计算机科学
COMPUTER SCIENCE

PKHOI:利用先验知识增强人-物交互检测算法

赵文豪, 梅萌, 王小平, 罗航宇

引用本文

赵文豪, 梅萌, 王小平, 罗航宇. PKHOI:利用先验知识增强人-物交互检测算法[J]. 计算机科学, 2026, 53(1): 141-152.

ZHAO Wenhao, MEI Meng, WANG Xiaoping, LUO Hangyu. PKHOI:Enhancing Human-Object Interaction Detection Algorithms with Prior Knowledge [J]. Computer Science, 2026, 53(1): 141-152.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250100086

PKHOI:利用先验知识增强人-物交互检测算法

赵文豪　梅　萌　王小平　罗航宇
同济大学计算机科学与技术学院　上海 200092
(zwh1625@tongji.edu.cn)

摘　要　人-物交互检测(Human-Object Interaction,HOI)在视觉场景理解中起着至关重要的作用,随着深度学习技术的发展,基于视觉的交互检测模型已经能够获得良好的性能。然而,现有方法大多缺乏对先验的逻辑知识的运用,有时会推导出不合理的结果。其次,一些方法将空间信息和人体姿态信息用于推理,但它们仅在推理结果和标注之间构造损失,导致解码器无法学习到准确的隐含关系。因此,提出一种利用先验知识增强现有人-物交互检测算法的方法PKHOI,该方法能够有效增强现有人-物交互检测算法的准确性。具体而言,从训练集中构建了一个包含物品功能性、空间关系、人体姿态和动词共现的逻辑规则表,将其形式化为一阶逻辑并映射到连续空间中,在训练阶段和推理阶段分别以损失函数和矩阵乘法的形式将先验的逻辑规则融入神经网络,提升模型的准确性。此外,提出一种通过融合多模态信息(空间、语义和人体姿态信息)生成人-物对查询的方法,结合逻辑损失函数,可以引导解码器学习到更多的隐含知识。利用提出的方法增强了两个主流的人-物交互检测算法UPT和PViC,并在V-COCO,HICO-DET和Flickr30k数据集上进行了评估,实验结果表明,提出的方法可以有效提高现有方法的性能。

关键词:人-物交互检测;先验知识;一阶逻辑;姿态信息;多模态信息融合
中图分类号　TP391

PKHOI:Enhancing Human-Object Interaction Detection Algorithms with Prior Knowledge

ZHAO Wenhao,MEI Meng,WANG Xiaoping and LUO Hangyu
College of Computer Science and Technology,Tongji University,Shanghai 200092,China

Abstract　HOI detection plays a crucial role in visual scene understanding. With the advancement of deep learning technologies, vision-based interaction detection models have achieved promising performance. However, most existing methods lack the utilization of prior logical knowledge, sometimes leading to unreasonable predictions. Additionally, while some methods employ spatial information and human pose information for reasoning, they only construct losses between inference results and annotations, preventing decoders from learning accurate implicit relationships. Therefore, this paper proposes the PKHOI method, which enhances existing HOI detection algorithms by leveraging prior knowledge, effectively improving the accuracy of current HOI detection algorithms. Specifically, it constructs a logical rule table from the training set, encompassing object functionality, spatial relationships, human poses, and verb co-occurrence. These rules are transformed into first-order logic and mapped to continuous space. The prior logical rules are then incorporated into neural networks through loss functions during training and matrix multiplication during inference, enhancing model accuracy. Furthermore, this paper proposes a method to generate human-object pair queries by fusing multimodal information (spatial, semantic, and human pose information). Combined with logical loss functions, this approach guides the decoder to learn more implicit knowledge. The proposed method enhances two mainstream HOI detection algorithms, UPT and PViC, and evaluates them on V-COCO, HICO-DET, and Flickr30k datasets. Experimental results demonstrate that the proposed method can effectively improve the performance of existing approaches.

Keywords　Human-object interaction detection, Prior knowledge, First-order logic, Pose information, Multi-modal information fusion

1　引言

人-物交互检测作为一项重要的计算机视觉任务,致力于定位图像中的人-物对〈人类,物体〉并识别它们之间的交互动作。HOI检测系统执行以人为中心的场景理解,在现实世界有着广泛的应用。例如,人机交互、增强/虚拟现实(AR/VR)

到稿日期:2025-01-14　返修日期:2025-03-30
基金项目:国家重点研发计划(2022YFB4300504-4);上海市经济和信息化委员会专项基金(202201034)
This work was supported by the National Key Research and Development Program of China(2022YFB4300504-4)and Special Fund Project supported by Shanghai Municipal Commission of Economy and Information Technology(202201034).
通信作者:王小平(xpwang6510@tongji.edu.cn)

和视频监控等。

本文专注于图像领域的 HOI 检测，根据架构的不同，现有的 HOI 检测器可以分为单阶段和双阶段方法。为了提高检测效率，单阶段方法[1-7]直接使用联合区域或交互点来检测人-物对，同时并行识别交互动作的类别，这通常是一个端到端的过程。然而在单阶段方法中，模型收敛十分具有挑战性，由于三元组标签提供的监督信息较为稀疏，并且实体检测器和交互头需要联合训练，因此模型的收敛通常需要数百个 GPU 小时[6]。与此相比，双阶段方法[8-21]通过使用具有预训练权重的实体检测器[8-9]来定位图片中的人和物体，然后使用交互头预测人-物之间的交互动作。这种方法专注于训练交互头，从而加速了收敛过程。此外，双阶段检测器能够遍历所有潜在的人-物交互对，在视觉推理场景中针对特定交互对的识别具有很强的灵活性。

然而，现有的检测器在实际应用中仍面临着一系列挑战。首先，现有方法缺乏对数据集中先验知识的利用，完全依靠数据驱动，这种方法可能导致不合理的推理结果[15]。例如，当检测到的人体姿态和物体类别与可能的交互动作存在明显矛盾时，网络仍然可能输出违反常识的交互动作，缺乏先验的逻辑约束限制了模型的可解释性和鲁棒性。其次，一些方法试图结合空间信息[10-11]和人体姿态信息[12-13]进行推理，但它们仅在推理结果与标注信息之间构建损失来使推理结果与标注对齐，试图让解码器能借此学习到各种信息与推理结果之间的关系，但这种隐式的学习方式难以确保模型充分理解多模态信息与交互动作之间的复杂关联。这种设计使得模型无法充分利用这些多模态信息（如空间、语义和人体姿态信息），从而影响推理的准确性和合理性。

针对上述问题，本文提出了一种利用先验的逻辑知识增强现有人-物交互检测算法的方法 PKHOI，其与现有技术存在 3 个核心差异：1）区别于 UPT[17]等纯数据驱动方法，本文首次将逻辑规则系统引入 HOI 检测框架，通过一阶逻辑的连续空间映射实现常识约束的数学表达；2）与 PViC[18]等视觉组合性方法相比，本文提出的多模态查询生成机制实现了空间、语义和姿态信息的动态融合，而非简单的特征拼接；3）创新性地设计了“训练时逻辑监督＋推理时规则抑制”的双重约束机制，在保持端到端训练优势的同时提升了预测结果的合理性。

具体而言，本文从训练集中构建了一个包含物品功能性、空间关系、人体姿态和动词共现性的逻辑规则表，将其转换为一阶逻辑并映射到连续空间中，在训练阶段将先验的逻辑知识以损失函数的形式融入交互解码器，从而提升推理结果的准确性。在推理阶段，基于逻辑规则表对互斥的推理结果进行抑制，只需要简单快速的矩阵乘法即可提升模型的准确性。此外，本文提出了一种通过融合多模态信息（空间、语义和人体姿态信息）生成人-物对查询的方法，该方法可以引导解码器的注意力集中于信息更丰富的区域，并且结合逻辑损失函数，可以有效地构建出不同信息与推理结果之间的联系，引导解码器学习到更多的隐含知识。

本文的主要贡献如下：

1）提出了一种利用先验的逻辑知识增强现有人-物交互检测算法的方法 PKHOI，该方法在训练阶段和推理阶段以逻辑损失和矩阵乘法的形式将逻辑知识注入神经网络，通过双重约束机制，在保持模型端到端可训练性的同时实现常识约束。相较于 UPT[17]等纯数据驱动方法，该方法在推理合理性和可解释性方面取得了提升。

2）提出了一种通过融合多模态信息来生成人-物对查询的方法，可以引导解码器的注意力关注信息更丰富的区域，并且该方法与逻辑损失函数相结合，可以有效地构建出各模态信息和推理结果之间的联系，这使得交互解码器能够有效地学习到更多的隐含知识。

3）通过大量的实验验证，PKHOI 可以有效地提高现有方法在广泛使用的数据集上的准确性。在 HICO-DET[22]数据集上，PKHOI 在使用 ResNet-50 和 Swin-Large 作为骨干网络时，mAP 分别达到了 35.44%和 45.79%。

2 相关工作

2.1 人-物交互检测

2.1.1 单阶段方法

单阶段检测器旨在通过单次前向传播来定位图片中的人和物，并且同时完成人-物之间交互动作类型的识别。这些方法通常利用预定义的交互点或锚点[1-5]。例如，PPDM[3]提出了一种简单而有效的策略，将人与物体之间的中点作为交互点。QAHOI[4]提出了基于查询的锚点方法，通过可变形 Transformer 解码器生成参考点。FGAHOI[5]进一步丰富了这一概念，通过生成细粒度的锚点来指导复杂任务中的 HOI 特征提取。除了锚点策略外，一些研究还专注于在单阶段方法中优化检测过程。例如，ERNet[6]在其分类头中集成了预测不确定性估计框架，从而提高了预测的鲁棒性。CDN[7]则以级联方式解耦人-物检测和交互分类，将两阶段检测器的优势引入单阶段方法。

2.1.2 两阶段方法

在近期的研究中，两阶段方法的优势逐渐显现。两阶段方法将 HOI 检测任务分解为实体检测和交互推理，显著提升了效率和灵活性。两阶段方法利用现有的实体检测器[8-9]获取检测结果，并通过有效利用人、物查询或结合额外信息（如空间[10-11]、姿态[12-13]和图特征[14-15]）来丰富 HOI 交互特征，同时引入以实例为中心的注意力机制[16-18]和语言线索[19]。例如，iCAN[16]提出了以实例为中心的注意力网络，利用上下文图像特征增强人-物对的表示。UPT[17]指出，对一元特征和交互对使用自注意力机制能有效提升正样本的置信度。PViC[18]在 UPT[17]的基础上通过交叉注意力将图像特征重新引入人-物对表示中，以弥补相关上下文信息的不足。RLIPv1[19]结合了语言特征，提升了少样本 HOI 检测的效果，展示了对比视觉-语言预训练的有效性。此外，一些研究通过数据增强[20-21]来解决标注稀疏的问题。例如，RLIPv2[20]通过构建平衡数据集，为上万张图像添加 HOI 三元组标注，显著提升了数据多样性。

2.2 先验知识

先验通常是指在模型训练之前已经存在的、有助于模型学习的知识或信息，一般来自数据本身的结构特征、人类的

直觉、领域知识或历史经验。

现有研究表明[23-25]，结合先验知识可以有效地提高视觉模型的推理能力。例如，Diligenti 等[23]提出了一种通用的整合先验知识的方法，该方法利用基于语义的正则化来表示先验知识，将这些知识以一阶逻辑（First-Order Logic，FOL）的形式表达，并将其转换为一组约束，通过反向传播整合到学习过程中，可以有效提升神经网络在图像分类任务上的准确性。还有一些研究[26-28]利用知识图谱（Knowledge Graphs，KGs）来改进视觉推理任务，知识图谱提供了物体、动作和关系等语义先验知识，能够帮助视觉模型进行推理，提升模型在复杂场景下的性能。除此之外，一些 HOI 检测算法[29-31]尝试引入先验知识。例如，KGT[29]利用先验知识指导的三分支 Transformer 来进行 HOI 检测，该方法设计了基础三分支解码器，分别执行 HOI 检测的子任务。在训练阶段，先验知识网络利用强大的语义和空间先验知识，指导网络生成具有强特征表示能力的解码器输出。GEN-VLKT[30]从 CLIP[32]模型中迁移知识来引导位置嵌入和实体嵌入的生成。

3 本文方法

3.1 整体结构

本文具有两种设置，分别对应两个主流的两阶段人-物交互检测算法 UPT[17]和 PViC[18]，在此基础上实现了所提出的方法 PKHOI。以 PKHOI（PViC）为例，如图 1 所示，Pairwise Query Qenerator 和 Prior knowledge 是本文所提出的内容。模型整体可以分为两个阶段：一阶段完成对图片中人体、物体的识别，二阶段使用一系列后处理步骤生成若干个人-物对，然后完成人-物之间交互类型的识别。

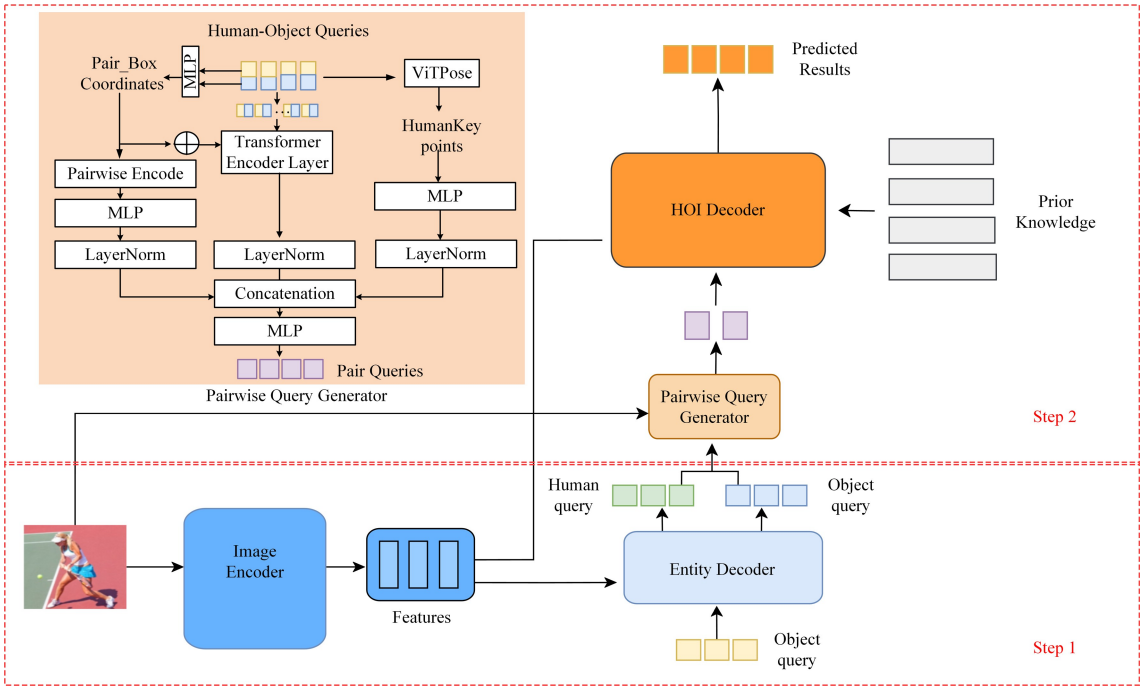


图 1 PKHOI 整体结构

Fig. 1 Overall structure of PKHOI

一阶段是目标检测阶段，主体是 Transformer 编解码器，使用预训练的 DETR[8]模型进行参数初始化，并在对应的人-物交互检测数据集[22-33]上进行了微调，输出人体和物体的特征。

二阶段是人-物之间交互动作的推理阶段，主要包含成对查询生成器和交互解码器。成对查询生成器通过一系列后处理步骤筛选出符合要求的人、物实体输出，这些实体成对交互组合为若干个人-物对，然后融合空间、语义和人体姿态信息生成新的查询，每一个查询代表一个潜在的人-物交互对，具体见 3.2 节。交互解码器由一个 Transformer 解码器构成，实现对人-物之间的交互关系的分类，具体见 3.3 节。本文所提出的先验逻辑规则详见 3.4 节。在训练阶段和推理阶段，将先验的逻辑规则以损失函数和矩阵乘法的形式融入神经网络，具体见 3.5 节和 3.6 节。

3.2 成对查询生成

在构建人-物对查询之前，通过置信度阈值过滤一阶段检测出的实体，然后将剩下的实体按照不同类别（人物和物体）分类，接着两两组合，枚举出所有的人-物对。

如图 2 所示，对于每一个人-物对，本文先使用分类器得到人和物的坐标，进而将中心坐标、高度和宽度三者的正弦嵌入连接起来，再利用一个 MLP 进行编码，得到人-物对的空间信息。形式上，用 $\phi:\mathbb{R}\rightarrow\mathbb{R}^d$ 表示从标量到正弦嵌入的映射，计算方法如下：

$$\phi(x)=\sin\left(\frac{x}{T}\right) \tag{1}$$

其中，T 是温度参数，这里设置为 1。

对于人-物对对应的人体和物体的特征，先将其拼接，然后执行自注意力以优化实体特征[34]。

对每一个人-物对，将原始图片和人体所在边界框输入一个姿态识别模型 ViTPose[35]中，得到 k 个人体姿态关键点。本文强调人类在互动中的主导作用，通过考虑人体关键点和物体之间的几何关系，可以更精细地表示人体姿态特征。对

于 k 个人体关键点，沿 X 轴和 Y 轴扩展关键点坐标定义出 k 个人体关键部位区域。一个人体关键部位区域由一个中心关键点坐标 (x_i, y_i) 和一个相邻的辅助关键点坐标 (x_j, y_j) 指定。生成的人体关键部位区域边界框的左上角和右下角顶点分别由 $(x_i-\Delta x,\ y_i-\Delta y)$ 和 $(x_i+\Delta x,\ y_i+\Delta y)$ 决定。这意味着这些区域以 (x_i, y_i) 为中心，宽度为 $2\times\Delta x$，高度为 $2\times\Delta y$。其中 Δx 和 Δy 分别代表沿 X 轴和 Y 轴的自适应偏移，计算方法如下：

$$\begin{cases}\Delta x=\left(\alpha+\beta\dfrac{\tau_{i,j}}{1+\tau_{i,j}}\right)\times d_{i,j}\\ \Delta y=\left(\alpha+\beta\dfrac{1}{1+\tau_{i,j}}\right)\times d_{i,j}\end{cases} \tag{2}$$

其中，α,β 是固定的超参数。$\tau_{i,j}$ 和 $d_{i,j}$ 的计算方法分别如下：

$$\tau_{i,j}=\frac{|x_i-x_j|}{|y_i-y_j|} \tag{3}$$

$$d_{i,j}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \tag{4}$$

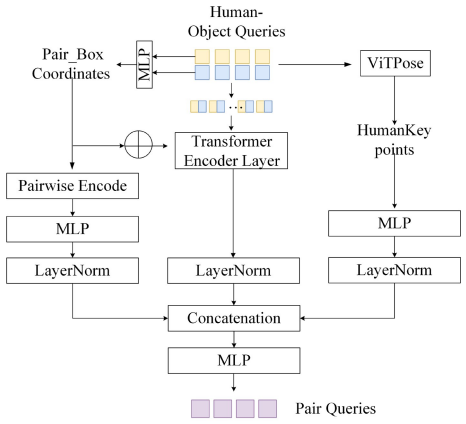


图 2 成对查询生成器

Fig. 2 Pairwise query generation

生成人体关键部位区域后，通过计算所有人体部位与检测到的物体之间的人体部位区域交集（Intersection over Body-part Area，IOB）来衡量每个人体部位在互动中的贡献。当人类执行“握”“拉”和“挥”等动作时，手部的 IOB 显著增加，而在执行“看”和“飞”等远距离互动中，所有 IOB 都趋近于零。这表明，IOB 可以捕捉不同的行为模式。如图 3 所示，第一行展示了 ViTPose[35] 检测到的人体关键点和物体边界框，第二行显示了生成的人体关键部位区域。

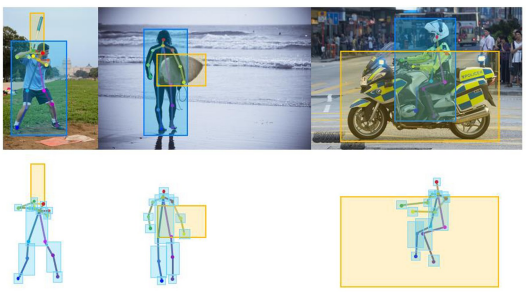


图 3 人体关键部位示意图

Fig. 3 Schematic diagram of body key parts

接着，连接（Concat）人体姿态关键点（中心点坐标）、关键部位区域（中心点坐标与偏移量）和对应的 IOB 为单个向量，再经过一个 MLP 处理得到最终的人体姿态信息。

最后，对得到的空间信息、语义信息和人体姿态信息应用一次 LayerNorm[36]，来稳定训练过程并防止数值溢出，再将 3 个模态的信息拼接并输入 MLP，得到更新的人-物对查询。

3.3 交互解码器

以 PKHOI（PViC）为例，交互解码器使用 Transformer 解码器将生成的人-物对查询作为 query，第一阶段中将编码器的输出特征作为全局特征输入 cross-attention 中，并且使用成对查询生成器中生成的空间信息作为位置嵌入。尽管生成的人-物对表示已经包含了空间先验知识，但位置嵌入仍十分重要，因为它在注意力权重上起到了空间偏差的作用。

本文使用人-物对的中心坐标和人体姿态关键点的中心坐标拼接得到位置嵌入，将键和查询分别表示为 k_c 和 q_c，它们各自的位置嵌入表示为 k_p 和 q_p。为简化，省略线性变换和归一化。注意力的点积计算式如下：

$$(k_c+k_p)^{\mathrm{T}}\cdot(q_c+q_p)=k_c^{\mathrm{T}}q_c+\cdots+k_p^{\mathrm{T}}q_p \tag{5}$$

右侧第一项 $k_c^{\mathrm{T}}q_c$ 衡量了 key（代表图像特征）和 query（代表内容特征）之间的相似性，而最后一项 $k_p^{\mathrm{T}}q_p$ 则衡量了图像的位置编码与位置嵌入之间的相似性，所以位置嵌入的引入可以让交互解码器的注意力权重产生偏差项 $k_p^{\mathrm{T}}q_p$，经过学习之后就可以使得注意力关注信息更丰富的区域（如人-物接触区域）。具体而言，对于一个具有归一化空间索引 (i,j) 的图像标记和一个具有归一化坐标 (x,y) 的物体中心点，最后一项可以展开为坐标之间相似性的简单求和，其计算式如下：

$$k_p^{\mathrm{T}}q_p=\phi(i)^{\mathrm{T}}\phi(x)+\phi(j)^{\mathrm{T}}\phi(y) \tag{6}$$

再将 q_p 拓展到人、物中心点坐标和 k 个人体关键点坐标，最后得到的注意力偏差为：

$$k_p^{\mathrm{T}}q_p=\sum_{i=1}^{k+2}(\phi(i)^{\mathrm{T}}\phi(x_i)+\phi(j)^{\mathrm{T}}\phi(y_i)) \tag{7}$$

3.4 先验规则的构建

从数据集中构建了 4 种先验的逻辑规则以指导交互解码器的训练和推理，这 4 种逻辑规则分别是物品功能性、空间关系属性、动词共现性和人体姿态属性。尽管目前已经有一些研究[10-13]使用到了这些属性，但是它们通常从统计角度分析可能性，例如计算动作与对象共现的分布来修改预测[37]，或者简单地将位置编码或人体姿态集成到网络特征中[10-13]。而本文从受限子集（推理结果）是预给定对象（物品类别）和条件（空间关系、姿态）的逻辑结果这一角度来实现。以下是 4 种先验逻辑规则的具体解释。

物品功能性：当物品种类确定时，可能出现的动词种类也就确定了，其他的动词均与当前物品种类互斥，如物品类别“书”与动词“骑”互斥。

空间关系属性：当人-物之间呈现某种相对位置关系时，可能出现的动词种类是确定的。本文以人为参考，定义了 5 种相对位置关系，分别是上方（如风筝在人的上方）、下方（如滑板在人的下方）、周围（如长颈鹿在人的周围）、内部（如手提

包在人的内部）、包含（如公交车包含人）。

动词共现性：由于人-物交互检测是一个多分类的任务，对于同一个人-物对，当某个动词出现时，有些动词会同时出现，而另一些动词则与之互斥，不同时出现。比如对于摩托车，“坐在上方”是伴随着“骑乘”同时发生的，而此时动作“清洗”不会发生。

人体姿态属性：人-物交互检测是以人体为中心的任务，所以人体的姿态对于交互动作有着指导作用，当人体呈现某个姿态时，一些动词是与之互斥的。例如人体姿态为“躺”时，动词“骑乘”不会发生。本文定义了如下姿态类别：躺、骑乘、蹲坐、弯腰和站立。

本文从训练集中构建出先验的逻辑规则，以互斥矩阵的形式保存。以物品功能性为例，算法 1 是构建逻辑规则的算法。

算法 1 逻辑规则构建

输入：训练集标注文件

输出：逻辑规则表

1. 加载标注文件；
2. 初始化互斥矩阵(80,117)；//物品类别 80，动词类别 117
3. for annotation in data['annotation']：
4. objects=annotation['object']；
5. verbs=annotation['verb']；
6. for i in range(len(objects))：
7. obj=objects[i]；
8. verb=verbs[i]；
9. if 索引 obj 和索引 verb 在有效范围(80,117)内：
10. 互斥矩阵[obj,verb]值置为 1；
11. end if
12. end for
13. end for
14. 保存互斥矩阵为逻辑规则表。

同理可以构建出其他的逻辑规则表。对于空间关系和人体姿态，使用算法 2 和算法 3 计算出对应的空间关系类别及姿态类别。

算法 2 空间关系计算

输入：人物坐标，物品坐标

输出：人-物相对空间关系

1. 计算重叠区域的面积、人的区域面积、物体的区域面积；
2. 计算人和物体的中心点坐标；
3. 计算重叠区域占人和物体区域的比例；
4. if 人体最低点＞物品最低点 and 人体中心点坐标＞物品最高点 and 人体中心点横坐标不超出物品框宽度：
5. 输出位置关系“在上方”；
6. end if
7. if 人体最高点＜物品最高点 and 人体中心点坐标＜物品最低点 and 人体中心点横坐标不超出物品框宽度：
8. 输出位置关系“在下方”；
9. end if
10. if 物体面积＜人体面积 and 物体重叠比例＞0.9：
11. 输出位置关系“内部”；
12. end if
13. if 物体面积＞人体面积 and 人体重叠比例＞0.9：
14. 输出位置关系“包含”；
15. end if
16. 输出位置关系“环绕”；

算法 3 姿态类别计算

输入：人体姿态关键点坐标

输出：人体姿态类别

1. 计算各个关键点之间的角度；
2. if 45＜左髋-左膝-左踝角度＜135 and 45＜右髋-右膝-右踝角度＜135 and 左肩-左髋-左膝角度＜135 or 右肩-右髋-右膝角度＜135：
3. 输出姿态“骑乘”；
4. end if
5. if 0＜左髋-左膝-左踝角度＜110 and 0＜右髋-右膝-右踝角度＜110 and 左肩-左髋-左膝角度＜110 or 右肩-右髋-右膝角度＜110：
6. 输出姿态“蹲坐”；
7. end if
8. if 70＜左髋-左膝-左踝角度＜110 or 70＜右髋-右膝-右踝角度＜135 and 左肘、右肘纵坐标＜左髋、右髋纵坐标：
9. 输出姿态“弯腰”；
10. end if
11. if 90＜左肩-左髋-左膝角度＜180 or 90＜右肩-右髋-右膝角度＜180 and 左肩和左髋高度差小于阈值 or 右肩和右髋高度差小于阈值：
12. 输出姿态“躺”；
13. end if
14. 输出姿态“站立”；

3.5 先验逻辑监督的训练过程

以空间关系为例，对于一个人-物对 x，给定物品类别 o 和空间关系 p，可以推理出与当前条件互斥的动词集合 $\{h_1,\cdots,h_M\}$，如下所示：

$$\forall x(o(x)\wedge p(x))\rightarrow\neg h_1(x)\wedge\neg h_2(x)\wedge\cdots\wedge\neg h_M(x) \tag{8}$$

例如，对于人-物对〈人，风筝〉，$o(x)$ 为风筝，当 $p(x)$ 为内部时，可以推理出 $h(x)$：放飞与之互斥。

接着，将定义在离散布尔变量上的逻辑连接词（$\rightarrow$，$\neg$，$\vee$，$\wedge$）通过乘积逻辑[38]映射为连续变量上的函数：

$$\begin{aligned}&\psi\rightarrow\phi=1-\psi+\psi\cdot\phi\\&\neg\psi=1-\psi\\&\psi\vee\phi=\psi+\phi-\psi\cdot\phi\\&\psi\wedge\phi=\psi\cdot\phi\end{aligned} \tag{9}$$

而量化器（$\exists$，$\forall$）则按照文献[39]的广义均值方式实现映射：

$$\begin{aligned}&\exists x(\psi(x))=\left(\frac{1}{K}\sum_{k=1}^{K}\psi(x_k)^q\right)^{\frac{1}{q}}\\&\forall x(\psi(x))=1-\left(\frac{1}{K}\sum_{k=1}^{K}(1-\psi(x_k))^q\right)^{\frac{1}{q}}\end{aligned} \tag{10}$$

根据式(8)－式(10)可以将该一阶逻辑转换为连续空间中的函数：

$$\delta_{o,p}=1-\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}(s_k[o]\cdot s_k[h_m]\cdot s_k[p])\right) \tag{11}$$

其中，M 是由本文构建的逻辑规则表所确定的；K 是训练样本的数量（放宽至小批量的数量）；$s_k[o]$，$s_k[p]$，$s_k[h_m]$ 分别是对于样本 x_k、物品类别 o、空间关系 p 和动词类别 h_m 的预测分数。

至此，可以得到空间关系损失函数 $\ell_{o,p}$，将其定义为：

$$\ell_{o,p}=1-\delta_{o,p} \tag{12}$$

$\delta_{o,p}$ 评估预测结果对式(8)中定义的规则的满足程度。例如，给定物品类别“风筝”是高概率（即 $s_k[o]$ 是高值），如果位置关系为“包含”且互斥动作“放飞”也是高概率（即 $s_k[h_m]$ 是高值），那么 $\delta_{o,p}$ 将得到一个低值（不满足该规则），$\ell_{o,p}$ 将得到一个高值，进而惩罚推理出动作“放飞”。通过该损失函数，可以建立起输入条件（空间信息）与输出（交互类别）之间的关系。

同理可以推理出其他函数：

$$\delta_o=1-\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}(s_k[o]\cdot s_k[h_m])\right) \tag{13}$$

$$\delta_{o,g}=1-\frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}(s_k[o]\cdot s_k[h_m]\cdot s_k[g])\right) \tag{14}$$

$$\delta_v=1-\frac{1}{M}\sum_{m=1}^{M}\frac{1}{M}\sum_{n=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}(s_k[h_m]\cdot s_k[h_n])\right) \tag{15}$$

进而可以得到物品功能性损失 ℓ_o、动词共现性损失 ℓ_v 和人体姿态损失 $\ell_{o,g}$，分别定义为：

$$\ell_o=1-\delta_o \tag{16}$$

$$\ell_v=1-\delta_v \tag{17}$$

$$\ell_{o,g}=1-\delta_{o,g} \tag{18}$$

最终定义总的损失为：

$$\ell=\ell_f+\alpha\cdot(\ell_{o,p}+\ell_{o,g}+\ell_o+\ell_v) \tag{19}$$

其中，ℓ_f 是原始模型（PViC[18]，UPT[17]），用于推理结果与标注信息之间的 Focal Loss；α 为固定的超参数。

训练过程的伪代码描述如算法 4 所示。

算法 4 模型训练过程

输入：训练集样本及标注文件

输出：无

```
for step in epochs:
    使用微调的目标检测模型定位图片中的实体；//一阶段
    使用分类器获得每个实体对应的类别、预测框及置信度；
    基于置信度阈值过滤一阶段得到的实体；//二阶段
    按照实体类别（人、物品）分类并两两组合得到若干个人-物对 regions；
    for region in regions:
        使用成对查询生成器生成对应的 query 并保存为集合 queries；
    end for
    for query in queries:
        使用交互解码器推理出每个 query 对应的交互动词的概率；
    end for
    使用式(12)、式(16)－式(18)计算逻辑损失；
    使用式(19)计算总损失；
    根据损失更新网络参数。
```

3.6 先验逻辑引导的推理过程

虽然在训练阶段使用了先验的逻辑规则对训练过程进行监督，但是由于数据集存在长尾分布问题，有些交互类型出现的次数很少，因此训练过程不足以很好地学习到对应的逻辑规则。因此，在推理阶段以矩阵乘法的形式利用逻辑规则对推理结果进行引导。

具体来说，对于一个人-物对 x，在得到物品类别 o、空间关系 p、人体姿态 g 和推理结果 h 之后，可以根据逻辑规则表计算出对交互类别 h_m 的惩罚值。对交互类别 h_m，基于空间关系的惩罚值 $\zeta_{o,p}$ 计算方法如下：

$$\zeta_{o,p}=\begin{cases}s[o]\cdot s[p]\cdot s[h_m], & o\wedge p\wedge h_m=\varnothing\\ 0, & o\wedge p\wedge h_m\neq\varnothing\end{cases} \tag{20}$$

其中，$s[o]$，$s[p]$，$s[h_m]$ 分别代表物品类别 o、空间关系 p 和推理结果 h_m 的预测概率。具体应用中，该过程是一个矩阵计算，同时计算出所有交互类别对应的惩罚值。同理，可以计算出其余 3 个规则对应的惩罚值：

$$\zeta_{o,g}=\begin{cases}s[o]\cdot s[g]\cdot s[h_m], & o\wedge g\wedge h_m=\varnothing\\ 0, & o\wedge g\wedge h_m\neq\varnothing\end{cases} \tag{21}$$

$$\zeta_o=\begin{cases}s[o]\cdot s[h_m], & o\wedge h_m=\varnothing\\ 0, & o\wedge h_m\neq\varnothing\end{cases} \tag{22}$$

$$\zeta_o=\sum_{\substack{i\\ h_i\wedge h_m=\varnothing}}s[h_m]\cdot s[h_i] \tag{23}$$

其中，$\zeta_{o,g}$，ζ_o，ζ_v 分别代表人体姿态、物品功能性和动词共现性这 3 种规则的惩罚值。最终，结合 4 种逻辑规则更新推理结果的公式如下：

$$s[h_m]'=s[h_m]-\mu\cdot\zeta_{o,p}-\bar{\omega}\cdot\zeta_{o,g}-\theta\cdot\zeta_o-\vartheta\cdot\zeta_v \tag{24}$$

其中，权重 μ，$\bar{\omega}$，θ，ϑ 是固定的超参数。推理过程的伪代码描述如算法 5 所示。

算法 5 模型推理过程

输入：图片样本

输出：人-物对位置及其对应的各交互动作的概率

```
读取图片；
使用微调的目标检测模型定位图片中的实体；//一阶段
使用分类器获得每个实体对应的类别、预测框及置信度；
使用成对查询生成器生成人-物对 query；
使用交互解码器推理出每个 query 对应的交互动作的概率；//二阶段
使用式(20)－式(23)计算出概率惩罚值；
使用式(24)计算出更新之后的动作概率。
```

4 实验

4.1 实验设置

4.1.1 实验数据集

本文的数据集使用了公开数据集 HiCO-Det[22]，V-COCO[33] 和 Flickr30K[40]。HiCO-Det[22] 包含 47 776 幅图像，包括 80 个对象类别和 117 个动词类别，从而产生了 600 种 HOI 三元组。该数据集包括 3 个子集：1）完整，包括所有 600 种 HOI 三元组；2）罕见，包括 138 个训练样本少于 10 个的 HOI 三元组；3）非罕见，包括其他 462 个 HOI 三元组。其图片示例如图 4 所示。

V-COCO[33] 是 MS-COCO 的子集，与 HICO-DET[22] 相比规模较小，包括 10 346 幅图像，包含 24 个动词类别和 80 个对象类别。其图片示例如图 5 所示。

Flickr30K[40] 数据集是一个未进行三元组标注的真实世界数据集，包含 31 783 张图像，每张图像配有 5 个独立的英文

描述句子，主要用于图像描述生成领域。其图片示例如图 6 所示。

图 4 HiCO-Det 数据集部分图片示例

Fig. 4 Examples of some images from the HiCO-Det dataset

图 5 V-COCO 数据集部分图片示例

Fig. 4 Examples of some images from the V-COCO dataset

图 6 Flickr30K 数据集部分图片示例

Fig. 6 Examples of some images from the Flickr30K dataset

4.1.2 实验环境

本文的实验环境设置如表 1 所列，操作系统为 CentOS Linux release 8.5.2111，处理器为 AMD EPYC 7763，包含 3 张 80 GB 内存的 A800 显卡。

表 1 实验环境

Table 1 Experimental environment

实验环境	环境配置
操作系统	CentOS Linux release 8.5.2111
处理器	AMD EPYC 7763
显卡	3×NVIDIA Tesla A800 80 GB PCIE
显存	80GB
深度学习框架	PyTorch 1.10.0

4.2 评价指标

对于 HiCO-DET[22] 数据集，由于其是多分类任务，按照惯例使用平均精度均值（mean Average Precision，mAP）作为评价指标，并且对其 3 个子集分别进行计算。mAP 是目标检测领域最常用的评价指标之一，通过计算 Precision-Recall 曲线下的面积，能够综合反映模型在精度（Precision）和召回率（Recall）上的表现。

对于 V-COCO[33] 数据集，按照惯例使用 AP 作为评价指标，且按照数据集要求在两个场景下进行测试。场景一为缺少人物注释的测试用例，如果动作正确且人物框的重叠度大于 0.5 且对应角色为空（如[0,0,0,0]），则代表预测是正确的，此场景适用于遮挡导致的角色缺失，评价指标为 AP^{S1}_{role}。场景二为缺少人物注释的测试用例，如果动作正确且人物框之间的重叠度大于 0.5（忽略相应人物），则代表预测是正确的，此场景适用于角色超出 COCO 类别的情况，评价指标为 AP^{S2}_{role}。

4.3 对比实验

为衡量提出模型的性能，本文复现了近几年主流的人-物交互检测模型，即 FCL[1]，QPIC[41]，UPT[17]，STIP[42]，PViC[18] 和 GEN-VLKT[30] 等，分别使用 ResNet-50 和 Swin-Large 作为特征提取网络进行实验，利用提出的 PKHOI 增强了其中两个算法 UPT[17] 和 PViC[18]，并与这些模型进行对比。

1）FCL[1] 模型：将提取的人体特征与物体特征进行融合，并通过学习机制提高对交互信息的区分能力。

2）QPIC[41]：基于 DETR 的端到端的 HOI 检测器。它对 DETR 的 Transformer 结构进行了改进，添加额外的检测头分别用于定位人体和物体，并预测它们之间的动作。

3）UPT[17]：一种两阶段模型，用额外的 Transformer 层来细化 DETR 的输出特征，用于人-物交互分类，以一元表示和成对表示两种方式编码实例信息，这些表示提供了正交信息，提高了准确性。

4）STIP[42]：将交互动作预测的过程再次细分为两个后续阶段，首先生成交互建议，然后通过结构感知的 Transformer 将非参数交互建议转换为人-物交互动作的预测。

5）PViC[18]：设计了一个高效的交互头用于 HOI 检测，通过空间引导的交叉注意力精确定位与相关身体部位对应的图像区域，将图像特征重新引入人-物对表示中。

6）GEN-VLKT[30]：提出了一种基于图像描述的 HOI 检测预训练策略，旨在从预训练的视觉-语言模型中获取知识。该方法利用了 CLIP[32] 中的视觉-语言知识来增强关系分类，通过迁移 CLIP 中蕴含的知识，增强对交互关系的理解。

7）FGAHOI[5]：一种基于多尺度采样（MSS）、分层空间感知合并（HSAM）和任务感知合并（TAM）机制的 HOI 检测模型。多尺度采样从噪声背景中提取人体、物体和交互区域的特征，用于不同尺度的 HOI 实例；然后利用层次空间感知和任务感知的合并机制将提取的特征与查询嵌入进行语义对齐和合并。

4.3.1 定量实验

本文通过引入先验的逻辑知识对神经网络进行监督，从而增强现有的 HOI 检测算法，在两个公开数据集 HiCO-DET[22] 和 V-COCO[33] 上均有效提升了原始模型的性能。

表 2 列出了在 HiCO-DET[22] 数据集上，使用 ResNet-50 和 Swin-Large 作为 backbone 的实验结果。使用 ResNet-50 作为 backbone，PKHOI（UPT）在原始模型的基础上 mAP 提升了 1.45%；PKHOI（PViC）在原始模型的基础上 mAP 提升了 1.17%。使用 Swin-Large 作为 backbone，PKHOI（PViC）在原始模型的基础上 mAP 提升了 1.47%。HICO-DET 数据集上“罕见类”样本量少，模型易过拟合到错误关联（如“书＋骑”）。PKHOI 通过先验的逻辑规则（如“书不可骑”）直接抑制不合理预测，减少对噪声标签的依赖，从而提升罕见类别的泛化能力，在“Rare”子集（罕见类）上，PKHOI（PViC）提升显著。

表 2 HiCO-DET 数据集实验结果

Table 2 Experimental results on HiCO-DET dataset

(%)

Method	Backbone	Default			Known Object		
		Full	Rare	No-Rare	Full	Rare	No-Rare
FCL[1]	Resnet-50	24.68	20.03	26.07	26.80	21.61	28.35
QPIC[41]	Resnet-50	29.07	21.85	31.23	31.68	24.14	33.93
UPT[17]	Resnet-50	31.66	25.94	33.36	35.05	29.27	36.77
STIP[41]	Resnet-50	32.22	28.15	33.43	35.29	31.43	36.45
PViC[18]	Resnet-50	34.27	31.56	35.10	37.92	35.36	38.95
GEN-VLKT[30]	Resnet-50	33.75	29.25	35.10	36.78	32.75	37.99
PKHOI(UPT)	Resnet-50	33.11(+1.45)	29.51	34.18	37.10	34.79	37.62
PKHOI(PViC)	Resnet-50	**35.44(+1.17)**	**32.26**	**36.62**	**39.48**	**36.10**	**40.49**
FGAHOI[5]	Swin-Large	37.18	30.71	39.11	38.93	31.93	41.02
PViC[18]	Swin-Large	44.32	44.61	44.24	47.81	48.38	47.64
PKHOI(PViC)	Swin-Large	**45.79(+1.47)**	**46.01**	**45.80**	**49.27**	**50.59**	**49.03**

表 3 列出了在 V-COCO[33] 数据集上，使用 ResNet-50 和 Swin-Large 作为 backbone 的实验结果。使用 ResNet-50 作为 backbone，PKHOI(UPT)在原始模型的基础上 AP^{S1}_{role} 提升了 0.7%，AP^{S2}_{role} 提升了 0.5%；PKHOI(PViC)在原始模型的基础上 AP^{S1}_{role} 提升了 0.5%，AP^{S2}_{role} 提升了 1.2%。使用 Swin-Large 作为 backbone，PKHOI(PViC)在原始模型的基础上 AP^{S1}_{role} 提升了 1.2%，AP^{S2}_{role} 提升了 0.6%。V-COCO 数据集规模较小，仅有 10000 多张图像(相比 HICO-DET 的 47000 多张图像)，且动词类别仅 24 种(相比 HICO-DET 的 117 种)。PKHOI 的性能受限主要有以下几个原因：1)PKHOI 的逻辑规则表基于训练集标注构建，V-COCO 中低频动作的样本量较少，导致规则表无法充分覆盖所有可能的合理/不合理组合；2)该数据集的动词类别更为宽泛，如“hold”与“carry”在 V-COCO 中统一用“hold”表示，这使得逻辑规则的互斥约束效果减弱；3)V-COCO 数据集的两种场景设置包含了物品缺失的情况，而 PKHOI 的逻辑规则与物品类别预测密切相关，当物品类别缺失时会影响模型对互斥动词的抑制能力。这些因素共同导致 PKHOI(PViC)在该数据集上的提升幅度不及在 HICO-DET 数据集上。

在 V-COCO 数据集中，PKHOI(PViC)相比于原始模型有一定的提升，但与 STIP[42] 和 DiffHOI[43] 的实验结果仍存在差距。这主要是因为：STIP 将交互检测分为“交互建议生成”和“结构感知 Transformer 优化”两阶段，其多阶段细化机制更适应于 V-COCO 这类小规模数据集上的细粒度优化；V-COCO 数据集的场景一存在遮挡导致的物品缺失情况，此时 DiffHOI 基于 CLIP 模型的分类器能够利用额外数据作为支撑，从而在遮挡场景下展现出更好的处理性能。

表 3 V-COCO 数据集上的实验结果

Table 3 Experimental results on V-COCO dataset

(%)

Method	Backbone	AP^{S1}_{role}	AP^{S2}_{role}
QPIC[41]	Resnet-50	58.8	61.0
UPT[17]	Resnet-50	59.0	64.5
STIP[42]	Resnet-50	**66.0**	**70.7**
PViC[18]	Resnet-50	59.7	65.4
PKHOI(UPT)	Resnet-50	59.7(+0.7)	65.0(+0.5)
PKHOI(PViC)	Resnet-50	60.6(+0.5)	66.6(+1.2)
DiffHOI[43]	Swin-Large	**65.7**	68.2
PViC[18]	Swin-Large	61.7	68.0
PKHOI(PViC)	Swin-Large	62.9(+1.2)	68.6(+0.6)

4.3.2 定性实验

本文通过先验的逻辑规则对互斥的动词预测概率进行抑制，如图 7 所示。图 7(a)、图 7(b)和图 7(c)、图 7(d)分别是同一个人-物对中不同交互动词的预测概率，前者根据空间关系规则，判定动词“握”是互斥动词，因为此时“人在马上方”；后者根据人体姿态规则，判定动词“坐”是互斥动词，因为此时“人体姿态为躺”。所以模型可以对预测结果进行修正，对互斥的动词预测概率进行抑制。如图 7(a)、图 7(d)所示，相比于原始模型，PKHOI 将互斥动词的预测概率进行了有效的抑制；如图 7(b)、图 7(c)所示，PKHOI 对正确动词的预测也有着有效的提升。

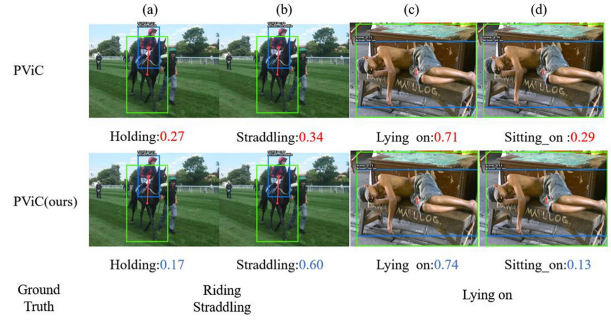


图 7 部分图片的预测结果

Fig. 7 Prediction results for some images

通过可视化交互解码器中 cross-attention 的最后一层注意力权重来探寻融合多模态信息的查询和融合空间信息与姿态信息的位置编码对于推理结果的提升。现有研究[44]证明，从目标检测器中提取的人-物特征往往汇集了边界框的信息，有利于定位人-物对。但如图 8 第一行所示，UPT[17] 模型简单地应用这些特征，虽然能够使得注意力关注到人-物所在区域，但是无法使注意力自主聚焦于信息更丰富的区域，因为这种方法缺乏适应性。本文考虑到人-物交互中人体才是主体，充分考虑了参与互动的人物肢体的重要性，通过融合人体姿态信息生成人-物对查询和将人体姿态关键点融入位置编码的形式，引导注意力集中到人-物所接触的关键区域。提出人体部位区域交集(IOB)的概念，模型在不断的训练过程中学习到 IOB 与预测动词之间的关系，将注意力引导至与动词预测有关的身体部位上。例如，在图 8(a)中，人体做出动词“洗”时手部的 IOB 较高，模型通过将注意力集中在人体手上

和与自行车靠近的水管上，推理出动作“洗”；在图 8(c)中，髋部与膝盖的 IOB 较高，模型将注意力集中在人体髋部和马背上，推理出动作“骑”；在图 8(e)上，阅读时除了拿书的手，其他部分的 IOB 均趋向于 0，模型将注意力集中在人的眼睛和拿书的手上，有效提高了动词“阅读”的预测概率，而原始模型由于未引入姿态信息，注意力则错误地集中在书本与人体中心点(脖颈处)上。

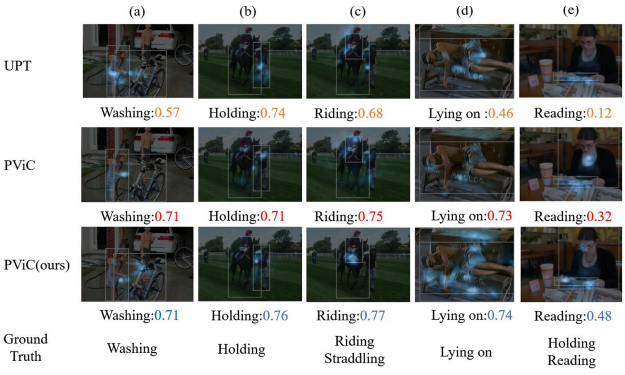


图 8 注意力均值可视化结果

Fig. 8 Results of attention mean visualization

4.4 跨数据集泛化性验证

4.4.1 实验设置

为验证在不同数据分布上的泛化能力，在 Flickr30K[40] 数据集上进行了跨数据集泛化性验证。为了进行定性分析，本文随机选择了 150 张图像，确保涵盖不同场景、光照条件和交互类型，以便全面评估模型的表现。模型选取在 HiCO-Det 数据集上训练的 PKHOI(PViC)模型，backbone 为 ResNet-50，通过人工检查模型的预测结果，识别成功和失败的案例，并对案例进行分析。

4.4.2 定性实验结果

图 9 展示了模型在测试数据集上的成功识别案例。如图 9(a)和图 9(b)所示，对于 HiCO-Det 数据集中已存在的交互类型(如“骑自行车”“坐椅子”和“躺椅子”等)，本文方法表现出较高的识别准确率。值得注意的是，如图 9(c)和图 9(d)所示，即使对于 HiCO-Det 数据集中未出现的物品及其相关交互(如“穿旱冰鞋”“抓绳索”和“操作纹身枪”等)，模型仍能进行准确识别，这表明模型具有良好的泛化能力。

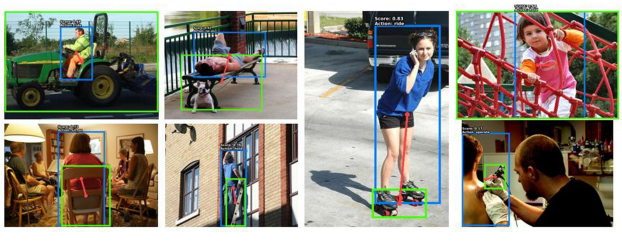


图 9 识别成功案例

Fig. 9 Successful recognition cases

图 10 展示了典型的识别失败案例。如图 10(a)和图 10(d)所示，对于模型已学习过的交互类别(如“操作手机”“持有海报”等)，当目标物体存在部分遮挡时，PKHOI(PViC)虽能识别出正确的交互类型，但输出的置信度较低。这种现象主要是由于模型的最终置信度计算与物体检测置信度密切相关。如图 10(b)和图 10(c)所示，对于模型未见过的特定交互类别(如“骑在人上”“操作游戏机”等)，PKHOI(PViC)未能实现准确识别，这反映了模型在处理未见过的复杂交互场景时的局限性。

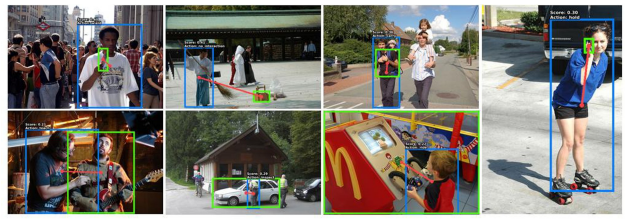


图 10 识别失败案例

Fig. 10 Failed recognition cases

4.4.3 分析与讨论

实验结果表明，本文方法在跨数据集场景下对已知交互类型(如“骑自行车”)和新颖交互类型(如“穿旱冰鞋”)均表现出一定的泛化能力。这表明模型成功学习到了基本的人-物交互模式，特别是通过多个逻辑规则的组合设计，模型展现出优秀的组合泛化性，使其能够在新的数据分布上保持良好的识别性能。然而，本文方法也存在一定局限性：由于逻辑规则对互斥动词的抑制能力与物品识别的准确率密切相关，在物体遮挡条件下，模型对交互类型的推测不确定性显著增加。此外，对于训练集中未出现的复杂交互类型，尤其是涉及非常规交互方式时，模型的识别准确率出现明显下降。这些现象反映了模型在处理新颖、独特交互模式时的局限性。

4.5 消融实验

4.5.1 不同模块的有效性研究

本小节研究成对查询生成器、逻辑损失和逻辑推理的有效性。消融实验均使用 ResNet-50 作为 backbone，在 HiCO-DET[22] 数据集上使用 PKHOI(PViC)模型进行实验，均训练 30 个 epoch。

如表 4 所列，使用成对查询生成器融合多模态的信息，然后使用逻辑损失构建出这些信息与推理结果之间的桥梁，同时使用修改后的查询与逻辑损失，可以有效地提升预测的准确性。

表 4 不同模块消融实验结果

Table 4 Ablation experiment results for different modules

(%)

Query Generator	Logic Loss	Logic Inference	Full	Rare	No-Rare
—	—	—	34.27	31.56	35.10
√	—	—	34.69	32.14	35.45
—	√	—	34.91	32.18	36.08
—	—	√	34.59	32.09	35.24
√	√	—	35.29	32.18	36.45
√	—	√	34.79	32.14	36.14
—	√	√	35.05	32.21	36.11
√	√	√	**35.44**	**32.26**	**36.62**

从表 4 中可以看出，由于先验逻辑引导的推理过程不需要训练，逻辑推理模块无需训练即可修正长尾分布问题，因此其对罕见类型的交互类别也有可观的提升(从 31.56%提升

至 32.09%)。例如,当模型对“人+滑板”预测动词“坐”时,根据人体姿态规则(人体姿态不能为“站立”),抑制动词“坐”并提升动词“滑行”的预测概率。

4.5.2 成对查询生成器中 MLP 层数的研究

本小节研究成对查询生成器中,对不同模态信息进行线性映射的多层感知机(MLP)的全连接层数对推理结果的影响。实验仅使用成对查询生成器与逻辑损失函数,训练 30 个 epoch。如表 5 所列,当 MLP 层数为 3 时获得最优值,层数为 1 时效果不佳,这说明单个全连接层无法有效学习到各种隐含的知识。

表 5 MLP 不同层数的实验结果

Table 5 Experimental results for different numbers of layers in MLP

(%)

Layers	Full	Rare	No-Rare
1	35.12	32.02	36.17
2	35.23	32.18	36.40
3	**35.29**	**32.18**	**36.45**

4.5.3 不同查询模块的有效性

本小节研究成对查询生成器中,融合不同查询模块对推理结果的影响。实验仅使用成对查询生成器,不使用逻辑损失与逻辑推理,训练 30 个 epoch。结果如表 6 所列,其中第 4 种设置条件与 PViC[18] 相同。实验结果表明,空间信息的引入对于预测的准确性至关重要,当去掉空间信息后,准确率大幅降低。

表 6 不同查询模块的影响

Table 6 Impact of different query modules

(%)

Spatial	Content	Pose	Full	Rare	No-Rare
√	—	—	33.54	30.17	34.55
—	√	—	33.04	28.31	34.43
—	—	√	33.59	29.65	34.76
√	√	—	34.27	31.56	35.10
√	—	√	33.91	29.28	35.29
—	√	√	33.72	29.14	34.92
√	√	√	**34.69**	**32.14**	**35.45**

4.5.4 逻辑推理中不同规则权重的影响

本小节研究先验知识引导的逻辑推理过程中,不同规则的权重 $\mu,\omega,\theta,\vartheta$ 对推理结果的影响。实验在原始 PViC[18] 模型基础上进行,仅使用逻辑推理,不进行额外训练。结果如图 11 所示,$\mu,\omega,\theta,\vartheta$ 分别对应 alpha_posi, alpha_pose, alpha_obj, alpha_verb。权重在分别取值为 0.25, 0.15, 0.3, 0.225 时,mAP 取得最大值。从实验结果可以看出,物品功能性(alpha_obj)对结果的提升很小,仅提升 0.3%,这表明现有模型已经能够根据获取到的物品的特征准确推理出相应的动词;空间关系规则(alpha_posi)对结果提升显著(+1.2%),这是因为空间关系是 HOI 检测的核心线索,其直接约束交互的物理可行性;姿态信息规则(alpha_pose)因关键点检测噪声(ViTPose 误差)和姿态-动作多义性(如“站立”对应多种动作),提升效果波动较大(±0.5%)。未来工作需结合知识增强与鲁棒姿态估计,以平衡各规则的贡献。

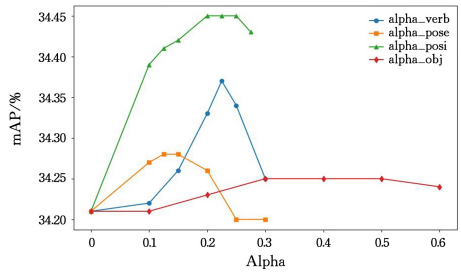


图 11 推理结果随参数 $\mu,\omega,\theta,\vartheta$ 的取值变化

Fig. 11 Variation of inference results with values of parameter $\mu,\omega,\theta,\vartheta$

结束语 本文提出一种利用先验知识增强现有人-物交互检测算法的方法 PKHOI,旨在增强现有二阶段检测器的准确性。具体而言,从训练集中构建了一个包含物品功能性、空间关系、人体姿态和动词共现的逻辑规则表,将其形式化为一阶逻辑并映射到连续空间中。在训练阶段和推理阶段分别以损失函数的形式和矩阵乘法的形式融入神经网络,增强模型的准确性。此外,提出一种通过融合多模态信息(空间、语义和人体姿态信息)生成人-物对查询的方法,通过逻辑损失建立不同信息与推理结果之间的联系。利用提出的方法增强了两个主流的人-物交互检测算法 UPT 和 PViC,并在 V-COCO 和 HICO-DET 数据集上进行了评估。

本文还存在以下不足:1)根据人体姿态关键点获取姿态类别的规则缺乏鲁棒性,使得该规则可能产生噪声,未来可以利用神经网络的方式训练一个分类器,而不是采用固定的规则;2)逻辑规则的提取只使用了训练集的标注文件,无法提取不存在于训练集中的规则,未来可以考虑使用知识图谱推理的方式增强获取到的逻辑规则。

参 考 文 献

[1] FANG H S, XIE Y C, SHAO D, et al. DIRV: Dense interaction region voting for end-to-end human-object interaction detection [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 1291-1299.

[2] KIM B, CHOI T, KANG J, et al. UnionDet: Union-level detector towards real-time human-object interaction detection[C]// Computer Vision ECCV 2020: 16th European Conference. 2020: 498-514.

[3] LIAO Y, LIU S, WANG F, et al. PPDM: Parallel point detection and matching for real-time human-object interaction detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 482-490.

[4] CHEN J, YANAI K. QAHOI: Query-based anchors for human-object interaction detection[C]// 2023 18th International Conference on Machine Vision and Applications (MVA). New York: IEEE, 2023: 1-5.

[5] MA S, WANG Y, WANG S, et al. FGAHOI: Fine-grained anchors for human-object interaction detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(4): 2415-2429.

[6] LIM J Y, BASKARAN V M, LIM J M Y, et al. ERNet: An effi-

cient and reliable human-object interaction detection network [J]. IEEE Transactions on Image Processing，2023，32：964-979.

[7] ZHANG A，LIAO Y，LIU S，et al. Mining the benefits of two-stage and one-stage HOI detection [J]. Advances in Neural Information Processing Systems，2021，34：17209-17220.

[8] CARION N，MASSA F，SYNNAEVE G，et al. End-to-end object detection with transformers[C]// Computer Vision ECCV 2020：16th European Conference. Berlin：Springer，2020：213-229.

[9] GIRSHICK R. Fast R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. New York：IEEE，2015：1440-1448.

[10] BANSAL A，RAMBHATLA S S，SHRIVASTAVA A，et al. Detecting human-object interactions via functional generalization [C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York：AAAI，2020：10460-10469.

[11] LI Y L，LIU X P，LU H，et al. Detailed 2D-3D joint representation for human-object interaction [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2020：10166-10175.

[12] GUPTA T，SCHWING A，HOIEM D. No-frills human-object interaction detection：Factorization，layout encodings，and training techniques[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York：IEEE，2019：9677-9685.

[13] WU E Z Y，LI Y，WANG Y，et al. Exploring pose-aware human-object interaction via hybrid learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2024：17815-17825.

[14] PARK J，PARK J W，LEE J S. VIPLO：Vision transformer based pose-conditioned self-loop graph for human-object interaction detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2023：17152-17162.

[15] ZHANG F Z，CAMPBELL D，GOULD S. Spatially conditioned graphs for detecting human-object interactions[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York：IEEE，2021：13319-13327.

[16] WANG T，ANWER R M，KHAN M H，et al. Deep contextual attention for human-object interaction detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York：IEEE，2019：5694-5702.

[17] ZHANG F Z，CAMPBELL D，GOULD S. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2022：20072-20080.

[18] ZHANG F Z，YUAN Y，CAMPBELL D，et al. Exploring predicate visual context in detecting human-object interactions[C]// 2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE，2023：10377-10387.

[19] YUAN H，JIANG J，ALBANIE S，et al. RLIP：Relational language-image pre-training for human-object interaction detection [J]. Advances in Neural Information Processing Systems，2022，35：37416-37431.

[20] YUAN H，ZHANG S W，WANG X，et al. RLIPv2：Fast scaling of relational language-image pre-training[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York：IEEE，2023：21649-21661.

[21] NING S，QIU L，LIU Y，et al. HOICLIP：Efficient knowledge transfer for HOI detection with vision-language models[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2023：23507-23517.

[22] 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). New York：IEEE，2018：381-389.

[23] DILIGENTI M，ROYCHOWDHURY S，GORI M. Integrating prior knowledge into deep learning[C]// 2017 16th IEEE International Conference on Machine Learning and Applications(ICMLA). New York：IEEE，2017：920-923.

[24] CHEN S，LENG Y，LABI S. A deep learning algorithm for simulating autonomous driving considering prior knowledge and temporal information [J]. Computer-Aided Civil and Infrastructure Engineering，2020，35(4)：305-321.

[25] DING X，LUO Y，LI Q，et al. Prior knowledge-based deep learning method for indoor object recognition and application [J]. Systems Science & Control Engineering，2018，6(1)：249-257.

[26] ZHENG S，MAI S，SUN Y，et al. Subgraph-aware few-shot inductive link prediction via meta-learning [J]. IEEE Transactions on Knowledge and Data Engineering，2022，35(6)：6512-6517.

[27] GENG Y，CHEN J，PAN J Z，et al. Relational message passing for fully inductive knowledge graph completion[C]// 2023 IEEE 39th International Conference on Data Engineering(ICDE). New York：IEEE，2023：1221-1233.

[28] ZHANG Y，YAO Q. Knowledge graph reasoning with relational digraph[C]// Proceedings of the ACM Web Conference 2022. New York：ACM，2022：912-924.

[29] CHEN D，LAI H，GAO G，et al. Prior knowledge guided three-branch transformer for HOI detection [EB/OL]. https://papers. ssrn. com/sol3/papers. cfm?abstract_id=4608308.

[30] LIAO Y，ZHANG A，LU M，et al. Gen-VLKT：Simplify association and enhance interaction understanding for HOI detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2022：20123-20132.

[31] GAO J，LIANG K，WEI T，et al. Dual-prior augmented decoding network for long tail distribution in HOI detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York：AAAI，2024：1806-1814.

[32] RADFORD A，KIM J W，HALLACY C，et al. Learning transferable visual models from natural language supervision[C]// International Conference on Machine Learning. PMLR，2021：8748-8763.

[33] GUPTA S，MALIK J. Visual semantic role labeling [J]. arXiv：1505. 04474，2015.

[34] ZHANG F Z, CAMPBELL D, GOULD S. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 20104-20112.

[35] XU Y, ZHANG J, ZHANG Q, et al. ViTPose: Simple vision transformer baselines for human pose estimation [J]. Advances in Neural Information Processing Systems, 2022, 35: 38571-38584.

[36] BA J L, KIROS J R, HINTON G E. Layer normalization [J]. arXiv: 1607.06450, 2016.

[37] KIM D J, SUN X, CHOI J, et al. Detecting human-object interactions with action co-occurrence priors[C]// Computer Vision-ECCV 2020: 16th European Conference. Berlin: Springer, 2020: 718-736.

[38] VAN KRIEKEN E, ACAR E, VAN HARMELEN F. Analyzing differentiable fuzzy logic operators [J]. Artificial Intelligence, 2022, 302: 103602.

[39] SERAFINI L, D'AVILA GARCEZ A, BADREDDINE S, et al. Logic tensor networks: Theory and applications[M]// Neuro-Symbolic Artificial Intelligence: The State of the Art. Amsterdam: IOS, 2021: 370-394.

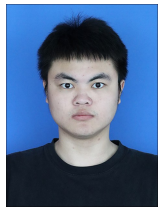
[40] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions [J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67-78.

[41] TAMURA M, OHASHI H, YOSHINAGA T. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2021: 10410-10419.

[42] ZHANG Y, PAN Y, YAO T, et al. Exploring structure-aware transformer over interaction proposals for human-object interaction detection[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 19526-19535.

[43] YANG J, LI B, YANG F, et al. Boosting human-object interaction detection with text-to-image diffusion model [J]. arXiv: 2305.12252, 2023.

[44] ZHANG F Z, YUAN Y, CAMPBELL D, et al. Exploring predicate visual context in detecting human-object interactions[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 10411-10421.



ZHAO Wenhao, born in 2001, postgraduate. His main research interests include computer vision and HOI detection.

WANG Xiaoping, born in 1965, Ph. D, professor. His main research interests include AI algorithms, deep learning and computer vision.

（责任编辑：李亚辉）