计算机科学
COMPUTER SCIENCE

# 融合多层次视觉信息的人物交互动作识别

李宝珍, 张晋, 王宝录, 余平

引用本文

李宝珍, 张晋, 王宝录, 余平.融合多层次视觉信息的人物交互动作识别[J]. 计算机科学, 2022, 49(11A): 220700012-8.

LI Bao-zhen, ZHANG Jin, WANG Bao-lu, YU Ping. Human-Object Interaction Recognition Integrating Multi-level Visual Features [J]. Computer Science, 2022, 49(11A): 220700012-8.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700012

# 融合多层次视觉信息的人物交互动作识别

李宝珍[1]　张　晋[1]　王宝录[1]　余　平[2]
1 国家能源集团神东锦界煤矿　陕西 神木 719319
2 国能网信科技(北京)有限公司　北京 100011
(10015446@chnenergy.com.cn)

**摘　要**　基于计算机视觉的人体动作识别技术在视频监控、智能驾驶、人机交互、多媒体内容审核等领域均有着广阔的应用前景,其中人体动作中的人物交互是动作识别的核心内容之一。现有的人物交互动作识别模型对人物关系的提取仅仅停留在表层视觉特征之上,并未充分挖掘人体关键区域以及人物之间的深层语义关系。针对此问题,文中提出了层次化的图神经网络模型(HGNN)对人物交互动作建模。HGNN模型从局部到整体显式地对人体关键区域以及人和物构成的场景图进行建模,并利用注意力图池化机制(AttPool)剔除层次图中冗余的信息和噪声,再通过图卷积网络提取图结点之间的深层语义关系,对卷积网络提取的特征进行聚合与优化,从而得到反映人物交互动作本质的特征表示。另外,HGNN模型在中层图进行的临时监督分类也能够约束网络更好地学习到交互动作的人体模式,避免网络对交互对象产生"偏见"。最后,针对HGNN模型,设计了多任务损失函数,用于有效进行模型训练。为了验证HGNN模型的有效性,在公开的大型数据集V-COCO上进行了广泛的实验,结果均显示所提出的HGNN模型对常见的人物交互动作具有广泛的适应性和鲁棒性,精度(mAP)超过了现有的基于图神经网络的模型,同时领先于大部分最新的多流卷积模型。

**关键词**:计算机视觉;人体动作识别;人物交互;深度学习;图神经网络
**中图法分类号**　TP391

## Human-Object Interaction Recognition Integrating Multi-level Visual Features

LI Bao-zhen[1],ZHANG Jin[1],WANG Bao-lu[1] and YU Ping[2]
1 Shendong Jinjie Colliery,Chn Energy,Shenmu,Shaanxi 719319,China
2 Chn Energy Network Infomation Technology(Beijing) CO., LTD.,Beijing 100011,China

**Abstract**　Computer vision based human action recognition technique has a broad application in the fields of video surveillance,intelligent driving,human-computer interaction,multimedia content audit,etc. More importantly,human-object interaction is one of the core components in human action recognition. Most of the existing human-object interaction action recognition models,which are based on multi-stream convolutional neural networks,only capturing the visual features superficially. They fail to fully explore the key areas of human body and the deep semantic relationship between human and objects. To solve this problem,this paper proposes a hierarchical graph neural network(HGNN) model. HGNN explicitly models the critical areas of the human body and the interaction of human-object in the scene from local to global,and uses an attention pooling mechanism(AttPool) to eliminate redundant information and noise in the graph. Then,the deep semantic relationship between graph nodes are captured by the graph convolution network,and the initial features extracted by convolutional neural network are aggregated and optimized. In this way,the feature representation which reflects the essential character of human-object interaction can be obtained. In addition,the interim supervised classification in the middle graph can also constrain the model to better learn the human patterns of interactive actions,and avoid the model to produce "bias" on the interactive objects. Finally,a multi-task loss function is designed for the HGNN to effectively train the model. To test and verify the effectiveness of the proposed HGNN model,extensive experimental evaluations on the famous public benchmark V-COCO have been conducted. The results show that the proposed HGNN model is adaptive and robust for human-object interaction detection,which outperforms the previous graph neural network based methods by a large margin,and also performs better than most of the latest convolutional neural network based models.

**Keywords**　Computer vision,Human action recognition,Human-Object interaction,Deeplearning,Graph neural network

## 1　引言

人体动作识别是计算机视觉的一大研究方向,其与简单的视觉理解不同,人体动作识别往往包含着对图像或视频内容更深层次的理解。例如人体动作通常包含人与人、人与物以及人与环境之间的交互,同时人体动作也包含着姿态、手势、眼神等细粒度语义信息。近年来,国内外学者也围绕着人物交互动作识别设计出了多种有效的模型。以Chao等提出的多流卷积神经网络模型HO-RCNN[1]为例,其在目标检测网络R-CNN的基础上,通过堆叠的卷积层提取图像视觉

通信作者:余平(20049948@ceic.com)

特征，并设计了多流结构分别用于提取人体、物体和交互关系3种语义特征。多流卷积神经网络可以有效地提取图像的表层视觉特征，从而对图像中的人物交互动作信息作出较为准确的识别。但传统卷积神经网络对图像内容的理解往往停留在表观视觉特征，并没有显式地对人物交互关系进行建模。因此，本文的创新点之一便是引入层次化的图结构显式地对人物关系进行建模，并利用图神经网络在图结构上提取深层语义特征。在现有研究中，Qi等提出了GPNN模型，首次将图神经网络用于人物交互动作检测问题[2]，其彻底抛弃了原有的多流卷积结构，直接对图像中的所有实体构建不规则图结构。尽管GPNN模型在人物交互关系建模上更符合常识，但过于复杂的场景图使其难以训练，也影响了模型的精度表现。目前，国内也有很多文献聚焦于人体动作识别研究[3-6]，这些模型都充分挖掘了卷积神经网络的潜力，但针对人物交互动作检测这个特定的问题却较少提及，也尚未出现利用图神经网络对人物交互关系进行建模的相关研究。基于现有研究，本文提出了层次化图神经网络模型(HGNN)，其结合了多流卷积结构和图结构的优势，从局部到整体逐步对单个人-物对进行建模，避免了规模过大的场景图构建，同时有效提升了模型精度。另外，本文针对HGNN模型设计了多任务损失函数，以有效进行模型训练。

## 2 相关技术

### 2.1 图神经网络

在现实世界中，更多的数据都处于不规则的非欧氏空间中，即图结构(graph)数据，例如化学分子结构[7]、交通网络[8]、社交网络[9]以及图像中的实体交互关系[10-11]。因此，如何在非欧氏空间中发挥出深度模型的信息提取能力，是近年来深度学习的研究热点。2009年，Scarselli等提出了可以在图结构数据上学习的神经网络模型，首次提出了图神经网络(GNN)的概念，其是一种在图结构上按照一定规则循环聚合结点信息的可训练的模型[12]。2016年，Defferrard等利用图空间的傅里叶变换将卷积操作拓展到了非欧氏空间中，提出了图卷积模型(GCN)[13]，随后Kipf等提出了利用图卷积进行半监督学习的深度模型[14]。但以上图卷积方法都是在光谱域(spectral-domain)对图结构进行操作，因此计算效率通常很低。2017年，Monti等对光谱域图卷积进行进一步简化，将其定义在了空间域(spatial-domain)，直接对图结点及其相邻结点进行操作，免去了傅里叶变换和切比雪夫多项式近似等复杂步骤[15]。目前，图神经网络和卷积神经网络相结合的深度模型被广泛应用于处理计算机视觉相关任务，尤其是具备非欧氏空间数据结构的人体骨架动作识别[16]、视觉问答[17]和实体关系识别[10]。

定义图$G=(V,E)$，其中$V$为结点，$E$为边，则一层GNN操作可以表示为：

$$G_{l+1}=Update(Aggregate(G_l,W_l^{agg}),W_l^{update}) \tag{1}$$

其中，$G_l$和$G_{l+1}$分别表示第$l$层GNN输入和输出的图。$Update(\cdot)$和$Aggregate(\cdot)$分别表示GNN的聚合和更新函数，$W_l^{agg}$和$W_l^{update}$是可训练的参数。式(1)是GNN的一般形式，现有研究大多围绕聚合和更新函数展开。本文也在一般形式的基础上，设计了更符合人物交互动作检测的聚合函数和更新函数。

类比于CNN中的池化(pooling)操作，近年来也出现了部分研究聚焦于图结构上的池化。早期的工作通常使用简单的全局池化直接将图结点的个数降至1[18]，或是利用固定的规则对大量图结点聚类，再针对每一类进行全局池化[19-20]。这些方法在减少图结点个数的同时，并未考虑结点的重要性，因此会造成重要信息的丢失，也容易受到噪声的影响。近期，一些研究提出了可学习的图池化模块[21-24]。作为一种数据驱动的图池化方式，这类方法可以更好地适应训练数据的分布，从而在减少结点数量的同时，充分考虑结点的重要程度。针对人物交互动作检测问题，本文也提出了一种基于注意力机制的图池化模块。与现有方法不同，本文并非采用自注意力模型(self-attention)，而是通过例如空间结构等其他信息生成结点的注意力映射，因此更加适应于动作识别。

### 2.2 人物交互动作识别

尽管在深度卷积神经网络的帮助下，图像表层语义特征的提取与识别已经达到较高水准，但如何识别图像更深层次的语义内涵，仍旧是一项具有挑战性的任务。人物交互动作便是最重要的图像深层语义之一。Gupta等学者最早提出人物交互动作识别问题，并建立贝叶斯模型在人工提取的图像特征上解决该问题[25]。Saurabh等于2015年发布了该领域重要的大型数据集之一V-COCO，并改进了R-CNN模型用于检测人物交互[26]。同年，Chao等也发布了另一个重要的大型数据集HICO，同时指出深度语义信息在人物交互动作识别中的重要性[27]。后续其又发布了HICO-DET作为HICO数据集的拓展[5]。在大规模数据集的支持下，基于深度网络的人物交互动作识别模型也快速发展。Chao等提出了多流卷积神经网络模型HO-RCNN，将人体、物体和相互关系的语义特征放在3个卷积流中提取，最后再将三者融合，从而将算法精度和泛化能力提升到了新的高度[1]。后续有不少工作在多流卷积神经网络的基础上进行改进，例如基于人体特征预测交互物体目标区域的InteractNe[28]，加入了场景全局上下文特征的iCAN[29]，加入了可跨数据集学习的人物交互性评价模型Transferable Interactiveness Network(TIN)[30]和加入了人体姿态细节以及注意力机制的PMFNet[31]。上述方法[25,31]均利用深度卷积神经网络在图像层面提取视觉语义特征，然后进行动作分类，但他们都忽略了人物交互视觉场景中人和物以及人体各关键部位和物之间的深层语义关系。因此，Qi等于2018年提出了用图神经网络对人物交互关系进行建模的新方法GPNN[2]。GPNN将场景中的人和物实体都视为图结构中的结点，以图的边表示结点之间的交互关系，分别用连接(link)、消息(message)、更新(update)和输出(readout)函数定义图神经网络的信息传播规则。结果表明，图神经网络可以很好地对深层人物交互关系建模，从而帮助计算机理解交互场景。之后，Zhou等将图神经网络和多流网络结构进行进一步融合，提出了RPNN，充分发挥了二者在处理欧氏图像信息和非欧氏关系信息上的性能[32]。Xu等又将外部知识图(Knowledge Graph)引入模型，以解决现有的大型数据集中存在的长尾分布问题[33]。综上所述，人物交互动作识别模型经历了从传统方法到深度学习方法的演变，近期又由深度卷积神经网络模型转向了图神经网络和卷积神经网络的融合模型。同时，更多细节且高层次的语义信息被合理利用，也提升了模型识别的精度。

## 3 层次化图神经网络模型

### 3.1 总体结构

给定一张 RGB 图像 $I \in \mathbb{R}^{C\times W\times H}$，人物交互动作检测的目标是检测出图像上所有的人体和物体以及其之间的交互关系。本文模型主要包括以下 5 个模块：1）基于 Mask R-CNN 的目标检测模块；2）基于 ResNet 的表观视觉特征提取模块；3）基于层次化 GNN 的深层语义特征提取模块；4）基于注意力机制的图池化模块；5）多任务特征分类模块。其中，模块 4）是模块 3）的子模块，二者共同构成层次化图神经网络。各模块间的关系如图 1 所示。


图 1 模型的总体结构

Fig. 1 Overall structure of our model

图像 $I$ 首先经过 Mask R-CNN[34] 检测出其中的人体区域推荐框（Proposals）和检测得分（Detection Scores）$\{\langle x_h, S_h\rangle\}$，以及物体区域推荐框、物体类别和检测得分 $\{\langle x_o, c_o, S_o\rangle\}$。之后利用感兴趣区域匹配池化操作（ROI Align），在 ResNet[15] 提取的图像特征上得到人体，物体和合并区域相应的特征 $F_h$，$F_o$ 和 $F_u$，池化后特征长宽为 $R_b$。同时，利用 Mask R-CNN 也可以得到人体的关键点位置，其包括了 17 个关键点的坐标 $\{p_h^k \mid k\in\{1,2,\cdots,17\}\}$，分别表示人体的鼻子、眼睛、左右手、左右脚等关键部位。在关键点坐标的基础上，定义关键区域 $x_{pk}\in\mathbb{R}^4$。其是以点 $p_h^k$ 为中心，长宽分别为人体边界框 $x_h$ 的 $\gamma$ 倍的边界框。之后同样利用感兴趣区域匹配池化操作，在图像特征上提取人体关键区域的特征 $\{F_{pk} \mid k\in\{1,2,\cdots,17\}\}$，池化后特征长宽为 $R_p$。利用人体关键点、人体和物体的边界框构造出人物空间结构图 $I'\in\mathbb{R}^{3\times W'\times H'}$[30]（Spatial Configuration，见图 2）。之后利用全连接操作提取空间结构图的特征向量 $F_s$。至此，图像 $I$ 中所有的表观视觉特征提取完成，其包括了人体特征 $F_h$、物体特征 $F_o$、合并区域特征 $F_u$、人体关键区域特征 $\{F_{pk} \mid k\in\{1,2,\cdots,17\}\}$ 以及空间结构特征 $F_s$。得到表观视觉特征之后，将人体和物体两两组合构成相应的人物对，按照一定的顺序，将上述特征分别通过全连接层优化后输入层次化图神经网络之中，利用图神经网络对特征进行进一步融合与优化，从而得到图像深层的语义特征。最后利用多任务特征分类模块，对深层次语义特征进行分类，从而得到该人物对最终的人物交互动作检测结果。对图像 $I$ 中所有的人物对执行上述检测过程，即得到图像中包含的所有 $\{x_h, x_o, c_o, a_{h,o}\}$ 元组。其中 $x_h\in\mathbb{R}^4$ 为人体边界框，$x_o\in\mathbb{R}^4$ 为物体边界框，$c_o\in\{1,\cdots,C\}$ 为物体类别，$a_{h,o}\in\{1,\cdots,A\}$ 为人物交互动作类别。下面详细介绍各个模块的内部结构与原理。

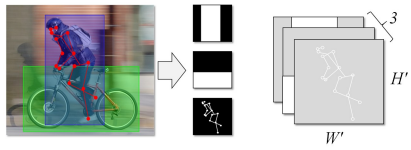


图 2 人物空间结构图的构造方式

Fig. 2 Construction of figure space structure diagram

### 3.2 基于 Mask R-CNN 的目标检测模块

本文在传统模型的基础上，引入了人体关键点信息。如何高效地完成目标检测和人体关键点检测是本文的首要任务。Mask R-CNN 是目前主流的二阶目标检测模型之一，其在共享特征图的基础上可以同时用于实体分割和关键点检测等多种任务中。因此，本文利用 Mask R-CNN[34] 进行前期的目标检测与人体关键点检测。

利用 Mask R-CNN 可以得到图像中所有的人体边界框、物体边界框、物体类别以及人体关键点。需要说明的是，在提取相关实体的表观视觉特征时，本文并没有重复利用 Mask R-CNN 的中间特征，而是重新构造 ResNet 网络在原始图像上提取新的特征，以此来保证特征对于人物交互动作识别任务的适应性，从而提高模型精度。

### 3.3 基于 ResNet 的表观视觉特征提取模块

利用 Mask R-CNN 得到了图像中目标的边界框之后，本文采用主流的 ResNet50 提取图像的表观视觉特征。通过 ResNet50 可以得到通道数为 512 的特征映射，本文在此基础上添加一个卷积层，使其输出通道数为 256，该卷积层的卷积核尺寸为 1×1，卷积步长为 1。为了增强人与物之间的相对位置特征，针对每一个人物对，本文在 256 通道的特征映射之后拼接 2 个通道的相对位置掩膜，其每个位置的特征值分别表示相对于物体中心位置的长宽坐标偏移量。至此，针对一张图片 $I$，我们得到了其目标边界框和通道数为 258 的特征映射。之后，利用感兴趣区域匹配池化操作（ROI Align），可以得到每个目标的特征映射，其包括了人体特征 $F_h$、物体特征 $F_o$、合并区域特征 $F_u$ 和所有的人体关键区域特征 $\{F_{pk} \mid k\in\{1,2,\cdots,17\}\}$。另外，需要特别说明的是，空间结构特征 $F_s$ 并不是直接在图像中获取的，而是通过构造人物空间结构图（见图 3），再经过 2 层全连接层提取得到。

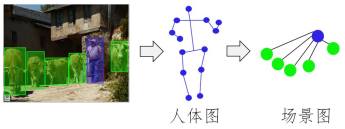


图 3 “农夫放牛”场景解析

Fig. 3 Analysis of “farmer herding cattle” scenario

### 3.4 基于层次化 GNN 的深层语义特征提取模块

现有基于多流卷积网络的人物交互动作识别模型[1] 在上述表观视觉特征的基础上直接对交互动作进行分类。尽管在加入了注意力机制[28]、上下文信息[29] 等技巧之后，模型精度有所提升，但其始终停留在浅层视觉特征的运用，没有挖掘实体之间的语义关系。深度卷积神经网络通过像素级的运算在图像表观视觉特征的提取上可以取得不错的结果，但无论如何扩大卷积层的感受野，增加模型深度，其对图像的理解也只停留在规则像素组成的欧氏空间中，缺乏对非欧氏空间信息的处理能力。在现实场景中，人体各部位之间以及人和物体

之间广泛存在的语义关系呈现出了不规则且有层次的图结构。以“农夫放牛”这个简单的场景为例(见图 3),农夫自身身体各部位之间以人的物理形态为关联,组成了“人体图”结构。农夫与牛之间以“放牛”这个动作为关联,组成了“场景图”结构。牛和牛之间也存在着空间上的位置关系。这些不规则图结构的语义关系是在表观视觉特征之上更深层的图像内涵。如何学习这些语义关系,是人物交互动作识别任务所面对的关键问题之一,也是本文的出发点。因此,本文通过显式地对图像中层次化的图结构关系建模,并提出层次化图神经网络来提取实体间的语义关系特征,以弥补卷积网络在这方面的缺陷,从而让模型学习到更好的人物交互动作特征表示。需要指出的是,GPNN 网络[2]和 RPNN 网络[32]也利用图结构对实体关系建模,但二者并没有层次化的结构,且限制了图的邻接矩阵,未能广泛表示实体关系。本文模型与两者有着很大区别,且模型精度也大幅领先。下面详细介绍基于层次化 GNN 的深层语义特征提取模块。

基于层次化 GNN 的深层语义特征提取模块的整体结构如图 4 所示,为了简单起见,后文称之为 HGNN(Hierarchical Graph Neural Network)。类比于 CNN 网络,浅层卷积在像素层面提取局部信息,深层卷积由于感受野的扩大,在更大范围内提取全局信息,这种从局部到整体的逻辑结构也符合多层次的场景图。因此,HGNN 首先利用人体关键区域特征 $F_{pk}$ 构建局部图结构(图 4 中 local graph)。之后利用图卷积操作 GCN(Graph Convolutional Network)在局部图上进行特征优化。然后利用图注意力池化(Graph Attention Pooling)操作减小局部图结点个数,之后加入空间结构特征 $F_s$ 和人体特征 $F_h$ 构造中层图结构(图 4 中 middle graph)。用 GCN 操作在中层图结构上进行特征优化,之后加入物体特征 $F_o$ 与合并区域特征 $F_u$,构造全局图结构,并用图注意力池化策略进一步减少结点个数(图 4 中 global graph)。最后将图结点特征与初始特征合并起来构造全局特征 $F_{global}$ 进行特征分类并计算损失函数(Loss Function)。另外,在中层图阶段,我们也将结点特征合并起来进行临时的特征分类,并参与 Loss 计算。这个临时的特征分类结果不仅可以提升模型对于数据长尾分布的适应性,还可以避免模型在交互动作识别过程中对交互对象产生依赖,同时参与图池化模块的注意力计算。整个过程中,HGNN 模型通过注意力图池化不断减少人体关键点结点个数并加入新的更大尺度的结点,本文构造出了由局部到整体三级层次化的图结构,并利用三层 GCN 进行特征优化,从而得到更高级的人物关系特征表示。下面,本文针对局部图、中层图以及全局图上的具体构造和图卷积操作细节进行逐一介绍。

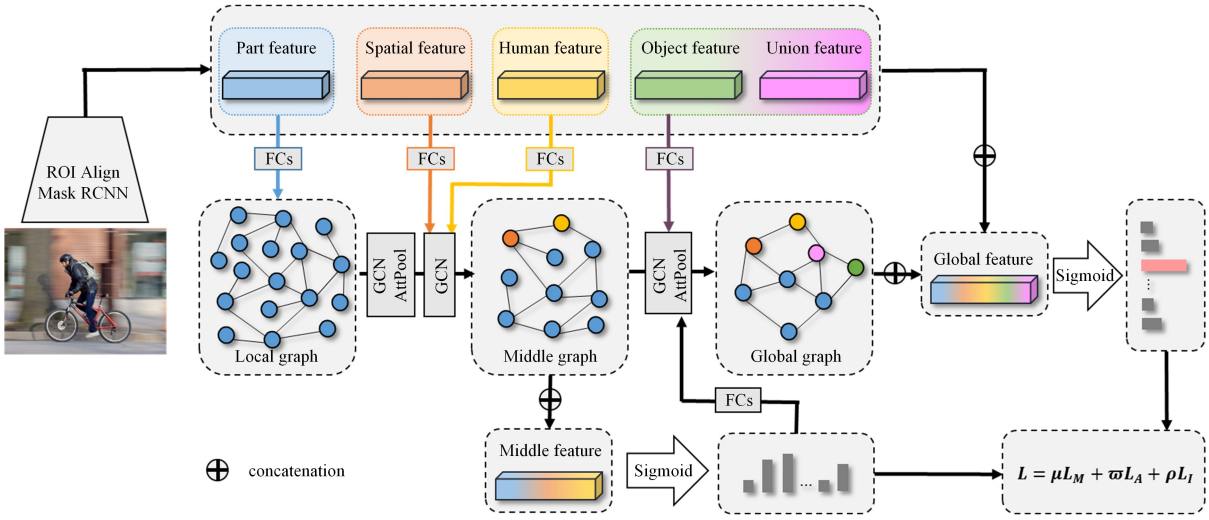


图 4 基于层次化 GNN 的深层语义特征提取模块结构示意图

Fig. 4 Structure of deep semantic feature extraction module based on hierarchical GNN

### 3.4.1 局部图、中层图及全局图构造

定义局部图 $\mathcal{G}_l=(V_l,\mathcal{E}_l)$,其中 $V_l$ 为人体关键点结点,$\mathcal{E}_l$ 为边。Mask R-CNN 提取的人体关键点数量为 17,因此局部图包含 17 个结点,结点 $v_{lk}\in V_l$ 的初始特征为全连接投影之后相应的人体关键区域特征 $FC(\{F_{pk}\})$,$FC(\cdot)$ 表示全连接操作(包括激活函数)。边 $\mathcal{E}_l$ 可用邻接矩阵 $\boldsymbol{A}_l$ 表示,现有研究通常利用结点之间的直观关系人工指定邻接矩阵 $\boldsymbol{A}_l$[32],但这种方式并不利于图上的信息传播,例如人体的左右手在物理结构上并不相连,但在很多交互动作中,左右手之间的关系是其显著特征。因此,本文使用自学习的方式确定更加泛化的邻接矩阵 $\boldsymbol{A}_l$:

$$\boldsymbol{A}_{l,ij}=\sigma(\mathcal{F}_l(\boldsymbol{F}_{li}\oplus\boldsymbol{F}_{lj}))+\boldsymbol{A}_{c,ij} \tag{2}$$

其中,$\boldsymbol{A}_{l,ij}$ 表示 $\boldsymbol{A}_l$ 在 $(i,j)$ 位置处的值;$\boldsymbol{A}_c$ 表示人体关键点之间的物理结构邻接矩阵;$\sigma(\cdot)$ 表示 sigmoid 激活函数;$\mathcal{F}_l(\cdot)$ 表示两个全连接层,中间添加 Relu 激活函数;$\oplus$ 表示向量拼接;$\boldsymbol{F}_{li}$ 表示 $\mathcal{G}_l$ 上结点 $i$ 的特征向量。

定义中层图 $\mathcal{G}_m=(V_m,\mathcal{E}_m)$,其中 $V_m$ 包含池化后的人体关键区域结点以及空间结构特征结点和人体特征结点,$\mathcal{E}_m$ 为边。空间结构特征结点的初始特征为 $FC(F_s)$,人体特征结点的初始特征为 $FC(F_h)$。定义 $\mathcal{G}_m$ 的邻接矩阵 $\boldsymbol{A}_m$,其计算式如下:

$$\boldsymbol{A}_{m,ij}=\sigma(\mathcal{F}_m(\boldsymbol{F}_{mi}\oplus\boldsymbol{F}_{mj}))+\mathbf{I}_{ij} \tag{3}$$

与局部图不同,中层图中的结点没有明显的物理邻接结构,因此用单位矩阵 $\mathbf{I}$ 替换 $\boldsymbol{A}_c$ 矩阵。

定义全局图 $\mathcal{G}_g=(V_g,\mathcal{E}_g)$,其中 $V_g$ 在 $V_m$ 的基础上,加入了物体特征结点和合并区域特征结点,并分别用相应的全连接投影后的特征向量进行初始化。邻接矩阵 $\boldsymbol{A}_g$ 的计算式与 $\boldsymbol{A}_m$ 相同。

$$\boldsymbol{A}_{g,ij}=\sigma(\mathcal{F}_g(\boldsymbol{F}_{gi}\oplus\boldsymbol{F}_{gj}))+\mathbf{I}_{ij} \tag{4}$$

### 3.4.2 图卷积(GCN)

图卷积继承于图神经网络,是对图结点特征的聚合和

更新过程。本文使用的图卷积的具体形式如下：

$$\boldsymbol{F}^{l+1}=\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{F}^{l}\boldsymbol{W}+\boldsymbol{F}^{l} \tag{5}$$

其中，$\boldsymbol{F}^{l+1}\in\mathbb{R}^{N\times D^{l+1}}$ 表示第 $l$ 层输出的结点特征；$\boldsymbol{A}\in\mathbb{R}^{N\times N}$ 表示邻接矩阵；$\boldsymbol{F}\in\mathbb{R}^{N\times D^{l}}$ 表示第 $l$ 层输入的结点特征；$\boldsymbol{W}\in\mathbb{R}^{D^{l}\times D^{l+1}}$ 是可学习的参数矩阵；$\boldsymbol{D}\in\mathbb{R}^{N\times N}$ 为 $\boldsymbol{A}$ 的度矩阵，其中 $\boldsymbol{D}_{ii}=\sum_{j=1}^{N}\boldsymbol{A}_{ij}$。

### 3.5 基于注意力机制的图池化模块

在卷积神经网络中，池化(pooling)可以有效减少数据量并扩大卷积感受野，同时提高网络的泛化能力。随着图神经网络的兴起，也出现了不少研究将池化模块引入图卷积中，从而减少图结点的数量[18-24]。参考 Huang 等[24]提出的注意力图池化方法(AttPool)，本文针对人体关键点结点进行注意力图池化操作。这项操作具有两个重要目的：1)减少高层图结构中人体的关键点结点的数量，从而平衡各尺度结点在图中的占比；2)利用注意力机制筛选出人体的重要部位，舍弃冗余特征，并进行特征增强。下面具体介绍操作过程，如图 5 所示。

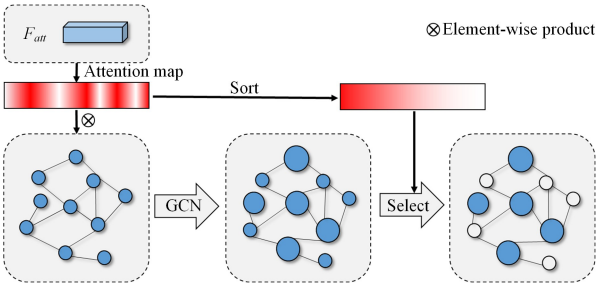


图 5 基于注意力机制的图池化策略

Fig. 5 Graph pooling strategy based on attention mechanism

首先针对所有的人体关键点结点(17 个)计算注意力映射(attention map)：

$$Att=\sigma(\boldsymbol{F}_{\mathrm{att}}(\boldsymbol{F}_{\mathrm{att}})) \tag{6}$$

其中，$Att\in\mathbb{R}^{17}$ 表示 17 个结点的注意力；$\boldsymbol{F}_{\mathrm{att}}(\cdot)$ 表示注意力映射函数，包含两个全连接层，中间添加 Relu 激活函数。$\boldsymbol{F}_{\mathrm{att}}$ 表示参与计算的特征向量，在两次 AttPool 中有着不同的 $\boldsymbol{F}_{\mathrm{att}}$。需要注意的是，为了进行特征增强，我们在 GCN 操作之前先将 $Att$ 作用于相应的人体关键点特征，由此得到了新的 GCN 形式：

$$\boldsymbol{X}_{i}^{l}=Att_{i}\boldsymbol{F}_{i}^{l} \tag{7}$$

$$\boldsymbol{F}^{l+1}=\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{X}^{l}\boldsymbol{W}+\boldsymbol{X}^{l} \tag{8}$$

其中，$Att_i$ 表示第 $i$ 个结点的注意力权重。$F_i^l$ 表示第 l 层第 $i$ 个结点的特征向量。在 HGNN 的第一层和第三层 GCN 中，我们使用式(7)、式(8)所示的 GCN 形式，而在第二层中我们使用式(5)所示的形式，因为第二层 GCN 没有添加 AttPool 操作。

之后对 $Att$ 进行降序排列，并筛选出前 $k$ 个注意力权重较高的人体关键点结点进行保留，舍去其余结点。本文 HGNN 在由局部图得到中层图的过程中(第一层 GCN)$k$ 取 8，由中层图得到全局图的过程中(第三层 GCN)$k$ 取 4。需要注意的是，在 AttPool 过程中被舍弃的结点特征并没有完全损失，而是通过 GCN 聚合在了保留结点之上。另外，尽管第二次池化的原始结点个数为 8，我们在计算注意力映射 $Att$ 时依旧要计算 17 个结点，之后根据第一次筛选结点的索引首先筛选出 8 个相应注意力映射，再对这 8 个进行降序排列及二次筛选。其原因是本文 AttPool 模块使用的不是自注意力模型(self-attention)，即在计算 $Att$ 时使用的特征向量 $\boldsymbol{F}_{\mathrm{att}}$ 并不是结点特征本身，因此得到的注意力映射对结点顺序敏感。

本文 HGNN 中，第一次图池化操作计算注意力映射用到的特征向量 $\boldsymbol{F}_{\mathrm{att}}$ 为空间结构特征 $\mathrm{F_s}$。其原因是从直观上来讲，通过人物空间位置关系，我们可以大致判断交互动作所涉及的人体部位，并给模型提供初步的注意力映射。第二次计算注意力映射用到的特征向量是中层图之后临时的特征分类结果，其原因是该分类结果可以提供近似的人物交互动作类型，从而根据动作类型判断人体各部位的重要程度。

### 3.6 多任务特征分类及损失函数设计

传统的人物交互动作识别模型仅对交互动作进行分类，其利用最终的特征表示直接计算人物对之间包含各类动作的概率[1]。文献[30]首次提出了人物交互性(Interactiveness)概念，其在构造多任务分类模块的同时对人物交互性和交互类别进行分类，从而提升模型精度。本文借鉴了这种多任务特征分类思想，利用 HGNN 最终输出的特征向量 $\boldsymbol{F}_{\mathrm{global}}$ 完成两项分类任务：1)交互性二值分类；2)动作类别分类。其中，交互性二值分类可以得到人物对之间存在交互行为的概率：

$$P_{\mathrm{Int}}=\sigma(\boldsymbol{F}_{\mathrm{Int}}(\boldsymbol{F}_{\mathrm{global}})) \tag{9}$$

其中，$P_{\mathrm{Int}}\in(0,1)$ 表示交互性概率，$\boldsymbol{F}_{\mathrm{Int}}(\cdot)$ 由两个全连接层组成，中间添加 Relu 激活函数。设置阈值 $H_{\mathrm{Int}}$，若 $P_{\mathrm{Int}}<H_{\mathrm{Int}}$，则认为该人物对之间无交互，不再进行交互动作分类；否则继续对交互动作类别进行分类。

$$S_{g}^{a}=\sigma(\boldsymbol{F}_{\mathrm{act}}(\boldsymbol{F}_{\mathrm{global}})) \tag{10}$$

$$P_{h,o}^{a}=S_{g}^{a}S_{h}S_{o}P_{\mathrm{Int}} \tag{11}$$

其中，$S_g^a\in(0,1)$ 为人物对之间存在动作 $a$ 的得分；$\mathscr{F}_{\mathrm{act}}(\cdot)$ 由两个全连接层组成，中间添加 Relu 激活函数；$S_h$ 为 Mask R-CNN 输出的人体检测概率；$S_o$ 为物体检测概率；$P_{h,o}^{a}$ 为人物对之间存在动作 a 的概率。

上文提到，HGNN 在中层图阶段将图结点特征合并为 $F_{\mathrm{middle}}$ 进行初步的特征分类：

$$S_{m}^{a}=\sigma(\mathscr{F}'_{\mathrm{act}}(F_{\mathrm{middle}})) \tag{12}$$

其中，$S_m^a\in(0,1)$ 为人物对之间动作 a 的得分，$\mathscr{F}'_{\mathrm{act}}(\cdot)$ 由两个全连接层组成，中间添加 Relu 激活函数。$S_m^a$ 仅用于计算注意力映射和损失函数，并不代表最终的分类结果。

针对多任务分类，本文设计如下的损失函数：

$$L=\frac{1}{N}\sum_{i=1}^{N}\{\sum_{a=1}^{A}(\mu L_{\mathrm{CrossEntropy}}(y^{a,i},\boldsymbol{S}_{m}^{a,i})+\bar{\omega}L_{\mathrm{CrossEntropy}}(y^{a,i},\boldsymbol{S}_{g}^{a,i}))+\rho L_{\mathrm{CrossEntropy}}(z^{i},\boldsymbol{P}_{\mathrm{int}}^{i})\} \tag{13}$$

其中，$L_{\mathrm{CrossEntropy}}(a,b)=a\log(b)+(1-a)\log(b-1)$；$\mu$，$\bar{\omega}$ 和 $\rho$ 为超参数；$y$ 和 $z$ 分别为人物交互动作标签和交互性标签。

## 4 实验

### 4.1 模型实现细节

本文选择 PyTorch 实现提出的深度模型[35]。在模型实现过程中，人物空间结构图的长宽取 64 像素。人体特征、物体特征和合并区域特征的池化长宽 $R_b$ 取 7 像素。在获取人体关键部位边界框时参数 $\gamma$ 取 0.1，池化长宽 $R_p$ 取 5 像素。池化之后的人体特征、物体特征、合并区域特征和人体关键部位特征经过 FC(258×W×H-256)-Relu 操作后作为图结点的

初始特征向量。人物空间结构图经过 FC(3×64×64-512)-Relu-FC(512-256)-Relu 操作后作为图结点的初始特征向量。在 GCN 操作中,图结点特征向量的维度都为 256。最终全局图的结点个数为 8,每个结点的特征向量维度为 256。我们对其中的 4 个人体关键区域结点特征进行合并,并利用全连接层将特征维度投影为 256,同时将 17 个初始的人体关键区域结点特征合并,利用全连接层将特征维度投影为 256。最后,我们将 256 维的人体特征、物体特征、合并区域特征、空间结构特征以及人体关键区域特征与其初始的 256 维特征合并,构造 512×5 维的全局特征向量,用于交互性分类与交互动作分类。在 Loss 函数中,超参数 $\mu$,$\omega$ 和 $\rho$ 分别取 0.5,1 和 0.1。

### 4.2 数据集与评价标准

微软 MS-COCO 数据集[36]于 2014 年发布,2015 年和 2017 年又陆续更新,如今已成为图像目标检测、语义分割、人体关键点检测等任务的标准测试平台,2015 年 Saurabh 等在 MS-COCO 数据集的基础上,提取其包含人物动作交互的图像子集,并重新进行精细的标注,构建了专门用于人物交互动作检测的标准数据集 V-COCO[26]。V-COCO 包含了 10 346 张图像,其中标注了超过 16 000 个人体实体以及他们的 26 种动作信息。

现有研究均采用 mAP(Mean Average Precision)作为人物交互动作检测的评价指标[1,2,30-33]。当人和物的检测框与标签边界框的 IOU(Interaction Over Union)[34]大于 0.5,且动作类别正确时,被视为一个正确的正样本(True Positive)。

### 4.3 模型训练

本实验使用一张 Titan XP GPU。使用动量 SGD 优化器训练模型,初始学习率设置为 $7.5\times10^{-3}$,权重衰减设置为 $1\times10^{-4}$,动量设置为 0.9。每次迭代的图像数量设置为 3,迭代次数设置为 69 333,训练时间为 12 h。每 200 次迭代的 loss 变化情况如图 6 所示。可以看到,模型的 loss 值在训练过程中逐步收敛,最终趋于平稳。

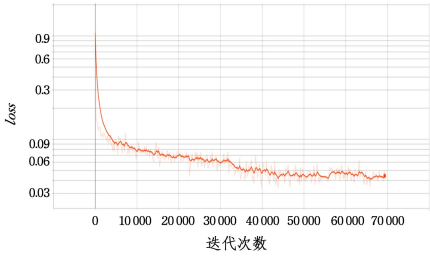


图 6 每 200 次迭代的 loss 变化情况

Fig. 6 Changes in loss per 200 iterations

### 4.4 模型测试

针对 V-COCO 测试集,本实验测试了 24 种人物交互动作的 mAP(除去无交互对象的人体动作,如"smile")。现有基于图神经网络的模型 GPNN[2]和 RPNN[32]在 V-COCO 上的 mAP 分别达到了 44.0%和 47.53%。相比这两者,本文设计的层次化图神经网络结构可以更好地提取人体关键区域以及人和物之间的交互关系,从而得到更有效的特征表示。本文提出的 HGNN 模型的 mAP 达到了 50.99%,超过了 GPNN 模型的 6.99%以及 RPNN 模型的 3.46%,同时超越了大部分现有的基于多流卷积网络的模型。目前主流模型的 mAP 对比结果如表 1 所列,可以看到本文 HGNN 模型的表现处于领先水平。

表 1 现有主流模型 mAP 的对比

Table 1 mAP comparison of mainstream models

| Models | mAPs/% |
|---|---|
| InteractNet(CVPR2018)[28] | 40.00 |
| iCAN(BMVC2018)[29] | 44.70 |
| TIN(CVPR2019)[30] | 47.80 |
| Bingjie Xu et al.(CVPR2019)[33] | 45.90 |
| GPNN(ECCV2018)[2] | 44.00 |
| RPNN(ICCV2019)[32] | 47.53 |
| HGNN(Ours) | 50.99 |

注:灰色单元格表示基于 GNN 的模型

针对实验中将人物交互动作区分的 24 种,其详细的 AP(Average Precision)如图 7 所示。AP 值最高的动作为"skateboard-instr",其 AP 达到了 86.52%,主要原因是该动作的数据质量较好,交互行为清晰。AP 值最低的动作为"eat-instr",其 AP 为 6.13%,主要原因是数据集中样本不足,且"eat-instr"与"hold-instr"等动作易混淆。另外,一半以上的交互动作 AP 值均在 50%以上,其说明了本文 HGNN 模型对大多数常见的人物交互动作和交互场景均具有适应性。

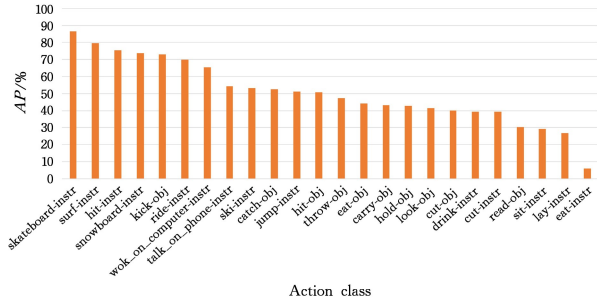


图 7 V-COCO 测试集中 24 类人物交互动作的 AP 对比

Fig. 7 AP comparison of 24 types of interactive actions in V-COCO test set

图 8 随机给出了 V-COCO 测试集中一些人物交互动作的检测结果。其中,红色框表示人体边界框,绿色框表示与其交互的物体边界框,下方文字为动作置信度和动作类别。可以看到,本文提出的 HGNN 模型可以在复杂现实场景中准确检测并识别常见的人物交互动作,并且可以有效检测出同一个主体对不同客体的不同交互动作(例如,图 8 右上角"sit"和"eat")。


图 8 V-COCO 测试集中部分人物交互动作检测结果(电子版为彩图)

Fig. 8 Part of detection results of interactive actions in V-COCO test set

### 4.5 消融实验与可视化实验

为了验证 HGNN 模型各部分的有效性,现对其进行消融实验。

4.5.1　人物空间结构图(H-O SCM)

为了突出利用人物空间位置关系以及人体姿态信息，本文 HGNN 模型构造了人物空间结构图并通过卷积层在其上提取特征。人物空间结构特征可以帮助 HGNN 网络更好地学习人物交互模式，以提升动作识别的准确性。实验结果表明，人物空间结构图帮助模型提升了 0.6% mAP(见表 2)。

表 2　HGNN 模型在 V-COCO 测试集上的消融实验

Table 2　Ablation experiment of HGNN model on V-COCO testset

(单位：%)

| Models | | | | | mAP |
|---|---|---|---|---|---|
| Basis | H-O SCM | GCN | AttPool+MC | IC | |
| √ | | | | | 47.50 |
| √ | √ | | | | 48.10 |
| √ | √ | √ | | | 49.85 |
| √ | √ | √ | √ | | 50.65 |
| √ | √ | √ | √ | √ | 50.99 |

4.5.2　图卷积神经网络(GCN)

在构建场景图结构的基础上，利用 GCN 进行特征聚合与优化是 HGNN 模型学习人物交互模式的重要步骤。实验结果表明，图结构与 GCN 模块帮助模型提升了 1.75% mAP(见表 2)。

4.5.3　基于注意力机制的图池化和中层图分类(AttPool+MC)

图池化操作可以减少图结点个数，有效去除噪声与冗余信息。基于注意力机制的图池化模块帮助模型自学习得到重要的人体交互部位，并利用监督中层图分类使模型聚焦于重要的人体特征，缓解了物体特征对动作识别造成的“偏见”。该模块帮助模型提升了 0.8 mAP(见表 2)。

4.5.4　交互性分类(IC)

在具体动作类别分类之前，首先进行人物交互性分类，以判断人物对之间是否存在交互。该模块可以有效过滤掉不存在交互状态的人和物，避免模型做出不必要的动作类别分类。该模块帮助模型提升了 0.34 mAP(见表 2)。

为验证 HGNN 模型中注意力机制的有效性，现对其进行可视化实验。图 9 随机给出了 V-COCO 测试集中两张图像的注意力权重，红色越深表示人体该区域注意力权重越大。如图 9 左侧所示，对于“look”和“skateboard”两个动作，注意力模块可以正确筛选出人体下半身区域和头部区域。如图 9 右侧所示，对于“hold”和“hit”两个动作，注意力模块正确筛选出了人体手臂区域。由此可见，基于注意力机制的图池化模块可以帮助模型正确筛选出人体关键区域，并剔除冗余信息，从而提升模型对动作细节的感知能力，提高动作分类精度。

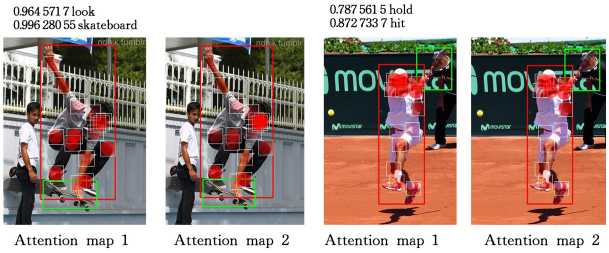


图 9　V-COCO 测试集中部分人物交互动作注意力权重可视化(电子版为彩图)

Fig. 9 Visualization of part of the attention weight of interactive actions in V-COCO testset

**结束语**　人物交互动作识别是计算机视觉领域的重要研究课题，其在智能监控、运动分析、人机交互、辅助医疗、多媒体内容审核等领域有着重大实践意义，同时具有很高的应用前景和市场价值，近年来也吸引了众多国内外学者的关注。然而，由于实际环境中难以避免复杂因素，包括运动背景变化、遮挡、视角变化等，传统的基于多流卷积的人物交互动作识别模型只在视觉层面提取人物的表观特征，难以对人物交互的深层关系特征进行挖掘。本文针对多流卷积模型的缺陷，提出了用层次图模型 HGNN 从局部到整体对人物交互场景建模，利用注意力图池化机制剔除层次图中冗余的信息和噪声，再通过图卷积网络提取图结点之间的深层语义关系，对卷积网络提取的特征进行聚合与优化，从而得到反映人物交互动作本质的特征表示。最后，对中层图的监督分类也约束网络更好地学习到交互动作的人体模式，避免网络对交互对象产生“偏见”。为了验证 HGNN 模型的有效性，在公开数据集 V-COCO 上进行了完整的实验，包括精度对比、消融实验与可视化实验。实验结果均显示本文提出的 HGNN 模型对常见的人物交互动作具有广泛的适应性和鲁棒性，mAP 超过了现有的基于图神经网络的模型，同时领先大部分最新的多流卷积模型。

## 参 考 文 献

[1] CHAO Y W, LIU Y, LIU X, et al. Learning to detect human-object interactions[C]// 2018 IEEE Winter Conference on Applications of Computer Vision(WACV). IEEE, 2018: 381-389.

[2] QI S, WANG W, JIA B, et al. Learning human-object interactions by graph parsing neural networks[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 401-417.

[3] TANG C, WANG W J, ZHANG C, et al. Human Action Recognition Using RGB-D Image Features[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(10): 901-908.

[4] LI X, XIAO Q K. Human Action Recognition Using Auto-encode and PNN Neural Network[J]. Software Guide, 2018, 17(1).

[5] ZHANG J, JIA Y, XIE W, et al. Zoom Transformer for Skeleton-based Group Activity Recognition[C]// IEEE Transactions on Circuits and Systems for Video Technology. 2022.

[6] ZHANG J, YE G, TU Z, et al. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition[J]. CAAI Transactions on Intelligence Technology, 2022, 7(1): 46-55.

[7] ZITNIK M, LESKOVEC J. Predicting multicellular function through multi-layer tissue networks[J]. Bioinformatics, 2017, 33(14): 190-198.

[8] ZHOU Z, WANG Y, XIE X, et al. RiskOracle: A Minute-level Citywide Traffic Accident Forecasting Framework[J]. arXiv: 2003.00819, 2020.

[9] TANG L, LIU H. Relational learning via latent social dimensions[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009: 817-826.

[10] SUHAIL M, SIGAL L. Mixture-Kernel Graph Attention Network for Situation Recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 10363-

10372.
[11] WU J, WANG L, WANG L, et al. Learning actor relation graphs for group activity recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9964-9974.
[12] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks, 2008, 20(1): 61-80.
[13] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]// Advances in Neural Information Processing Systems. 2016: 3844-3852.
[14] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv: 1609. 02907, 2016.
[15] MONTI F, BOSCAINI D, MASCI J, et al. Geometric deep learning on graphs and manifolds using mixture model cnns[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5115-5124.
[16] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Thirty-second AAAI Conference on Artificial Intelligence. 2018.
[17] LI L, GAN Z, CHENG Y, et al. Relation-aware graph attention network for visual question answering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 10313-10322.
[18] HENAFF M, BRUNA J, LECUN Y. Deep convolutional networks on graph-structured data[J]. arXiv: 1506. 05163, 2015.
[19] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[C]// Advances in Neural Information Processing Systems. 2016: 3844-3852.
[20] JOAN B, WOJCIECH Z, ARTHUR S, et al. Spectral networks and locally connected networks ongraphs[J]. arXiv: 1312. 6203, 2013.
[21] GAO H, JI S. Graph u-nets[J]. arXiv: 1905. 05178, 2019.
[22] YING Z, YOU J, MORRIS C, et al. Hierarchicalgraph representation learning with differentiable pooling[C]// Advances in Neural Information Processing Systems. 2018: 4800-4810.
[23] LEE J, LEE I, KANG J. Self-attention graph pooling[J]. arXiv: 1904. 08082, 2019.
[24] HUANG J, LI Z, LI N, et al. AttPool: Towards Hierarchical Feature Representation in Graph Convolutional Networks via Attention Mechanism[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 6480-6489.
[25] GUPTA A, KEMBHAVI A, DAVIS L S. Observing human-object interactions: Using spatial and functional compatibility for recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(10): 1775-1789.
[26] SAURABH G, JITENDRA M. Visual semantic role labeling [J]. arXiv: 1505. 04474, 2015.
[27] CHAO Y W, WANG Z, HE Y, et al. Hico: A benchmark for recognizing human-object interactions in images[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 1017-1025.
[28] GKIOXARI G, GIRSHICK R, DOLLAR P, et al. Detecting and recognizing human-object interactions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8359-8367.
[29] GAO C, ZOU Y, HUANG J B. ican: Instance-centric attention network for human-object interaction detection[J]. arXiv: 1808. 10437, 2018.
[30] LI Y L, ZHOU S, HUANG X, et al. Transferable interactiveness knowledge for human-object interaction detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3585-3594.
[31] WAN B, ZHOU D, LIU Y, et al. Pose-aware Multi-level Feature Network for Human Object Interaction Detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 9469-9478.
[32] ZHOU P, CHI M. Relation Parsing Neural Network for Human-Object Interaction Detection[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 843-851.
[33] XU B, WONG Y, LI J, et al. Learning to detect human-object interactions with knowledge[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[34] HE K, GKIOXARI G, DOLLAR P, et al. Mask r-cnn[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.
[35] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high-performance deep learning library[C]// Advances in Neural Information Processing Systems. 2019: 8024-8035.
[36] RAHMAN M A, WANG Y. Optimizing intersection-over-union in deep neural networks for image segmentation[C]// International symposium on visual computing. Cham: Springer, 2016: 234-244.

**LI Bao-zhen**, born in 1983, engineer. His main research interests include mechanical engineering and automation.

**YU Ping**, born in 1978, graduate. His main research interests include deep learning and computer vision.