时空语义驱动的渐进多视角行为去偏置研究

钟忺[1]，陈亮[1]，刘文璇[1*]，叶舒[1]，江奎[2]，王正[3]，林嘉文[4]

（1. 武汉理工大学计算机与人工智能学院，湖北 武汉 430070；
2. 哈尔滨工业大学计算机科学与技术学院，黑龙江 哈尔滨 150001；
3. 武汉大学计算机学院，湖北 武汉 430072；4. 台湾清华大学电机工程系，台湾 新竹 300044）

摘 要： 在实际应用中，单视角摄像头采集数据由于物体存在遮挡而失去对某些区域的可见性，因此结合多个视角下的数据进行行为分析对于维护社会稳定及民生安全至关重要。针对多视角行为识别中存在的偏置问题，即不同视角下空间语义不一致导致的视角间行为表征差异以及同一行为执行过程中的时序语义不一致导致的行为表征差异，提出一种渐进去偏置的多视角方法。首先，在多视角下的同一行为样本中以证据理论为引导，结合不同视角下的行为同构性进行视角间行为去偏置，优化不同视角下关注的行为特征权重，以获得更全面的无偏行为表示。其次，结合多粒度解耦策略，分析不同粒度对行为特征无偏表达的影响，准确分离行为相关和行为无关特征，以避免视角内行为无关信息扰乱行为表征导致的显著差异。最后，在时序维度上构建不同行为特征权重，增强同一视角内行为特征一致性，减弱同一行为的行为表征差异。在多个数据集上的实验结果验证了所提方法的有效性，在 N-UCLA 和 NTU-RGB+D 数据集上的跨视角准确率分别达到了 97.4%和 96.4%，并且所提方法在满足多视角下对行为识别进行准确分析应用需求的同时通过一种新的去偏置思路为多视角行为识别问题提供了一种有效的解决方案。

关键词： 多视角行为识别；渐进式去偏置；证据理论；解耦；多粒度

中图分类号： TP391.41 **文献标志码：** A **DOI：** 10.19678/j.issn.1000-3428.0069307

Research on Temporal-Spatial Semantic-Driven Progressive Multiview Action Debiasing

ZHONG Xian[1], CHEN Liang[1], LIU Wenxuan[1*], YE Shu[1], JIANG Kui[2], WANG Zheng[3], LIN Chia-Wen[4]

(1. School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, Hubei, China;
2. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, Heilongjiang, China;
3. School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China;
4. Department of Electrical Engineering, Taiwan Tsing Hua University, Hsinchu 300044, Taiwan, China)

【Abstract】 In practical applications, data collected from single-view cameras often lose visibility in certain areas due to object occlusion. Therefore, analyzing data from multiple views is crucial for maintaining social stability and public safety. To address the bias in multiview action recognition, arising from spatial semantic inconsistencies among different views and temporal semantic disparities during the execution of the same action, a multiview progressive debiasing method is proposed. First, guided by evidence theory within the context of multiple views for the same sample, the method leverages isomorphism across different views to mitigate inter-view bias. This involves optimizing the weights of features from different views to obtain a more comprehensive and unbiased representation. Second, employing a multi-granularity decoupling strategy, the method analyzes the impact of different granularities on debiased expression of features, thereby accurately separating relevant and irrelevant features while avoiding significant differences in representation caused by irrelevant information within a single view. Finally, the method constructs different feature weights along the temporal dimension, enhancing consistency in features within the same view and mitigating representation disparities for the same sample. The effectiveness of the proposed method is validated on multiple datasets, achieving cross-view accuracy rates of 97.4% and 96.4% on the N-UCLA and NTU-RGB+D datasets, respectively. This method not only meets the requirements for accurate recognition analysis under multiple views but also provides an effective solution to the bias problem in multiview recognition from a novel debiasing perspective.

【Key words】 multiview action recognition; progressive debiasing; evidence theory; decoupling; multi-granularity

收稿日期： 2024-01-26 **修回日期：** 2024-04-19

基金项目： 国家自然科学基金(62271361)。

通信作者 E-mail： * lwxfight@whut.edu.cn

0 引言

不同于传统的行为识别任务[1-3]，多视角行为识别旨在利用多个视角的信息对行为进行有效识别。随着视频监控在公安、人机交互等领域的普及，不同摄像头采集的信息包含着多样的观察角度和行为特征，有效利用这些信息可以提升各种场景中对行为的理解和识别能力，同时具备为相关行业赋能的潜力，由此创造更安全、智能的生活和工作环境[4-5]。

先前的一些多视角行为识别方法[6]聚焦于行为本身，通过学习多个视角下行为的聚类特征，力图构建视角不变的行为表示。基于同样的想法，部分研究[7]利用不同模态数据的特点来学习更精确的行为特征。然而，这些方法注重于捕获行为信息，忽视了视角变化对行为表征的影响。这可能会导致关键信息的丢失，从而影响模型的整体性能。因此，近期的研究[8-9]开始关注除行为信息以外的视角信息，通过应用解耦策略分离行为和视角信息，显著提升了识别效果。尽管如此，这些方法仍面临挑战，即如何应对多视角任务中时空、场景变化导致的行为表征差异带来的偏置问题。

受到偏置问题在其他任务定义中的启发，需要重新从时空、语义双角度对多视角行为识别中的偏置问题进行定义。在不同视角下采集的同一行为数据涵盖多个角度，场景空间变化降低了模型对行为的关注能力。同时，对于同一视角内的数据，人类行为的高度复杂性和连续性导致不同时刻行为的表现也不同，模型对视角内不同时序片段下行为的辨别能力存在差异。结合多视角任务的本质，将多视角行为识别中的偏置定义为空间语义不一致导致的视角间偏置和时序语义不一致导致的视角内偏置，例如，在使用两个不同视角的数据进行训练时，模型可能会偏向于场景语义信息已经记住的视角，这导致学习到的行为信息实际上包括场景的空间语义信息，且在进行测试时，无法有效地识别视角切换后未知场景下的行为，这限制了模型的应用范围和准确性。视角内的偏置是指模型在处理特定视角下的数据时，无法处理在执行过程中时序变化导致的行为差异。

由于多视角行为识别中的偏置现象与多视角的场景空间语义信息和单视角下的行为时序语义信息密切相关，这些时空语义信息既可以辅助行为表征，又可能对模型的判断力造成干扰，因此去除多视角行为识别的偏置是本文研究的核心内容。

基于此，本文提出时空语义驱动的渐进多视角行为去偏置方法。利用不同视角下同一行为的同构性，以证据理论为引导，优化行为特征权重，以获得全面的无偏行为表示。其中，证据理论和不确定性分布将各视角下的无偏特征进行动态加权融合，通过互相学习不同视角下的空间语义信息以加强空间一致性进而减弱视角间的偏置。结合粗细粒度的解耦方法，分析不同粒度对行为表征的影响，使模型关注行为本身的同时准确分离行为相关和行为无关特征，以避免视角内行为无关信息扰乱行为表征导致的显著差异，尽可能消除视角内行为无关信息带来的偏置，并基于此在时序维度上构建不同时刻下行为特征的权重，增强同一视角内行为时序语义特征一致性，减弱同一行为的表征差异。时空语义去偏通过协同提升时空语义一致性，解决了多视角中的偏置问题，提高了行为识别的准确率。

综上所述，本文的贡献可概括为以下 3 个方面：

1）重新定义了多视角行为识别任务中的偏置问题，分析了由时空语义变化导致的行为表征差异带来的偏置对行为识别产生的负面影响，为多视角下的其他任务提供了一种去偏置的新思路。

2）以去偏置为基础构建了一种时空语义驱动的渐进多视角行为去偏置方法，协同并行解决视角间、视角内的偏置问题。通过证据理论综合不同视角下的无偏特征，解决了视角间的偏置问题。结合解耦与时序一致性，避免了视角内行为无关信息带来的偏置，加强了视角内行为相关信息的时空一致性。

3）通过在多个数据集上进行广泛验证，结果表明所提方法在多视角行为识别的 N-UCLA 和 NTU-RGB+D 数据集上的准确率取得明显提升，证明了其有效性和优越性。

1 相关工作

1.1 多视角行为识别

多视角行为识别作为行为识别领域的延伸已成为研究的热点。ZHAO 等[6]通过最大化跨视角的对比性互信息实现行为表示的视角不变性，确保在特征空间中不同视角的行为表示得到对齐。LIU 等[9]通过解耦的方式提取出视角特定信息和行为特定信息，提高了识别的准确率。詹健浩等[10]利用多教师知识蒸馏的方法来融合多模态数据以提高识别准确率。施海勇等[11]通过质心运动路径松弛算法降低深度图像冗余，并提出新的时空特征表示方法，以有效融合深度和骨骼数据。LI 等[12]引入了对比学习的思想，通过比较同一行为在不同视角下的差异，使模型能够识别不同行为并形成视角不变的特征表达。SHAO 等[13]引入了“骨骼自相似性”概念，

以捕捉不同视角下相同行为时的骨骼结构的相似性。BAHRAMPOUR 等[14]提出了联合字典学习的方法，使用联合稀疏约束来平衡不同视角特征的贡献。在此基础上，LIU 等[15]引入了任务驱动的方法，其中字典和分类器在联合稀疏约束下同时进行训练，以实现更优的模型构建。然而，这些方法忽略了多视角任务中固有的偏置问题。

1.2 行为识别中的偏置

偏置是行为识别领域的关键挑战之一。LI 等[16]正式定义了表征偏置的概念，并提出了一种在数据集校准过程中最小化偏置的方法。KIM 等[17]提出了一种创新的正则化算法，通过最小化特征提取器提取的特征与偏置信息之间的互信息，让网络学习不受偏置影响的特征表示。CHOI 等[18]在基于对抗学习的方法中引入对抗性损失，促使网络忽略场景信息，专注于行为。此外，BAHNG 等[19]通过学习偏置来减少偏置，激励模型远离一系列有意识偏置的表现。基于此，BAO 等[20]引入了对比学习，通过对比时间和空间偏置特征，网络学习了强调视角的无偏特征。DUAN 等[21]通过对抗性训练和对空间行为的重新加权降低偏置，选择性利用网络数据以中和偏置。不同于此，LI 等[22]考虑了 RGB 和骨骼数据的特点，通过骨骼数据引导特征融合模块使模型关注行为信息。在这些研究的基础上，本文研究专注于解决多视角行为识别中的偏置问题。空间上视角的变化和时间上行为的变化带来的行为表征差异所引起的偏置问题导致多视角行为识别任务面临额外的挑战。

2 渐进式去偏置方法

2.1 概述

如图 1 所示，构建一个渐进式时空语义驱动的行为去偏置方法，通过逐层递进的方式解决多视角行为识别中的偏置问题。模型学习各个视角的无偏特征并融合以减弱视角间的偏置，通过解耦分离各视角内的行为相关信息和行为无关信息，减轻视角内行为无关信息产生的偏置对模型的误导，在解耦的基础上，对获得的行为相关信息进行时序语义一致性的加强训练和加权融合，降低视角内由行为相关信息产生的偏置影响。其中，解耦模块及之后的无偏特征优化(UFO)模块和证据融合(ET)模块有效解决了视角内的偏置问题。

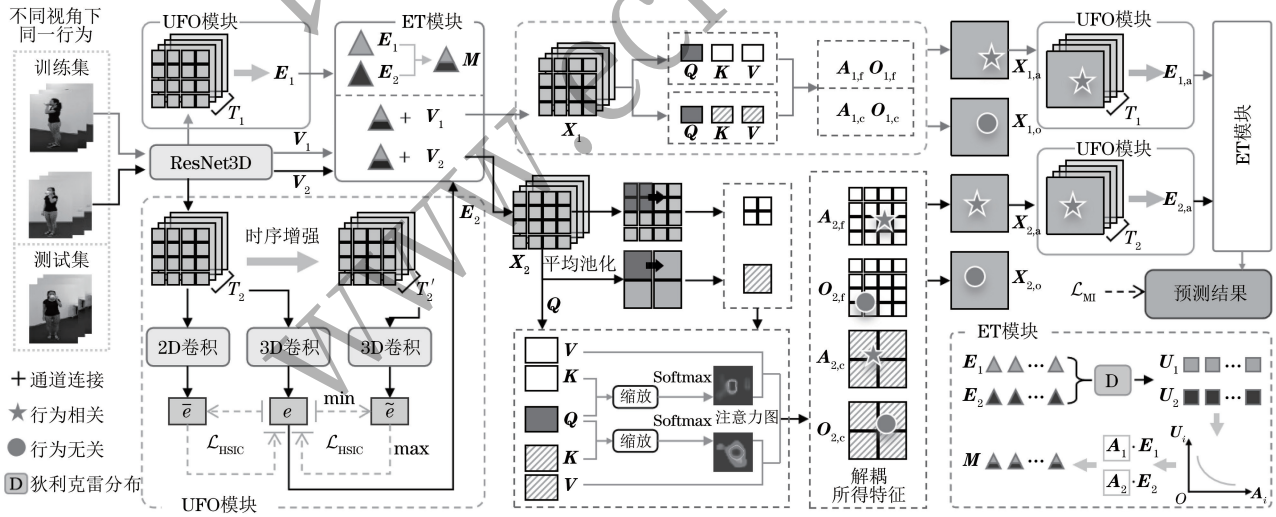


图 1 渐进式去偏置方法框架

Fig. 1 Framework of progressive debiasing method

2.2 视角间空间语义一致性去偏置

为了避免对某个视角信息过于依赖，使模型能够有效学习不同视角下的特征，提高泛化能力，提出了 UFO 和 ET 模块。UFO 通过学习有偏特征为每个视角获取无偏特征，ET 为来自不同视角的无偏特征动态地分配权重，并通过加权融合实现全面的特征表达。UFO 和 ET 的配合有效地减少了视角间的偏置。

2.2.1 UFO 模块

在视频识别任务中，偏置不仅源自行为本身，还受到上下文时序变化的影响。为此，将 UFO 模块设计为三分支结构：一个分支用于学习无偏特征，其他两个分支分别负责学习与时间序列和行为本身相关的有偏特征。最终目标是挖掘不同视角下行为的同构性，促进无偏特征的学习。

在三分支结构中：中间的 Conv3D 分支用于学习无偏特征 e；为了捕获视频数据中由时间动态导致的有偏特征 $\tilde{e}$，UFO 在另一个 Conv3D 分支中采用了数据增强技术，通过随机输入的方法将时间序

列进行重排，提高模型学习时间多样化信息的能力；第 3 个分支利用 Conv2D 网络层来捕获行为本身固有的有偏特征 $\bar{e}$，该分支与中间分支共享相同的输入，但为了适应 Conv2D 结构，将训练批次与时间序列进行了合并。

使用 Hilbert-Schmidt 独立性准则(HSIC)[23]平衡无偏特征和有偏特征之间的独立性和关联性。HSIC 用于评估两个变量之间的独立性，其值与变量间的独立程度成反比，并通过式(1)对特征进行约束：

$$\begin{aligned}&\mathrm{HSIC}(e,e^{\mathrm{b}};\theta_1)=\\&\mathrm{HSIC}(e,\bar{e};\theta_1)+\mathrm{HSIC}(e,\tilde{e};\theta_1)\\&\mathrm{HSIC}(e^{\mathrm{b}},e;\theta_2)=\\&\mathrm{HSIC}(\bar{e},e;\theta_2)+\mathrm{HSIC}(\tilde{e},e;\theta_2)\end{aligned}\tag{1}$$

式中：e^{b} 表示有偏特征，$e^{\mathrm{b}}\in\{\bar{e},\tilde{e}\}$；$\theta_i$ 代表独立性参数。

式(1)中的第 1 个公式的值应尽可能小，使所学的无偏特征 e 远离有偏特征 $\bar{e}$ 和 $\tilde{e}$，保证无偏特征的独立性；第 2 个公式的值应尽可能大，使得有偏特征 $\bar{e}$ 和 $\tilde{e}$ 与无偏特征 e 具有强关联性，保证模型在有偏特征中加强对于视角间空间一致性的判断。

2.2.2 证据融合模块

由于不同视角对同一行为的观察存在不一致性导致每个视角对行为识别的贡献不均衡，因此在处理不同视角的特征时不能简单平等地对待每个视角。该模块结合了狄利克雷证据理论和不确定性分布的思想，通过狄利克雷分布引导证据的不确定性分布，并进一步利用这种不确定性动态地为特征分配权重。ET 模块首先将从 UFO 模块获得的无偏特征 $\boldsymbol{E}=[e_1,e_2,\cdots,e_n]$ 作为证据源，对第 i 个视角的样本 $\boldsymbol{V}_i$ 的无偏特征 e_i，α_i 通过等式 $\alpha_i=e_i+1$ 与证据 e_i 相关联，并使用 α_i 的总和代表狄利克雷强度 D_i，不确定性 u_i 的计算如式(2)所示：

$$u_i=\frac{K}{D_i}\tag{2}$$

ET 根据 u_i 为每个视角动态地分配权重，不确定性越高的视角权重越低，反之亦然。然后对来自不同视角的证据进行加权融合并通过级联获得最终行为特征的全面表达 $\boldsymbol{M}_i$，如式(3)所示：

$$\boldsymbol{M}_i=\sum_{i=1}^{n}\frac{1-u_i}{n-\sum_{i=1}^{n}u_i}\boldsymbol{V}_i\tag{3}$$

受到证据深度学习(EDL)[24]的启发，引入以下损失作为约束，以促进证据学习，如式(4)所示：

$$\mathcal{L}_{\mathrm{ET}}=\sum_{k=1}^{K}y_i(\ln D_i-\ln\alpha_i)\tag{4}$$

式中：y_i 是样本 $\boldsymbol{V}_i$ 的一维 K 类标签。

2.2.3 空间一致性协同约束

为了更有效地解决视角间的偏置问题，对 UFO 和 ET 进行了联合优化，以促进模块间的协同协作。此阶段的目标是最小化 $\mathcal{L}_{\mathrm{db}}$，如式(5)所示：

$$\mathcal{L}_{\mathrm{db}}=\mathcal{L}_{\mathrm{ET}}+\mathrm{HSIC}(e,e^{\mathrm{b}};\theta_1)-\mathrm{HSIC}(e^{\mathrm{b}},e;\theta_2)\tag{5}$$

2.3 视角内行为无关信息去偏置

最近的一些研究表明：多尺度信息的结合可以有效地增强特征的表示能力[25-26]。首先结合粗粒度和细粒度策略对多视角下的特征进行解耦，将视角内特征视为行为相关和行为无关信息的组合特征，然后从这些特征中提取行为相关信息，以减弱由行为无关信息引起的偏置的影响。通过解耦将行为相关和行为无关特征从初始特征中分离出来，以不同粒度保留特征的多尺度完整性。

本文引入滑动窗口和多头自注意力机制。滑动窗口大小设置为 2×2 像素，以不重叠的方式移动，同时使用零填充的方法保证滑动窗口能完全遍历特征，避免边缘信息的丢失。对于多头自注意力，将头的数量设置为 H，以 r 作为头部分割比将其分为两组，其中，rH 用于粗粒度分支，$(1-r)H$ 用于细粒度分支。为了在解耦过程中更准确地提取与行为相关的特征，使用从 ET 处获得的共同聚类特征 $\boldsymbol{M}_i$ 作为引导，将 $\boldsymbol{M}_i$ 与 $\boldsymbol{V}_i$ 在通道上进行拼接得到的 $\boldsymbol{X}_i$ 用作模块的输入。

在细粒度和粗粒度分支中，虽然使用相同的滑动窗口，但自注意力机制的实现存在差异。在细粒度分支中，滑动窗口使模型能够更精确地关注复杂的局部特征。从特征 $\boldsymbol{X}_i$ 中获得自注意力机制的查询 $\boldsymbol{Q}$，从滑动窗口内的特征中获得键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$。这种设计有效地捕获了原始特征图中潜藏的复杂细节，有助于更精细地提取局部特征。在粗粒度分支中，在每个滑动窗口内执行平均池化操作，以更好地考虑局部特征的全局依赖性。值得注意的是，查询 $\boldsymbol{Q}$ 同样来自特征 $\boldsymbol{X}_i$，但键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 则来自经过窗口内平均池化后的特征。细粒度和粗粒度特征的注意力输出的计算如式(6)所示：

$$S(\boldsymbol{X}_i)=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{D_{\mathrm{h}}}}\right)\boldsymbol{V}\tag{6}$$

式中：D_{h} 代表隐藏的头数量；$S(\boldsymbol{X}_i)$ 代表经过自注意力处理后获得的特征，它是通过对 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的点积进行缩放后应用 Softmax 计算得出的，然后通过 $\boldsymbol{V}$ 计算的注意力权重加权；$S_{\mathrm{f}}(\boldsymbol{X}_i)$ 和 $S_{\mathrm{c}}(\boldsymbol{X}_i)$ 分别表示第 i 个视角的细粒度和粗粒度特征。

通过对相关性图和相应的原始特征 $\boldsymbol{V}_i$ 进行逐元素乘法进一步得到行为特征的相关性图，获得的细粒度行为相关特征和粗粒度行为相关特征分别用 $\boldsymbol{A}_{i,\mathrm{f}}$ 和 $\boldsymbol{A}_{i,\mathrm{c}}$ 表示，如式(7)所示：

$$\begin{aligned}\boldsymbol{A}_{i,\mathrm{f}}&=\sigma(W_{\mathrm{f}}(S_{\mathrm{f}}(\boldsymbol{X}_i)))\odot\boldsymbol{V}_i\\ \boldsymbol{A}_{i,\mathrm{c}}&=\sigma(W_{\mathrm{c}}(S_{\mathrm{c}}(\boldsymbol{X}_i)))\odot\boldsymbol{V}_i\end{aligned}\tag{7}$$

式中：σ 表示 Sigmoid 激活函数；W_{f}、W_{c} 是经过批量标准化的细粒度和粗粒度信息流的权重，经过线性整流函数(ReLU)层和 1×1 卷积层的精炼处理；⊙表示逐元素乘法。

对 $\boldsymbol{A}_{i,\mathrm{f}}$ 和 $\boldsymbol{A}_{i,\mathrm{c}}$ 进行取反操作得到细粒度行为无关特征 $\boldsymbol{O}_{i,\mathrm{f}}$ 和粗粒度行为无关特征 $\boldsymbol{O}_{i,\mathrm{c}}$，并通过以下公式分别融合粗细粒度分支中对应的行为相关特征和行为无关特征，得到结合两种粒度的行为相关特征 $\boldsymbol{X}_{i,\mathrm{a}}$ 和行为无关特征 $\boldsymbol{X}_{i,\mathrm{o}}$，如式(8)所示：

$$\begin{aligned}\boldsymbol{X}_{i,\mathrm{a}}&=\boldsymbol{A}_{i,\mathrm{f}}\oplus\boldsymbol{A}_{i,\mathrm{c}}\\ \boldsymbol{X}_{i,\mathrm{o}}&=\boldsymbol{O}_{i,\mathrm{f}}\oplus\boldsymbol{O}_{i,\mathrm{c}}\end{aligned}\tag{8}$$

式中：⊕表示对应位置的逐元素加法。

2.4 视角内时序语义一致性去偏置

在这一阶段，对 UFO 和 ET 模块进行了复用，但它们的目的有所不同。在去视角间偏置时，UFO 输入包括来自多个视角的信息，通过学习不同视角下无偏特征并将其整合，解决了视角间的偏置。在这里，输入是来自解耦后得到的视角内行为相关特征，通过从时序上学习这些特征的无偏表示，实现了同一视角内行为特征的准确表征，解决了视角内的偏置问题。ET 整合来自不同视角的无偏特征以增强同一视角内行为的一致性并进行最终预测。

为了增强解耦模块提取行为相关信息的能力，采用基于互信息的方法施加约束，连接同一视角内的行为相关特征和行为无关特征作为正样本 $\boldsymbol{X}_{i,\mathrm{pos}}$，连接行为相关特征和不同视角的行为无关特征作为负样本 $\boldsymbol{X}_{i,\mathrm{neg}}$。互信息约束 $\mathcal{L}_{\mathrm{MI}}$ 如式(9)所示：

$$\mathcal{L}_{\mathrm{MI}}=-\frac{1}{N}\sum_{i=1}^{N}(\ln\boldsymbol{X}_{i,\mathrm{pos}}+\ln(1-\boldsymbol{X}_{i,\mathrm{neg}}))\tag{9}$$

2.5 训练和推理

约束 $\mathcal{L}$ 由 3 个部分组成：$\mathcal{L}_{\mathrm{db}}$，$\mathcal{L}_{\mathrm{MI}}$ 和 $\mathcal{L}_{\mathrm{ce}}$。首先使用 $\mathcal{L}_{\mathrm{db}}$ 有助于更好地消除视角间偏置和视角内行为相关信息带来的偏置，然后使用 $\mathcal{L}_{\mathrm{MI}}$ 协助解耦模块更准确地提取行为相关特征，最后使用交叉熵损失 $\mathcal{L}_{\mathrm{ce}}$ 来全面衡量模型输出概率分布与实际标签之间的差异。关系如式(10)所示：

$$\mathcal{L}=\lambda_1\mathcal{L}_{\mathrm{db}}+\lambda_2\mathcal{L}_{\mathrm{MI}}+\lambda_3\mathcal{L}_{\mathrm{ce}}\tag{10}$$

式中：λ 表示对应约束所占的权重。

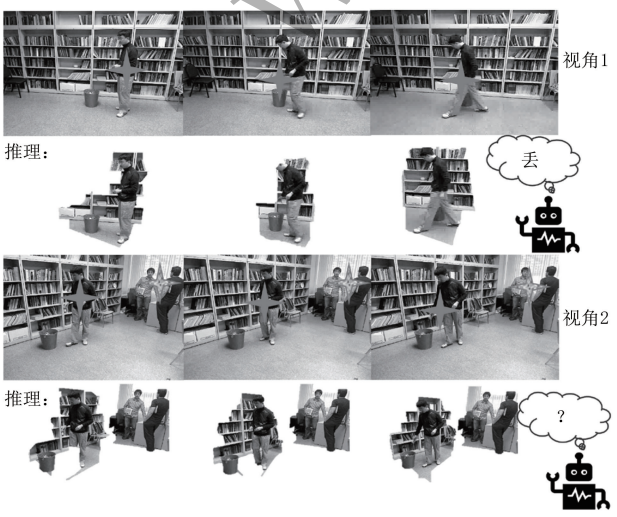
3 实验结果与分析

3.1 数据集和评价指标

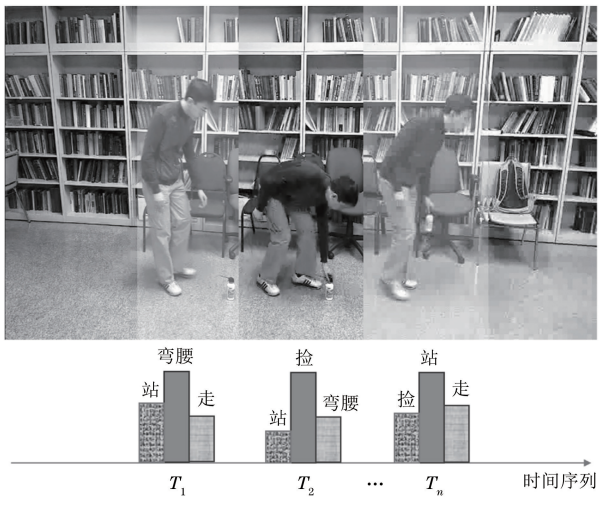
N-UCLA[27] 数据集包括从 3 个不同视角的 Kinect 摄像头同时捕获的 RGB、深度和人体骨架数据。该数据集包含约 1 500 个视频，展示了由 10 名志愿者执行的 10 个不同行为类别。

NTU-RGB+D[28] 数据集包含了从各种摄像头捕获的视频、深度图、骨架关节点位置、RGB 图像和红外序列。该数据集包含约 56 880 个样本，涵盖了由 40 名志愿者执行的 60 个不同行为类别，这些行为大致分为 3 类，即 40 个日常活动(如“喝水”、“吃饭”等)、9 个与健康相关的行为(如“打喷嚏”、“摔倒”等)和 11 个人与人之间的互动(如“打拳”、“拥抱”等)。在实验中，使用 RGB 视频作为训练数据。

如图 2(a)和图 2(b)所示，视角间的场景空间变



(a)视角间偏置



(b)视角内偏置

图 2 视角间及视角内的偏置示例

Fig. 2 Examples of inter-view and intra-view biasing

化和视角内的时序复杂性导致的行为变化使模型难以准确关注和辨别不同角度及时刻下的行为。因此,本文利用其中 2 个视角数据进行训练,第 3 个视角数据进行测试以验证本文方法的有效性。与文献[29-30]的工作一致,使用跨视角的 Top-1 准确率作为评估指标。

3.2 实现细节

实验使用 PyTorch 实现,在一块 NVIDIA Tesla V100 GPU 上进行训练,基于 SlowOnly[31],利用 ResNet3D[32]作为骨干网络构建了一个双流网络作为基线模型。在具体实现中,使用基于 SlowOnly 的预训练模型进行训练,从每个视频中均匀采样 8 帧,在整个训练过程中,训练批次设置为 8,在 256～320 像素的范围内随机调整输入帧的短边,然后使用随机裁剪策略,裁剪尺寸设计为 224×224 像素。对于测试,使用训练批次大小为 32、尺寸为 256×256 像素的随机裁剪,选择 Adam[33]作为优化器,初始学习率为 5×10^{-4},权重衰减为 1×10^{-6},一共训练 60 个轮次。

3.3 对比实验

表 1 展示了所提方法在 N-UCLA 和 NTU-RGB+D 数据集上的实验结果,并将其分别与以下方法进行比较:

1)使用骨架模态作为输入的方法,JEDM+JEAs[34]、AGC-LSTM[35]、Shift-GCN[36]、CrosSCLR[37]、MSNN[13]、AMV-GCNs[38]、3s-AimCLR[39]、EfficientGCN[40]和 HiCLR[41]。

2)使用 RGB+深度模态作为输入的方法,Hybrid[7]、CVAM[42]、CAT[43]和 CMAT[44]。

3)使用 RGB 模态作为输入的方法,VCD[8]、DRDN[9]、CVAM[42]、Conflux LSTM[45]、MotionFormer[46]、3D Geometric[47]和 ViewCLR[48]。

需要注意的是,CVAM 方法同时使用了 RGB+深度、RGB 两种模态的实验设置,在表 1 中“—”表示原文献中没有该指标值。

从整体上看,所提方法在 N-UCLA 和 NTU-RGB+D 数据集上的准确率分别达到了 97.4%和 96.4%,与表中其他方法相比均有着明显的优势,显示了所提方法的有效性和卓越性。在使用骨架模态作为输入时,Shift-GCN[36]在两个数据集上分别达到了 94.6%和 96.5%的最高准确率,所提方法与之相比在 NTU-RGB+D 数据集上的准确率降低了 0.1 百分点,但在 N-UCLA 数据集上提升了 2.8 百分点,主要原因是与 N-UCLA 数据集相比,NTU-RGB+D 数据集的环境更为纯净导致偏置的因素较少,说明所提方法能有效解决偏置问题。在使用 RGB+深度模态作为输入时,CMAT[44]在两个数据集上分别达到了 94.2%和 93.9%的最高准确率,所提方法与之相比提高 3.2 和 2.5 百分点。在同样仅使用 RGB 模态的方法中,所提方法相较对比方法中的最佳方法 DRDN[9]和 ViewCLR[48],在 N-UCLA 和 NTU-RGB+D 数据集上分别提高了 3.5 和 2.3 百分点。这些结果表明所提方法能有效缓解由不同视角下空间语义不一致及同一视角下行为执行过程中时序语义不一致带来的行为表征差异所导致的偏置问题。

表 1 不同方法的性能比较

Table 1 Performance comparison of different methods %

模态	方法	N-UCLA	NTU-RGB+D
骨架	JEDM+JEAs	94.4	91.8
	AGC-LSTM	93.3	95.0
	Shift-GCN	94.6	96.5
	CrosSCLR	—	83.4
	MSNN	89.4	88.7
	AMV-GCNs	93.9	92.2
	3s-AimCLR	—	92.8
	EfficientGCN	—	96.1
	HiCLR	—	95.7
RGB+深度	Hybrid	89.5	84.1
	CVAM	—	77.5
	CAT	87.8	—
	CMAT	94.2	93.9
RGB	VCD	93.8	92.3
	DRDN	93.9	92.9
	CVAM	83.1	86.3
	Conflux LSTM	88.9	—
	MotionFormer	—	91.6
	3D Geometric	—	93.7
	ViewCLR	89.1	94.1
	所提方法	97.4	96.4

3.4 消融实验

在 N-UCLA 数据集的跨视角设置下进行了多种不同的消融实验,使用 Top-1 准确率作为评估指标,以展示所提方法的有效性。

3.4.1 阶段消融

为了验证本文提出的渐进式去偏置方法的有效性,本实验通过逐步增加不同阶段来进行消融,结果如表 2 所示,其中,“√”表示具备当前方法和阶段,“×”表示不具备当前方法和阶段。在阶段 1 中,加入 UFO 和 ET 模块后,Top-1 准确率由 88.7%提升至 93.9%,增幅达 5.2 百分点,这表明行为特征在不同视角数据间的分布存在不均衡的现象,而阶段 1 能很好地平衡不同视角间的特征,避免了模型

对某一视角数据的偏重，有效减少了视角间的偏置问题。在阶段 2 中，加入解耦模块后，Top-1 准确率提升了 2.8 百分点，达到 96.7%，这说明视角内部存在由行为无关信息产生的偏置，且此偏置极易被模型利用，从而严重影响模型的识别性能，验证了将行为相关信息和行为无关信息分离能有效减少行为无关信息带来的偏置。在阶段 3 中，再次加入 UFO 和 ET 模块后，Top-1 准确率进一步提高，证实了该阶段缓解了行为本身固有的偏置问题。总体而言，这些结果表明偏置不仅存在于视角之间，而且也存在于视角内部，且所提方法能够有效缓解这些偏置问题。

表 2　不同阶段的性能比较

Table 2　Comparison of different stages　%

基线方法	阶段 1	阶段 2	阶段 3	Top-1 准确率
√	×	×	×	88.7
√	√	×	×	93.9
√	√	√	×	96.7
√	√	√	√	97.4

3.4.2　解耦模块中不同头部分割比的比较

在图 3 中展示了不同头部分割比 r 对解耦的影响。通过改变 r 的取值研究粗粒度和细粒度分支中不同头部数量对解耦的影响。需要注意的是，为了避免模块间的相互干扰，此实验中没有加入 UFO 模块。当 $r=0$ 或 $r=1.0$ 时，模型将所有的头部数量都分配给某一分支，但此时 Top-1 准确率并没有达到峰值，说明单独使用某一分支并不能保证行为相关信息的完整性。当 $r=0.3$ 或 $r=0.7$ 时，Top-1 准确率达到最高值 95.2%，这表明多粒度结合的策略能在解耦中减少行为信息的丢失，在保证提取信息完整性的同时有效避免了行为无关信息带来的偏置的影响。

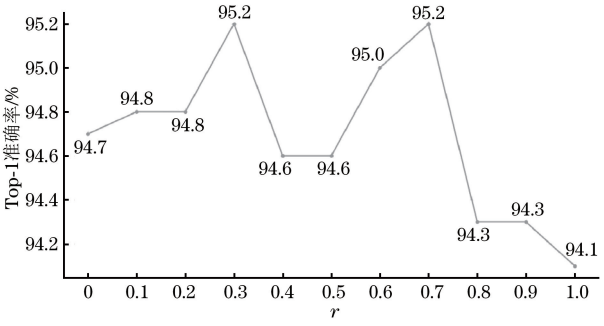


图 3　不同头部分割比的比较

Fig.3　Comparison of different head segmentation ratios

3.4.3　UFO 中不同独立性参数的比较

本实验在图 3 实验的基础上加入 UFO 模块，以研究 HSIC 中不同独立性参数 θ 对模型的影响，并寻找最佳的头部分割比 r。由图 4 中数据可知，与 $r=0.3$ 相比，当 $r=0.7$ 时，模型的整体表现较为优秀，这可能是因为当 $r=0.3$ 时，模型在解耦中过于关注行为的局部导致行为本身产生较大的偏置。当 $\theta=0$ 或 $\theta=1.0$ 时，Top-1 准确率并没有达到最高点。当 $\theta=0$ 时，HSIC 的值最小，此时有偏特征与无偏特征之间的独立性最强，学习到的有偏特征并没有很好地与无偏特征产生联系导致学习到的无偏特征仍然包含了部分偏置信息。当 $\theta=1.0$ 时，HSIC 的值最大，有偏特征与无偏特征之间的关联性最强，在去偏置的过程中有偏特征可能包含了与无偏特征较多相似的地方导致学习到的无偏特征丢失了部分重要信息。当 $\theta=0.3$ 时，有偏特征与无偏特征的独立性达到了平衡点，此时 Top-1 准确率最高为 95.7%。

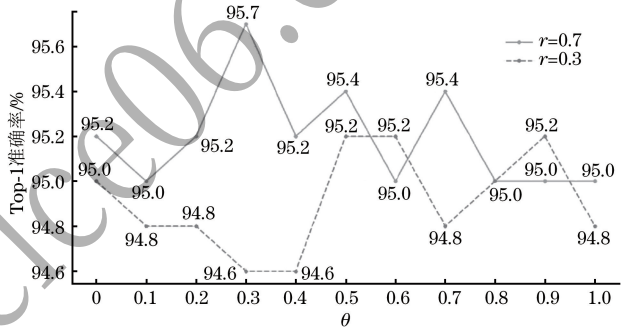


图 4　不同独立性参数的比较

Fig.4　Comparison of different independence parameters

3.4.4　不同 loss 超参数的比较

为了寻找最佳的 λ_1、λ_2、λ_3 值，本实验采用了吉布斯采样[49]的方法。首先在实验 1 中将 λ_2 和 λ_3 的值默认设为 1，然后对 λ_1 采用不同的数值进行多次实验，以此找到在这些条件下的最佳 λ_1 值，确定 λ_1 的最佳值后，在实验 2 中保持 λ_1 的这一最佳值和 λ_3 的默认值不变，对 λ_2 进行多轮实验，从而确定 λ_2 的最佳取值，接着在实验 3 中固定已找到的最佳 λ_1 和 λ_2 值，进行同样的多次实验确定最佳 λ_3 值，最后在实验 4 中使用最佳的 λ_2 和 λ_3 对 λ_1 进行新一轮的迭代，验证模型的准确率是否收敛，避免三者取值相互影响产生局部最优解。由表 3 可知，在前 3 组实验中，依次确定 λ_1、λ_2、λ_3 的最佳取值分别为 1、1 和 10，此时模型的准确率达到最高，为 97.4%。在实验 4 中，模型的准确率并没有进一步的提升，说明 λ_2 和 λ_3 并没有对 λ_1 的取值产生影响，由此可见，λ_1、λ_2、λ_3 的最佳取值仍然为 1、1 和 10。其中，交叉熵损失 $\mathcal{L}_{ce}$ 对应的权重比例最大，这说明模型更关注整体的优化和去偏置效果，以提高对不同视角、行为和场景的泛化能力。

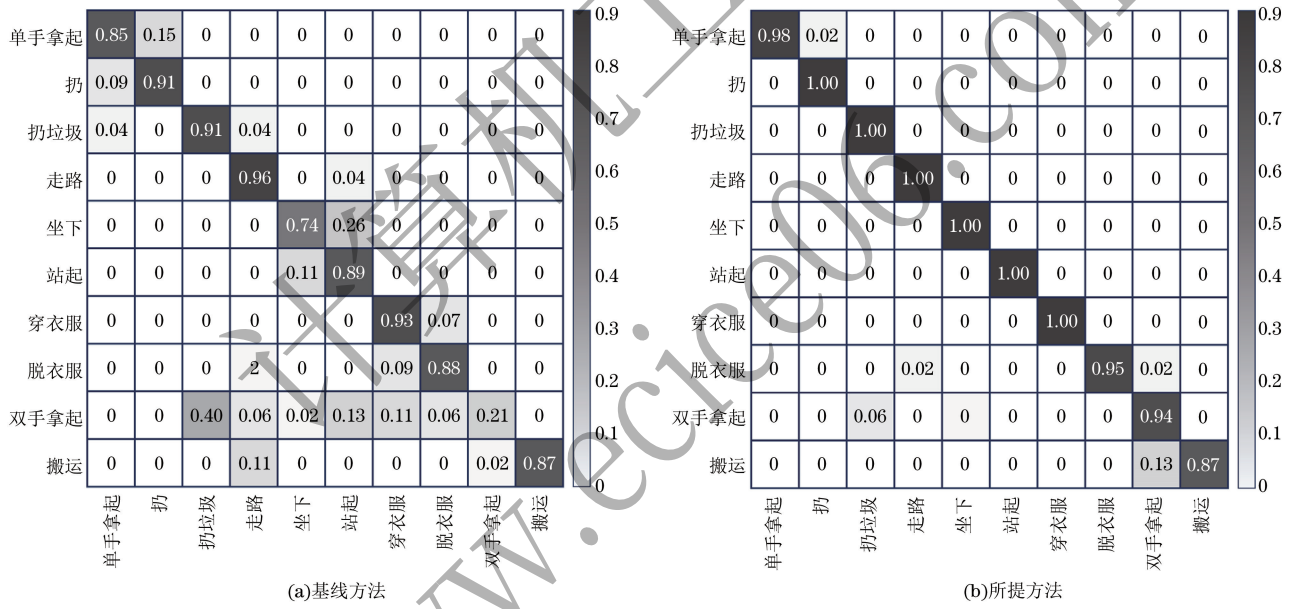
表 3 不同 loss 超参数的比较

Table 3 Comparison of different loss hyperparameters

实验	准确率/%					最佳值
	$\lambda=0.01$	$\lambda=0.1$	$\lambda=1$	$\lambda=10$	$\lambda=100$	
实验 1	95.0	95.2	95.7	95.0	95.2	$\lambda_1=1$
实验 2	94.8	95.4	95.7	95.2	93.3	$\lambda_2=1$
实验 3	27.6	81.5	95.7	97.4	65.0	$\lambda_3=10$
实验 4	95.2	96.1	97.4	96.1	95.2	$\lambda_1=1$

3.5 可视化结果

为了更直观地显示所提方法的有效性，本节在图 5(a)和图 5(b)中分别展示了基线方法与所提方法在 N-UCLA 数据集上混淆矩阵的可视化。由图 5(a)可见，基线方法在相似行为分类上表现不准确，例如，由于“坐下”和“站起”在特征表征上的相似性，模型在单一视角下难以明确区分这两个行为。此外，基线方法在“双手拿起”识别方面表现出极低的准确率，主要原因在于该行为在起始阶段与其他行为存在显著的混淆性，模型若未能充分考虑行为的时序关系，容易误分类。由图 5(b)可知，所提方法在分类准确率上得到了明显提升，不同行为之间的混淆明显减少。该实验结果表明所提方法能够有效处理不同视角之间的空间语义和同一视角内时序语义所带来的影响，显著减少了多视角情况下的偏置。



(a)基线方法 (b)所提方法

图 5 混淆矩阵可视化

Fig. 5 Confusion matrix visualization

4 结束语

本文深入探讨了多视角行为识别任务中，由不同视角间空间语义不一致及同一视角内行为时序语义不一致带来的表征差异所导致的偏置问题，并据此提出了一种时空语义驱动的渐进式去偏置方法。该方法通过证据理论动态融合不同视角的无偏特征缓解了视角间的偏置问题，在视角内应用解耦策略分离行为有关与行为无关信息，抑制了行为无关信息导致的偏置问题带来的不利影响，同时在时序上优化行为相关信息的特征权重，增强了同一视角内行为的时空一致性，有效减少了视角内的偏置问题。实验结果表明，该方法在多视角行为识别常用数据集 N-UCLA 和 NTU-RGB+D 上分别取得了 97.4%和 96.4%的跨视角准确率。在未来的工作中，将深入研究更高效的证据理论方法，以全面且准确地整合不同视角的特征，从而更有效地缓解视角间的偏置问题，同时提高模型对未知视角的泛化能力，提升其在更多实际场景中的应用潜力。

参考文献

[1] 刘思进，朱小飞，彭展望. 联合多任务学习的对话情感分类和行为识别[J]. 计算机学报，2023，46(9)：1947-1960.
LIU S J，ZHU X F，PENG Z W. Dialogue sentiment classification and act recognition based on multi-task learning [J]. Chinese Journal of Computers，2023，46(9)：1947-1960. (in Chinese)

[2] 蒲瞻星，葛永新. 基于多特征融合的小样本视频行为识别算法[J]. 计算机学报，2023，46(3)：594-608.
PU Z X，GE Y X. Few-shot action recognition in video based on multi-feature fusion[J]. Chinese Journal of Computers，2023，46(3)：594-608. (in Chinese)

[3] YOU H，ZHONG X，LIU W X，et al. Converting artificial neural networks to ultralow-latency spiking neural networks for action recognition[J]. IEEE Transactions on Cognitive and Developmental Systems，2024，16(4)：1533-1545.

[4] 张洋，姚登峰，江铭虎，等. 基于 EfficientDet 网络的细粒度吸烟行为识别[J]. 计算机工程，2022，48(3)：302-309，314.
ZHANG Y，YAO D F，JIANG M H，et al. Fine-grained

smoking behavior recognition based on EfficientDet network [J]. Computer Engineering, 2022, 48(3): 302-309, 314. (in Chinese)
[5] 闫兴亚, 匡娅茜, 白光睿, 等. 基于深度学习的学生课堂行为识别方法[J]. 计算机工程, 2023, 49(7): 251-258.
YAN X Y, KUANG Y Q, BAI G R, et al. Student classroom behavior recognition method based on deep learning [J]. Computer Engineering, 2023, 49(7): 251-258. (in Chinese)
[6] ZHAO L, WANG Y X, ZHAO J P, et al. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 12793-12802.
[7] DHIMAN C, VISHWAKARMA D K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics[J]. IEEE Transactions on Image Processing, 2020, 29: 3835-3844.
[8] ZHONG X, ZHOU Z, LIU W X, et al. VCD: view-constraint disentanglement for action recognition [C]// Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2022: 2170-2174.
[9] LIU W X, ZHONG X, ZHOU Z, et al. Dual-recommendation disentanglement network for view fuzz in action recognition [J]. IEEE Transactions on Image Processing, 2023, 32: 2719-2733.
[10] 詹健浩, 甘利鹏, 毕永辉, 等. 基于知识蒸馏的多模态融合行为识别方法[J]. 计算机工程, 2023, 49(10): 280-288, 297.
ZHAN J H, GAN L P, BI Y H, et al. Action recognition method with multi-modality fusion based on knowledge distillation[J]. Computer Engineering, 2023, 49(10): 280-288, 297. (in Chinese)
[11] 施海勇, 侯振杰, 巢新, 等. 多模态时空特征表示及其在行为识别中的应用[J]. 中国图象图形学报, 2023, 28(4): 1041-1055.
SHI H Y, HOU Z J, CHAO X, et al. Multimodal spatial-temporal feature representation and its application in action recognition [J]. Journal of Image and Graphics, 2023, 28(4): 1041-1055. (in Chinese)
[12] LI J N, WONG Y, ZHAO Q, et al. Unsupervised learning of view-invariant action representations[EB/OL]. [2024-01-02]. http://arxiv.org/abs/1809.01844.
[13] SHAO Z P, LI Y F, ZHANG H. Learning representations from skeletal self-similarities for cross-view action recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 160-174.
[14] BAHRAMPOUR S, NASRABADI N M, RAY A, et al. Multimodal task-driven dictionary learning for image classification[J]. IEEE Transactions on Image Processing, 2016, 25(1): 24-38.
[15] LIU Z G, WANG L, YIN Z Y, et al. Task-driven joint dictionary learning model for multi-view human action recognition [J]. Digital Signal Processing, 2022, 126: 103487.
[16] LI Y W, LI Y, VASCONCELOS N. RESOUND: towards action recognition without representation bias [C]// Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 520-535.
[17] KIM B, KIM H, KIM K, et al. Learning not to learn: training deep neural networks with biased data [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2019: 9012-9020.
[18] CHOI J, GAO C, MESSOU J C E, et al. Why can't I dance in the mall? Learning to mitigate scene bias in action recognition[EB/OL]. [2024-01-02]. http://arxiv.org/abs/1912.05534.
[19] BAHNG H, CHUN S, YUN S, et al. Learning de-biased representations with biased representations[EB/OL]. [2024-01-02]. http://arxiv.org/abs/1910.02806.
[20] BAO W T, YU Q, KONG Y. Evidential deep learning for open set action recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2021: 13349-13358.
[21] DUAN H D, ZHAO Y, CHEN K, et al. Mitigating representation bias in action recognition: algorithms and benchmarks[C]// Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2023: 557-575.
[22] LI Q K, HUANG X L, LUO Y W, et al. Mitigating context bias in action recognition via skeleton-dominated two-stream network[C]//Proceedings of the 2023 Workshop on Advanced Multimedia Computing for Smart Manufacturing and Engineering. New York, USA: ACM Press, 2023: 65-70.
[23] GRETTON A, BOUSQUET O, SMOLA A, et al. Measuring statistical dependence with Hilbert-Schmidt norms [M]. Berlin, Germany: Springer, 2005.
[24] SENSOY M, KAPLAN L, KANDEMIR M. Evidential deep learning to quantify classification uncertainty [EB/OL]. [2024-01-02]. http://arxiv.org/abs/1806.01768.
[25] TIAN C W, ZHENG M H, ZUO W M, et al. A cross Transformer for image denoising[J]. Information Fusion, 2024, 102: 102043.
[26] TIAN C W, ZHENG M H, LI B, et al. Perceptive self-supervised learning network for noisy image watermark removal[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 7069-7079.
[27] WANG J, NIE X H, XIA Y, et al. Cross-view action modeling, learning, and recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 2649-2656.
[28] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2016: 1010-1019.
[29] CHEN Y X, ZHANG Z Q, YUAN C F, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2021: 13359-13368.
[30] CHI H G, HA M H, CHI S, et al. InfoGCN: representation learning for human skeleton-based action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2022: 20186-20196.
[31] FEICHTENHOFER C, FAN H Q, MALIK J, et al. SlowFast networks for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Washington D. C., USA: IEEE Press, 2019: 6202-6211.
[32] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 6546-6555.
[33] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2024-01-02]. http://arxiv.org/abs/1412.6980.
[34] NIE Q, WANG J L, WANG X, et al. View-invariant human

action recognition based on a 3D bio-constrained skeleton model[J]. IEEE Transactions on Image Processing, 2019, 28(8): 3959-3972.
[35] SI C Y, CHEN W T, WANG W, et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2019: 1227-1236.
[36] CHENG K, ZHANG Y F, HE X Y, et al. Skeleton-based action recognition with shift graph convolutional network [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2020: 183-192.
[37] LI L G, WANG M S, NI B B, et al. 3D human action representation learning via cross-view consistency pursuit[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 4741-4750.
[38] LIU X, LI Y, XIA R. Adaptive multi-view graph convolutional networks for skeleton-based action recognition [J]. Neurocomputing, 2021, 444: 288-300.
[39] GUO T Y, LIU H, CHEN Z, et al. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2022: 762-770.
[40] SONG Y F, ZHANG Z, SHAN C F, et al. Constructing stronger and faster baselines for skeleton-based action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 1474-1488.
[41] ZHANG J H, LIN L L, LIU J Y. Hierarchical consistent contrastive learning for skeleton-based action recognition with growing augmentations [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2023: 3427-3435.
[42] VYAS S, RAWAT Y S, SHAH M. Multi-view action recognition using cross-view video prediction[C]//Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 427-444.
[43] XU C, WU X, LI Y C, et al. Cross-modality online distillation for multi-view action recognition [J]. Neurocomputing, 2021, 456: 384-393.
[44] PARK Y, WOO S, LEE S M, et al. Cross-modal alignment and translation for missing modality action recognition[J]. Computer Vision and Image Understanding, 2023, 236: 103805.
[45] ULLAH A, MUHAMMAD K, HUSSAIN T, et al. Conflux LSTMs network: a novel approach for multi-view action recognition[J]. Neurocomputing, 2021, 435: 321-329.
[46] PATRICK M, CAMPBELL D, ASANO Y, et al. Keeping your eye on the ball: trajectory attention in video Transformers [J]. Advances in Neural Information Processing Systems, 2021, 34: 12493-12506.
[47] PIERGIOVANNI A J, RYOO M S. Recognizing actions in videos from unseen viewpoints[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Washington D. C., USA: IEEE Press, 2021: 4124-4132.
[48] DAS S, RYOO M S. ViewCLR: learning self-supervised video representation for unseen viewpoints[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Washington D. C., USA: IEEE Press, 2023: 5573-5583.
[49] CASELLA G, GEORGE E I. Explaining the Gibbs sampler [J]. The American Statistician, 1992, 46(3): 167-174.

编辑 陆燕菲