本文引用格式：钱政华,林剑峰,潘星,等.基于注视引导的复杂环境视频人-物交互检测模型[J].自动化与信息工程,2026,47(1):67-74.

QIAN Zhenghua, LIN Jianfeng, PAN Xing, et al. Video human-object interaction detection model with gaze guidance for complex environments[J]. Automation & Information Engineering, 2026,47(1):67-74.

基于注视引导的复杂环境视频人-物交互检测模型

钱政华[1] 林剑峰[1] 潘星[1] 陈辉[1] 陈静霖[1] 林宜轩[1] 王玉琛[2]

（1.华电新能源集团股份有限公司福建分公司，福建 福州 350000

2.南京南自信息技术有限公司，江苏 南京 210032）

摘要：针对基于视频的人-物交互检测存在的鲁棒性不足问题，提出基于注视引导的复杂环境视频人-物交互检测模型。该模型利用基于特征相似度的头身匹配算法，提高了头身匹配准确率；设计特征融合模块，将图像、文本和注视热力图特征进行有效融合。在 VidHOI 数据集上，本文模型全部类的 mAP 比性能最好的对比模型 HOIGF 提高了 0.84；在 Action Genome 数据集上，本文模型全部类的 Rec@10 比性能最好的对比模型 HOIGF 提高了 0.7。该模型提升了复杂环境下视频人-物交互检测的准确性，具有较好的鲁棒性。

关键词：人-物交互检测；注视引导；特征融合；特征相似度匹配

中图分类号：TP391.41 **文献标志码：**A **文章编号：**1674-2605(2026)01-0007-08

DOI：10.12475/aie.20260107 开放获取

Video Human-object Interaction Detection Model with Gaze Guidance for Complex Environments

QIAN Zhenghua[1] LIN Jianfeng[1] PAN Xing[1] CHEN Hui[1]

CHEN Jinglin[1] LIN Yixuan[1] WANG Yuchen[2]

(1.Huadian New Energy Group Co., Ltd., Fujian Branch, Fuzhou 350000, China

2.Nanjing Nanzi Information Technology Co., Ltd., Nanjing 210032, China)

Abstract: To address the robustness issues in video-based human-object interaction detection, a gaze-guided video human-object interaction detection model for complex environments is proposed. This model utilizes a head-body matching algorithm based on feature similarity to improve the accuracy of head-body matching. A feature fusion module is designed to effectively integrate image, text, and gaze heatmap features. On the VidHOI dataset, the proposed model achieved an improvement of 0.84 in mAP across all classes compared to the best-performing baseline model, HOIGF. On the Action Genome dataset, the proposed model achieved an improvement of 0.7 in Rec@10 across all classes compared to the best-performing baseline model, HOIGF. The model enhances the accuracy of video HOI detection in complex environments and demonstrates good robustness.

Keywords: human-object interaction recognition; gaze guidance; feature fusion; feature similarity matching

0 引言

在复杂环境中准确识别动态实体（如人类、动物等）与静态实体（如建筑物、家具等）之间的交互关系，对计算机视觉理解至关重要。人-物交互（human-object interaction, HOI）检测作为计算机视觉领域的一项关键技术，旨在定位图像或视频中的人类及其交互物体，并识别其交互类别，以深入理解人类行为。尽管基于图像的 HOI 检测技术[1-3]已较为成熟，但将其直接应用于视频 HOI 检测，效果往往有限。而现有基于视频的 HOI 检测技术[4-6]在复杂场景下仍面临诸多

AUTOMATION & INFORMATION ENGINEERING

挑战，如注视信息等交互线索未得到充分挖掘利用。已有研究表明[7-8]，注视信息在很多视觉检测任务中具有辅助作用，可进一步缩小人类与物体之间的定位偏差。

为此，本文提出基于注视引导的复杂环境视频 HOI 检测模型（简称复杂环境视频 HOI 检测模型）。该模型在空间-时间变换器（spatial-temporal transformer, STTran）[9]模型中引入文本和注视热力图信息，并设计特征融合模块，将注视热力图特征与目标检测特征有效融合；利用基于特征相似度的头身匹配算法，提高目光注视检测中头身匹配的准确率，从而解决复杂环境下视频 HOI 检测鲁棒性不足的问题。

1 相关工作

1.1 基于视频的 HOI 检测

在基于视频的 HOI 检测任务中，理解人-物关系随时间的演变至关重要。文献[10]将人-物关系表示为时空图，并采用结构递归神经网络（structural recurrent neural network, S-RNN）推断其交互类别，检测效果优于传统的线性处理方法，但需要预先定义并计算节点与边的结构。文献[11]在 S-RNN 基础上引入物体间的关联关系，进一步提升了检测准确率，但在复杂多物体场景下仍存在计算复杂度过高的问题。文献[12]提出一种基于实例的推理架构，通过单独推理每个人-物对实例，并利用人类骨架作为 HOI 检测的额外线索，提升了检测准确率，但其长时序建模能力有限。文献[4,13]提出的时空人-物交互（spatial-temporal human-object interaction, ST-HOI）模型将人体姿态特征与三维卷积网络提取的时空实例特征融合，并在 VidHOI 数据集上验证了其有效性，但对复杂环境下多个人-物目标的检测效果不佳。文献[6]提出的原型嵌入优化模型通过引入原型嵌入优化机制，缓解了直播中 HOI 检测出现的对象偏差问题，但其依赖复杂的原型库，导致模型泛化性能不足。

基于上述问题，本文利用基于特征相似度的头身匹配算法，即利用特征相似度构建代价矩阵，基于同一物体不同部位的特征相似性高于其他物体的特性，采用匈牙利二分匹配算法提高头身匹配精度，增强了人-物匹配的准确率，提升了复杂环境下视频 HOI 检测的鲁棒性。

1.2 基于 STTran 模型的 HOI 检测

近年来，研究人员[9,12,14]提出了基于 STTran 模型的视频 HOI 检测方法。文献[14]通过一个具有密集线性变换和时间编码器的 HOI 变换器，融合视觉外观特征和人体姿态特征，实现了基于 Transformer 的端到端 HOI 检测，提高了模型的检测精度；但其对大型数据集和计算资源依赖性较高，且收敛速度较传统 HOI 检测方法慢。STTran 模型将每个人-物实例的视觉外观特征与空间关系特征、语义特征进行串联，虽然加快了收敛速度，但由于输入模态单一，导致检测精度及鲁棒性不足。现有基于 STTran 模型的 HOI 检测模型通常仅以原始视频帧等单一模态特征作为输入，未能有效融合物体类别文本、注视热力图等多源特征，限制了模型在复杂环境下的适应能力。

2 复杂环境视频 HOI 检测模型

本文提出的复杂环境视频 HOI 检测模型主要包括目标检测器、目光注视检测器、特征融合模块、STTran 时空编码器-解码器 4 个部分，其结构如图 1 所示。

视频帧<I_1,I_2,…,I_n>被同时输入至目标检测器和目光注视检测器；目标检测器识别视频中的人和对应的物体，从每帧视频中获取人-物目标检测框{$b_{t,i}$}、目标类别{$cls_{t,j}$}、目标追踪特征O_j，其中，t 为时间帧，i、j 分别为第 i 个目标物体、物体类别 j；目光注视检测器利用基于特征相似度的头身匹配算法进行头身匹配，并通过注视跟踪模块获取注视热力图特征 $g_{t,i}$，从而得到人眼与物体的交互信息；特征融合模块将目标检测器与目光注视检测器输出的特征进行有效融合；STTran 时空编码器-解码器对融合特征进行处理，并生成 HOI 检测结果。

目标检测器采用单阶段目标检测模型 YOLOv5[15]进行目标检测，利用基于深度关联度量的简单在线实时跟踪（simple online and realtime tracking with a deep

AUTOMATION & INFORMATION ENGINEERING

association metric, DeepSORT）模型[16]实现目标追踪；STTran 时空编码器-解码器除在编码器输入端增加特征融合层外，其他主体结构与文献[9]基本一致。因此，本文主要介绍目光注视检测器和特征融合模块。

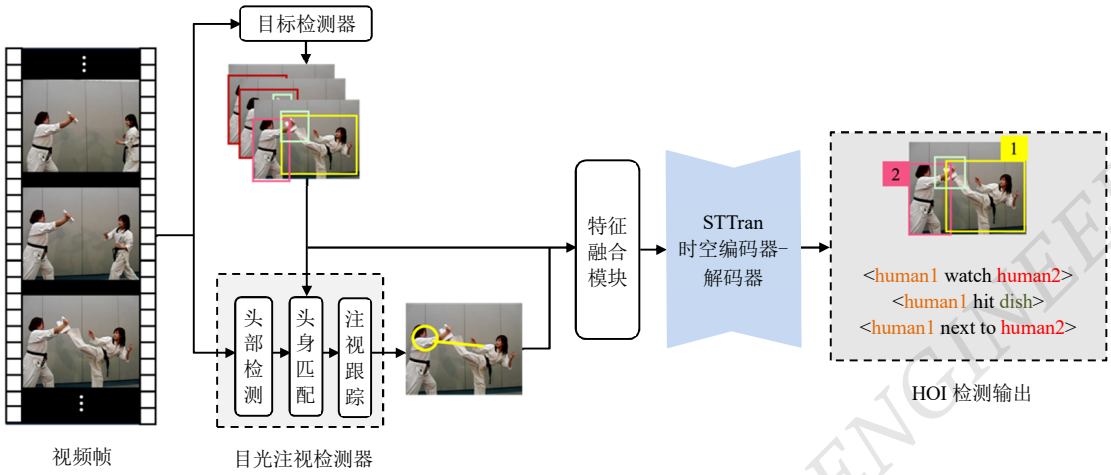


图 1　复杂环境视频 HOI 检测模型结构图

2.1　目光注视检测器

目光注视检测器主要包括头部检测、头身匹配、注视跟踪 3 个模块。其中，头部检测模块通过头部检测器识别单帧视频中的人体头部；头身匹配模块利用基于特征相似度的头身匹配算法获取头部与身体的匹配关系；注视跟踪模块采用文献[17]提出的目光跟踪方法为单帧视频中的每个人生成注视热力图特征。

目前，头身匹配算法主要分为两种：一种是在目标检测器的目标检测框内直接进行二次查找，但该算法易出现头身混淆，尤其在单帧视频中存在多人且相互遮挡时，可能头身出现匹配错误；另一种是先在单帧视频中获取头部检测框，再将其与身体检测框匹配，但匹配准则通常基于启发式方法，如以头部与身体检测框重叠区域的比值作为判断依据，鲁棒性不足，存在头部与身体错误匹配的风险。

本文通过实验发现，同一目标不同位置的特征相似度高于不同目标的特征相似度。基于该特性，本文提出一种基于特征相似度的头身匹配算法。该算法将头部与身体的检测框匹配看作二分匹配，并采用匈牙利二分匹配算法[18]计算头部特征与不同身体特征的相似度，从而构建代价矩阵，具体流程如下：

1） 利用目标检测器分别得到单帧视频中的头部检测框 $\{b_{t,i}^h\}$ 和身体检测框 $b_{t,k}$；

2） 分别提取头部特征 $f_i^h = F\left(b_{t,i}^h\right)$ 和人体去除头部的身体特征 $f_k = F\left(b_{t,k} - \{b_{t,i}^h\}\right)$；

3） 构建基于特征相似度的头身匹配算法的代价矩阵 $\boldsymbol{Cost} \in \mathbb{R}^{n\times m}$，其中，$n$为身体预测框数量，$m$为头部预测框数量，$Cost_{k,j} = 1 - \cos(f_k, f_j^h)$ 为利用匈牙利二分匹配算法得到的最佳头身匹配关系；

4） 根据文献[17]的方法，采用卷积长短期记忆网络[19]将匹配后的头部预测框与全局特征进行融合编码，以估计每个人的注视热力图特征 $g_{t,i}$。

2.2　特征融合模块

在每个时间帧 t 中，目标检测器对每个人-物对 $<b_{t,i},b_{t,j}>$ 输出一个目标检测框 $(cls_{t,j},b_{t,i},O_j)$；目光注视检测器输出其注视热力图特征 $g_{t,i}$。由于 STTran 时空编码器-解码器的输入包括图像、文本和注视热力图等 3 种不同模态的特征，且这些特征分布在不同的特征空间，因此需设计合适的特征融合模块进行编码。受 STTran 启发，本文设计特征映射与特征拼接相结合的特征融合模块，其结构及处理流程如图 2 所示。

AUTOMATION & INFORMATION ENGINEERING

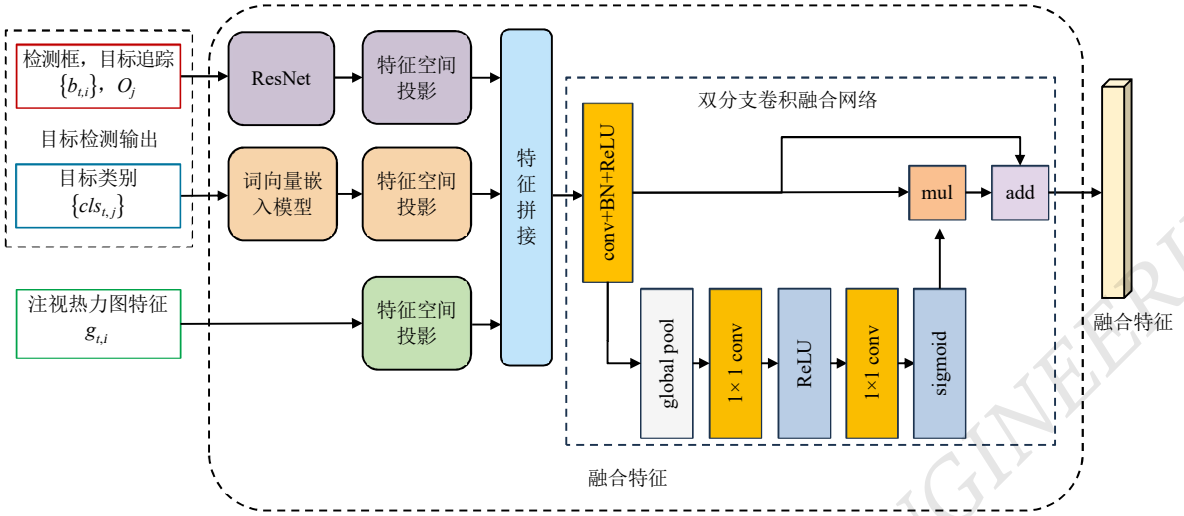


图 2　特征融合模块结构及处理流程

特征融合模块利用 ResNet[20]提取目标检测框$\{b_{t,i}\}$、目标追踪特征O_j的图像特征，并通过特征空间投影得到$x_{\text{img}} \in R^{d_i}$；利用词向量嵌入模型对目标类别$\{cls_{t,j}\}$文本特征进行编码，并通过特征空间投影得到$x_{\text{text}} \in R^{d_t}$；注视热力图特征$g_{t,i}$直接通过特征空间投影得到$x_{\text{gaze}} \in R^{d_g}$；将上述三类特征进行特征拼接，得到混合特征$x_{\text{fuse}} \in R^{d_i+d_t+d_g}$；通过本文设计的双分支（侧向分支与主干分支）卷积融合网络融合混合特征。其中，侧向分支主要由全局池化层（global pool）与1×1 卷积层（conv）等组成，用于生成特征动态权重，以强调重要特征并抑制不相关特征；主干分支主要由1×1 卷积层（conv）、归一化层（BN）和 ReLU 激活函数等组成，用于分析和转换混合特征，并利用侧向分支的特征动态权重进行特征筛选，以融合不同模态特征并降低特征维度。最终，将筛选后的特征与混合特征$x_{\text{fuse}} \in R^{d_i+d_t+d_g}$进行残差连接，输出当前视频帧中人-物对$< b_{t,i}, b_{t,j} >$的特征$X_{t,i,j} \in R^d$。通过对视频逐帧进行上述操作，可获取视频所有帧的人-物对的融合特征。

2.3　损失函数

在数据集中，一个人-物对$< b_{t,i}, b_{t,j} >$可能被同时标注多种交互关系，如<human1, next to & watch & wave bat>，属于典型的多类多标签分类问题。传统的多类多标签分类通常采用二元交叉熵损失[21]。然而，由于数据集呈长尾分布，存在严重的类别不平衡问题，本文采用类平衡 Focal 损失：

$$\text{CB}_{\text{Focal}}\left(p_j, y_j\right) = -\frac{1-\beta}{1-\beta^{n_j}}\left(1-p_{y_j}\right)^{\gamma}\log\left(p_{y_j}\right)$$

$$p_{y_j} = \begin{cases} p_j & y_j = 1 \\ 1-p_j & \text{otherwise} \end{cases}$$

式中：p_j为预测为第 j 个类别的概率，$y_j \in \{0,1\}$为真实标签，$\beta \in [0,1]$为一个可学习的参数，n_j为第j个类别的采样样本数，γ 为平衡超参数。

3　实验

3.1　数据集

3.1.1　VidHOI 数据集

VidHOI 数据集[4]收录了人类在无预定义脚本、复杂环境下的真实日常活动视频，并提供了完整的 HOI 三元组（<人，动作，物>）标注。该数据集采用关键帧注释，每秒采样 1 帧，共包含 78 个物体类别和 50 个谓词类别，其中 8 个谓词类别为空间关系，其余 42 个为动作关系。

AUTOMATION & INFORMATION ENGINEERING

VidHOI 数据集根据类别标签的数量可划分为：1） 全部，验证集涵盖所有 557 个 HOI 三元组类别；2） 稀有，验证集中少于 25 个实例的 315 个类别；3） 非稀有，验证集中不少于 25 个实例的 242 个类别。本文在这 3 个类别集上分别进行实验，以验证所提模型的可靠性。

3.1.2 Action Genome 数据集

Action Genome[22]是一个包含 35 个物体类别和 25 种交互类别的视频数据集，每个视频中仅标注了单人的 HOI 三元组。

3.2 评估指标

对于 VidHOI 数据集，本文采用平均精度均值（mAP）作为模型的评估指标，mAP 越高，表明模型的检测精度越高。

对于 Action Genome 数据集，为了与现有基准模型进行公平对比，本文采用召回率（Rec@K）作为评估指标，Rec 越高，表明模型的检测精度越高。

预测的 HOI 三元组被视为真正正例的条件包括：

1） 检测到的人-物体目标检测框与真实标注检测框的交并比（IoU）大于 0.5；

2） 预测的物体类别正确；

3） 预测的 HOI 关系正确。

本文遵循 ST-HOI 基准中定义的两种评估模式：一种为真实值模式（Oracle），即利用真实物体边界框进行评估；另一种为预测模式（Detection），即利用 YOLOv5 模型预测输出的物体边界框进行评估。在预测模式下，ST-HOI 基准去除未检测到物体的视频帧以提高召回率。但本文不去除任何视频帧，在未检测到物体的视频帧中，所有真实 HOI 均被视为假负例。

3.3 实验环境

本文选取 YOLOv5[15]模型作为目标检测器，用于提取单帧视频中所有目标检测框及其对应的类别，基于 COCO 数据集预训练其权重并进行初始化，且通过 VidHOI 数据集进行微调；采用预训练的 DeepSORT 模型[16]作为目标追踪器，追踪不同视频帧之间的相同目标。

在目光注视检测器中，本文采用 YOLOv5 模型从单帧视频中检测人体头部，头部检测器在 Crowdhuman 数据集[23]上进行预训练；采用文献[17]的目光跟踪方法生成注视热力图特征。

在特征融合模块中，采用 ResNet-101 模型[20]提取目标检测框$\{b_{t,i}\}$、O_j的图像特征；利用 GloVe 模型[24]进行词向量嵌入；特征空间投影层利用 1×1 卷积进行映射，损失函数采用类平衡 Focal 损失。

STTran 时空编码器-解码器的前馈网络维度为 2 048，多头注意力层的头数为 8，空间编码器为 1 层，时间解码器为 3 层，时间窗口长度设为 6。$\mathrm{CB}_{\mathrm{Focal}}$中的$\gamma = 0.5$、$\beta = 0.999\,9$，以缓解数据样本分布不均衡的问题。为加快模型训练速度，本文使用小批量学习，模型采用 AdamW 优化器，设定初始学习率为1×10^{-8}，峰值学习率为1×10^{-4}，指数衰减为 0.1，预热阶段为 3 轮，权重衰减因子为1×10^{-2}，Dropout 率为 0.1，所有训练运行 25 个 epoch，并设置固定的随机种子，以确保实验结果的可重复性。

此外，本文通过随机水平翻转增强数据样本。本文模型的超参数在 VidHOI 数据集上进行了微调。在 Action Genome 数据集的实验中，直接沿用与 VidHOI 数据集相同的设置。

实验基于 Python 3.9 环境，在 NVIDIA RTX 4090 GPU 上完成，采用 PyTorch 框架构建与训练模型，YOLOv5 模型基于 Ultralytics 实现。

3.4 实验过程及结果

真实值模式（Oracle）可排除目标检测误差的干扰，以验证本文模型 HOI 检测的有效性；预测模式（Detection）可模拟真实的视频分析场景，用于评估本文模型在复杂环境下端到端的综合检测性能。在 VidHOI 数据集上，本文模型与图解析神经网络（graph parsing neural network, GPNN[25]）、人-物交互变换器（human-object interaction transformer, HOTR[14]）、ST-

AUTOMATION & INFORMATION ENGINEERING

HOI[4]、STTran[9]、人-物注视交互（human-object interaction with gaze following, HOIGF[26]）模型在真实值模式（Oracle）、预测模式（Detection）下进行对比实验，实验结果如表 1、2 所示。

表 1　VidHOI 数据集上 Oracle 模式下的实验结果

模型	mAP		
	全部类	非稀有类	稀有类
GPNN	18.47	24.50	16.41
HOTR	21.14	30.75	19.83
ST-HOI	17.60	27.20	17.30
STTran	28.32	42.08	17.74
HOIGF	38.61	52.44	27.99
本文模型	**39.45**	**54.17**	**28.16**
	（+0.84）	**（+1.73）**	**（+0.17）**

由表 1 可知，在 Oracle 模式下，本文模型全部类的 mAP 比性能最好的对比模型 HOIGF 提高了 0.84，表明在排除目标检测误差的情况下，本文模型 HOI 检测有效。

表 2　VidHOI 数据集上 Detection 模式下的实验结果

模型	目标检测器	mAP		
		全部类	非稀有类	稀有类
ST-HOI	Detectron2	3.10	5.90	2.10
STTran	YOLOv5	7.61	13.18	3.33
HOIGF	YOLOv5	10.40	16.83	5.46
本文模型	Detectron2	9.33	15.44	4.62
	YOLOv5	**11.46**	**18.61**	**5.97**
		(+1.06)	**(+1.78)**	**(+0.51)**

由表 2 可知，在 Detection 模式下，本文模型全部类的 mAP 比性能最好的对比模型 YOLOv5 提高了 1.06，表明目标检测器在 HOI 检测中起着重要作用。

为进一步验证本文模型在不同数据分布下的泛化能力，在 Action Genome 数据集上进行对比实验。实验以关系检测网络（relationship detection network，RelDN[27]）、图属性感知网络（graph property sensing network，GPS-Net[28]）等模型作为对比模型，在 Oracle 模式下进行评估，实验结果如表 3 所示。

表 3　Action Genome 数据集上 Oracle 模式的实验结果

模型	Rec		
	Rec@10	Rec@20	Rec@50
RelDN	70.7	78.8	80.3
GPS-Net	71.3	81.2	82.0
STTran	73.2	83.1	84.0
HOIGF	75.4	83.7	84.3
本文模型	**76.1**	**84.1**	**84.5**
	（+0.7）	**（+0.4）**	**（+0.2）**

由表 3 可知，在 Action Genome 数据集上，本文模型的 Rec@10 比性能最好的对比模型 HOIGF 提高了 0.7，表明本文模型在不同数据源上具有较好的泛化性能。

为验证不同模块的作用，本文在 VidHOI 数据集、Oracle 模式、HOIGF 模型上进行消融实验，分别利用基于特征相似度代价矩阵、特征融合模块替换 HOIGF 模型的对应区域，实验结果如表 4 所示。

表 4　VidHOI 数据集上 Oracle 模式下不同模块效果

设定	mAP		
	全部类	非稀有类	稀有类
HOIGF	38.61	52.44	27.99
+特征相似度代价矩阵	38.98	54.16	28.11
+特征融合	38.84	52.96	28.03
HOIGF +特征相似度代价矩阵+特征融合	**39.45**	**54.17**	**28.16**

由表 4 可知，特征相似度代价矩阵、特征融合模块均能有效提升模型检测精度，且通过基于特征相似

AUTOMATION & INFORMATION ENGINEERING

度代价矩阵得到的注视热力图特征，经特征融合模块与目标检测特征融合后，模型检测精度提升更为显著，表明本文提出的特征相似度代价矩阵、特征融合模块均是有效的。

4 结论

本文针对复杂环境下 HOI 检测鲁棒性不足的问题，提出基于注视引导的复杂环境视频 HOI 检测模型。该模型通过引入基于特征相似度代价矩阵与特征融合模块，提升了模型在复杂环境下的鲁棒性，在 VidHOI、Action Genome 数据集上均取得了较好的性能。同时，通过在 Oracle 和 Detection 模式下的实验结果发现，模型性能受目标检测器影响较大，未来可进一步探索如何利用更高效的目标检测器来提升 HOI 检测性能。

©The author(s) 2024. This is an open access article under the CC BY-NC-ND 4.0 License (https://creativecommons.org/licenses/by-nc-nd/4.0/)

参考文献

[1] GUPTA S, MALIK J. Visual semantic role labeling[J]. arXiv preprint arXiv:1505.04474,2015.

[2] LU C, KRISHNA R, BERNSTEIN M, et al. Visual relationship detection with language priors[C]//European conference on computer vision. Cham: Springer International Publishing,, 2016:852-869.

[3] 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.

[4] CHIOU M J, LIAO C Y, WANG L W, et al. ST-HOI: A spatial-temporal baseline for human-object interaction detection in videos[C]//Proceedings of the 2021 ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, 2021:9-17.

[5] TU D, SUN W, MIN X, et al. Video-based human-object interaction detection from tubelet tokens[J]. Advances in Neural Information Processing Systems, 2022,35:23345-23357.

[6] ZHANG M, ZHANG J, CHEN L, et al. Prototype embedding optimization for human-object interaction detection in livestrea-ming[J]. arXiv preprint arXiv:2505.22011,2025.

[7] SONG Y, WANG X, YAO J, et al. Vitgaze: gaze following with interaction features in vision transformers[J]. Visual Intelligence, 2024,2(1):31.

[8] GENG P, YANG J, ZHANG S. HORP: Human-object relation priors guided HOI detection[C]//Proceedings of the Computer Vision and Pattern Recognition Conference, 2025: 25325-25335.

[9] CONG Y, LIAO W, ACKERMANN H, et al. Spatial-temporal transformer for dynamic scene graph generation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021:16372-16382.

[10] JAIN A, ZAMIR A R, SAVARESE S, et al. Structural-RNN: Deep learning on spatio-temporal graphs[C]//Proceedings of the IEEE Conference On Computer Vision And Pattern Recog-nition, 2016:5308-5317.

[11] TRUONG A M, YOSHITAKA A. Structured LSTM for human-object interaction detection and anticipation[C]//2017 14th IEEE international Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2017:1-6.

[12] JI J, DESAI R, NIEBLES J C. Detecting human-object relationships in videos[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 8106-8116.

[13] SUN X, HE Y, REN T, et al. Spatial-temporal human-object interaction detection[J]. 2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021:1-6.

[14] 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.

[15] JOCHER G, CHAURASIA A, STOKEN A, et al. ultralytics/yolov5: v7.0 YOLOv5 SOTA Realtime Instance Segmentation:v7.0[DB/OL].(2022-11-23).https://doi.org/10.5281/zenodo.7347926.

[16] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017:3645-3649.

[17] CHONG E, WANG Y, RUIZ N, et al. Detecting attended visual targets in video[C]//Proceedings of the IEEE/CVF Conference

AUTOMATION & INFORMATION ENGINEERING

on Computer Vision and Pattern Recognition, 2020:5396-5406.

[18] KUHN H W. The Hungarian method for the assignment problem[J]. Naval research logistics quarterly, 1955, 2(1-2): 83-97.

[19] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: A machine learning approach for precipitation now-casting[J]. Advances in neural information processing systems, 2015,28.

[20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-778.

[21] CUI Y, JIA M, LIN T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition, 2019:9268-9277.

[22] JI J, KRISHNA R, FEI-FEI L, et al. Action genome: Actions as compositions of spatio-temporal scene graphs[C]//Proceed-ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:10236-10247.

[23] SHAO S, ZHAO Z, LI B, et al. Crowdhuman: A benchmark for detecting human in a crowd[J]. arXiv preprint arXiv: 1805.00123,2018.

[24] 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.

[25] 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.

[26] NI Z, MASCARÓ E V, AHN H, et al. Human–object interaction prediction in videos through gaze following[J]. Computer Vision and Image Understanding, 2023, 233: 103741.

[27] ZHANG J, SHIH K J, ELGAMMAL A, et al. Graphical contrastive losses for scene graph parsing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:11535-11543.

[28] LIN X, DING C, ZENG J, et al. GPS-net: Graph property sensing network for scene graph generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:3746-3753.

作者简介：

钱政华，女，1986 年生，本科，高级工程师，主要研究方向：电力生产技术管理、通讯网络工程技术和电力自动化控制系统等。E-mail: 12048172@qq.com

林剑峰，男，1972 年生，专科，工程师，主要研究方向：安全生产管理。E-mail: 1240349599@qq.com

潘星，女，1995 年生，本科，工程师，主要研究方向：电气工程与自动化、集控运行。E-mail: 527274387@qq.com

陈辉，男，1990 年生，本科，工程师，主要研究方向：安全管理。E-mail: 1240349599@qq.com

陈静霖，女，1990 年生，本科，工程师，主要研究方向：信息运维。E-mail: 502886290@qq.com

林宜轩，男，2002 年生，本科，助理工程师，主要研究方向：电气工程及其自动化。E-mail: chenjinglin@chd.com.cn

王玉琛（通信作者），男，1999 年生，硕士研究生，工程师，主要研究方向：计算机视觉。E-mail: 369802364@qq.com

声 明

本刊从 2024 年 7 月 1 日开始为开放获取期刊，所有录用和发表的论文可以立即并永久供读者免费阅读、下载、链接。本刊采用知识共享-署名-非商业-禁止演绎许可协议（Creative Commons Attribution-NonCommercial-No--Derivs），允许第三方用户按照（CC-BY-NC-ND 4.0），即署名（BY）、非商业用途（NC）、禁止演绎（ND）的方式共享，前提是原作未以任何方式进行更改，并且被适当引用。详情请访问 https://creativecommons.org/ licenses/by--nc-nd/4.0/。

AUTOMATION & INFORMATION ENGINEERING