Hierarchical Reasoning Network for Human-Object Interaction Detection

Yiming Gao, Zhanghui Kuang, *Member, IEEE*, Guanbin Li, *Member, IEEE*, Wayne Zhang, and Liang Lin, *Senior Member, IEEE*

***Abstract*— Human-object interaction detection that aims at detecting <human, verb, object> triplets is critical for the holistic human-centric scene understanding. Existing approaches ignore the modeling of correlations among hierarchical human parts and objects. In this work, we introduce a Hierarchical Reasoning Network (HRNet) to capture relations among human parts at multiple scales (including the holistic human, human region, and human keypoint levels) and objects via a unified graph. In particular, HRNet first constructs one multi-level human parts graph, each level of which consists of human parts at one specific scale, objects, and the unions of human part-object pairs as nodes, and their mutual visual and spatial layout relations as intra-level reasoning. To also capture the relations across scales, we further introduce inter-level reasoning between the nodes of two consecutive levels based on the prior of human body structure. The representations of graph nodes are propagated along intra-level and inter-level reasoning in turn during reasoning. Extensive experiments demonstrate our HRNet obtains new state-of-the-art results on three challenging HICO-DET, V-COCO and HOI-A benchmarks, validating the compelling effectiveness of the proposed method.**

***Index Terms*— Human-object interaction, hierarchical reasoning network, graph neural network.**

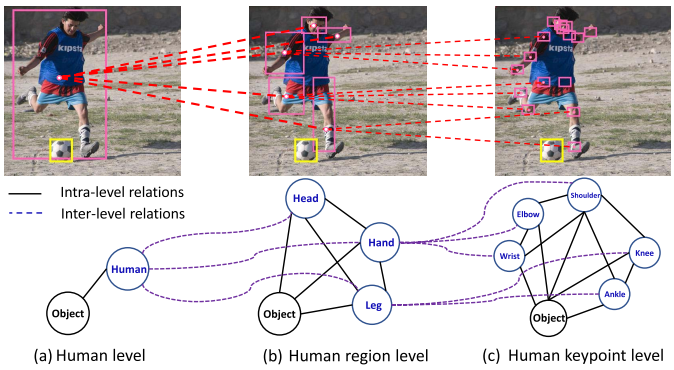

Fig. 1. How do human recognize human-object interactions such as kicking a football? We conjecture that the detection of human-object interaction depends on not only the human and object appearances but also underlying relations among human parts and objects at different scales and those across scales from coarse to fine and vice versa. Thus, we propose to model human-object interactions as one hierarchical graph consisting of three levels which correspond to human, human region, and human keypoint scales respectively.

I. Introduction

WITNESSING the great progress in instance understanding such as object detection [1]–[3] and human pose estimation [4]–[6] in recent years, the vision community steps forward to comprehend visual relationships between individual instances. Human-Object Interaction (HOI) detection aims at detecting and recognizing triplets of human, object and their relationships in images, which plays an important role in the image holistic comprehension. It benefits various vision tasks (*e.g.*, visual question answering [7], visual grounding [8], image and video caption [9] and activity recognition [10]) and is being widely studied.

Recent works [11]–[14] make attempt to explore human part representation for HOI detection to distinguish fine-grained interactions. They either directly extract representations of human poses [11], [12] or pairwise body part interaction [14], or use human poses to align human part representations via one zoom-in module [13]. Although they achieve impressive results, they all utilize single-scale human part representations only as additional local clues, and overlook the hierarchical structure of human bodies, which leads to sub-optimal performance.

How do humans recognize fine-grained human object interactions (*e.g.*, distinguishing person kicking a football from person stopping a football, as shown in Fig. 1) ? We conjecture that humans recognize action from the perspective of analyzing both coarse and fine level simultaneously. Humans first take a glance at the human and object appearance and their holistic positional relations. But just recognizing from single human level is not sufficient. Humans also focus on the fine level of relative human parts, and analyze the appearance and spatial relations among the parts and object. Finally they determine the interactions by considering both coarse and fine

Manuscript received December 17, 2020; revised May 16, 2021; accepted June 22, 2021. Date of publication September 29, 2021; date of current version October 5, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFC0830103, in part by the National Natural Science Foundation of China under Grant 61976250 and Grant U1811463, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2020B1515020048, in part by Guangzhou Science and Technology Project under Grant 202102020633, and in part by CCF-Tencent Open Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peihua Li. *(Corresponding author: Guanbin Li.)*

Yiming Gao, Guanbin Li, and Liang Lin are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: liguanbin@mail.sysu.edu.cn).

Zhanghui Kuang and Wayne Zhang are with SenseTime Research, Hong Kong.

Digital Object Identifier 10.1109/TIP.2021.3093784

level together. Thus, we need to model the relations of both human-object at multiple scales and human-region-keypoint.

Inspired by the procedure described above, we step forward to propose a novel Hierarchical Reasoning Network (HRNet) to simultaneously capture and represent the above cues via a unified hierarchical graph. As described above, the cues need to represent appearance and spatial information at both coarse and fine level. The cues also contain relationships among human parts and object with various levels, including appearance and spatial relationships. To represent multiple levels information, we build the graph which consists of three levels from top to bottom named human, human region and human keypoint level, as shown in Fig 1. Each sub-graph at one level represents the visual appearance and spatial information of corresponding body parts and object. The three levels can model the relationships among human parts and objects in terms of visual appearance correlation and spatial layouts at human, human regions and human keypoint level respectively. The consecutive graph levels are connected with the prior of human body structure. In this way, HRNet can learn multi-level human object interaction representations between coarse and fine simultaneously. Multi-level representations can be mutually enhanced via messages propagation along the intra-level and inter-level connections in it. We finally concatenate multi-scale representations to classify the category of human object interactions.

Different from previous methods [11], [12], [14], which sophisticatedly and individually design one network subbranch for each kind of information, our proposed HRNet equivalently considers the information of humans, human parts, human poses, objects and union boxes of humans and objects. The proposed HRNet provide the techniques of representing multi-level information as graph nodes and their visual and spatial relations as graph edges in one unified graph structure. Thanks to unified graph structure, our framework is extensible, and easily incorporates other information. All the representations of nodes and edges are learnable without handcrafting. Specifically, HRNet can be applied to not only the human-object interaction detection, but also the challenging fine-grained classification via representing the local subtle regions and capturing their relationships.

We conduct extensive experiments on three human-object interaction benchmarks, V-COCO [16], HICO-DET [17] and HOI-A [18]. Our proposed HRNet obviously outperforms its counterpart without hierarchical reasoning even with a stronger baseline. It achieves new state-of-the-art results on HICO-DET, V-COCO and HOI-A in terms of average precision (AP). Specifically, it obtains an AP of 21.93%, 53.1% and 67.26% on HICO-DET with default setting, V-COCO and HOI-A respectively. Moreover, the experiments are also conducted on a challenging fine-grained classification dataset, which demonstrate the superiority of our HRNet in both accuracy and generality, and also validate the effectiveness of each component.

Our main contributions are summarized in three aspects, as follows:

- We propose the tailored hierarchical reasoning network (HRNet) for human-object interaction detection, in which we incorporate the relations among body parts and object at multi-scale levels with the appearance and spatial information in a unified graph.
- We provide extensive experiments on several challenging human-object interaction detection benchmarks and show that our proposed HRNet achieves new state-of-the-art performances on three challenging benchmarks, *i.e*., HICO-DET, V-COCO and HOI-A.
- We also demonstrate the effectiveness of HRNet on fine-grained classification, in which we model the relations among regions at multiple levels in a unified graph. We validate its effectiveness on different baselines and show that it can improve the fine-grained classification performance obviously.

II. Related Work

A. Visual Relationship Detection

Visual relationship detection [19]–[24] targets at detecting generic objects and their mutual relationships simultaneously in images. Recent work explore languages priors [20], visual translation embedding [21], deep feature interaction [23], and attention mechanism [24] to improve its performance. Different from visual relationship detection, we aim at a related but different task, human-object interaction detection, which has different challenges. *e.g*., dramatic human pose variations.

B. Human-Object Interaction Detection

Human-Object Interaction (HOI) detection [16], [17] aims at detecting the human-centric relationship, which is essential for human-centric understanding in the vision. Comparing with the visual relationship detection, HOI detection is more fine-grained and a multi-label problem, *e.g*., a person is eating and holding an apple, and talking on a phone while lying on a bed. Earlier methods [16], [17], [25] employ a multi-stream framework to directly use human appearances, object appearances and spatial configurations separately to detect HOIs. Recently, some studies [11]–[15] make attempt to use local representations of human parts as auxiliary clues to enhance the global representation for HOI detection, which has achieved impressive performance improvements. Specifically, PMFNet [13] concatenates local representations around human keypoints weighted by attention with the global one. Some recent studies [18], [26], [27] propose a stronger detector to improve the HOI detection performance. DJ-RN [28] proposes to incorporate detailed 3D information Neither of the two methods model any relation among human local parts and objects.

Another solutions [11], [15], [29]–[31] for HOI detection are to explore the relations among human and objects. Fang *et al*. [14] explore the correlations of human keypoint pairs to focus on crucial parts. No-frills [11] explicitly encodes the spatial relative position without considering their visual correlations. ACP [29] explicitly leverage the action co-occurrence priors for HOI detection. GPNN [31], DRG [30] and RPNN [15] model the relations between objects and human instances or human regions such as the head, hands and legs, and those between human regions and human. Different

TABLE I

COMPARISON BETWEEN APPROACHES IN TERMS OF WHETHER MODELING HUMAN REGIONS, HUMAN KEYPOINTS, HUMAN PART INTRA-LEVEL/INTER-LEVEL RELATIONS, AND HUMAN PART OBJECT RELATIONS. NO-FRILLS [11] MODELS THE SPATIAL RELATIONS BETWEEN HUMAN PARTS AND OBJECTS ONLY WITHOUT THEIR VISUAL CORRELATIONS. OUR HRNET ENCODES ALL THE INFORMATION IN ONE UNIFIED GRAPH WITHOUT SOPHISTICATED DESIGN FOR EACH ONE

Methods	TIN [10]	PMFNet [11]	PBPA [12]	RPNN [13]	No-frills [9]	Ours
Human regions				✓		✓
Human keypoints	✓	✓	✓		✓	✓
Intra-level relations			✓			✓
Inter-level relations				✓		✓
Part-object relations				✓	✓	✓
Spatial information	✓	✓			✓	✓

from all above mentioned methods, our proposed HRNet not only encodes the spatial information and multi-scale visual representations, but also models intra-level and inter-level relations among human parts at multi-scales and objects. More detailed comparisons between our method with previous approaches can be found in Table I. Though some of the cues are addressed in existing methods, we step forward to explore a novel hierarchical graph structure to model the relations within and across the hierarchical levels. The proposed hierarchical graph module can uniformly represent the multi-level appearance and spatial features and capture the relations among them.

C. Graph Reasoning

Graph reasoning has shown to have substantial practical merits for many tasks, such as node and graph classification [32]–[34], object detection & parsing [35]–[37], visual grounding [38]–[40], visual question answering [41], fashion retrieval [42], [43] and action recognition [44]–[48]. Previous works [44], [48] build the graph structure for action recognition where the nodes are represented by human or objects. Some works [33], [34] build the hierarchical graph structure through adaptive merging or differentiable graph pooling. However, they ignore the prior knowledge such as human skeleton to build the hierarchical structure. Recently, GPNN [31] and RPNN [15] construct the graph structure using single scale of human regions and objects as nodes to model the relations between human and objects. However, Neither of them model the relations among human parts at multi-scale levels. In contrast, we customize the graph structure to address the HOI detection. Comparing with normal Graph Neural Network pipeline, we emphasize the different as follows: 1) incorporation of prior knowledge (e.g., human skeleton) to build the hierarchical structure; 2) construction of three sub-graphs which models the relations between regions of specific human level and the object; 3) the design of inter-level reasoning to connect the sub-graphs among different human structure levels.

D. Fine-Grained Classification

Our work is also related to fine-grained image classification. Fine-grained image classification aims at recognizing sub-classes of some given object categories, such as species of birds. Previous works tackle this problem with extra annotations, such as bounding box [49], [50]. Recent studies [51]–[53] mainly focus on locating the local discriminative regions with weak supervision. Some recent approaches propose to extract more discriminative representations via the pooling operation [54]–[57] and the bilinear operation [58]. Some recent approaches [59], [60] exploit extra explicit knowledge to model the relations among categories. Instead, we explore to learn the relations of local regions without extra explicit knowledge.

III. METHODOLOGY

A. Overview

The goal of HOI detection is to recognize the relationship of each pair of detected human and object in the image. Given an image $\mathbf{I}$, we first use an off-the-shelf object detector [1] to obtain the human/object boxes $\mathbf{b}_h, \mathbf{b}_o$, where $\mathbf{b}_h, \mathbf{b}_h \in \mathbb{R}^4$. Then, we can easily obtain the union box b_u. Second, we further denote the bounding boxes of human parts at level l by $\mathbf{B}_p^l = [\mathbf{b}_{p_l}^1, \ldots, \mathbf{b}_{p_l}^{N_l}] \in \mathbb{R}^{4\times N_l}$ for the detected human $\mathbf{b}_h$, where N_l is the number of human parts at level l and $l \in \{1, \ldots, L\}$ indicates the level index from top to down. We treat the whole human as the human part at level 1 for simplification. Thus, $N_1 = 1$ and $\mathbf{b}_{p_l}^1 = \mathbf{b}_h$. For each pair of the human and object $< \mathbf{b}_h, \mathbf{b}_o >$ in an image, the vector score of interactions is given by

$$\mathbf{s}_{hoi} = f(\mathbf{I}, \mathbf{b}_h, \mathbf{b}_o, \mathbf{b}_u, \{\mathbf{B}_p^l\}_{l=1}^L), \tag{1}$$

where f is the prediction network. Let $\mathbf{f}_{p_l}^i$, $\mathbf{f}_o$ and $\mathbf{f}_u$ denote the appearance features corresponding to the human part $\mathbf{b}_{p_l}^i$, the object $\mathbf{b}_o$ and the human-object union region $\mathbf{b}_u$ respectively. They are extracted by RoI-Align [2] from the output feature maps of the backbone stream based on their bounding boxes. We propose the hierarchical reasoning module to model the relations among human parts with bounding boxes $\mathbf{b}_p^{l,i}$ at multiple scales and objects with bounding box $\mathbf{b}_o$. It can refine the human object interaction representation via message propagation, and output multi-scale feature vectors. We stack the hierarchical reasoning module upon the convolutional neural network, and obtain the Hierarchical Reasoning Network (HRNet). In this section, we first introduce our Hierarchical Reasoning Module in Section III-B, and then describe the overall architecture of HRNet in Section III-C.

B. Hierarchical Reasoning Module

Our proposed hierarchical reasoning module targets at learning discriminative representation of human object interaction.

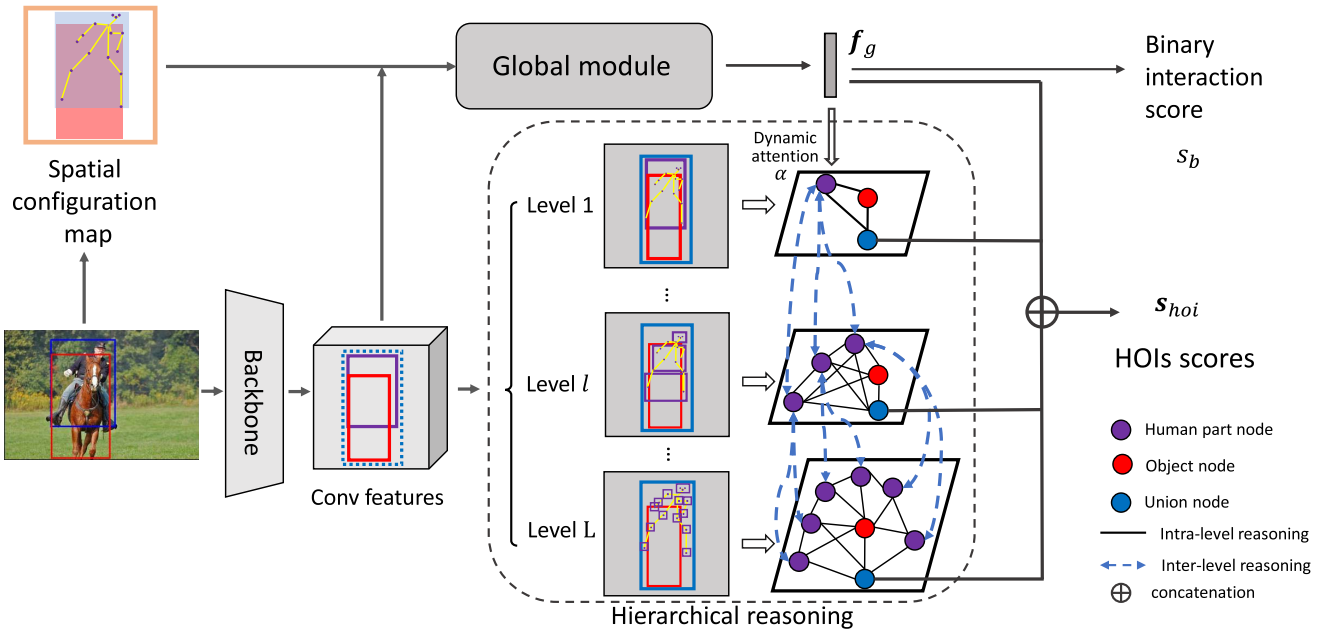

Fig. 2. The overall framework of our proposed HRNet. Given a pair of detected human and object proposals, their features extracted by the backbone stream, and their spatial configuration map encoding human and object bounding boxes and human pose, are fed into the global module to extract the global feature $\mathbf{f}_g$, which predicts one interaction scalar score s_b. The features of proposals are also fed into the hierarchical reasoning module to build one multi-level graph, each level of which consists of human part nodes at one specific scale, one object node, and one human part-object union node. Features in the graph are attentively enhanced by the dynamic attention α generated by $\mathbf{f}_g$, and then propagated along the intra-level and inter-level edges. Finally, the concatenation of features of the union nodes at multiple levels and the global feature, is classified into HOI categories, obtaining one human object interaction category vector score $\mathbf{s}_{hoi}$.

It is designed as multi-level graph, in which the l^{th} level consists of human parts $\mathbf{b}_p^{l,i}$, the object $\mathbf{b_o}$, and the human-object union $\mathbf{b_u}$ as nodes, and their mutual visual correlations and spatial layouts as edges. Each node in the graph encodes not only its corresponding appearance feature (*i.e.*, $\mathbf{f}_p^{l,i}$, $\mathbf{f}_o$ and $\mathbf{f}_u$), but also spatial information. To represent the spatial information, we denote the $\mathbf{s}_p^{l,i}$, $\mathbf{s}_o$ and $\mathbf{s}_u$ as the spatial information for $\mathbf{b}_p^{l,i}$, $\mathbf{b_o}$, and $\mathbf{b_u}$ respectively. Specifically, we represent and embed the spatial information for each node by one 6-dimensional vector consisting of corresponding bounding box center coordinate, width, height, and bounding box area normalized w.r.t the whole input image, and the bounding box aspect ratio. The node representations of human parts, the object and the human-object union are given by

$$\begin{aligned}\mathbf{x}_p^{l,i} &= [\mathbf{f}_p^{l,i} || f_s(\mathbf{s}_p^{l,i})], i = 1, \cdots, N_l,\\ \mathbf{x}_o^{l} &= [\mathbf{f}_o^{l} || f_s(\mathbf{s}_o)],\\ \mathbf{x}_u^{l} &= [\mathbf{f}_u^{l} || f_s(\mathbf{s}_u)],\end{aligned} \tag{2}$$

where $||$ is the concatenation operation, and $\mathbf{f}_o^l$ and $\mathbf{f}_u^l$ are feature vectors of the object $\mathbf{b}_o$ and the human-object union $\mathbf{b}_u$ at the l^{th} level after pooling. f_s is one MLP with two layers and one ReLU between them for embedding the spatial information. In this way, we finally obtain the graph node representation of the l^{th} level $\mathbf{X}^l = [\mathbf{x}_p^{l,1}; \cdots; \mathbf{x}_p^{l,N_l}; \mathbf{x}_o^l; \mathbf{x}_u^l] \in \mathbb{R}^{C\times(N_l+2)}$ where C is the feature dimension.

There are two kinds of reasoning in our multi-level graph. Namely, the intra-level reasoning and inter-level reasoning. The intra-level reasoning targets at modeling the intra-level relations among human parts at one specific scale, the object and the human-object pair. And the inter-level reasoning models the inter-level relations those among human parts across scales.

1) Intra-Level Reasoning: We represent the relations between nodes at the same level via their appearance correlations and spatial layouts. We establish appearance relations between nodes using the attention mechanism [61]. Formally, the edge $e_a^l(i, j)$ between node i and j at level l is formulated as

$$e_a^l(i, j) = \frac{\exp(f_a(\mathbf{x}_i^l, \mathbf{x}_j^l))}{\sum_k \exp(f_a(\mathbf{x}_i^l, \mathbf{x}_k^l))}, \tag{3}$$

where f_a is one MLP with two layers and one LeakyReLU followed by them, as done in [61]. The negative input slope of the LeakyReLU is set to 0.2. We denote the appearance relations at level l by the adjacency matrix $\mathbf{E}_a^l \in \mathbb{R}^{(N_l+2)\times(N_l+2)}$ with $e_a^l(i, j)$ being its (i, j) element. To obtain the spatial layout, we establish the spatial relations between nodes via computing their normalized l2-distance:

$$e_s^l(i, j) = 1 - \frac{d(i, j)}{\sqrt{w_{ij}^2 + h_{ij}^2}}, \tag{4}$$

where $d(i, j)$ denotes the l2-distance between the center coordinates of node i and j. w_{ij} and h_{ij} indicates the width and height of the human box $\mathbf{b}_h$ when both node i and i correspond to human parts, or the width and the height of union box $\mathbf{b}_u$ when either node i or node j refers to the object. We construct the adjacency matrix encoding spatial

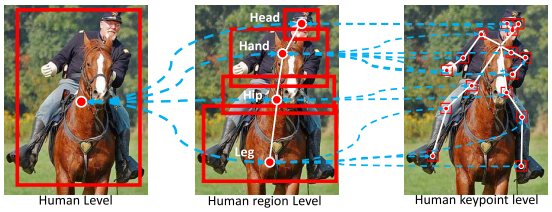

Fig. 3. The illustration of the inter-level edges. We connect the nodes of two consecutive levels based on the human body structure. The head node is connected with five keypoint nodes. *i.e.*, the nodes of the nose, the left eye, the right eye, the left ear, and the right ear; The torso node is connected with the nodes of the left hand, the right hand, the left wrist, the right wrist, the left elbow and the right elbow. The hip node is connected with the nodes of the left hip and the right hip. The leg node is connected with the left knee, the right knee, the left ankle and the right ankle.

layouts as $\widehat{\mathbf{E}}_s^l \in \mathbb{R}^{(N_l+2)\times(N_l+2)}$ with $e_s^l(i,j)$ being its (i,j) element. We normalize $\widehat{\mathbf{E}}_s^l$ following [32] to stabilize the training, obtaining $\mathbf{E}_s^l = \widehat{\mathbf{D}}^{-\frac{1}{2}}\widehat{\mathbf{E}}_s^l\widehat{\mathbf{D}}^{-\frac{1}{2}}$, where $\widehat{\mathbf{D}}$ is one diagonal matrix with $\widehat{\mathbf{D}}_{ii} = \sum_{j\neq i} e_s^l(i,j)$. $\mathbf{E}_s^l$ is a normalized adjacency matrix encoding spatial relations between nodes at level l.

With the appearance relation $\mathbf{E}_a^l$ and the spatial relation $\mathbf{E}_s^l$, we refine the node features by message propagation in our proposed multi-level graph at each level l independently. Formally, we perform graph reasoning to evolve the node representation $\mathbf{X}^l$, which can be given by

$$\widehat{\mathbf{X}}^l = \mathrm{ReLU}(\mathbf{W}_a^p(\mathbf{X}^l\mathbf{E}_a^l)) + \mathrm{ReLU}(\mathbf{W}_s^p(\mathbf{X}^l\mathbf{E}_s^l)), \tag{5}$$

where $\mathbf{W}_a^p \in \mathbb{R}^{C'\times C}$ and $\mathbf{W}_s^p \in \mathbb{R}^{C'\times C}$ are the learnable weight. $\widehat{\mathbf{X}}^l \in \mathbb{R}^{C'\times(N_l+2)}$ is the updated node representation of the l^{th} level after once message propagation.

2) Inter-Level Reasoning: Besides capturing the relations among human parts, the object, and the union of human and object region at each level, we also model those across scale levels via constructing inter-level edges. We establish the inter-level edges between two consecutive levels according to the prior knowledge of the body structure. As shown in in Fig 3, our designed three levels of the graph correspond to scales of human, human regions, and human keypoints from coarse to fine respectively. The human is represented by one node at the top level, while four human regions (*i.e.*, head, torso, hip, and leg) at the middle level and 17 keypoints at the bottom level. Four human regions are connected to the human node at the top level while each node in the human region level is connected to its corresponding keypoints in the third level (see the connections in Fig 3). For any inter-level connection between $\mathbf{x}_i^{l_1}$ and $\mathbf{x}_j^{l_2}$, we define a scalar edge weight $e_h^{l_1,l_2}(i,j)$, which is given by

$$e_h^{l_1,l_2}(i,j) = \frac{\exp(f_h(\mathbf{x}_i^{l_1}, \mathbf{x}_j^{l_2}))}{\sum_{(l,k)\in\mathbb{N}_i^{l_1}} \exp(f_h(\mathbf{x}_i^{l_1}, \mathbf{x}_k^{l}))}, \tag{6}$$

where f_h is one MLP with two layers and one ReLU between them. $\mathbb{N}_i^{l_1}$ represents the index set of the connected nodes of $\mathbf{x}_i^{l_1}$ at different levels. It consists of tuples (l,k) with l and k being the level index and the node index of $\mathbf{x}_k^l$ respectively.

Similarly, we refine node representations via propagating and aggregating them across different levels. In this way, our node representations contain multi-scale clues for human object interaction detection. Formally, the refined node feature $\widehat{\mathbf{x}}_i^{l_1}$ is given by

$$\mathbf{x'}_i^{l_1} = \mathbf{x}_i^{l_1} + \sum_{(l,k)\in\mathbb{N}_i^{l_1}} e_h^{l_1,l}(i,k)\mathbf{x}_k^l \tag{7}$$

and

$$\widehat{\mathbf{x}}_i^{l_1} = \mathrm{ReLU}(\mathbf{W}_h\mathbf{x'}_i^{l_1}), \tag{8}$$

where $\mathbf{W}_h \in \mathbb{R}^{D'\times D}$ is the learnable matrix.

3) Sequential Reasoning: There are two kinds of reasoning to propagate information at one specific level or across different levels. Namely, intra-level reasoning models the relations of both appearance correlations and spatial layouts at one specific level, which is described in Eq. 5. And inter-level reasoning models the relations across different levels with the prior knowledge of human body structure, as described in Eq. 8. Since two kinds of reasoning models different relations and each sub-graph $\mathbf{X}^l$ has its specific level of appearance and spatial information, we choose to conduct the two reasoning modules in an alternate sequential manner for better optimization. In each reasoning iteration of our hierarchial reasoning module, we sequentially conduct the two propagations. Specifically, we first perform the intra-level reasoning via Eq. 5, and then perform the inter-level reasoning via Eq. 8. We perform graph reasoning for T times to capture the complex relations among human parts at three scales, the object and the human-object pair union.

C. Network Architecture

As shown in Figure 2, our proposed Hierarchical Reasoning Network (HRNet) is composed of the feature extraction backbone, the hierarchical reasoning module, and the global module.

1) Feature Extraction Backbone: We adopt ResNet-50 [62] with FPN [63] as our backbone following [13]. Given the human and object bounding boxes detected by an off-the-shelf detector, the feature extraction backbone extracts the regional features by employing ROI-Align [2] on the output feature maps with highest resolution of FPN based on the bounding boxes.

2) Global Module: Following [13], the global module first takes the pooled regional appearance features of the human, the object, the union of human-object pair and the spatial configuration as input, as shown in Figure 4 (a). We adopt four non-shared MLPs with 2 layers and ReLU to extract the features of the human, the object, the union region of human-object and the spatial configuration map respectively, and concatenate all output features to obtain the global representation $\mathbf{f}_g$. Specifically, the spatial configuration is encoded by one tensor with size of $M\times M\times 3$, as shown in Figure 4 (b). Its first two channels contain the binary masks of the human and the object, where the value is set to 1 within the human (or the object) bounding box and 0 in other areas. Its third channel contains the human pose mask represented by one line-graph, which plots lines to connect neighbor joints based on the human skeleton configuration of COCO [64].

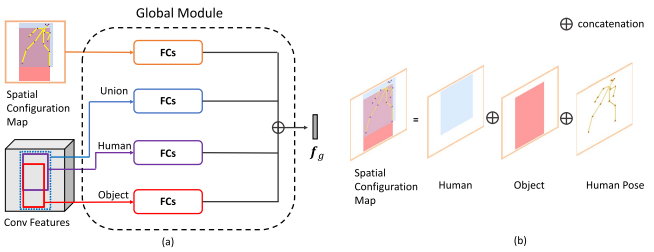

Fig. 4. The structure of the global module and the spatial configuration map. (a) The global module takes the pooled region appearance features of the human, the object, the union region of human-object and the spatial configuration map as input to obtain the global representation $\mathbf{f}_g$. (b) The spatial configuration map contains the binary masks of the human and the object respectively, and a human pose mask encoded by an line-graph.

3) Dynamic Attention: Motivated by the observation of recognizing interactions by human, it can be noticed that not all the human parts are benefit for recognizing interactions in one particular human-object pair. Thus, we would like to focus on the important human parts to guide the reasoning. We utilize the global information $\mathbf{f}_g$ to endow the weight for graph nodes via one MLP with two layers and one RELU between them. Specifically, the network takes the global features $\mathbf{f}_g$ as input, and outputs the attention weight vector $\alpha^l \in \mathbb{R}^{N_l+2}, l \in \{1, \ldots, L\}$. The graph node representations are enhanced attentively, and we obtain $\alpha_i^l \mathbf{x}_i^l$.

4) Loss: Following [12], [13], given a pair of human and object, we predict the binary score s_b of whether interaction exists between the human and object, and the score vector $\mathbf{s}_{hoi}$ of HOI categories simultaneously. For inference, we take the global representation $\mathbf{f}_g$ as input, and then obtain the binary score s_b as follows

$$s_b = \text{sigmoid}(f_b(\mathbf{f}_g)), \tag{9}$$

where f_b is one MLP with two layers and one ReLU between them.

For the HOI category prediction task, we concatenate the global feature from the global module, and representations of the human-object pair union at multi-level together, and then predict the category score vector. Formally, we have

$$\mathbf{s}_{hoi} = \text{sigmoid}(f_{hoi}([\mathbf{f}_g || \mathbf{x}_u^1 || \mathbf{x}_u^2 || \mathbf{x}_u^3])), \tag{10}$$

where f_{hoi} is one MLP with two layers and one ReLU between them. $\mathbf{x}_u^l$ is the feature of the human-object pair union node at the l^{th} level. Finally, we can obtain the Binary Cross Entropy (BCE) loss over $\mathbf{s}_{hoi}$ and s_b as follows:

$$L = L_{BCE}(\mathbf{s}_{hoi}, \bar{\mathbf{s}}_{hoi}) + \mu L_{BCE}(s_b, \bar{s}_b), \tag{11}$$

where $\bar{\mathbf{s}}_{hoi}$ and $\bar{s}_b$ are the ground-true of multi-label HOI category label and binary interaction label respectively, and μ is a hyperparameter used to balance the two loss terms.

IV. Experiments

A. Experimental Settings

1) Implementation Details: We use the Faster R-CNN and AlphaPose [65] pretrained on COCO as the off-the-shelf object detector and pose estimator to obtain the object detection and pose estimation, respectively. The output of the pose estimation is the same as COCO setting with 17 keypoints. For estimating the bounding box of each human region at the second level in our proposed graph, we first estimate the minimum rectangle that contains its connected keypoints and then expand it along the horizontal and vertical direction by 10% of the human bounding box $\mathbf{b}_h$ width and height respectively. Specifically, we build four human parts, named as "head", "hand", "hip" and "leg", in the human region level following [15]. The "head" part contains the keypoints of "left/right eye", "left/right ear" and "nose". The "hand" part contains the keypoints of "left/right hand", "left/right wrist" and "left/right elbow". The "hip" part contains the keypoints of "left/right hip". The "leg" part contains the keypoints of "left/right knee" and "left/right ankle". We set the rectangle around a keypoint with size of 20% $\mathbf{b}_h$ height $\times$ 20% $\mathbf{b}_h$ width to be the keypoint's bounding box in the human keypoint level.

Following [13], we employ ResNet-50 [62] with FPN [63] as feature extractor, and extract 7×7 regional features from the feature maps of highest resolution in FPN for human parts at multi-level.

The dimension of node representation is set to 128, *i.e.*, $C' = D = D' = 128$, and the dimension of the hidden layer of the MLPs in the global module and dynamic attention module is set to 256 while the dimension of hidden layer of MLPs f_b and f_{hoi} is set to 1024. The number of sequential reasoning iteration is set to $T = 2$. The spatial configuration in the global module is encoded by one tensor with size of $64 \times 64 \times 3$. *i.e.*, $M = 64$. The hyper-parameter μ is 0.1.

During training, we adopt the SGD optimizer with the momentum of 0.9 and the weight decay of 0.0001. We set the batch size to 4 with initial learning rate being 0.04. For V-COCO, we train our model for 30 epochs and decrease learning rate by $10\times$ every 15 epochs. For HICO-Det, we train our model for 35 epochs and decrease learning rate by $10\times$ every 25 epochs. The weights of the feature extractor ResNet-50 pretrained on COCO are frozen to avoid overfitting during training. When training the classifier to determine if interaction exists, we select the positive and negative samples with ratio 1:3. During testing, for fair comparison, we use pre-detected object results from [66], and discard the detected human/object results of detection confidence lower than 0.5/0.4.

2) Datasets: We evaluate the performance of our HRNet on three challenging HOI detection benchmarks, V-COCO, HICO-DET and HOI-A. V-COCO [16], which is a subset of COCO [64], contains 10,346 images (2,533 for training, 2,867 for validation and 4,946 for testing) including 16,199 human instances and annotates 26 HOI categories. HICO-Det [17] contains 47,776 images (38,118 for training, 9,658 for testing) and annotates 600 HOI categories on 80 object categories (same as COCO) and 117 verb categories. HOI-A [18] contains 38,668 images (29,842 for training, 8787 for testing), 11 kinds of objects and 10 action categories.

3) Evaluation Metric: Following [16], [17], we adopt the mean average precision to evaluate the performance. The prediction is regarded as a true positive only when the detected bounding boxes of both the human and the object have IoUs

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON HICO-DET TEST SET. '−' INDICATES NO REPORT. 'COCO' IN THE 'DETECTOR' COLUMN INDICATES THAT THE OBJECT DETECTORS ARE FIRST PRE-TRAINED ON MS-COCO [64] AND THEN FIXED WHEN TRAINING HOI MODELS, WHILE 'HICO-DET' INDICATES THAT THE OBJECT DETECTORS ARE FINE-TUNED WITH THE HICO-DET DATASET. HRNET^{ACP} INDICATES OUR HRNET USE ACP [29] AS THE BASELINE

Methods	Detector	Backbone	Default			Know Object		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
HO-RCNN [15] (WACV 18)	COCO	VGG16	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [23] (CVPR 18)	COCO	ResNet50-FPN	9.94	7.16	10.77	-	-	-
GPNN [29] (ECCV 18)	COCO	ResNet101	13.11	9.34	14.23	-	-	-
iCAN [57] (BMVC 18)	COCO	ResNet50	14.84	10.45	16.15	-	-	-
TIN [10] (RP_DC_D) (CVPR 19)	COCO	ResNet50	17.03	13.42	18.11	19.17	15.51	20.26
Wang et al. [45] (ICCV 19)	COCO	ResNet50	16.24	11.16	17.75	17.73	12.78	19.21
No-Frills [9] (ICCV 19)	COCO	ResNet152	17.18	12.17	18.68	-	-	-
RPNN [13] (ICCV 19)	COCO	ResNet50	17.35	12.78	18.71	-	-	-
PMFNet [11] (ICCV 19)	COCO	ResNet50-FPN	17.46	15.65	18.00	20.34	17.47	21.20
VSGNet [58] (CVPR 20)	COCO	ResNet152	19.80	16.05	20.91	-	-	-
DJ-RN [26] (CVPR 20)	COCO	ResNet50	21.34	**18.53**	22.18	23.69	20.64	24.60
UnionDet [24] (ECCV 20)	COCO	ResNet50-FPN	14.25	10.23	15.46	-	-	-
DRG [28] (ECCV 20)	COCO	ResNet50-FPN	19.26	17.74	19.71	23.40	**21.75**	23.89
VCL [59] (ECCV 20)	COCO	ResNet50-FPN	19.43	16.55	20.29	22.00	19.09	22.87
ACP [27] (ECCV 20)	COCO	ResNet152	20.59	15.92	21.98	-	-	-
PD-Net [60] (ECCV 20)	COCO	ResNet152	20.81	15.90	22.28	24.78	18.88	26.54
PPDM [16] (CVPR 20)	HICO-DET	Modified-DLA34 [16]	21.10	14.46	23.09	23.09	16.14	25.17
UnionDet [24] (ECCV 20)	HICO-DET	ResNet50-FPN	17.58	11.72	19.33	19.76	14.68	21.27
Baseline	COCO	ResNet50-FPN	15.47	12.29	16.42	18.98	15.63	19.98
HRNet	COCO	ResNet50-FPN	18.10	15.89	18.76	21.12	18.20	21.99
HRNet^{ACP}	COCO	ResNet152	**21.93**	16.30	**23.62**	**25.22**	18.75	**27.15**

TABLE III

COMPARISON WITH STATE-OF-THE-ART METHODS ON V-COCO TEST SET [16]

Methods	Backbone	AP_{role}
Gputa et al. [14]	VGG16	31.8
InteractNet [23] (CVPR 18)	ResNet50-FPN	40.0
GPNN [29] (ECCV 18)	ResNet101	44.0
iCAN [57] (BMVC 18)	ResNet50	44.7
TIN [10] (RP_DC_D) (CVPR 19)	ResNet50	47.8
Wang et al. [45] (ICCV 19)	ResNet50	47.3
RPNN [13] (ICCV 19)	ResNet50	36.7
PMFNet [11] (ICCV 19)	ResNet50-FPN	52.0
Cascade-HOI [25] (CVPR 20)	ResNet50	48.9
VSGNet [58] (CVPR 20)	ResNet152	51.8
UnionDet [24] (ECCV 20)	ResNet50-FPN	47.5
DRG [28] (ECCV 20)	ResNet50-FPN	51.0
VCL [59] (ECCV 20)	ResNet50-FPN	48.3
PD-Net [60] (ECCV 20)	ResNet152	52.6
Baseline	ResNet50-FPN	49.4
HRNet	ResNet50-FPN	**53.1**

w.r.t. the ground true larger than 0.5 and the HOI classification result is correct.

B. Comparison With State-of-the-Art Methods

We compare our proposed method with existing methods on HICO-DET [17], V-COCO [16] and HOI-A [18] benchmarks, which is shown in Table II, Table III and Table IV respectively.

1) Results on HICO-DET Dataset: Table II shows the results on the HICO-DET dataset. Previous state-of-the-art methods achieve high performance thanks to the stronger backbone [11], [29], [69], stronger detector [18], [26] and detailed 3D information [28]. Some previous works also leverage pose clues, *e.g.*, concatenating pose information [12], [13], passing features from parts to human and object [15]. Instead, our HRNet outperforms the existing methods benefiting from the hierarchical reasoning module that captures the relations among parts and object, and propagates parts information across different levels via inter-level reasoning. Specifically, comparing with our baseline which achieves 15.47 mAP on the Full classes on the Default setting, our HRNet module can improve its baseline by a sizeable margin of 2.63/3.6/3.76 mAP on the Full/Rare/Non-Rare classes under Default setting. Moreover, we also simply incorporate our light-weight hierarchical reasoning module into ACP [29] to endow the ability of capturing both intra- and inter-relations. We name it by "HRNet^{ACP}". As shown in Table II, it consistently improves ACP by a sizeable margin of 1.34 mAP on Full classes under the Default setting. The result of our HRNet^{ACP} surpasses the current state-of-the-art methods. The results demonstrate the effectiveness of our proposed hierarchical reasoning module, even w.r.t a strong baseline method of ACP. It also demonstrates that our proposed hierarchical reasoning module is general-purpose and extensible for two-stage HOI detection methods.

2) Results on V-COCO Dataset: We also perform experiments on V-COCO dataset, as shown in Table III. Comparing with other state-of-the-art methods, our method achieves a new state-of-the-art result. Specifically, our HRNet achieves 53.1 AP_{role} on V-COCO test set, which also improves our baseline by 3.7 AP_{role}.

3) Results on HOI-A Dataset: To assess the generalization capability of the proposed method, we also conduct experiments on the new HOI-A dataset, as shown in Table IV. Previous state-of-the-art methods [18], [27] achieve high performance thanks to their stronger detectors, such as cascaded detectors [27], [70] and training human-object interaction classifier with detectors [18], [27]. We incorporate our hierarchical

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON HOI-A TEST SET [18]. HRNET^{C-HOI} INDICATES OUR HRNET REPLACE THE BASELINE WITH CASCADE-HOI [27]

Methods	Backbone	mAP
iCAN [57]	ResNet50	44.23
TIN [10]	ResNet50	48.64
Cascade-HOI [25]	ResNet50	66.04
PPDM [16]	Modified-DLA34 [16]	67.03
HRNet^{C-HOI}	ResNet50	**67.26**

reasoning module into Cascade-HOI [27], and name it by "HRNet^{C-HOI}". As shown in Table IV, it consistently improves Cascade-HOI by a margin of 1.22 mAP on HOI-A test set and achieves a new state-of-the-art result on HOI-A dataset.

C. Ablation Studies

We further investigate the effectiveness of each component in our proposed method on V-COCO dataset, which is mainly shown in Table V, Table VI, Table VII and Table VIII.

1) Intra-Level Reasoning: As reported in Table V, our Intra-level edges can acquire 1.7 AP_{role} improvement compared to the baseline (# 1 vs # 4) by encoding the relations of appearance correlations and spatial layouts. Furthermore, we also discuss all kinds of relations in intra-level reasoning to validate the effects of relations. Comparing with # 1, # 2, # 3 and # 4, we observe that the relations of appearance correlations or spatial layouts and both consistently improve the performance by 0.6, 1.3 and 1.7 AP_{role} respectively. Moreover, we also report the result that without the intra-level reasoning (# 7) which decreases AP_{role} by 0.7 (# 7 vs # 8), and note that we still keep the edges between union node and the other parts nodes in # 7. The results demonstrate that the effectiveness of our Intra-level reasoning which captures the relations of both appearance correlations and spatial layouts among human parts and object at each level.

2) Spatial Information: Comparing # 4 with # 5, it can be observed that adding the spatial information brings 0.5 AP_{role} improvement. It shows that encoding the spatial information into graph nodes enhances the representation of each region.

3) Inter-Level Reasoning: We investigate the effectiveness of the inter-level reasoning, which uses adaptive edges to dynamically fuse the information between different levels based on their features. Comparing reasoning within each level separately, it brings 1.1 AP_{role} improvement (#5 vs # 8), which shows that both coarse and fine contextual information can be effectively captured simultaneously through the propagation of information between different levels.

4) Prior of the Hierarchical Body Structure: We validate the effectiveness of the prior knowledge of the hierarchical human body structure. Comparing with fully connecting nodes across levels, the connections based on hierarchical body structure acquire 0.9 AP_{role} improvement (# 6 vs # 8), which demonstrates the benefit of human body structure. It shows that using body structure eliminates the invalid connections and captures the correlative parts information as a context clue.

5) Dynamic Attention: To dynamically focus on the important body parts at each level and suppress other irrelevant parts, the dynamic attention improves the performance by 0.4 AP_{role}. The result demonstrates that dynamic attention helps to tweak the performance (# 8 vs # 9).

6) Number of the Sequential Reasoning Iteration: As described in above, one iteration of sequential reasoning contains the intra-level reasoning and the inter-level reasoning sequentially. We conduct experiments with different iteration number of our proposed hierarchical reasoning. The AP_{role} is 51.6, 53.1, 53.0 when the number of iteration is set to 1, 2 and 3, respectively. We observe the performance slightly drop if the layer number increases further due to over-fitting.

7) Level Number of the Human Parts: We further compare the results of different number and granularity of parts level, which is reported in Table VI. Comparing with the baseline, using the single level of human level, human region level and human keypoint level can acquire 0.7, 0.8 and 1.1 AP_{role} improvement respectively. This suggests that using the local features with intra-level reasoning leads to a better understanding of human-object interaction. Comparing the single part level, using two level of part and pose level brings further performance gain. We further analyze the different granularity of two levels (human level + human region level vs human region level + human keypoint level). The result shows the fine-grained level with a performance boost of 0.9 AP_{role}, since it can provide more fine-grained information to assist the interaction detection.

8) Ways of Extracting Graph Representation: Since we use the node corresponding to the union region to embed the whole graph information at each level, we also compare the way of global average pooling to extract the whole graph information. We conduct the experiment that uses global average pooling to extract the graph representation and then concatenates with $\mathbf{f}_g$. The result is 52.7% which is lower than 53.1% demonstrating that the human-object pair union nodes can reliably represent the human object interaction after graph reasoning.

9) Dimension Number and Hyperparameter μ: We report the results with different numbers of the features dimension on each node in our graph and the value of hyperparameter μ, as shown in Table VII. It has been observed that our HRNet is insensitive to the features dimension. Thus, we empirically fix the dimension number of each node to 256 to be a tradeoff between the performance and memory cost. For the hyperparameter μ used in Eq. 11, we observe that the performance of AP_{role} is overwhelmed when we bias the hyperparameter μ which is the weight of the loss term of binary interaction score. It also shows that the larger number of μ, the more severe the performance degradation, because it is much simpler to determine whether an intersection exists than to identify what kind of intersections are there.

10) Edge Weight: We conduct experiments to study the impact of the edge weight on our graph structure, as shown in Table VIII. Specifically, the edge weight of Eq. 3 is similar with Eq. 6, and they are both used to capture the appearance relations. The Eq. 4 in intra-level reasoning is used to capture the spatial relations within one level. The major difference between them is the activation function and where it is placed.

TABLE V

ABLATION EXPERIMENTS ON V-COCO. IN 'CONNECTION WAY', THE 'FULLY CONNECTION' AND 'BODY STRUCTURE' INDICATE THE FULLY CONNECTING NODES ACROSS LEVELS AND CONNECTING NODES BASED ON THE HIERARCHICAL BODY STRUCTURE IN THE INTER-LEVEL CONNECTIONS, RESPECTIVELY

#	Intra-Level Reasoning			Inter-Level Reasoning		Dynamic Attention	Accuracy AP_{role}
	Relations		Nodes	Connection Way			
	Appearance (Eq. 3)	Spatial (Eq. 4)	Spatial Information	Fully Connection	Body Structure		
1	-	-	-	-	-	-	49.4
2	✓	-	-	-	-	-	50.0
3	-	✓	-	-	-	-	50.7
4	✓	✓	-	-	-	-	51.1
5	✓	✓	✓	-	-	-	51.6
6	✓	✓	✓	✓	-	-	51.8
7	-	-	✓	-	✓	-	52.0
8	✓	✓	✓	-	✓	-	52.7
9	✓	✓	✓	-	✓	✓	53.1

TABLE VI

EVALUATION RESULTS OF OUR HIERARCHICAL REASONING MODULE WHEN INCORPORATING DIFFERENT NUMBER AND GRANULARITY OF PARTS LEVEL ON V-COCO DATASET. HUMAN, REGION AND JOINT INDICATE THE HUMAN LEVEL, HUMAN REGION LEVEL AND HUMAN KEYPOINT LEVEL RESPECTIVELY

Level number	Human parts	mAP_{role}
Baseline (Global module only)		49.4
1	Human level	50.1
	Region level	50.2
	Region level	50.2
2	Human + Region level	51.9
	Region + Keypoint level	52.8
3	Human + Region + Keypoint level	53.1

TABLE VII

IMPACTS OF THE DIMENSION AND HYPER-PARAMETER μ

Hyperparameter μ	Features dimension number of each node	Accuracy AP_{role}
0.1	128	53.0
0.1	256	53.1
0.1	512	53.2
0.5	256	52.8
1.0	256	52.7

To study the impact, we conduct experiments via exchanging these two edge weights. We can observe that the results #1-#3 in Table VIII are comparable, which demonstrates that our method is insensitive with the particular architecture of edge weight for capturing appearance relations.

11) Sequential Reasoning: We conduct experiments to validate the effectiveness of the sequential reasoning. As shown in Table VIII. Comparing with #4 and #1, we observe a performance drop when the graph performs a simultaneous reasoning, where we perform the intra-level and inter-level message propagation at the same time. Thus, we employ the sequential reasoning in all our experiments except as otherwise noted.

12) Computational Complexity: Comparing with the 23M parameters of backbone (ResNet50), the parameters of our hierarchical reasoning module is 1.84 M only, which shows that our proposed module is light-weight. The average runtime of our proposed hierarchical reasoning module for a pair of human-object features is 54.5 ms. And the average runtime for extracting part and keypoint features used in our method is 1014 ms. We evaluate the runtime on one GTX1080Ti and measure it by calculating the mean runtime over 5k iterations. It shows that the runtime of our method can be negligible.

TABLE VIII

IMPACTS OF REASONING WAY AND EDGE WEIGHT IN OUR HRNET ON THE V-COCO DATASET

#	Reasoning way	Edge weight		AP_{role}
		intra-level	inter-level	
1	Sequential	Eq. 3& Eq. 4	Eq. 6	53.1
2	Sequential	Eq. 6& Eq. 4	Eq. 6	53.0
3	Sequential	Eq. 3& Eq. 4	Eq. 3	53.1
4	Simultaneous	Eq. 3& Eq. 4	Eq. 6	52.7

D. Fine-Grained Classification

In addition to human-object interaction detection, we further conduct experiments on fine-grained image classification to assess the generalization capability of HRNet. Since fine-grained classification faces the challenging of the inherently subtle difference between different categories, it is required to consider the local discriminative parts and more challenging than general image classification. Thus, we validate the generalized ability of HRNet to capture relations among local regions in the task of fine-grained classification.

1) Datasets: We conduct experiments on the challenging CUB dataset [71], which consists of 11,788 images of 200 bird species with detail annotations of parts and bounding boxes.

2) Implementation Details: In the experiments, we adopt our hierarchical reasoning module with the baselines, B-CNN [58] and iSQRT-COV-Net [72], following their implementation setting for fair comparison. Specifically, we build a hierarchical reasoning module containing the entire object and their part region level as two level, and then concatenates the output with the features from baseline as the final features.

Table IX shows the results on the CUB dataset. The results demonstrate that the proposed hierarchical reasoning module can improve B-CNN [58] and iSQRT-COV-Net [72] by 0.6% and 0.8% respectively. It shows that the generalization ability of the proposed hierarchical reasoning module which can capture and utilize the relations among different levels on fine-grained classification.

TABLE IX

THE RESULTS ON THE CUB-200-2011 DATASET. '+ HRNET' INDICATES THAT INCORPORATING OUR PROPOSED HIERARCHICAL REASONING MODULE INTO THE SPECIFIC METHOD

Method	Backbone	Accuracy (%)
CBP [48]	ResNet50	81.6
RA-CNN [42]	VGG19	85.3
SMSO [47]	ResNet50	85.8
MA-CNN [43]	VGG19	86.5
MPN-COV-Net [46]	VGG16	86.7
KERL [51] w/bbox	VGG16	86.8
DFL [44]	ResNet50	87.4
HSE [50]	ResNet50	88.1
B-CNN (250k-dims) [49] w/bbox	VGG16	85.1
B-CNN (250k-dims) [49] w/bbox + HRNet	VGG16	85.7
iSQRT-COV-Net (8k) [63]	ResNet50	87.3
iSQRT-COV-Net (8k) [63] + HRNet	ResNet50	**88.1**

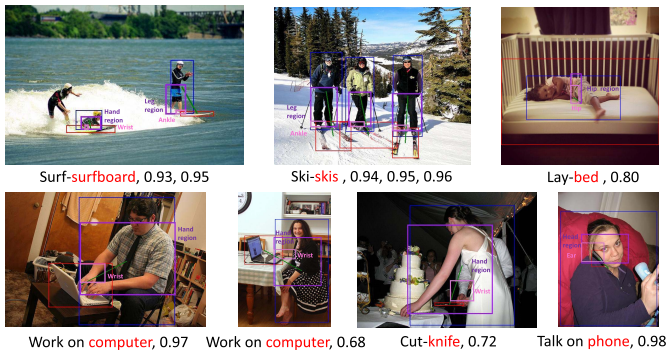

Fig. 5. Visualization results on the V-COCO test set. The number denotes the score of the interaction from left to right on the image predicted by our HRNet. The pink boxes indicate the most relevant body parts at the human region level and keypoint level.

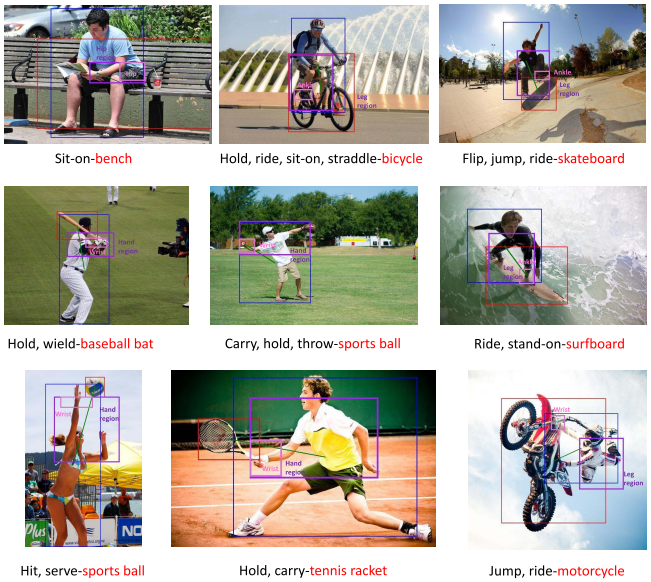

Fig. 6. Visualization results on the HICO-DET test set. The pink boxes indicate the most relevant body parts at the human region level and keypoint level.

E. Qualitative Results

Figure 5 and Figure 6 visualize the HOIs detection results on the V-COCO and HICO-Det test set respectively. We present the visual results of one interaction, its score and its most relevant human part at the human region and keypoint level. These relevant regions are obtain by the largest edges outgoings to the union node. Visualization results show that our HRNet learns to focus on the relative human region and keypoints (*e.g.*, the keypoint of wrist in the hand region) for HOI detection based on the object and the human pose.

V. CONCLUSION

In this work, we propose a novel Hierarchical Reasoning Network (HRNet) for human-object interaction detection. We emphasize the importance of hierarchical multi-scale human parts and propose to build the multi-level graphs to represent parts at multi-level via their appearance and spatial information. Furthermore, we also propose the intra-level reasoning to model relations within the specific level based on their appearance correlation and spatial layout, and inter-level reasoning to dynamically propagate the information among parts at different levels with the prior of body structure. Our proposed lightweight HRNet also have extensible ability, which can be easily incorporated into the two-stage HOI detection models (*e.g.*, our baseline, ACP [29] and Cascade-HOI [27]) and the common fine-grained classification models (*e.g.*, B-CNN [58] and iSQRT-COV-Net [72]). Experimental results show that our HRNet can not only achieve state-of-the-art performance on V-COCO, HICO-DET and HOI-A benchmarks, but also boost the performance of the fine-grained methods.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[3] G. Li, Y. Gan, H. Wu, N. Xiao, and L. Lin, "Cross-modal attentional context learning for RGB-D object detection," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1591–1601, Apr. 2019.
[4] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: http://arxiv.org/abs/1812.08008
[5] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10863–10872.
[6] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*. [Online]. Available: http://arxiv.org/abs/1902.09212
[7] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6904–6913.
[8] S. Yang, G. Li, and Y. Yu, "Relationship-embedded representation learning for grounding referring expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2765–2779, Aug. 2021.
[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
[10] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.
[11] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9677–9685.

[12] Y.-L. Li *et al.*, "Transferable interactiveness knowledge for human-object interaction detection," 2018, *arXiv:1811.08264*. [Online]. Available: http://arxiv.org/abs/1811.08264

[13] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9469–9478.

[14] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 51–67.

[15] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 843–851.

[16] S. Gupta and J. Malik, "Visual semantic role labeling," 2015, *arXiv:1505.04474*. [Online]. Available: http://arxiv.org/abs/1505.04474

[17] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 381–389.

[18] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "PPDM: Parallel point detection and matching for real-time human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 482–490.

[19] R. Krishna *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[20] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 852–869.

[21] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5532–5540.

[22] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Detecting unseen visual relations using analogies," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1981–1990.

[23] G. Yin *et al.*, "Zoom-Net: Mining deep feature interactions for visual relationship recognition," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 322–338.

[24] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 670–685.

[25] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8359–8367.

[26] B. Kim, T. Choi, J. Kang, and H. J. Kim, "UnionDet: Union-level detector towards real-time human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 498–514.

[27] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4263–4272.

[28] Y.-L. Li *et al.*, "Detailed 2D-3D joint representation for human-object interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10166–10175.

[29] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 718–736.

[30] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "DRG: Dual relation graph for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 696–712.

[31] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 401–417.

[32] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*. [Online]. Available: http://arxiv.org/abs/1609.02907

[33] F. Hu, Y. Zhu, S. Wu, L. Wang, and T. Tan, "Hierarchical graph convolutional networks for semi-supervised node classification," 2019, *arXiv:1902.06667*. [Online]. Available: http://arxiv.org/abs/1902.06667

[34] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 4800–4810.

[35] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7239–7248.

[36] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang, and L. Lin, "Graphonomy: Universal human parsing via graph transfer learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7450–7459.

[37] L. Lin, Y. Gao, K. Gong, M. Wang, and X. Liang, "Graphonomy: Universal image parsing via graph reasoning and transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 8, 2020, doi: 10.1109/TPAMI.2020.3043268.

[38] S. Yang, G. Li, and Y. Yu, "Dynamic graph attention for referring expression comprehension," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4644–4653.

[39] S. Yang, G. Li, and Y. Yu, "Graph-structured referring expression reasoning in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9952–9961.

[40] S. Yang, M. Xia, G. Li, H.-Y. Zhou, and Y. Yu, "Bottom-up shift and reasoning for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11266–11275.

[41] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6700–6709.

[42] Y. Gao *et al.*, "Fashion retrieval via graph reasoning networks on a similarity pyramid," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 18, 2020, doi: 10.1109/TPAMI.2020.3025062.

[43] Z. Kuang *et al.*, "Fashion retrieval via graph reasoning networks on a similarity pyramid," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3066–3075.

[44] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 399–417.

[45] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, and S. Avidan, "Graph embedded pose clustering for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10539–10547.

[46] F. Baradel, N. Neverova, C. Wolf, J. Mille, and G. Mori, "Object level visual reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 105–121.

[47] R. Herzig *et al.*, "Spatio-temporal action graph networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[48] J. Materzynska, T. Xiao, R. Herzig, H. Xu, X. Wang, and T. Darrell, "Something-else: Compositional action recognition with spatial-temporal interaction networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1049–1059.

[49] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1173–1182.

[50] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 834–849.

[51] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4438–4446.

[52] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5209–5217.

[53] Y. Wang, V. I. Morariu, and L. S. Davis, "Learning a discriminative filter bank within a CNN for fine-grained recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.

[54] T. Wang *et al.*, "Deep contextual attention for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5694–5702.

[55] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2070–2078.

[56] K. Yu and M. Salzmann, "Statistically-motivated second-order pooling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 600–616.

[57] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 317–326.

[58] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1449–1457.

[59] T. Chen, W. Wu, Y. Gao, L. Dong, X. Luo, and L. Lin, "Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 2023–2031.

[60] T. Chen, L. Lin, R. Chen, Y. Wu, and X. Luo, "Knowledge-embedded representation learning for fine-grained image recognition," 2018, *arXiv:1807.00505*. [Online]. Available: http://arxiv.org/abs/1807.00505

[61] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*. [Online]. Available: http://arxiv.org/abs/1710.10903
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[63] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[64] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[65] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.
[66] C. Gao, Y. Zou, and J.-B. Huang, "ICAN: Instance-centric attention network for human-object interaction detection," 2018, *arXiv:1808.10437*. [Online]. Available: http://arxiv.org/abs/1808.10437
[67] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13617–13626.
[68] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 584–600.
[69] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis*, 2020, pp. 69–85.
[70] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.
[71] P. Welinder *et al.*, "Caltech-UCSD birds 200," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2010-001, 2010.
[72] Q. Wang, J. Xie, W. Zuo, L. Zhang, and P. Li, "Deep CNNs meet global covariance pooling: Better representation and generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2582–2597, Aug. 2021.

Yiming Gao received the B.S. degree from the School of Mathematics, South China University of Technology, China. He is currently pursuing the master's degree with the School of Data and Computer Science, Sun Yat-sen University. His current research interests include computer vision and machine learning.

Zhanghui Kuang (Member, IEEE) received the B.S. degree from Sun Yat-sen University, Guangzhou, China, in 2009, and the Ph.D. degree from The University of Hong Kong in 2014. He is currently a Research Director with SenseTime Group Ltd. He has authored and coauthored more than ten papers in top-tier academic journals and conferences. His research interests include deep learning and computer vision.

Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University. He has authored and coauthored more than 70 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning. He was a recipient of ICCV 2019 Best Paper Nomination Award. He also serves as an Area Chair for the conference of VISAPP. He has been serving as a Reviewer for numerous academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *IJCV*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, ICCV, ECCV, and NeurIPS.

Wayne Zhang received the B.Eng. degree in electronic engineering from Tsinghua University, Beijing, China, in 2007, and the M.Phil. and Ph.D. degrees in information engineering from The Chinese University of Hong Kong in 2009 and 2012, respectively. He is currently a Senior Research Director with SenseTime Group Ltd. He also serves as an EXCO Member for AI Specialist Group, Hong Kong Computer Society. His research interests include deep learning and computer vision.

Liang Lin (Senior Member, IEEE) is currently a Full Professor with Sun Yat-sen University and the CEO of Dark Matter AI. He worked as the Executive Director of SenseTime Group from 2016 to 2018, leading the research and development teams in developing cutting-edge, deliverable solutions in computer vision, data analysis and mining, and intelligent robotic systems. He has authored or coauthored more than 200 papers in leading academic journals and conferences. He is a fellow of IET. He was a recipient of the Annual Best Paper Award by *Pattern Recognition* (Elsevier) in 2018, the Diamond Award for Best Paper in IEEE ICME in 2017, the Award for the Best Student Paper in IEEE ICME in 2014, Hong Kong Scholars Award in 2014, Google Faculty Award in 2012, and ACM NPAR Best Paper Runners-Up Award in 2010. He is also an Associate Editor of IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and *IET Computer Vision*, and served as the Area/Session Chair for numerous conferences, such as CVPR, ICME, and ICCV.