中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-16

论文引用格式:Du Haitao, Zhao Feng, Liu Hanqiang, Tang Yao. Dense-Sparse Skeleton Joint Representation-Guided RGB Image ROI Localization for Multimodal Action Recognition[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260177.(杜海涛, 赵凤, 刘汉强, 唐垚. 稠密-稀疏骨骼联合表征引导RGB图像ROI定位的多模态行为识别[J/OL]. 中国图象图形学报,XXXX:1-16. DOI: 10.11834/jig.260177.)
[DOI:10.11834/jig.260177]

稠密–稀疏骨骼联合表征引导RGB图像ROI定位的多模态行为识别

杜海涛[1],赵凤[1*],刘汉强[2],唐垚[1]

1. 西安邮电大学通信与信息工程学院,陕西西安　710121; 2. 陕西师范大学人工智能与计算机学院,　陕西西安　710119

摘　要:**目的**　人体行为识别在智能监控、人机交互和辅助医疗等场景中具有重要应用价值。针对现有图卷积方法中骨骼模态难以同时建模局部运动模式与远距离全局运动模式的缺陷,RGB模态在复杂背景下难以聚焦动作关键区域以及多模态融合过程中难以充分挖掘模态间互补信息的问题,本文提出一种基于稠密-稀疏骨骼联合表征引导RGB图像ROI(Region of Interest)定位的多模态行为识别网络。**方法**　首先,构建稠密-稀疏骨骼联合表征框架,同时建模相邻关节的局部运动模式与远距离关键关节的全局运动模式,进而增强骨骼模态的时空表征能力;其次,设计基于跨模态注意力的粗细粒度骨骼引导RGB ROI定位策略,利用稀疏骨骼特征和稠密骨骼特征对RGB模态分别进行粗和细粒度两阶段引导ROI,逐步强化动作相关视觉区域并抑制背景噪声;最后,提出骨骼主导的跨模态门控融合模块,以骨骼特征生成门控权重对RGB特征进行自适应加权,进而与骨骼特征拼接实现两种模态互补信息的有效融合。**结果**　在NTU-RGB+D 60、NTU-RGB+D 120和UAV-Human数据集上的实验结果表明,所提方法在多个评估基准下均取得了理想的识别性能。具体来说,本文方法在NTU-RGB+D 60数据集的X-Sub和X-View两个基准上的准确率分别达到94.70%和98.30%,在NTU-RGB+D 120数据集的X-Sub和X-Set两个基准上的准确率分别达到92.82%和93.91%,在低空场景下无人机拍摄的UAV-Human数据集的CSv1和CSv2两个基准上的准确率分别达到53.60%和76.90%。**结论**　所提方法能够充分挖掘骨骼模态与RGB模态之间的互补关系,既增强了骨骼对局部与全局运动关系的建模能力,又提升了RGB模态对动作关键区域的关注程度,从而有效提高复杂场景下人体行为识别的准确性与鲁棒性。

关键词:人类行为识别;多模态融合;图卷积网络;跨模态注意力;ROI定位

Dense-Sparse Skeleton Joint Representation-Guided RGB Image ROI Localization for Multimodal Action Recognition

Du Haitao[1], Zhao Feng[1*], Liu Hanqiang[2], Tang Yao[1]

1. School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China; 2. School of Artificial Intelligence and Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710119, China

Abstract: **Objective** Human action recognition plays a critical role in intelligent surveillance, human–computer interaction, and assisted healthcare. Multimodal action recognition based on skeleton sequences and RGB videos has attracted increasing attention in recent years due to its ability to integrate complementary structural motion information and visual

收稿日期:2026-04-04;**修回日期**:2026-06-11

*通信作者:赵凤 zhaofeng201@xupt.edu.cn

基金项目:国家自然科学基金(62071379);陕西高校青年创新团队。

Supported by: National Natural Science Foundation of China(62071379); The Youth Innovation Team of Shaanxi Universities.

appearance cues. Skeleton data provide relatively stable structural priors of the human body and are less sensitive to background clutter and appearance variations, while RGB videos contain rich texture, context, and human-object interaction information that cannot be fully captured by skeletons alone. However, existing methods still face several challenges. First, most skeleton-based approaches built on graph convolutional networks rely on dense natural body topology, which is effective for modeling local motion patterns among adjacent joints but insufficient for capturing long-range dependencies among semantically related yet spatially distant joints. In many actions, discriminative cues depend not only on local movements but also on global coordination across distant body parts. Second, RGB-based methods are highly susceptible to background clutter, viewpoint variations, scale changes, and irrelevant regions, making it difficult to focus on action-relevant areas, especially in complex scenes. Third, many multimodal fusion methods adopt simple feature concatenation or modality-symmetric fusion strategies, which fail to fully exploit the complementary information across modalities and do not adequately reflect the structural stability of the skeleton modality. **Method** To address these issues, this paper proposes a Dense-Sparse Skeleton Joint Representation-Guided RGB Image ROI Localization for Multimodal Action Recognition network a dense-sparse skeleton joint representation framework is constructed. The dense skeleton branch preserves the complete physical topology of the human body and is responsible for modeling local motion patterns among adjacent joints. Because it retains full joint connectivity, this branch is effective for describing local posture transitions and fine-grained motion dynamics. In contrast, the sparse skeleton branch retains only key joints, including the head, hands, elbows, knees, feet, and a trunk center. By compressing topological paths, it effectively enhances long-range dependency modeling across distant body parts. Compared with the dense branch, the sparse branch focuses more on global coordination relationships and the overall action structure. Through jointly learning dense and sparse skeleton representations, the model achieves unified modeling of local motion dynamics and global coordination relationships. In this way, the proposed framework improves the spatiotemporal representation capability of the skeleton modality by combining local structural continuity with long-range semantic interaction. Second, a Cross-Modal Attention-Based Coarse-to-Fine Skeleton-Guided RGB ROI Localization Strategy is proposed to enhance the discriminative capability of the RGB modality. Specifically, a two-stage coarse-to-fine guided ROI localization mechanism is designed. In the first stage, sparse skeleton features are used to perform coarse-grained ROI localization, where the overall action structure serves as a prior to guide the RGB branch to focus on the human body and major motion regions while suppressing large-scale background noise. This stage mainly emphasizes the action subject and broad movement-related regions, enabling the visual branch to quickly concentrate on the main action area. In the second stage, dense skeleton features are employed to conduct fine-grained ROI localization on the enhanced visual features, further emphasizing discriminative local regions such as hands, elbows, knees, and human-object interaction areas. Since dense skeleton features preserve richer local topology, they provide more precise structural guidance for highlighting subtle but important regions for action discrimination. The two stages therefore serve different but complementary purposes: the first provides global localization, while the second refines local details. Through this progressive process, the RGB modality is guided from coarse subject-level focus to fine-grained action-related region enhancement, which helps reduce the influence of irrelevant background information. In addition, the proposed ROI localization is performed in feature space rather than by explicitly cropping raw RGB frames, making the visual enhancement process more stable and flexible. Finally, a skeleton-dominant cross-modal gated fusion module is designed to achieve effective multimodal integration. Unlike conventional modality-symmetric fusion strategies, the proposed method treats skeleton features as the dominant representation and generates gating weights to adaptively reweight RGB features along the channel dimension. The rationale is that skeleton data provide more stable and task-relevant structural cues, whereas RGB information, although rich in appearance semantics, is also more vulnerable to environmental noise. Under this mechanism, RGB information is incorporated as a complementary modality conditioned on skeleton semantics, enabling the model to selectively utilize visual appearance cues that are beneficial for distinguishing fine-grained actions while suppressing irrelevant noise. In this way, the complementary relationship between skeleton and RGB modalities can be exploited more effectively. In addition, the entire framework is trained in a joint optimization manner with both primary and auxiliary supervision. The fusion branch is supervised by the main classification loss, while the skeleton and RGB branches are equipped with auxiliary losses to enhance feature discriminability and improve training stability. This training strategy helps optimize

both fused representation learning and branch-specific representation quality. **Result** Extensive experiments are conducted on three public benchmarks, including NTU-RGB+D 60, NTU-RGB+D 120, and UAV-Human. Experimental results show that the proposed method achieves 94. 7% and 98. 3% accuracy under the X-Sub and X-View protocols on NTU-RGB+D 60, and 92. 82% and 93. 91% under the X-Sub and X-Set protocols on NTU-RGB+D 120. On the UAV-Human dataset, it achieves 53. 60% and 76. 90% under the CSv1 and CSv2 protocols, respectively. Compared with existing skeleton-based and multimodal methods, the proposed approach demonstrates superior performance and stronger robustness under various evaluation settings. The results on the NTU datasets indicate that the proposed framework can effectively exploit complementary structural and visual information in standard benchmark scenarios. More importantly, its performance on UAV-Human further demonstrates its robustness in low-altitude UAV-view scenes, where the task is more challenging because of viewpoint changes, target scale variation, occlusion, and complex backgrounds. These results verify that the proposed framework is effective not only in relatively controlled indoor environments but also in more difficult open-view conditions. Ablation studies further validate the effectiveness of each component. The dense-sparse skeleton joint representation significantly outperforms single dense skeleton modeling, showing that combining dense topology with sparse key-joint topology is beneficial for capturing both local motion patterns and long-range dependency relationships. The coarse-to-fine guided ROI localization consistently achieves better performance than single-stage guidance, confirming that progressive visual guidance from global action-subject localization to local detail enhancement is more effective for action-related region modeling. Moreover, the proposed skeleton-dominant cross-modal gated fusion module outperforms commonly used fusion strategies, including feature concatenation, matrix multiplication, and cross-attention. This comparison indicates that using the skeleton modality to adaptively regulate RGB contribution is more suitable for multimodal action recognition than treating both modalities equally. Category-level analysis shows that the RGB complementary branch provides more significant improvements for fine-grained and easily confused actions, further demonstrating the effectiveness of exploiting modality complementarity. In particular, actions involving subtle hand motion, object interaction, or visually similar body configurations benefit more from the proposed framework, because RGB information supplements details that skeleton data alone cannot adequately represent. **Conclusion** The proposed method enhances skeleton modeling by jointly capturing local motion patterns and global coordination relationships through dense-sparse skeleton joint representation, improves the focus of the RGB modality on action-relevant regions via coarse-to-fine guided ROI localization, and achieves effective multimodal integration through a skeleton-dominant cross-modal gated fusion module. By integrating these three components into a unified framework, the model is able to make fuller use of the complementary advantages of skeleton sequences and RGB videos. Experimental results demonstrate that the proposed framework not only achieves superior performance on standard indoor benchmarks but also maintains strong robustness and generalization ability in complex UAV-view scenarios, providing an effective solution for multimodal human action recognition. Overall, the study shows that strengthening skeleton representation, progressively guiding visual attention, and designing a skeleton-dominant fusion mechanism are all important for improving multimodal action recognition performance in complex scenes.

Key words: human action recognition; multimodal fusion; graph convolutional network; cross-modal attention; ROI localization

论文引用格式：Du H T, Zhao F, Liu H Q and Tang Y. 2026. Dense-sparse skeleton joint representation-guided RGB image ROI localization for multimodal action recognition. Journal of Image and Graphics, XX: XXX-XXX(杜海涛, 赵凤, 刘汉强, 唐垚. 2026. 稠密-稀疏骨骼联合表征引导RGB图像ROI定位的多模态行为识别. 中国图象图形学报, XX: XXX-XXX)[DOI: 10. 11834/jig. 260177]

0 引 言

人体行为识别作为计算机视觉领域的核心研究方向，在智能监控(姜权晏等，2022)、人机交互(Gu等，2018)、辅助医疗(程勇等，2025)等场景中具有广泛应用价值。随着深度学习技术的快速发展，基于多模态数据的行为识别方法逐渐成为研究热点。其中，骨骼序列与RGB视频作为两种互补性较强的模

态，能够分别提供人体结构动态信息与丰富的外观细节（施海勇等，2023），如何提高基于骨骼和RGB的多模态行为识别方法识别性能成为当前亟待解决的关键问题。

在基于骨骼模态的行为识别方面，现有研究已经取得了显著进展。早期方法主要依赖人工设计的几何特征。随着深度学习技术的兴起，循环神经网络被广泛应用于骨骼序列的时序建模，显著增强了对动作动态特性的表征能力。近年来，图卷积网络的发展进一步推动了骨骼动作识别性能提升。Yan等（2018）首次提出时空图卷积网络，将骨骼序列表示为时空图结构，通过图卷积与时间卷积联合建模人体动作；Chen等（2021）通过自适应图学习增强了骨骼拓扑表达能力；Liu等（2020）引入注意力机制以突出关键关节的重要性；Chen等（2021）则利用多尺度图结构提升了模型对不同粒度动作特征的建模效果。总体而言，基于图卷积的骨骼行为识别方法已经成为人体行为识别的重要技术路线。

尽管基于图卷积的骨骼行为识别方法具备较强的结构建模能力，但其局限性同样明显。首先，骨骼数据本质上是对人体关节位置的稀疏表示（吴志泽等，2025），虽能刻画人体结构与运动轨迹，却缺乏纹理、外观以及交互对象等语义信息，因此在区分“阅读”与“书写”、“打字”与“书写”等具有相似骨骼模式的细粒度动作时容易产生混淆（Shahroudy等，2016）。其次，现有基于图卷积的骨骼行为识别方法大多仍建立在稠密骨骼拓扑之上，其信息传播主要依赖相邻关节之间的局部连接关系，导致模型更擅长建模局部运动模式，而与远距离关键关节之间的全局运动模式建模不足（Wu等，2025）。然而，许多动作的判别信息并不只体现在局部关节运动上，还体现在手—头、手—脚、双手之间等跨身体部位的全局协同关系中。因此，如何同时建模骨骼时空的拓扑局部特征和全局依赖关系，仍是基于图卷积的骨骼动作识别方法需要进一步解决的重要问题。

与骨骼模态相比，RGB视频能够提供更加丰富的视觉语义信息，因此也是人体行为识别的重要研究对象。早期方法通常采用手工设计的时空特征进行动作描述。随着深度学习的发展，基于卷积神经网络的视频行为识别方法逐渐成为主流。Girdhar等（2017）基于双流框架从采样的外观和运动帧中提取特征，并使用动作词词汇将特征聚合到单个视频级表示中进行分类；Diba等（2017）通过元素乘法对分割后的特征进行聚合；Feichtenhofer等（2017）在特征级别将外观残差特征与运动信息相乘；Pan等（2019）提出一种双流网络，分别对RGB帧与光流进行建模；Alomar等（2025）将每个视频分成三个片段，并使用双流网络对每个片段进行处理，然后，通过整合三个片段的分类分数进行融合以产生视频级预测。综上，RGB视频在动作语义表达和场景理解方面展现出较强优势。

然而，RGB模态在行为识别中同样面临挑战。一方面，RGB视频虽然包含丰富的外观和上下文信息，但也不可避免地引入大量背景、光照变化、视角变化以及无关区域干扰，模型若直接对整帧视频进行建模，容易将注意力分散到与动作无关的区域，从而削弱对关键动作部位的关注。尤其在室内或复杂背景场景中，动作判别往往依赖手部、上肢或人体与交互物体之间的局部细节，而这些区域在整帧视觉表征中容易被背景信息淹没。另一方面，现有RGB动作识别方法大多关注通用时空表征学习，但对于“如何更有针对性地定位动作相关区域”考虑不足，导致视觉模态的判别优势未能被充分发挥。因此，如何利用更强的结构先验对RGB特征进行有效约束和引导，使模型从粗到细地聚焦于动作关键区域，是提升RGB模态有效性的关键所在。

为充分发挥骨骼与RGB两种模态的互补优势，多模态融合方法逐渐成为研究重点（Sedaghati等，2025；AlShami等，2025）。已有研究表明，骨骼模态能够为视觉模态提供稳定的人体结构先验，因此部分工作尝试利用骨骼信息引导视觉特征学习。例如根据骨骼关节位置定位手部区域以捕获人与物体的交互信息（Wu等，2016），或进一步扩展至多个身体部位的区域建模（Garcia等，2018；Das等，2020）。此外，也有研究通过跨模态表示对齐或注意力机制增强骨骼与视频之间的信息交互（Wei等，2017）。Yu（2023）等提出MMNet，将RGB视频转换为围绕头部、双手和双脚等关键身体部位的时空ROI，并利用骨骼图卷积分支学习得到的关节权重对不同ROI区域进行加权，使RGB分支能够更加关注对骨骼模态具有补充作用的外观信息。Chiang等（2025）则进一步针对固定ROI采样容易引入无关区域的问题，依据关节运动幅度动态选择区域，从而提升RGB区域采样的针对性，这些方法在一定程度上提升了多模

态行为识别性能，但其研究重点主要集中在RGB输入区域的构造或采样策略上，缺少骨骼特征与RGB特征之间更深层的跨模态语义交互。

从融合层次来看，现有多模态方法大致可分为数据级融合、特征级融合和决策级融合（Pan等，2019；Zhao等，2025）。其中，特征级融合通过在中间语义空间中整合不同模态特征，实现更深层次的信息交互（Baradel等，2018）；决策级融合则对不同模态的预测结果进行加权聚合（Luo等，2018）。然而，在实际融合过程中，许多方法仍采用简单特征拼接或模态对等融合策略（Shahroudy等，2018），未能充分考虑骨骼模态在人体结构表达上的稳定性优势，也缺乏根据动作语义自适应调节RGB模态贡献程度的机制。因此，如何基于骨骼模态设计跨模态引导策略，使RGB模态逐步聚焦于动作关键区域，并进一步构建由骨骼主导的自适应融合方式，仍然是多模态行为识别研究中的关键问题。

针对上述问题，本文提出一种面向骨骼与RGB多模态行为识别的统一框架。本文的主要贡献如下：

1）提出稠密-稀疏骨骼联合表征框架。该框架在保留稠密骨骼自然拓扑以描述相邻关节局部运动关系的基础上，引入由关键关节构成的稀疏骨骼图以缩短远距离关节间的信息传播路径，从而增强骨骼模态对局部运动模式与全局协同关系的联合建模能力。

2）提出基于跨模态注意力的粗细粒度骨骼引导RGB ROI定位策略。该策略首先利用稀疏骨骼特征引导RGB模态完成动作主体区域的粗粒度定位，再利用稠密骨骼特征对局部判别区域进行细粒度增强，使RGB模态能够在复杂背景下由整体到局部逐步聚焦于动作相关视觉区域。

3）提出骨骼主导的跨模态门控融合模块。该模块以骨骼特征生成门控权重，对RGB特征进行通道级自适应调节，使RGB模态作为互补信息参与融合，从而在保持骨骼结构表达稳定性的同时，更充分挖掘两种模态之间的互补关系。

1 方　法

1.1 总体框架

为了充分利用骨骼结构信息与RGB视觉信息在人体行为识别中的互补优势，本文提出一种基于稠密-稀疏骨骼联合表征引导RGB图像ROI定位的多模态行为识别网络。该框架以骨骼模态作为结构先验，通过骨骼特征逐步引导RGB模态定位动作相关区域，并在高层语义空间中实现骨骼主导的跨模态融合，从而获得更加鲁棒且判别性更强的动作表示。

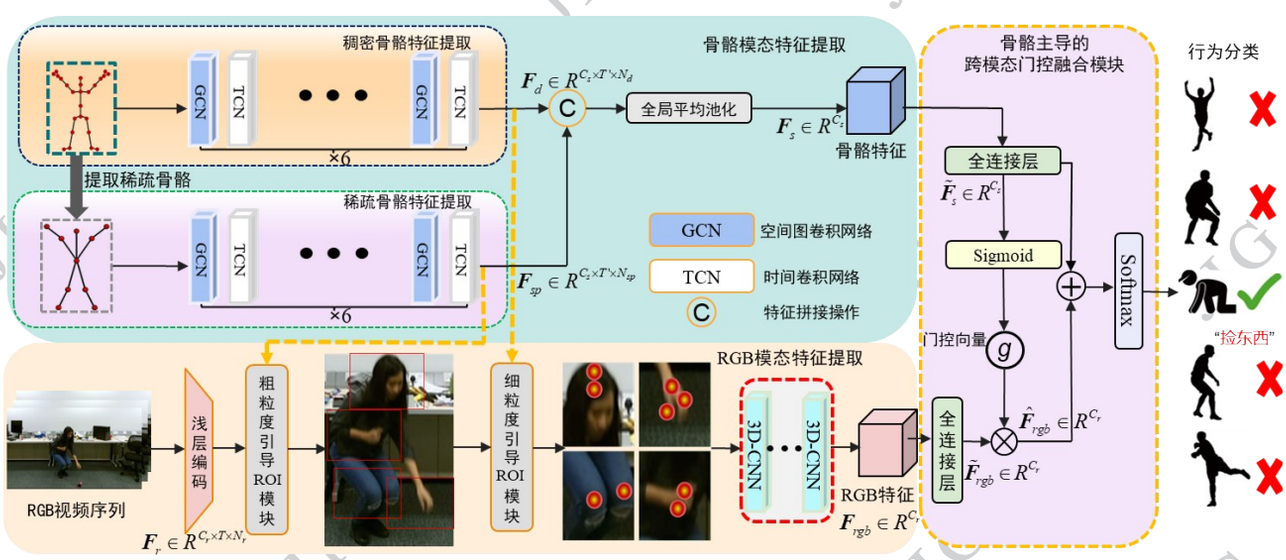

图1　方法总体框架图

Fig. 1　Overall framework of the proposed method

图1展示了本文所提多模态行为识别框架的整体流程。首先，根据原始骨骼序列构造稀疏骨骼序列，构造稠密骨骼特征提取分支与稀疏骨骼特征提取分支，以提取相邻关节的局部运动模式与远距离关键关节的全局运动模式；同时，RGB视频经过浅层视觉编码后，在骨骼特征引导下依次进行粗粒度和细粒度的动作相关ROI区域定位，得到抑制背景干扰的视觉表征；最后，骨骼特征与增强后的RGB特征通过骨骼主导的跨模态门控融合模块进行整合，并输入分类头完成动作类别预测。

通过上述三个模块的协同作用，所提出的框架能够在骨骼结构先验的引导下实现跨模态信息的高效交互，使模型既能够捕获人体动作的结构动态，又能够利用视觉信息提供的细粒度语义，从而实现更加准确和鲁棒的人体行为识别。

1.2 稠密-稀疏骨骼联合表征框架

骨骼序列能够直接刻画人体关节之间的空间拓扑关系及其随时间变化的动态信息，因此在人体行为识别任务中具有重要作用。现有基于图卷积的骨骼行为识别方法通常采用稠密骨骼图对人体结构进行建模，即依据人体物理连接关系构建邻接矩阵，并结合图卷积与时间建模对动作进行时空特征提取。骨骼序列包含T帧，每一帧由N个关节构成。通过人体骨骼的自然连接关系可以构建骨骼图结构，基于该图结构，空间关系由图卷积进行建模，而时间维度的动态变化则通过时间卷积网络进行建模，从而实现对人体动作时空特征的联合学习。其过程可表示为：

$$\boldsymbol{H}_s^{(l)}=\sigma(\tilde{\boldsymbol{D}}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}\boldsymbol{H}^{(l)}\boldsymbol{W}_s^{(l)}) \tag{1}$$

$$\boldsymbol{H}^{(l+1)}=\mathrm{TCN}(\boldsymbol{H}_s^{(l)};\boldsymbol{W}_t^{(l)}) \tag{2}$$

式中，$\boldsymbol{A}$为加入自连接的邻接矩阵，$\boldsymbol{A}=\boldsymbol{A}_{phy}+\boldsymbol{I}$，$\boldsymbol{A}_{phy}$为由人体自然物理连接关系构建的邻接矩阵，$\boldsymbol{I}$为$N\times N$单位矩阵；$\boldsymbol{D}$为对应的度矩阵，$\boldsymbol{H}^{(l)}$与$\boldsymbol{H}^{(l+1)}$分别表示第$l$层与第$l$+1层的节点特征表示，$\boldsymbol{W}_s^{(l)}$和$\boldsymbol{W}_t^{(l)}$分别为空间图卷积与时间卷积的可学习参数，$\sigma(\cdot)$表示非线性激活函数，TCN(·)表示沿时间维进行的一维卷积操作。该操作本质上通过邻域聚合实现节点特征的逐层传播。为了兼顾模型表达能力与计算复杂度，本文在稠密骨骼分支与稀疏骨骼分支中均采用6层空间图卷积+时间卷积堆叠结构进行时空特征提取。

然而，仅依赖稠密骨骼图进行动作建模仍存在明显局限。首先，图卷积本质上是一种局部邻域聚合机制，节点特征主要在物理相邻关节之间传播。当网络层数增加时，远距离关节之间的信息交互需要经过多跳传递，传播路径较长，容易导致关键信息在传递过程中逐渐衰减。因此，传统图卷积方法通常更擅长建模局部关节运动模式，而难以捕捉远距离关节的长程依赖。然而，在许多复杂动作中，判别性信息往往依赖于远距离关节之间的全局特征，这种长程依赖在传统稠密骨骼图中难以被有效建模。

为缓解上述问题，本文在骨骼建模过程中进一步引入稀疏骨骼图结构，作为对稠密骨骼图的补充。与保留全部关节节点的稠密骨骼图不同，稀疏骨骼图仅保留对动作判别更为关键的少量关节节点，包括头部、左右手肘、左右手、左右膝、左右脚以及一个躯干中心点，共计10个关键关节。该设计具有两方面优势：一方面，头部、手部及足部等末端关节通常具有更大的运动幅度，并与交互行为密切相关；另一方面，通过减少节点数量，骨骼拓扑结构更加紧凑，远距离身体部位之间的连接路径被显著缩短，从而有助于提升长程信息传播效率。在稀疏骨骼分支中，首先，基于关键关节集合重新构建邻接矩阵，形成新的骨骼图结构；图2展示了稠密骨架和所对应的稀疏骨架构造图，其中绿色的骨骼点即从稠密骨骼中选中的构造为稀疏骨骼的骨骼点。

随后，采用与稠密骨骼分支一致的时空图卷积

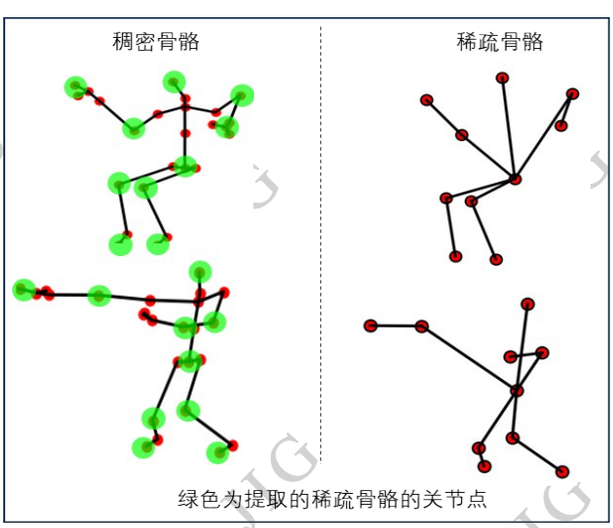

图2 稠密和稀疏骨骼图构造示意图

Fig. 2 Schematic diagram of dense and sparse skeletal graph construction.

和时间卷积组合对其进行时空特征建模，以提取关键关节之间的全局协同关系与整体动作结构。相比之下，稠密骨骼分支保留完整的人体自然拓扑，更适合描述相邻关节之间的局部连接关系、关节微运动模式以及连续姿态变化；而稀疏骨骼分支通过压缩拓扑路径，强化跨身体部位的长程依赖建模，更关注动作整体轮廓及远距离关节运动变化。两种分支在空间建模上形成显著互补，而非简单结合，从而提升骨骼模态的整体时空表达能力。最终，将两个分支提取的特征进行融合，得到骨骼模态的时空表示：

$$\boldsymbol{F}_s = \mathrm{Concat}(\boldsymbol{F}_d, \boldsymbol{F}_{sp}) \tag{3}$$

式中，$\boldsymbol{F}_d \in R^{C_s \times T' \times N_d}$ 表示稠密骨骼分支特征，$\boldsymbol{F}_{sp} \in R^{C_s \times T' \times N_{sp}}$ 表示稀疏骨骼分支特征，其中 C_s 表示通道数，N_d 和 N_{sp} 分别表示稠密骨骼图和稀疏骨骼图中的关节节点数。T' 为时空图卷积输出的时间维度。Concat(·)表示特征拼接操作。通过引入稀疏骨骼结构，模型能够在保持完整人体拓扑信息的同时强化全局动作关系建模能力，从而弥补基于图卷积的骨骼行为识别在长程依赖建模方面的不足。

1.3 基于跨模态注意力的粗细粒度骨骼引导RGB ROI定位策略

在多模态人体行为识别任务中，RGB视频能够提供丰富的外观细节、场景上下文以及交互目标信息，但同时也容易受到复杂背景、光照变化和无关区域的干扰。若直接对原始视频进行时空特征提取，模型往往难以持续聚焦于与动作判别最相关的视觉区域，从而削弱RGB模态的有效性。相比之下，骨骼序列能够稳定表征人体姿态结构及其动态变化，具有明确的物理语义和较强的结构先验。因此，本文利用骨骼模态对RGB模态进行引导，提出一种基于跨模态注意力的粗细粒度骨骼引导RGB ROI定位策略，通过由粗到细的两阶段定位过程，逐步强化动作相关区域的响应并抑制背景噪声。需要说明的是，本文并不对原始RGB图像进行显式裁剪，而是在特征空间内对动作相关区域进行加权增强，从而实现更稳定、更灵活的ROI建模。

设输入RGB视频序列 $\boldsymbol{I}_t \in R^{H \times W \times 3}$。经浅层视觉编码后，可得到RGB特征表示 $\boldsymbol{F}_r \in R^{C_r \times T \times N_r}$，$T$ 表示视频帧数，N_r 表示每帧对应的空间位置即视觉token数，C_r 表示RGB特征维度。为建立骨骼结构与RGB空间位置之间的对应关系，本文借助OpenPose提取图像平面中的二维关节坐标，仅用于辅助构建骨骼到视觉特征的空间对齐关系。模型的骨骼模态输入仍采用数据集原始提供的骨骼序列。为了便于描述粗细粒度引导ROI定位的注意力计算过程，下面以粗粒度ROI定位阶段为例，给出骨骼引导RGB特征增强的跨模态注意力计算形式，用其获得粗粒度增强后的RGB特征 $\boldsymbol{F}_{cg}$：

$$\boldsymbol{F}_{cg} = \mathrm{Attn}(\boldsymbol{Q}_s, \boldsymbol{K}_r, \boldsymbol{V}_r) = \mathrm{Soft\,max}\left(\frac{\boldsymbol{Q}_s \boldsymbol{K}_r^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}_r \tag{4}$$

$$\boldsymbol{Q}_s = \boldsymbol{F}_{sp}\boldsymbol{W}_Q, \boldsymbol{K}_r = \boldsymbol{F}_r\boldsymbol{W}_K, \boldsymbol{V}_r = \boldsymbol{F}_r\boldsymbol{W}_V \tag{5}$$

式中，$\boldsymbol{F}_{sp}$ 为稀疏骨骼特征，$\boldsymbol{Q}_s$ 由稀疏骨骼特征投影得到，$\boldsymbol{K}_r$ 和 $\boldsymbol{V}_r$ 由RGB浅层编码特征投影得到。d 表示特征维度，$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$ 共享线性投影矩阵，目的是通过可学习线性映射将骨骼特征和RGB特征投影到同一特征空间。其计算结构图如图3所示。

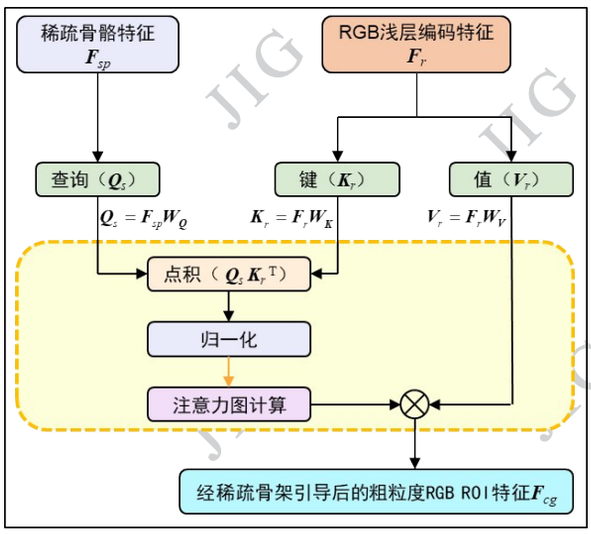

图3 粗粒度引导ROI模块结构图

Fig. 3 Architecture of the coarse-grained skeleton-guided ROI module

在上述基础上，本文还设计了细粒度ROI定位，其在计算方式上与粗粒度ROI定位保持一致，即采用公式(4)和(5)，不同之处在于其输入骨骼特征是稠密骨骼特征 $\boldsymbol{F}_d$，其输入视觉特征是通过粗粒度增强后的RGB特征 $\boldsymbol{F}_{cg}$，最终获得细粒度ROI特征 $\boldsymbol{F}_{fg}$。细粒度引导ROI并非重新在全局视觉空间中搜索动作区域，而是在已经完成粗粒度ROI定位的基础上，对动作判别相关的细节区域进行进一步强化。例如，手部、肘部、膝部以及人体与交互目标之间的局部区域，都能够在这一阶段获得更高的响应权重。

通过上述由粗到细粒度的定位过程，本文实现

了骨骼模态对RGB模态的ROI区域定位：粗粒度ROI定位利用稀疏骨骼特征完成主体动作区域的粗定位，细粒度ROI定位利用稠密骨骼特征对关键局部区域进行精细化增强。最终得到的细粒度ROI特征$\boldsymbol{F}_{fg}$能够更加突出与人体动作判别相关的视觉信息，同时有效抑制复杂背景带来的噪声干扰。最后，将增强后的RGB特征送入3D-CNN主干网络进行时空特征提取，从而获得更具判别性的视觉动作表示。通过这种特征层面的ROI定位方式，模型无需对原始图像进行显式裁剪，即可在骨骼结构先验的约束下实现对动作关键区域的稳定建模。

1.4 骨骼主导的跨模态门控融合模块

在获得骨骼模态特征与RGB模态特征后，需要对两种模态信息进行有效融合以完成最终动作识别。然而，骨骼数据与RGB视频在信息属性上存在显著差异。骨骼序列主要刻画人体结构及其运动轨迹，具有较强的几何稳定性；而RGB视频虽然包含丰富的外观细节和交互信息，但同时也更容易受到背景噪声和环境变化的影响。若采用简单的特征拼接或固定加权融合策略，往往难以充分发挥两种模态之间的互补优势。为此，本文提出一种骨骼主导的跨模态门控融合模块。该模块以骨骼模态作为融合主导，通过骨骼特征生成门控权重，对RGB特征进行自适应调节，从而实现更加稳健的跨模态信息整合。

设由骨骼双分支得到的骨骼时空特征为$\boldsymbol{F}_s \in R^{C_s}$，由骨骼引导ROI通过3D-CNN网络得到的RGB动作特征表示为$\boldsymbol{F}_{rgb} \in R^{C_r}$，其中$C_r$表示特征维度。为了便于跨模态交互，首先将两种模态特征映射到统一的语义空间：

$$\tilde{\boldsymbol{F}}_s = \boldsymbol{W}_s \boldsymbol{F}_s, \tilde{\boldsymbol{F}}_{rgb} = \boldsymbol{W}_r \boldsymbol{F}_{rgb} \tag{6}$$

式中，$\boldsymbol{W}_s$和$\boldsymbol{W}_r$为可学习线性映射参数。与传统融合方式不同，本文并不直接对两种模态进行对等融合，而是将RGB特征视为对骨骼特征的补充信息。具体而言，首先根据骨骼特征生成门控权重：

$$\boldsymbol{g} = \delta(\boldsymbol{W}_g \tilde{\boldsymbol{F}}_s) \tag{7}$$

式中，W_g为可学习门控参数，$\boldsymbol{g} \in R^C$是由骨骼特征生成的通道门控权重，用于表示RGB特征在对应语义通道上的贡献程度。$\delta(\cdot)$表示Sigmoid函数，用于将各通道权重限制在[0,1]范围内。在此基础上，对RGB特征进行门控加权：

$$\hat{\boldsymbol{F}}_{rgb} = \boldsymbol{g} \odot \tilde{\boldsymbol{F}}_{rgb} \tag{8}$$

式中，$\odot$表示逐元素乘法。该操作根据骨骼语义对RGB信息进行筛选，保留与当前动作结构相关的视觉信息，而抑制无关或噪声信息。最终，融合特征表示为

$$\boldsymbol{F}_{fus} = \tilde{\boldsymbol{F}}_s + \hat{\boldsymbol{F}}_{rgb} \tag{9}$$

式中，骨骼特征作为融合结果的主体部分被直接保留，而RGB特征则以门控加权后的补充信息形式参与融合。该设计从结构上保证了骨骼模态在融合过程中的主导地位，同时允许RGB模态在骨骼表达不足的语义维度上提供必要的补充信息。通过此门控融合模块，模型能够根据骨骼模态所反映的人体姿态与运动模式，自适应地调节RGB模态的贡献程度，从而在不同动作场景下实现更加合理的跨模态信息协同。相比简单拼接或平均加权融合，该策略能够更有效地利用骨骼结构先验，提升模型在复杂背景环境下的人体行为识别性能。最终，融合特征$\boldsymbol{F}_{fus}$被输入分类层进行动作类别预测。

1.5 损失函数

为实现骨骼分支、RGB分支与融合分支的联合优化，本文采用端到端的多分支监督训练策略。具体而言，稠密骨骼分支与稀疏骨骼分支首先在特征层进行融合，形成骨骼模态表示；随后，骨骼模态特征与经骨骼引导ROI增强后的RGB特征通过跨模态门控融合模块进行整合，并由融合分类头输出最终预测结果。为了增强各分支的判别能力并提高训练稳定性，本文除对融合分支施加主监督外，还分别在骨骼分支输出和RGB分支输出后引入辅助监督。模型总损失函数定义为：

$$L = L_{fus} + L_{sk} + L_{rgb} \tag{10}$$

式中，L_{fus}表示融合特征的主分类损失，L_{sk}和L_{rgb}分别表示骨骼分支和RGB分支的辅助分类损失。所有损失均采用交叉熵形式。融合分支损失用于直接约束最终跨模态表示的类别判别能力，是模型优化的主要目标；骨骼分支辅助损失用于保持骨骼模态对人体结构与动作动态的稳定建模能力；RGB分支辅助损失则用于增强视觉模态对动作相关区域的判别学习。通过这种“主监督+辅助监督”的联合优化方式，模型能够在保证融合性能的同时，提高各模态分支的特征表达质量，从而进一步提升整体识别性能与训练稳定性。

2 实 验

2.1 实验设置

本文在NTU-RGB+D 60(Shahroudy等，2016)、NTU-RGB+D 120(Liu等，2020)和UAV-Human(Li等，2021)三个数据集上验证所提方法的有效性。三个数据集可提供RGB视频和骨骼序列。其中，NTU-RGB+D 60采用跨主体(X-Sub)和跨视角(X-View)两种评估基准；NTU-RGB+D 120采用跨主体(X-Sub)和跨设置(X-Set)两种评估基准。UAV-Human采集于低空无人机场景，有CSv1和CSv2基准测试，具有目标尺度变化大、视角变化剧烈、遮挡频繁和背景复杂等特点，可用于进一步检验模型在开放环境中的鲁棒性。

在优化策略上，本文采用SGD优化器，总训练轮数设为80个epoch，初始学习率设为0.1，并在训练后期按计划衰减。所有结果均以Top-1准确率作为评价指标。

2.2 与现有方法的比较

为了全面验证本文方法的有效性，本文将其与多种具有代表性的人体行为识别方法进行了比较。对比方法按照输入模态可分为仅使用骨骼模态的方法以及同时融合骨骼与RGB信息的多模态方法。表1和表2分别给出了本文方法在NTU-RGB+D 60和NTU-RGB+D120数据集上的实验结果。

由表1可见，在NTU-RGB+D 60数据集上，仅依赖骨骼信息的先进方法已经能够取得较高识别精度，例如HD-GCN(Hierarchically Decomposed Graph Convolutional Networks)和BlockGCN(Block Graph Convolutional Networks)在X-Sub与X-View基准下均表现出较高性能，说明骨骼模态在人体动作建模中具有较强的结构表达能力。然而，多模态方法整体上仍优于单模态方法，这表明RGB视频所提供的外观细节与环境上下文信息能够对骨骼结构信息形成有效补充。本文方法在该数据集上分别取得94.7%的X-Sub准确率和98.3%的X-View准确率，相比MMNet(Model-based Multimodal Network)的94.2%和97.8%均取得进一步提升，也优于VPN(Video-Pose Embedding Networks)、MS-ROI和TransMODAL等多模态方法。该结果说明，本文提出的基于跨模态注意力的骨骼引导RGB ROI定位策略能够更有效地抑制视觉背景干扰，并引导RGB分支聚焦于动作相关区域，从而提升视觉模态对动作语义的建模能力；与此同时，骨骼主导的跨模态门控融合模块能够更合理地利用两种模态之间的互补信息，进一步增强最终分类性能。

表1 本文方法与其他方法在NTU-RGB+D 60数据集上的精度对比

Table 1 Comparison of the accuracy of the proposed method with other methods on the NTU-RGB+D 60 dataset.

方法	S	R	X-Sub	X-View
ST-GCN(Yan等，2018)	√	-	81.57	88.33
2AS-GCN(Shi等，2019)	√	-	86.89	94.25
MS-G3D(Liu等，2020)	√	-	91.55	96.24
CTR-GCN(Chen等，2021)	√	-	92.20	96.10
InfoGCN(Chi等，2022)	√	-	89.80	91.20
HD-GCN(Lee等，2023)	√	-	93.40	97.20
BlockGCN(Zhou等，2024)	√	-	93.10	97.00
VPN(Das等，2020)	√	√	93.50	96.20
TSN(Bruce等，2021)	√	√	92.50	97.40
MMNet(Yu等，2023)	√	√	94.20	97.80
MS-ROI(Ming等，2025)	√	√	94.55	98.21
TransMODAL(Majid等，2025)	√	√	92.10	93.23
本文	√	√	**94.70**	**98.30**

注：加粗字体为最优值。S为骨骼模态，R为RGB模态。

在更具挑战性的NTU-RGB+D 120数据集上，本文方法同样取得了具有竞争力的结果。由于该数据集包含更多动作类别且动作间相似性更强，从表2可以看出，各方法的识别精度相较NTU-RGB+D60普遍有所下降，但本文方法仍在X-Sub与X-Set基准下分别达到92.82%和93.91%，优于MMNet、MS-ROI、TransMODAL等多模态方法。值得注意的是，ProtoGCN(Prototypical Graph Convolutional Networks)在该数据集上的骨骼识别性能已经较高，但本文方法在融合RGB信息后仍实现了进一步提升，这说明当动作类别数量增加、类别边界更加模糊时，仅依赖骨骼拓扑关系仍存在一定局限，而本文所引入的骨骼引导视觉补充信息能够为细粒度动作判别提供更充分依据。

表3给出了本文方法在UAV-Human数据集上

表2　本文方法与其他方法在NTU-RGB+D 120数据集上的精度对比

Table 2　Comparison of the accuracy of the proposed method with other methods on the NTU-RGB+D120 dataset.

方法	S	R	X-Sub	X-Set
ST-GCN(Yan等,2018)	√	–	70.74	73.20
2AS-GCN(Shi等,2019)	√	–	82.50	84.20
MS-G3D(Liu等,2020)	√	–	86.90	88.40
CTR-GCN(Chen等,2021)	√	–	88.90	90.60
InfoGCN(Chi等,2022)	√	–	90.10	91.60
HD-GCN(Lee等,2023)	√	–	90.30	91.50
BlockGCN(Zhou等,2024)	√	–	90.90	92.20
ProtoGCN(Liu等,2025)	√	–	92.00	93.80
VPN(Das等,2020)	√	√	86.30	87.80
TSN(Bruce等,2021)	√	√	87.00	89.10
MMNet(Yu等,2023)	√	√	91.90	93.40
MS-ROI(Ming等,2025)	√	√	92.55	93.51
TransMODAL(Majid等,2025)	√	√	90.22	91.75
本文	√	√	**92.82**	**93.91**

注:加粗字体为最优值。S为骨骼模态,R为RGB模态。

的实验结果。为保证对比实验的公平性,HD-GCN、BlockGCN和ProtoGCN四种方法的结果均采用其开源代码运行获得,实验数据划分、训练轮数、优化器设置均在与本文方法一致。其他方法结果来自于文献公开报导。从表中可以看出,在低空无人机视角下,本文方法在CSv1和CSv2基准下分别取得53.60%和76.90%的识别精度,在所列对比方法中均达到最优。与NTU-RGB+D 60和NTU-RGB+D 120数据集相比,UAV-Human的数据采集条件更加复杂:一方面,低空俯视视角导致人体目标尺度变化更大、局部细节更难稳定保留;另一方面,复杂背景、遮挡以及拍摄平台运动会进一步干扰RGB分支的判别能力,同时也会降低骨骼估计的稳定性。在这种条件下,本文方法仍能保持较优性能,这表明本文方法不仅能够在标准室内数据集上获得稳定提升,而且在无人机低空拍摄的背景复杂的场景下仍具有较好的泛化能力。

图4和图5分别展示了本文方法在NTU-RGB+D 60和UAV-Human数据集上仅使用骨骼分支与引入RGB补充分支后的分类精度对比。从图中可以看出,显著的性能增益主要集中在不明显且易混淆的动作类别上。具体而言,在NTU-RGB+D 60数据集中,“喝水”“吃饭”“刷牙”“穿/脱夹克”“打电话/玩手机”等动作提升较为明显;在UAV-Human数据集中,“喝水”“吃零食”“阅读”“书写”“穿/脱外套”等动作同样获得了显著增益。

这些动作仅依赖骨骼序列通常难以有效区分,而RGB模态能够补充交互物体外观、局部纹理细节以及场景线索等骨骼难以表达的判别信息,从而有效缓解相似动作之间的混淆。尤其在UAV-Human低空复杂场景中,由于存在尺度变化大、视角变化剧烈、遮挡频繁和背景复杂等问题,RGB模态带来的提升更为明显,进一步说明本文方法能够充分利用骨骼结构先验与视觉外观信息之间的互补性,提升复杂场景下细粒度行为识别的准确性。

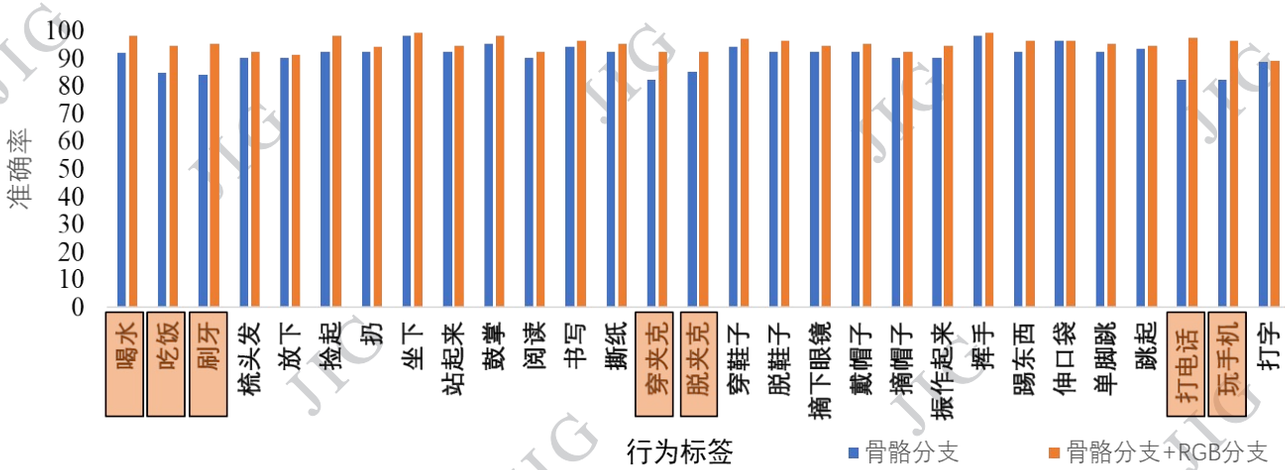

图4　在NTU-RGB+D 60上每个动作的识别准确率

Fig. 4　Per-class recognition accuracy on the NTU-RGB+D 60 datasets

表3　本文方法与其他方法在UAV-Human数据集上精度对比

Table 3　Comparison of the accuracy of the proposed method with other methods on the UAV-Human dataset.

方法	S	R	CSv1	CSv2
ST-GCN(Yan等,2018)	√	–	30.25	56.14
2s-AGCN(Shi等,2019)	√	–	34.84	66.68
MS-G3D(Liu等,2020)	√	–	37.91	70.10
CTR-GCN(Chen等,2021)	√	–	37.11	69.23
HD-GCN(Lee等,2023)	√	–	45.43	72.86
BlockGCN(Zhou等,2024)	√	–	43.21	45.60
ProtoGCN(Liu等,2025)	√	–	31.00	60.80
VPN(Das等,2020)	√	√	49.32	75.65
TSN(Bruce等,2021)	√	√	51.33	72.91
MMNet(Yu等,2023)	√	√	52.66	75.00
MS-ROI(Ming等,2025)	√	√	51.55	76.51
TransMODAL(Majid等,2025)	√	√	53.22	74.75
本文	√	√	**53.60**	**76.90**

注:加粗字体为最优值。S为骨骼模态,R为RGB模态。

2.3　消融实验

为了进一步分析本文方法中各关键设计对最终性能的影响,本文在NTU-RGB+D 60和NTU-RGB+D 120数据集上围绕稠密-稀疏骨骼联合表征框架、基于跨模态注意力的粗细粒度骨骼引导RGB ROI定位策略以及骨骼主导的跨模态门控融合模块进行了消融实验,每组消融实验除被消融模块外,其余设置保持与本文网络一致。

表4的实验结果可以观察到,相比仅使用稠密骨骼分支,引入稀疏骨骼分支后本文方法的完整模型在各评估基准下均获得明显提升。例如,在NTU-RGB+D 60上,模型由基线结构提升至94.70%和98.30%;在NTU-RGB+D 120上,相应性能也由79.50%和81.20%提升至92.82%和93.91%。这一现象说明,传统稠密骨骼图虽然能够保留完整人体物理连接关系,但其图卷积传播路径较长,难以高效建模远距离关节之间的依赖关系;相比之下,稀疏骨骼图通过保留关键关节并压缩拓扑路径,能够显著强化不同身体部位之间的全局交互能力,从而弥补稠密骨骼图在长程依赖建模上的不足。因此,将稠密骨骼分支与稀疏骨骼分支结合,有助于同时捕获局部运动模式与全局依赖关系,进而提升骨骼模态的整体表达能力。

表5给出了基于跨模态注意力的粗细粒度骨骼

表4　稠密-稀疏骨骼联合表征框架的消融结果

Table 4　Ablation study of the dense–sparse skeleton joint representation framework

方法	S	R	NTU-RGB+D 60		NTU-RGB+D 120	
			X-Sub	X-View	X-Sub	X-Set
仅RGB分支		√	87.21	88.45	85.36	87.54
仅稠密骨骼分支	√		78.57	85.33	79.50	81.20
稠密+稀疏骨骼分支	√		90.33	92.35	89.10	91.32
本文方法	√	√	**94.70**	**98.30**	**92.82**	**93.91**

注:加粗字体为最优值。S为骨骼模态,R为RGB模态。

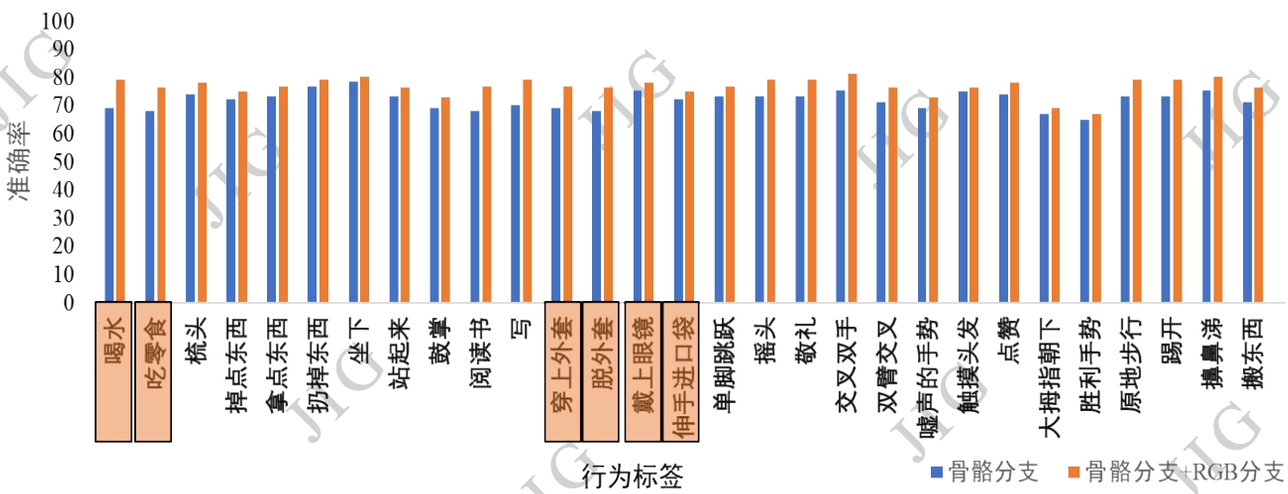

图5　在UAV-Human数据集上每个动作的识别准确率

Fig. 5　Per-class recognition accuracy on the UAV-Human datasets

引导RGB ROI定位策略的消融结果，从结果可以看出，当仅采用粗粒度引导时，模型在各基准下的识别性能分别为89.61%、92.80%、88.50%和89.20%；而在引入第二阶段细粒度引导后，对应结果提升至94.70%、98.30%、92.82%和93.91%。这一结果表明，粗细粒度引导ROI在视觉区域建模中是必要的。其原因在于，稀疏骨骼特征更侧重全局动作结构与关键身体部位之间的远距离长程依赖关系，因此适合作为粗粒度ROI定位的动作先验，用于快速定位主体区域；而稠密骨骼特征保留了完整的人体拓扑连接，更有利于描述局部关节运动与细粒度动作差异，因此能够通过细粒度ROI定位进一步细化动作显著区域。两者协同作用，使RGB分支获得了更具判别性的区域特征表达。

图6展示了粗粒度ROI区域与细粒度ROI热力图可视化结果。从图中可以观察到，在第一阶段粗粒度引导中，模型的响应主要集中于人体主体及其主要运动区域，大范围背景噪声已被明显抑制；而在第二阶段细粒度引导后，模型的响应区域进一步收缩并聚焦到与动作判别最密切相关的局部部位，使视觉关注区域更加紧凑、更加具有针对性。

表5　基于跨模态注意力的粗细粒度骨骼引导RGB ROI定位策略的消融结果

Table 5　Ablation study of the cross-modal attention-based skeleton-guided RGB ROI localization strategy

方法	NTU-RGB+D 60		NTU-RGB+D 120	
	X-Sub	X-View	X-Sub	X-Set
仅粗粒度ROI定位	89.61	92.80	88.50	89.20
粗粒度ROI定位+细粒度ROI定位	**94.70**	**98.30**	**92.82**	**93.91**

注：加粗字体为最优值。

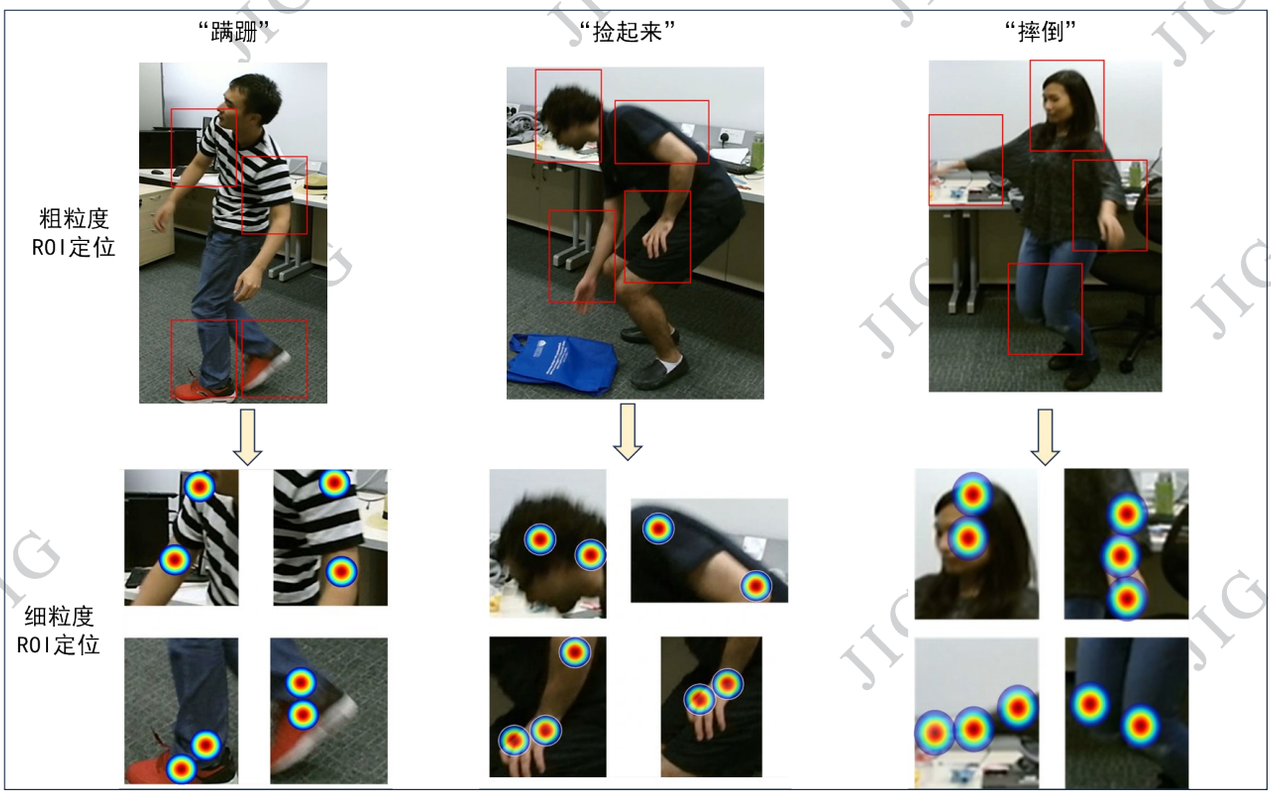

图6　粗粒度ROI区域与细粒度ROI热力图可视化结果

Fig. 6　Visualization results of coarse-grained ROI regions and fine-grained ROI heatmaps

表6对比了不同多模态融合策略的实验结果，包括特征拼接、矩阵相乘、交叉注意力(Zhang等，2025)以及本文提出的骨骼主导的跨模态门控融合模块。由结果可见，矩阵相乘在所有基准下的表现最弱，说明简单的乘性耦合难以稳定建模骨骼与RGB两种模态之间的复杂关系；特征拼接和交叉注意力相较之下能够取得更好的识别性能，但仍然不如本文提出的骨骼主导的跨模态门控融合模块。该

结果验证了本文融合设计的合理性。与对等融合方式不同，骨骼主导的跨模态门控融合模块以骨骼模态作为融合主导，通过骨骼特征生成条件门控权重，对RGB特征进行通道级自适应加权，这种方式能够在保留骨骼模态结构稳定性的同时，有效引入RGB模态中与当前动作结构相关的外观细节信息，因此适合骨骼与RGB双模态的行为识别任务。

表6　骨骼主导的跨模态门控融合模块与其他融合方式的准确率对比

Table 6　Comparison of recognition accuracy between the proposed skeleton-driven cross-modal gated fusion module and other fusion methods

方法	S	R	NTU-RGB+D 60		NTU-RGB+D 120	
			X-Sub	X-View	X-Sub	X-Set
仅RGB分支		√	87.21	88.45	85.36	87.54
仅骨骼分支	√		90.33	92.35	89.10	91.32
特征拼接	√	√	92.92	97.61	89.91	91.75
矩阵相乘融合	√	√	82.11	82.20	81.20	82.25
交叉注意力融合	√	√	93.82	94.60	90.10	92.17
骨骼主导的门控融合(本文方法)	√	√	**94.70**	**98.30**	**92.82**	**93.91**

注：加粗字体为最优值。S为骨骼模态，R为RGB模态

为进一步分析骨骼主导门控融合模块的可解释性，本文对测试集中不同动作类别的门控权重进行统计。对每个测试样本生成的门控权重在通道维度上计算门控权重均值来分析RGB模态贡献与门控权重大小之间的关系。其统计数据如图7所示同动作类别下的门控权重存在明显差异。对于喝水、吃饭、打电话/玩手机、刷牙、阅读、书写、穿/脱夹克等动作，仅依赖骨骼序列时容易因手部运动模式相近、交互物体缺失或局部细节不足而产生混淆，而RGB模态能够补充物体外观、手部局部区域和纹理信息，因此这些类别在引入RGB后获得更明显的精度提升，对应的门控权重均值也相对较高。相反，对于站起来、坐下、跳跃等人体整体姿态变化较明显的动作，骨骼模态本身已具有较强判别性，RGB模态带来的额外增益较小，对应门控权重均值也相对较低。

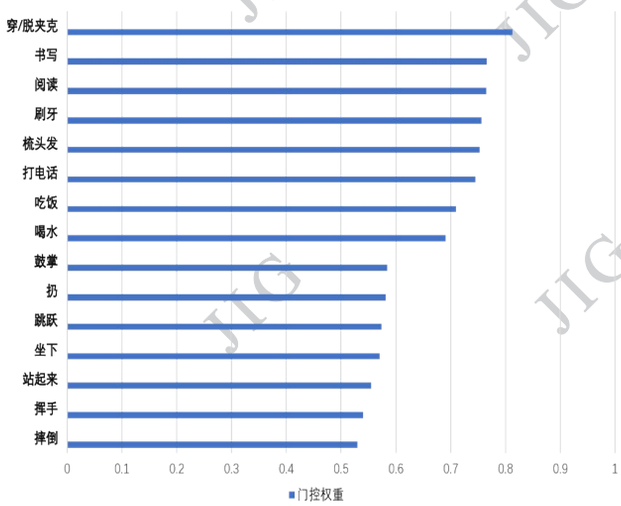

图7　不同动作类别的门控权重均值统计

Fig. 7　Mean statistics of gating weights for different action categories

3　结　论

本文针对RGB与骨骼多模态人体行为识别中存在的图卷积方法难以同时捕捉骨骼模态的局部与全局运动模式、RGB模态易受复杂背景干扰而难以聚焦动作相关区域以及跨模态互补特征融合不充分等问题，提出了一种基于稠密-稀疏骨骼联合表征引导RGB图像ROI定位的多模态行为识别方法。首先，在骨骼建模方面，本文通过构建稠密骨骼分支与稀疏骨骼分支的联合表征框架，在保留人体自然拓扑连接关系的基础上，进一步强化了图卷积网络同时捕捉相邻关节的局部运动模式与远距离关键关节的全局运动模式的能力。其次，在视觉建模方面，本文提出基于跨模态注意力的粗细粒度骨骼引导RGB ROI定位策略，以稀疏骨骼特征实现对动作主体区域的粗粒度定位，再以稠密骨骼特征对局部判

别区域进行细粒度增强，使RGB分支能够在复杂背景下由整体到局部逐步聚焦于动作相关区域。最后，在跨模态融合方面，设计了骨骼主导的跨模态门控融合模块，以骨骼模态作为融合主导，根据动作结构语义自适应调节RGB模态的信息贡献，从而更有效地挖掘两种模态之间的互补性。实验结果表明，所提方法在NTU-RGB+D 60、NTU-RGB+D120以及UAV-Human数据集上均取得了优秀的识别性能，尤其在低空无人机场景下仍表现出理想的泛化能力，验证了所提方法在复杂场景人体行为识别任务中的有效性。

需要指出的是，本文所设计的粗细粒度引导ROI定位机制在提升识别精度的同时，也带来了一定的计算开销。后续工作将进一步围绕高效的ROI定位机制展开研究，以提升模型的实际部署能力。此外，还将引入语言模态指导视觉特征提取，开展视觉-语言双驱动的多模态行为识别方法研究。

参考文献(References)

Alomar K, Aysel H I and Cai X. 2025. CNNs, RNNs and transformers in human action recognition: a survey and a hybrid model. Artificial Intelligence Review, 58(12): 387 [DOI: 10.1007/s10462-025-11388-3]

AlShami A K, Rabinowitz R, Lam K, Shleibik Y, Mersha M, Boult T and Kalita J. 2025. SMART-vision: survey of modern action recognition techniques in vision. Multimedia Tools and Applications, 84(27): 32705-32776 [DOI:10.1007/s11042-024-20484-5]

Baradel F, Wolf C, Mille J and Taylor G W. 2018. Glimpse clouds: human activity recognition from unstructured feature points//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 469-478 [DOI:10.1109/CVPR.2018.00056]

Bruce X B, Liu Y and Chan K C C. 2021. Multimodal fusion via teacher-student network for indoor action recognition//Proceedings of the AAAI Conference on Artificial Intelligence, 35(4): 3199-3207 [DOI:10.1609/aaai.v35i4.16430]

Chen Y X, Zhang Z Q, Yuan C F, Li B, Deng Y and Hu W M. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13359-13368 [DOI:10.1109/ICCV48922.2021.01311]

Chen Z, Li S, Yang B, Li Q and Liu H. 2021. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition//Proceedings of the AAAI Conference on Artificial Intelligence, 35(2): 1113-1122 [DOI:10.1609/aaai.v35i2.16197]

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 180-189 [DOI:10.1109/CVPR42600.2020.00026]

Cheng Y, Gao Y Y, Wang J, Yang L, Xu X L, Cheng Y and Zhang K H. 2025. Combining dilated convolution and multiscale fusion temporal action detection. Journal of Image and Graphics, 30(2): 406-420 (程勇, 高园元, 王军, 杨玲, 许小龙, 程遥, 张开华. 2025. 结合扩张卷积与多尺度融合的实时时空动作检测. 中国图象图形学报, 30(2): 406-420) [DOI:10.11834/jig.240098]

Chi H G, Ha M H, Chi S, Lee S W, Huang Q X and Ramani K. 2022. InfoGCN: representation learning for human skeleton-based action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE: 20154-20164 [DOI:10.1109/CVPR52688.2022.01955]

Chiang M L, Huang C M and Liao J F. 2025. Multimodal human action recognition base on sparse joints graph convolutional network and dynamic image ROI sampling. IEEE Sensors Journal, : 1-1 [DOI: 10.1109/JSEN.2025.3587646]

Das S, Sharma S, Dai R, Brémond F and Thonnat M. 2020. VPN: learning video-pose embedding for activities of daily living//Computer Vision - ECCV 2020. Cham: Springer: 72-90 [DOI:10.1007/978-3-030-58545-7_5]

Diba A, Sharma V and Van Gool L. 2017. Deep temporal linear encoding networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 1541-1550 [DOI:10.1109/CVPR.2017.168]

Feichtenhofer C, Pinz A and Wildes R P. 2017. Spatiotemporal multiplier networks for video action recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 7445-7454 [DOI: 10.1109/CVPR.2017.787]

Garcia N C, Morerio P and Murino V. 2018. Modality distillation with multiple stream networks for action recognition//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 106-121 [DOI:10.1007/978-3-030-01237-3_7]

Girdhar R, Ramanan D, Gupta A, Sivic J and Russell B. 2017. ActionVLAD: learning spatio-temporal aggregation for action classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 3165-3174 [DOI:10.1109/CVPR.2017.337]

Gu C, Sun C, Ross D A, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, Schmid C and Malik J. 2018. AVA: a video dataset of spatio-temporally localized atomic visual actions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 6047-6056 [DOI:10.1109/CVPR.2018.00633]

Jiang Q Y, Wu X J and Xu T Y. 2022. M2FA: multi-dimensional feature fusion attention mechanism for skeleton-based action recognition. Journal of Image and Graphics, 27(8): 2391-2403 (姜权晏, 吴小俊, 徐天阳. 2022. 用于骨骼行为识别的多维特征嵌合注意力机制. 中国图象图形学报, 27(8): 2391-2403) [DOI: 10.11834/jig.210091]

Lee J, Lee M, Lee D and Lee S. 2023. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 10444-10453

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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 16266-16275 [DOI: 10.1109/CVPR46437.2021.01600]

Liu H D, Liu Y F, Ren M, Wang H, Wang Y L and Sun Z A. 2025. Revealing key details to see differences: a novel prototypical perspective for skeleton-based action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 29248-29257

Liu J, Shahroudy A, Perez M, Wang G, Duan L Y and Kot A C. 2020. NTU-RGB+D 120: a large-scale benchmark for 3D human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(10): 2684-2701 [DOI: 10.1109/TPAMI.2019.2916873]

Liu J F, Wang X S, Wang C, Gao Y and Liu M Y. 2024. Temporal decoupling graph convolutional network for skeleton-based gesture recognition. IEEE Transactions on Multimedia, 26: 811-823 [DOI: 10.1109/TMM.2023.3271811]

Liu Z Y, Zhang H W, Chen Z H, Wang Z Y and Ouyang W L. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 143-152 [DOI: 10.1109/CVPR42600.2020.00022]

Luo Z, Hsieh J T, Jiang L, Niebles J C and Fei-Fei L. 2018. Graph distillation for action detection with privileged modalities//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 174-192 [DOI: 10.1007/978-3-030-01264-9_11]

Joudaki M, Imani M and Arabnia H R. 2025. TransMODAL: a dual-stream transformer with adaptive co-attention for efficient human action recognition. Electronics, 14(16): 3326 [DOI: 10.3390/electronics14163326]

Pan B, Sun J K, Lin W W, Wang L M and Lin W Y. 2019. Cross-stream selective networks for action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, CA, USA: IEEE: 454-460 [DOI: 10.1109/CVPRW.2019.00059]

Sedaghati N, Ardebili S and Ghaffari A. 2025. Application of human activity/action recognition: a review. Multimedia Tools and Applications, 84 (28): 33475-33504 [DOI: 10.1007/s11042-024-20576-2]

Shahroudy A, Liu J, Ng T T and Wang G. 2016. NTU RGB+D: a large scale dataset for 3D human activity analysis//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE: 1010-1019 [DOI: 10.1109/CVPR.2016.115]

Shahroudy A, Ng T T, Gong Y H and Wang G. 2018. Deep multimodal feature analysis for action recognition in RGB+D videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5): 1045-1058 [DOI: 10.1109/TPAMI.2017.2691321]

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, 钟卓锟. 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, Cheng J and Lu H Q. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, CA, USA: IEEE: 12026-12035 [DOI: 10.1109/CVPR.2019.01230]

Shi W Z, Li D, Wen Y and Yang W. 2023. Occlusion-aware graph neural networks for skeleton action recognition. IEEE Transactions on Industrial Informatics, 19(10): 10288-10298 [DOI: 10.1109/TII.2022.3229140]

Wei P, Zhao Y B, Zheng N N and Zhu S C. 2017. Modeling 4D human-object interactions for joint event segmentation, recognition, and object localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6): 1165-1179 [DOI: 10.1109/TPAMI.2016.2574712]

Wu D, Pigou L, Kindermans P J, Le N, Shao L, Dambre J and Odobez J M. 2016. Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8): 1583-1597 [DOI: 10.1109/TPAMI.2016.2537340]

Wu Z, Ding Y, Wan L, Li T and Nian F. 2025. Local and global self-attention enhanced graph convolutional network for skeleton-based action recognition. Pattern Recognition, 159: 111106 [DOI: 10.1016/j.patcog.2024.111106]

Wu Z Z, Chen X, Xu T, Nian F D, Wang X F and Li T. 2025. Dynamic multi-granularity graph convolutional networks for skeleton based action recognition. Journal of Image and Graphics, 30(8): 2822-2834 (吴志泽, 陈鑫, 徐童, 年福东, 王晓峰, 李腾. 2025. 基于动态多粒度图卷积网络的人体骨架行为识别. 中国图象图形学报, 30(8): 2822-2834) [DOI: 10.11834/jig.240352]

Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolu-

tional networks for skeleton-based action recognition //Proceedings of the AAAI Conference on Artificial Intelligence, 32(1): 7444-7452[DOI:10.1609/aaai.v32i1.12328]

Yu B X B, Liu Y, Zhang X, Zhong S H and Chan K C C. 2023. MMNet: a model-based multimodal network for human action recognition in RGB-D videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 3522-3538[DOI: 10.1109/TPAMI.2022.3177813]

Zhang Z Y, Cai W R, Liu Q J and Wang Y H. 2025. SkeletonX: data-efficient skeleton-based action recognition via cross-sample feature aggregation. IEEE Transactions on Multimedia, 27: 9646-9658[DOI:10.1109/TMM.2025.3618561]

Zhao X, Tang C, Hu H S, Wang W J, Qiao S and Tong A Y. 2025. Attention mechanism based multimodal feature fusion network for human action recognition. Journal of Visual Communication and Image Representation, 110: 104459[DOI: 10.1016/j.jvcir.2025.104459]

Zhou Y X, Yan X D, Cheng Z Q, Yan Y, Dai Q and Hua X S. 2024. BlockGCN: redefine topology awareness for skeleton-based action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 2049-2058[DOI:10.1109/CVPR52733.2024.00200]

作者简介

杜海涛,男,硕士研究生,主要研究方向为模式识别与计算机视觉。E-mail:1582235456@qq.com

赵凤,通讯作者,女,教授,主要研究方向为模式识别、图像处理与机器学习。E-mail:zhaofeng201@xupt.edu.cn

刘汉强,男,副教授,主要研究方向为模式识别与图像处理。E-mail:liuhq@snnu.edu.cn

唐垚,男,副研究员,主要研究方向为视频与信号处理。E-mail:tangyao77@163.com