SKELETON-BASED INTERACTIVE GRAPH NETWORK FOR HUMAN OBJECT INTERACTION DETECTION

*Sipeng Zheng, Shizhe Chen, Qin Jin**

{zhengsipeng, cszhe1, qjin}@ruc.edu.cn

ABSTRACT

The human-object interaction detection (HOI) task aims to localize human and objects in an input image and predict their relationships, which is essential for understanding human behaviors in complex scenes. Due to the human-centric nature of the HOI task, it is beneficial to make use of human-related knowledge such as human skeletons to infer fine-grained human-object interactions. However, previous works simply embed skeletons via convolutional networks, which fail to capture structured connections in human skeletons and ignore the object influence. In this work, we propose a Skeleton-based Interactive Graph Network (SIGN) to capture fine-grained human-object interactions via encoding interactive graphs between keypoints in human skeletons and object from spatial and appearance aspects. Experimental results demonstrate the effectiveness of our SIGN model, which achieves significant improvement over baselines and outperforms other state-of-the-art methods on two benchmarks.

***Index Terms*—** human object interaction, graph convolutional network, graph attention mechanism

1. INTRODUCTION

Holistic image understanding is far beyond merely recognizing the global scene [1] or detecting individual objects [2]. The interaction between different objects plays an essential role in understanding semantic contents in the image. Among all types of interactions, human-object interaction has received particular attentions due to its human-centric nature related to our daily lives. Therefore, we focus on the human-object interaction (HOI) task [3] which aims to detect human and objects in an image and predict their relationships.

Since the subject in the HOI task is fixed as human, it can be beneficial to acquire human-related knowledge such as human poses for distinguishing fine-grained human-object interactions. For example, though the overall visual appearance and spatial locations between the human and bicycle are similar in Figure 1 (a) and (b), human poses in the two images contain clear difference and thus provide strong evidences to discriminate the "ride" and the "stand on" interactions. Most previous works [4, 5] simply embed human poses via convolutional networks, which can be insufficient to capture structured connections in fine-grained human skeleton and ignore the influence of object on the skeleton encoding. For instance, the awareness of the object "cup" can implicitly determine that more attention should be paid to hand and head areas than to less relevant body parts such as the feet. Moreover, the structured connections between different body parts can be very important to distinguish interactions with subtle differences. For example, how the hand, head and cup are connected is vital to discriminate "drink from a cup" (Figure 1 (c)) from "hold a cup" (Figure 1 (d)) as both interactions involve the same object and body parts.

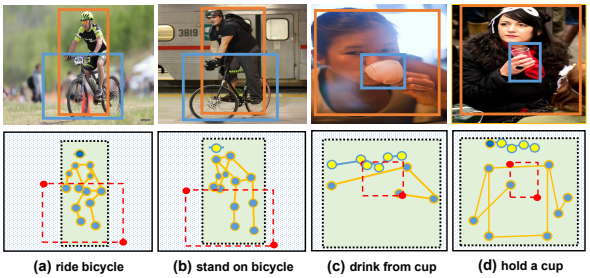

Fig. 1. Human skeletons provide fine-grained details to infer human-object interactions. The blue dots refer to the skeleton keypoints, and red dots refer to the object keypoints.

In this work, we propose a Skeleton-based Interactive Graph Network (SIGN) to employ human skeletons more effectively for HOI detection, which not only considers the influence of the interactive objects on human skeleton encoding, but also captures their structured connections with graphs. To be specific, we construct two types of graphs based on the human skeleton, namely spatial interactive graph (SIG) and appearance interactive graph (AIG). The SIG focuses on spatial aspect via learning spatial associations between the skeleton and the object keypoints, while the AIG aims to capture fine-grained visual appearance features from the interaction between the human body parts and the object. We utilize graph attention networks to encode SIG and AIG respectively and combine their graph embeddings as fine-grained skeleton representation to predict human-object interactions. Extensive experiments are carried out on two HOI benchmark

*Qin Jin is the corresponding author.

978-1-7281-1331-9/20/$31.00 2020 IEEE

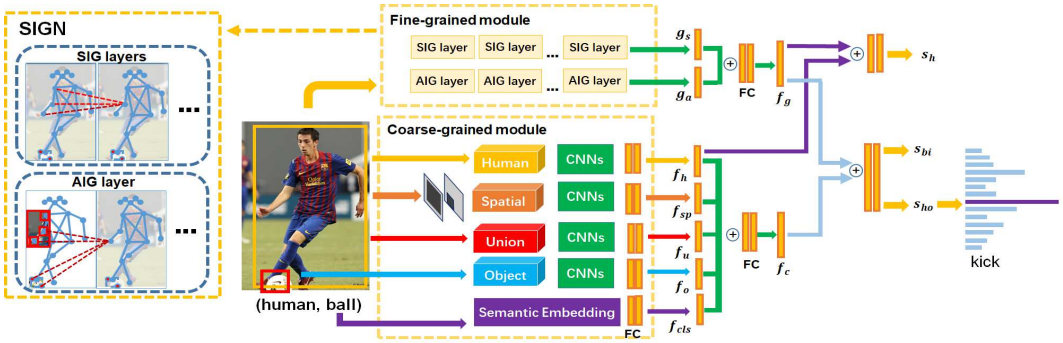

Fig. 2. The overview of our HOI prediction framework, consisting of a coarse-grained module and a fine-grained prediction module which is our proposed SIGN network. SIGN involves SIG and AIG layers for fine-grained association extraction.

datasets: HICO-DET and V-COCO. The proposed SIGN model achieves significant improvement over coarse-grained skeleton encoding baselines and outperforms other state-of-the-art approaches.

The main contributions of this work are three-fold: (1) We propose a Skeleton-based Interactive Graph Network (SIGN) to encode human skeletons associated with the object for fine-grained HOI detection. (2) We generate skeleton-based interactive graphs from spatial and appearance aspects, and employ graph attention networks to focus on salient skeleton keypoints and their association with the object. (3) The effectiveness of our SIGN model is demonstrated on two benchmark HOI datasets. The proposed model achieves the state-of-the-art performance.

2. RELATED WORK

Human-object interaction (HOI) [6, 7, 8] is a specific visual relation detection problem [9] focusing on human-centric pairs, which is crucial for understanding human behaviors under complex scenes. With human as the subject, it is more challenging to predict the interaction categories as there are more diverse interactions with subtle differences. In recent years, several large-scale datasets such as HICO-DET [3], V-COCO [10] and HCVRD [11] have been developed to benchmark this task. Many works receive significant improvement based on DNN architectures. Gao *et al.* [12] propose an instance-centric attention network which dynamically learns from salient regions relevant to the interaction. Shen *et al.* [13] explore the zero-shot learning method by disentangling the reasoning on verbs and objects. In [4], an interactiveness network is proposed to determine whether the human and object interact with each other.

Inspired by graph neural networks [14], some methods employ graph-based architectures to encode the human-object relationship. Among them, Qi *et al.* [15] incorporate structural knowledge and pass message by a graph parsing neural network (GPNN). Kato *et al.* [16] further design a novel graph-based model for compositional learning to explore the zero-shot problem. More recent studies start to focus on the human pose which provides fine-grained information between the human and object. Fang *et al.* [17] propose a pairwise body-part attention model to focus on crucial parts of interaction. Wan *et al.* [5] propose a zoom-in module to extract local skeleton features based on semantic attention. Generally, these graph-based models are constructed by instance-level nodes like human and objects, which ignores intra association among skeletons and their inter connection with the object. In this paper, we propose the skeleton-based interactive network to overcome these drawbacks.

3. OVERALL FRAMEWORK

As illustrated in Figure 2, our HOI framework consists of two stages: object detection and relation classification. In the object detection stage, we detect human and objects in an input image, and form human-object pairs as $\{b_h, b_o, p_h, p_o\}$, where b_h, b_o denote the detected human and object bounding boxes and p_h, p_o represent their detection confidences. Then in the relation classification stage, we predict interaction categories for each human-object pair via a coarse-grained prediction module and a fine-grained prediction module.

In the coarse-grained prediction module, we represent the human-object pair by global appearance, spatial features and semantic embeddings, same as in general visual relationship detection approaches [18, 19]. For appearance features, we extract region-level visual features f_h, f_o, f_u from the CNN block for human, object and their union bounding box respectively to obtain sufficient contextual information. For spatial features, we first generate spatial binary masks [18] for the human-object pair, which consists of a human channel and an object channel. The value of each channel is 1 inside the bounding box and 0 for anywhere else. Then the spatial binary masks are fed into a spatial CNN block to generate the spatial feature f_{sp}. For semantic embeddings, we extract semantic representations from BERT [20] and feed them into fully-connected layers to generate f_{cls}. Finally we concatenate above $f_h, f_o, f_u, f_{sp}, f_{cls}$ to obtain f_c for coarse-grained

prediction module.

The fine-grained prediction module aims to employ more human-related knowledge such as human skeletons to extract fine-grained representations for HOI prediction. We propose the SIGN model for this purpose, which will be described in Section 4. Two types of graphs are proposed to generate fine-grained spatial feature g_s and appearance feature g_a respectively. We combine g_s and g_a as the output feature f_g for fine-grained prediction module. Combining the outputs of coarse-grained and fine-grained modules together, the relation prediction score can be written as:

$$s_{ho} = \sigma(\mathrm{FC}(f_c \oplus f_g)) \tag{1}$$

where σ is the sigmoid function, $\oplus$ denotes vector concatenation and FC represents fully-connected layers. The final score for the HOI pair is determined by both object detection probabilities and relation prediction score as follows:

$$S(h, o, r) = p_h \cdot p_o \cdot s_{ho} \tag{2}$$

4. SKELETON-BASED INTERACTIVE GRAPH NETWORK

In this section, we introduce our Skeleton-based Interactive Graph Network (SIGN), which captures fine-grained correlations between human skeletons and object. As shown in Figure 3, it contains two types of graphs: a spatial interactive graph (SIG) and appearance interactive graph (AIG). Section 4.1 describes the SIG for learning fine-grained spatial relationship, followed by the description of AIG in Section 4.2 for learning fine-grained appearance skeleton features. Finally, Section 4.3 presents training objectives for our model.

4.1. Spatial Interactive Graph

Previous spatial binary masks [18] only represent coarse-grained spatial relationship between human and object bounding boxes, which may be ambiguous for fine-grained interaction prediction as shown in Figure 1. Therefore, we propose to learn detailed spatial associations between keypoints in human skeletons and objects.

SIG Construction. As illustrated in Figure 3 (a), the spatial graph of a human-object pair is represented as $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, where $\mathcal{V}_s$ denotes keypoints of human skeletons (blue dots) and the object (red dots), $\mathcal{E}_s$ represents edges among these keypoints. We utilize 17 human skeleton keypoints and the top-left and bottom-right corners of the object. Thus the number of nodes N_s is 19. Each node is represented as its spatial location and confidence score $(x_h, y_h, x_o, y_o, s_k)$, where s_k represents the detection score. The (x_h, y_h) and (x_o, y_o) are relative locations to the human gravity center and object center accordingly. $\mathcal{E}_s$ contains edges between keypoints in human skeletons according to human skeleton connectivity (orange solid lines) and edges connecting 5 major skeleton joints

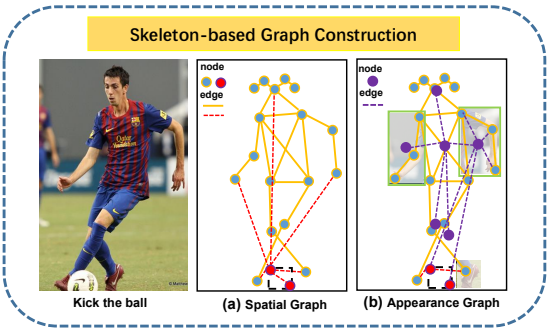

Fig. 3. The construction of our skeleton-based graph including the spatial graph (a) and appearance graph (b).

(including nose, both hands and both feet) with the object corners (red dashed lines).

Attentional SIG Encoding. In order to effectively encode SIG, we employ the graph convolutional network [14]. For each node $v_i \in \mathcal{V}_s$, the graph convolutional kernel considers its neighboring node set S_i as context to update its representation. In order to take spatial arrangement of neighboring nodes into account in the convolution, we follow the spatial partitioning strategy proposed in [21] which divides S_i into three subsets $\{S_i^k\}_{k=0}^2$. The subset S_i^0 denotes the node v_i itself; S_i^1 contains neighboring nodes that are closer to the skeleton gravity center than v_i; and S_i^2 contains the remained neighboring nodes. We utilize different weight matrices W_k^l for convolution of different subsets with residual connection, which is:

$$g_i^{l+1} = g_i^l + \frac{1}{|S_i|}\delta(\sum_{k=0}^{2}\sum_{v_j \in S_i^k} W_k^l g_j^l) \tag{3}$$

where g_i^l is the output in the l-th layer for graph node v_i, and δ is the elu activation function. However, different neighbours can contribute differently to a node especially with the existence of object, which implicitly determines which part of skeletons are more relevant in the interaction. Therefore, we improve the graph convolution with graph attention mechanism [22] to dynamically select more important neighbouring nodes in graph encoding, which is:

$$\hat{\beta}_{ij}^l = \phi(a^l[W_i^l g_i^l \oplus W_j^l g_j^l]) \tag{4}$$

$$\beta_{ij}^l = \frac{\exp(\hat{\beta}_{ij}^l)}{\sum_{v_j \in S_i}\exp(\hat{\beta}_{ij}^l)} \tag{5}$$

where ϕ is the Leaky ReLU activation layer, W_i^l, W_j^l and a^l are learnable parameters at l-th layer. The output of our attentional graph encoding for the node v_i at l-th layer then is computed as:

$$g_i^{l+1} = g_i^l + \delta(\sum_{k=0}^{2} \sum_{v_j \in S_i^k} \beta_{ij}^l W_k^l g_j^l) \quad (6)$$

We stack the attentional graph convolution for multiple layers to propagate sufficient contexts for each node. The final output of SIG is denoted as g_s^i for i-th node. We feed the concatenation of all node representations into a fully-connected layer to obtain the spatial graph feature f_s:

$$f_s = \text{FC}(g_s^1 \oplus g_s^2 ... \oplus g_s^{N_s}) \quad (7)$$

4.2. Appearance Interactive Graph

We propose the appearance interactive graph (AIG) to capture fine-grained visual appearance of how human interacts with object, which shares similar principles as in SIG.

AIG Construction. We denote the AIG as $\mathcal{G}_a = (\mathcal{V}_a, \mathcal{E}_a)$ as illustrated in Figure 3 (b). The nodes $\mathcal{V}_a$ include skeleton nodes to cover local visual regions in human skeleton and object node that represents visual feature of the object. We propose two types of skeleton nodes including keypoint nodes and bodypart nodes. The keypoint nodes are located in the same location as $\mathcal{V}_s$ in SIG. We extract local appearance features from cropped bounding boxes centered at keypoints in $\mathcal{V}_s$ as initial node inputs. The size of the cropped region is in 0.1 fraction of the human bounding box b_h. However, since visual keypoint nodes are disconnected and their relative features are not as meaningful as spatial features in SIG, simply connecting visual keypoint nodes still suffers from the lack of visual interaction contexts. To address this problem, we further introduce bodypart nodes comprised of relevant keypoints. For example, the bodypart "head" consists of the keypoint nodes of eyes, ears and nose. The bodypart nodes cover minimal regions of corresponding keypoints, which are shown in purple dots of Figure 3 (b). For the edges $\mathcal{E}_a$ in AIG, we firstly add intra-connections among keypoint nodes and object nodes according to human skeleton structure in the same way as SIG. Then we link connections among bodypart nodes (purple lines) to add visual contexts.

Attentional AIG Encoding. We employ the same graph attention network architecture as in SIG for AIG encoding. The final embedding for the i-th node is denoted as g_a^i. We concatenate embeddings of all nodes and utilize fully-connected layers to obtain the fine-grained graph appearance representation, which is:

$$f_a = \text{FC}(g_a^1 \oplus g_a^2 ... \oplus g_a^{N_a}) \quad (8)$$

4.3. Joint Training

We train the overall model with multi-task objective losses in order to improve the discrimination ability of each individual prediction module.

Binary interactive loss $\mathcal{L}^{bi}$ is from the binary interactiveness classification, which shares the same feature f_c, f_g with the HOI prediction and learns to predict whether the human interacts with the object based on the classification score s_{bi}.

Human interactive loss $\mathcal{L}^h$ trains the capability of our fine-grained prediction module to encode human-related knowledge, which only employs fine-grained human skeleton embedding f_g and human appearance feature f_h to predict interaction score s_h.

Human-Object interactive loss $\mathcal{L}^{ho}$ is the cross entropy loss of the HOI relationship prediction (s_{ho}), which combines features from both coarse-grained and fine-grained modules f_c, f_g Finally, the joint objective functions for a human-object pair is as follows:

$$\mathcal{L} = \mathcal{L}^{ho} + \lambda_1 \mathcal{L}^{bi} + \lambda_2 \mathcal{L}^h \quad (9)$$

where λ_1, λ_2 are hyper-parameters. We average the loss over all pairs in training.

5. EXPERIMENTS

To demonstrate the effectiveness of our proposed SIGN model, we carry out extensive experiments on two benchmark datasets, HICO-DET [3] and V-COCO [10]. We first describe our experimental settings and then compare our model with the state-of-the-arts. Finally, we carry out ablation studies to evaluate contributions of different components in our model.

5.1. Experimental Settings

Datasets. The HICO-DET dataset contains 47,776 images (38,118 for training and 9,658 for testing) and more than 150K human-object pairs. It consists of 600 HOI categories, 80 object categories and 117 verbs. The V-COCO dataset is a subset of MS-COCO [23], which includes 10,346 images (2,533 for training, 2867 for validation and 4926 for testing) and 16,199 person instances with 29 unique action categories.

Metrics. We follow the same settings as in [3] to evaluate the HOI detection results and use the mean average precision (mAP) metric. An HOI prediction is true positive only when IoUs of both human and object bounding boxes are greater than 0.5 and the interaction category prediction is correct. For V-COCO, we utilize the mAP for actions with roles as metrics. For HICO-DET, we follow previous works [24] with two modes in evaluation: Default and Knowledge. The annotations in Knowledge mode are cleaner than those in Default mode. For each mode, besides mAP on all categories, we also separately evaluate mAP on rare and non-rare classes.

Implementation Details. We use Faster R-CNN with ResNet50-FPN [25] as the backbone network. During training, we freeze ResNet50 backbone and finetune parameters of FPN component. The human pose consists of 17 skeleton keypoints and is detected by AlphaPose [26] pretrained on MS-COCO dataset, which is the same as [5] for fair comparison. We use SGD for training and set the weight decay as

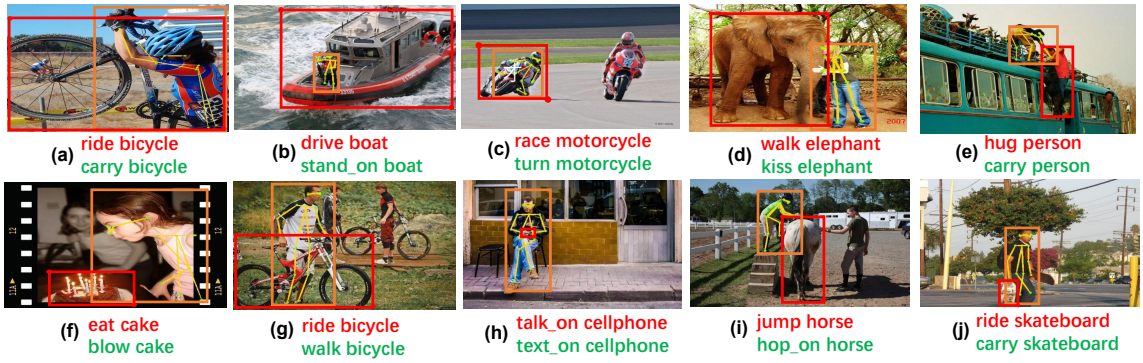

Fig. 4. Examples from HICO-DET dataset. The baseline results are in red and results of our SIGN model are in green.

Table 1. Comparison Results on HICO-DET and V-COCO.

Method	V-COCO	HICO-DET					
	AP_{role}	Default			Knowledge		
		Full	Rare	Non-Rare	Full	Rare	Non-Rare
HO-RCNN [3]	-	7.81	5.37	8.54	10.41	8.94	10.95
InteractNet [27]	40.0	9.94	7.16	10.77	-	-	-
GPNN [15]	44.0	13.11	9.34	14.23	-	-	-
iCAN [12]	45.3	14.84	10.05	16.15	16.26	11.33	17.73
TIN [4]	48.2	17.03	13.42	18.11	19.17	15.51	20.26
PMF-Net [5]	52.0	17.46	**15.65**	18.00	20.34	17.47	21.20
baseline	49.9	17.14	13.96	18.25	19.44	16.02	20.58
baseline+PM	50.5	17.36	14.57	18.27	19.89	16.33	20.90
SIGN	**53.1**	**17.51**	15.31	**18.53**	**20.49**	**17.53**	**21.51**

1e-4 and momentum as 0.9. We initialize the learning rate as 4e-2 for V-COCO with 72k iteration and 1e-4 for HICO-DET with 25 training epochs. Finally, the ratio between positive and negative samples is 1 : 3. Please note that each experiment is carried out three times and the average performance is reported is this paper.

5.2. Comparison with State-of-the-arts

In this section, we compare our proposed model with the following state-of-the-art methods: 1) **iCAN** [12]: an instance-centric attention network exploring attention mechanism to highlight salient interactions regions. 2) **TIN** [4]: a two-stage HOI system with Non Interaction Supression (NIS) to remove irrelevant human-object pairs. 3) **PMF-Net** [5]: including a zoom-in module to extract local skeleton-centric features with semantic-based attention on more relevant skeletons. 4) **baseline**: consisting of the coarse-grained prediction module using the same training and inference as the SIGN model. 5) **baseline+PM**: combining our baseline with pose map (PM) encoding similar to [4, 5].

For the V-COCO dataset, our baseline achieves 49.9 AP_{role} which is already comparable with some recent works like TIN [4]. Combining the baseline with simple human skeleton feature in baseline+PM improves the HOI prediction performance with 0.6 absolute gains, which shows that the fine-grained human skeleton knowledge is useful. The proposed SIGN model achieves significant improvement over the two strong baselines, and outperforms TIN [4] and PMF-Net [5] with 4.9 and 1.1 gains respectively. It demonstrates that our skeleton-based interactive graph can employ human skeletons more effectively than previous works. The trend is similar on the HICO-DET dataset. We achieve the state-of-the-art performance across all metrics under the cleaner Knowledge evaluation setting as well.

In Figure 4, we visualize some example results. Our SIGN model can distinguish fine-grained interactions, which benefits from the skeleton-based interactive knowledge. For example, in Figure 4(h), our model can predict the "text_on cellphone" instead of "talk_on cellphone" with the awareness of the fine-grained interaction between human skeletons and the object cellphone.

Table 2. Ablation studies of individual component in the proposed SIGN model. The acronym K denotes using skeleton keypoints, KB represents using skeleton keypoints along with bodyparts and O denotes adding object nodes.

	SIG			AIG			joint train	NIS	AP_{role}
	nodes		attn	nodes		attn			
	K	K+O		K+O	KB+O				
1	×	×	×	×	×	×	×	×	47.8
2	×	×	×	×	×	×	✓	×	49.0
3	×	×	×	×	×	×	✓	✓	49.9
4	✓	×	×	×	×	×	✓	×	51.1
5	×	✓	×	×	×	×	✓	×	51.3
6	×	✓	✓	×	×	×	✓	×	51.6
7	×	×	×	✓	×	×	✓	×	49.8
8	×	×	×	×	✓	×	✓	×	50.0
9	×	×	×	×	✓	✓	✓	×	50.4
10	×	✓	✓	×	✓	✓	✓	×	52.3
11	×	✓	✓	×	✓	✓	✓	✓	53.1

5.3. Ablation Studies

In order to analyze contributions from different components in our model, we carry out ablation studies on the V-COCO dataset. Table 2 presents the experimental results.

Spatial Interactive Graph (SIG). In Row 4-6, we compare different variants of SIG encoding and they all outperform baselines in Row 1-3, which do not consider fine-grained spatial relationships between human and object. Comparing Row

4 and Row 5, we can see that the awareness of object in the graph benefits the encoding of human skeletons. The introduce of attention mechanism in Row 6 further improves the performance by 0.3, which demonstrates the importance of focusing on salient keypoints in association with the object.

Appearance Interactive Graph (AIG). The Row 7-9 presents the performance of different AIG encoding approaches. The AIG with bodypart nodes achieves better performance than AIG with only keypoint nodes as shown in Row 7&8. The graph attention is also beneficial to the AIG encoding with 0.4 absolute gains comparing Row 8 and 9.

Complementarity of SIG and AIG. In Row 10, we combine the SIG and AIG for HOI prediction, which significantly improves the performance with 0.7 absolute gains than the single best model in Row 6. It proves that the SIG and AIG are complementary since they capture the fine-grained interactions between skeleton and object from different aspects.

Joint Training and Non-Interactive Suppression (NIS). The joint training strategy with binary and human interactive loss in Eq (9) can improve both baseline and our SIGN model. Moreover, utilizing the NIS [4] further boosts the prediction performance and we achieve the state-of-the-art results with all the components.

6. CONCLUSION

We propose a skeleton-based interactive graph network (SIGN) to encode fine-grained human skeletons as well as their interaction with the object from both spatial and appearance aspects. As a result, our model is capable of recognizing fine-grained human-object interactions with subtle difference. We conduct extensive experiments to demonstrate the model effectiveness and achieve the state-of-the-art performance on both HICO-DET and V-COCO datasets.

7. ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (No. 61772535), Beijing Natural Science Foundation (No. 4192028), and National Key Research and Development Plan (No. 2016YFB1001202).

8. REFERENCES

[1] J Deng, W Dong, R Socher, L Li, K Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR 2009*.

[2] S Ren, K He, R Girshick, and J Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[3] Y Chao, Y Liu, X Liu, H Zeng, and J Deng, "Learning to detect human-object interactions," in *WACV 2018*. IEEE, 2018.

[4] Y Li, S Zhou, X Huang, L Xu, Z Ma, H Fang, Y Wang, and C Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *CVPR 2019*, 2019.

[5] B Wan, D Zhou, Y Liu, R Li, and X He, "Pose-aware multi-level feature network for human object interaction detection," in *ICCV 2019*, 2019.

[6] Y Wang, H Jiang, Mark S Drew, Z Li, and G Mori, "Unsupervised discovery of action classes," in *CVPR 2006*.

[7] N Ikizler, R G Cinbis, S Pehlivan, and P Duygulu, "Recognizing actions from still images," in *ICPR 2008*. IEEE, 2008.

[8] W Yang, Y Wang, and G Mori, "Recognizing human actions from still images with latent poses," in *CVPR 2010*.

[9] C Lu, R Krishna, M Bernstein, and Li Fei-Fei, "Visual relationship detection with language priors," in *ECCV 2016*. Springer, 2016.

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

[11] B Zhuang, Q Wu, C Shen, I Reid, and Anton van den Hengel, "Care about you: towards large-scale human-centric visual relationship detection," *arXiv preprint arXiv:1705.09892*, 2017.

[12] C Gao, Y Zou, and Jia-B Huang, "ican: Instance-centric attention network for human-object interaction detection," in *BMVC 2018*, 2018.

[13] L Shen, S Yeung, J Hoffman, G Mori, and Li Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *WACV 2018*. IEEE, 2018.

[14] Thomas N. K and Max W, "Semi-supervised classification with graph convolutional networks," in *ICLR 2017*, 2017.

[15] S Qi, W Wang, B Jia, J Shen, and S Zhu, "Learning human-object interactions by graph parsing neural networks," in *ECCV 2018*, 2018.

[16] K Kato, Y Li, and A Gupta, "Compositional learning for human object interaction," in *ECCV 2018*, 2018.

[17] H Fang, J Cao, Y Tai, and C Lu, "Pairwise body-part attention for recognizing human-object interactions," in *ECCV 2018*, 2018.

[18] B Dai, Y Zhang, and D Lin, "Detecting visual relationships with deep relational networks," in *CVPR 2017*, 2017.

[19] S Zheng, S Chen, and Q Jin, "Visual relation detection with multi-level attention," in *ACM Multimedia 2019*.

[20] J Devlin, M Chang, K Lee, and K Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*. 2019, ACL 2019.

[21] S Yan, Y Xiong, and D Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI 2018*.

[22] Petar V, Guillem C, Arantxa C, Adriana R, Pietro L, and Yoshua B, "Graph attention networks," in *ICLR 2018*.

[23] T 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 2014*. Springer, 2014.

[24] Y Chao, Z Wang, Y He, J Wang, and J Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *ICCV 2015*, 2015.

[25] T Lin, P Dollár, R Girshick, Kaiming He, B Hariharan, and S Belongie, "Feature pyramid networks for object detection," in *CVPR 2017*, 2017.

[26] H Fang, S Xie, Y Tai, and C Lu, "Rmpe: Regional multi-person pose estimation," in *ICCV 2017*, 2017.

[27] G Gkioxari, R Girshick, P Dollár, and Kaiming He, "Detecting and recognizing human-object interactions," in *CVPR 2018*, 2018.