Opto-Electronic Engineering
光 电 工 程

Article
2022 年, 第 49 卷, 第 7 期

DOI: 10.12086/oee.2022.210429

基于交互实例推荐网络的人-物交互检测方法研究

薛丽霞，尹凯建，汪荣贵，杨　娟*

合肥工业大学计算机与信息学院，安徽 合肥 230031

摘要： 人-物交互检测任务的目标是定位并且识别图像中人与其周围物体的交互关系。该任务的挑战在于机器无法知道人具体和哪些物体存在交互关系，现有方法大多对人和物进行完全配对来解决这个问题。与他们不同，本文提出了一种基于关系推理的交互实例推荐网络来适应人-物交互检测任务，主要想法是利用人和物体的视觉关系中潜在的交互关系来推荐人-物对。此外，本文还设计了一个跨模态信息融合模块，对不同的上下文信息根据其对检测结果的影响程度进行融合，以此提高检测精度。本文在 HICO-DET 和 V-COCO 数据集上进行了充分的实验来验证所提出的方法，结果表明，本文方法在 HICO-DET 和 V-COCO 数据集上的 mAP 达到了 19.90% 和 50.3%，分别比基准网络高了 4.5% 和 2.8%。

关键词： 人-物交互检测；行为检测；注意力机制；图神经网络；图像理解

中图分类号： TP391.4　　**文献标志码：** A

薛丽霞，尹凯建，汪荣贵，等. 基于交互实例推荐网络的人-物交互检测方法研究 [J]. 光电工程，2022，**49**(7): 210429
Xue L X, Yin K J, Wang R G, et al. Interactive instance proposal network for HOI detection[J]. *Opto-Electron Eng*, 2022, **49**(7): 210429

Interactive instance proposal network for HOI detection

Xue Lixia, Yin Kaijian, Wang Ronggui, Yang Juan*

School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230031, China

Abstract: Human-object interaction detection is to locate and identify the interactive relationship between humans and objects in an image. The challenge is that the machine cannot know which object the person is interacting in. Most existing methods try to solve this problem by matching humans and objects exactly. Different from them, this paper proposes an interactive instance proposal network based on relational reasoning to adapt to the task. Our main idea is to recommend human-object pairs by using the potential interaction relationships in the visual relationship between humans and objects. In addition, a cross-modal information fusion module is designed to fuse different context information according to its influence on the detection result, so as to improve the detection accuracy. To evaluate the proposed method, we performed sufficient experiments on two large-scale datasets: HICO-DET and V-COCO. Results show that our method achieves 19.90% and 50.3% mAP on HICO-DET and V-COCO, which are 4.5% and 2.8% higher than our baseline, respectively.

Keywords: Human-object interaction detection; action detection; attention; GNNs; image understanding

收稿日期：2022-01-10；收到修改稿日期：2022-04-06
基金项目：重点研发计划 (2020YFC1512601)；国家自然科学基金资助项目 (62106064)
*通信作者：杨娟，yangjuan6985@163.com。

1 引 言

近年来，视觉关系检测在目标检测[1-2]，动作识别[3-4]，和场景分割[5-6]等领域取得了长足的发展。但是为了更深层次地理解图像，识别图像中的场景，不仅需要定位单个对象实例，还需要识别对象之间的交互关系。因此学者们开辟了视觉关系检测的一个重要分支领域：人-物交互检测 (Human-object interaction detection，HOI)，该项任务旨在检索图像中人和物的位置，并且识别存在于两者之间的交互动作。这项工作对行为理解至关重要，引起了越来越多研究人员的关注，最近在该领域利用深度神经网络进行的研究已经取得了令人瞩目的进展[7-16]。

HOI 的具体任务是推断场景中的三元组<主语，谓语，宾语>。例如，在图 1 中，首先检测定位出人和物，接下来推断他们的关系，最终得出一个三元组<human, ride, motorcycle>。一般地，该领域采用双阶段检测方法，对于给定图像和它的目标检测结果，首先将人和物完全配对，接着模型将这些人-物对归类为不同交互类别[8-9,17]。由于一个人可以同时和多个物体发生交互，同时，一个物体也可以与多个人发生交互，因此 HOI 检测本质上是一个多标签分类任务[18]。Chao 等人[9]首先提出了一种利用人-物视觉特征和空间特征的多流方法来检测 HOI。随后 Gao 等人[8]在此工作的基础上提出了一种以实例为中心的多流网络来检测人物交互，利用注意力机制聚焦图片中对 HOI 检测有帮助的区域，将 HOI 的检测效果提高到了一个新的高度。之后，Li 等人[19]扩展了 Gao 等人[8]的方法，利用姿势信息进一步强化表达了人物交互之间的细粒度上的区别，学习了一种可迁移的 HOI 知识表达方法。此外，受到图模型在场景理解[20-23]领域成功应用的启发，不少学者尝试将图模型和神经网络相结合来解决 HOI 检测问题[11,17,24]。最近的工作大多通过引入额外信息进行检测，如先验知识[25]，语义嵌入[17]，人体骨骼[26]等，也有学者尝试利用 Transformer 的自注意力机制来改进 HOI 模型[12,14]，还有一些工作[27-30]致力于解决 HOI 检测中的长尾分布问题。

然而，基于一种朴素的直觉，一张图片中不可能所有的人-物对之间都存在交互关系，上述方法中都对大量的非交互对也进行了推理，这无疑增加了检测结果错误的可能性。尽管 Li 等人[19]对非交互对抑制做了一些尝试，但效果并不理想，这是因为没有充分利用到人-物之间的交互关系。

我们进一步发现，现有 HOI 检测相关工作[8-11,17,19,24,27,31-32]都使用现成的目标检测模型[1]来生成人和物的推荐框，然而目标检测模型的检测目标与 HOI 检测的需求并不匹配：前者定位图像中所有实例，后者希望只定位存在交互关系的实例。受到 Wang 等人[28]的启发，本文提出一种基于关系推理的适用于 HOI 检测的目标检测器，充分利用图像中人-物间的交互关系进行人、物推荐，尽可能减少非交互人-物对的出现。本文方法遵循 Faster RCNN[1]的流程，用交互实例推荐网络 (interactive instance proposal network，IIPN) 替换了原始的区域推荐网络 (region proposal network，RPN)，IIPN 根据人-物之间的交互可能性，通过图神经网络 (graph neural networks，简写为 GNNs) 的迭代推理，筛选出存在交互关系的人和物作为其输出，由此减少后续关系推理的数量。

此外，现有方法在推理环节均为简单将几种特征相加或拼接[8,17,19,28]，如：人外观特征、物体外观特征、人-物的空间特征。这种做法忽略了在不同动作中各类特征的影响程度不同，例如：区分<human, ride, bike>和<human, hold, bike>两个三元组中的 ride (骑) 和 hold (推) 动作更多的取决于 human 和 bike 之间的空间关系，而要区分<human, eat, pizza>和<human,

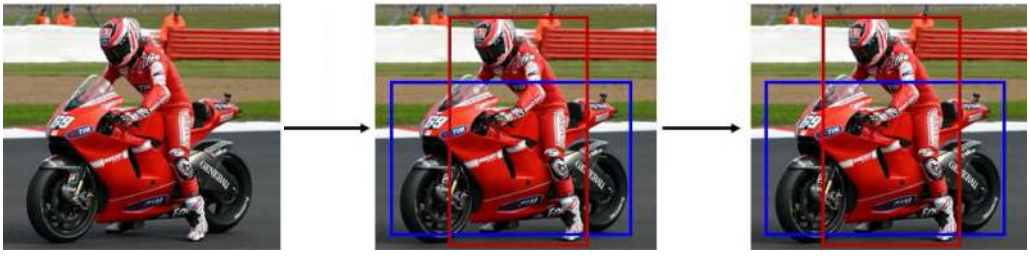


图 1 人-物交互检测流程

Fig. 1 Pipeline of human object interaction detection

cut, pizza>中的 eat 和 cut 动作则更多取决于人的姿势，即外观特征。同时，本文认为语义先验对 HOI 检测也有帮助，例如：图像中同时出现人和苹果 (apple)，那么来自苹果的语义先验将会提示模型去关注与苹果相关的动作，如吃 (eat) 或拿 (hold)，而非骑 (ride) 或站 (stand) 这种毫不相关的动作。因此，本文设计了一种多模态信息融合模块 (cross-modal information fusion module，CIFM)，根据不同特征对检测结果的影响程度计算融合注意力，并进行加权融合。

综上所述，本文的贡献可以总结为：1) 本文提出一种基于关系推理的交互实例推荐网络，推理图像中人和物之间的交互关系，并依此筛选出正确的人-物对，极大地减少了模型对非交互人-物对的推理；2) 本文注意到不同特征对动作预测结果的影响程度不同，提出了一个新的多模态信息融合模块，基于融合注意力计算不同特征对预测结果的影响值并加权融合；3) 本文在 HICO-DET 和 V-COCO 数据集上进行了完备的实验，证明了所提方法的有效性，并与若干种现有方法进行比较，结果表明，本文方法在 HICO-DET 数据集上取得了最优的效果。

2 本文方法

2.1 概述

本文提出的检测方法主要分为两步：1) 交互实例推荐；2) HOI 检测。与大多数现有方法不同[8-11,17,19,24,27,31-33]，本文方法第一步使用交互实例推荐网络 (IIPN) 推荐图片中的交互人-物对。我们用$b_h \in H$来表示检测出的人边界框，用$b_o \in O$表示检测到的物边界框，其中H和O分别表示检测出来的人和物的集合。此外，对于检测到的人和物，用s_h和s_o来分别表示它们的置信度。第二步，将 IIPN 的输出输入到跨模态信息融合模块 (CIFM) 来融合特征并预测交互得分。图 2 描述了本文方法的总体结构。

2.2 交互实例推荐网络

如上所述，本文方法首先通过 IIPN 推荐交互对，在 Li 等人[19] 的研究中表明，图片场景中充斥着大量非交互对。IIPN 的目标是关注并推荐场景中的交互对，为了达到这一目的，我们设计了两个分支分别用来选择得分较高的M个人的锚盒和N个物的锚盒。接着将它们建模为图，利用图神经网络在图中传递信息，从而推荐存在交互关系的人物对。

相较于原始的 RPN，本文的做法有着明显的优势。原始 RPN 中，锚盒的得分仅仅与它覆盖某个物体的程度相关，一个锚盒覆盖一个物体的部分越多，它的得分也越高。然而在 HOI 检测任务中，不仅需要准确地检测物体，更重要的是检测出具有交互关系的人和物，我们希望，存在交互关系的物锚盒能够获得更高的分数。IIPN 的详细结构如图 3 所示。

2.2.1 人、物选择分支

遵循 Wang 等人[28] 的设计，本文同样使用了一个人选择分支和一个物选择分支，目的是从 RPN 的隐藏层特征中，计算出得分较高的锚盒，为后续的图建模提供输入。在人选择分支中，首先计算得出前 M 个人隐藏层输出，接着根据输出获得对应的隐藏层特征，由此便获得了前 M 个人的隐藏层特征。物选择分支的计算过程类似。于是，我们得到了M个人的隐

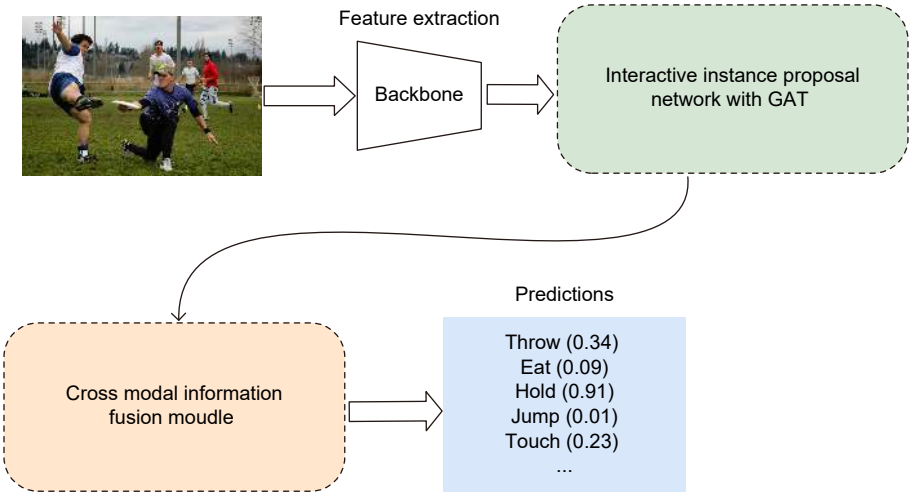


图 2 基于交互实例推荐网络的 HOI 检测方法纵览

Fig. 2 Overview of human object interaction detection based on interactive instance proposal network

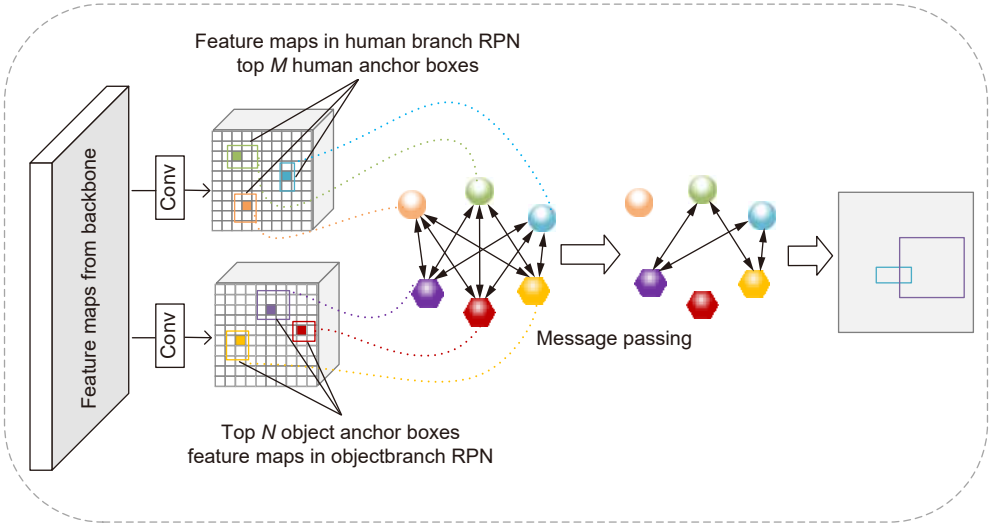


图 3 交互实例推荐网络

Fig. 3 Interactive instance proposal network

藏层特征和N个物的隐藏层特征。下面开始介绍用于推理关系的注意力图神经网络。

2.2.2 基于注意力图神经网络的关系推理

1) 图定义。一个图的定义为$G=(V,E)$，它由一组V个节点和E条边组成。我们用F_v表示节点的特征，用F_e表示边上的特征。令$v_i \in V$表示图中的第i个节点，$e_{ij}=(v_i,v_j))$表示从节点v_i指向节点v_j的边。对于一个具有n个节点和m条边的图，用$\boldsymbol{X}_v \in \boldsymbol{R}^{n\times d}$表示节点特征矩阵，用$\boldsymbol{X}_e \in \boldsymbol{R}^{m\times c}$表示边特征矩阵，其中，节点$i$的特征向量表示为$\boldsymbol{F}_{v_i}$，$(i,j)$之间边的特征向量表示为$\boldsymbol{F}_{e_{ij}}$。

2) 注意力图神经网络。为了通过图结构在上下文中传递信息，本文使用了注意力图神经网络。在描述本模块之前，先来回顾一下基本的图注意力网络。图注意力网络利用图结构和节点特征来更新节点的表征。假设某个节点v_i的特征表示为$z_{v_i} \in \boldsymbol{R}^{d_v}$，其相邻节点的特征$\{z_{v_i} \big| v_j \in N(v_i)\}$首先通过一层可学习的线性映射$\boldsymbol{W}$，接着将映射后的特征以预定义的权重经过$k$层图卷积以及聚合函数来进行聚合，那么每一层中节点$v_i$的聚合特征$a_{v_i}$计算如下：

$$a_{v_i}=f_{\text{aggregate}}^{(k)}(\{a_{ij}\boldsymbol{W}z_{v_j}^{(k-1)}:v_j\in N_i\}), \tag{1}$$

$$z_{v_i}^{(k)}=f_{\text{update}}^{(k)}(\{z_{v_i}^{(k-1)},a_{v_i}^{(k)}\}). \tag{2}$$

初始化时，$z_{v_i}^{(0)}=F_{v_i}$，N_i表示节点v_i的所有邻接节点，a_{ij}表示预定义的权重。对于$f_{\text{aggregate}}^{(k)}(\cdot)$和$f_{\text{update}}^{(k)}(\cdot)$，存在多种方法，通常对于前者使用均值化的聚合策略，对于后者采用相加的更新策略[17]，于是可以将式 (1) 展开如下：

$$a_{v_i}=\frac{1}{|N_i|}\sum_{v_j\in N_i}z_{v_j}, \tag{3}$$

$$z_{v_i}^{(k)}=\sigma(z_{v_i}^{(k-1)}+a_{v_i}^{(k)}). \tag{4}$$

接着将两式合并，并将节点的表征整合成矩阵的形式，上面的等式就可以合并如下：

$$z_{v_i}^{(k)}=\sigma\left(\boldsymbol{W}\boldsymbol{Z}^{(k-1)}\boldsymbol{\alpha}_i\right), \tag{5}$$

其中：σ表示非线性映射函数 ReLU，$Z^{(k-1)} \in \boldsymbol{R}^{d_v\times Tn}$。对于不与$v_i$相邻的节点$v_j$来说$\boldsymbol{\alpha}_{ij}$值为 0，对于节点$v_i$允许自连接，此时$\boldsymbol{\alpha}_{ii}=1$。对于普通的 GNNs 来说，图中的节点连接是已知的，其系数向量$\boldsymbol{\alpha}_i$是基于特征的对称归一化邻接矩阵预设的。

3) 关系推理。经过人和物选择分支的计算，得到了一组M个人的特征和N个物的特征，这里暂时将其表示为$h_i \in \boldsymbol{R}^d, o_i \in \boldsymbol{R}^d$，其中在本文的实验中$d=512$。将$M$个人和$N$个物体的隐藏层特征构建成特征矩阵$\boldsymbol{F} \in \boldsymbol{R}^{(m+n)\times d}$，相应的邻接矩阵形式为$\boldsymbol{E}^{k\times k}$，其中$k=m+n$。通常由于一个物体可以与多个人发生交互，一个人也可以同时与多个物体发生交互，同时，本文在实验中规定，不考虑人和人、物和物之间的交互关系。于是得到初始邻接矩阵式：

$$\boldsymbol{E}=\begin{bmatrix} 0 & 0 & \cdots & 1 & 1 & 1 \\ 0 & 0 & & 1 & 1 & 1 \\ \vdots & & \ddots & & \vdots & \\ 1 & 1 & & 0 & 0 & 0 \\ 1 & 1 & \cdots & 0 & 0 & 0 \\ 1 & 1 & & 0 & 0 & 0 \end{bmatrix}. \tag{6}$$

从上面注意到，由于缺乏完善的监督关系标签，

模型很难直接去计算人物对之间的交互连接。为了解决这个问题，我们引入了一种注意力机制来隐式地学习人物对之间的交互相关性，用$\alpha_{o,h}$表示，同时，用α_{ij}来动态地更新$\boldsymbol{E}[i,j]$，使得存在相关性的人物对之间可以获得充分的信息交互。为了计算这种潜在的交互关系，受到文献 [24] 的启发，我们利用一个多层感知机$R_r(f_o,f_h)$来做这个工作，R_r的输入是f_o和f_h拼接之后的向量。于是，这种交互关系可以表示为

$$s_r = R_r(f_o, f_h)\,. \tag{7}$$

这里的交互关系得分就表示一对人物对之间的潜在交互可能性，由于每次只计算一对人物对，因此这个交互概率可以通过一个 softmax 函数进行归一化，如下式：

$$\alpha_{o,h} = \frac{\exp(s_r)}{\sum_{n=1}^{N}\exp(s_{r_n})}\,, \tag{8}$$

其中：N表示一张图片中人物对的个数。

在计算出$\alpha_{o,h}$之后，接着就可以获得动态更新的邻接矩阵，通过带有信息的邻接矩阵，就可以更新f_h和f_o，其过程可以用下式表示：

$$f_o^k = f_o^{k-1} + \sum_{o=1}^{N_o}\alpha_{o,h}\cdot f_h\,, \tag{9}$$

$$f_h^k = f_h^{k-1} + \sum_{h=1}^{N_h}\alpha_{h,o}\cdot f_o\,, \tag{10}$$

其中：N_o和N_h分别表示图像中的人和物的个数。与 Zhang 等人[34] 的实验设置一样，在本文的实验中$k=2$，即会有两轮的信息传递和汇集过程。IIPN 进一步通过处理f_h和f_o得到人、物边界框 (后续处理流程与 Faster RCNN[1] 相同)。

2.3 跨模态信息融合模块

在 2.2 节中介绍了关系推理模块，并且利用推理模块获得了只存在交互关系的人、物边界框。具体地，将 IIPN 的输出表示为，$\boldsymbol{p}=\{p^1,p^2,\ldots,p^m\}$,$\boldsymbol{o}=\{o^1,o^2,\ldots,o^m\}$，其中，每个边界框都包含了$[x,y,w,h,s,c]$。其中$x$和$y$分别表示边界框中心点的横、纵坐标，$w$和$h$分别表示边界框的宽和高，$c$表示分类，$s$表示其置信度。上文中提到，现有的 HOI 方法利用多种信息来进行推理，如人和物的视觉信息、空间信息、编码之后的距离和位置信息、甚至是语义信息。然而这些工作中对这些信息仅仅是粗糙的相乘或者相加，并没有挖掘出更深层次的隐含信息。我们认为，不同信息对于 HOI 检测结果的贡献是不一样的，为了进一步挖掘这些隐含信息，本文设计了一个跨模态信息融合模块，其结构如图 4 所示。

2.3.1 外观特征

本文的外观特征分为人和物两种，通过带有特征金字塔[35] 的 ResNet[36] 作为骨干网络提取图片特征，接着利用 ROI Pooling 和 IIPN 生成的人物边界框，提取到具体的人和物的外观特征，分别将它们表示f_h和f_o，以便后续的分类。

2.3.2 联合空间特征

尽管视觉信息中已经包含了相当的线索可供动作识别，但这还远不够[9]。仅仅利用单独的视觉信息往往会导致错误的预测。例如在<human, eat, apple>和<human, hold, apple>这两个三元组中，视觉信息是非常接近的，只利用视觉信息无法做出正确的预测。为

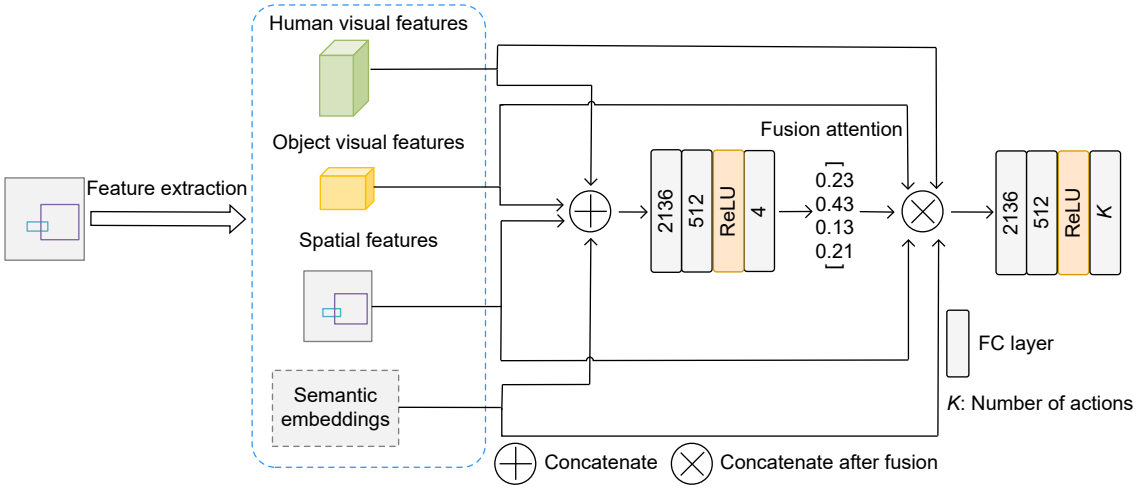


图 4 跨模态信息融合模块详细结构

Fig. 4 Structure of cross-modal information fusion module

了消除这种错误，提高模型精度，本文借鉴了文献 [9] 中的做法，对人物对的空间位置关系也进行了编码，由于更加关注人和物的空间位置关系，本部分的输入应该忽略像素值而只利用边界框的位置信息。为达到这一目的，本文利用了 Chao 等人[9] 设计的一种双通道的二进制图像。具体做法是，对于每一对边界框，第一个通道中位于人边界框之内的值全为 1，否则全为 0，第二个通道中位于物边界框之内的值全为 1，否则全为 0。接着利用卷积神经网络从这个双通道图片中来提取空间特征。不同的是，本文还引入了联合外观特征来优化模型的精度，具体的做法是，对于一对$b_{\rm h}=(x_1^{\rm h},y_1^{\rm h},x_2^{\rm h},y_2^{\rm h})$和$b_{\rm o}=(x_1^{\rm o},y_1^{\rm o},x_2^{\rm o},y_2^{\rm o})$，先将其构建为联合框，如下式：

$$b_{\rm u}=\begin{cases}x_1^{\rm u}=\min(x_1^{\rm h},x_1^{\rm o})\\ y_1^{\rm u}=\min(y_1^{\rm h},y_1^{\rm o})\\ {\rm x}_2^{\rm u}=\max(x_2^{\rm h},x_2^{\rm o})\\ y_2^{\rm u}=\max(y_2^{\rm h},y_2^{\rm o})\end{cases}. \tag{11}$$

接着利用 ROI Pooling 提取出联合外观特征，我们认为联合的特征可能含有一些有用的上下文信息，在实验中证明本文这么做的优点。最后，将提取到的空间编码特征和联合外观特征进行按元素相加，得到联合空间特征，并将其表示为$f_{\rm u}$。

2.3.3 语义特征

在许多工作中都证明了语义信息在 HOI 检测中的有效性[17,37]。为了消除 HOI 检测的歧义预测，进一步提高模型的精度。本文同样引入了语义特征。具体地，本文使用目前流行的 Glove[38] 来提取词嵌入，它接受文本输入，输出文本的向量表示，这种向量表示潜在地保留了文本的语义和语法特征。本文使用了公开的预训练过的 Glove 模型[38]，该模型对于输入的单词和短语产生 300 维的向量。HICO-DET 中的所有三元组都用来获得词向量表示，本文用$f_{\rm sem}$表示，并且用来生成特征融合的注意力系数。

2.3.4 特征融合

至此，得到了包括人和物的外观特征$f_{\rm h}$和$f_{\rm o}$，空间外观特征$f_{\rm u}$，动作的语义特征$f_{\rm sem}$。为了确定四种特征对最终预测结果的影响程度，本文设计了一个简单而高效的融合模块来动态地计算每次预测时不同因素的注意力。具体地，利用一个多层感知机 (multilayer perceptron，MLP) 来完成这项任务。MLP 包含 3 层全连接层，其中一层为 2136 维，另一层为 512 维，最后一层为 4 维。首先将$f_{\rm h},f_{\rm o},f_{\rm u}$和$f_{\rm sem}$进行拼接，接着通过 MLP 产生一个 4 维的注意力向量，如下式：

$$A=\sigma(f_{\rm A-MLP}(f_{\rm h},f_{\rm o},f_{\rm u},f_{\rm sem}))\,. \tag{12}$$

最后利用注意力进行加权融合得到预测向量。再通过一个 MLP 输出最终的预测向量。MLP 同样包含 3 层全连接层，分别为 2136、512 和 117(在 V-COCO 数据集上为 26) 维。对于一组人-物对，模型计算某个动作v的概率如下：

$$P(v|(b_{\rm h},b_{\rm o}))=\sigma(f_{\rm P-MLP}(a_if_{\rm h},a_if_{\rm o},a_if_{\rm u},a_if_{\rm sem}))\,, \tag{13}$$

其中：$\sigma(\cdot)$表示 sigmoid 函数，$(\cdot,\cdot,\cdot,\cdot)$表示对特征向量的拼接。

2.4 训练

本文模型的训练分为两部分，第一部分为对 IIPN 端到端的训练。损失函数与 Faster RCNN[1] 相同，包括分类损失和回归损失。如下式：

$$\begin{aligned}L(p_i,t_i)=&\frac{1}{N_{\rm cls}}\sum_i L_{\rm cls}(p_i,p_i^*)\\&+\lambda\frac{1}{N_{\rm reg}}\sum_i p_i^*L_{\rm reg}(t_i,t_i^*)\,.\end{aligned} \tag{14}$$

训练的第二阶段是对 HOI 检测任务进行端到端的训练，此时目标函数为一个多标签分类的二元交叉熵 (binary cross entrop，BCE) 损失函数，公式如下：

$$L=\frac{1}{N\times K}\sum_{i=1}^{N}\sum_{j=1}^{K}f_{\rm BCE}(P_{ij},y_{ij}^{\rm label})\,, \tag{15}$$

其中：N为人-物对的个数，K为预测的动作的个数，在 HICO-DET 数据集上$K=117$，在 V-COCO 数据集上$K=26$。

3 实 验

为了验证本文方法的有效性，本文在两个大型 HOI 数据集上进行了一系列严谨的实验，并与目前的若干个方法做了对比。本文还对模型检测效果做了一些可视化展示以证明本文方法的有效性和优越性。

3.1 数据集和评估指标

3.1.1 数据集

本文所选取的数据集分别是 HICO-DET[9] 和 V-COCO[39]。HICO-DET 是当下最流行的大型 HOI 数据集，它包含了 47776 张图片 (其中训练集 38118 张，测试集 9658 张)，117 个动词，80 种物品，以及 600 个 HOI 类别。此外，HICO-DET 还进一步将 600 个 HOI 划分为 462 个常见类别和 138 个稀有类别，稀有类别是指训练样本少于 10 个类别。与 HICO-DET 相

比，V-COCO 要小一些，它是 MS-COCO[40] 的一个子集，总计包含 10346 张图片 (其中训练集 2553 张，验证集 2867 张，测试集 4946 张)，25 个动词和 80 种物品。

3.1.2 评估指标

本文采用 Chao 等人[9] 的标准来评估本文模型的检测效果。具体地，当且仅当一个三元组满足以下条件时才被认为是正确的正样本：1) 检测的人框和物框与真实值的 IoU 大于 0.5；2) 检测出的 HOI 类别与真实值相同。对于 V-COCO，与 Gupta 等人[39] 一样计算$\mathrm{mAP}_{\mathrm{role}}$来评估模型。而对于 HICO-DET，本文评估三个方面的 mAP：1) 所有 600 个 HOI 类别 (记为 Full)；2) 138 个稀有的 HOI 类别 (记为 Rare)；3) 462 个常见的 HOI 类别 (记为 Non-Rare)。

3.2 实验细节

本文提出的模型利用公开的 PyTorch 框架编程，基于 Detectron2 和带有特征金字塔[35] 的 ResNet-50[36] 构建。训练阶段，本文采用 Detectron2 在 COCO 上的预训练参数来初始化模型，并采用了双阶段的训练方式，首先在 V-COCO 的训练集上对 IIPN 进行 10K 次的迭代训练，对于 IIPN 的输出，保留$s_{\mathrm{h}} > 0.8$的人边界框和$s_{\mathrm{o}} > 0.3$的物边界框；接着在 HICO-DET 的训练集上对整体模型进行100K次的迭代训练。最后分别在它们的测试集上进行测试。模型训练使用一张 NVIDIA RTX 1070 GPU，一个批次训练两张图片，采用 SGD 方法训练模型，初始学习率为 0.005，每 10 K次迭代学习率降低 0.0001，动量设为 0.9。

3.3 实验结果

本小节中，我们将所提方法与数个现有的方法进行对比评估。数据显示，本文的模型在 HICO-DET 上的三种测试方法均取得了最优的效果。如下分别在表 1 和表 2 中展示了在 HICO-DET 和 V-COCO 上的对比结果。

如表 1 所示，对于 HICO-DET 数据集，本文方

表 1　不同方法在 HICO-DET 测试集上的效果对比

Table 1　Experimental results on HICO-DET test set of different approaches

Method	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
Shen et al.[27]	6.46	4.24	7.12	-	-	-
HO-RCNN[9]	7.81	5.37	8.54	10.41	8.94	10.85
iCAN[8]	14.84	10.45	16.15	16.26	11.33	17.73
RPNN[11]	17.35	12.78	18.71	-	-	-
PMFNet[31]	17.46	15.65	18.00	20.34	17.47	21.20
DRG[7]	19.26	**17.74**	19.71	**23.40**	21.75	23.89
Peyre et al.[32]	19.40	14.60	20.90	-	-	-
VCL[10]	19.43	16.55	20.29	22.00	19.09	22.87
BaseLine	19.05	14.72	20.35	21.22	18.06	22.88
IIPN(Ours)	**19.90**	15.84	**21.12**	23.13	**19.08**	**24.33**

表 2　不同方法在 V-COCO 测试集上的效果对比

Table 2　Experimental results on V-COCO test set of different approaches

Method	$\mathrm{mAPa}_{\mathrm{role}}$
Gupta et al.[39]	31.8
GPNN[24]	44.0
iCAN[8]	45.3
RPNN[11]	47.5
VCL[10]	48.3
PMFNet w/o human pose[31]	48.6
PMFNet[31]	52.0
IIPN(Ours)	**50.3**

法超过了 PMFNet[31]，一种利用了人身体姿势特征来检测 HOI 的方法，然而本文没有利用这种额外特征。同时，本文方法还超过了 Pryre 等人[32] 所提出的一种学习单独的词嵌入结合相似对象之间类比，用以检测人-物交互的方法。更值得一提的是，该项工作的目的是改进对 Rare 模式下的模型精度，然而无论是在 Default 模式或是 Known Object 模式下，本文的模型在 Rare 下的测试结果都优于该方法。最后，本文所提出的方法在 Default 模式和 Known Object 模式下的三种测试结果也均超过了本文的基线方法。其中，在 Default 模式下比基线分别提高了+0.85 (4.5%)，+1.12 (7.6%)，+0.77 (3.8%)，在 Known Object 模式下比基线分别提高了+1.91 (9%)，+1.02 (5.7%)，+1.45 (6.3%)。

对于 V-COCO 数据集，本文方法也取得了不俗的效果。正如表 2 所示，本文所提出的方法效果超过了除 PMFNet[31] 之外的所有方法，这是由于在 V-COCO 数据集上动词较少，姿势特征对于检测结果的提升有较大的影响。同时，PMFNet[31] 在论文中给出了去除姿势信息后的模型精度为 48.6，比本文的方法低了 1.7，充分说明了本文方法的优越性。

3.4 消融实验

在 2.2 节中已详细介绍了所提出的交互实例推荐网络，利用图神经网络来推理人-物交互关系，利用注意力使得网络更加聚焦。本文分别对比了基线网络，基线网络+移除注意力的图神经网络以及基线网络+IIPN 的效果。此外，在 2.3 节中介绍了跨模态信息融合模块 (CIFM)，本文分别对比了基线网络，基线网络+移除注意力的 CIFM 和基线网络+完整 CIFM。具体数据如表 3 所示，方便起见，在表中将 without 写做 w/o。同时，为了更直观地显示本文模型的实验数据，只展示了 Default 模式和 Known Object 模式在 Full 下的测试结果。

3.4.1 有/无 IIPN 对比

IIPN 是为了能推荐出存在交互关系的人-物对。根据表 3 中的数据显示，在不加注意力的情况下，本文的 IIPN 在 Default 模式下，比基线网络提高了 0.28 (1.5%)，加上了注意力之后，模型精度提升到了 19.72，比基线网络提升了 0.67，提升百分比达到了 3.5%，这充分显示了本文所提出的 IIPN 的有效性。通过将人-物构建为图模型，IIPN 学习了人-物之间的隐含交互关系，并利用注意力进一步强化学习到了这种关系，最终推理得出了存在交互关系的人框和物框，提高了模型的检测精度。

3.4.2 有/无 CIFM 效果对比

CIFM 是为了计算不同特征对 HOI 检测结果的贡献程度。根据表 3 的数据，基线网络+移除注意力计算的 CIFM 检测精度为 19.26，只比基线网络提高了 0.21 (1.1%)。然而，引入注意力之后，对基线网络的提升达到了 0.55(2.9%)，这印证了 CIFM 挖掘到了各特征对检测结果的影响方式，并充分证明了 CIFM 在评估各种特征对 HOI 检测影响程度的有效性。

3.4.3 M和N数量的对比分析

在 IIPN 中，本文分别选择了M个人的隐藏层特征和N个物的隐藏层特征，因此本小节对M和N的不同取值所得到的效果进行了验证，结果如表 4 所示。

对于M的选择，本文参考了 Wang 等人[28] 中的参数设置为 8。同时，本文也尝试了将M设置为 6，数据显示M为 6 的情况总体比M为 8 下降了一个台阶。经过对数据集的分析，认为这是由于在数据集中，有相当一部分的图片中是人群密集的，将人隐藏层特征的数量设置的稍大有助于模型去适应这一部分数据集。此外，还分别尝试了将N设置为 2, 3, 4，在完整模型的实验下，发现将N设置为 3 时模型的表现达到了最

表 3 本文所提各模块在 HICO-DET 上的消融实验

Table 3 Ablation studies of the proposed module on HICO-DET

Method	Default(Full)	Known Object(Full)
BaseLine	19.05	21.22
BaseLine + IIPN w/o attention	19.33	21.92
BaseLine + IIPN with attention	19.72	22.33
BaseLine + CIFM w/o attention	19.26	21.74
BaseLine + CIFM with attention	19.60	21.89
BaseLine + IIPN + CIFM (Ours)	19.90	23.13

表 4 不同的 *M* 和 *N* 对实验结果的影响

Table 4 Effects of different *M* and *N* on experimental results

	Default		
	Full	Rare	Non-Rare
M=6, *N*=2	19.53	15.12	20.43
M=6, *N*=3	19.77	15.74	20.99
M=6, *N*=4	19.56	15.71	20.03
M=8, *N*=2	19.66	15.67	20.86
M=8, *N*=3	19.90	15.84	21.12
M=8, *N*=4	19.60	15.70	20.76

好的水平。这或许和人不能同时和太多的物体交互有关，3 种已经是极限。

3.4.4 不同特征的对比

在本文所提出的跨模态信息融合模块中，用到了人的外观特征 f_h，物的外观特征 f_o，联合空间特征 f_u，以及语义特征 f_{sem}，本文也分别进行了实验，以证明各个特征对最终检测结果的有效性。结果如表 5 所示。

通过分析表 5，可以清晰地显示各个特征流的有效性。特别地，发现当仅融入 f_h+f_o+f_u 时，模型的检测结果为 19.63，将SP替换为 f_u 后，模型精度达到了 19.78，提升了 0.15。这恰好显示了 f_u 的有效性。

3.4.5 不同信息传递轮次*k*的对比

在本文的图神经网络信息传递时，尽管本文遵循了其他作者[34]的设置，我们仍然决定进行一组对比实验，以确定不同的*k*对本文实验结果有何影响。实验结果如表 6 所示。通过表 6 可以发现，当$k=2$时，本文的实验效果在 V-COCO 和 HICO-DET 数据集均取得了最好的效果。

3.5 可视化展示

由于本文方法基于 Faster-RCNN[1]，我们对 IIPN 的人-物对推荐效果和 Faster-RCNN 进行了对比，并挑选了部分图片进行可视化展示。如图 5 所示，其中第一行为 Faster-RCNN 的效果展示，第二行为 IIPN 的效果展示。通过图 5 可以清晰地看到，Faster-RCNN 对不相关的人或物也进行了推荐，而 IIPN 只推荐了存在交互关系的人-物对。在图 6 和图 7 中，本文对模型的最终检测结果进行了可视化来直观地感受模型的检测效果。图 6 中将 HICO-DET 数据集上的部分检测结果进行了可视化展示，其中人用红色方框框出，物体用蓝色方框框出，其中分别展示了简单场景下和复杂场景下，本文所提出模型的检测效果。图 7 则将 CIFM 的融合注意力进行了可视化展示，其中每幅图片右侧数据从上到下分别为：f_h、f_o、f_u、f_{sem} 对检测结果的贡献分数。

表 5 不同特征对实验结果的影响

Table 5 Influence of different characteristics on experimental results

f_h	f_o	SP	f_u	f_{sem}	mAP
✓	✗	✗	✗	✗	19.03
✓	✓	✗	✗	✗	19.35
✓	✓	✓	✗	✗	19.63
✓	✓	✗	✓	✗	19.78
✓	✓	✗	✓	✓	19.90

表 6 不同迭代次数 *k* 对实验结果的影响

Table 6 Influence of different *k* on experimental results

Number of *k*	Result on V-COCO	Result on HICO-DET
k=1	49.4	19.81
k=2	**50.3**	**19.90**
k=3	49.9	19.85

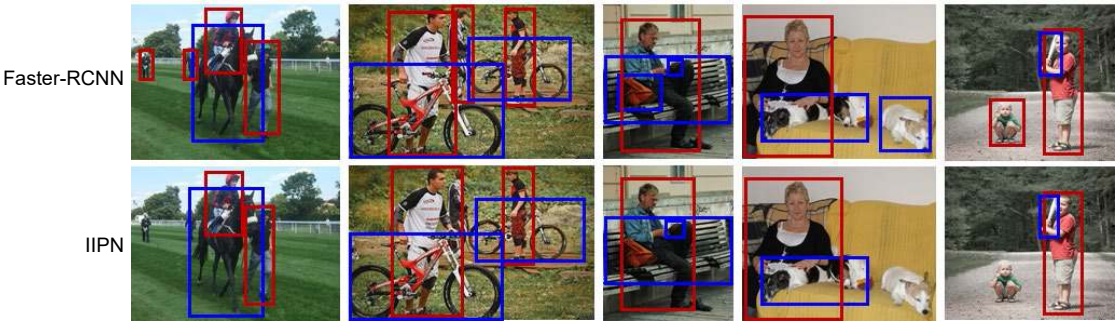


图 5 IIPN 与 Faster-RCNN 的人-物推荐效果展示

Fig. 5 Comparison of IIPN (bottom row) with Faster-RCNN (upper row)

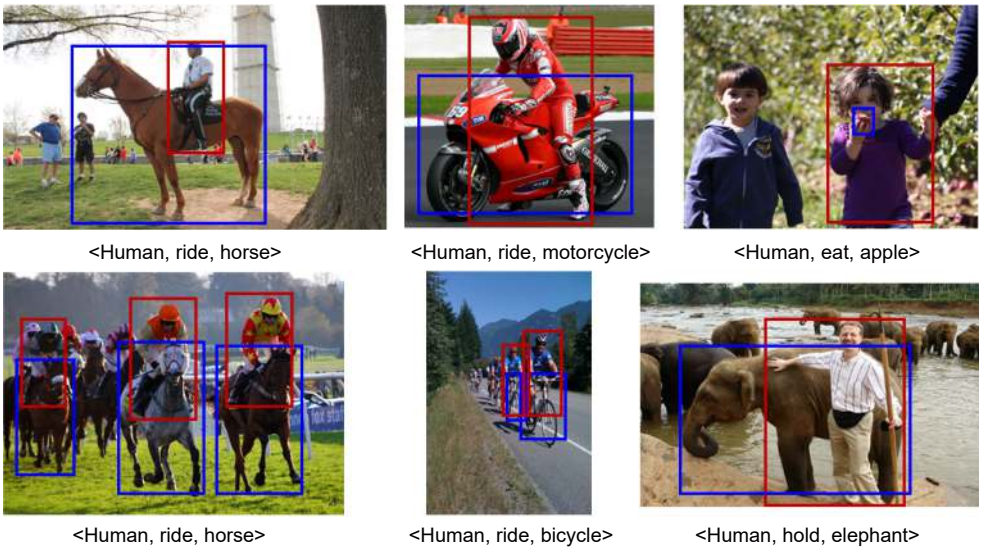


图 6 HICO-DET 数据集上的检测结果可视化

Fig. 6 Visualization of detection results on HICO-DET

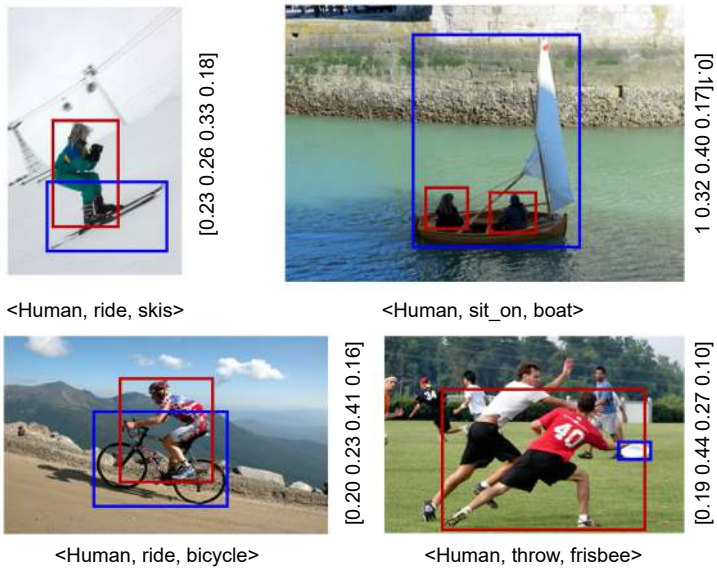


图 7 融合注意力可视化展示

Fig. 7 Visualization of fusion attention

4 结 论

本文提出了一种全新的双阶段人-物交互检测模型，首先利用交互实例推荐网络 (IIPN) 来推荐存在交互关系的人-物对。IIPN 根据视觉特征以及图像中人-物之间的交互关系进行交互对推荐。在实验中证明了 IIPN 能够推荐出正确的交互对来提高检测效果。此外，本文设计了一个跨模态信息融合模块 (CIFM)，通过引入融合注意力，来动态计算各种特征对检测结果的影响程度，本文的实验证明了该模块的有效性。最后，在两个流行的大型数据集上，本文所提出的方法都取得了不俗的效果。

参考文献

[1] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015: 91–99.

[2] Girshick R. Fast R-CNN[C]//*Proceedings of the IEEE International Conference on Computer Vision*, 2015: 1440–1448.

[3] Yang C Y, Xu Y H, Shi J P, et al. Temporal pyramid network for action recognition[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 588–597.

[4] Li M S, Chen S H, Chen X, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 3590–3598.

[5] Kirillov A, He K M, Girshick R, et al. Panoptic segmentation[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 9396–9405.

[6] Sofiiuk K, Sofiyuk K, Barinova O, et al. AdaptIS: adaptive instance selection network[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 7354–7362.

[7] Gao C, Xu J R, Zou Y L, et al. DRG: dual relation graph for human-object interaction detection[C]//*16th European Conference on Computer Vision*, 2020: 696–712.

[8] Gao C, Zou Y L, Huang J B. iCAN: instance-centric attention network for human-object interaction detection[C]//*British Machine Vision Conference 2018*, 2018.

[9] Chao Y W, Liu Y F, Liu X Y, et al. Learning to detect human-object interactions[C]//*2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018: 381–389.

[10] Hou Z, Peng X J, Qiao Y, et al. Visual compositional learning for human-object interaction detection[C]//*16th European Conference on Computer Vision*, 2020: 584–600.

[11] Zhou P H, Chi M M. Relation parsing neural network for human-object interaction detection[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 843–851.

[12] Kim B, Lee J, Kang J, et al. HOTR: end-to-end human-object interaction detection with transformers[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 74–83.

[13] Zhang A X, Liao Y, Liu S, et al. Mining the benefits of two-stage and one-stage HOI detection[C]//*Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[14] Zou C, Wang B H, Hu Y, et al. End-to-end human object interaction detection with HOI transformer[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 11820–11829.

[15] Chen M F, Liao Y, Liu S, et al. Reformulating HOI detection as adaptive set prediction[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 9000–9009.

[16] Kamath A, Clark C, Gupta T, et al. Webly supervised concept expansion for general purpose vision models[Z]. arXiv: 2202.02317, 2022. https://arxiv.org/abs/2202.02317v1.

[17] Li Z M, Zou C, Zhao Y, et al. Improving human-object interaction detection via phrase learning and label composition [Z]. arXiv: 2112.07383, 2021. https://doi.org/10.48550/arXiv.2112.07383.

[18] Xue L X, Jiang D, Wang R G, et al. Learning semantic dependencies with channel correlation for multi-label classification[J]. *Vis Comput*, 2020, **36**(7): 1325−1335.

[19] Li Y L, Zhou S Y, Huang X J, et al. Transferable interactiveness knowledge for human-object interaction detection[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 3580–3589.

[20] Yang J W, Lu J S, Lee S, et al. Graph R-CNN for scene graph generation[C]//*Proceedings of the 15th European Conference on Computer Vision (ECCV)*, 2018: 690–706.

[21] Chen T S, Yu W H, Chen R Q, et al. Knowledge-embedded routing network for scene graph generation[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 6156–6164.

[22] Yang J, Sun X, Wang R G, et al. PTPGC: pedestrian trajectory prediction by graph attention network with ConvLSTM[J]. *Robot Auton Syst*, 2022, **148**: 103931.

[23] Liang W X, Jiang Y H, Liu Z X. GraghVQA: language-guided graph neural networks for graph-based visual question answering[Z]. arXiv: 2104.10283, 2021. https://arxiv.org/abs/2104.10283v2.

[24] Qi S Y, Wang W G, Jia B X, et al. Learning human-object interactions by graph parsing neural networks[C]//*Proceedings of the 15th European Conference on Computer Vision (ECCV)*, 2018: 407–423.

[25] Xu B J, Wong Y K, Li J N, et al. Learning to detect human-object interactions with knowledge[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 2019–2028.

[26] Zheng S P, Chen S Z, Jin Q. Skeleton-based interactive graph network for human object interaction detection[C]//*2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020: 1–6.

[27] Shen L Y, Yeung S, Hoffman J, et al. Scaling human-object interaction recognition through zero-shot learning[C]//*2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018: 1568–1576.

[28] Wang S C, Yap K H, Yuan J S, et al. Discovering human interactions with novel objects via zero-shot learning[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 11649–11658.

[29] Fang H S, Xie Y C, Shao D, et al. DecAug: augmenting HOI detection via decomposition[C]//*Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021: 1300–1308.

[30] Sarullo A, Mu T T. Zero-shot human-object interaction recognition via affordance graphs[Z]. arXiv: 2009.01039, 2020. https://doi.org/10.48550/arXiv.2009.01039.

[31] Wan B, Zhou D S, Liu Y F, et al. Pose-aware multi-level feature network for human object interaction detection[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 9468–9477.

[32] Peyre J, Sivic J, Laptev I, et al. Detecting unseen visual relations using analogies[C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 1981–1990.

[33] Liu Y, Chen Q C, Zisserman A. Amplifying key cues for human-object-interaction detection[C]//*16th European Conference on Computer Vision*, 2020: 248–265.

[34] Zhang F Z, Campbell D, Gould S. Spatio-attentive graphs for human-object interaction detection[Z]. arXiv: 2012.06060, 2020. https://arxiv.org/abs/2012.06060v1.

[35] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 936–944.

[36] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[37] Chen L, Zhang H W, Xiao J, et al. Zero-shot visual recognition using semantics-preserving adversarial embedding networks[C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1043–1052.

[38] Pennington J, Socher R, Manning C D. GloVe: global vectors for word representation[C]//*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014: 1532–1543.

[39] Gupta S, Malik J. Visual semantic role labeling[Z]. arXiv: 1505.04474, 2015. https://arxiv.org/abs/1505.04474v1.

[40] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]//*13th European Conference on Computer Vision*, 2014: 740–755.

作者简介

薛丽霞 (1976-)，女，博士，副教授，硕士生导师，主要从事神经网络与深度学习技术、智能视频图像处理与分析等的研究。

E-mail: 51003239@qq.com

尹凯建 (1997-)，男，合肥工业大学在读研究生，本科毕业于江苏科技大学，主要研究方向是计算机视觉和深度学习。

E-mail: kajan_ykj@foxmail.com

汪荣贵 (1966-)，男，博士，教授，博士生导师，主要从事深度学习理论与应用、视频大数据与云计算、智能视频监控与公共安全、嵌入式多媒体技术等领域的研究，主持完成国家自然科学基金面上项目等多个纵向研究课题。

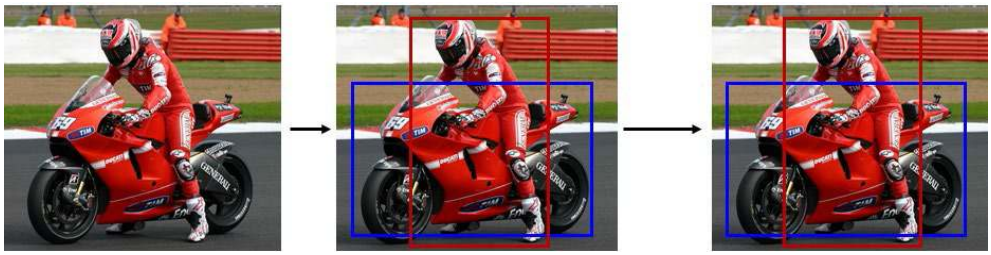
E-mail: wangrgui@hfut.edu.cn

【通信作者】杨娟 (1983-)，女，博士，讲师，硕士生导师，主要从事神经网络与深度学习技术、智能视频图像处理与分析、WEB 数据智能分析软件的研究与实现等的研究。

E-mail: yangjuan6985@163.com

Interactive instance proposal network for HOI detection

Xue Lixia, Yin Kaijian, Wang Ronggui, Yang Juan*



Pipeline of human-object interaction detection

Overview: With the development of computer vision, people increasingly need to understand images, including recognizing the scenes and the human behaviors in images. The task of HOI detection is to locate humans and objects in images and infer their relationships. This requires not only locating a single object instance, but also identifying the interaction between the objects. However, machines cannot know which object humans are interacts in. Most of the existing methods solve this problem by completely pairing the people and objects. They use off-the-shelf object detectors to detect instances, but this does not meet the requirements of the HOI task. This paper proposes an object detector suitable for HOI detection based on relational reasoning, which makes use of the interactive relationship between humans and objects in the images to recommend human-object pairs, so as to reduce the occurrence of non-interactive human-object pairs as much as possible. Our method follows the two-stage detection like most works. Firstly, the interactive instance proposal network (IIPN) is used to recommend human-object pairs. The IIPN follows the pipeline of faster RCNN, but replaces the region proposal network (RPN) with the IIPN. The IIPN selects human-object pairs based on the interaction possibility between humans and objects using the visual information in the picture. It passes the message through the iterative reasoning of the graph neural networks (GNNS), only human-object pairs that include interactive relationships are selected as the IIPN's outputs. Secondly, we design a cross-modal information fusion module (CIFM), which calculates the fusion attention according to the influence of different features on the detection results, and performs weighted fusion. This is because the existing methods simply add or splice several features such as human visual features, object visual features, and human-object spatial features in the reasoning part. The different influence degrees of various features in different actions are ignored. For example, the verbs like ride and hold in < human, ride bike> and < human, hold, bike > depend more on the spatial relationships, while eat and cut in <human, eat, pizza> and <human, cut, pizza> depend more on human's postures, that is, visual features. Meanwhile, this paper believes that semantic prior knowledge is also helpful to HOI detection. For example, if we have apples in an image, the probability of predicting the human's action as eating or holding is greater than others. Finally, complete experiments are performed on two popular large-scale HOI datasets, HICO-DET and V-COCO. The experimental results show the effectiveness of the proposed method.

Xue L X, Yin K J, Wang R G, et al. Interactive instance proposal network for HOI detection[J]. *Opto-Electron Eng*, 2022, **49**(7): 210429; DOI: 10.12086/oee.2022.210429

Foundation item: the National Key Research and Development Program of China (2020YFC1512601) and National Natural Science Foundation of China (62106064)

School of Computer and Information, Hefei University of Technology, Hefei, Anhui 230031, China

* E-mail: yangjuan6985@163.com