Visual-Semantic Graph Attention Networks for Human-Object Interaction Detection

Zhijun Liang[1†], Juan Rojas[2†*], Junfa Liu[1], and Yisheng Guan[1*]

***Abstract*— In scene understanding, robotics benefit from not only detecting individual scene instances but also from learning their possible interactions. Human-Object Interaction (HOI) Detection infers the action predicate on a <human, predicate, object> triplet. Contextual information has been found critical in inferring interactions. However, most works only use local features from single human-object pair for inference. Few works have studied the disambiguating contribution of subsidiary relations made available via graph networks. Similarly, few have learned to effectively leverage visual cues along with the intrinsic semantic regularities contained in HOIs. We contribute a dual-graph attention network that effectively aggregates contextual visual, spatial, and semantic information dynamically from primary human-object relations as well as subsidiary relations through attention mechanisms for strong disambiguating power. We achieve comparable results on two benchmarks: V-COCO and HICO-DET. Code is available at https://github.com/birlrobotics/vs-gats.**

I. INTRODUCTION

Human-Object Interaction (HOI) detection has recently gained important traction and has pushed forward robot's abilities to understand the visual world. Whilst computer vision has experienced extraordinary advances in object detection [1]–[3], human pose estimation [4], [5] and scene segmentation [6]; the harder problem of HOI detection has made less progress. Generally, HOI detection starts with instance detection and continue with interaction inference as illustrated in Fig. 1(a). The goal is to infer an interaction predicate for the <human, predicate, object> triplet. Note that humans can simultaneously interact with different objects and have different interactions with the same object. *I.e.* for Fig. 1(b) on the right, the HOIs could be *<human,cut,cake>*, *<human,hold,knife>* and *<human,cut_with,knife>*. Therefore, HOI detection is a multi-label problem, which requires better understanding of contextual information for better inference.

Over time, researchers have exploited a variety of contextual cues including visual, spatial, semantic, human pose to better understand a scene [7]–[15]. Researchers have also used a variety of architectures (Sec. II). But most works have only leveraged local-primary relations in the scene to infer interactions. Very recently graph attention nets [16] have been considered. However, they just use limited contextual cues in a signal graph with complicated features updating mechanism.

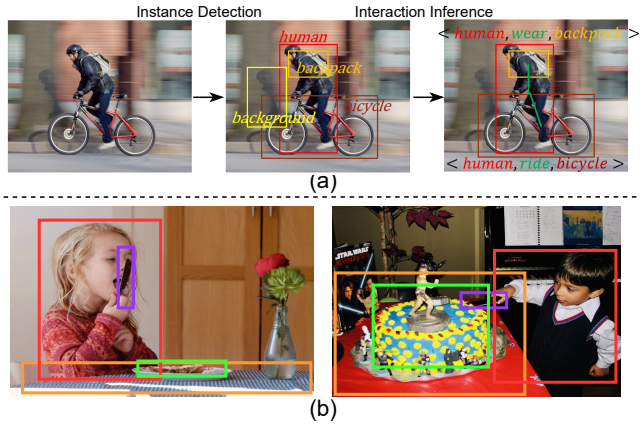

Fig. 1. (a) A general framework of HOI detection. (b) How subsidiary relations facilitate HOI detection: On the left, with the features from *[human-knife]*, the model can easily infer *"hold"* and *"lick"* predicates for this tuple, while the spatial message from subsidiary relations *[knife-table]* inhibits the model from choosing *"cut_with"*. On the right, if we just focus on the features from *[human-cake]*, the model may output similar scores for the *"cut"* and *"light"* predicates since they share similar visual embedding features. However, messages from subsidiary relations *[human-knife]* and *[knife-cake]* promote *<human,cut,cake>*.

Under a graph-based structure, image instance proposals yield graph nodes connected by edges. A primary relation is defined as the immediate human-object relation under consideration; whilst subsidiary relations are all other connections in the graph. In this manner, primary and subsidiary relations are relative. One key insight of this work is that leveraging various contextual cues from subsidiary relations aid to disambiguate in HOI detection. For example, in Fig. 1 consider the *[human-knife]* the primary relation on the left and the *[human-cake]* the primary relation on the right. On the left, this primary relation's visual and spatial cues might predict *"hold"* and *"cut_with"*. But spatial cues from the subsidiary relations *[knife-bread]* inhibit the system from choosing *"cut_with"*. On the right, the primary relation's cues might predict *"cut"* or *"light"* as these actions share similar visual embeddings. However, only when the system pays attention to the *[knife-cake]* and *[human-knife]* subsidiary contextual cues can it easily infer that *"cut"* is the right interaction. Another key insight of our work is that HOIs also posses intrinsic semantic regularities that aid detection despite diverse scenes. For instance, semantic cues

[1]The Biomimetic and Intelligent Robotics Lab (BIRL), School of Electromechanical Engineering, Guangdong University of Technology, 510006 Guangzhou, China. [2]Dept. of Mechanical and Automation Engineering, Chinese University of Hong Kong, Hong Kong, China.

† Equal contribution. * Corresponding authors (ysguan@gdut.edu.cn and juan.rojas@cuhk.edu.cn). The work in this paper is in part supported by the Frontier and Key Technology Innovation Special Funds of Guangdong Province (Grant No. 2017B050506008), the Key R&D Program of Guangdong Province (Grant No. 2019B090915001).

from *human* and *knife*, may help the model focus on the actions related to the *knife* instead of actions like *"ride"*.

In this paper, we study the disambiguating power of subsidiary scene relations and intrinsic semantic regularities via a double graph attention network that aggregates visual-spatial and semantic information in parallel. This graph-based attention network structure explicitly enables the model to leverage rich information by integrating and broadcasting information through graph structure with attention mechanism. We call our system: Visual-Semantic Graph Attention Networks (VS-GATs). As shown in Fig. 2, our method begins by using instance detection to yield bounding-boxes with visual features and semantic categories. From this, a pair of scene graph are created. The first graph's nodes are instantiated from the bounding-box visual features; while the edges are instantiated from corresponding spatial features. The second graph's nodes are instantiated from corresponding word embedding features. Two graph attention networks then update the node features of each graph via message passing. A combined graph is created by concatenating both graph's node updated features. Then inference is done through a readout step on paired human-object nodes. Please see Sec. III for more details.

On HICO-DET dataset, our method achieves comparable results for the Full, Rare and Non-Rare categories with mAP of **20.27**, **16.03** and **21.54** respectively. On V-COCO dataset, our model also obtains promising results with mAP of **50.6**.

II. RELATED WORK

In this section, we present the related works by keying in on the architecture type: multi-streams DNN and GNN.

*a) **Multi DNN Streams with Various Contextual Cues**:* A primary way to do HOI detection has been to extract visual features from instance detectors along with spatial information to instantiate multi-streams of DNNs. Each stream contains detected human, object, and other contextual features. A final fusion step is designed for inference. [7]–[9]. Lu *et al.* [17] considered semantic information under the multi-stream DNN setting stating that interaction relationships are also semantically related to each other. Peyre *et al.* [15] used a concept of visual analogies. They instantiated a stream using a visual-semantic embedding of the triplet resulting in a trigram. Gupta *et al.* [14] and Li *et al.* [10] used fine-grained layouts of the human pose and leverage relation elimination or interactiveness modules to improve inference. Wan *et al.* [13] further considered not only human pose but also human body part features to enhance inference. Recently, some works have been parallelly developed with ours. Hou *et al.* [18] proposed a framework to perform compositional learning by sharing visual components across images for new interaction samples generation. Kim *et al.* [19] extended [14] by explicitly leveraging the action co-occurrence priors for HOI detection. Li *et al.* further explored detailed 2D-3D joint representation [20] and detailed human body part states [21] for better HOI detection. **However**, these works are limited to local features for inference, which do not consider subsidiary relations. In this work, we explore using graph structure network to take the subsidiary relations into account for learning rich contextual information to facilitate HOI detection.

*b) **Graph Neural Networks**:* GNNs have been used to model scene relations and knowledge structures. Yang *et al.* [22] proposed an attentional graph convolution network to aggregate global context for scene graph generation. Sun *et al.* [23], do multi-person action forecasting in video. They use a recursive GNN on visual and spatio-temporal features to update the graph. Kato *et al.* [24] use an architecture that consists of one stream of convolutional features and another stream composed of a semantic graph for HOI classification. Leaning on the concept of semantic regularities, Xu *et al.* [11] similarly use a visual stream with convolutional features for human and object instances and a parallel knowledge graph for HOI detection.

To data, only Qi *et al.* [16] have used GAT architecture for HOI detection. Their method (GPNN) creates nodes and edges from visual features. The graph structure is set by an adjacency matrix and features is updated by a weighted sum of the messages of the other nodes. Finally, a node readout function is used for interaction inference. Our method is similar, but different. **First**, as illustrated in Fig. 2, instead of using single graph our model uses a novel parallel dual-attention graph architecture which also takes semantic cues into account. Furthermore, we identify spatial features as critical in the final inference step. **Second**, we leverage a simpler but more effective node features updating mechanism without multiple iterations using in [16]. A **final** difference is that in [16], Qi *et al.* use a node readout function to *separately* infer actions for each node. We find it more reasonable to jointly infer actions with the combined features of the human and object; as such, we use an edge readout function (Eqtn. 9) to infer the interaction from the edges connected to the human. Overall, our model outperforms GPNN by a great margin on both datasets.

III. VISUAL-SEMANTIC GRAPH ATTENTION NETWORK

In this section, we first define graphs and then describe the contextual features and our VS-GATs.

A. Graphs

A graph G is defined as $G = (V, E)$, where V is a set of n nodes and E is a set of m edges. Node features and edge features are denoted by $\mathbf{h}_v$ and $\mathbf{h}_e$ respectively. Let $v_i \in V$ be the ith node and $e_{i,j} = (v_i, v_j) \in E$ be the directed edge from v_i to v_j. A graph with n nodes and m edges has a node features matrix $\mathbf{X}_v \in \mathcal{R}^{n \times d}$ and an edge feature matrix $\mathbf{X}_e \in \mathcal{R}^{m \times c}$ where $\mathbf{h}_{v_i} \in \mathcal{R}^d$ is the feature vector of node i and $\mathbf{h}_{e_{i,j}} \in \mathcal{R}^c$ is the feature vector of edge (i, j). Fully connected edges imply $e_{i,j} \neq e_{j,i}$.

B. Contextual Features

*a) **Visual Features**:* Visual features are extracted from human and object proposals generated from Faster R-CNN [1]. As with [14], first, the RPN generates human and object

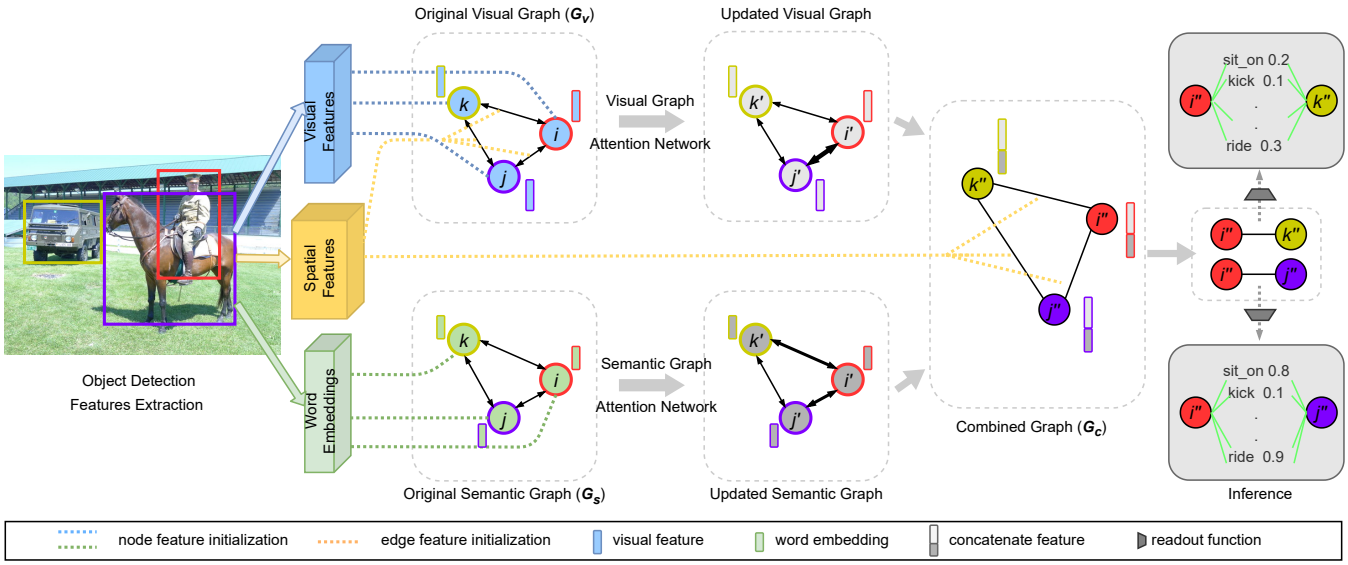

Fig. 2. Visual-Semantic Graph Attention Network: After instance detection, a visual-spatial graph and a semantic graph are created. Node features are dynamically updated through graph attention networks (Sec. III-C). We combine these updated graphs and then perform a readout step on box-pairs to infer all possible predicates between one human and one object.

proposals. Thus, for an image I, the ith human bounding-box b_h^i and the jth object bounding-box b_o^j are used to extract latent features from Faster R-CNNs last fully-connected layer ($FC7$ after the ROI pooling layer) to instantiate the visual graph (G_v) nodes as illustrated in Fig. 2.

*b) **Spatial Features**:* Instance spatial features such as bounding box locations and relative locations are informative about the relationship that proposals have with each other [25]–[28]. Consider the "ride" predicate, we can deduce that human is above the object.

Given a pair of bounding boxes, their paired-coordinates are given by (x_i, y_i, x_j, y_j) and $(x_i^{'}, y_i^{'}, x_j^{'}, y_j^{'})$ and centres are denoted as (x_c, y_c) and $(x_c^{'}, y_c^{'})$. Along with respective areas A and $A^{'}$ and an image area A^I of size (W, H).

Instance spatial features $\mathbf{h}_{f_{ij}} = \mathbf{h}_{f_{ij}^{rs}} \bigcup \mathbf{h}_{f_{ij}^{rp}}$ can be grouped into relative scale features $\mathbf{h}_{f_{ij}^{rs}} = \left[\frac{x_i}{W}, \frac{y_i}{H}, \frac{x_j}{W}, \frac{y_j}{H}, \frac{A}{A^I}\right]$, and relative position features $\mathbf{h}_{f_{ij}^{rp}} = \left[(\frac{x_i - x_i^{'}}{x_j^{'} - x_i^{'}}), (\frac{y_i - y_i^{'}}{y_j^{'} - y_i^{'}}), \log(\frac{x_j - x_i}{x_j^{'} - x_i^{'}}), \log(\frac{y_j - y_i}{y_j^{'} - y_i^{'}}), \frac{x_c - x_c^{'}}{W}, \frac{y_c - y_c^{'}}{H}\right]$. Spatial features are used to: (i) instantiate the edges in the Visual graph (G_v) (ii) and in the Combined Graph (G_c) as illustrated in Fig. 2.

*c) **Semantic Features**:* In this work, we use Word2vec embeddings [29] as semantic features. We use the publicly available Word2vec vectors pre-trained on the Google News dataset (about 100 billion words) [30]. All existing object classes in the dataset are used to obtain the 300-dimensional Word2vec latent vector representations offline. These semantic features are used to instantiate the nodes in the semantic graph (G_s) as illustrated in Fig. 2.

C. Graph Attention Networks

In graph neural networks, a node's features are updated by aggregating its neighboring nodes' features. The node updated features $\tilde{\mathbf{h}}_{v_i}$ for node v_i are generically defined as:

$$\mathbf{a}_{v_i} = f_{aggregate}\left(\{\mathbf{h}_{v_j} : v_j \in \mathcal{N}_i\}\right) \quad (1)$$

$$\tilde{\mathbf{h}}_{v_i} = f_{update}\left(\mathbf{h}_{v_i}, \mathbf{a}_{v_i}\right). \quad (2)$$

Where $\mathcal{N}_i$ is the set of nodes adjacent to v_i. Also, the common $f_{aggregate}(\cdot)$ is averaging:

$$\mathbf{a}_{v_i} = \frac{1}{|\mathcal{N}_i|} \sum_{v_j \in \mathcal{N}_i} \mathbf{h}_{v_j}. \quad (3)$$

1) Visual Graph Attention Network: The visual graph is constructed with visual features and instance spatial features illustrated in Sec. III-B. We first use an edge function $f_{edge}(\cdot)$ to *encode the relation features between two connected nodes*:

$$\mathbf{h}_{e_{ij}} = f_{edge}([\mathbf{h}_{v_i}, \mathbf{h}_{f_{ij}}, \mathbf{h}_{v_j}]). \quad (4)$$

Note that in the first step, the object detector may yield irrelevant instances, then using Eqtn. 3 for node features aggregation might introduces significant noise. Instead, we leverage an attention mechanism to mitigate this problem:

$$\alpha_{ij} = \frac{exp(f_{attn}(\mathbf{h}_{e_{ij}}))}{\sum_{v_o \in \mathcal{N}_i} exp(f_{attn}(\mathbf{h}_{e_{io}}))}. \quad (5)$$

where α_{ij} is the soft weight indicated the importance of node v_j to node v_i via this softmax operation. Then we apply a weighted sum and use the updated function $f_{update}(\cdot)$ to update each node's features:

$$\mathbf{z}_{v_i} = \sum_{v_j \in \mathcal{N}_i} \alpha_{ij}(\mathbf{h}_{v_j} + \mathbf{h}_{e_{ij}}) \quad (6)$$

$$\tilde{\mathbf{h}}_{v_i} = f_{update}([\mathbf{h}_{v_i}, \mathbf{z}_{v_i}]). \quad (7)$$

Note that $\mathbf{z}_{v_i}$ consists of the accumulated latent relation features of all the neighboring node connected to v_i.

At this point, we can get an "updated visual graph" with new features as illustrated in Fig. 2. The different edge thickness' represent the soft weight distributions. Note that in our method, we implement $f_{attn}(\cdot)$, $f_{update}(\cdot)$, and $f_{edge}(\cdot)$ as a single fully-connected layer network with node dimensions of 1, 1024, and 1024 respectively.

2) Semantic Graph Attention Network: In the semantic graph, Word2vec latent representations of the class labels of detected objects are used to instantiate the graph's nodes. We denote $\mathbf{w}_{v_i}$ as the word embedding for node i. As with the visual graph, we use an $f'_{edge}(\cdot)$ function and an $f'_{attn}(\cdot)$ function to compute the distributions of soft weights α'_{ij} on each edge $\alpha'_{ij} = softmax(f'_{attn}(f'_{edge}([\mathbf{w}_{v_i}, \mathbf{w}_{v_j}])))$. Then, the global semantic features for each node are computed through the linear weighted sum:

$$\mathbf{z}'_{v_i} = \sum_{v_j \in N_i} \alpha'_{ij} \mathbf{w}_{v_j}. \tag{8}$$

After that, we update the node's features as $\tilde{\mathbf{w}}_{v_i} = f'_{update}([\mathbf{w}_{v_i}, \mathbf{z}'_{v_i}])$. As with the visual graph, we output an "updated visual graph" with new features as shown in Fig. 2. Similarly, $f'_{edge}(\cdot)$, $f'_{attn}(\cdot)$, and $f'_{update}(\cdot)$ are designed in the same way as with the visual graph.

D. Combined Graph.

To jointly leverage the dynamic information of both the visual (G_v) and the semantic (G_s) GATs, it is necessary to fuse them as illustrated in the "Combined Graph" (G_c) of Fig. 2. We concatenate the features of each of the updated nodes to produce new nodes and initialize the edges with the original $\mathbf{h}_f$ described in Sec. III-B. We denote the combined node features as $\mathbf{h}^c_{v_i}$ for node i, where $\mathbf{h}^c_{v_i} = [\tilde{\mathbf{h}}_{v_i}, \tilde{\mathbf{w}}_{v_i}]$.

E. Readout and Inference.

Through above graph attention networks, *I.e.* for Fig. 1 (b) on the right, human combined node features have encoded relation cues of [human, cake], [human, knife] and [human, table], while knife's have encoded relation cues of [knife, cake], [knife, table] and [knife, human]. Then we box-pair all specific human-object as illustrated in Fig. 2. *In doing so, when predicting the relation between human and cake, the model can not only leverage the cues of [human, cake] but other subsidiary cues from [human, cake], [cake, knife].*

With box-pairing, we finally construct the concatenated representation $\zeta_{ij} = [\mathbf{h}^c_{v_i}, \mathbf{h}_{f_{ij}}, \mathbf{h}^c_{v_j}]$ for prediction. To compute the action category score $\mathbf{s}^a \in \mathcal{R}^k$, where k denotes the total number of possible actions, we apply an edge readout function $f_{readout}(\cdot)$ [1], and then apply a binary sigmoid classifier for each action category:

$$\mathbf{s}^a = sigmoid(f_{readout}(\zeta)). \tag{9}$$

The final score of a triplet's predicate $\mathbf{S_R}$ can be computed through the chain multiplication of the action score $\mathbf{s}^a$, the detected human score s_h and the detected object score s_o from object detection as: $\mathbf{S_R} = s_h * s_o * \mathbf{s}^a$.

[1]A multi-layer perceptron with 2 layers of dimensions 1024 and 117 for HICO-DET, and 1024 and 24 for V-COCO.

Training. For training, we use a multi-class cross-entropy loss that is minimized between action scores and the ground truth action label:

$$\mathcal{L} = \frac{1}{N \times k} \sum_{i=1}^{N} \sum_{j=1}^{k} BCE(s_{ij}, y_{ij}^{label}) \tag{10}$$

where N is the number of all box-pairs in each mini-batch and $s_{ij} \in \mathbf{s}^a_i$. See Sec. IV-A for more training details.

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

Datasets. In this work, we use two common benchmarks: V-COCO [31] and HICO-DET [7]. V-COCO has 2,533, 2,867, 4,946 training, validating, and testing images respectively and 16,199 human instances annotated with 26 action categories. Compared to V-COCO, HICO-DET is much larger and diverse. HICO-DET has 38,118 and 9,658 training and testing images. The 117 interaction classes and 80 objects yield 600 HOI total categories. There are 150K annotated human-object pair instances. HICO-DET is divided in three classes: Full: all 600 categories; Rare: 138 categories with less than 10 training instances, and Non-Rare: 462 categories.

Evaluation Metrics. As with prior works, we use the standard mean average precision (mAP) metric to evaluate the model's detection performance. In this case, we consider a detected result with the form <human, predicate, object> is positive when the predicted verb is true and both the detected human and object bounding boxes have the intersection-of-union (IoU) exceed 0.5 with respect to the corresponding ground truth.

Implementation Details. Our architecture is built on Pytorch and the DGL library [32]. For object detection we use Pytorch's re-implemented Faster R-CNN API [1]. Faster R-CNN use a ResNet-50-FPN backbone [33], [34] trained on the COCO dataset [35]. The object detector and Word2vec vectors are frozen during training. We keep the human bounding-boxes whose detection score exceeds 0.8, while for objects we use a 0.3 score threshold.

All neural network layers in VS-GAT are constructed as MLPs as mentioned in previons. Training on HICO-DET, we use batch size of 32 and a dropout rate of 0.3. We use an Adam optimizer with a initial learning rate of 1e-5. We reduce the learning rate to 3e-6 at 200 epochs and stop training at 250 epochs. As for the activation function, we use a LeakyReLU in all attention network layers and a ReLU elsewhere. For V-COCO dataset, we train the model with the same hyperparameters except for the dropout rate (from 0.3 to 0.5) and the training epoch (from 250 to 600).

B. Results

1) Quantitative Results and Comparisons: Our experiments show our model achieves the promising results on both HICO-DET and V-COCO. For fair comparison, results of all compared baselines are chosen with their object detectors trained on COCO dataset.

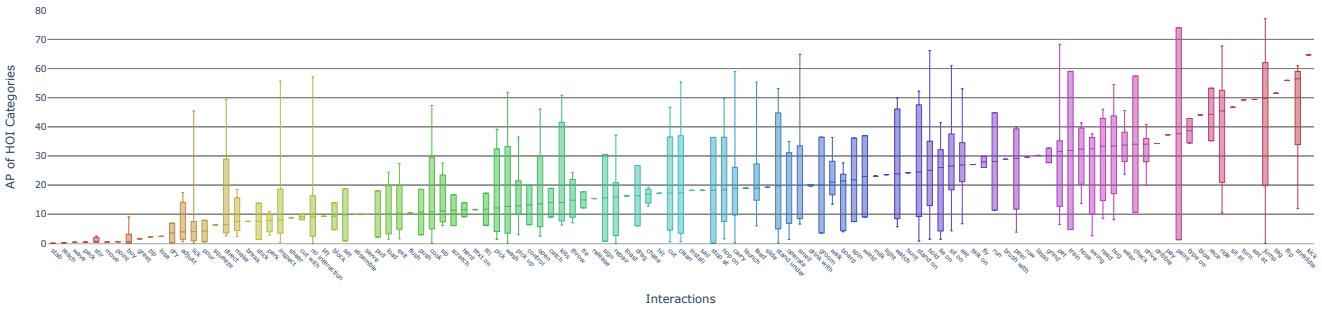

Fig. 3. Spread of performance across objects for a given interaction on HICO-DET. The horizontal axis is sorted by median AP.

TABLE I

MAP PERFORMANCE COMPARISON ON HICO-DET TEST SET.

Method	Full(600)↑	Rare(138)↑	Non-Rare(462)↑
InteractNet [8]	9.94	7.16	10.77
GPNN [16]	13.11	9.34	14.23
iCAN [9]	14.84	10.45	16.15
Xu *et al.* [11]	14.70	13.26	15.13
Bansal *et al.* [12]	16.96	11.73	18.52
Gupta *et al.* [14]	17.18	12.17	18.68
Li *et al.* [10]	17.22	13.51	18.32
PMFNet [13]	17.46	15.65	18.00
Peyre *et al.* [15]	19.40	14.60	20.90
VCL [18]	19.43	16.55	20.29
PaStaNet*-Linear [21]	19.52	17.29	20.19
ACP [19]	20.59	15.92	21.98
DJ-RN [20]	21.34	18.53	22.18
Ours(VS-GATs)	20.27±0.10	16.03±0.42	21.54±0.02

TABLE II

MAP PERFORMANCE COMPARISON ON V-COCO TEST SET.

Method	AP_{role} (Scenario 1)
Gupta *et al.* [31]	31.8
InteractNet [8]	40.0
GPNN [16]	44.0
iCAN [9]	45.3
Xu *et al.* [11]	45.9
Li *et al.* [10]	47.8
VCL [18]	48.3
PMFNet [13]	52.0
PMFNet (w/o pose) [13]	48.6
ACP [19]	53.0
Ours (VS-GATs)	50.6

On HICO-DET, we train our model three times to get the more convincing results. Our model obtains comparable results in all three categories with map of 20.27, 16.03, 21.54 respectively. We surpass GPNN that also uses GNN by 7.16 mAP gain. We also outperform most of multi-streams works [8]–[15], [18], [21] including works that leveraged human pose [10], [13], [14], [21] except [19] and [20]. However, DJ-RN [20] is resource-consuming for further leveraging the expensive 2D even 3D pose estimator for detailed joint representation. And ACP [19] is labor-intensive for explicitly constructing the action co-occurrence prior which is various in different datasets. Meanwhile, it also inserts the pose detector. However, our method tries to learn co-occurrence relationship implicitly by data-driven, and also demonstrates better disambiguating power with 20.27 mAP favorable result even without extra pose cues.

On V-COCO, VS-GATs achieves 50.6 mAP which exceeds GPNN by 6.6 mAP and also outperforms most of other STOAs using multi-streams network except [13] and [20]. [13] develops a well-defined *Zoom-in Module* which utilizes *fine-grained human pose* as well as *body part features* to extract detailed local appearance cues, which makes their model surpass [10] by a great margin on the small-scale dataset. However, [13] and [10] have a similar performance on the more diverse dataset HICO-DET. Without the *Zoom-in Module*, [13] obtains the result of 48.6, which worse than our model. ACP [19] which also use pose cues set up a new STOA result 53.0 mAP, which indicates explicitly leveraging the co-occurrence prior is more effective on the small-scale dataset than all other methods that explore this cues implicitly.

In Fig. 3, we also visualize the performance distribution of our model across objects for a given interaction. As mentioned in [14], it still holds that interactions that occur with just a single object (e.g. 'kick ball' or 'flip skateboard') are easier to detect than those predicates that interact with various objects. Compared to [14], the median AP of those interactions like 'cut' and 'clean' shown in Fig. 3 outperform those in [14] by a considerable margin. We hold that gains from our works are due to learning the rich contextual relationship by the dual attention graphs which enable each node to leverage contextual cues from a wide-spread set of (primary and subsidiary) other nodes as illustrated in III-E.

By comparation, we note that human pose cues are informative for HOI inferring [13], [19], [20] and the object detector also play a important role in HOI detection [12], [18]. We will try to export those in our future works.

2) Qualitative Results: Fig. 4 shows some <human, predicate, object> triplets' detection results on HICO-DET test set. From the results, our proposed model is able to detect various kinds of HOIs such as: single person-single object, multi person-same object, and multi person-multi objects.

C. Ablation Studies

In this section, we choose HICO-DET as training dataset to study the impact of each component in our model. We

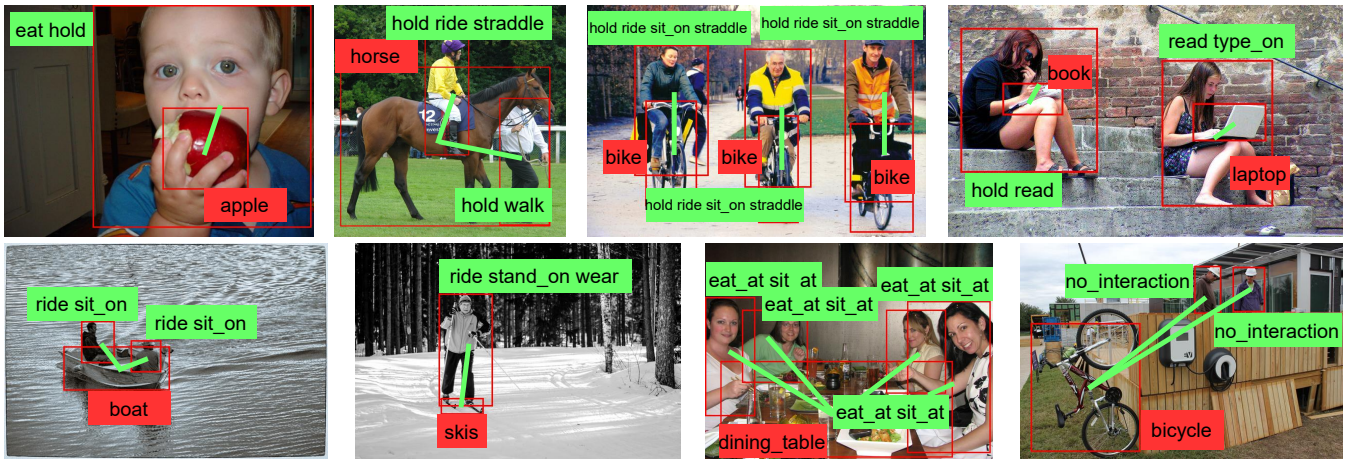

Fig. 4. HOI detection examples on the HICO-DET testing dataset. Humans and objects are shown in red bounding boxes. Interaction classes are shown on the human bounding box. Interactive objects are linked with the green line. We show all triplets whose inferred *action score* (9) is greater than 0.3.

TABLE III
MAP PERFORMANCE FOR VARIOUS ABLATION STUDIES.

Method	Full↑	Rare↑	Non-Rare↑
Ours(VS-GATs)	**20.27**±0.10	**16.03**±0.42	**21.54**±0.02
01 G_V only	19.68±0.05	15.42±0.25	20.96±0.03
02 G_S only	14.34±0.02	10.91±0.16	15.36±0.03
03 w/o attention	19.75±0.05	14.84±0.31	21.21±0.05
04 w/o $\mathbf{h}_f$ in G_C	19.32±0.06	15.54±0.15	20.45±0.06
05 Message passing in G_C	20.04±0.14	15.24±0.22	21.47±0.18
06 Unified V-S graph	19.92±0.07	15.46±0.66	21.25±0.12

conduct the following six tests (we run each test three times to get the results):

01 Visual Graph Only: G_V only. In this test, we remove the Semantic-GAT and keep the Visual-GAT, attention, and inference the same. This study will show the importance of leveraging the semantic cues.

02 Semantic Graph Only: G_S only. In this test, we remove the Visual-GAT and keep the Semantic-GAT, attention, and inference the same. This study will show the importance of aggregating visual and spatial cues.

03 Without Attention. In this test, we use the averaging mechanism (Eqtn. 3) to aggregate features instead of the weighted sum attention mechanism.

04 Without Spatial Features ($\mathbf{h}_f$) in G_C. In this test, we remove spatial features from the edges of the combined graph G_C to study the role that spatial features can play after the aggregation of features across nodes.

05 Message Passing in G_C. In this test, we leverage an additional graph attention network to process the combined graph which is similar to what we do to the original visual-spatial graph. We examine if there would be a gain from an additional message passing on G_C with combined feature from G_V and G_S.

06 Unified V-S Graph. In this test, we choose to start with a single graph in which visual and semantic features are concatenated in the nodes from the start. Spatial features are still used to instantiate edges. This test examines if there would be a gain from using combined visual-semantic features from the start instead of through separate streams.

We now report on the ablation test results. For the Full category, study 01 yields an mAP of 19.68 which is a large portion of our full model mAP result suggesting that semantic cues can promote HOI detection but less effective than visual cues. When only using the Semantic graph in test 02, the effect is less marked, which also indicates the visual and spatial features play a primary role in inferring HOI. When combining these 3 contextual cues in a graph but not using the attention mechanism in test 03, the map results drop to 19.75. This suggests that attention mechanism assists the model for better HOI detection. Afterwards, inserting attention but removing spatial features at the end in test 04 hurts. This indicates that spatial features, even after the aggregation stage, are helpful. By inserting spatial features in the combined graph we are basically using a skip connection step in neural networks which has also shown to help classification. In test 05, we learn that additional message passing in the combined graph does not confer additional benefits. Similarly with test 06, a combined V-S graph is still not as effective as separating cues early on. This suggests that visual cues and semantic cues may have some degree of orthogonality to them even though they are related to each other.

V. CONCLUSION

In this paper we present a novel HOI detection architecture that studies and leverages the role of not only primary human-object contextual cues in interaction, but also the role of subsidiary relations. We show that multi-modal contextual cues from visual, semantics, and spatial data can be graphically represented through graph attention networks to leverage primary and subsidiary contextual relations. Our work have a promising results on both HICO-DET and V-COCO dataset.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *In NIPS*, pp. 91–99, 2015. 1, 2, 4

[2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *In ECCV*, pp. 21–37, 2016. 1

[3] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," *In NIPS*, pp. 379–387, 2016. 1

[4] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, "Learning 3D human pose from structure and motion," *In ECCV*, pp. 668–683, 2018. 1

[5] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," *In CVPR*, pp. 7753–7762, 2019. 1

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *In ICCV*, pp. 2961–2969, 2017. 1

[7] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," *In WACV*, pp. 381–389, 2018. 1, 2, 4

[8] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and Recognizing Human-Object Interactions," *In CVPR*, pp. 8359–8367, 2018. 1, 2, 5

[9] C. Gao, Y. Zou, and J. B. Huang, "ICAN: Instance-centric attention network for human-object interaction detection," *In BMVC*, 2018. 1, 2, 5

[10] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y.-F. Wang, and C. Lu, "Transferable Interactiveness Knowledge for Human-Object Interaction Detection," *In CVPR*, 2018. 1, 2, 5

[11] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to Detect Human-Object Interactions with Knowledge," *In CVPR*, pp. 2019–2028, 2019. 1, 2, 5

[12] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting Human-Object Interactions via Functional Generalization," *In arXiv preprint arXiv:1904.03181v1*, 2019. 1, 5

[13] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware Multi-level Feature Network for Human Object Interaction Detection," *In ICCV*, pp. 9469–9478, 2019. 1, 2, 5

[14] T. Gupta, A. Schwing, and D. Hoiem, "No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques," *In ICCV*, 2019. 1, 2, 5

[15] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," *In ICCV*, pp. 1981–1990, 2019. 1, 2, 5

[16] S. Qi, W. Wang, B. Jia, J. Shen, and S. C. Zhu, "Learning human-object interactions by graph parsing neural networks," *In ECCV*, pp. 407–423, 2018. 1, 2, 5

[17] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," *In ECCV*, pp. 852–869, 2016. 2

[18] H. Zhi, P. Xiaojiang, Q. Yu, and T. Dacheng, "Visual compositional learning for human-object interaction detection," *In ECCV*, 2020. 2, 5

[19] K. Dong-Jin, S. Xiao, C. Jinsoo, L. Stephen, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," *In ECCV*, 2020. 2, 5

[20] L. Yong-Lu, L. Xinpeng, L. Han, W. Shiyi, L. Junqi, L. Jiefeng, and C. Lu, "Detailed 2d-3d joint representation for human-object interaction," *In CVPR*, 2020. 2, 5

[21] L. Yong-Lu, X. Liang, L. Xinpeng, H. Xijie, X. Yue, W. Shiyi, F. Hao-Shu, M. Ze, C. Mingyang, and L. Cewu, "Pastanet: Toward human activity knowledge engine," *In CVPR*, 2020. 2, 5

[22] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for Scene Graph Generation," *In ECCV*, pp. 690–706, 2018. 2

[23] C. Sun, A. Shrivastava, C. Vondrick, R. Sukthankar, K. Murphy, and C. Schmid, "Relational Action Forecasting," *In CVPR*, pp. 273–283, 2019. 2

[24] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," *In ECCV*, pp. 234–251, 2018. 2

[25] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards Context-Aware Interaction Recognition for Visual Relationship Detection," *In ICCV*, pp. 589–598, 10 2017. 3

[26] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," *In CVPR*, pp. 1115–1124, 2017. 3

[27] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, "Phrase localization and visual relationship detection with comprehensive image-language cues," *In ICCV*, pp. 1928–1937, 2017. 3

[28] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," *In CVPR*, pp. 5532–5540, 2017. 3

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *In NIPS*, pp. 3111–3119, 2013. 3

[30] Google, "Google Code Archive - Long-term storage for Google Code Project Hosting." 2013. [Online]. Available: https://code.google.com/archive/p/word2vec/ 3

[31] S. Gupta and J. Malik, "Visual semantic role labeling," *In arXiv preprint arXiv 1505.04474*, 2015. 4, 5

[32] M. Wang, L. Yu, Z. Da, G. Quan, G. Yu, Y. Zihao, L. Mufei, Z. Jinjing, H. Qi, M. Chao, H. Ziyue, G. Qipeng, Z. Hao, L. Haibin, Z. Junbo, L. Jinyang, S. Alexander, and Z. Zheng, "Deep graph library: Towards efficient and scalable deep learning on graphs," *In ICLR Workshop on RepresentationLearning on Graphs and Manifolds*, 2019. 4

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In CVPR*, pp. 770–778, 2016. 4

[34] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *In CVPR*, pp. 936–944, 2017. 4

[35] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," *In ECCV*, pp. 740–755, 2014. 4