

中图法分类号: TP391 文献标识码: A 文章编号: 1006-8961(2025)01-0130-18

论文引用格式: Liu W X, Zhong X, Xu X Y, Zhou Z, Jiang K, Wang Z and Bai X. 2025. Discriminative information incremental learning for Air-ground multi-view action recognition. Journal of Image and Graphics, 30(01):0130-0147(刘文璇, 钟欣, 徐晓玉, 周卓, 江奎, 王正, 白翔. 2025. 空一地多视角行为识别的判别信息增量学习方法. 中国图象图形学报, 30(01):0130-0147)[DOI: 10. 11834/jig. 230815]

空一地多视角行为识别的判别信息增量学习方法

刘文璇¹, 钟欣^{1,2*}, 徐晓玉¹, 周卓³, 江奎⁴, 王正³, 白翔⁵

1. 武汉理工大学计算机与人工智能学院, 武汉 430070; 2. 水路交通控制全国重点实验室, 武汉 430063;
3. 武汉大学计算机学院, 武汉 430072; 4. 哈尔滨工业大学计算机科学与技术学院, 哈尔滨 150001;
5. 华中科技大学软件学院, 武汉 430074

摘要: 随着社会民生需求的日益增长, 地面设备联合无人机等空中设备进行空一地多视角场景下的行为识别应运而生。现有方法只关注水平空间视角变化时的视角关系, 忽视了垂直空间变化带来的巨大判别行为差异。由于高度不同, 同一物体的观测特征差异显著, 对传统的多视角动作识别方法在应对垂直空间视角变化时构成了重大挑战。**方法** 本文将显著的动作外观差异定义为辨别动作信息的差异, 提出了一种基于判别行为信息增量学习(discriminative action information incremental learning, DAIL)的空一地场景多视角行为识别方法, 根据视角高度和信息量判别地面视角和空中视角。引入类脑“由易到难, 循序渐进”的思想, 分别蒸馏不同视角的判别行为信息, 将地面视角(简单)样本的判别行为信息增量到空中视角(困难)样本中, 辅助网络学习空中视角样本。**结果** 在Drone-Action和UAV(unmanned aerial vehicle)两个数据集上进行了实验, 对比于当前先进方法SBP(stochastic backpropagation), 在两个数据集上准确率分别提高了18.0%和16.2%; 对比于强基线方法, 本文方法在UAV数据集上参数量减少了2.4 M, 计算量减少了6.9 G。**结论** 所提出的方法表明通过使用简单样本增强复杂样本, 显著提高了网络的特征学习能力。相反, 尝试反向操作会导致准确性下降。本文从新角度讨论多视角行为识别任务, 兼具效果和性能, 在常见的高空视角数据集中优于代表性的方法且可拓展到其他多视角任务。

关键词: 多视角行为识别; 增量学习; 样本分类; 判别行为信息; 蒸馏学习

Discriminative information incremental learning for Air-ground multi-view action recognition

Liu Wenxuan¹, Zhong Xian^{1,2*}, Xu Xiaoyu¹, Zhou Zhuo³, Jiang Kui⁴, Wang Zheng³, Bai Xiang⁵

1. School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China;
2. State Key Laboratory of Maritime Technology and Safety, Wuhan 430063, China; 3. School of Computer Science, Wuhan University, Wuhan 430072, China; 4. School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China;
5. School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

Abstract: Objective With the increasing demand for urban security of people, ground devices are combined with air devices, such as drones, for identifying action in air-ground scenarios. Meanwhile, the extensive ground-based camera networks and a wealth of ground surveillance data can offer reliable support to these aerial surveillance devices. How to effectively utilize the mobility of these aerial devices is a topic that warrants further research. Existing multi-view action recogni-

收稿日期: 2023-11-27; 修回日期: 2024-08-19; 预印本日期: 2024-08-26

* 通信作者: 钟欣 zhongx@whut.edu.cn

基金项目: 国家自然科学基金项目(62271361); 湖北省科技计划项目(2024BAB039)

Supported by: National Natural Science Foundation of China(62271361); Hubei Provincial Key R&D Program, China (2024BAB039)

tion methods focus only on the difference in discriminative action information when the horizontal spatial view changes, but do not consider the difference in discriminative action information when the vertical spatial view changes. The high mobility of aerial perspectives can lead to changes in the vertical spatial perspective. According to the principles of perspective, observing the same object from different heights results in a significant change in appearance. This, in turn, causes substantial differences in the appearance of the same person's actions when observed from high-altitude and ground-level perspectives. These significant variations in action appearance are referred to as differences in discriminative action information, and they pose a challenge for traditional multi-view action recognition methods in effectively addressing the issue of vertical spatial perspective changes. **Method** When the viewing perspective aligns with the objects being observed in the same horizontal spatial plane, the most comprehensive and rich discriminative action information can be observed. Networks can easily learn and comprehend this information. However, when the viewing perspective is in a different horizontal spatial plane from the observed objects, inclined perspective occurs, resulting in a significant change in action appearance. This transition from a ground-level perspective to an aerial perspective leads to insufficiently observed information and a reduction in the discriminative action information. When networks attempt to learn and understand this information, misclassifications are more likely to occur. Therefore, on the basis of the amount of discriminative action information, ground-level perspective information can be considered as easily learned and understood simple information, while aerial perspective information can be seen as complex information that is challenging to learn and understand. In fact, the human brain follows a progressive learning process when dealing with various types of information, prioritizing the processing of simple information and using the learned simple information to assist in learning complex information. In the vertical spatial multi-view action recognition task, differences in perspectives and environmental influences lead to varying amounts of discriminative action information observed at different heights. In this chapter, we adopt a brain-like approach. We rank samples from the aerial perspective on the basis of the amount of discriminative action information they contain. Complex samples contain less discriminative action information, and networks find them challenging to learn and understand. Simple samples contain more discriminative action information and are easier for networks to learn and comprehend. We then distill discriminative action information separately from simple and complex samples. Within the same action category, despite differences in the amount of discriminative action information between simple and complex samples, the represented action categories should have commonalities. Therefore, by using the discriminative action information incremental learning method, we incrementally inject the rich discriminative action information learned from simple samples into the feature information of complex samples. This approach addresses the issue of complex samples carrying insufficient discriminative action information, allowing complex samples to learn more discriminative action information with the assistance of simple samples. Thus, networks can learn and understand complex samples easily. This paper proposes a discriminative action information incremental learning (DAIL) for multi-view action recognition in complex air-ground scenes and to distinguish the ground view from the air view on the basis of the view height and the amount of information. This paper utilizes a neuromorphic learning knowledge referred to as "ordered incremental progression" to distill discriminative action information for different views separately. Discriminative action information is incremented from the ground-view (simple) samples into the air-view (complex) samples to assist the network in learning and understanding the air-view samples. **Result** The method is experimentally validated on two datasets, namely, Drone-Action and unmanned aerial vehicle (UAV). The accuracy of the two datasets is improved by 18.0% and 16.2%, respectively, compared with that of the current state-of-the-art method SBP. Compared with the strong baseline method, our method reduces the parameters by 2.4 M and the FLOPS by 6.9 G on the UAV dataset. To validate the effectiveness of our proposed method in scenarios involving both ground-level and aerial perspectives, we introduced two datasets: N-UCLA (comprising samples exclusively from ground-based cameras with rich discriminative behavior information) and Drone-Action (comprising a mix of ground-level and aerial samples, where aerial samples contain relatively limited discriminative behavior information). A joint analysis of discriminative behavior information ranking was conducted on these datasets. Our findings indicate that enhancing complex samples using simpler ones significantly improves the network's feature learning capacity. Conversely, attempting the reverse can lead to reduce the accuracy. This observation aligns with the way the human brain processes information, embodying the concept of progressive learning. **Conclusion** In this study, we proposed DAIL for multi-view action recognition in complex air-

ground scenes and to distinguish the ground view from the air view on the basis of the view height and the amount of information.

Experiment results show that our model outperforms several state-of-the-art multi-view approaches and improves the performance.

Key words: multi-view action recognition; incremental learning; sample classification; discriminative action information; distillation learning

0 引言

高空监控设备(如无人机、高空摄像头等)由于具有机动性和广域性优势受到城市安防应用的青睐,高空视角数据比重逐渐上升。然而,高空视角数据在现实场景中由于存在空间分布不均衡等问题导致行为携带的有效信息不充分,难以准确识别。与高空视角不同,地面相机拍摄的行为样本则具有丰富的判别行为信息。因此,如何有效利用地面视角样本来增强对高空视角样本的行为识别能力是一个值得研究的问题。

现有的多视角行为识别工作往往只关注水平空间视角变化带来的挑战,旨在通过融合多视角特征挖掘出具有视角不变性的判别行为信息(Geng等, 2021; Zhao等, 2021; Wang等, 2022; Zhong等, 2022; Chen等, 2020)。然而,这些方法忽略了视角垂直空间变化带来的问题:同一行为在不同高度下的外观差异和信息量差异。如图1(a)所示,当观察视角与被观察物体处于同一水平空间时,可以观察到完整且简单的判别行为信息,网络获得较多有效行为信息。然而,当观察视角与被观察物体处于不同垂直空间时,会引起倾斜透视,行为外观显著变化,如图1(b)所示。高空视角会产生有效行为信息不充分问题,观察到的判别行为信息复杂且冗余,网络在学习和理解这些信息时,更容易出现误判。行为外观差异反映了判别行为信息量的差异,也使得传统的多视角行为识别方法难以有效地应对空一地场景中垂直空间视角变化的问题。本文提出一种渐进式学习策略,有效利用地面视角样本中丰富的有效行为信息来指导高空视角样本的学习,提升了空一地多视角行为识别的性能。

事实上,人脑在处理各式各样的信息时会遵循一种“由易到难,循序渐进”的学习方式(Ho等, 2022; Hardy, 2010),优先处理简单信息,将简单信息作为理解复杂事物的知识基础。此外,国内学者也针对视觉知识及智能进化进行相关探讨(杨易等,

2022),通过类比联想的研究将视觉知识中的关系建模推广到更为抽象的关系。受到该论文的启发,本文中不同视角下的行为信息类比于知识,以简单样本引导复杂样本进行识别。针对空一地场景下垂直空间视角变化,尤其是高空视角样本判别行为信息复杂的难点,本文采用类脑的处理策略。根据判别行为信息的含量对样本进行复杂度排序,高空视角样本的判别行为信息较为稀疏,网络难以学习和理解,因此可视为困难样本。地面视角样本的判别行为特征较为丰富,网络容易学习和理解,可视为简单样本。然后从简单样本和困难样本中分别蒸馏出判别行为信息。对于同一个行为类别,尽管简单样本和困难样本所包含的判别行为信息量存在差异,但是判别行为信息所表征的行为类别具有共性。困难样本可以利用行为共性从简单样本的丰富特征中学习到的更多相关行为特征。因此,通过判别行为信息增量方法,将从简单样本中学习到的有效行为信息增量到困难样本的特征信息中,充分利用丰富的地面视角样本以提升网络对高空视角样本的行为识别能力,有效地弥补了困难样本携带判别行为信息量不足的问题。

针对现有的多视角行为识别无法有效处理空一

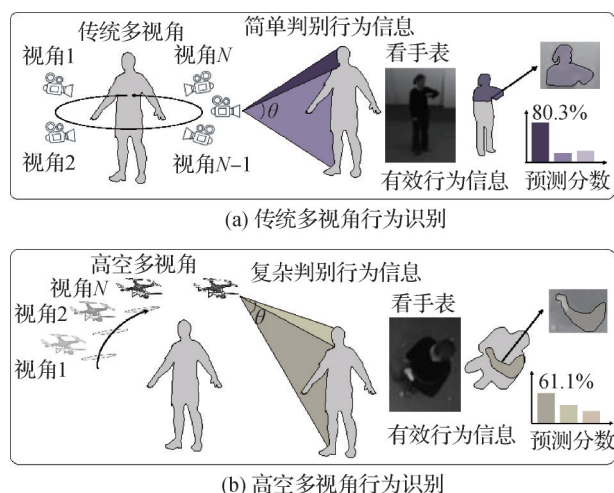


图1 传统与高空多视角行为识别的有效行为信息差别

Fig. 1 Difference in effective action information between traditional and high-altitude multi-view action recognition

((a) traditional multi-view; (b) our air-ground multi-view)

地场景下垂直空间视角差异带来的行为类内差异过大的问题,本文提出了一种判别行为信息增量学习(discriminative action information incremental learning, DAIL)网络。首先处理无人机行为识别数据集中的样本划分问题,通过考虑不同视角的高度差异和计算得到的样本置信度,将地面视角划分为判别行为信息充足的简单样本,高空视角划分为判别行为信息不足的困难样本;其次,采用双流网络并使用蒸馏方法提取出简单样本和困难样本的判别行为信息。对于同一种行为类别,即使从不同视角观测到的行为外观不同,但其行为特征存在共性。因此,本文设计了一种判别行为信息增量学习的方法,利用简单样本学习到的判别行为信息指导网络学习困难样本,以解决判别行为信息利用不充分的问题,以及垂直视角跨度过大造成的判别行为信息灾难性遗忘的问题。

本文的主要贡献如下:1)发现了一个具有挑战的多视角行为识别领域,即空一地多视角行为识别。该领域主要关注传统的多视角行为识别方法未能考虑到空一地场景垂直空间视角变化带来的类内差距过大的问题。本文首次在多视角行为识别任务中使用类脑的知识,并提出了一种“由易到难,循序渐进”的思想,学习空一地多视角场景中的困难样本。2)设计了一种判别行为信息有序蒸馏方法,解决了空一地场景下高空视角样本比例过大,判别行为信息混淆导致识别效果不佳的问题。区别于传统方法平等对待所有样本,本文方法深入挖掘不同视角样本包含的判别信息量的有序关系。3)提出了一种跨视角判别行为信息增量学习框架,通过从简单样本到困难样本的有序关系指导增量,解决类内判别行为信息量差距过大导致的灾难性遗忘问题。该方法为空一地多视角及相关领域的研究提供了一种新颖的解决思路。

本文提出了一个判别行为信息增量学习网络,并在两个公开的无人机行为识别数据集进行实验,辅以地面多视角数据集验证,与现有多种基于时间和空间的显著性模型的行为识别方法进行比较。实验结果表明,本文方法表现优于现有的行为识别方法,对比于SBP(stochastic backpropagation)(Cheng等,2022),在Drone-Action上提高了18.0%,在UAV(unmanned aerial vehicle)上提高了16.2%。

1 相关工作

对行为信息进行有效提取是多视角行为识别的关键,相较于传统的地面多视角行为信息,高空视角的行为信息更复杂,包含的判别行为信息更少,因此更难被辨别。

1.1 多视角行为识别

行为识别(Simonyan和Zisserman,2014;周波和李俊峰,2020;施海勇等,2023)一直以来都是模式识别领域的关注焦点。传统工作(张冰冰等,2021;李俊峰和张飞燕,2015;Liu等,2022;Dave等,2022)大多关注单视角行为识别。I3D(inflated 3D)(Carreira和Zisserman,2017)融合了光流特征对样本进行学习,Slow-Fast(Feichtenhofer等,2019)通过双路径学习样本的特征,Fayyaz等人(2021)学习时间特征的相似性并将相似的特征分组,SBP(Cheng等,2022)使用随机反向传播的方法学习行为特征。Li等人(2023b)提出了图卷积神经网络,实现基于骨架的双人行为识别。Mazzia等人(2022)使用注意力机制精确有效地提供了一个低延迟行为识别方法。尽管这些工作在单视角样本取得比较好的识别效果,但当视角发生变化时性能会显著下降。因此,多视角行为识别逐渐成为当前行为识别任务中的研究热点。目前的一些多视角行为识别的方法关注于学习多视角特征的视角不变特征。其中,Xiao等人(2019)提出了一个经典的方法,通过放置大量的虚拟摄像机来列举尽可能多的视角以合成不同视角的训练数据。而另外一些工作采用深度神经网络来处理所有视角,并通过特征聚类来学习视角无关特征(Ullah等,2021;Vyas等,2020)。由于没有准确的约束条件,这种方法可能会受到未学习过的视角影响。一些工作(Ullah等,2021;Bin等,2020;Xu等,2021;Shao等,2021)使用人类骨架数据作为额外的补充信息,构建可以克服视觉噪声影响的网络。基于骨架的多视角行为识别方法的效果会受到骨架数据预处理结果的限制。Kong等人(2017)旨在通过边缘化自动编码器学习特定视角特征和共享视角特征。而对场景动态进行编码,可以从多个视角中学习视角不变特征(Vyas等,2020)。然而,空一地联合视角样本没有详细的视角划分,因此传统的多视角行为识别工作不能很好地适用于空一地联合行为识别任务。

1.2 样本分类

此外,大型标注数据集的建立也推动了模式识别的发展。普遍认为,数据集的标注越详细,对于学习不同视角下行为的有效性就越大。然而,在无人机行为识别数据集中,对不同视角下行为进行详细区分仍然面临巨大的挑战。一个有效的样本选择器(Luo等,2021),以及对目标样本进行评分并设置阈值选择那些会导致分类变化的样本(Lifshitz和Wolf,2021)可以一定程度优化分类结果。Swayamdipta等人(2020)提出了一种对大型数据集进行可视化和诊断的自动方法来区分样本。然而,这些方法都没有对困难样本进行充分学习,从而未能有效解决在空一地多视角场景下,难以学习的高空视角困难样本比重过大导致分类效果差的问题。

1.3 增量学习

增量学习是当前模式识别社区比较火热的一个话题。增量学习一种学习范式,旨在在稳定性—可塑性困境中寻找效用最大的平衡点,其中可塑性代表整合新知识的能力,稳定性则代表在学习新知识过程中保留旧知识的能力。增量学习方法可以分为3个类别:正则化,回放和参数更新。Liu等人(2018)通过正则化网络确保前一个网络和当前网络的输出相似来解决学习新知识的灾难性遗忘问题。Rebuffi等人(2017)存储前一个任务包含判别知识的训练样本,并通过网络参数更新的方法确保之后的网络可以学习到旧知识。Shin等人(2017)和Wu

等人(2021)训练生成网络来重新训练旧数据,将增量学习转换为传统的监督学习。Yu等人(2020)提出近似训练新任务后的语义漂移,补充了几个原本为分类网络设计的增量学习方法。Wei等人(2021)使用零样本学习来弥补不同任务之间的语义差距,缓解灾难性遗忘的问题。此外,增量学习可与其他模式识别任务相结合,如零样本学习、语义分割等,这可以帮助弥补模式识别领域和真实世界之间的语义差距。与上述方法相比,本文提出一种跨视角的增量学习的方法,通过对不同视角的补充来弱化差异,用简单样本来指导困难样本学习判别行为信息。

2 本文方法

本文以类脑学习为出发点,提出了一种“由易到难,循序渐进”思想的判别行为信息增量学习方法DAIL,具体的框架如图2所示。首先,与传统分类方法剔除困难样本不同,本文根据样本包含的判别行为信息量将无人机数据样本分类为容易学习的地面视角简单样本和难以学习的高空视角困难样本。其次,使用基线模型SlowOnly(Feichtenhofer等,2019),分别对简单样本 E (easy samples)和困难样本 H (hard samples)提取判别行为信息,以学习判别性行为特征。最后,利用简单样本提取的判别行为信息指导困难样本的学习,帮助网络理解困难样本。网络的总损失函数 L_{all} 由3个部分组成,即增量学习损

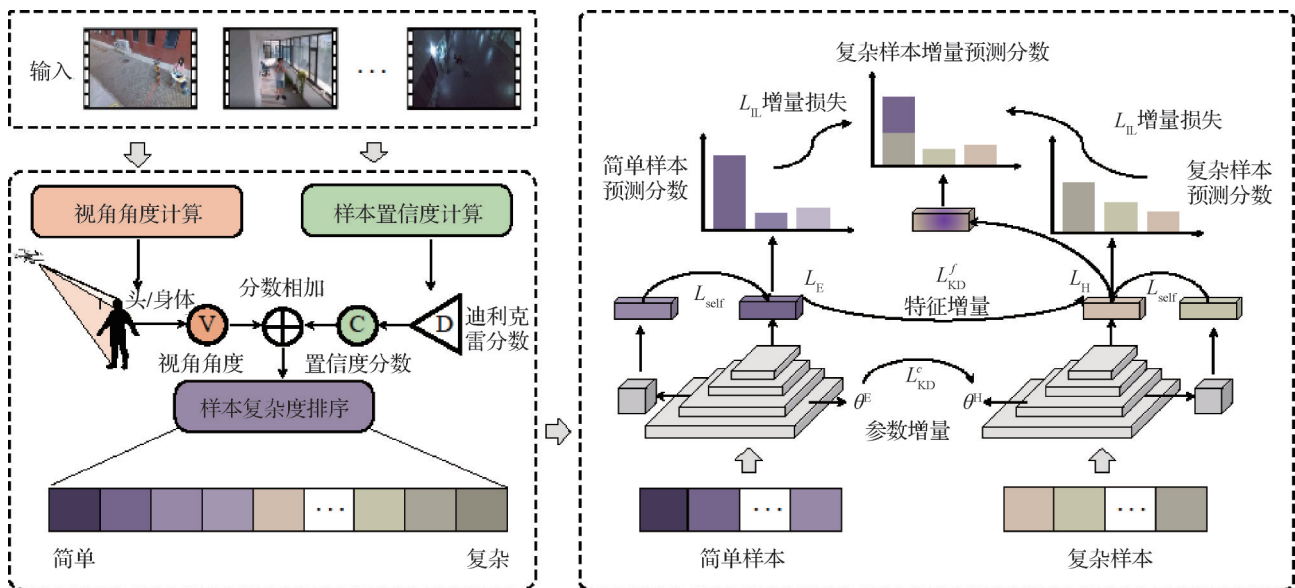


图2 本文提出方法DAIL的框架

Fig. 2 Framework of our proposed DAIL

失函数 L_{ll} , 简单样本学习损失函数 L_E 以及困难样本学习损失函数 L_{hl} 。

2.1 判别行为信息有序蒸馏

研究空一地多视角场景下行为识别的判别行为信息提取问题, 并根据头身比和置信度来度量判别行为信息, 从判别行为信息量的角度对样本进行有序排列。

在行为识别任务中, 网络应该关注行为主体本身, 因此本节的目标是通过观察到的人的判别行为信息量来区分视角。基于透视学基本原理, 将无人机样本根据视角高度划分为地面视角和高空视角。对于每个样本, 从不同高度视角看到的头身比不同, 高空观察的样本由于俯拍导致头在整个身体中所占的比例比较大, 而地面视角由于平视拍摄, 头部对于整个身体所占比例较小, 头部检测框对角线和人体检测框对角线的比会随着视角高度变化发生改变, 因此使用头身比作为度量视角高度的一种方法。具体来说, 使用 YOLO (you only look once) 系列 (<https://github.com/ultralytics/yoloV5>) 检测网络来检测人的头部和身体的检测框。检测框的4个坐标可以用来计算检测框的高度和宽度。头部检测框的对角线长度 (B_h) 与身体检测框的对角线长度 (B_b) 的比例会随着视角改变发生变化, 因此可以作为度量视角的方法。视角的计算为

$$\cos \frac{v_i}{2} = \frac{B_h}{B_b}, \quad v_i = 2 \arccos \frac{B_h}{B_b} \quad (1)$$

式中, v_i 代表第 i 个样本的视角度量分数, B_h 代表头部检测框的对角线长度, B_b 代表身体检测框的对角线长度。其中, 视角度量参数的大小决定了人的行为信息被观察到的程度, 头与身体的比例 (v_i) 越小, 视角越平, 表示可以观察到的信息越多。但仅通过视角度量参数来区分样本是不可靠的, 这可能导致样本偏差和样本不平衡的问题。本节考虑计算样本的置信度分数, 并将其与视角度量参数结合起来以增加样本分类的可信度。

同时, 自适应调整分类阈值确保样本平衡分布。一个合理的分类阈值可以使样本均匀分布, 避免地面视角样本和高空视角样本数量的分布差异过大导致网络过拟合。计算不同样本的置信度分数可以判断样本中包含的有效信息量的可信度, 在视角度量参数一致的情况下, 可以揭示有效信息量大小。置信度分数越大, 表明样本包含更多的行为信息。在

多分类任务中, 主观逻辑 (Jøsang, 2016) 将迪利克雷分布的参数与证据分布联系起来, 其中迪利克雷分布可以被认为是多分类分布的共轭先验。主观逻辑定义了一个理论框架, 从数据中收集的预测分数来获得样本置信度分数。具体计算为

$$c_i = 1 - \frac{K}{D_j}, \quad D_j = \sum_{j=1}^K (d_j + 1) = \sum_{j=1}^K \alpha_j \quad (2)$$

式中, c_i 表示第 i 个样本的置信度分数, K 表示所有行为类别的数量, D_j 表示迪利克雷强度, d_j 表示使用线性层和线性整流函数层 (rectified linear unit, ReLU) 计算的预测分数, 迪利克雷分布 $\alpha_j = d_j + 1$ 。在每10轮后保存所有样本的 v_i 和 c_i , 并考虑样本平衡, 使用最大类间方差来确定分类的阈值 μ 。这种方法根据选定的优先级阈值将样本分为简单样本和困难样本。考虑到 v_i 和 c_i 计算规则, 样本优先级计算为

$$p_i = v_i + c_i \quad (3)$$

式中, p_i 表示第 i 个样本的优先级分数。 v_i 代表第 i 个样本的视角度量分数, c_i 表示第 i 个样本的置信度分数。根据 p_i 可以得到样本优先级, 后计算分类阈值 μ 的自适应规则为

$$\mu = \frac{1}{N} \sum_{i=1}^N p_i (y_i | p_i^r) \quad (4)$$

式中, μ 为阈值, N 为样本数, y_i 和 p_i^r 分别代表第 i 个样本的真实标签和预测标签。图3可视化了样本分布与网络训练的变化趋势, 反映不同组的样本分布对网络学习特征能力的贡献。

具体做法为使用完整数据集训练网络, 将样本分类到不同的区域进行分布内评估。每经过10轮训练收集所有样本的分数, 并根据阈值 μ 将其分为简单样本和困难样本。样本分布规则为

$$u_i = \begin{cases} 0 & p_i > \mu \\ 1 & p_i \leq \mu \end{cases} \quad (5)$$

式中, u_i 表示第 i 个样本的分布类别, $u_i = 0$ 表示样本分类为简单样本, $u_i = 1$ 表示分类为困难样本。

本节采用自蒸馏的方法, 通过 KL (Kullback-Leibler) 散度约束判别行为信息提升模型的识别能力。自蒸馏利用深层特征作为浅层监督信号, 使得网络在初期关注深层次语义, 保证网络整体上的语义一致性, 有助于浅层网络提取更具有判别性的语义特征。给定一个样本, 用骨干网络 ResNet3D (Hara 等, 2018) 提取包含更多细粒度特征信息的浅层网络特征 Z_i^s 和包含丰富语义信息的深层网络特

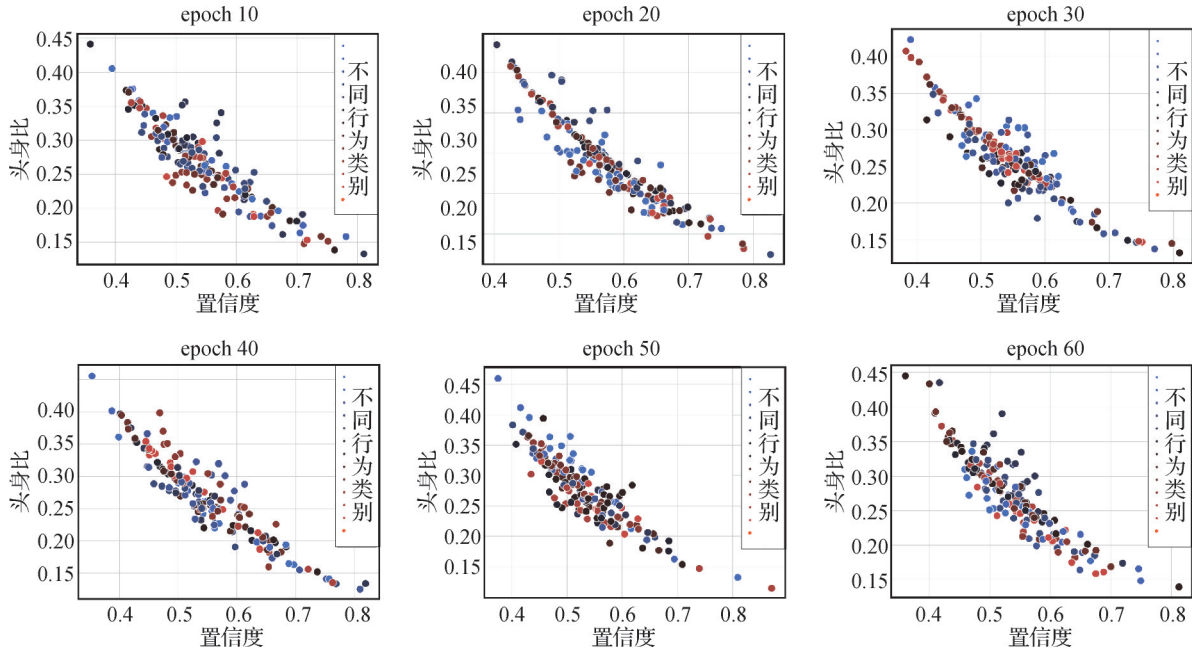


图3 样本分布图

Fig. 3 Sample distribution

征 Z_i^d 。样本自蒸馏损失 L_{dis} 定义为

$$L_{\text{dis}} = \frac{1}{N} \sum_{i=1}^N ((1 - \beta_i) L_{\text{CE}}(Z_i^d, y_i) + \beta_i L_{\text{KL}}(Z_i^d, Z_i^s)) \quad (6)$$

式中, L_{dis} 为样本自蒸馏损失, L_{CE} 表示交叉熵损失, L_{KL} 表示 KL 散度, N 表示当前样本的总数量。浅层特征为 Z_i^s , 深层特征为 Z_i^d 。 $\beta_i \in [0, 1]$, $i = 1, \dots, N$, 表示平衡不同优先级样本重要性比例的超参数。此外, 受到 FitNet 的启发, 对浅层特征增加约束, 以平衡 Z_i^s 和 Z_i^d , 即

$$L_s = \frac{1}{N} \sum_{i=1}^N \lambda_i l_2(Z_i^d, Z_i^s) \quad (7)$$

式中, L_s 为浅层特征约束, $\lambda_i > 0$, $i = 1, \dots, N$, 表示平衡不同优先级样本重要性的超参数, l_2 表示二范数损失函数。综上所述, 自蒸馏的总损失函数为

$$L_{\text{self}} = L_{\text{dis}} + L_s \quad (8)$$

式中, L_{self} 表示自蒸馏的总损失函数。

2.2 判别行为信息增量学习

本节采用类脑的思想学习和理解困难信息, 使用参数和特征增量的方法来通过简单样本指导挖掘困难样本的判别行为信息。对简单样本和困难样本采取相同的自蒸馏方法提取判别行为信息。

$$L_{\text{KD}}(E) = L_{\text{KD}}^f(\theta^E, \theta^H, E) + L_{\text{KD}}^c(\theta^E, \theta^H, E) \quad (9)$$

式中, E 表示简单样本, H 表示困难样本。 θ^E 和 θ^H 分别表示简单样本和困难样本网络的参数。利用知识蒸馏损失 L_{KD} 在特征提取器和分类器中保留旧样

本的信息, L_{KD}^f 和 L_{KD}^c 分别代表特征增量损失和参数增量损失。在网络训练样本时简单样本和困难样本的判别行为信息差通过式(10)进行评估, 具体为

$$L_{\text{KD}}^f(\theta^E, \theta^H, E) = D(F(\theta^E, E), F(\theta^H, E)) \quad (10)$$

式中, D 衡量 $F(\theta^E, E)$ 和 $F(\theta^H, E)$ 之间的距离。 F 为特征提取网络, 考虑到实际执行中的计算复杂性。以参数 θ 为例, 当网络在困难样本上训练时, 不同网络参数 θ 的差表示为 $\Delta\theta$, 其数学形式为

$$\Delta\theta = \theta^E - \theta^H \quad (11)$$

为评估 $\Delta F(\theta^E, \theta^H, E)$, 本节计算 θ^E 和 θ^H 的欧氏距离, 具体为

$$L_{\text{KD}}^f(\theta^E, \theta^H, E) \approx |L_{\text{CE}}(\theta^E, E) - L_{\text{CE}}(\theta^H, E) + L_{\text{CE}}(\Delta\theta, E)|^2 \quad (12)$$

式中, $|\cdot|$ 为绝对值, L_{KD}^c 的一般实现形式可描述为

$$L_{\text{KD}}^c(\theta^E, \theta^H, E) = \sum \sum -\delta(\theta^E, E) \log(\delta(\theta^H, E)) \quad (13)$$

式中, $\delta(\cdot)$ 表示用余弦定理度量相似度。利用特征增量损失约束和参数增量损失约束的协同优化, 实现由简单网络到困难网络的参数增量和特征增量, 有助于困难样本网络利用相同行为共性从简单样本网络中学习到更多有效参数和有效特征, 增强模型对困难样本的行为信息的提取与判别能力, 同时保证对简单样本的判别能力, 从而提高模型对不同视角样本的识别能力。

通过有序关系增量学习, 判别行为信息从简单

样本增量到困难样本,弥补了困难样本包含判别行为信息不足的问题,网络在简单样本的帮助下学习到困难样本中最具判别力的行为特征。简单判别行为信息到困难判别行为信息的增量学习损失函数 L_{IL} 表示为

$$L_{IL}(\hat{H}) = L_{KD}(H) + L_{KD}(E) + L_{self} \quad (14)$$

式中, $\hat{H}$ 为增量判别行为信息后的困难样本。

2.3 损失函数

在训练过程中,使用双流网络来分别训练不同优先级的样本。网络的一个分支学习包含丰富判别行为信息的简单样本。通过对特征和网络参数的增量学习,可以帮助另一个分支更好地学习困难样本。

最终的总损失函数 L_{all} 由3个部分组成,即增量学习损失函数 L_{IL} 、简单样本学习损失函数 L_E 和困难样本学习损失函数 L_H ,具体表述为

$$L_{all} = \varepsilon_1 L_{IL} + \varepsilon_2 L_E + \varepsilon_3 L_H \quad (15)$$

式中, ε_1 、 ε_2 和 ε_3 代表不同损失的权重。 L_E 、 L_H 为简单、困难特征提取网络的分类损失,根据权重分析实验结果, ε_1 、 ε_2 、 ε_3 取 1、1、0.01 可达到最佳效果。

3 实验结果

3.1 数据集

为了评估本文提出的 DAIL 网络的有效性,在两个包括空中视角和地面视角的空一地多视角场景数据集 Drone-Action (Perera 等, 2019) 和 UAV (unmanned aerial vehicle) (Li 等, 2021) 上,以及只包括地面视角的数据集 N-UCLA (Wang 等, 2014) 上进行实验。

Drone-Action 包含 13 种常见的户外人类行为。这些行为由 10 名参与者在无人机处于悬停或跟随模式时进行拍摄。

UAV 是基于无人机的人类行为识别大规模基准数据集,包含 119 个行为主体,428 个带标签的视频。

N-UCLA 由位于不同视角的 3 台摄像机收集。总共包含约 1 500 个视频,涉及 10 个行为主体和 10 个行为类别。

MOD20 数据集 (Perera 等, 2020) 提供包括航空和地面多视角的户外视频,共 2 324 个。视频的分辨率为 720×720 像素,以 29.97 帧/s 的速度录制。

3.2 训练过程

本文使用 ResNet3D (Hara 等, 2018) 作为骨干网络,并选择 SlowOnly (Feichtenhofer 等, 2019) 作为基线模型。在每个三维空间卷积模块之后插入一个卷积核大小为 3 的二维时间卷积模块,然后进行批量归一化和非线性激活。在全局池化层和最后一个全连接层之间使用滤波值为 0.2 的滤波器。初始学习率为 $1E-2$,衰减率为 $1E-4$ 。分别在 Drone-Action、UAV 数据集上使用 PyTorch 框架在 32 GB 内存的 NVIDIA Tesla P100 GPU 上训练 60 轮,在 MOD20 数据集上训练 100 轮。

3.2.1 评价指标

本文遵循传统行为识别的实验设置选择 Top-1 识别准确率评估实验效果。对于增量分析,本文遵循 Wei 等人 (2021) 文中引用的增量评价指标,用平均增量准确率来计算额外增加类别的识别准确率,用平均遗忘率来评估遗忘的信息量。对于增量实验,每次向输入网络添加一个行为类别的样本,设置 1 到 13 表示样本中包含的行为类别的数量。

3.3 对比方法及结果

表 1 展示了本文方法与当前最优方法在 Drone-Action 和 UAV 数据集上的指标结果。除了可见光,还考虑了骨架模态,以进一步验证本文方法在不引入额外数据情况的有效性。

在 Drone-Action 上,本文方法表现次优,分析可能是因为该数据集的视角差距表现并不明显,更加关注模型对视频的时空表征能力。本文方法在 ResNet+ATFR (adaptive temporal feature resolutions) (Fayyaz 等, 2021) 和 SBP (Cheng 等, 2022) 上分别实现了 19.7% 和 18.0% 的精度提升。

UAV 数据集中不同视角差异较大、背景嘈杂,本文方法可以较为明显地利用头身比区分行为信息,所以达到当前最优表现。对于可见光模态,本文方法准确率比 ResNet + ATFR 和 SBP 方法分别高 16.9% 和 16.2%。而对于骨架模态,本文方法比 Shift-GCN (shift-graph convolutional network) (Cheng 等, 2020) 和 HARD-Net (Li 等, 2020) 分别高出 7.2% 和 9.1%。实验证明,本文方法能充分利用判别行为信息,且有效地解决空一地多视角场景下行为识别任务的难点。

3.3.1 判别行为信息有序蒸馏分析

N-UCLA 样本都来自地面摄像机,包含丰富的

表1 本文DAIL与其他当前最优方法在Drone-Action和UAV数据集的比较

Table 1 Comparison of DAIL with other state-of-the-art methods on the Drone-Action and UAV datasets

		/%	
模态	对比方法	Drone-Action	UAV
骨架	HELP-L(Jhuang等,2013)	64.3	–
	P-CNN(Chéron等,2015)	75.9	–
	ST-GCN(Yan等,2018)	–	30.3
	DGNN(Shi等,2019a)	–	29.9
	2s-AGCN(Shi等,2019b)	–	34.8
	HARD-Net(Li等,2020)	–	36.0
	Shift-GCN(Cheng等,2020)	–	37.9
可见光	TSN(Wang等,2016)†	46.4	21.7
	I3D(Carreira和Zisserman,2017)†	50.5	23.9
	X3D(Feichtenhofer,2020)†	52.8	27.3
	ResNet+ATFR(Fayyaz等,2021)†	61.5	28.2
	TSM(Lin等,2022)†	55.4	24.3
	SBP(Cheng等,2022)†	63.2	28.9
	UniFormerV2(Li等,2023a)†	81.9	32.4
	ASAT(Shi等,2023)	–	39.7
	基线方法	61.1	33.9
	本文方法	81.2 (+20.1)	45.1 (+11.2)

注:加粗字体表示各列最优结果,“–”表示数据为空,“†”表示复现结果。

判别行为信息。Drone-Action由地面视角和高空视角组成,高空视角样本包含的判别行为信息不足。在有序关系增量学习中,为了区分数据流中的样本,本文将先输入的样本称为旧样本,后输入的样本称为新样本。为评估简单样本对困难样本的指导效果,本文设置了3种实验环境。第1种设置将Drone-Action数据集的地面视角和高空视角样本作为旧样本和新样本;第2种设置将N-UCLA作为旧样本,而Drone-Action整体作为新样本;第3种设置将Drone-Action作为旧样本,而N-UCLA作为新样本。通过这些设置,本文可以深入研究空地多视角样本对行为识别任务的影响,从而更加全面地评估本文方法的有效性。

如表2所示,本文方法在第2种设置中的准确度比第1种设置高出1.1%,表明网络优先学习简单样

本有助于提高网络的识别能力,因为使用N-UCLA作为简单样本可以提炼出更丰富的判别行为信息,有助于网络学习和理解Drone-Action数据集样本,同时验证了本文“由易到难,循序渐进”的类脑学习思想的有效性。另一方面,第3种设置的准确度比第1种设置低1.4%,表明网络优先学习困难样本会降低网络的识别能力。因为困难样本中稀疏的判别行为信息难以支持简单样本学到更有效的行为信息。通过以上的实验分析,可以发现使用简单样本增强困难样本会给网络学习特征的能力带来明显的改善,反之则会导致准确度下降,这与人脑处理信息的方式是一致的,体现了“由易到难,循序渐进”的思想。

表2 判别行为信息分析实验

Table 2 Discriminative action information analysis

		/%	
对比方法	旧样本	新样本	准确度
基线方法	Drone	Drone	61.1
	N-UCLA	Drone	63.2 (+2.1)
	Drone	N-UCLA	60.5 (–0.6)
本文方法	Drone	Drone	81.2
	N-UCLA	Drone	82.3 (+1.1)
	Drone	N-UCLA	79.8 (–1.4)

注:加粗字体表示每组各列最优结果。

3.3.2 判别行为信息增量学习分析

图4显示了本文方法和其他方法在Drone-Action数据集上的平均增量准确度。结果表明,与当前最优方法相比,本文方法可以弥补不同样本之间判别行为信息的差距。这说明DAIL在学习不同类别的判别行为信息时,能够更好地处理样本之间的差异,从而在新类别的学习中表现出更好的性能。突出本文方法在解决不同样本之间判别行为信息差距的问题上的有效性。

图5显示了本文方法和其他方法在Drone-Action数据集上的平均遗忘。结果表明,随着类别的增加,所有方法的平均遗忘量都显著增加,表明网络在嵌入新类别时存在判别行为信息遗忘的问题。数据显示相对于传统的行为识别方法,本文方法受灾难性遗忘影响减小,表明本文方法在处理新类别的情况下,能够减少遗忘现象的发生,保留更多的旧知识。这证明了本文方法可以有效地缓解判别行为

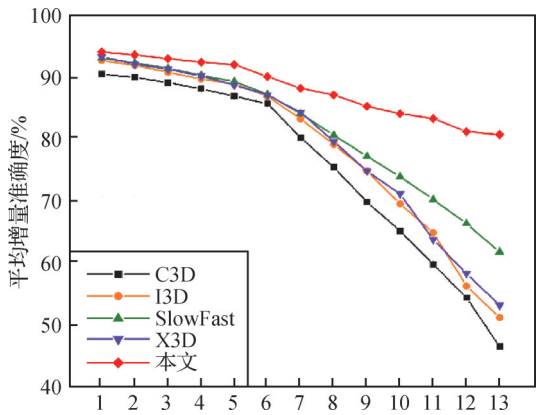


图 4 平均增量准确度示意图
Fig. 4 Illustration of average incremental accuracy

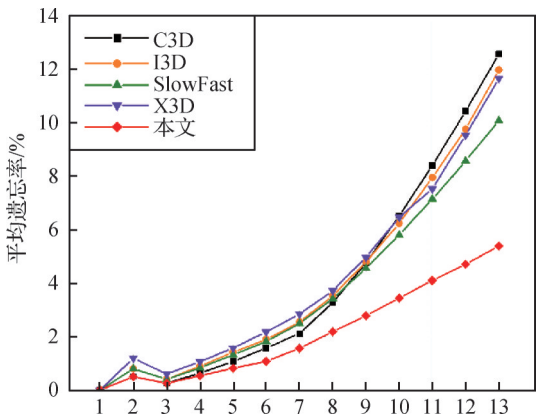


图 5 平均遗忘率示意图
Fig. 5 Illustration of average forgetting rate

特征的灾难性遗忘问题,对于未知或者困难样本具备更好的准确度和鲁棒性。

在 Drone-Action 数据集上,平均遗忘结果见表 3。第 j 类的遗忘为 $f_j = \max_{l \in 1, \dots, k-1} (p_l - p_k)$, 第 k 类的平均遗忘 F_k 的平均增量精度 P_k 定义为

$$F_k = \frac{1}{k-1} \sum_{j=1}^{k-1} f_j, \quad P_k = \sum_{j=1}^k p_j \quad (16)$$

随着类别的增加,所有当前最优方法的平均遗忘率均显著增加。与 X3D (expand 3D) (Feichtenhofer, 2020) 和 SlowFast (Feichtenhofer 等, 2019) 相比,本文方法分别减少了 6.3% 和 4.7% 的遗忘率。证明本文方法比传统的行为识别方法遗忘知识更少,而且可以充分地判别行为信息进行增量。

表 4 总结了其他方法和本文方法在 Drone-

表 3 Drone-Action 数据集上的平均增量遗忘率

Table 3 Average incremental forgetting rate on the Drone-Action dataset

方法	/%												
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
C3D(Tran 等, 2015)	0.00	0.50	0.25	0.63	1.08	1.58	2.12	3.29	4.76	6.53	8.42	10.45	12.60
I3D(Carreira 和 Zisserman, 2017)	0.00	0.80	0.40	0.90	1.43	1.90	2.55	3.52	4.80	6.26	7.97	9.78	12.00
SlowFast(Feichtenhofer 等, 2019)	0.00	0.80	0.40	0.83	1.33	1.82	2.50	3.43	4.56	5.82	7.16	8.59	10.11
X3D(Feichtenhofer, 2020)	0.00	1.20	0.60	1.07	1.58	2.18	2.85	3.73	4.98	6.48	7.55	9.55	11.68
SBP(Cheng 等, 2022)	0.00	0.83	0.50	0.94	1.42	1.99	2.74	3.62	4.75	6.27	7.38	9.25	10.56
本文	0.00	0.50	0.25	0.53	0.83	1.08	1.57	2.19	2.79	3.46	4.11	4.72	5.40

注:加粗字体表示各列最优结果。

表 4 Drone-Action 数据集上的平均增量准确度

Table 4 Average incremental accuracy on the Drone-Action dataset

方法	/%												
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
C3D(Tran 等, 2015)	90.5	90.0	89.1	88.1	86.9	85.7	80.2	75.4	69.8	65.1	59.7	54.3	46.0
I3D(Carreira 和 Zisserman, 2017)	92.7	91.9	90.8	89.7	88.9	86.9	83.2	79.1	74.8	69.5	64.8	56.2	50.5
SlowFast(Feichtenhofer 等, 2019)	93.1	92.3	91.4	90.3	89.3	87.2	84.1	80.6	77.2	73.9	70.2	66.3	61.1
X3D(Feichtenhofer, 2020)	93.3	92.1	91.3	90.2	88.7	87.1	84.3	79.6	74.8	76.1	63.7	55.2	52.8
SBP(Cheng 等, 2022)	93.8	92.3	91.7	91.2	89.5	88.2	85.1	80.7	75.6	74.3	68.1	66.1	63.2
本文	94.1	93.6	93.0	92.4	92.0	90.1	88.2	87.1	85.3	84.1	83.3	81.9	81.2

注:加粗字体表示各列最优结果。

Action 数据集上的平均增量精度结果。与 I3D (Carreira 和 Zisserman, 2017) 和 C3D (Tran 等, 2015) 相比, 本文方法可以分别实现 29.3% 和 34.3% 的精度提升。结果表明, 与当前最优的方法相比, 本文方法可以弥补不同样本之间判别行为信息的差距。

3.4 消融分析

3.4.1 自适应分类阈值的定量消融分析

本文首先使用固定分类阈值在 Drone-Action 数据集上做实验。如表 5 所示, 当使用头身比例作为分类标准时, 头身比例的阈值越小, 简单样本的比例越高, 简单样本对网络学习和理解困难样本的指导作用会越强, 例如阈值为 0.19 时会比 0.21 时的识别效果更好。当使用置信度作为评判标准时, 置信度阈值越低, 表示简单样本的比例越高, 网络学习和理解困难样本的能力会越强。然而, 当样本数量偏差过大时会对网络学习困难样本产生负面影响, 自适应阈值的方法可以避免这一问题。数据显示, 通过自适应地结合头身比例和置信度, 可以达到最佳效果。这表明本文提出的自适应优先级计算方法是有效的, 可以根据不同的数据集和样本分布动态地调整分类阈值, 从而在增量学习中更好地处理困难样本和简单样本, 提高网络的学习性能。

表 5 Drone-Action 数据集上自适应阈值的准确度

Table 5 Accuracy of adaptive thresholds on Drone-Action dataset		
	阈值	准确度/%
头身比	0.19	64.1
	0.20	70.2
	0.21	60.3
	自适应	79.1
置信度	0.4	65.2
	0.5	69.5
	0.6	60.7
	自适应	78.4
本文方法	自适应	81.2

注: 加粗字体表示最优结果。

3.4.2 模块定量消融分析

表 6 展示了 DAIL 中不同模块的作用, 判别行为信息有序蒸馏模块的加入在 Drone-Action 上比基线模型提高了 13.3%, 在 UAV 上比基线模型提高了

6.5%。而使用判别行为信息增量学习模块的网络比基线模型分别提高 17.3% 和 8.7%, 表明信息增量模块的加入可以帮助网络在不忘记旧行为信息的同时积极捕捉新的行为知识。总之, 判别行为信息增量学习模块比判别行为信息有序蒸馏模块对网络有更大的提升, 网络可以在简单样本的指导下学习困难样本。通过增量模块和蒸馏模块的结合使用, 本文方法可以达到最佳效果。

表 6 Drone-Action 和 UAV 数据集上的模块消融实验

Table 6 Module ablation studies on Drone-Action and UAV datasets				
			/%	
基线模型	有序蒸馏	有序增量	Drone-Action	UAV
√	×	×	61.1	33.9
√	√	×	74.4	40.4
√	×	√	78.4	42.6
√	√	√	81.2(+20.1)	45.1(+11.2)

注: 加粗字体表示各列最优结果, “√”表示采用, “×”表示未采用。

3.4.3 损失函数的定量消融分析

表 7 显示了在 Drone-Action 和 UAV 数据集上进行的损失函数参数采样消融实验。本文使用吉布斯采样寻找最佳的 ε_1 、 ε_2 和 ε_3 。首先设定 ε_2 和 ε_3 初始值为 1, 通过多次采样找到最优 ε_1 。随后以类似手法寻找最优 ε_2 和 ε_3 。最优结果中 L_{IL} 和 L_E 占比较大, 表明网络学习判别行为信息对理解困难样本非常重要。

图 6 展示了本文方法在 Drone-Action 和 UAV 数据集上准确率和损失的变化趋势。由于样本数量存在差异, 网络分别在第 30 和 50 轮趋于平稳, 在后续

表 7 Drone-Action 和 UAV 数据集上的损失函数参数采样分析

Table 7 Loss function parameter sampling analysis on Drone-Action and UAV datasets							
						/%	
数据集	参数	0	0.001	0.01	0.1	1	10
Drone-Action	ε_1	69.9	75.6	77.2	79.5	81.2	78.4
	ε_2	65.7	76.4	77.6	79.1	81.2	78.2
	ε_3	71.1	78.5	81.2	79.9	80.3	75.4
	ε_1	37.5	41.4	42.8	44.3	45.1	43.7
UAV	ε_2	35.2	40.8	42.1	43.6	45.1	43.3
	ε_3	40.2	44.3	45.1	43.7	42.9	41.6

注: 加粗字体表示各行最优结果。

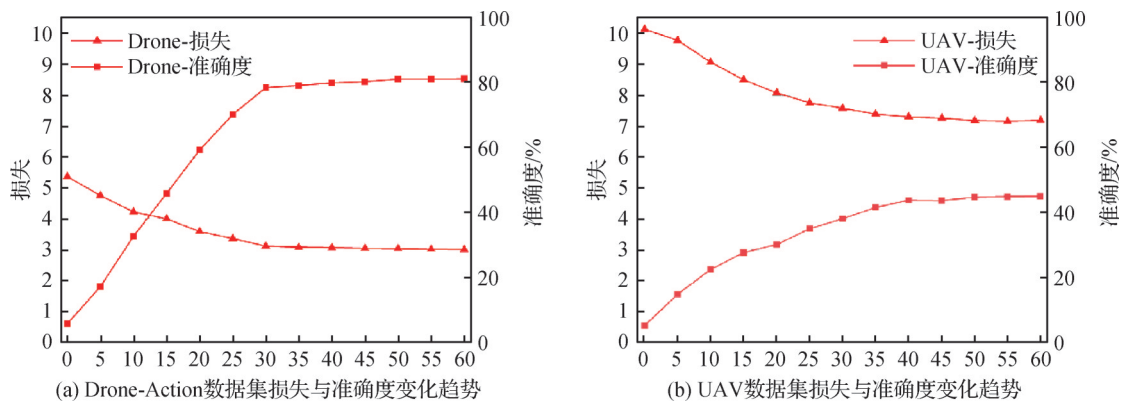


图 6 损失函数和准确率变化曲线示意图

Fig. 6 Illustration of loss function and accuracy variation curves
((a) tendency on Drone-Action dataset; (b) tendency on UAV dataset)

的工作中可以考虑结合早停等训练手法节约训练代价。

3.4.4 时间复杂度分析

对本文方法以及基于可见光的方法 X3D(Feichtenhofer, 2020), 强基线模型 SlowFast (Feichtenhofer 等, 2019) 和基于骨架的方法 Shift-GCN (Cheng 等, 2020) 在 UAV 数据集上做出时间复杂度对比。具体如表 8 所示, 与强基线方法 SlowFast 相比, 本文方法不仅提升了 18.0% 的准确率, 还降低了 2.4 M 的参数量和 6.9 G 的计算量。这得益于本文提出的判别行为信息增量的方法, 避免了重复学习已经学习过的样本, 对于减少时间复杂度有很大的帮助。同时, 本文使用了 SlowOnly 的低采样率作为视频特征处理方法, 因此对比于 SlowFast 在时间复杂度上有进一步的下降。此外, 与基于可见光的方法 X3D 和基于骨架的方法 Shift-GCN 相比, 本文方法在提高识别准确度的同时, 也有效地平衡了时间开销。这表明本文方法在效率和效果方面达到了良好的平衡。综上所述, 本文方法在大幅度提升识别准确度的同时, 并没有过多地增加时间复杂度, 从实验证明了其在效率和效果方面的有效平衡。

3.5 可视化分析

3.5.1 判别行为信息的可视化分析

根据图 7 中的样本结果, 本文方法在处理空一地多视角场景中的人体行为信息时, 能够集中关注于判别行为信息, 并帮助网络学习更多的判别性行为表征。

图 7(a) 表示输入样本的原始帧, 图 7(b) 中, 较浅颜色的区域表示判别行为信息集中的区域, 表明

表 8 UAV 数据集上时间复杂度分析

Table 8 Time complexity analysis on UAV dataset

模型	参数量/M	计算量/G	准确率/%
X3D (Feichtenhofer, 2020)	20.6	20.7	27.3
SlowFast (Feichtenhofer 等, 2019)	34.1	34.2	27.1
Shift-GCN (Cheng 等, 2020)	15.5	10.0	37.9
本文	31.7	27.3	45.1(+7.2)

注: 加粗字体表示各列最优结果。



图 7 判别行为信息注意力示意图

Fig. 7 Illustration of discriminative action information attention
((a) original samples; (b) feature maps;
(c) attention sketch maps in original samples)

本文方法能从视频中提取出关键的行为信息。图7(c)中,红色区域表示最活跃的部分,其次是绿色区域,说明本文方法能够有效地捕捉到人体行为中的重要信息并进行判别。这些样本结果支持了本文方法在提取和学习行为信息方面的优势,以及在复杂场景下的应用潜力。

3.5.2 预测结果的可视化分析

图8展示了本文从UAV数据集中得到的定性结果,绿色框表示预测正确,红色框表示预测错误。如样本识别结果所示,有许多因素会影响样本的识别精度,如自遮挡、视角变化、光线变化和相机晃动等。“捡拾”等行为运动信息量大,光照明亮且视频相对稳定,这对网络识别这些样本起到积极的影响,因此可以观察到这些行为类别的判别行为信息是相对充足的。

但是对于“戴耳机”等行为的判别性特征并不明

显,传统方法很难学习到判别性信息,因此不能正确识别。本文提出的方法通过增量判别行为信息来帮助网络学习判别性特征,可以比其他方法更准确地识别这些困难样本。对于“鼓掌”等行为,高空摄像机拍摄导致样本画面模糊,拍摄场景光线昏暗导致难以学习判别行为信息,识别效果不佳。

图9展示了混淆矩阵以评估本文方法的行为识别准确度。由于“跑步”的不同形态表现较为相似,基线模型的准确率较低。而本文方法表现优异,通过使用判别行为信息可以更好地捕捉相似行为的细粒度信息。DAIL对于“挥舞瓶子”行为识别困难,可能原因是瓶子颜色与背景颜色接近,会对网络造成干扰。此外,发现不同网络还可能混淆“慢跑”和“跑步”,网络对于不同行为的速率捕捉能力欠缺。

3.5.3 特征分布的可视化分析

UAV数据集的行为类别过于丰富,很难分析特



图8 Drone-Action数据集样本识别准确度示意图

Fig. 8 Illustration of sample recognition accuracy on Drone-Action dataset

(a) 基线模型

(b) 本文方法

Fig. 9 Confusion matrix for the Drone-Action dataset ((a)confusion matrix of baseline; (b)ours)

线方法。然而,从图中可以发现仍然存在一些特征重叠的现象,这可能与现实场景中的光照、遮挡和无人机拍摄画面抖动等因素有关。这些现实问题也是今后关注和继续深入研究的方向。

与基线模型相比,本文方法显著提高了类间差异,使得分类边界更清晰。同时,从图9中可以发现,本文方法对于不同子类的识别准确度均高于基

为了验证本文方法在不同数据集上的泛化性,引入MOD20数据集进行实验,如表9所示。实验中挑选几种不同时期的强基线网络进行对比,发现本

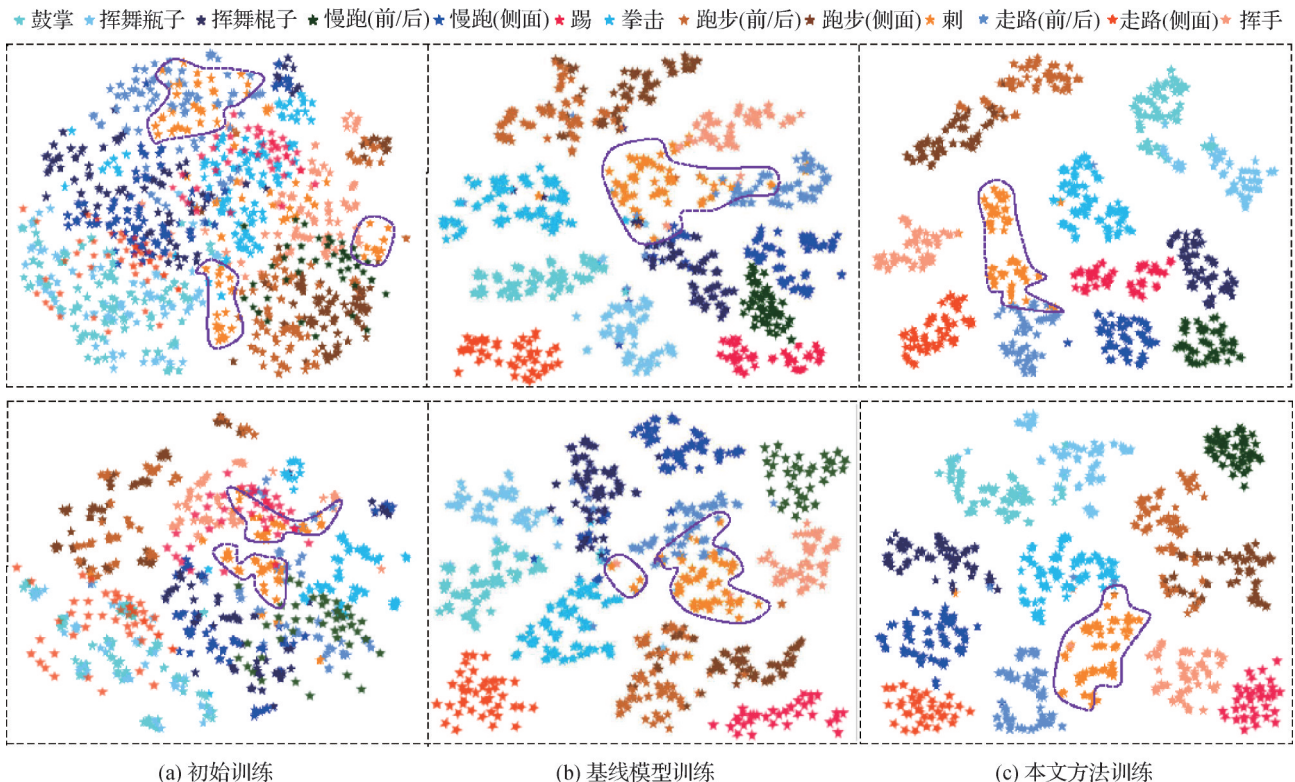


Fig. 10 Illustration of feature distribution ((a) initial training; (b) baseline training; (c) ours)

文方法在 Top-1 准确率达到最高为 95.1%。在 SlowOnly 和 X3D 的基线上增加本文提出训练设置后发现都有不同程度的提升,表明本文方法在不同基线上同样具有较好的可移植性。I3D 对于 MOD20 数据集的表现较差,表明该网络对于处理视角差异较大的场景中的能力欠缺。

表 9 MOD20数据集上泛化性分析
Table 9 Generalization analysis on MOD20 dataset
/%

模型	Top-1	Top-5
I3D(Carreira 和 Zisserman, 2017)	71.2	94.5
TSN(Wang 等, 2016)	88.8	98.2
SlowOnly(Feichtenhofer 等, 2019)	69.9	94.6
本文方法(SlowOnly)	95.1(+25.2)	97.6(+3.0)
X3D(Feichtenhofer, 2020)	92.4	99.1
本文方法(X3D)	93.2(+0.8)	99.3(+0.2)

注:加粗字体表示各列最优结果。

4 结 论

本文提出了一种“由易到难,循序渐进”的类脑学习思想和一个判别行为信息增量学习网络,针对空一地多视角场景下的行为识别任务,有效地解决了高空视角下的行为信息获取和利用难题,取得了优异的表现。与传统分类方法提出困难样本不同,本文设计的判别行为信息有序蒸馏方法通过类脑的方式对样本由简单到复杂进行排序,并将样本分类为容易学习的地面视角简单样本和难以学习的高空视角困难样本,分别提取判别行为信息。此外,本文设计了一种判别行为信息增量方法对网络的参数以及判别性行为特征进行增量,实现判别行为信息的跨视角增量,解决了跨视角判别行为信息的灾难性遗忘问题。

实验结果表明,本文方法对比于当前最优方法 SBP,在 Drone-Action 上提高了 18.0%,在 UAV 上提高了 16.2%。与强基线方法相比,本文方法在 UAV 数据集上参数量减少了 2.4 M,计算量减少了 6.9 G。同时与动作识别典型方法 I3D、SlowOnly 等相比,本文在 MOD 数据集上有较强的泛化能力。

然而,值得注意的是,虽然本文方法在空天多视

角行为识别中封闭数据集中取得了出色的结果,但是在现实应用中仍存在缺乏训练数据、应对复杂环境能力较弱的缺点。因此,未来的工作可以在以下几个方面进行拓展。

首先,将继续探讨视觉知识智能进化中知识重建及新老知识的融合,以进一步提升本文新旧知识交互的合理性。同时结合现实复杂应用场景中新知识、老知识中可能存在的噪声信息,采用主动学习方式自适应汲取更纯净的有效行为信息。此外,在开放环境中应进一步考虑新类别与旧类别中可能存在的灾难性遗忘问题,确保模型能够有效更新。同时应及时跟进多模态大模型,实现不同模态之间的知识交互与融合。

参考文献(References)

Bin Y R, Cao X, Chen X Y, Ge Y H, Tai Y, Wang C J, Li J L, Huang F Y, Gao C X and Sang N. 2020. Adversarial semantic data augmentation for human pose estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 606-622 [DOI: 10.1007/978-3-030-58529-7_36]

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chen P H, Yang H C, Chen K W and Chen Y S. 2020. MVSNet++: learning depth-based attention pyramid features for multi-view stereo. IEEE Transactions on Image Processing, 29: 7261-7273 [DOI: 10.1109/TIP.2020.3000611]

Cheng F, Xu M Z, Xiong Y J, Chen H, Li X Y, Li W and Xia W. 2022. Stochastic backpropagation: a memory efficient strategy for training video models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8291-8300 [DOI: 10.1109/CVPR52688.2022.00812]

Cheng K, Zhang Y F, He X Y, Chen W H, Cheng J and Lu H Q. 2020. Skeleton-based action recognition with shift graph convolutional network//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 180-189 [DOI: 10.1109/CVPR42600.2020.00026]

Chéron G, Laptev I and Schmid C. 2015. P-CNN: pose-based CNN features for action recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3218-3226 [DOI: 10.1109/ICCV.2015.368]

Dave I R, Chen C and Shah M. 2022. SPAct: self-supervised privacy preservation for action recognition//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New

- Orleans, USA: IEEE: 20132-20141 [DOI: 10.1109/CVPR52688.2022.01953]
- Fayyaz M, Bahrami E, Diba A, Noroozi M, Adeli E, van Gool L and Gall J. 2021. 3D CNNs with adaptive temporal feature resolutions// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4729-4738 [DOI: 10.1109/CVPR46437.2021.00470]
- Feichtenhofer C. 2020. X3D: expanding architectures for efficient video recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 200-210 [DOI: 10.1109/CVPR42600.2020.00028]
- Feichtenhofer C, Fan H Q, Malik J and He K M. 2019. SlowFast networks for video recognition//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6201-6210 [DOI: 10.1109/ICCV.2019.00630]
- Geng Y, Han Z B, Zhang C Q and Hu Q H. 2021. Uncertainty-aware multi-view representation learning//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI Press: 7545-7553 [DOI: 10.1609/AAAI.V35I9.16924]
- Hara K, Kataoka H and Satoh Y. 2018. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?//Proceedings of 2018 IEEE/CVF conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 6546-6555 [DOI: 10.1109/CVPR.2018.00685]
- Hardy M. 2010. Pareto's law. The Mathematical Intelligencer, 32(3): 38-43 [DOI: 10.1007/s00283-010-9159-2]
- Ho M K, Abel D, Correa C G, Littman M L, Cohen J D and Griffiths T L. 2022. People construct simplified mental representations to plan. Nature, 606(7912): 129-136 [DOI: 10.1038/s41586-022-04743-9]
- Jhuang H, Gall J, Zuffi S, Schmid C and Black M J. 2013. Towards understanding action recognition//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 3192-3199 [DOI: 10.1109/ICCV.2013.396]
- Jøsang A. 2016. Subjective Logic: A Formalism for Reasoning under Uncertainty. Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer: 1-326 [DOI: 10.1007/978-3-319-42337-1]
- Kong Y, Ding Z M, Li J and Fu Y. 2017. Deeply learned view-invariant features for cross-view action recognition. IEEE Transactions on Image Processing, 26(6): 3028-3037 [DOI: 10.1109/TCSVT.2018.28-68123]
- Li J F and Zhang F Y. 2015. Human behavior recognition based on directional weighting local space-time features. Journal of Image and Graphics, 20(3): 320-331 (李俊峰, 张飞燕. 2015. 基于局部时空特征方向加权的人体行为识别. 中国图象图形学报, 20(3): 320-331) [DOI: 10.11834/jig.20150303]
- Li K, Wang Y, He Y, Li Y Z, Wang Y, Wang L M and Qiao Y. 2023a. UniFormerV2: spatiotemporal learning by arming image ViTs with video UniFormer//Proceedings of 2023 International Conference on Computer Vision. Paris, France: IEEE: 1-24 [DOI: 10.48550/arXiv.2211.09552]
- Li S, He X X, Song W F, Hao A M and Qin H. 2023b. Graph diffusion convolutional network for skeleton based semantic recognition of two-person actions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7): 8477-8493 [DOI: 10.1109/TPAMI.2023.3238411]
- Li T J, Liu J, Zhang W and Duan L Y. 2020. HARD-Net: hardness-aware discrimination network for 3D early activity prediction//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 420-436 [DOI: 10.1007/978-3-030-58621-8_25]
- Li T J, Liu J, Zhang W, Ni Y, Wang W Q and Li Z H. 2021. UAV-Human: a large benchmark for human behavior understanding with unmanned aerial vehicles//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16261-16270 [DOI: 10.1109/CVPR46437.2021.01600]
- Lifshitz O and Wolf L. 2021. Sample selection for universal domain adaptation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI Press: 8592-8600 [DOI: 10.1609/AAAI.v35i10.17042]
- Lin J, Gan C, Wang K and Han S. 2022. TSM: temporal shift module for efficient and scalable video understanding on edge devices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(5): 2760-2774 [DOI: 10.1109/TPAMI.2020.3029799]
- Liu W X, Zhong X, Jia X M, Jiang K and Lin C W. 2022. Actor-aware alignment network for action recognition. IEEE Signal Processing Letters, 29: 2597-2601 [DOI: 10.1109/LSP.2022.3229646]
- Liu X L, Masana M, Herranz L, Van De Weijer J, Lopez A M and Bagdanov A D. 2018. Rotate your networks: better weight consolidation and less catastrophic forgetting//Proceedings of the 24th International Conference on Pattern Recognition. Beijing, China: IEEE: 2262-2268 [DOI: 10.1109/ICPR.2018.8545895]
- Luo C, Zhao P, Chen C, Qiao B, Du C, Zhang H Y, Wu W, Cai S W, He B, Rajmohan S and Lin Q W. 2021. PULNS: positive-unlabeled learning with effective negative sample selector//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI Press: 8784-8792 [DOI: 10.1609/AAAI.v35i10.17064]
- Mazzia V, Angarano S, Salvetti F, Angelini F and Chiaberge M. 2022. Action Transformer: a self-attention model for short-time pose-based human action recognition. Pattern Recognition, 124: #108487 [DOI: 10.1016/J.PATCOG.2021.108487]
- Perera A G, Law Y W and Chahl J. 2019. Drone-Action: an outdoor recorded drone video dataset for action recognition. Drones, 3(4): #82 [DOI: 10.3390/DRONES3040082]
- Perera A G, Law Y W, Ogunwa T T and Chahl J. 2020. A multiview-point outdoor dataset for human action recognition. IEEE Transac-

- tions on Human-Machine Systems, 50(5): 405-413 [DOI: 10.1109/THMS.2020.2971958]
- Rebuffi S A, Kolesnikov A, Sperl G and Lampert C H. 2017. iCaRL: incremental classifier and representation learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5533-5542 [DOI: 10.1109/CVPR.2017.587]
- Shao Z P, Li Y F and Zhang H. 2021. Learning representations from skeletal self-similarities for cross-view action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 31(1): 160-174 [DOI: 10.1109/TCSVT.2020.2965574]
- Shi G G, Fu X Y, Cao C Z and Zha Z J. 2023. Alleviating spatial misalignment and motion interference for UAV-based video recognition//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 193-202 [DOI: 10.1145/3581783.3611799]
- Shi H Y, Hou Z J, Chao X and Zhong Z K. 2023. Multimodal spatial-temporal feature representation and its application in action recognition. Journal of Image and Graphics, 28(4): 1041-1055 (施海勇, 侯振杰, 巢新, 钟卓鲲. 2023. 多模态时空特征表示及其在行为识别中的应用. 中国图象图形学报, 28(4): 1041-1055) [DOI: 10.11834/jig.211217]
- Shi L, Zhang Y F, Cheng J and Lu H Q. 2019a. Skeleton-based action recognition with directed graph neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7904-7913 [DOI: 10.1109/CVPR.2019.00810]
- Shi L, Zhang Y F, Cheng J and Lu H Q. 2019b. Two-stream adaptive graph convolutional networks for skeleton-based action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12018-12027 [DOI: 10.1109/CVPR.2019.01230]
- Shin H, Lee J K, Kim J and Kim J. 2017. Continual learning with deep generative replay//Advances in 31st Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 2994-3003
- Simonyan K and Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos//Advances in 27th Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 568-576
- Swayamdipta S, Schwartz R, Lourie N, Wang Y Z, Hajishirzi H, Smith N A and Choi Y. 2020. Dataset cartography: mapping and diagnosing datasets with training dynamics//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Online: ACL: 9275-9293 [DOI: 10.18653/v1/2020.emnlp-main.746]
- Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]
- Ullah A, Muhammad K, Hussain T and Baik S W. 2021. Conflux LSTMs network: a novel approach for multi-view action recognition. Neurocomputing, 435: 321-329 [DOI: 10.1016/J.NEUCOM.20-19.12.151]
- Vyas S, Rawat Y S and Shah M. 2020. Multi-view action recognition using cross-view video prediction//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 427-444 [DOI: 10.1007/978-3-030-58583-9_26]
- Wang J, Nie X H, Xia Y, Wu Y and Zhu S C. 2014. Cross-view action modeling, learning, and recognition//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 2649-2656 [DOI: 10.1109/CVPR.2014.339]
- Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O and van Gool L. 2016. Temporal segment networks: towards good practices for deep action recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 20-36 [DOI: 10.1007/978-3-319-46484-8_2]
- Wang Q, Sun G, Dong J H, Wang Q Q and Ding Z M. 2022. Continuous multi-view human action recognition. IEEE Transactions on Circuits and Systems for Video Technology, 32(6): 3603-3614 [DOI: 10.1109/TCSVT.2021.3112214]
- Wei K, Deng C, Yang X and Li M S. 2021. Incremental embedding learning via zero-shot translation//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI Press: 10254-10262 [DOI: 10.1609/AAAI.v35i11.17229]
- Wu W H, He D L, Lin T W, Li F, Gan C and Ding E R. 2021. MVF-Net: multi-view fusion network for efficient video recognition//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI Press: 2943-2951 [DOI: 10.1609/AAAI.v35i4.16401]
- Xiao Y, Chen J, Wang Y C, Cao Z G, Zhou J T and Bai X. 2019. Action recognition for depth video using multi-view dynamic images. Information Sciences, 480: 287-304 [DOI: 10.1016/J.INS.2018.12.050]
- Xu C, Wu X, Li Y C, Jin Y N, Wang M M and Liu Y. 2021. Cross-modality online distillation for multi-view action recognition. Neurocomputing, 456: 384-393 [DOI: 10.1016/J.NEUCOM.2021.05.077]
- Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA: AAAI Press: 7444-7452 [DOI: 10.1609/AAAI.v32i1.12328]
- Yang Y, Zhuang Y T and Pan Y H. 2022. The review of visual knowledge: a new pivot for cross-media intelligence evolution. Journal of Image and Graphics, 27(9): 2574-2588 (杨易, 庄越挺, 潘云鹤. 2022. 视觉知识: 跨媒体智能进化的新支点. 中国图象图形学报, 27(9): 2574-2588) [DOI: 10.11834/jig.211264]

- Yu L, Twardowski B, Liu X L, Herranz L, Wang K, Cheng Y M, Jui S and Van De Weijer J. 2020. Semantic drift compensation for class-incremental learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE: 6980-6989 [DOI: 10.1109/CVPR42600.2020.00701]
- Zhang B B, Ge S Y, Wang Q L and Li P H. 2021. Multi-order information fusion method for human action recognition. *Acta Automatica Sinica*, 47(3): 609-619 (张冰冰, 葛疏雨, 王旗龙, 李培华). 2021. 基于多阶信息融合的行为识别方法研究. *自动化学报*, 47(3): 609-619 [DOI: 10.16383/j.aas.2018.c180265]
- Zhao L, Wang Y X, Zhao J P, Yuan L Z, Sun J J, Schroff F, Adam H, Peng X, Metaxas D and Liu T. 2021. Learning view-disentangled human pose representation by contrastive cross-view mutual information maximization//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE: 12788-12797 [DOI: 10.1109/CVPR46437.2021.01260]
- Zhong X, Zhou Z, Liu W X, Jiang K, Jia X M, Huang W X and Wang Z. 2022. VCD: view-constraint disentanglement for action recognition//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE: 2170-2174 [DOI: 10.1109/ICASSP43922.2022.9747610]
- Zhou B and Li J F. 2020. Human action recognition combined with

object detection. *Acta Automatica Sinica*, 46(9): 1961-1970 (周波, 李俊峰). 2020. 结合目标检测的人体行为识别. *自动化学报*, 46(9): 1961-1970 [DOI: 10.16383/j.aas.c180848]

作者简介

刘文璇,女,博士研究生,主要研究方向为视频理解和行为识别。E-mail:lwxfight@126.com

钟欣,通信作者,男,教授,主要研究方向为计算机视觉、多媒体信息处理和类脑视觉。E-mail:zhongx@whut.edu.cn

徐晓玉,女,硕士研究生,主要研究方向为行为识别和计算机视觉。E-mail:xiaoyuchsn@foxmail.com

周卓,男,博士研究生,主要研究方向为动作识别和隐秘意图识别。E-mail:zhouz_aim@whu.edu.cn

江奎,男,教授,主要研究方向为视频图像处理增强、多模态融合感知、数字人生成和视觉驱动。

E-mail:jiangkui@hit.edu.cn

王正,男,教授,主要研究方向为多媒体内容分析与检索、大数据分析 and 自然行为与社会理解。

E-mail:wangzwhu@whu.edu.cn

白翔,男,教授,主要研究方向为计算机视觉与模式识别。

E-mail:xbai@hust.edu.cn