Visual-Semantic-Pose Graph Mixture Networks for Human-Object Interaction Detection

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

***Abstract*— Human-Object Interaction (HOI) Detection infers the action predicate on a <subject, predicate, object> triplet. Whilst contextual information has been found critical in this task, even with the advent of deep learning, researchers still grapple to understand how to best leverage contextual cues for inference. What is the best way to integrate visual, spatial, semantic, and pose information? Many works have used a subset of cues or limited their analysis to single subject-object pair for inference. Few works have studied the disambiguating contribution of subsidiary relations made available via graph networks. In this work, we contribute a two-stream (multi-branched) network that effectively aggregates a series of contextual cues. In a first study, we propose a dual graph attention network to dynamically aggregate the visual, instance spatial, and semantic cues from primary subject-object relations as well as subsidiary ones to enhance inference. Subsequently, we incorporate human pose features and propose a second network stream that runs a pose-based modular network. The latter is composed of dual branches that run a graph convolutional network and multi-layer perceptrons to improve detection in crowded scenes. The result is a graph mixture network that processes a wide set of contextual cues effectively. We call our model: Visual-Semantic-Pose Graph Mixture Networks (VSP-GMNs). Our final model outperforms state-of-the-art on the challenging HICO-DET dataset by significant margins of almost 10%, especially in long-tail cases that are harder to interpret. We also achieve a competitive performance on the smaller V-COCO dataset. Code, video, and supplementary material information are available at `www.juanrojas.net/VSPGMN`.**

I. INTRODUCTION

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. HOI detection typically begins with instance detection followed by interaction inference. The goal is to infer an interaction predicate for the <subject, predicate, object> triplet in a multi-label setting. Humans can simultaneously interact with different objects and can also have different interactions with the same object. *I.e.* in Fig. 1a, the HOI triplets could be <*human, lick, knife*>, <*human, hold, knife*> and <*human, cut_with, knife*>. Researchers have exploited a variety of contextual cues including visual, spatial, semantic and human pose to better understand a scene [7]–[14]. Researchers have also used a variety of network architectures (Sec. II). However, most works have only leveraged local-primary relations in the scene to infer interactions. Very recently graph attention networks (GATs) [15] have been considered; however, they have been used with a limited set of contextual cues with single-branch networks.

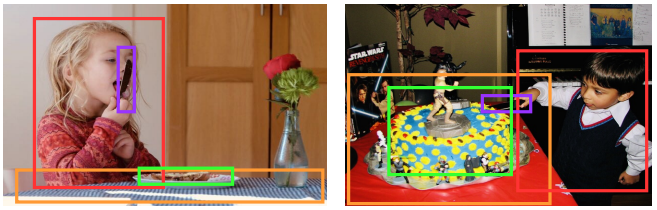

Fig. 1. Examples of HOI. The left image depicts a child, a knife, and a cake. Licking and holding a knife are easily detected? How about cutting the bread? Subsidiary spatial cues from [knife-bread] predict *"cut"* is unlikely.

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"* or *"cut"*. But cues from the subsidiary relations *[knife-bread]*, *[human-table]* inhibit the system from choosing *"cut"*. On the right, the primary relation's cues might predict *"cut"* or *"light"* as these actions share similar embeddings. However, when the system considers the *[knife-cake]* and *[human-knife]* subsidiary contextual cues does it 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 that might seem likely from visual cues such as *"ride"*. In this work, we leverage them through another independent graph attention network. We note that the visual-spatial cues and semantic ones behave as being orthogonal to each other (Sec. IV-C). As such, instead of processing the cues in a single GAT branch, we design a dual GAT network

[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, 2019A050510040), and the National Science Foundation of China (Grant No. 61950410758).

for inference yielding enhanced performance.

Furthermore, recent work in human pose related information has shown to posses informative context [10], [12], [13]. Particularly, to improve inference results in crowded scenes, we set up a follow-up study, in which we incorporate fine-grained human pose information. An additional stream (attached in parallel to the visual-semantic GAT branch) is introduced to processes absolute and relative human pose features via a dual-branch consisting of a graph convolutional network and multi-layer perceptrons.

As such we contribute the Visual-Semantic-Pose Graph Mixture Network which consists of two streams: one with a Visual-Semantic Graph Attention network (VS-GAT) and one with the Pose-based Modular Network (PMN). This paper makes three major contributions. **First**, we propose a novel dual-graph attention network to leverage rich relation information by taking human-object subsidiary relations into account inclusive of intrinsic semantic regularities to improve HOI detection. Additionally, the dual network design allows embeddings to reach a more appropriate level of abstraction before fusing them to yield better inference. **Second**, we also propose a simple but effective network to leverage the pose cues to aid the system to better perform in the crowded scenes. **Third**, our whole network processes a wide set of contextual cues effectively and surpasses state-of-the-art results on HICO-DET by almost 10% and also get promising results on V-COCO.

II. RELATED WORK

*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.* [16] considered semantic information under the multi-stream DNN setting stating that interaction relationships are also semantically related to each other. Peyre *et al.* [14] used a concept of visual analogies. They instantiated a stream using a visual-semantic embedding of the triplet resulting in a triagram. Gupta *et al.* [13] 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.* [12] further considered not only human pose but also human body part features to enhance inference. However, these works are limited to local features for inference, which do not consider subsidiary relations. In our first study, 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.* [17] proposed an attentional graph convolution network to aggregate global context for scene graph generation. Sun *et al.* [18], do multi-person action forecasting in video. They use a RecGNN based on visual and spatio-temporal features to create and update the graph. Kato *et al.* [19] 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 date, only Qi *et al.* [15] 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 message updates leverage attention mechanisms via a weighted sum of the messages of the other nodes. Finally, a node readout function is used for interaction inference.

VS-GAT is similar, but different. First, as illustrated in Fig. 2, instead of using a single graph, the first stream of our network, 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-feature updating mechanism that does not require multiple iterations as in [15]. Finally, in contrast to [15], which uses 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 the object. As such, we use an edge readout function (Eqtn. 9) to infer the interaction from the edges connected to the human. Overall, VS-GAT outperforms GPNN by a great margin on both HICO-DET and V-COCO. Moreover, in our second study, we incorporate pose cues which improve inference accuracy.

III. STREAM 1: VISUAL-SEMANTIC GRAPH ATTENTION NETWORK

In our first study, we introduce a visual-semantic graph attention network. We will describe the visual and semantic instantiations, attention mechanisms, fusion step, inference, training, and implementation details.

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 subject and object proposals generated from Faster-RCNN [1]. First, the RPN generates subject and 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-RCNNs last fully-connected layer ($FC7$ after the ROI pooling layer) to instantiate the visual graph (G_v) nodes as illustrated in Fig. 2.

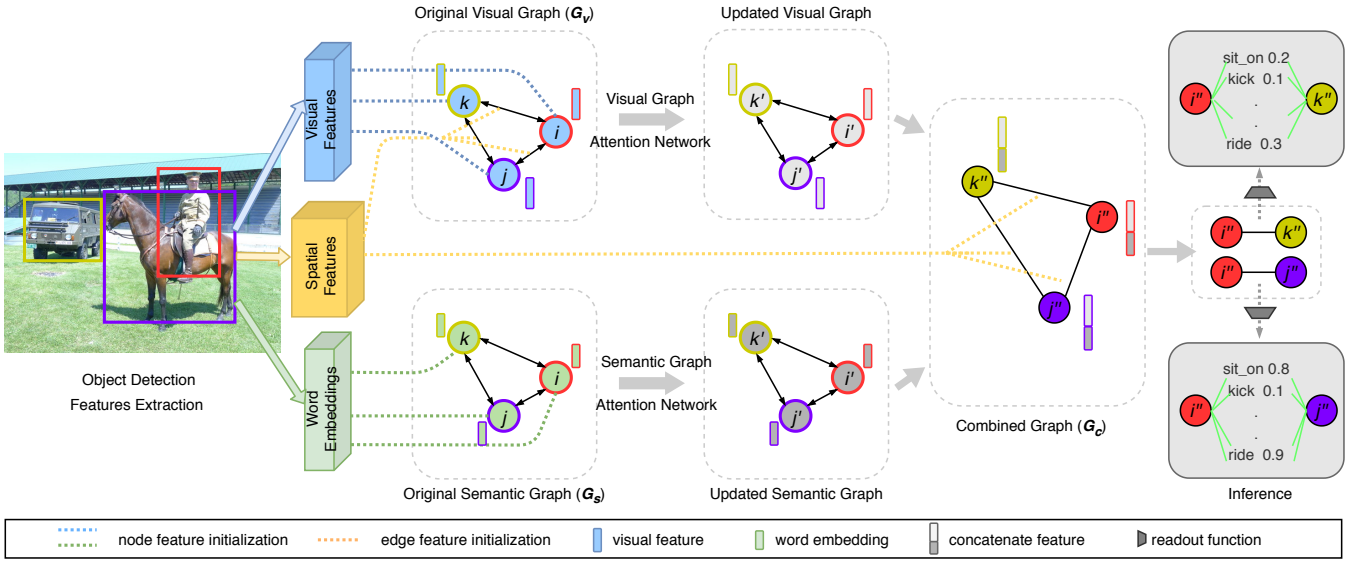

Fig. 2. Stream 1: The Visual-Semantic Graph Attention Network. After instance detection, a visual-spatial and a semantic graph are created. Node features are dynamically updated through attention network (Sec. III-C). We combine these updated graphs and then perform an edge readout step on box-pairs to infer all possible predicates between one subject and one object.

b) ***Instance Spatial Features****:* Instance spatial features such as bounding box locations and relative locations are informative about the relationship that proposals have with each other [20]–[23]. Consider the "ride" predicate, we can deduce that subject 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 $s = s_{rs} \bigcup s_{rp}$ can be grouped into relative scale features $s_{rs} = \left[\frac{x_i}{W}, \frac{y_i}{H}, \frac{x_j}{W}, \frac{y_j}{H}, \frac{A}{A^I}\right]$, and relative position features $s_{rp} = \left[(\frac{x_i - x_i^{'}}{x_j^{'} - x_i^{'}}), (\frac{y_i - y_i^{'}}{y_j^{'} - y_i^{'}}), \log(\frac{x_j - x_i}{x_j^{'} - x_i^{'}}), \log(\frac{y_j - y_i}{y_j^{'} - y_i^{'}}), \frac{x_c - x_c^{'}}{W}, \frac{y_c - y_c^{'}}{H}\right]$. 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 [24] as semantic features. We use the publicly available Word2vec vectors pre-trained on the Google News dataset (about 100 billion words) [25]. All existing object classes in the HICO-DET 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 $\tilde{h}_{v_i}$ 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*:

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

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

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

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 custom weighted sum and use the updated function $f_{update}(\cdot)$ to update each node's features:

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

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

Where $\oplus$ means element-wise summation operation. Note that $\mathbf{z}_{v_i}$ consists of the accumulated latent features of all the neighboring node connected to v_i (*i.e.* the subsidiary cues).

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 hidden 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 s described in Sec. III-B. We denote the combined node features as γ_i for node i, where $\gamma_i = [\tilde{\mathbf{h}}_{v_i}, \tilde{\mathbf{w}}_{v_i}]$.

E. Readout and Inference.

The last step is to infer the interaction label for a predicate as part of our original triplet <subject, predicate, object>. We box-pair all specific subject-object as illustrated in Fig. 2. With box-pairing, we finally construct the concatenated representation $\zeta_{ij} = [\gamma_i, s_{ij}, \gamma_j]$ 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: $\mathbf{S_R} = s_h * s_o * \mathbf{s}^a$.

Training. Note that HOI detection is a multi-label classification problem [9]. This network is jointly optimized end-to-end, with a multi-class cross-entropy loss that is minimized between action scores and the ground truth action label:

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

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

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

IV. EXPERIMENTS AND RESULTS

We now evaluate the performance of the first stream: VS-GATs on HICO-DET [7] and V-COCO [26] in Table I and Table II respectively. Ablation studies evaluate the impact of the proposed techniques (Table III). Several detection visualization results are shown in Fig. 3. Our supplementary material also shows the distribution of our model across objects for a given interaction in Fig. 1.More visualization results are available in our website.

A. Experimental Setup

Datasets. In this work, we use two common benchmarks: V-COCO [26] 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. We use the standard mean average precision (mAP) metric to evaluate the model's detection performance. mAP is calculated with recall and precision which is common used for the detection task. In this case, we consider a detected result with the form <subject, 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 [27]. For object detection we use Pytorch's re-implemented Faster R-CNN API [1]. Faster-RCNN use a RestNet-50-FPN backbone [28], [29] trained on the COCO dataset [30]. 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 hyperparameter except for the dropout rate (from 0.3 to 0.5) and the training epoch (from 250 to 600).

B. Results

Experiments show that VS-GAT sets the SOTA results in all three categories on HICO-DET. We achieve gains of

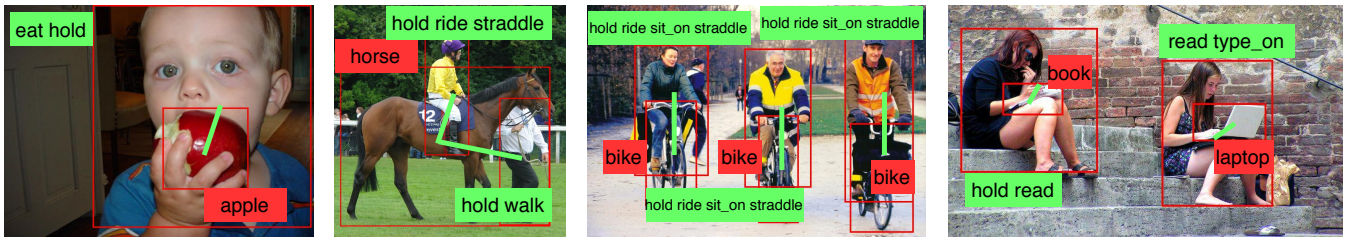

Fig. 3. HOI detection examples on the HICO-DET testing dataset. Subjects and objects are shown in red bounding boxes. Interaction classes are shown on the subject bounding box. Interactive objects are linked with the green line. We show all triplets whose inferred *action score* (9) is greater than 0.3.

TABLE I

MAP PERFORMANCE COMPARISON ON HICO-DET TEST SET.

Method	Full(600)↑	Rare(138)↑	Non-Rare(462)↑
iCAN [9]	14.84	10.45	16.15
Xu *et al.* [11]	14.70	13.26	15.13
Wang *et al.* [31]	16.24	11.16	17.75
Gupta *et al.* [13]	17.18	12.17	18.68
Li *et al.* [10]	17.22	13.51	18.32
PMFNet [12]	17.46	15.65	18.00
Peyre *et al.* [14]	19.40	14.60	20.90
VS-GAT	**20.38**	**16.54**	**21.52**

+0.98 (5.1%), +0.89 (5.7%), and +0.62 (2.96%). Our model also obtains comparable results of 50.6 mAP on V-COCO.

On HICO-DET, VS-GAT outperforms Peyre *et al.* [14], which exploits functional approximation through visual similarity enabling disambiguation between same-action but different-object scenarios. We also outperform works that leveraged human pose [10], [12], [13] which provides informative cues for better inference (we will study human pose subsequently). We note that Bansal *et al.* [32] achieved an mAP of 21.96 for the Full category. We do not include these results since Bansal *et al.* pre-trained their Faster-RCNN net directly on HICO-DET instead of COCO (as the rest of the works have done).This gives their system an unfair advantage compared to systems that train on the more general COCO dataset. Such training, helps their system refine instance proposals and reduce uninformative instances and noise. Recently, Hou [33] also verify that retrain the object detector on HICO-DET can possibly improve more than **4 mAP**. We chose not to re-train Faster-RCNN directly on HICO-DET for fair comparison.

TABLE II

MAP PERFORMANCE COMPARISON ON V-COCO TEST SET.

Method	AP_{role} (Scenario 1)
InteractNet [8]	40.0
GPNN [15]	44.0
iCAN [9]	45.3
Xu *et al.* [11]	45.9
Wang *et al.* [31]	47.3
Li *et al.* [10]	47.8
PMFNet [12]	52.0
VS-GAT	**50.6**

We also perform a qualitative analysis. As Fig. 3 shows, VS-GAT detect various kinds of HOIs such as: single person-single object, multi person-same object, and multi person-multi objects.

On V-COCO, VS-GAT outperforms most SOTA results. Wan *et al.* [12], report a 52.0 mAP. They develop a well-defined *Zoom-in Module* which utilizes *human pose* as well as *body part features* to extract detailed local appearance cues helping their model surpass [10] by a great margin on the small-scale dataset. However, [12] and [10] have a similar performance on the more diverse HICO-DET dataset. Without the pose estimator, [12] obtains 48.6 mAP, worse than our model. In our subsequent study, we will explore the impact of integrating pose-cues into our architecture.

C. Ablation Studies

In our ablation studies, we choose HICO-DET and just train the model on the *train_set* and directly test on the *test_set* without retraining on the *trainval_set*. We conduct the following six tests :

01 Visual Graph Only: G_V only. Here we remove the Semantic-GAT. We keep the Visual-GAT, attention, and inference the same. This study will show the importance of aggregating visual and spatial cues.

02 Semantic Graph Only: G_S only. Here we remove the Visual-GAT and keep the Semantic-GAT, attention, and inference the same. This study will show the importance of only working with Semantic cues.

03 Without Attention. Here, we use the averaging attention mechanism of Sec. III-C instead of the weighted sum mechanism. We still combine the graphs and infer similarly.

04 Without Spatial Features in G_C. Here, we remove spatial features from the combined graph edges G_C to study the role that spatial features play after the aggregation of

TABLE III

MAP PERFORMANCE FOR VARIOUS ABLATION STUDIES.

Method	Full↑	Rare↑	Non-Rare↑
VS-GAT	**19.66**	**15.79**	**20.81**
01 G_V only	18.81	13.96	20.26
02 G_S only	14.61	11.76	15.46
03 w/o attention	19.01	14.12	20.47
04 w/o spatial features in G_C	18.52	14.28	19.78
05 Message passing in G_C	19.23	14.31	20.70
06 Unified V-S graph	19.39	14.84	20.75

features across nodes.

05 Message Passing in G_C. Here, we leverage an additional graph attention network to process the combined graph (similar to our original visual-spatial graph). We examine if there would be a gain from an additional message passing on G_C with combined features from G_V and G_S.

06 Unified V-S Graph. Here, our models uses a single graph in which visual and semantic features are concatenated in the nodes from the start. Spatial features still instantiate edges. We examine if there is a gain from combining visual-semantic features from the start.

We now report on the ablation test results under the Full category. Study 01 yields an mAP of 18.81 which is a large portion of our mAP result. This suggests that the visuo-spatial features play a key role in inference. When only using the Semantic graph in 02, the effect is still considerable though less marked for this single contextual semantic cue. When combining these 3 contextual cues in a dual graph but not using the attention mechanism in test 03, yields a gain bringing the mAP to 19.01. This suggests that edge relations with multi-contextual cues are helpful even without attention. In test 04, using attention but removing spatial features at the end hurts. By inserting spatial features in the combined graph we are effectively using a skip connection step in neural networks which has been shown to help classification. In test 05, additional attention in the combined graph confers slight benefits. Rather, it is the attention mechanisms for the independent visual-spatial and semantic features are more informative. In test 06, a combined V-S graph is not as effective as separating cues early on. This suggests that different contextual cues need to reach a proper level of abstraction in the latent space before being fused together.

V. STREAM 2: POSE-BASED MODULAR NETWORK

In our second study, we subsequently studied the effects of fine-grained human poses for HOI via relative and absolute human pose features (Fig. 4). It is clear that a human's posture information plays a key role in disambiguating interactions (particularly in crowded scenes). Relative spatial pose features are those between each human joint and the target object bounding box center and provide informative cues for inference. E.g. as in Fig. 4, the entirety of the human joints above the skateboard centroid strongly indicate a *ride* interaction. Absolute human pose features consist of the human joint coordinates normalized with respect to the center of the human bounding box and are indicative of posture. Such features help to distinguish between say a human and person standing or sitting on top of an object.

In this subsequent study, we integrate our pose features within a second stream that runs in parallel to VS-GAT. Different from previous works [10], [12], [13] which do not consider the relative and absolute pose features separately, this stream is composed of two branches that process the relative spatial pose features of each joint independently via two hidden layers and the absolute pose features of each joint via graph convolutions. The branches features are fused and their output summed with the VS-GAT stream before performing inference. We refer to the second network stream as the pose-based modular network (PMN) shown in Fig. 4.

VI. METHODOLOGY

In this stream, we use an off-the-shelf pose detector [6] extracts human pose keypoints from which relative and absolute pose features are constructed and fed to the network. PMN's score factors (are defined as with VS-GAT) p_2 are generated and summed with VS-GAT's before fed into a sigmoid function to predict the score for each action/predicate.

A. Pose Features

1) Relative Spatial Pose Features: In previous works [7], [9], [10], [12], spatial relationships have been encoded in two particular ways. Once is the instance spatial relationships we used in VS-GAT, others adopt interactive representations that extract relative positions in the instance bounding boxes [7]. In our work, we extract more nuanced spatial cues from the human pose as illustrated in the skateboard images of Fig. 4. Our relative spatial pose features consist of the coordinate offset between each person's keypoints and the center of (the candidate) object bounding box. We define the *i*th human joint (keypoint) coordinates as (x_i, y_i) and the relative spatial features as: $f^i_{rp} : (x^{'}_i, y^{'}_i) = (\frac{x_i - x^o_c}{W}, \frac{y_i - y^o_c}{H})$. Here, (x^o_c, y^o_c) are the object's bounding box center and (W, H) are the image size . We denote the final 17×2 relative spatial pose features as $f_{rp} \in \mathcal{R}^{17\times 2}$.

2) Absolute Pose Features: We construct absolute keypoint pose features by normalizing with the center of the human bounding box as done in [13]: $f^i_{ap} : (x^{''}_i, y^{''}_i) = (\frac{x_i}{x^h_c}, \frac{y_i}{y^h_c})$, where (x^h_c, y^h_c) denotes the center of the human bounding box. We denote the final 17×2 dimensional absolute pose features of all keypoints as $f_{ap} \in \mathcal{R}^{17\times 2}$.

B. Pose-based Modular Network

An overview of the PMN stream is shown in Fig. 4 and described below. An equivalent description with mathematical definitions can be found in our supplementary material. The stream's two branches project the relative and absolute pose features to higher dimensional features respectively before getting concatenated, flattened and classified.

The first branch encodes relative spatial pose features via two fully-connected layers. For the second branch, inspired from [34], [35], we use a fully-connected layer followed by a GCN layer to process the absolute pose features. As with stream 1, a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is used in PMN but in this case to describe the human pose. Additionally, an adjacency matrix $A \in \mathcal{R}^{V\times V}$ indicates joint connections whilst a degree matrix $D_{ii} = \sum_j A_{ij}$ indicates the number of local joint connections to a given joint.

Once the relative and absolute pose features are fed through the network streams, the processed features h_1 and h_2 are concatenated and fed into a fully-connected layer. We then reshape the foregoing features: $h \in \mathcal{R}^{N\times 17\times 64} \rightarrow h^{'} \in \mathcal{R}^{N\times 1088}$, where N are the box-pairs per mini-batch.

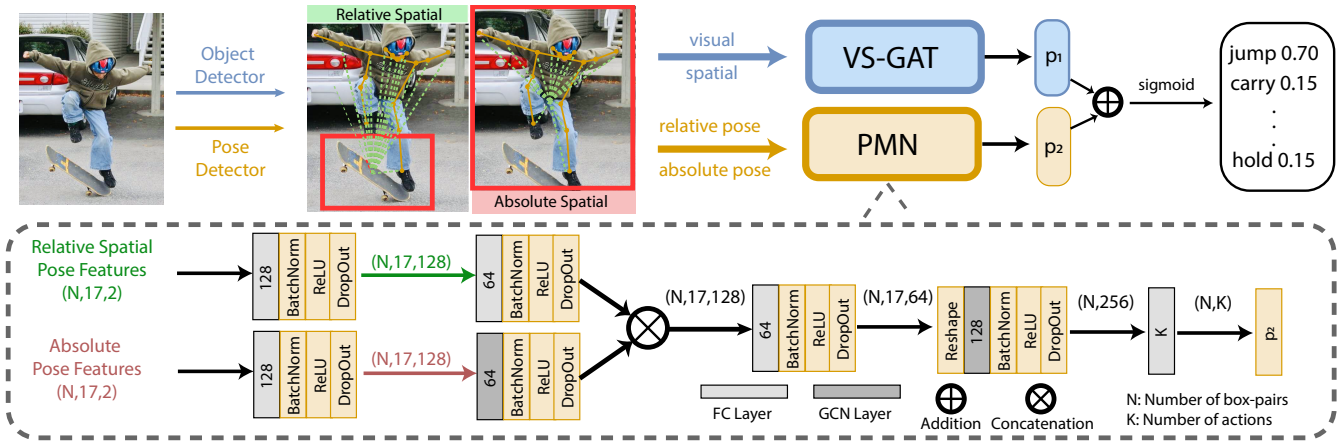

Fig. 4. PMN runs parallel to VS-GAT. In this dual branch, relative spatial pose features are processed by 2 hidden layers & absolute pose features by a GCN. Later layers are fused and the PMN output is summed with VS-GAT before applying a sigmoid.

Next, we use two fully-connected layers to get the action score factor. After obtaining the action score factors from VS-GAT and PMN, we sum p_1 and p_2 and then apply a binary sigmoid classifier to the box pairs as in Sec. III-E.

Training of the PMN is done is jointly trained end-to-end in a supervised manner by minimizing the multi-label binary cross-entropy loss $BCE(\cdot)$ between the inferred action score s and the ground truth action label y^{label} for each action category. Note, that at this stage training we fix the VS-GAT stream and just train the PMN[2].

VII. Experiments And Results

The same datasets, evaluation metrics, and network implementation are used as with VS-GAT. In this section, we report our best quantitative and qualitative results for our full VSP-GMN along with additional ablation studies.

1) Quantitative Results and Comparisons: Table IV lists both the HICO-Det and V-COCO results under the same setups of Tables I & II. Results show we surpasses all SOTA metrics on HICO-DET improves the result from VS-GATs by 0.98 mAP (∼4.6%), 1.57 mAP (∼9.8%), 0.75 mAP (∼3.5%) for the Full, Rare and Non-Rare categories respectively. As compared to other works, we achieve important gains of ∼**9.33%** (1.81 mAP), ∼**12.50%** (1.95 mAP), and ∼**9.49%** (1.39 mAP) respectively for an average gain of ∼9.49%.

On V-COCO, the **51.8 mAP** improves VS-GATs by **2 mAP** (∼4.0%) and yields competitive result, surpassing [10] which also leverages human pose. Note that PMFNet [12] considers not only human pose but also human body part features and outperforms previous works by a considerable margin. Our framework has comparable performance without the use of complicated human body part features.

2) Qualitative Results.: The visualization results of Fig. 5 compare VSP-GMN with VS-GATs along on the V-COCO *test* set. For instance, VS-GATs alone may output false positives when multiple persons and objects are close to each other. In the first image, the 2nd, 4th, and 6th person from left-to-right perform the action of "sking" on their *neighbors*' skis. Not so with VSP-GMNs.

[2]Future work will conduct training of both streams in a joint manner. More details can be found in our website.

TABLE IV
MAP FOR VSP-GMN ON HICO-DET AND V-COCO *test* SETS.

HICO-DET	Full(600)↑	Rare(138)↑	Non-Rare(462)↑
VS-GAT + PMN	**21.21**	**17.60**	**22.29**
V-COCO		AP_{role} (Sce. 1)	
VS-GAT + PMN		**51.8**	

Fig. 5. VS-GAT attributes actions to neighbor's objects while VSP-GMN does not. More visualization results can be found in our website.

A. Ablation Studies

These studies use the same setup as in VS-GATs. Two tests are conducted.

*a) **PMN vs. NFPN.**:* In [13], Gupta *et al.* design their fine-grained layout factor network as three MLP layers to encode the pose features. Here, we also construct a No-Frills Pose Network (NFPN) implemented by a 3-layer MLP with (128,128,117) neurons respectively. The first two layers

use batch normalization, ReLU activation, and dropout. We flatten and concatenate our relative spatial and absolute pose features as the 68 (= 17 x 2 + 17 x 2) dimensional input features. As shown in the table below, NFPN improves VS-GAT but the PMN stream performs better.

Method	Full↑	Rare↑	Non-Rare↑
VS-GAT + NFPN	20.88	17.12	22.01
VS-GAT + PMN	**21.12**	**17.59**	**22.18**

b) ***Relative spatial pose features v.s. Absolute pose features.****:* In this study we quantify the contributions of both relative and absolute pose features in our work. The table below shows the results and indicates that relative features provide greater gains than absolute ones; nonetheless both together provide the greatest gain.

Relative	Absolute	Full↑	Rare↑	Non-Rare↑
–	–	20.38	16.54	21.52
–	√	20.55	16.65	21.71
√	–	20.94	16.91	22.15
√	√	**21.12**	**17.59**	**22.18**

VIII. CONCLUSION

Visual-Semantic-Pose Graph Mixture Networks learned to exploit primary and subsidiary relations in visuo-spatial-semantic space through its VS-GAT network stream. It also learned to leverage absolute and relative posture information through its PMN stream. The dual-branching design of the networks supported better fusing of feature for better inference. Our network processed a wide range of contextual cues resulting that outperformed the state-of-the-art, particularly in datasets with fewer training classes that are harder to predict. This work's key insights assess how best to fuse contextual information via graph structure network to enhance HOI detection.

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.

[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.

[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.

[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.

[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.

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

[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.

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

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

[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.

[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.

[12] 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.

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

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

[15] 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.

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

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

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

[27] 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.

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

[29] 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.

[30] 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.

[31] T. Wang, R. Anwer, K. Muhammad, M. Haris, F. S. Khan, Y. Pang, L. Shao, and J. Laaksonen, "Deep contextual attention for human-object interaction detection," *In ICCV*, 2019.

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

[33] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," *In ECCV*, 2020.

[34] J. Liu, Z. Liang, Y. Li, Y. Guan, and J. Rojas, "A Graph Attention Spatio-temporal Convolutional Networks for 3D Human Pose Estimation in Video," mar 2020. [Online]. Available: http://arxiv.org/abs/2003.14179

[35] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," *In CVPR*, 2019.