零样本人-物交互检测综述

胡家铭[1)]

[1)](西安交通大学 电信学部，陕西 西安, 710049)

摘 要 作为零样本学习、行为识别、视觉关系检测的交叉学科，零样本人物交互（Zero-Shot Human-Object Interaction，HOI）检测旨在识别场景中已见和未见的人与物体的相互关系。本文对零样本人物交互检测研究成果进行了系统总结及论述。首先，从提升对未见样本识别率的不同方法，把人物交互检测方法分为基于语义属性、基于生成模型、基于迁移学习和基于注意力机制四类，并对代表性方法进行了详细阐述和分析；然后，讨论了零样本人物交互检测在动态视频识别、辅助机器人中的应用；最后，从语义鸿沟、数据长尾分布以及多样性三方面出发，总结了零样本人物交互检测面临的挑战，并指出语义图谱，数据增强，多模态学习可以作为未来进一步发展的方向。

关键词 零样本学习；人物交互；视觉关系；目标检测；动作识别

A Survey of Zero-Shot HOI Detection

HU Jiaming[1)]

[1)](Faculty of Electronic and Information Engineering ,Xi'an Jiaotong University, Xi'an 710049，China)

Abstract As an interdisciplinary field encompassing Zero-Shot learning, action recognition, and visual relationship detection, zero-shot Human-Object Interaction (HOI) detection aims to discern the relationship between individuals and objects within a given scene. This paper provides a comprehensive summary and analysis of research findings in zero-shot human interaction detection. Firstly, various methods for enhancing the recognition accuracy of unseen samples are categorized into four groups: those based on semantic attributes, generative models, transfer learning, and attention mechanisms. Representative approaches within each category are elaborated upon and thoroughly examined. Subsequently, the applications of zero-shot human interaction detection in dynamic video recognition and assistant robotics are discussed. Finally, this paper outlines the challenges faced by zero-shot human interaction detection from three perspectives: semantic gap, long-tail data distribution and diversity issues; furthermore suggesting that future development directions may involve leveraging techniques such as semantic graph modeling, data augmentation strategies,and multimodal learning.

Key words zero-shot learning; human-object interaction; visual relationship; object detection; action recognition

1 引言

人-物交互（Human-Object Interaction，HOI）检测在近些年的研究中已经逐步完善，通过空间条件化和概率矩阵等方法，已经可以实现对经过训练的人-物交互情景有着较好的学习率。但是这种方法在应用中有一个无法回避的问题，即严重依赖于带有预定义 HOI 类别标注的训练数据集。这会导致当遭遇训练数据外的人-物交互情形时，没有办法准确识别，只是将其归类至与之最相似的预设类别中，并且如果模型需要扩展更多的数据类别，将需要极大的训练成本。

因此，将零样本（Zero-Shot）学习技术加入到 HOI 检测中是十分重要的，零样本人-物交互 检测不仅能够检测出训练集中经过训练的 HOI 情形，还能以很高的准确率检测出从未见过的 HOI 情形。而且，其优势不只是能使模型正确识别更多 HOI 情形，还能大大减轻模型的训练成本和扩展成本。

1.1 人-物交互检测研究

人-物交互检测区别于传统的目标检测，旨在利

用人体、物体以及人-物对的特征将人与物体之间的交互进行关联，从而实现对图像或视频中动作的定位及分类。

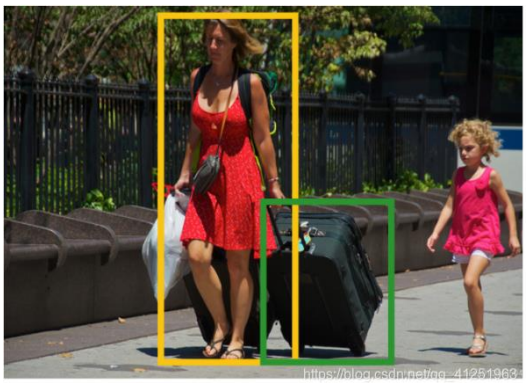

图 1 人-物交互检测示例（女人 拉着 行李箱）

HOICS[1]是 2024 年提出的端到端并行 HOI 检测框架，通过兼容性自学习策略来解决零样本 HOI 检测问题。该方法利用组合概率矩阵和 HOI 兼容性得分来提取 HOI 兼容性数据的洞察，并优化交互预测头。

HTOR[2]是 2021 年发表，使用变换器（Transformer）结构进行端到端的 HOI 检测，该方法利用变换器的自注意力机制来捕捉人类和物体之间的复杂关系。

GEN-VLKT[3]是 2022 年提出的通过简化关联并增强交互理解来提高 HOI 检测的性能，该方法通过学习交互的短语表示和标签组合来改善检测结果。

1.2 零样本学习

零样本学习领域主要有基于图神经网络的学习方法[4]（Schmitt 等人，2020），通过学习类别之间的图结构关系来提高分类性能；通过跨域翻译和元学习来实现零样本学习[5]（Xu 等人，2021），通过在源域和目标域之间建立映射来提高对未见类别的识别能力；基于语义引导的图卷积网络的零样本学习方法[错误!未找到引用源。]（X Wang 等人，2018），通过学习类别之间的语义关系来提高分类性能；使用生成类别感知嵌入的零样本学习方法[7]（P Zhao 等人，2021），通过生成与类别相关的嵌入来提高对未见类别的识别能力。

1.3 零样本人-物交互检测

零样本人-物交互检测是一个新兴的研究领域，允许模型识别在训练阶段未遇到的新类别，这能增强模型的泛化能力，同时对于稀有或难以获取的交互类别，零样本检测可以减少对大量标注数据的需求。Weiying Xue 等人于 2024 年提出了一个名为 Knowledge Integration to HOI (KI2HOI)的新框架[8]，该框架通过视觉-语言模型的知识整合来提高零样本 HOI 检测的性能。同时通过交叉注意力机制整合空间和视觉特征信息，有效地提取信息丰富的区域。而对于低数据量的零样本学习，其采用了 CLIP 文本编码器的先验知识初始化线性分类器，以增强交互理解。并在 HICO-DET 和 V-COCO 数据集上进行了广泛实验，证明了模型的优越性 。

Daoming Zong 等人于 2023 年提出了一种相似性传播（SP）方案[9]，利用余弦相似度距离来调节已见和未见对象之间的预测边界。并基于类别语义相似性，在训练期间为未见对象引入伪监督。还将语义感知的实例级和交互级对比损失与 Transformer 结合，增强了类内紧凑性和类间可分性，从而改善了视觉表示。

2 零样本人-物交互检测分类

在零样本人-物交互检测中，主要通过将视觉等信息映射到语义空间，实现对未见事物的语义表达，从而达到识别未见事物的目的。将图像输入到模型中后，先使用深度学习模型从输入图像中提取特征，并进行表示。然后在嵌入语义信息时，对于已知类别，系统使用预训练的语义嵌入（如 Word2Vec 或 GloVe）来表示类别的语义信息。对于未见过的类别，可以使用辅助数据源或预训练的语言模型（如 BERT 或 CLIP）来获取或推断语义信息。随后模型通过分析图像中的特征表示和类别语义嵌入来生成可能的交互候选。这可能涉及到匹配人物和物体之间的空间关系，以及它们与语义类别的对应关系。最终模型通过评估交互候选，确定哪些候选代表真实的人物交互，并输出最终的检测结果。

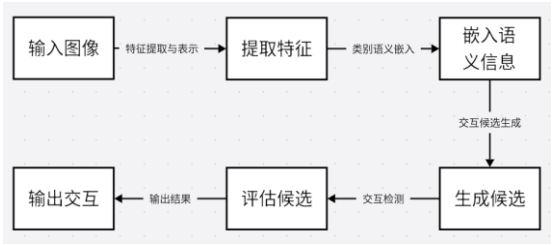

图 2 零样本人-物交互检测流程

零样本人-物交互检测研究目前主要有分析语义属性、加入生成模型、迁移学习和融合注意力机制等方法。

2.1 基于语义属性

Weiying Xue 等人[8]提出了一种新的框架KI2HOI，用于零样本人-物交互检测。该框架通过视觉-语言集成，利用 CLIP 模型的视觉语言知识来增强交互表示和识别未知的 HOI 实例。框架主要由四个部分组成：视觉编码器、动词特征学习、实例交互器和交互语义表示。

视觉编码器作为框架的基础，负责从输入图像中提取特征图，并生成全局视觉特征。这为后续的动词特征学习和交互表示提供了必要的视觉信息。

动词特征学习模块通过动词提取解码器进行动词识别，作为辅助信息，帮助模型理解不同的交互类型。这与视觉编码器提取的特征相结合，增强了对人-物体交互的理解。

实例交互器计算人类查询和物体查询的均值，进一步处理这些查询以生成交互的预测。这一过程依赖于前两个模块提供的特征信息。

交互语义表示通过将视觉特征与 CLIP 的文本嵌入结合，模型能够更好地转移知识，从而提高对未知 HOI 类别的检测能力。

他们使用预训练的 DETR 与 ResNet-50 作为骨干网络，视觉编码器基于 ViT-32/B CLIP，训练过程中 CLIP 的参数保持不变。模型的编码器和解码器有 3 层，每层有 64 个查询，除了动词提取解码器有 1 层。使用 AdamW 优化器进行训练，并在 60 个 epoch 后逐渐降低学习率。所有实验在 4 个 NVIDIA A6000 GPU 上进行，每批大小为 8。

图 3 KI2HOI 实验结果（从左至右分别为 HOI 提取，动词提取解码器，注意力解码器的可视化结果）

在 HICO-DET 数据集上进行的各种零样本设置实验中，KI2HOI 模型超越了现有的最先进方法，表现出强大的性能竞争力。特别是在 RF-UC 和 NF-UC 设置下，模型的相对平均精度（mAP）分别超过了 EOID 23.26%和 7.91%。在 UA 设置中，模型超过了最新工作 HOICLIP 1.03 mAP，用于未见类型的检测。在全部和未见类别中，模型相对于 EOID 实现了 1.07 和 2.5 mAP 的提升。在未见物体的情况下，模型的性能超过了 GEN-VLKT 3.21 mAP。

2.2 基于生成模型

Ting Lei 等人[10]提出了一种名为条件多模态提示（CMMP）的新技术，旨在通过条件多模态提示来适应大型基础模型，以应对零样本人-物交互检测的挑战。CMMP 将零样本人-物交互检测分为两个子任务：提取空间感知的视觉特征和交互分类，这种设计与传统的提示学习方法不同，强调了视觉和语言提示的独立性。

同时 CMMP 又通过结合视觉和语言的条件多模态提示，利用先验知识来增强模型的理解能力，从而实现更好的零样本人-物交互检测。条件视觉提示和语言提示的解耦设计使得模型能够在处理已见和未见的交互类别时，保持较高的泛化能力和准确性。

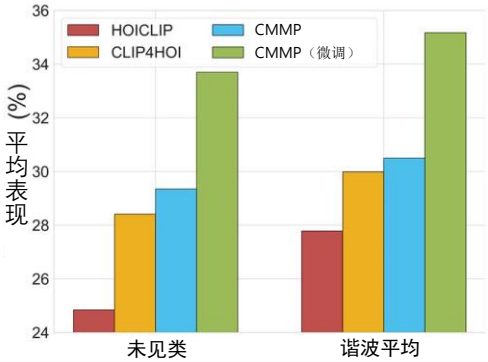

图 4 CCMP 的模型表现

CMMP 在各种零样本设置下的性能表现出色，超越了以往的方法。在未见类别上，CMMP 相对于以前的最先进方法实现了平均 mAP 的显著提升，分别在五个零样本设置上实现了 6.82%、3.44%、2.07%、6.20%和 0.81%的相对 mAP 增益。此外，通过使用 ViT-L/14 骨干网络扩展 CMMP，实现了在所有分割上的性能提升，展示了 CMMP 在空间关系提取和原型学习方面的优势。

2.3 基于知识迁移

Hangjie Yuan 等人[11]提出了一种新的预训练策略——关系语言-图像预训练（Relational Language-Image Pre-training）。RLIP 利用大规模的图像和文本数据进行预训练，学习跨模态的关联表示，然后将这些知识迁移到零样本人-物交互检测任务中，以识别训练期间未见过的交互类别。另外，通过引入关系质量标签（RQL）和关系伪标签（RPL）来解决关系语义模糊性的问题。

关系质量标签（RQL）通过标签平滑的方法来评估实体检测的质量，从而调整关系标签的置信

度。这一方法基于实体检测的准确性，旨在提高关系检测的可靠性。关系伪标签（RPL）通过伪标签策略来处理同义词问题，利用文本嵌入的语义相似性来选择高相似度的标签，从而增强模型的学习能力。RQL 和 RPL 都是为了解决关系语义模糊性而提出的策略，前者通过评估检测质量来调整标签置信度，后者则通过伪标签来增强模型对同义词的理解。这两者共同作用，提升了模型在复杂场景下的表现。

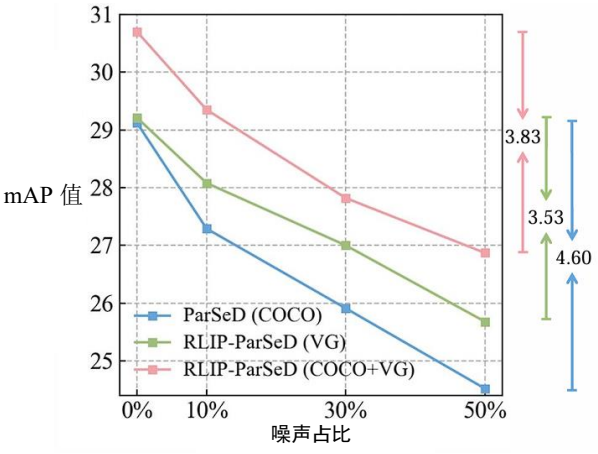

图 5 RLIP 预训练的抗噪声表现

在研究标签噪声对模型影响的实验中，他们展示了随着标签噪声的增加，使用 RLIP 进行预训练的模型，相比使用 COCO 进行预训练的模型性能下降幅度更小，这说明在少样本迁移学习和标签噪声的影响下，RLIP 表现出更强的鲁棒性。

2.4 基于注意力机制的方法

现有的两阶段人-物交互检测模型在视觉特征的使用上存在主要弱点，即往往依赖于缺乏细粒度上下文信息的对象特征，这限制了对复杂或模糊交互行为的识别能力。为此，Frederic Z. Zhang 等人[12]提出了一种改进的设计，通过跨注意力机制重新引入图像特征，以增强谓词视觉上下文（PViC），从而提高模型对 HOI 的识别能力。

经过广泛的实验，他们发现引入图像特征和位置嵌入作为空间指导显著提升了模型的性能，尤其是在需要细粒度视觉特征（如人类姿态）和额外上下文（如参与交互的其他物体）时。实验中还发现现有模型在处理复杂或模糊的交互时，往往由于缺乏细致的上下文信息而表现不佳，但是通过交叉注意机制引入的视觉上下文能够显著改善对稀有交互类别的检测性能。这表明，视觉上下文在提高人类-物体交互检测的准确性和鲁棒性方面起到了重要作用。

3 其他新技术

Yuchen Zhou 等人收集了一个名为 IG 的新颖注视点数据集，包含 530,000 个注视点，覆盖 740 种不同的交互类别，捕捉人类观察者在认知交互过程中的视觉注意力。接着，他们引入了零样本交互导向的注意力预测任务（ZeroIA），挑战模型在训练中未遇到的交互的视觉线索预测能力。此外，论文提出了交互注意力模型（IA），旨在模仿人类观察者的认知过程来解决 ZeroIA 问题。

Alessio Sarullo 等人通过使用可供性图谱进行零样本人-物交互检测，利用图像中外部知识模拟动作和物体之间的可供性关系，以提高对未见类别的识别能力。

Hangjie Yuan 等人提出了一个基于 transformer 的模型，用于实时检测和预测视频中的人-物交互，以帮助辅助机器人更有效地与人类合作。模型结合了预训练模型和时空一致性，以提高检测和预测的准确性和速度。

Jocher 设计了一个框架，通过跟踪人类视线来预测视频中的人-物交互。框架利用人类在与物体交互前往往会注视该物体的倾向，将视线信息与场景上下文和人-物对的视觉外观融合，通过 transformer 对时间及空间信息进行处理，以提高交互预测的准确性。

文献	发表时间	主要工作概述
文献[13]	2024	通过观察者视线进行零样本注意力预测，以人-物交互为导向
文献[13]	2020	通过使用可供性图谱进行零样本人-物交互识别
文献[13]	2023	针对辅助机器人的人-物交互预测
文献[13]	2023	通过视线跟踪预测视频中的人-物交互

表 1 其他新技术总结

4 挑战与展望

近年来，随着深度学习和自然语言处理技术的发展，零样本人-物交互检测取得了显著进展。研究者们通过语义嵌入、知识迁移、多模态学习等技术，开发了多种方法来解决这一问题。尽管如此，该领域距离实际应用还有很大距离，仍面临以下一些挑战：

1. 语义鸿沟

零样本学习的核心挑战之一是训练集和测试集之间的语义不一致性。例如已知的交互（如"人-骑-自行车"）和未知的交互（如"人-推-自行车"）尽管共享某些特征，但其语义描述和视觉表示可能存在显著差异，导致模型难以泛化。

2. 数据不均衡与长尾分布

长尾分布，顾名思义，数据的分布表现和图中类似，纵轴为样本数量，横轴为不同的样本种类，；看起来就像有一条长长的“尾巴”。常见种类的数据（如"人-拿-杯子"）有大量样本支持而少见的交互（如"人-抛-球"）样本稀少甚至完全缺失。这会导致模型在一些种类的数据上训练得过拟合但是在罕见种类的数据上却欠学习，不利于模型的训练。

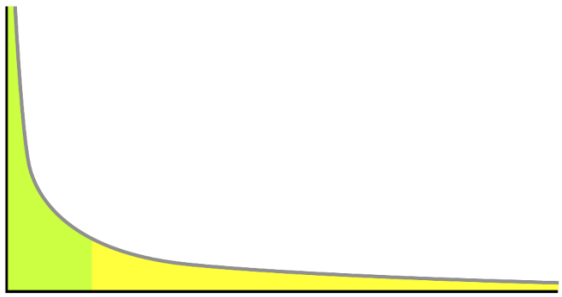

图 6 长尾分布表现

3. 多样性与复杂性

人类的行为和交互具有高度的多样性和复杂性。同一交互（如"人-推-物体"）可能在不同的环境、角度或姿态下呈现出截然不同的视觉特征。此外，不同交互之间可能存在细微的视觉差异，进一步增加了模型学习的难度。

图 7 交互的复杂性

例如在图 7 中，8 号手提箱和 10 号女人实际没有动作联系，但是在这幅图上产生了一定的错位效果，容易被机器以为女人提着手提箱，这就提高了模型训练的难度。

综上所述，零样本人物交互检测方法的分析与研究还处于起步阶段，有很多困难需要去探索和认识，未来研究可以从如下几方面开展：

1. 改进语义嵌入与知识表示

利用知识图谱（Knowledge Graph）构建更加丰富的语义表示，从而捕捉动作、物体和交互之间的复杂关系。以及结合大规模语言模型（如 BERT、GPT）从文本中提取更细粒度的语义信息。

2. 数据增强

针对数据不均衡问题，可以通过生成对抗网络（GAN）技术生成未见交互的合成样本。也可以对长尾分布数据进行采样或加权，平衡常见和罕见交互类型的学习过程。

3. 引入多模态学习

可以利用多模态学习提升零样本人-物交互检测性能。例如利用视觉-语言对比学习（如 CLIP）来捕捉交互的多模态特征，学习动态信息（如动作轨迹）提升对复杂交互的理解，并在视觉、文本和其他模态之间建立统一的特征空间。

参 考 文 献

[1] Jiang M , Li M , Ren J ,et al.HOICS: Zero-Shot Hoi Detection via Compatibility Self-Learning[C]//ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).0[2024-11-29].DOI:10.1109/ICASSP48485.2024.10446999.

[2] Kim B , Lee J , Kang J ,et al.HOTR: End-to-End Human-Object Interaction Detection with Transformers[J]. 2021.DOI:10.1109/CVPR46437.2021.00014.

[3] Liao Y , Zhang A , Lu M ,et al.GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection[C]//arXiv.arXiv, 2022.DOI:10.48550/arXiv.2203.13954.

[4] Y. Liu, Y. Dang, X. Gao, J. Han and L. Shao, "Zero-Shot Learning With Attentive Region Embedding and Enhanced Semantics," in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 3, pp. 4220-4231, March 2024, doi: 10.1109/TNNLS.2022.3202014.

[5] Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, Isabelle Augenstein, Zero-Shot Cross-Lingual Transfer with Meta Learning[1]

[6] X. Wang, Y. Ye and A. Gupta, "Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6857-6866, doi: 10.1109/CVPR.2018.00717.

[7] Zhao P , Zhang S , Liu J ,et al.Zero-shot Learning via the fusion of generation and embedding for image recognition[J].Information Sciences, 2021, 578:831-847.DOI:10.1016/J.INS.2021.08.061.

[8] Xue, Weiying, Qi Liu, Qiwei Xiong, Yuxiao Wang, Zhenao Wei, Xiaofen Xing and Xiangmin Xu. “Towards Zero-shot Human-Object Interaction Detection via Vision-Language Integration.” ArXiv abs/2403.07246 (2024): n. pag.

[9] D. Zong and S. Sun, "Zero-Shot Human–Object Interaction Detection via Similarity Propagation," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2023.3309104.

[10] Lei T , Yin S , Peng Y ,et al.Exploring Conditional Multi-modal Prompts forZero-Shot HOI Detection[C]//European Conference on Computer Vision.Springer, Cham, 2024.DOI:10.1007/978-3-031-73007-8_1.

[11] Hangjie Yuan, Jianwen Jiang, Samuel Albanie, Tao Feng, Ziyuan Huang, Dong Ni, and Mingqian Tang. 2024. RLIP: relational language-image pre-training for human-object interaction detection. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 2712, 37416–37431.

[12] F. Z. Zhang, Y. Yuan, D. Campbell, Z. Zhong and S. Gould, "Exploring Predicate Visual Context in Detecting of Human–Object Interactions," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 10377-10387, doi: 10.1109/ICCV51070.2023.00955.

[13] Y. Zhou, L. Liu and C. Gou, "Learning from Observer Gaze: Zero-Shot Attention Prediction Oriented by Human-Object Interaction Recognition," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 28390-28400, doi: 10.1109/CVPR52733.2024.02682.

[14] Sarullo A , Mu T .Zero-Shot Human-Object Interaction Recognition via Affordance Graphs[J]. 2020.DOI:10.48550/arXiv.2009.01039.

[15] Mascaro, E.V., Sliwowski, D., & Lee, D. (2023). HOI4ABOT: Human-Object Interaction Anticipation for Human Intention Reading Collaborative roBOTs. Conference on Robot Learning.

[16] Ni, Zhifan, Esteve Valls Mascar'o, Hyemin Ahn and Dongheui Lee. “Human-Object Interaction Prediction in Videos through Gaze Following.” Comput. Vis. Image Underst. 233 (2023): 103741.

1 Zero-Shot Cross-Lingual Transfer with Meta Learning, https://arxiv.org/abs/2003.02739