核心观点：人的pose自身以及人的关节点与目标物品之间相对位置是重要的 che

Pose-based Modular Network for Human-Object Interaction Detection

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

[1] Guangdong University of Technology
[2] Chinese University of Hong Kong

Abstract

Human-object interaction(HOI) detection is a critical tasks in scene understanding. The goal is to infer the triplet <subject, predicate, object> in a scene. In this work we note that the human pose itself as well as the relative spatial information of the human pose with respect to the target object can provide informative cues for HOI detection. We contribute a Pose-based Modular Network (PMN) which explores the absolute pose features and relative spatial pose features to improve HOI detection and is fully compatible with existing networks. Our module consists of a branch that first processes the relative spatial pose features of each joint independently. Another branch updates the absolute pose features via fully-connected graph structures. The processed pose features are then fed into an action classifier. To evaluate our proposed method, we combine the module with the state-of-the-art model named VS-GATs and obtain significant improvement on two public benchmarks: V-COCO and HICO-DET, which shows its efficacy and flexibility. Code is available at https://github.com/birlrobotics/PMN.

1. Introduction

Recently, great progress has been made in computer vision, including object detection [3, 6, 21, 26], human pose estimation [2, 20, 23, 35], action recognition [15, 34] and scene segmentation [11]. However, to better understand the visual world, a robot should not only detect the individual instances in a scene but also further comprehend how a person interact with the world. A subclass of that interaction is with objects. As such, human-object interaction (HOI) detection has recently attracted increasing attention in the field of computer vision.

Human-object interaction detection infers the triplet *<subject, predicate, object>* in a scene. For example, in Fig. 1, we first detect the *human* and object (*skateboard*) instances. We finally infer the interaction *ride* between them, yielding the triplet *<human, ride, skateboard>*. Note that some images may contain multiple humans simultaneously interacting with various objects. One person may also have different interactions with a single object. For instance, Fig. 1 contains the set of ground-truth triplets: *<human, ride, skateboard>*, *<human, jump, skateboard>* and *<human, stand_on, skateboard>*.

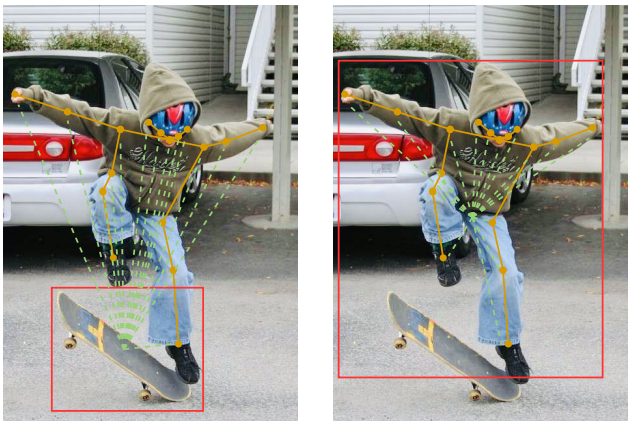

Figure 1. **Two constructed pose features we use in our method**. The *relative spatial pose features* (left) are the offset between each joint of human pose and the target object, which provides more detailed spatial information. The *absolute pose features* (right) are the normalized keypoint features with respect to the human bounding box, which offers the pose intrinsic properties cues to the model.

Recently, researchers have proposed a variety of networks for HOI detection [1, 5, 7, 9, 16, 17]. The first works were multi-stream neural networks that leveraged visual and spatial cues for HOI detection [1,5,7]. Others have considered human pose or human part features and have outperformed previous works by a great margin showing that HOI detection system benefit from relevant context [9, 16, 30]. More recently, Liang *et al.* [17] propose a dual-graph attention network which enables the model to leverage the

rich information by integrating and broadcasting information through the graph structure. However, they don't consider the useful human pose cues.

In this paper, we study fine-grained human poses via relative and absolute pose features (Fig. 1) to aid HOI detection. The models receives detailed spatial information in the form of relative spatial pose features between each human's keypoint coordinates (*i.e*. the joint) and the center of the target object bounding box. Moreover, the human pose intrinsic properties can also provide useful cues. For example, <*human, eat, apple*> and <*human, drink_with, bottle*> may have the similar posture. So we also use the absolute pose features, which consist of the keypoint coordinates normalized to the center of the human bounding box.

Furthermore, we propose a Pose-based Modular Network (PMN) which explores the constructed pose features (Fig. 1) and is fully compatible with existing networks for HOI detection. The module consists of one branch that processes the relative spatial pose features of each joint independently and another branch which uses graph convolutions to update the absolute pose features of each joint. We then fuse the processed features followed by an action classifier as depicted in Fig. 2.

We evaluate our proposed module on two public benchmarks V-COCO [8] and HICO-DET [1]. Our method consistently improves the state-of-the-art method [17]. On V-COCO, our method improves SOTA by **2 mAP** (∼4.0%). On the more challenging HICO-DET, our method improves SOTA 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. The addition of human pose cues to visual, spatial, and semantic cues; whilst being attended with attention, aided to further reduce false positives in the crowded scenes in general (Fig. 3). The improvements indicate our method is efficient and flexible.

2. Related work

Object Detection and Pose Estimation. In scene understanding, object detection [3, 6, 21, 26] identifies, localizes, and classifies object instances in a scene. Pose estimation [2, 20, 23] computes 2D or 3D coordinates of human skeleton keypoints (body joints like shoulders, eyes, and knees often 17 in total). For HOI detection, researchers have used off-the-shelf object detector to localize people and objects. They have also adopted pose estimators to obtain fine-grained human poses. Then, instances and human pose features are leveraged by neural architectures for HOI inference.

Graph Neural Network. Graph neural networks (GNN) [10, 14, 22, 29, 31, 33] have recently grained increasing attention. Kipf *et al.* [14] proposed a variant of graph convolutional networks (GCN) by introducing a first-order approximation to spectral graph convolutions. Velivckovic *et al.* [29] introduced graph attention networks (GATs) which leveraged masked self-attentional layers to enable its nodes to attend their neighborhood features with varying dynamic weights. Lately, GNNs have been successfully applied to pose estimation [20, 35]. Inspired by [20, 35], we also use GCNs to encode absolute pose features.

Human-Object Interaction Detection. Improving HOI detection requires the model to better leverage contextual information in complex scenes. Chao *et al.* [1] contributed the HICO-DET dataset and proposed a novel DNN input named Interaction Pattern to represent the spatial relations. Gkioxari *et al.* [7] designed InteractNet that predicts a density over target object locations based on the appearance of a detected person. Gao *et al.* [5] extended the methods in [1, 7] by introducing an instance-centric attention module to dynamically highlight the region of interest in an image. Different from these multi-stream neural networks, Qi *et al.* [25] introduced the Graph Parsing Neural Network (GPNN) which iteratively updates features over a graph structure. Recently, Xu *et al.* [32] considered the intrinsic semantic regularities across the scene to facilitate HOI detection. Liang *et al.* [17] contributed a dual-graph attention network which takes visual, spatial and semantic cues to learn rich relations across scene instances over the novel graph network. Li *et al.* [16] and Wan *et al.* [30] further combine the fine-grained human pose and interaction pattern [1] as the spatial configuration map followed by a MLP and later concatenate all the processed features from multiple branches. However, this design can not be fully transferred to existing networks. Lately, Gupta *et al.* [9] designed a factored model which considered human appearance features, boxed-pair configurations, and fine-grained human poses as isolated factors for HOI detection. However, in their fine-grained layout factor network, they simply flatten and concatenate the pose information (including relative and absolute pose features) and feed them into an MLP. In our method, we first process the relative and absolute pose features separately with different networks and later fuse and flatten them before feeding into the classifier (Fig. 2). Our proposed pose-based module is fully compatible with existing networks and yields a significant gain in performance.

3. Method

In this section, we start with an overview of the entire system (Sec. 3.1) followed by an introduction to the various pose features considered in our work (Sec. 3.2). Then, we outline our pose-based modular neural network structure (Sec. 3.3). Finally, we describe inference and learning for our model (Sec. 3.4).

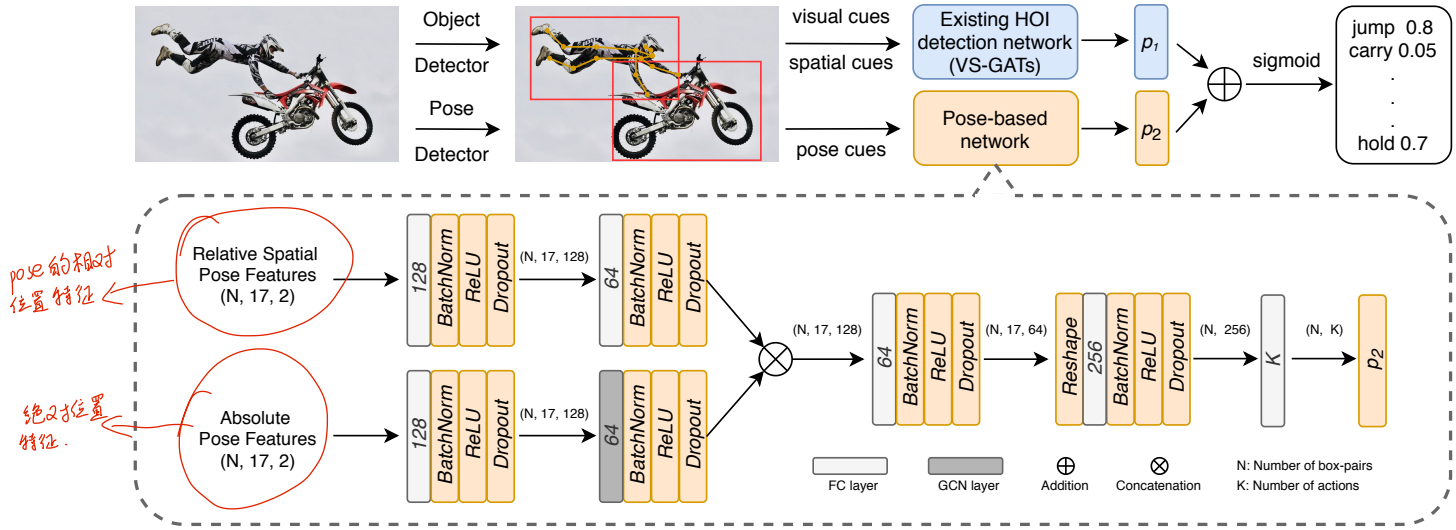

Figure 2. **Framework Overview**. Our system consists of two streams: a) an existing HOI detection network for inference based on supplied cues *(e.g.* visual, spatial, semantic.); b) our proposed pose-based modular network that extends the top branch for better contextualization with absolute and relative pose cues. The bottom half of the diagram depicts the pose-based network design in detail.

3.1. Overview

As illustrated in Fig. 2, our system consists of two branches. Given an input image: (i) an off-the-shelf object detector [26] extracts instance bounding boxes along with their embedded features and (ii) an off-the-shelf pose detector [11] extracts the human pose keypoints. Suitable features are constructed and fed both into the existing visual-semantic graph attention HOI detection network and the proposed pose-based network. Score factors (the output of last layer of each stream) p_1 and p_2 are generated, summed and fed into a sigmoid function to predict the score for each action/predicate.

Specifically, for each human-object pair, we denote s_h and s_o as the confidence scores of the detected human and object instances respectively. We denote s^a as the score of action $a \in \{1, ..., K\}$, where K is the total number of possible actions. The final score of the HOI triplet <*subject, predicate, object*> is the product of the scores:

$$S = s_h * s_o * s^a. \tag{1}$$

We choose VS-GATs [17] as the existing HOI detection network in our framework as it has shown that by capturing visual-spatial and semantic cues via independent attention mechanisms that are later combined, the network is able better disambiguate hard detection cases and beating the state-of-the-art on the challenging HICO-DET dataset and ranking second place on the small-scale V-COCO dataset. Additionally, their code and preprocessed features are publicly available which enables testing and benchmarking for the community.

3.2. Pose Features

3.2.1 Relative Spatial Pose Features

Spatial features are able to provide informative cues to infer the predicate. For example, the *human* box above the *skateboard* box strongly indicates the *ride* interaction. In past works, there have been two main approaches to encode spatial relationship between person and object [1, 5, 16, 30]. Works like those of Chao *et al.* [1], adopt an interaction representation that extract the relative position of instance bounding boxes. Pixels within the human and object bounding boxes take a value of 1 and 0 elsewhere. A DNN can use this representation to learn 2D filters of human-object spatial configurations. Works like those of Liang *et al.* and Gupta *et al.* [9, 17], extract relative scale features and relative position features based on the coordinates of the instance bounding boxes. As for our work, we extract more nuanced spatial cues from the human pose as illustrated in the left image of Fig. 1.

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 employ He *et al.* pose detector [12] to estimate 17 keypoints for each person in the 2D image (in COCO [19] format). We define the ith human keypoint coordinates as (x_i, y_i) and the relative spatial features f^i_{rp} as:

$$f^i_{rp} : (x^{'}_i, y^{'}_i) = (\frac{x_i - x^o_c}{W}, \frac{y_i - y^o_c}{H}). \tag{2}$$

where (x^o_c, y^o_c) is the center of object bounding box, and (W, H) is the size of image. We denote the final 17×2 relative spatial pose features as $f_{rp} \in \mathcal{R}^{17\times 2}$.

3.2.2 Absolute Pose Features

Generally, a person will have different postures when performing different actions. For instance, the human pose when sitting <human, sit_on, chair> or when standing <human, stand_on, chair> are very different. Other times, similar postures may occur when a person acts with different objects (*e.g.* riding a horse or a bicycle). These intuitions indicate that a human's pose intrinsic properties are also useful for HOI detection.

Similar to [9], we construct absolute keypoint pose features f_{ap} by normalizing with the center of the human bounding box:

$$f_{ap}^{i} : (x_i^{''}, y_i^{''}) = (\frac{x_i}{x_c^h}, \frac{y_i}{y_c^h}). \quad (3)$$

where (x_c^h, y_c^h) denotes the center of the human bounding box. We denote the final 17×2 dimensional absolute pose features of all keypoints as $f_{ap} \in \mathcal{R}^{17\times 2}$.

3.3. Pose-based Modular Network

An overview of our pose-based module is shown in Fig. 2. The module's two streams, project the relative and absolute pose features to higher dimensional features respectively. Then we concatenate and flatten the features before classifying them.

The first stream encodes relative spatial pose features via two fully-connected layers with batch normalization, ReLU activations, and dropout. Eqtn. 4 defines the operation:

$$h_1 = \text{ReLU}(\text{ReLU}(f_{sp}\, W_0)\, W_1). \quad (4)$$

where $W_0 \in \mathcal{R}^{2\times 128}$ and $W_1 \in \mathcal{R}^{128\times 64}$ are the learnable weight matrices.

Inspired by [20, 35], we adopt a GCN [14] layer to process the absolute pose features. We define the human pose as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of V nodes and $\mathcal{E}$ is a set of E edges. $A \in \mathcal{R}^{V\times V}$ is the adjacent matrix that indicates the connection between joints and $D_{ii} = \sum_j A_{ij}$ is a degree matrix. In the second stream, we use a fully-connected layer followed by a GCN layer to process the absolute pose features as indicated in Eqtn. 5:

$$h_2 = \text{ReLU}(\hat{A}\ \text{ReLU}(f_{ap}\, W_2)\, W_3). \quad (5)$$

where $\hat{A} = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is the normalized adjacent matrix, with $\tilde{A} = A + I_N$ (I_N is the identity matrix) and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ [14]. Also, $W_2 \in \mathcal{R}^{2\times 128}$ and $W_3 \in \mathcal{R}^{128\times 64}$ denote the trainable weight matrices.

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:

$$h = \text{ReLU}((h_1 \otimes h_2)\, W_4). \quad (6)$$

where $W_4 \in \mathcal{R}^{128\times 64}$.

Later, we reshape the foregoing features: $h \in \mathcal{R}^{N\times 17\times 64} \rightarrow h^{'} \in \mathcal{R}^{N\times 1088}$, where N means the number of all box-pairs in each mini-batch. Then we adopt two fully-connected layers to get the final output:

$$p_2 = (\text{ReLU}(h^{'}\ W_5))\ W_6. \quad (7)$$

where $W_5 \in \mathcal{R}^{1088\times 256}$ and $W_6 \in \mathcal{R}^{256\times K}$. K denotes the total number of possible actions.

3.4. Inference and Learning.

After passing through the whole framework, we can get the action score factors (the output of last layer) p_1 and p_2 from the existing HOI detection network branch and our pose-based modular network branch, respectively. In the last inference phase, we directly sum p_1 and p_2 up, which make our module fully compatible with existing networks.

As mentioned in Sec. 1, HOI detection is a multi-label classification problem, where more than one action/predicate might be assigned to a $< subject, object >$ box-pair. Therefore, We apply a binary sigmoid classifier for each action category:

$$S^a = sigmoid(p_1 \oplus p_2). \quad (8)$$

where $S^a \in \mathcal{R}^{N\times K}$ contains the inferred score for each action categoy for each $< subject, object >$ box-pair.

Our framework is jointly trained end-to-end in a supervised manner by minimizing the multi-label binary cross-entropy loss $BCE(\cdot)$ between inferred the action score s and the ground truth action label y^{label} for each action category:

$$\mathcal{L} = \frac{1}{N \times K} \sum_{i=1}^{N} \sum_{j=1}^{K} BCE(s_{ij}, y_{ij}^{label}) \quad (9)$$

See Sec. 4.1 for more details on training procedures.

4. Experiments And Results

In this section, we first describe the experimental datasets and evaluation metrics, followed by more implementation details of our framework (Sec. 4.1). Then, we report the quantitative results (Sec. 4.2.1) compared with the state-of-the-art methods as well as qualitative detection visualization results (Sec. 4.2.2). Finally, we introduce ablation experiments (Sec. 4.3) which validate each component of the proposed module.

4.1. Experimental Setup

Datasets. We adopt two common benchmarks: V-COCO [8] and HICO-DET [1] to evaluate our framework. **V-COCO** is a subset of the MS-COCO [19] dataset with appended HOI annotations. It contains 10,346 images, where

Method	Object Detector	Full(600)↑	Rare(138)↑	Non-Rare(462)↑
Shen *et al.* [27]	Faster R-CNN with VGG19 [28]	6.46	4.24	7.12
HO-RCNN [1]	Fast R-CNN [6]	7.81	5.37	8.54
InteractNet [7]	Faster R-CNN with ResNet-50-FPN	9.94	7.16	10.77
GPNN [25]	Deformable ConvNets [4]	13.11	9.34	14.23
iCAN [5]	Faster R-CNN with ResNet-50-FPN	14.84	10.45	16.15
Xu *et al.* [32]	Faster R-CNN with ResNet-50-FPN	14.70	13.26	15.13
Gupta *et al.* [9]	Faster R-CNN with ResNet-152	17.18	12.17	18.68
$RP_{T2}C_D$ [16]	Faster R-CNN with ResNet-50-FPN	17.22	13.51	18.32
PMFNet [30]	Faster R-CNN with ResNet-50-FPN	17.46	15.65	18.00
Peyre *et al.* [24]	Faster R-CNN with ResNet-50-FPN	19.40	14.60	20.90
VS-GATs [17]	Faster R-CNN with ResNet-50-FPN	20.27	16.03	21.54
VS-GATs + PMN	Faster R-CNN with ResNet-50-FPN	**21.21**	**17.60**	**22.29**

Table 1. mAP performance comparison with SOTA on the HICO-DET *test* set.

2,533 form the *train* set, 2,867 form the *val* set, and 4,946 form the *test* set. It contains 16,199 human instances and 29 action annotation categories (five of them have no object interactions (*e.g. smiling*) which we do not consider for HOI detection). **HICO-DET** is a large-scale dataset which consists of 47,776 images in total (38,118 for training and 9658 for testing). It contains 150K annotated human-object pair instances and 600 HOI categories over 80 object categories (same as [19]) and 117 action categories. The 600 HOI categories are divided into: (i) Full: all 600 categories; (ii) Rare: 138 HOI categories with less than 10 training samples, and (iii) Non-Rare: 462 HOI categories with more than 10 training samples.

Evaluation Metrics. We adopt the mean average precision (mAP) to measure the detection performance. We consider a detected triplet as true positive when the predicted predicate is correct and both the detected human and object bounding boxes have the intersection-of-union (IoU) ≥ 0.5 with respect to the ground truth.

Implementation Details. We employ Faster R-CNN [26] with a RestNet-50-FPN backbone [13, 18] as the object detector. Mask R-CNN [11] serves as the human pose estimator pre-trained on COCO [19] [1]. As mentioned in Sec. 3.1, we choose VS-GATs [17] as the existing HOI detection network in our framework (Fig. 2). The architecture of our pose-based modular neural network is illustrated in Fig. 2. Note that the object detector, pose estimator and VS-GATs are frozen when training. That's to say, we just train the pose-based module.

We follow the same training scheme from previous works: select the hyperparameters on the *val* set and then retain the model on the *trainval* set (*train* set +*val* set) [2]. Following [17], we set the detection confidence threshold to 0.8 for humans and 0.3 for objects. When training, we use a batch size of 32 and dropout ratio of 0.2. We adopt an Adam optimizer with an initial learning rate of 3e-5. For V-COCO, we reduce the learning rate to 3e-6 at epoch 400 and stop training at epoch 600. For HICO-DET, we reduce the learning rate to 3e-6 at epoch 150 and stop training at epoch 200. We conduct our experiments on a single Quadro P3200 GPU.

Method	AP_{role} (Sce. 1)
Gupta *et al.* [8]	31.8
InteractNet [7]	40.0
GPNN [25]	44.0
iCAN [5]	45.3
Xu *et al.* [32]	45.9
Li *et al.* (RP_DC_D) [16]	47.8
PMFNet [30]	52.0
VS-GATs [17]	49.8
VS-GATs + PMN	**51.8**

Table 2. mAP performance comparison with SOTA on the V-COCO *test* set.

4.2. Results

4.2.1 Quantitative Results and Comparisons.

Our experiment results (Table 1 and Table 2) demonstrate that the proposed Pose-based Modular Network (PMN) combined with VS-GATs [17] beats all SOTA metrics on HICO-DET and achieves comparable result on V-COCO; thus showing the significance of pose cues showing its efficacy and flexibility.

[1]For the object detector and the pose estimator, we directly use Pytorch's re-implemented API https://pytorch.org/docs/stable/torchvision/models.html.

[2]We regard the original training set in HICO-DET as the *trainval* set and follow [17] to split it into the *train* set and the *val* set.

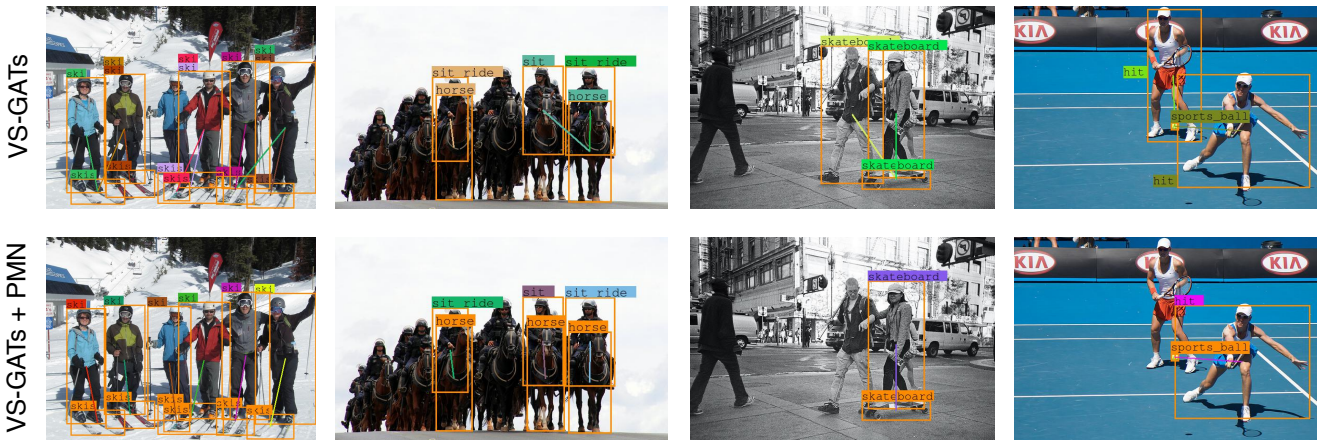

Figure 3. HOI detection results compared with VS-GATs on V-COCO *test* set. The first row is the detection results of original VS-GATs. The second row is the detection results of our framework (VS-GATs + PMN). Subjects and objects are shown in orange bounding boxes.The interaction classes are shown on the subject bounding box and the interactive objects are linked with the line in the same color. We show all triplets whose inferred *action score* exceeds 0.5.

On V-COCO, we achieve an **51.8 mAP**. Our method improves VS-GATs by **2 mAP** (∼4.0%) and also further surpasses most of SOTAs including [16] which also leverage human pose in their network. Note that PMFNet [30] considers not only human pose but also human body part features, which make it outperform previous works by a considerable margin. However, our framework still have a comparable performance without the complicated human body part features. On HICO-DET, our method improves 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, which makes VS-GATs [17] further outperform existing methods [1, 5, 7, 9, 16, 24, 25, 30, 32].

4.2.2 Qualitative Results.

Fig. 3 shows some Visualization results compared with VS-GATs on V-COCO *test* set. We find that VS-GATs tend to output the false positive detection when multiply persons and objects are close to each other. For example, in the first image, VS-GATs infers the wrong detection that the 2th, 4th, 6th person (from left to right) also ski their neighbors' skis. However, with the proposed pose-based module which explores the detailed spatial cues and intrinsic properties based on human pose, our framework (VS-GATs + PMN) performs better in the crowded scenes as shown in the second row.

4.3. Ablation Studies

In this section, we perform several ablation studies on HICO-DET. To simplify the training steps, as in [17], we train the model on the *train* set without further retraining on the *trainval* set.

PMN vs. NFPN. In [9], Gupta *et al.* design their fine-grained layout factor network as a simple three layers MLP to encode the pose features. Similarly, we also construct a No-Frills Pose Network (NFPN) implemented by a 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. From Table 3, NFPN also improves VS-GATs but our PMN performs better.

Method	Full↑	Rare↑	Non-Rare↑
VS-GATs	20.27	16.03	21.54
VS-GATs + NFPN	20.88	17.12	22.01
VS-GATs + PMN	**21.12**	**17.59**	**22.18**

Table 3. *PMN vs. NFPN.* Ablation studies results on HICO-DET *test* set.

Relative spatial pose features *vs*. Absolute pose features. Table 4 validates the importance of the pose features in our method. Both set of features facilitate HOI detection and the relative spatial pose features played a more dominant role in this task.

Relative	Absolute	Full↑	Rare↑	Non-Rare↑
–	–	20.27	16.03	21.54
–	√	20.55	16.65	21.66
√	–	20.94	16.91	21.15
√	√	**21.12**	**17.59**	**22.18**

Table 4. *Relative vs. Absolute pose features.* Ablation studies results on HICO-DET *test* set.

5. Conclusion

In this paper, we propose a pose-based modular network which studies the relative spatial pose feature as well as the absolute pose features to improve HOI detection. The module is easy to combine with existing networks. The experiment results show that our method facilitates the HOI detection system to perform better in the crowded scenes and consistently improves the state-off-the-art method VSGATs on both V-COCO and HICO-DET benchmarks.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *In WACV*, pages 381–389, 2018. 1, 2, 3, 4, 5, 6

[2] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3D human pose from structure and motion. *In ECCV*, pages 668–683, 2018. 1, 2

[3] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. *In NIPS*, pages 379–387, 2016. 1, 2

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable Convolutional Networks. *In ICCV*, pages 764–773, 2017. 5

[5] Chen Gao, Yuliang Zou, and Jia Bin Huang. ICAN: Instance-centric attention network for human-object interaction detection. *In BMVC*, 2018. 1, 2, 3, 5, 6

[6] Ross Girshick. Fast r-cnn. *In ICCV*, pages 1440–1448, 2015. 1, 2, 5

[7] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and Recognizing Human-Object Interactions. *In CVPR*, pages 8359–8367, 2018. 1, 2, 5, 6

[8] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv: 1505.04474*, 2015. 2, 4, 5

[9] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. *In ICCV*, 2019. 1, 2, 3, 4, 5, 6

[10] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *In NIPS*, 2017. 2

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. *In ICCV*, pages 2961–2969, 2017. 1, 3, 5

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *In ICCV*, 2017. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *In CVPR*, pages 770–778, 2016. 5

[14] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *In ICLR*, 2016. 2, 4

[15] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-Structural Graph Convolutional Networks for Skeleton-based Action Recognition. *In CVPR*, 2019. 1

[16] Yong Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan Feng Wang, and Cewu Lu. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. *In CVPR*, 2019. 1, 2, 3, 5, 6

[17] Zhijun Liang, Junfa Liu, Yisheng Guan, and Juan Rojas. Visual-semantic graph attention networks for human-object interaction detection. *arXiv: 2001.02302*, 2020. 1, 2, 3, 5, 6

[18] Tsung Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. *In CVPR*, pages 936–944, 2017. 5

[19] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *In ECCV*, pages 740–755, 2014. 3, 4, 5

[20] Junfa Liu, Zhijun Liang, Yihui Li, Yisheng Guan, and Juan Rojas. A graph attention spatio-temporal convolutional networks for 3d human pose estimation in video. *arXiv:2003.14179*, 2020. 1, 2, 4

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. *In ECCV*, pages 21–37, 2016. 1, 2

[22] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. *In ICML*, pages 2014–2023, 2016. 2

[23] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. *In CVPR*, pages 7753–7762, 2019. 1, 2

[24] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. *In ICCV*, pages 1981–1990, 2019. 5, 6

[25] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song Chun Zhu. Learning human-object interactions by graph parsing neural networks. *In ECCV*, pages 407–423, 2018. 2, 5, 6

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *In NIPS*, pages 91–99, 2015. 1, 2, 3, 5

[27] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. *In WACV*, 2018-Janua:1568–1576, 2018. 5

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *In ICLR*, 2015. 5

[29] Petar Veličković, Arantxa Casanova, Pietro Liò, Guillem Cucurull, Adriana Romero, and Yoshua Bengio. Graph attention networks. *In ICLR*, 2018. 2

[30] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. *In ICCV*, pages 9469–9478, 2019. 1, 2, 3, 5, 6

[31] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A Comprehensive Survey

on Graph Neural Networks. *Arxiv: 1901.00596*, pages 1–22, 2019. 2

[32] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to Detect Human-Object Interactions with Knowledge. *In CVPR*, pages 2019–2028, 2019. 2, 5, 6

[33] Keyulu Xu, Stefanie Jegelka, Weihua Hu, and Jure Leskovec. How powerful are graph neural networks? *In ICLR*, 2019. 2

[34] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *In AAAI*, pages 7444–7452, 2018. 1

[35] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. *In CVPR*, 2019. 1, 2, 4