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

Zhijun Liang¹, Junfa Liu¹, Yisheng Guan¹ and Juan Rojas²* 

Abstract—In scene understanding, robots 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 pairs for inference. Few works have studied the disambiguating contribution of subsidiary relations made available via graph networks. Similarly, few have leveraged visual cues with the intrinsic semantic regularities embedded in HOIs. We contribute Visual-Semantic Graph Attention Networks (VS-GATs): a dual-graph attention network that effectively aggregates visual, spatial, and semantic contextual information dynamically from primary human-object relations as well as subsidiary relations through attention mechanisms for strong disambiguating power. We achieve competitive results on two benchmarks: V-COCO and HICO-DET. The 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.

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 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 very competitive 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 very competitive 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 developed in parallel 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], did multi-person action forecasting in video. They used a recursive GNN on visual and spatio-temporal features to update the graph. Kato et al. [24] used an architecture that consisted 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 used a visual stream with convolutional features for human and object instances and a parallel knowledge graph for HOI detection.

To date, 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 are 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 differs in a few ways: (i) instead of a single graph, we use a parallel dual-attention graph to synergize visual and semantic information; (ii) we identify spatial features as critical in the final inference step; (iii) we leverage a simpler but more effective node features updating mechanism that does not require multiple compute iterations as in [16]; and finally (iv) as opposed to [16], which uses a node readout function to separately infer actions for each node; we jointly infer actions with the combined human and object features via an edge readout function (Eqt. 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 $h_v$ and $h_e$ respectively. Let $v_i \in V$ be the $i$th 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 feature matrix $X_v \in \mathbb{R}^{n \times d}$ and an edge feature matrix $X_e \in \mathbb{R}^{m \times d}$ where $h_{v_i} \in \mathbb{R}^d$ is the feature vector of node $i$ and $h_{e_{i,j}} \in \mathbb{R}^c$ 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
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.

As with [14], the RPN first generates human and object proposals. Thus, for an image $I$, the $i$th human bounding-box $b_h^i$ and the $j$th object bounding-box $b_o^j$ are used to extract latent features from Faster R-CNN’s last fully-connected layer ($FC^7$) 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 $h_{f_{ij}} = h_{f_{ij}}^{rs} \cup h_{f_{ij}}^{rp}$ can be grouped into relative scale features $h_{f_{ij}}^{rs} = \left[ \frac{x_j-x_i}{W}, \frac{y_j-y_i}{H}, \frac{x_j'-x_i'}{W}, \frac{y_j'-y_i'}{H} \right]$, and relative position features $h_{f_{ij}}^{rp} = \left[ (x_j-x_i), (y_j-y_i), \log(\frac{x_j-x_i}{y_j-y_i}), \log(\frac{x_j'-x_i'}{y_j'-y_i'}), \frac{x_j-x_i}{W}, \frac{y_j-y_i}{H} \right]$.

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 first aggregating the neighboring nodes features in an equivariant manner, and then using another function to update the resulting aggregation with the original node. In general, the updated features $h_{v_i}$ for node $v_i$ are defined as:

$$a_{v_i} = f_{aggregate} \left( \{ h_{v_j} : v_j \in N_i \} \right)$$  
$$h_{v_i} = f_{update}(h_{v_i}, a_{v_i})$$

Where $N_i$ is the set of nodes adjacent to $v_i$ and $f_{aggregate}(\cdot)$ is commonly defined as averaging:

$$a_{v_i} = \frac{1}{|N_i|} \sum_{v_j \in N_i} h_{v_j}.$$  

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

$$h_{v_{ij}} = f_{edge}(\{ h_{v_i}, h_{f_{ij}}, h_{w_j} \}).$$

Given that the object detector may yield irrelevant instances in step 1, using an averaging feature aggregator like Eqtn. 3 would introduces significant noise. Instead, we leverage a softmax-based attention mechanism to mitigate this problem:

$$\alpha_{ij} = \frac{\exp(f_{attn}(h_{v_{ij}}))}{\sum_{v_j \in N_i} \exp(f_{attn}(h_{v_{ij}}))},$$

where $\alpha_{ij}$, a soft weight, indicates the importance of node $v_j$ to node $v_i$. Then, for the update step, we first compute
a weighted sum which is later used in the update function \( f_{\text{update}}(\cdot) \) to update each node’s features:

\[
z_{v_i} = \sum_{v_j \in N_i} \alpha_{ij}(h_{v_j} + h_{v_{ij}}) \tag{6}
\]

\[
\tilde{h}_{v_i} = f_{\text{update}}(h_{v_i}, z_{v_i}). \tag{7}
\]

Note that \( z_{v_i} \) 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 \( f_{\text{attn}}(\cdot) \), \( f_{\text{update}}(\cdot) \), and \( f_{\text{edge}}(\cdot) \) as single fully-connected layer networks 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 \( w_{v_i} \) as the word embedding for node \( i \). As with the visual graph, we use \( f_{\text{edge}}(\cdot) \) function and an \( f_{\text{attn}}(\cdot) \) function to compute the distributions of soft weights \( \alpha_{ij}^{s} \) on each edge \( \alpha_{ij}^{s} = \text{softmax}(f_{\text{attn}}(f_{\text{edge}}([w_{v_i}, w_{v_j}]))) \). Then, the global semantic features for each node are computed through the linear weighted sum:

\[
z'_{v_i} = \sum_{v_j \in N_i} \alpha_{ij}^{s} w_{v_j}. \tag{8}
\]

After that, we update the node’s features as \( \tilde{w}_{v_i} = f'_{\text{update}}([w_{v_i}, z_{v_i}]) \). As with the visual graph, we output an “updated visual graph” with new features (see Fig. 2) and have networks \( f'_{\text{edge}}(\cdot), f'_{\text{attn}}(\cdot), \) and \( f'_{\text{update}}(\cdot) \) of same dimensions as in 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 \( h_f \) described in Sec. III-B. We denote the combined node features as \( h'_{v_i} \) for node \( i \), where \( h'_{v_i} = [h_{v_i}, \tilde{w}_{v_i}] \).

E. Readout and Inference.

After the combined graph, we box-pair all specific human-object relations as shown in Fig. 2. To illustrate the possibilities, consider the right side of Fig. 1 (b) where the combined nodes encode the human node relations: [human, cake], [human, knife], [human, table], and the cake node relations: [cake, knife], [cake, table] and [cake, human]. When the human-cake box-pair is set up, then for the HPI prediction, we will not only leverage the cues of [human, cake] but also from the subsidiary cues [human, knife] and [cake, knife].

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

\[
s^a = \text{sigmoid}(f_{\text{readout}}(\zeta)). \tag{9}
\]

The final score of a triplet’s predicate \( S_{R} \) can be computed through the chain multiplication of the action score \( s^a \), the detected human score \( s_h \) and the detected object score \( s_o \) from object detection as:

\[
S_{R} = s_h \ast s_o \ast s^a.
\]

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

\[
\mathcal{L} = \frac{1}{N \times k} \sum_{i=1}^{N} \sum_{j=1}^{k} \text{BCE}(s_{ij}, y^\text{label}_{ij}) \tag{10}
\]

where \( N \) is the number of all box-pairs in each mini-batch and \( s_{ij} \in \mathbb{R}^a \). 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 2533, 2867, and 4946 training, validating, and testing images respectively as well as 16,199 human instances annotated with 26 action categories. Compared to V-COCO, HICO-DET is much larger and diverse. HICO-DET has 38118 and 9658 training and testing images. There are 117 interaction classes and 80 objects. Together, 600 HOI categories are possible and there are a total of 150K annotated human-object pair instances. Furthermore, HICO-DET distinguishes amongst the classes with few training instances. A ‘Rare’ category is assigned when less than 10 training instances are available. This happens in 138 categories. ‘Non-Rare’ have more (462 categories). Together, you have the ‘Full’ set of 600 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 triplet \( <\text{human, predicate, object}> \) positive when the predicted verb is true and both the detected human and object bounding boxes have an intersection-of-union (IoU) that exceeds 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 previously) with a batch size of 32 and a dropout rate of 0.3. LeakyReLUs are used to activate all

\(^1\)A multi-layer perceptron with 2 layers of dimensions 1024 and 117 for HICO-DET, and 1024 and 24 for V-COCO.
attention network layers and ReLU elsewhere. Also an Adam optimizer starts with a learning rate of 1e-5 and decreased to 3e-6 at 200 epochs. Training stops at 250 epochs.

For the V-COCO dataset, the same hyperparameters are used for training except the dropout rate (0.5) and the number of training epochs (600).

B. Results

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

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 (which also uses GNNs) by 7.16 mAP points. 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 as it further leverages an expensive 2D and 3D pose estimator for detailed joint representation; and ACP [19] is labor-intensive as it explicitly constructs the action co-occurrence prior which varies across datasets and also uses the pose detector. Our method, on the other hand, learns co-occurrence implicitly from the data and demonstrates better disambiguating power with a favorable 20.27 mAP without the extra pose cues.

On V-COCO, VS-GATs achieves 50.6 mAP which exceeds GPNN by 6.6 mAP points and also outperforms most of the other SOTAs 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 is worse than our model. ACP [19] which also uses pose cues, achieve a new SOTA result of 53.0 mAP. The latter indicates that explicitly leveraging the co-occurrence prior is more effective on the smaller scale dataset than all other methods that explore this cue 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 comparison, we note that human pose cues are informative for HOI inference [13], [19], [20] and that the object detector also plays an important role in HOI detection [12], [18]. We will try to leverage those in our future works.

2) Qualitative Results: Fig. 4 shows some <human, predicate, object> triplet detection results on the HICO-DET test set. From the results, our model is detects: single person, single object, multi person-same object, and multi person-multi objects.

C. Ablation Studies

For ablations, we choose the HICO-DET training dataset to study the impact of each component in our model. We conduct the following six tests (each test runs three times to get the results):

| 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 | 14.60 | 20.90 |
| 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 |

| 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 |

TABLE I

MAP PERFORMANCE COMPARISON ON HICO-DET TEST SET.

TABLE II

MAP PERFORMANCE COMPARISON ON V-COCO TEST SET.

TABLE III

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

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.

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 ($h_f$) 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.

The ablation results for the Full category are as follows: study 01 yields an mAP of 19.68 which is a large portion of our full model mAP result. This suggests that semantic cues promote HOI detection but less so than visual cues. When only using the Semantic graph in test 02, the effect is less marked, indicating that the visual and spatial features play a primary role in inferring HOIs. When combining these 3 contextual cues in a graph but not using attention in test 03, the map results drop to 19.75. This suggests that attention assists in improving HOI detection. Afterwards, using attention but removing spatial features at the end (in test) 04 hurts. Spatial features, even after the aggregation stage, are still helpful. Inserting spatial features in the combined graph acts as inserting a skip connection step in the NN, which has been shown to help in classification. In test 05, 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.

VI. ACKNOWLEDGEMENTS

This work is supported by the generous support of the Hong Kong Center for Logistic Robotics, the Chinese University of Hong Kong Direct Grant No 4055141 and the National Natural Science Foundation of China Grant No 62176154.

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, 3, 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. Darbala, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, and A. Jain, “Learning 3D human pose from shape and motion,” in *ECCV*, pp. 668–683, 2018. 1
[5] D. Pavlo, 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 Interactivity 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, 3, 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. Dancheng, “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. Zhang, 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 Representation Learning 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