Computer Science and Application 计算机科学与应用, 2026, 16(11), 216-227
Published Online November 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.1611338

Hans汉斯

基于门控姿态融合与时序记忆加权的时空动作检测方法

朱　强

北京印刷学院信息工程学院，北京

收稿日期：2026年10月10日；录用日期：2026年11月10日；发布日期：2026年11月17日

摘　要

时空动作检测需要同时定位视频中人物动作的空间位置并判断动作类别。Holistic Interaction Transformer Network (HIT)通过RGB分支、姿态分支和时序记忆信息进行多模态交互建模，但原始结构对姿态分支可靠性和不同时序记忆特征相关性的区分仍不充分。针对该问题，本文在HIT基础上提出一种轻量改进方法：一方面设计门控姿态融合模块，根据RGB ROI特征与姿态交互特征生成门控权重，自适应调节姿态特征对动作表征的贡献；另一方面设计时序距离感知记忆加权模块，根据记忆帧与当前帧的时间距离对memory feature进行衰减加权，以增强近邻上下文并降低远距离特征干扰。在JHMDB-21数据集上的实验结果表明，最终组合模型在Video mAP@0.5和Video mAP@0.2上分别达到88.02%和89.55%，相较自训练基线分别提升1.88和2.15个百分点。实验结果说明，所提轻量模块能够提升HIT在视频级时空动作检测中的稳定性。

关键词

时空动作检测，HIT，门控姿态融合，时序记忆加权，JHMDB-21

Spatio-Temporal Action Detection Based on Gated Pose Fusion and Temporal Memory Weighting

Qiang Zhu

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: October 10, 2026; accepted: November 10, 2026; published: November 17, 2026

文章引用: 朱强. 基于门控姿态融合与时序记忆加权的时空动作检测方法[J]. 计算机科学与应用, 2026, 16(11): 216-227. DOI: 10.12677/csa.2026.1611338

Abstract

Spatio-temporal action detection requires recognizing action categories while localizing actors in video frames. The Holistic Interaction Transformer Network (HIT) provides a multimodal interaction framework with RGB features, pose features and temporal memory information, but its original structure still lacks explicit reliability modeling for pose fusion and temporal memory utilization. To address this problem, this paper proposes a lightweight improvement based on HIT. First, a gated pose fusion module is introduced to generate adaptive weights from RGB ROI features and pose interaction features, thereby controlling the contribution of the pose branch. Second, a temporal distance-aware memory weighting module is designed to reweight memory features according to their temporal distance from the current frame, to enhance nearby context and suppress the interference from distant features. Experiments on JHMDB-21 show that the final model achieves 88.02% Video mAP@0.5 and 89.55% Video mAP@0.2, improving the self-trained baseline by 1.88 and 2.15 percentage points, respectively. The results indicate that the proposed lightweight modules improve the video-level stability of HIT for spatio-temporal action detection.

Keywords

Spatio-Temporal Action Detection, HIT, Gated Pose Fusion, Temporal Memory Weighting, JHMDB-21

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

视频动作理解是计算机视觉领域的重要研究方向之一，其目标是从连续视频帧中识别人类动作状态，并进一步理解人与场景、物体以及时序上下文之间的关系。与普通视频分类任务不同，时空动作检测不仅需要判断视频中发生了何种动作，还需要定位动作实例在空间位置和时间维度上的范围[1]，因此对模型的空间定位能力、时序建模能力和多模态信息融合能力均提出了更高要求。该任务在智能监控、人机交互、体育视频分析和行为理解等场景中具有较高应用价值。

近年来，基于深度学习的视频动作检测方法逐渐从单一 RGB 外观建模转向时空特征建模和多实体交互建模。一方面，三维卷积网络、双流网络和 SlowFast 等[2]视频特征提取结构能够从连续帧中捕获运动信息，提高动作表征能力；另一方面，动作往往不仅由人体外观决定，还与人体姿态、手部区域、交互物体以及前后帧上下文密切相关。因此，如何充分利用多模态信息并对时序上下文进行有效建模，成为提升时空动作检测性能的重要方向。

Holistic Interaction Transformer Network (HIT) [3]通过构建人物、物体、手部、姿态和 memory feature 之间的交互关系，为多模态时空动作检测提供了较完整的建模框架。该模型利用 RGB 分支提取视频外观特征，同时引入人体关键点和记忆特征辅助动作分类，在 JHMDB-21 等数据集上表现出较好的检测性能。然而，在对 HIT 进行复现和实验分析的过程中可以发现，原始结构中仍存在两个可以进一步优化的环节。首先，RGB 特征与姿态交互特征之间通常采用固定加和或拼接方式进行融合，缺少对姿态分支可靠性的自适应判断；当关键点存在偏移、遮挡或检测误差时，直接融合可能引入辅助模态噪声。其次，memory 分支在利用前后时序上下文时，对不同时间距离的记忆特征缺少显式区分，距离当前帧较远的上下文可

能与当前动作状态不一致，从而影响视频级检测稳定性。

针对上述问题，本文在 HIT 基线模型基础上提出一种轻量改进方法。首先，在 RGB ROI 特征与姿态交互特征融合阶段引入门控姿态融合模块，通过由两路特征共同生成的门控权重自适应调节姿态特征注入强度。其次，在 memory feature 输入时序交互模块之前，提出时序距离感知的记忆特征加权方法，根据记忆帧与当前帧的时间距离构造衰减权重，以增强近邻上下文并减弱远距离上下文噪声。两个模块均不改变 HIT 主体检测流程，额外计算成本较低，便于在已有复现基础上进行训练和消融验证。

本文主要贡献包括以下三个方面。第一，针对 HIT 中姿态分支固定融合的问题，设计了门控姿态融合模块，使模型能够根据 RGB 与姿态交互特征动态调整辅助姿态信息的贡献。第二，针对 memory feature 等权使用的问题，提出时序距离感知记忆加权模块，以基于时间距离的衰减权重提升时序上下文利用的稳定性。第三，在 JHMDB-21 数据集[4]上完成自训练基线、单模块改进和组合模型的对比实验与消融分析，验证所提出轻量模块对视频级时空动作检测性能的提升作用。

2. 相关工作

2.1. 时空动作检测

时空动作检测任务要求模型在视频中同时完成动作类别识别和人物空间定位。早期视频动作识别方法多关注整段视频或短视频片段的分类，而时空动作检测进一步要求在帧级或视频级上输出人物动作框和类别得分。随着深度学习方法的发展，研究者逐渐将二维卷积特征、三维卷积特征和双流特征引入动作检测框架，通过融合外观信息和运动信息提升视频动作建模能力。近年来，SlowFast 等结构通过不同时间采样速率的分支同时建模慢速语义信息和快速运动变化，为动作检测提供了更有效的视频特征表示。

在动作检测框架上，现有方法大致可以分为两阶段方法和端到端方法。两阶段方法通常先生成候选人物框或动作候选区域，再对候选区域进行时空特征提取和分类；端到端方法则尝试在统一网络中同时完成特征学习、定位和分类。对于 JHMDB-21、UCF101-24 等数据集，帧级 mAP 和视频级 mAP 常被用于评价模型在空间定位和动作管道连续性上的表现。本文工作属于基于已有 HIT 框架的轻量改进，重点不在于重新设计检测器主体，而是围绕多模态融合和时序记忆利用两个环节提升检测稳定性。

2.2. 多模态动作理解与交互建模

人体动作通常由外观、姿态、交互物体和场景上下文共同决定。仅依赖 RGB 外观特征时，模型容易受到背景变化、遮挡、视角变化和类别间外观相似性的影响。为缓解这一问题，许多研究将人体关键点、光流、物体区域或人体部件区域引入动作理解模型中。人体姿态能够提供相对清晰的骨架结构信息，尤其适合描述挥手、跳跃、投掷等与身体形态密切相关的动作；物体和手部区域则有助于识别与交互对象有关的动作类别。

HIT 模型进一步将人物、物体、手部、姿态和时序记忆信息统一到交互建模框架中，通过不同实体之间的信息传递增强动作表征能力。这类方法的优势在于能够从动作发生机制出发，将“人正在做什么”和“人与什么发生交互”结合起来。然而，多模态信息并非在所有样本中都同等可靠。例如，关键点检测在遮挡或快速运动场景下可能出现偏差，姿态分支的辅助作用也会随样本变化而变化。如果模型始终以固定方式融合姿态特征，就可能在部分样本中引入额外噪声。因此，本文在 HIT 姿态分支之后引入门控融合机制，使模型能够根据当前 RGB 特征和姿态交互特征动态调节姿态信息贡献。

2.3. 时序上下文建模

视频动作具有明显的时序连续性，当前帧中的动作状态往往与前后帧密切相关。因此，时序上下文

建模是视频动作检测中的重要问题。常见方法包括使用三维卷积捕获局部时间窗口内的运动变化，使用循环结构或 Transformer 建模较长时间范围的依赖关系，以及利用 memory feature 保存前后帧的人物特征以辅助当前帧分类。对于时空动作检测而言，合理利用时序上下文不仅有助于帧级分类，还会影响视频级动作管道的连续性和稳定性。

尽管时序上下文通常能够提供辅助信息，但不同时间位置的上下文与当前帧并不具有相同相关性。距离当前帧较近的记忆特征往往对应动作的连续变化过程，而距离较远的记忆特征可能处于动作准备阶段、结束阶段或其他状态。如果将远近不同的 memory feature 等权输入交互模块，可能削弱当前动作表征的一致性。基于这一观察，本文提出时序距离感知的记忆特征加权模块，在不改变原有 memory 结构的前提下，根据时间距离对记忆特征进行衰减加权，使近邻上下文在时序交互中发挥更主要作用。

3. 改进方法

本文以 Holistic Interaction Transformer Network (HIT)为基线模型，针对其在多模态姿态特征融合和时序记忆特征利用中的可靠性建模不足问题进行轻量改进。整体思路是在不改变 HIT 主体交互结构和检测流程的前提下，分别从两个位置引入改进：一是在 RGB ROI 特征与姿态交互特征融合阶段加入门控姿态融合模块，使模型能够自适应调节姿态分支对最终动作特征的贡献；二是在 memory feature 输入时序交互模块之前，根据记忆帧与当前帧的时间距离对记忆特征进行加权，从而增强近邻上下文并抑制远距离上下文噪声。通过上述两个模块，本文希望在较小额外计算成本下提升模型在 JHMDB-21 数据集上的视频级时空动作检测性能。本文方法的整体框架如图 1 所示。

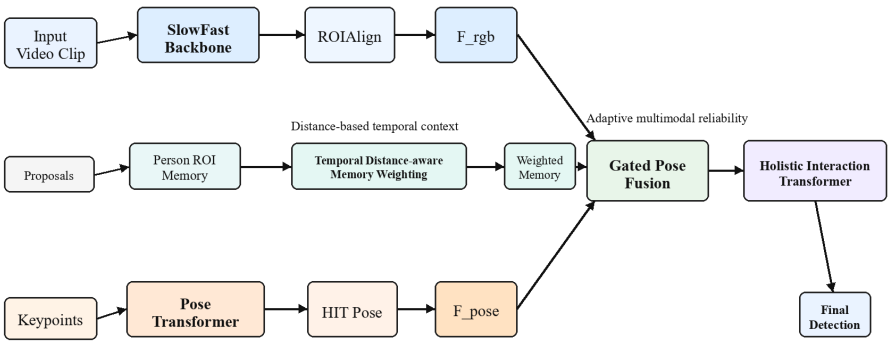


Figure 1. Overall framework of the improved HIT model
图 1. 模型整体框架

3.1. HIT 基线模型

HIT 是一种面向视频时空动作检测的多模态交互模型。对于输入视频片段，模型首先利用 SlowFast RGB Backbone 提取视频外观特征，并通过 ROIAlign [5] [6]获得以候选人物框为中心的 ROI 特征。与此同时，模型利用人体关键点信息构建姿态分支，通过 Pose Transformer [7]对关键点序列进行编码，得到与人体动作状态相关的姿态表征。随后，HIT 通过交互模块建模人物、手部、物体以及时序 memory feature 之间的关系，并将交互后的特征输入分类器，完成动作类别预测。

在原始 HIT 结构中，RGB 分支和姿态分支通常采用固定加和或拼接方式进行融合。这类融合方式实现简单，但默认不同模态特征始终可靠，缺少对辅助姿态信息质量的自适应判断。对于 JHMDB-21 这类规模较小且动作类别差异较明显的数据集，姿态信息通常有助于动作识别；但当关键点检测存在偏移、

遮挡或动作姿态变化较大时，直接引入姿态特征也可能带来噪声。此外，HIT 的 memory 分支会读取当前帧前后多个时间位置的记忆特征，但不同时间距离的记忆特征与当前动作状态的相关程度并不相同。基于上述观察，本文设计了门控姿态融合模块和时序距离感知记忆加权模块。

3.2. 门控姿态融合模块

具体而言，输入视频片段首先经过 SlowFast RGB Backbone 和 ROIAlign 得到 RGB ROI 特征，记为 F_{rgb}。人体关键点输入 Pose Transformer 后，再经过 HIT Pose 交互分支得到姿态交互特征，记为 F_{pose}。需要强调的是，本文的门控模块并不是直接作用于原始关键点坐标，而是作用于经过姿态编码和交互建模后的 F_{pose}，因此该特征已经包含一定的人体结构和动作交互信息。

为了获得姿态分支的自适应权重，本文首先将 F_{rgb} 与 F_{pose} 在通道维度进行拼接，然后通过 $1 \times 1 \times 1$ 的三维卷积进行特征压缩与映射，并使用 Sigmoid 函数生成门控权重 G。其计算过程如下：

$$G = \text{sigmoid}(\text{Conv3d}([F_{\text{rgb}}, F_{\text{pose}}])) \tag{1}$$

其中，$[F_{\text{rgb}}, F_{\text{pose}}]$表示两路特征在通道维度上的拼接，Conv3d 表示 $1 \times 1 \times 1$ 三维卷积，Sigmoid 函数将输出限制在 0 到 1 之间。随后，门控权重 G 与姿态交互特征 F_{pose} 进行逐元素相乘，得到加权后的姿态特征。最终输出特征 F_{out} 通过残差方式与 RGB ROI 特征相加：

$$F_{\text{out}} = F_{\text{rgb}} + G \cdot F_{\text{pose}} \tag{2}$$

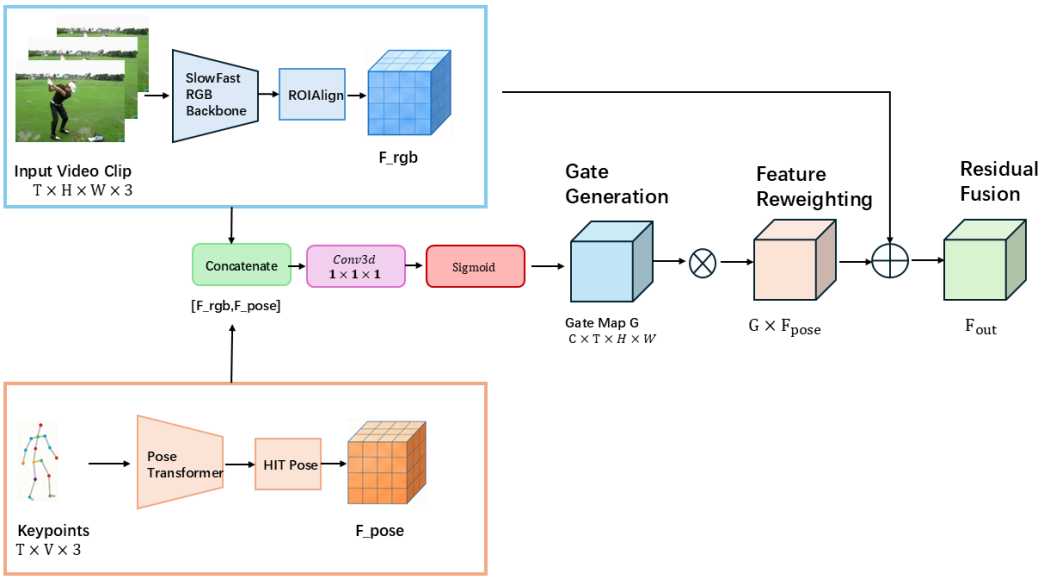


Figure 2. Gated pose fusion module
图 2. 门控姿态融合模块

上述设计具有两个作用。首先，当姿态分支对当前动作判断具有较强辅助作用时，模型可以通过较大的门控权重增强 F_{pose} 的贡献；其次，当姿态关键点受到遮挡、检测误差或动作姿态波动影响时，门控权重可以降低对应姿态特征的注入强度，从而减少不可靠辅助模态对 RGB 主分支的干扰。与直接加和相比，门控融合[8] [9]在保留残差结构稳定性的同时，引入了对姿态特征可靠性的自适应建模；与直接拼接相比，该模块不会显著增加后续分类层输入维度，参数增加主要来自一个 $1 \times 1 \times 1$ Conv3d，因此整体计算成本较低。

在实现上，当配置文件中的融合方式设置为 gated_add 时，模型在特征提取器中构建门控卷积层，并

在前向传播时执行式(1)和式(2)的计算。为保证训练初期不破坏原始加和融合结构，门控卷积层采用接近恒等残差融合的初始化方式，使模型初始状态接近原始 add 融合，再在训练过程中逐步学习不同样本和不同特征位置下的姿态注入强度。图 2 给出了本文设计的门控姿态融合模块(Gated Pose Fusion Module)。该模块的目标是在 RGB ROI 特征和姿态交互特征之间建立一种可靠性感知的融合机制，避免将姿态分支信息以固定强度直接加入 RGB 动作特征。

3.3. 时序距离感知记忆特征加权模块

设当前帧时间戳为 t，模型从 memory cache 中读取当前帧前后若干时间位置的人物 ROI 记忆特征，记为：

$$F_{\text{mem}} = \{F_{t-S}, \ldots, F_{t-1}, F_t, F_{t+1}, \ldots, F_{t+S}\} \tag{3}$$

其中，S 表示向前或向后读取的记忆范围，F_t 表示当前帧对应的人物 ROI 特征。对于任意一个记忆帧 t_i，本文首先计算其与当前帧之间的时间距离：

$$\Delta_i = |t_i - t| \tag{4}$$

在此基础上，本文采用指数衰减形式构造时序权重：

$$\omega_i = \omega_{\min} + (1 - \omega_{\min}) \cdot \exp(-\Delta_i/\tau) \tag{5}$$

其中，τ 为控制衰减速度的超参数，$\omega_{\min}$ 为最小保留权重。本文实验中设置 $\tau = 8.0$，$\omega_{\min} = 0.5$。该设置表示当前帧及近邻帧拥有较高权重，而距离当前帧较远的上下文特征会被适当衰减，但不会被完全丢弃。

根据式(5)得到权重后，对每一个记忆特征进行逐元素加权：

$$F'_i = \omega_i \cdot F_i \tag{6}$$

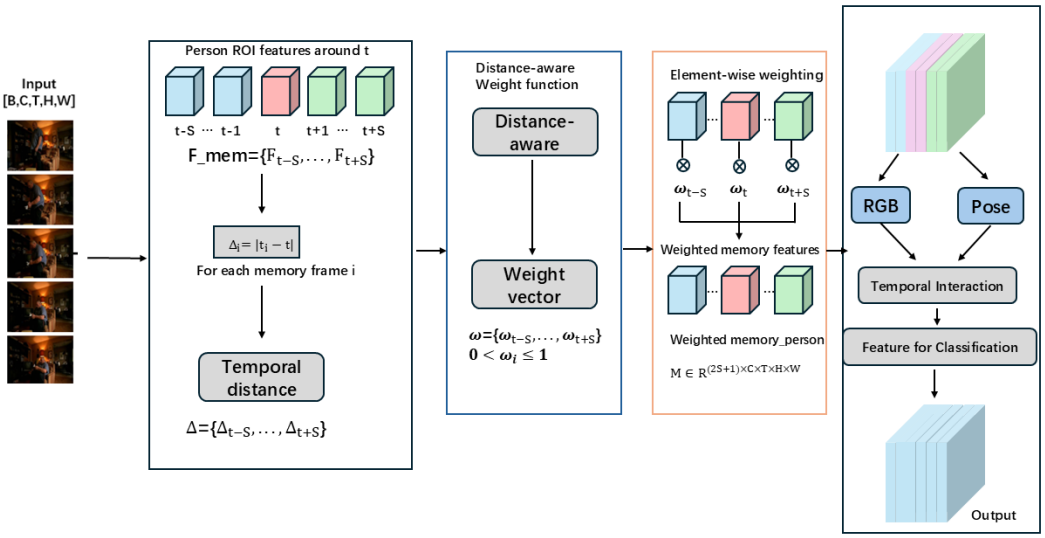


Figure 3. Temporal distance-aware memory weighting module
图 3. 时序距离感知记忆特征加权模块

加权后的记忆特征集合 F'_{mem} 被重新组织为 memory_person，并输入 HIT 的 temporal interaction 模块，与 RGB 分支和姿态分支的交互特征共同参与后续动作分类。与学习式注意力模块[10]相比，该方法不引入额外可学习参数，仅根据时间距离对已有 memory feature 进行重标定，因此实现简单、训练成本低，并且能够较直接地体现视频动作中“近邻帧通常更相关”的时序先验。

该模块的主要作用在于降低远距离上下文噪声。对于视频动作检测任务，动作状态和人体姿态会随

时间连续变化，距离当前帧过远的 memory feature 可能对应动作的准备阶段、结束阶段或背景状态，其语义与当前动作定位并不完全一致。如果将这些记忆特征与近邻帧特征等权输入时序交互模块，可能会降低动作管道评分的一致性。本文通过时间距离加权，使近邻 memory feature 在交互过程中占据更高贡献，从而提升视频级检测结果的稳定性。图 3 给出了本文设计的时序距离感知记忆特征加权模块(Temporal Distance-aware Memory Weighting Module)。该模块的目标是在 memory feature 进入 HIT 时序交互模块之前，根据记忆帧与当前帧之间的时间距离调整其特征强度，从而提升时序上下文利用的稳定性。

3.4. 总体模型与实现方式

本文最终模型同时启用门控姿态融合模块和时序距离感知记忆特征加权模块。整体流程可以概括为：首先利用 RGB 分支获得人物 ROI 外观特征 F_{rgb}，并利用关键点分支获得姿态交互特征 F_{pose}；随后通过门控姿态融合模块生成融合特征 F_{out}；与此同时，模型在读取 memory feature 时根据时间距离生成权重向量，对前后时序记忆特征进行重标定；最后，融合后的当前帧特征与加权后的记忆特征共同输入 HIT 交互结构和分类器，完成动作类别预测。

在实验配置中，门控姿态融合通过将 HIT_STRUCTURE.FUSION 设置为 gated_add 启用；时序距离感知记忆加权通过将 MEMORY_TEMPORAL_WEIGHT 设置为 True 启用。时序权重的两个超参数设置为 $\tau = 8.0$ 和 $\omega_{min} = 0.5$。两个模块均保持 HIT 原有主体结构不变，属于轻量级改进：门控姿态融合仅额外引入一个 $1 \times 1 \times 1$ Conv3d 用于生成门控图；时序记忆加权不引入额外可学习参数，只在 memory feature 送入交互模块前进行基于时间距离的乘法重标定。因此，本文方法适合在已有 HIT 复现基础上进行快速训练和消融验证。

需要说明的是，本文还对关键点置信度输入进行了探索，即在 Pose Transformer 输入中额外引入关键点置信度信息。实验结果显示，该策略能够提升部分帧级检测指标，但视频级 mAP 未能超过最终组合模型。因此，本文最终保留门控姿态融合和时序距离感知记忆加权两个模块作为主要改进，并在实验部分通过消融结果验证二者的有效性和互补性。

4. 实验与结果分析

为验证本文两个轻量改进模块的有效性，本章在 JHMDB-21 数据集上对原始 HIT 模型、不同特征融合方式以及时序记忆加权策略进行了对比实验。实验首先使用作者提供的预训练权重完成复现，以确认数据、环境和评价脚本能够正常运行；随后以本文重新训练得到的 baseline/add 模型作为公平比较基线，分别分析门控姿态融合模块和时序距离感知记忆加权模块对检测性能的影响。由于本文工作目标是在复现基础上完成面向硕士论文和中文科技期刊的轻量改进，因此实验重点放在同一数据集、同一训练配置和同一评价流程下的可比性，而不额外扩大到多数据集泛化实验。

4.1. 数据集与评价指标

本文实验采用 JHMDB-21 数据集。该数据集面向人体动作检测任务，包含 21 类动作，并为视频帧中的人体动作目标提供空间位置标注。相较于仅进行视频级分类的任务，时空动作检测需要同时判断动作类别和动作发生位置，因此模型不仅需要提取视频外观特征，还需要维持检测框和动作类别在时间维度上的一致性。

实验采用两类评价指标。第一类为 frame mAP@0.5，用于衡量模型在单帧层面的动作检测能力。当预测框与真实框的 IoU 达到 0.5 且动作类别预测正确时，该预测被视为有效检测。第二类为 video mAP，用于评估模型在视频动作管道层面的检测效果。本文分别报告 Video mAP@0.5 和 Video mAP@0.2，其中

较高 IoU 阈值更强调定位质量，较低 IoU 阈值更关注视频级动作轨迹的整体召回和类别判断能力。通过同时报告帧级和视频级指标，可以更全面地分析改进模块对空间定位与时序稳定性的影响。

4.2. 实验设置

本文以 HIT 模型为基础框架，RGB 分支采用 SlowFast-ResNet50-4x16 预训练特征提取结构，姿态分支用于建模人体关键点交互信息。为保证结果具有可比性，除被验证的模块开关或融合方式外，各组实验保持相同的数据路径、训练轮次和评价脚本。训练过程设置最大迭代次数为 7000，并在训练结束后使用同一测试流程生成检测结果。

评价阶段首先运行模型推理并生成 JHMDB-21 测试集检测结果，然后使用 Pascal VOC 风格评价脚本计算 frame mAP@0.5，再基于检测结果构建视频级动作管道并计算 Video mAP@0.5 和 Video mAP@0.2。需要说明的是，作者权重复现结果主要用于证明当前环境、数据组织和评价流程是正确的；后续消融实验的公平比较均以本文自训练 baseline/add 作为基线。

4.3. 主结果对比

表 1 给出了各组实验在 JHMDB-21 数据集上的整体结果。可以看到，作者权重复现模型取得 83.93% 的 frame mAP@0.5、87.16%的 Video mAP@0.5 和 88.75%的 Video mAP@0.2，说明本文搭建的环境和评价流程能够复现 HIT 模型的主要检测性能。在自训练设置下，baseline/add 模型取得 82.86%、86.14%和 87.40%的结果，该结果作为本文后续模块改进的主要对照。

Table 1. Overall experimental results on JHMDB-21
表 1. JHMDB-21 数据集整体实验结果

方法	融合方式	时序记忆加权	Frame mAP@0.5 (%)	Video mAP@0.5 (%)	Video mAP@0.2 (%)
作者权重复现	-	-	83.93	87.16	88.75
自训练 baseline/add	add	否	82.86	86.14	87.40
concat fusion	concat	否	83.16	86.75	88.17
gated fusion	gated_add	否	84.01	86.77	88.52
gated + pose confidence	gated_add	否	84.32	86.53	88.20
add + temporal memory	add	是	83.78	86.97	88.31
gated + temporal memory	gated_add	是	83.38	88.02	89.55

从整体结果看，最终组合模型 gated + temporal memory 在视频级评价指标上取得最佳结果。其中，Video mAP@0.5 从自训练 baseline/add 的 86.14%提升到 88.02%，提升 1.88 个百分点；Video mAP@0.2 从 87.40%提升到 89.55%，提升 2.15 个百分点。这说明本文的两个改进模块对动作管道级别的连续检测更有帮助。虽然最终组合模型的 frame mAP@0.5 低于单独使用 gated fusion 的结果，但其视频级性能更高，表明该组合更倾向于提升跨帧动作检测的稳定性，而不是仅优化单帧检测结果。

4.4. 门控姿态融合模块消融分析

为了验证姿态特征融合方式对模型性能的影响，本文比较了原始 add 融合、concat 拼接融合和 gated_add 门控融合三种策略，结果如表 2 所示。

Table 2. Ablation study of fusion strategies
表 2. 融合方式消融实验

融合策略	Frame mAP@0.5 (%)	Video mAP@0.5 (%)	Video mAP@0.2 (%)
add	82.86	86.14	87.40
concat	83.16	86.75	88.17
gated_add	84.01	86.77	88.52

与 baseline/add 相比，concat fusion 在 frame mAP@0.5、Video mAP@0.5 和 Video mAP@0.2 上分别提升 0.30、0.61 和 0.77 个百分点，说明增强 RGB 特征与姿态特征的联合表达对动作检测具有一定帮助。然而，concat fusion 只是将两类特征直接拼接，缺少对姿态模态可靠性的显式判断。当姿态关键点受到遮挡、尺度变化或估计误差影响时，直接引入姿态特征可能会带来额外噪声。

gated_add 融合在三项指标上均优于 baseline/add，其中 frame mAP@0.5 提升 1.15 个百分点，Video mAP@0.5 提升 0.63 个百分点，Video mAP@0.2 提升 1.12 个百分点。该结果表明，门控机制能够根据 RGB ROI 特征和姿态交互特征生成自适应权重，从而在姿态信息可靠时增强其贡献，在姿态信息不稳定时降低其影响。因此，相比固定加和或简单拼接，门控姿态融合更适合作为多模态特征融合方式。

4.5. 时序距离感知记忆加权模块消融分析

HIT 模型中的 memory 特征用于为当前动作判断提供邻近帧上下文信息，但不同时间位置的 memory feature 与当前帧之间的相关性并不完全相同。为验证时序距离感知记忆加权模块的有效性，本文分别在 add 融合和 gated_add 融合设置下加入该模块，结果如表 3 所示。

Table 3. Ablation study of temporal memory weighting
表 3. 时序记忆加权消融实验

方法	门控融合	时序记忆加权	Frame mAP@0.5 (%)	Video mAP@0.5 (%)	Video mAP@0.2 (%)
baseline/add	否	否	82.86	86.14	87.40
add + temporal memory	否	是	83.78	86.97	88.31
gated fusion	是	否	84.01	86.77	88.52
gated + temporal memory	是	是	83.38	88.02	89.55

在原始 add 融合设置下，引入时序距离感知记忆加权后，frame mAP@0.5 从 82.86%提升到 83.78%，Video mAP@0.5 从 86.14%提升到 86.97%，Video mAP@0.2 从 87.40%提升到 88.31%。该结果说明，即使不改变 RGB 与姿态特征的融合方式，仅对 memory feature 按时间距离进行权重调整，也能够提升模型对动作时序上下文的利用效率。

在 gated_add 融合基础上继续加入时序记忆加权后，模型取得最高的视频级检测结果。与自训练 baseline/add 相比，最终模型在 Video mAP@0.5 和 Video mAP@0.2 上分别提升 1.88 和 2.15 个百分点。与单独使用 gated fusion 相比，最终模型的 Video mAP@0.5 从 86.77%提升到 88.02%，Video mAP@0.2 从 88.52%提升到 89.55%。这表明门控姿态融合模块和时序距离感知记忆加权模块具有一定互补性：前者主要改善 RGB 与姿态特征之间的多模态融合质量，后者主要改善跨帧 memory feature 的时序利用方式。

4.6. 探索实验与结果讨论

除上述两个保留模块外，本文还进行了关键点置信度输入的探索实验。该实验尝试将姿态估计中的关键点置信度作为额外信息引入姿态分支，以期提升模型对姿态可靠性的判断能力，结果如表 4 所示。

Table 4. Exploratory experiment of pose confidence input
表 4. 关键点置信度输入探索实验

方法	Frame mAP@0.5 (%)	Video mAP@0.5 (%)	Video mAP@0.2 (%)
gated fusion	84.01	86.77	88.52
gated + pose confidence	84.32	86.53	88.20

实验结果显示，加入关键点置信度后，frame mAP@0.5 从 84.01%提升到 84.32%，说明该信息对单帧层面的动作定位和分类可能具有一定帮助。然而，该设置下 Video mAP@0.5 和 Video mAP@0.2 分别下降到 86.53%和 88.20%，低于 gated fusion。该现象说明，关键点置信度虽然可以增强单帧姿态质量判断，但其在连续帧中的波动可能影响视频动作管道的评分一致性。因此，本文未将该探索模块作为最终方法，而是保留视频级性能更稳定的时序距离感知记忆加权模块。

综合各组实验可以看到，本文方法的主要收益并非单纯来自模型参数规模增加，而是来自对多模态融合可靠性和时序上下文相关性的重新建模。门控姿态融合模块能够在 RGB 与姿态特征之间建立自适应融合关系，时序距离感知记忆加权模块则通过无需额外学习参数的方式调整不同时间位置 memory feature 的贡献。二者结合后，模型在视频级指标上取得更明显提升，符合时空动作检测任务对跨帧稳定性的要求。

4.7. 可视化定性分析

为了更直观地分析所提模块的作用，本文进一步对门控姿态融合模块和时序距离感知记忆加权模块进行可视化。对于门控姿态融合模块，本文选取测试集中具有代表性的动作片段，提取模型在推理过程中生成的门控权重 G，并将不同人物 ROI 的平均门控响应叠加到原始视频帧上，同时绘制门控权重在通道维度上的响应分布。需要说明的是，本文门控权重作用于 ROI 池化后的特征张量，而不是直接作用于原始图像像素，因此可视化结果主要反映模型在不同人物区域和不同特征通道上的姿态信息注入强度。

如图 4 所示，在姿态较清晰的动作样例中，门控模块能够为姿态交互特征分配较稳定的响应，使姿态分支更充分地参与动作表征；而在姿态存在遮挡、关键点不稳定或动作变化较大的样例中，不同 ROI 和不同通道上的门控响应出现差异，说明模型并非以固定强度引入姿态特征，而是能够根据 RGB ROI 特征和姿态交互特征的联合状态自适应调整姿态信息的贡献。该结果从定性角度说明，门控姿态融合模块能够缓解固定加和融合对姿态可靠性缺乏判断的问题。

对于时序距离感知记忆加权模块，本文选取一个动作实例，绘制当前帧前后不同 memory feature 的时间距离及其对应权重。可视化结果如图 5 所示，其中横轴表示记忆帧与当前帧之间的时间距离，纵轴表示根据式(5)计算得到的时序衰减权重。

从图 5 可以看到，当前帧及其邻近帧具有较高权重，随着记忆帧与当前帧之间的时间距离增大，权重逐渐衰减并趋近于最小保留权重。该设计使模型在利用前后文信息时更加关注与当前动作状态相关性更强的邻近帧，同时不会完全丢弃远距离记忆特征。该可视化结果直观说明，时序距离感知记忆加权模块能够在不增加额外可学习参数的情况下，对不同时间位置的 memory feature 进行差异化重标定，从而提升视频级动作检测结果的稳定性。

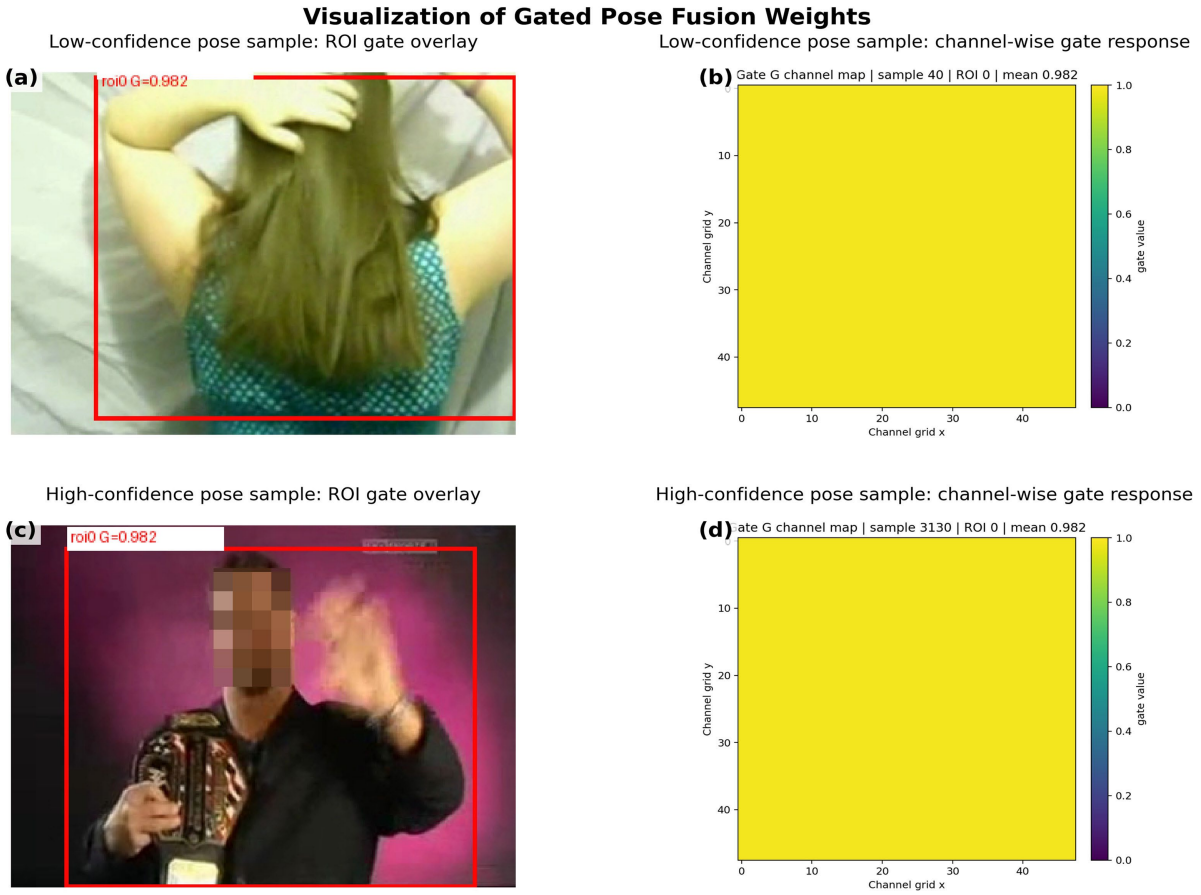


Figure 4. Visualization of weights in the gated pose fusion module
图 4. 门控姿态融合模块权重可视化结果

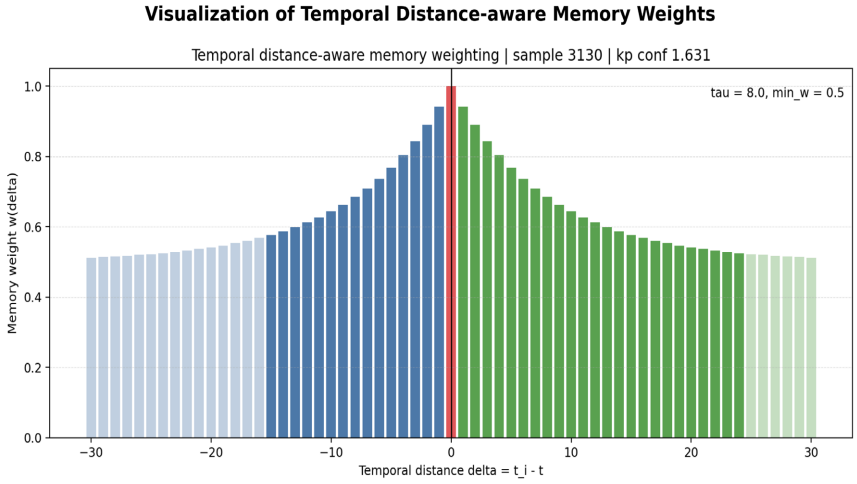


Figure 5. Visualization of adaptive weights of the temporal distance-aware memory weighting module
图 5. 时序距离感知记忆加权权重可视化结果

5. 总结

本文围绕 HIT 模型在多模态姿态融合和时序记忆特征利用方面的不足，提出了一种面向 JHMDB-21

数据集的轻量改进方法。该方法在不改变 HIT 主体交互结构和检测流程的前提下，引入门控姿态融合模块和时序距离感知记忆加权模块，分别从辅助姿态特征可靠性和时序上下文相关性两个角度改进模型。

实验结果表明，门控姿态融合模块能够提升 RGB 特征与姿态交互特征的融合效果，时序距离感知记忆加权模块能够改善不同时间位置 memory feature 的利用方式。二者结合后，最终模型在 JHMDB-21 数据集上取得 88.02%的 Video mAP@0.5 和 89.55%的 Video mAP@0.2，相较自训练基线分别提升 1.88 和 2.15 个百分点，说明所提方法对视频级动作检测稳定性具有积极作用。

本文方法的优势在于改动较小、实现简单，并能够在已有 HIT 复现基础上完成训练和消融验证。受限于实验成本，本文主要在 JHMDB-21 数据集上进行验证，后续可进一步在 UCF101-24、AVA 等更大规模数据集上测试模型泛化能力，并结合检测结果可视化分析不同动作类别下的改进效果。

参考文献

[1] Kalogeiton, V., Weinzaepfel, P., Ferrari, V. and Schmid, C. (2017) Action Tubelet Detector for Spatio-Temporal Action Localization. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 4415-4423. https://doi.org/10.1109/iccv.2017.472

[2] Feichtenhofer, C., Fan, H., Malik, J. and He, K. (2019) SlowFast Networks for Video Recognition. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 6201-6210. https://doi.org/10.1109/iccv.2019.00630

[3] Faure, G.J., Chen, M.H. and Lai, S.H. (2023) Holistic Interaction Transformer Network for Action Detection. 2023 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 2-7 January 2023, 3329-3339. https://doi.org/10.1109/wacv56688.2023.00334

[4] Jhuang, H., Gall, J., Zuffi, S., Schmid, C. and Black, M.J. (2013) Towards Understanding Action Recognition. 2013 *IEEE International Conference on Computer Vision*, Sydney, 1-8 December 2013, 3192-3199. https://doi.org/10.1109/iccv.2013.396

[5] Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems* 28, Montreal, 7-12 December 2015, 1137-1149.

[6] He, K., Gkioxari, G., Dollar, P. and Girshick, R. (2017) Mask R-CNN. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/iccv.2017.322

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems* 30, Long Beach, 4-9 December 2017, 5998-6008.

[8] Arevalo, J., Solorio, T., Montes, M. and Gonzalez, F.A. (2017) Gated Multimodal Units for Information Fusion. 2017 *ICLR Workshop*, Toulon, 24-26 April 2017, 1-12.

[9] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/cvpr.2018.00745

[10] Wang, X., Girshick, R., Gupta, A. and He, K. (2018) Non-Local Neural Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7794-7803. https://doi.org/10.1109/cvpr.2018.00813