-specific vector \mathbf{d}_r into the relation-specific hyperplane, instead of the same embedding space. Then denoted \mathbf{h}, \mathbf{r} , and \mathbf{t} as the corresponding embedding representation, the projection is denoted as \mathbf{h}_\perp and \mathbf{t}_\perp , respectively. If (h, r, t) is a golden triplet, we anticipate that the projection \mathbf{h}_\perp and \mathbf{t}_\perp can be linked through a translation vector, \mathbf{d}_r , with a short distance in the relation-specific hyperplane. The score function of the triplet is defined as

$$s_r(\mathbf{h}, \mathbf{t}) = \|(\mathbf{h} - \mathbf{w}_r^T \mathbf{h} \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{t} - \mathbf{w}_r^T \mathbf{t} \mathbf{w}_r)\|_2^2 \quad (1)$$

If the triplet is a valid one, we expect that the score $s_r(\mathbf{h}, \mathbf{t})$ would be high, whereas it would be low for an invalid one. The translational model encodes the semantic representation of the detected label. During the training of the translational model, the \mathbf{h} and \mathbf{t} aggregate the knowledge of the head and tail, then we will integer translational knowledge into the graph

embedding. Our HOI graph-based approach involves two key types of representations: 1) node embeddings and 2) edge embeddings.

Node Embedding We obtain the instance appearance feature \mathbf{f} from the ROIAlign, then passed the ROI pooled feature of the human/object into a MultiLayer Perceptron (MLP) to get a lower dimension. We denoted \mathbf{f}_h as the human appearance feature and \mathbf{f}_o as the object appearance feature. Then we formulate the node embedding by concatenating the appearance feature \mathbf{f} and entity projection \mathbf{h} (or \mathbf{t}) that we obtain from the translational model. The human node \mathcal{N}_h and object node \mathcal{N}_o could be written as

$$\mathcal{N}_h = \sigma(FC_1(\mathbf{f}_h \oplus \mathbf{h})) \quad (2)$$

$$\mathcal{N}_o = \sigma(FC_2(\mathbf{f}_o \oplus \mathbf{t})) \quad (3)$$

where $\mathbf{f}_h, \mathbf{f}_o \in \mathbb{R}^{1024}$ are linear projection of the appearance feature. The size of embedding \mathbf{h} and \mathbf{t} is determined by the hyper-parameter k . FC denotes the fully connected layer, two FC layers have independent weights, allowing for feature projection. σ is the ReLU activation function. The node embedding $\mathcal{N}_h, \mathcal{N}_o \in \mathbb{R}^{1024+k}$.

Edge Embedding We expect the edge embedding could learn the connection between the pair of human and object nodes. Following the approach proposed in SCG [36], we encode basic spatial information and pairwise information of bounding boxes and normalize them with respect to the corresponding image dimensions. Additionally, we enhance the pairwise connections by encoding the positions ratio, height ratio, and weight ratio of the pairwise bounding boxes, and including the area ratio of the pairwise bounding box as well.

C. Training and Inference

In the translational model training, we utilize a margin-based ranking loss to encourage distinction between correct triplets (golden triplets) and incorrect triplets (which are absent in the dataset). A correct triplet receives a positive translational score, while an incorrect triplet is assigned a negative translational score $s_r(\mathbf{h}, \mathbf{t})$. The following rules are used to determine whether a score is positive $s_r(\mathbf{h}, \mathbf{t})$ or negative $s_r(\mathbf{h}', \mathbf{t}')$:

$$\begin{cases} s_r(\mathbf{h}, \mathbf{t}), & \text{if } (h, t, r) \in \Delta \\ s_r(\mathbf{h}', \mathbf{t}'), & \text{if } (h, t, r) \notin \Delta \end{cases} \quad (4)$$

During the triplet score generation process, the number of negative scores is significantly higher than positive scores. This is because a single human-object pair has only a few interactions at the same time. However, when we developed the translational model, we proposed N triplets to represent N relation-specified relationship. Triplets that not occurring in the dataset are considered negative. However, in order to maintain an equivalent number of positive and negative scores, we randomly select negative triplets in the same quantity as the golden triplets during training. The loss of the translational model is defined as

$$\mathcal{L}_T = \sum_{(h, r, t) \in \Delta} \sum_{(h', r', t') \in \Delta'} [s_r(\mathbf{h}, \mathbf{t}) + \delta - s_r(\mathbf{h}', \mathbf{t}')]_+ \quad (5)$$

TABLE I

COMPARISON WITH STATE-OF-THE-ART ON HICO-DET TEST SET UNDER THE DEFAULT SETTING. SEE SECTION IV FOR THE DEFAULT SETTING. THE LETTERS ‘A’, ‘S’, ‘P’, ‘L’, AND ‘T’ CORRESPOND TO APPEARANCE, SPATIAL, HUMAN POSE, LANGUAGE, AND TRANSLATIONAL FEATURES, RESPECTIVELY. THE TRANSLATION FEATURE REFERS TO THE FEATURE OBTAINED FROM THE TRANSLATIONAL MODEL. FINE-TUNING DETECTION INVOLVES USING A DETECTOR THAT IS FIRST TRAINED ON THE MS-COCO DATASET, AND THEN FINE-TUNED ON THE HICO-DET DATASET.

Architecture	Method	Backbone	Fine-tuned Detection	Feature	Default		
					<i>full(mAP%)</i>	<i>Rare(mAP%)</i>	<i>Non-Rare(mAP%)</i>
Graph-Based	RPNN [40]	ResNet-50	✗	A+P	17.35	12.78	18.71
	VSGNet [29]	ResNet-152	✗	A+S	19.80	16.05	20.91
	SCG [36]	ResNet-50-FPN	✗	A+S	21.69	17.69	22.88
	SKGHOI	ResNet-50-FPN	✗	A+S+T	22.61	15.87	24.62
Interaction Points	PPDM [19]	Hourglass-104	✓	A	21.73	13.78	24.10
	GGNet [39]	Hourglass-104	✓	A	23.47	16.48	25.60
Graph-Based	DRG [7]	ResNet-50-FPN	✓	A+S+L	24.53	19.47	26.04
	SCG [36]	ResNet-50-FPN	✓	A+S	25.63	19.44	27.48
	SKGHOI	ResNet-50-FPN	✓	A+S+T	26.95	21.28	28.56

where $[x]_+ \triangleq \max(0, x)$, we applied the hyper-parameter δ as the margin to divide the set of golden triplets and negative triplets. By applying node and edge embedding, the Spatial-Semantic embedded nodes and edges will pass through the graph network (following the SCG [36] approach) and get the verb prediction scores set $\mathcal{C} = (c_1, \dots, c_j, \dots, c_{\mathcal{H} \times \mathcal{O}})$, where $\mathcal{H} \times \mathcal{O}$ are the total number of all possible pairs of humans and objects indexed by j . To supervise the existence of the corresponding human-object pair detection in all possible pairs, we define

$$p_j = (s_j^h)^\lambda (s_j^o)^\lambda \quad (6)$$

where $(s_j^h)^\lambda$ and $(s_j^o)^\lambda$ denote the detection scores for humans and objects in j -th pair. The λ is set to 1 during training and set to 2.8 during inference to overcome the effect of overconfidence detection scores. From (6), we obtain the interactiveness scores set $\mathcal{P} = (p_1, \dots, p_j, \dots, p_{\mathcal{H} \times \mathcal{O}})$ to indicates the detection probability of each pair. With \mathcal{P} and HOI verb prediction scores \mathcal{C} during the HOI graph model training. We have the final action score as follows:

$$\mathcal{V} = \mathcal{P} \cdot \mathcal{C} \quad (7)$$

Additionally, we have another supervisor $\mathcal{W} = (w_1, \dots, w_j, \dots, w_{\mathcal{H} \times \mathcal{O}})$ \mathcal{E} indicates interaction probability between all graph pairs. It will help to identify whether the j -th pair exists. We use binary focal loss [22] which formulated as:

$$FL(\hat{y}, y) \begin{cases} -\beta(1 - \hat{y})^\gamma \log(\hat{y}), & \text{if } y = 1 \\ -(1 - \beta)\hat{y}^\gamma \log(1 - \hat{y}), & \text{if } y = 0 \end{cases} \quad (8)$$

where \hat{y} is the prediction value between 0 and 1, and ground truth $y \in \{0, 1\}$. So we define the interactive loss as $\mathcal{L}_W = FL(\hat{\mathcal{W}}, \mathcal{W})$, where \mathcal{W} is the ground truth that if there is a pair exists, will be 1, otherwise 0. We also define verbal classification loss as $\mathcal{L}_V = FL(\hat{\mathcal{V}}, \mathcal{V})$, where \mathcal{V} is the ground truth of verb action. The objective function of our proposed network could be expressed as:

$$\mathcal{L} = \mathcal{L}_T + \mathcal{L}_W + \mathcal{L}_V \quad (9)$$

IV. EXPERIMENTS

A. Dataset and Metric

Dataset Our model was tested on the HICO-DET [3] datasets. The HICO-DET is a large-scale benchmark that contains 37,633 training and 9,546 test images, each annotated with a set of human-object interactions and the corresponding object bounding boxes. The interaction types are categorized into 600 classes, and the dataset contains a total of 80 object classes and 117 different action classes. The distribution of pairs across interaction classes is heavily imbalanced, with a long-tailed distribution. Specifically, there are 47 interaction categories that have only one training example.

Metric we use the mean average precision (mAP) metric to assess the accuracy of the predicted human-object interaction (HOI) pair. A predicted HOI instance is considered a true positive if the intersection over union (IoU) between the predicted human bounding box and the ground-truth human bounding box is greater than 0.5, and the IoU between the predicted object bounding box and the ground-truth object bounding box is also greater than 0.5. For every configuration, we provide the mean average precision (mAP) for three sets of HOI classes: the full set of 600 HOI classes (full), a set of 138 HOI classes with fewer than 10 training instances (rare), and a set of 462 HOI classes with 10 or more training instances (non-rare). The default setting for evaluating HOI (human-object interaction) detection involves testing on the entire set of images, including those that do not contain the object being targeted.

Implement Details We utilize the AdaMixer as the object detector with the pre-trained ResNet50-FPN [14] backbone. At the end of the first stage, we filter out the detections if their detected score is lower than 0.2, then perform non-maximum suppression(NMS) in a batched fashion. While training, we exclude pairs of the same person but retain pairs involving different individuals, since we regard the person as an object category as well. While in the second stage, we utilize the multi-scale feature maps obtained from various pyramid levels by employing FPN (Feature Pyramid Network) [21]. Subsequently, ROIAlign is applied to extract the box feature with an output size of 7×7 . The box features are projected into a 1024-dimensional space using a two-layer

MLP, combined with the transfer semantic feature from the translational model to form a node feature. Regardless of the size of the new node representation, it will map to a 1024-dimensional space before being passed into the HOI graph model. The edge representation undergoes the same operation, wherein the spatial feature is transformed into a 1024-dimensional space through a three-layer MLP. Finally, we set $\beta = 0.5$ and $\gamma = 0.2$ for the focal loss, and $\delta = 4$ for the margin-based rank loss.

We train the graph models for 12 epochs using the AdamW optimizer [23] with a batch size of 16, incorporating a momentum of 0.9 and weight decay of 10^{-4} . Also, the object detection was trained with the AdamW optimizer, and the learning rate is set to 2.5×10^{-5} , with 36 epochs and a weight decay of 10^{-4} .

B. Comparison and Analysis

The comparison of our results with state-of-the-art methods on the HICO-DET dataset was shown in the Table 1. Our best model for comparison with other state-of-the-art graph-based approaches is SKGHOI with $k = 50$. As most outperforming approaches require significant computational resources and longer convergence times due to their reliance on transformer architecture in the prediction head, we limited our comparison to graph-based HOI detection methods. While these models deliver good performance, they come with substantial training costs. Graph-based approaches, however, offer a more efficient and computationally tractable alternative, leveraging their structure to achieve high levels of accuracy without the same resource demands.

The present study examines the performance of the SKGHOI algorithm in comparison to other graph-based approaches, both with and without the utilization of a fine-tuned detector. Our findings demonstrate that SKGHOI outperforms other graph-based methods by achieving a mean average precision (mAP) of 0.92, without utilizing the fine-tuned detector. With the utilization of the fine-tuned detector, SKGHOI exhibits superior performance, achieving a significant 1.32 mAP increase over the state-of-the-art method. Additionally, when compared to the interaction point approach, SKGHOI outperforms with a margin of 3.48 mAP, further attesting to its effectiveness in this domain. These results emphasize the efficacy of the SKGHOI algorithm and highlight its potential in HOI detection tasks.

The performance of HOI approaches varies across different settings, revealing notable variations based on the use of a fine-tuned detector versus a detector without fine-tuning and the different feature utilization methods. This underscores the critical role of the object detector in two-stage HOI detections, as all works are built on the detections. In addition, the feature used for the HOI model is also vital, with more information utilized in the graph leading to better performance from the model. It is noted that the approach used in SKGHOI, which utilized not only the appearance and spatial features but also the translation feature, resulted in a higher mAP score of 1.32 compared to SCG [36] that only applied appearance and spatial features. Nonetheless, incorporating additional features into

TABLE II
ABLATION STUDY FOR THE DIFFERENT EMBEDDING SIZES k OF THE TRANSLATIONAL MODEL, THE ‘-’ MEANS WE DIDN’T UTILIZE THE TRANSLATIONAL MODEL, KINDLY REFER TO THE SECTION 4.2

k	full(mAP%)	rare(mAP%)	non-rare(mAP%)
-	26.08	19.93	27.92
30	26.45	20.63	28.19
50	26.95	21.28	28.56
70	26.48	20.20	28.26
100	26.56	21.03	28.21

the model leads to an upsurge in computational expenses. To mitigate this issue, we propose the application of Knowledge Graph Embedding (KGE) for feature translation. Notably, this method incurs minimal costs while providing a highly advantageous feature. In contrast to DRG [7], which leverages word embeddings of detected labels as a 300-dimensional semantic feature, SKGHOI requires only a 50-dimensional semantic feature that is translated using Knowledge Graph Embedding (KGE). In conclusion, our proposed approach not only provides a novel feature but also achieves a higher mAP score in a simple and efficient manner(KGE).

C. Ablation Study

The importance of the embedding size cannot be overstated in knowledge graph embedding models. The size determines the number of dimensions in the continuous vector space utilized to represent entities and relations as low-dimensional embeddings. In the translational model, each relation is linked to a hyperplane in the embedding space, and entity embeddings are projected onto these hyperplanes to generate their representations. To ensure exceptional performance on downstream tasks like link prediction and entity classification, selecting the ideal embedding size is critical. If the embedding size is too small, the model’s expressiveness may be limited, and performance may suffer. Conversely, a too-large embedding size may result in overfitting and slower training. Therefore, it is imperative to conduct an ablation study to examine the impact of different embedding sizes on the model’s performance.

We investigate the impact of different embedding sizes on the performance of a knowledge graph model and show the result in the Table 2. Specifically, we train a knowledge graph embedding model as a translational model on a large-scale image dataset and vary the embedding size from 30 to 100. The finding is that an embedding size of 50 yields the best performance. It is possible that an embedding size of 50 strikes a balance between the capacity of the translational model to capture the complexity of relation knowledge and the potential for overfitting. Notably, without utilizing the translational model to incorporate the supplementary feature, the model performs poorly, exhibiting an mAP point difference of nearly one compared to $k = 50$.

The proportion between the size of the relation embedding and the appearance feature in the node embedding is an important consideration in node embedding. As the size of the translation embedding increases, the proportion of the relation

feature in the node embedding also increases, as the node embedding is a composite of the translation embedding and appearance feature of the node embedding. However, if the proportion of the relation feature becomes too large, it may adversely affect the final performance of the model. Thus, finding the appropriate balance between the appearance and relation features in the node embedding is a crucial task for graph-based models. In our experiments, an embedding size of $k = 50$ resulted in the best performance, suggesting that this particular size strikes a favorable balance between the appearance and relation features in the node embedding.

V. CONCLUSION

This paper presents a novel graph-based approach for human-object interaction detection, which employs a knowledge graph embedding model as the translational model to embed the relation feature and integrate it into the node embedding. By doing so, our approach enhances not only the node representation but also the consistency between the node embedding and edge embedding, leading to improved detection performance. This approach provides a new avenue for researchers to explore the use of knowledge graph embedding models in graph-based approaches for human-object interaction detection.

Future work could explore improved methods for representing knowledge graph embeddings in graph-based approaches. Such methods have the potential to significantly reduce the computational requirements of graph models, making them more efficient and practical for real-world applications. One area in which this could be particularly useful is in transfer learning [17], [25] for human-object interaction detection. Many current datasets used for HOI detection are not representative of the complex interactions that occur in real-world settings. As such, it can be difficult to apply models trained on these datasets to real-world scenarios. To address this, we will explore more effective translational models for graph representation, with the aim of improving transfer learning for HOI detection. This research could ultimately lead to more practical and accurate HOI detection models that can be applied in a range of real-world contexts.

ACKNOWLEDGMENT

This work was supported by the College of Arts and Sciences and the Department of Computer Science at Bowling Green State University. The authors express their sincere gratitude to the colleges who contributed to this work. Specially, the authors thank Dr. Qing Tian, the Assistant Professor of Computer Science at Bowling Green State University, who generously provided computational resource and valuable advice.

REFERENCES

- [1] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26, 2013.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.
- [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)*, pages 381–389. IEEE, 2018.
- [4] Junwen Chen and Keiji Yanai. Qahoi: query-based anchors for human-object interaction detection. *arXiv preprint arXiv:2112.08647*, 2021.
- [5] 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*, pages 9004–9013, 2021.
- [6] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310*, 2018.
- [7] 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*, pages 696–712. Springer, 2020.
- [8] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5364–5373, 2022.
- [9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018.
- [11] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.
- [12] 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*, pages 9671–9685, 2019.
- [13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4507–4515, 2017.
- [16] 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*, pages 584–600. Springer, 2020.
- [17] Jaekoo Lee, Hyunjae Kim, Jongsun Lee, and Sungroh Yoon. Transfer learning for deep learning on graph-structured data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [18] 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*, pages 3585–3594, 2019.
- [19] 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*, pages 482–490, 2020.
- [20] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022.
- [21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

- [23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [24] Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. Fine-grained anchors for human-object interaction detection. *arXiv preprint arXiv:2301.04019*, 2023.
- [25] Shuteng Niu, Yongxin Liu, Jian Wang, and Houbing Song. A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, 1(2):151–166, 2020.
- [26] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 401–417, 2018.
- [27] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.
- [28] 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*, pages 10410–10419, 2021.
- [29] 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*, pages 13617–13626, 2020.
- [30] 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*, pages 9469–9478, 2019.
- [31] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 248–264. Springer, 2020.
- [32] 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*, pages 4116–4125, 2020.
- [33] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28, 2014.
- [34] Xiaoqian Wu, Yong-Lu Li, Xinpeng Liu, Junyi Zhang, Yuzhe Wu, and Cewu Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 121–136. Springer, 2022.
- [35] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose flow: Efficient online pose tracking. *arXiv preprint arXiv:1802.00977*, 2018.
- [36] 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*, pages 13319–13327, 2021.
- [37] 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*, pages 20104–20112, 2022.
- [38] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human-object interaction detection. *International Journal of Computer Vision*, 129:1910–1929, 2021.
- [39] 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*, pages 13234–13243, 2021.
- [40] 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*, pages 843–851, 2019.
- [41] 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*, pages 11825–11834, 2021.

VI. BIOGRAPHY SECTION



Lijing Zhu is a Data Science Ph.D. student at Bowling Green University, having previously obtained her Master's degree in Analysis from the same institution. Her research interests include Human-Object Interaction Detection and Knowledge Graph Embedding. She is dedicated to pursuing excellence in her field and contributing to its advancement. Her peers and mentors appreciate her technical skills and collaborative approach, and she is excited to continue exploring the possibilities of Data Science in the future.



Qizhen Lan is currently pursuing the Ph.D. degree in Data Science at Bowling Green State University, Bowling Green, OH, USA. He received the Master's degree in Intelligence and Analytics in 2019 and the Bachelor's degree in Business Analytics in 2018, both from Bowling Green State University, Bowling Green, OH, USA. He received the Bachelor's degree in Information Management and Information Systems in 2018 from Tiangong University, Tianjin, China. His research interests lie in object detection, knowledge distillation, and deep network pruning.



Alvaro Velasquez is a visiting professor of computer science at the University of Colorado Boulder and a program manager in the Innovation Information Office (I2O) of the Defense Advanced Research Projects Agency (DARPA), where he currently leads programs on neuro-symbolic AI. Before that, Alvaro oversaw the machine intelligence portfolio of investments for the Information Directorate of the Air Force Research Laboratory (AFRL). Alvaro received his PhD in Computer Science from the University of Central Florida in 2018 and is a recipient of the

National Science Foundation Graduate Research Fellowship Program (NSF GRFP) award, the University of Central Florida 30 Under 30 award, a distinguished paper award from AAAI, and best paper and patent awards from AFRL. He has co-authored over 60 papers and two patents and serves as Associate Editor of the IEEE Transactions on Artificial Intelligence. His research has been funded by the Air Force Office of Scientific Research.



Houbing Song (M'12–SM'14–F'23) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, in August 2012.

He is currently a Tenured Associate Professor, the Director of NSF Center for Aviation Big Data Analytics (Planning), and the Director of the Security and Optimization for Networked Globe Laboratory (SONG Lab, www.SONGLab.us), University of Maryland, Baltimore County (UMBC), Baltimore, MD. Prior to joining UMBC, he was a Tenured Associate Professor of Electrical Engineering and Computer Science at Embry-Riddle Aeronautical University, Daytona Beach, FL. He serves as an Associate Editor for IEEE Transactions on Artificial Intelligence (TAI) (2023-present), IEEE Internet of Things Journal (2020-present), IEEE Transactions on Intelligent Transportation Systems (2021-present), and IEEE Journal on Miniaturization for Air and Space Systems (J-MASS) (2020-present). He was an Associate Technical Editor for IEEE Communications Magazine (2017-2020). He is the editor of eight books, the author of more than 100 articles and the inventor of 2 patents. His research interests include cyber-physical systems/internet of things, cybersecurity and privacy, and AI/machine learning/big data analytics. His research has been sponsored by federal agencies (including National Science Foundation, US Department of Transportation, and Federal Aviation Administration, among others) and industry. His research has been featured by popular news media outlets, including IEEE GlobalSpec's Engineering360, Association for Uncrewed Vehicle Systems International (AUUSI), Security Magazine, CXOTech Magazine, Fox News, U.S. News & World Report, The Washington Times, and New Atlas.

Dr. Song is an IEEE Fellow, an ACM Distinguished Member, and an ACM Distinguished Speaker. Dr. Song is a Highly Cited Researcher identified by Clarivate™ (2021, 2022) and a Top 1000 Computer Scientist identified by Research.com. He received Research.com Rising Star of Science Award in 2022 (World Ranking: 82; US Ranking: 16). Dr. Song was a recipient of 10+ Best Paper Awards from major international conferences, including IEEE CPSCo-2019, IEEE ICII 2019, IEEE/AIAA ICNS 2019, IEEE CBDCo 2020, WASA 2020, AIAA/ IEEE DASC 2021, IEEE GLOBECOM 2021 and IEEE INFOCOM 2022.



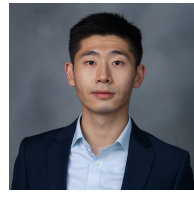
Kamal Acharya (Graduate Student Member, IEEE) received his Engineering degree in Electronics and Communication Engineering from Tribhuvan University, Kathmandu, Nepal in 2011 and Masters degree in Information System Engineering from Purbanchal University, Kathmandu, Nepal in 2019. Currently, he is pursuing PhD. in Electrical Engineering and Computer Science from Embry-Riddle Aeronautical University, Daytona Beach, FL.

He has been involved in teaching profession for about 7 years in the various universities of Nepal, Tribhuvan University and Purbanchal University were among few of them. He is mainly associated with the courses like programming(C,C++,Python), Computer Networks and Computer Architecture. He is working as Graduate Research Assistant in Embry-Riddle Aeronautical University. He is also serving as an reviewer for IEEE Transactions on Artificial Intelligence (TAI) and IEEE Transactions on Intelligent Transportation Systems. His preferred areas of research are Natural Language Processing(NLP), Deep Learning and Reinforcement Learning.



Qing Tian received the BEng degree in computer science and engineering from Yanshan University, Qinhuangdao, Hebei, China in 2011. He received the MEng and PhD degrees in electrical engineering in 2013 and 2021, respectively, from McGill University, Montreal, QC, Canada. He is currently an assistant professor at Bowling Green State University, Bowling Green, OH, USA. From 2019 to 2020, he was an applied scientist intern with Amazon.com, Inc. (Visual Search & AR) at Palo Alto, CA, USA. From 2013 to 2014, he worked as an software

developer at Nakisa Inc., Montreal, QC, Canada. His research interests include autonomous driving visual perception, deep network compression, efficient neural architecture search, and adversarial AI. His current research in efficient and robust self-driving perception is supported by the National Science Foundation.



Shuteng Niu (sniu@bgsu.edu) is an Assistant Professor of Computer Science at Bowling Green State University, Bowling Green, OH. Before joining BGSU, he was a Postdoc Research Fellow at School of Biomedical Informatics UTHealth, Houston, TX. Before that, He received his Ph.D. degree in Computer Science at Embry-Riddle Aeronautical University, Daytona Beach, FL. His research areas are machine learning, deep transfer learning, graph representation, and biomedical informatics.