计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(8): 2127−2142, 2026

RDT-MoE: 融合冗余削减与混合专家的暴力行为识别方法

杨建锋[1] 于大为[2] 曹 瑞[1] 任佳航[1] 王 平[1] 胡琦瑶[3]
[1](西北大学计算机学院 西安 710127)
[2](上海华为技术有限公司 上海 201206)
[3](西北大学电子信息学院 西安 710127)
(yjf@nwu.edu.cn)

RDT-MoE: Integrating Redundancy Reduction and Mixture-of-Experts for Violence Action Recognition

Yang Jianfeng[1], Yu Dawei[2], Cao Rui[1], Ren Jiahang[1], Wang Ping[1], and Hu Qiyao[3]
[1](*College of Computer Science*, *Northwest University*, *Xi'an* 710127)
[2](*Shanghai Huawei Technologies Co.*, *Shanghai* 201206)
[3](*School of Electronic Information*, *Northwest University*, *Xi'an* 710127)

Abstract To tackle the dual challenges of spatiotemporal feature redundancy and limited generalization in violent action recognition, we present RDT-MoE, a lightweight yet effective framework that couples redundancy-diminishing token (RDT) with a mixture-of-experts (MoE) architecture. The RDT module employs spatially and temporally decoupled attention to disentangle motion from appearance, and uses a dynamic Top-*k* token selection strategy to suppress nearly 50% of background redundancy, keeping only the most salient motion cues for subsequent encoding. Building on these compact tokens, we develop a LoRA-guided expert system and train it with a two-stage strategy, where we first jointly optimize the encoder and LoRA adapters to learn strong shared representations and stable adaptation signals, and then freeze the backbone to focus on refining the gating function and expert composition, which further improves robustness, convergence stability, and generalization across diverse scenes. To address limitations of existing benchmarks, we also introduce diverse violence action dataset (DVAD), a higher-resolution, multi-view violence dataset that is 50% larger than RWF-2000 and better reflects real-world variability. Extensive evaluations on RWF-2000 and DVAD, together with component-wise ablations, show that RDT-MoE outperforms VideoMAE by 2.00% and 2.34% in accuracy, respectively, while reducing parameters by about 50%, demonstrating strong efficiency, generalization, and practical applicability. This design supports deployment on resource-limited surveillance devices and maintains stable performance across diverse scenes consistently.

Key words violence action recognition; violence action dataset; LoRA; MoE; pre-training

摘 要 面对暴力行为识别中普遍存在的时空特征冗余与模型泛化不足的双重挑战,提出了一种融合稀疏词元选择与混合专家架构的高效识别模型——RDT-MoE。该模型由冗余特征去除模块 RDT(redundancy-diminishing token)与参数化混合专家网络(mixture-of-expert, MoE)组成。具体而言,RDT 模块基于时空分离注意力机制实现运动特征的解耦,并结合 Top-*k* 词元选择策略动态过滤约 50% 的背景冗余特征,从而显著提升特征表达效率;MoE 部分创新性地引入 LoRA 参数更新策略,采用 2 阶段训练范

收稿日期: 2025-10-14; **修回日期**: 2026-03-25
基金项目: 国家自然科学基金项目(62406247)
This work was supported by the National Natural Science Foundation of China (62406247).
通信作者: 曹瑞(cr@nwu.edu.cn)

式: 第 1 阶段训练 RDT 模块与编码器以获得 LoRA 参数, 第 2 阶段以 LoRA 参数初始化专家网络, 并冻结 RDT 及编码器, 仅微调门控与专家组合策略, 以增强模型的泛化能力与收敛稳定性。此外, 针对现有暴力行为数据集分辨率低、场景单一、样本量不足等问题, 构建了一个高分辨率、多视角的暴力行为数据集 DVAD (diverse violence action dataset), 其数据规模较 RWF-2000 提升约 50%, 在场景多样性与应用代表性方面具有显著优势。实验结果表明, RDT-MoE 在 RWF-2000 与 DVAD 数据集上均优于现有主流方法, 在参数量减少约 50% 的同时, 识别准确率分别提升 2.00 个百分点与 2.34 个百分点, 展现出较高的效率与性能优势。

关键词 暴力行为识别; 暴力行为数据集; LoRA; MoE; 预训练

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202550743 **CSTR:** 32373.14.issn1000-1239.202550743

由于存在时空冗余和现有模型的泛化能力差, 视频监控中的暴力行为识别仍然是一项关键但具有挑战性的任务。当前的方法往往难以从杂乱的场景中有效地提取出突出的运动线索, 并且无法适应不同的环境[1]。

以 Transformer[2] 为代表的框架已经在暴力行为识别领域取得了很好的效果, 其强大的时空建模能力推动了视频理解模型从卷积结构向自注意力架构的演进。经典方法 VideoMAE[3] 解决了迁移图像预训练模型的偏执影响, 为视频转换器提供了更好的预训练范式。然而现有暴力行为识别研究仍面临 3 重核心挑战: 数据瓶颈、特征冗余与类内交叉相似性, 如图 1 所示。首先, 暴力行为涵盖推搡、殴打等多样子类, 语义模糊性导致标注标准难以统一, 真实场景数据稀缺且分布失衡[4]。其次, 视频数据的时序冗余性不仅降低模型训练效率, 且干扰关键边缘特征的提取, 制约识别精度[5]。最后, 单一特征提取器因同时兼顾局部细节与全局上下文的解析需求, 往往弱化对判别性线索的聚焦, 进而引发跨类别误判[6]。

针对上述挑战, 现有工作主要从 3 个方向进行探索:

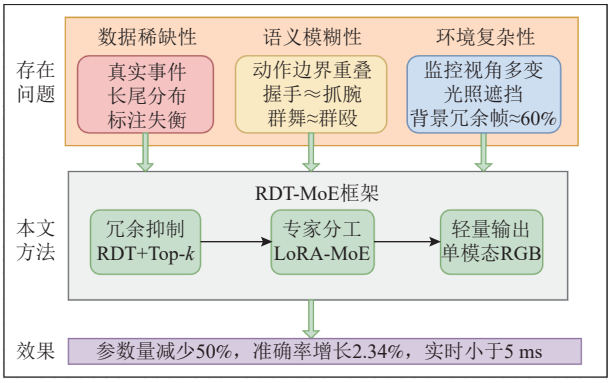


Fig. 1 Challenges in violence action recognition and the technical roadmap of RDT-MoE

图 1 暴力行为识别任务难点与 RDT-MoE 技术路径

1)数据增强与弱监督标注。通过时空切片[7]、GAN 合成样本[8] 等方式扩充数据, 并借助半监督或弱监督策略缓解标注稀缺与分布失衡问题[9]。然而, 这类方法通常依赖复杂生成模型或额外标注资源, 难以适用于大规模监控场景。

2)动态时空注意力与冗余抑制。例如基于自适应词元中止[5] 的机制可在一定程度上压缩冗余帧序列和降低计算开销, 但此类方法更侧重于“剪枝”与“跳过”, 容易忽略动作边缘细节, 造成关键信息丢失。

3)多分支网络与对比学习。多路特征提取或对比损失[10] 有助于增强跨类别判别能力, 缓解语义模糊导致的误判[11]。然而, 这些模型通常伴随显著的参数和算力开销, 难以兼顾轻量化与实时性需求。

因此, 亟需一种既能精确保留重要运动线索又能高效剔除冗余特征, 并在低资源条件下保持出色判别性能的统一框架, 以更好地应对视频监控中暴力行为识别的 3 重核心挑战。

针对上述 3 重挑战, 本文提出了一种基于 VideoMAE 基线并结合冗余感知动态词元模块 RDT (redundancy-diminishing token)与混合专家(mixture-of-expert, MoE)网络的统一识别框架, 并引入 2 阶段低秩自适应(LoRA)初始化策略, 既能精确保留重要运动线索, 又能高效剔除冗余特征, 提升暴力行为识别性能。首先, 以在 Kinetics-400[12] 上预训练的 VideoMAE 作为基线模型, 以避免从图像预训练模型迁移所带来的偏差。在此基础上, 引入 RDT 模块, 该模块结合分解时空注意力机制与 Top-k 词元选择策略, 剔除冗余特征并保留任务相关的重要信息, 显著提升计算效率。随后, 在编码器后接入 MoE 模块, 利用其专家多样性进一步增强特征表达能力[13]。针对目标领域数据规模有限的问题, 本文提出一种相关任务

驱动的 2 阶段训练策略：第 1 阶段在融合 ViT[14] 与 RDT 的骨干网络上进行训练，并采用 LoRA 对编码器进行参数高效微调[15]，得到 LoRA 适配参数；第 2 阶段冻结编码器与 RDT 部分，利用前一阶段获得的 LoRA 参数初始化 MoE 专家权重，并对该模块进行微调。

本文的主要贡献包括 4 个方面：

1）构建了一个更具代表性的暴力行为数据集（DVAD）。与现有公开数据集相比，该数据集分辨率更高、规模更大，其采集自监控场景与移动设备真实拍摄的暴力行为视频，同时涵盖多种暴力类型及易混淆的非暴力行为，能够有效提升模型泛化能力，具有较高的实际应用价值。

2）提出 RDT，融合时空分离注意力机制与 Top-k 词元选择策略，可有效识别并剔除视频中的冗余特征，使模型更聚焦于行为相关的重要线索。

3）提出一种结合 MoE 网络的 RDT-MoE 混合架构，通过门控网络自适应分配输入至不同专家模型，并对多个专家输出的特征进行加权融合，进一步增强模型表征能力。

4）提出一种基于相关任务驱动的 2 阶段 LoRA 参数初始化策略，可在低资源条件下促进专家模型的多样性并提升训练效率，该策略具备较强的迁移性与扩展性，可推广至其他数据稀缺的任务领域。

1 相关工作

1.1 行为识别方法：发展与挑战

视频行为识别技术的演进，从视频解析、长序列建模以及稀疏数据强化三大维度，经历了从局部特征提取到全局时空上下文理解的多阶段跃迁。

为了充分刻画主体动作信息，早期方法以 2D 卷积网络为基础，通过静态图像提取空间特征，但因无法捕捉运动变化，难以区分动态行为，如早期针对大规模视频分类的工作[16]。随后，双流卷积神经网络（two stream CNN）[17] 通过并行学习 RGB 图像和光流分支，引入运动线索，但光流计算的高延迟限制了实时性能。时序分段网络（temporal segment network, TSN）[18] 的稀疏时间采样和 SlowFast 网络[19] 的高低帧率路径设计在优化长序列模拟方面效果显著，但仍无法彻底摆脱光流带来的高耗时问题。3D 卷积网络（如 C3D[11], I3D[20]）直接在时空域进行卷积建模，提取联合特征。R(2+1)D[21] 将时空卷积分解为先空间、后时间的 2 步卷积，显著降低了计算复杂度，并在精度与效率之间取得平衡。

为捕捉更长的时域依赖，研究人员融合 LSTM 与 CNN，用以显式建模时间关联[22]。提出了诸如 LRCN[23], ConvLSTM[4], SepConvLSTM[24] 的经典方法，虽能够显式建模时间关联，但其串行计算瓶颈与长程建模缺陷促使研究者转向 Transformer 架构。随着 Transformer 在自然语言处理中的成功，ViViT[25] 和 TimeSformer[26] 通过分离时空注意力并行解析任意时空依赖，大幅提升了长程建模能力。与此同时，基于骨架的图卷积网络（graph convolution network, GCN）发展迅速，ST-GCN（spatial–temporal GCN）[27] 结合骨骼拓扑结构，有效抑制背景干扰并实现轻量化部署，CTR-GCN[28] 进一步通过通道层面的拓扑学习增强骨架特征的表示能力。此外，近年来还涌现出结构与动作融合的 GCN 变体（如 AS-GCN[29]）以及融合空间-时序注意力的 Transformer 架构（如 ST-TR[30]），进一步提升骨架数据处理能力，但这类方法非常依赖骨架检测的结果。

为了面对动作数据标注成本高与复杂场景的挑战，弱监督方法[31-32] 通过动作-背景分离与跨模态对齐挖掘有限标注信息，多模态融合模型[33-34] 结合音频、骨架等多源数据增强鲁棒性，但跨模态噪声和对齐难题仍待解决。与此同时，VideoMAE[3] 通过掩码重建的自监督预训练进一步降低了对标注数据的依赖，而适配器技术（如 ST-Adapter[35]）则支持视频领域预训练模型的高效任务迁移，为解决特定识别任务下数据稀缺问题提供了新思路。

尽管上述方法在各自领域取得显著成效，但在资源受限及动作样本高度相似的现实场景中，模型的泛化性能仍有待提升。为此，本文设计了一套结合数据扩增策略、动态路由机制与多专家协同架构的综合框架，旨在通过判别特征的高效解耦与集成，提升暴力行为识别准确率及跨域鲁棒性，并为复杂环境下的实时部署奠定基础。

1.2 暴力行为数据集：机遇与局限

暴力行为识别领域的数据集发展呈现出场景细化与实用化并行的趋势，从“环境多样性”“数据规模”“视频质量”“跨场景泛化”4 个维度的比较如表 1 所示。

早期数据集如 Hockey Fight Dataset 与 Movies Fight[36]，分别聚焦受控环境和影视场景的暴力行为建模。前者从冰球比赛中提取 1 000 段 1~2 s 的冲突片段，后者基于电影片段构建 200 段视频。二者虽为特定场景下的行为识别提供基准，但受限于单一行为

Table 1 Comparison of Violence Action Datasets

表 1 暴力行为数据集对比

发布年份	数据集名称	冲突片段数量	单条视频时长/s	平均分辨率	数据来源
2011	Hockey Fight Dataset[36]	1 000	1~2	360×288	曲棍球比赛
2011	Movies Fight[36]	200	1~2	720×480	电影
2015	ViolentCrowd[37]	246	1~6	320×240	自然场景
2020	RWF-2000[38]	2 000	5	724×445	监控、电影
2025	DVAD (本文)	3 000	5	763×592	自然场景

类型和人工修饰问题，难以迁移至真实场景。

随着公共安全需求升级，ViolentCrowd[37] 通过 246 段人群密集场景视频构建了复杂环境下的评估体系，但其数据规模不足且视频质量较低，对模型在遮挡、光照变化等干扰下的动作识别提出严峻挑战。近年来主流的 RWF-2000[38] 通过 2 000 段监控/电影片段大幅扩充数据规模，增强了模型对常见暴力场景（如打斗、推搡等）的学习能力。但其依然以单摄像头角度为主，分类粒度较粗，缺乏多视角、多行为混合的标注，限制了跨场景适应性。因此，现有数据集虽然在场景细分与实用化方面各有侧重，但均存在“样本规模或分辨率不足”“视角与行为类型单一”“正负样本易混淆场景缺失”等局限。

为此，本文构建了 DVAD（diverse violence action dataset）数据集，它以高分辨率、多视角采集真实监控和移动设备拍摄视频为基础，并专门包含大量与暴力行为易混淆的非暴力场景，从而更严苛地评估模型在复杂环境下的判别能力。

1.3 MoE 架构：演进与潜力

MoE 架构最初由 Jacobs 等人[39] 提出，通过为不同输入选择不同的专家网络，以提升模型表达能力。然而，其早期实现中所有专家同时激活，造成了显著的计算资源浪费。为此，后续研究引入了条件计算与稀疏门控机制，如稀疏激活策略[40]，通过门控网络选取 Top-k 个专家参与推理，在保证性能的同时大幅提升计算效率。

在自然语言处理（NLP）领域，MoE 架构迅速发展并成为大规模语言模型的关键技术之一。GShard[13] 通过自动分片机制支持跨设备训练，Switch Transformers[41] 进一步简化为 Top-1 路由策略，并引入负载均衡损失函数，提升了模型的训练效率和可扩展性。

视觉与多模态任务中的 MoE 扩展亦取得重要进展。V-MoE[42] 率先将稀疏门控引入 Vision Transformer 架构，在图像分类任务中实现了参数效率与性能之间的良好权衡。Swin-MoE[43] 利用自适应门控策略提升了视觉语言模型在跨模态场景中的泛化能力，而 ST-MoE[44] 则针对动态视频场景提出稳定可迁移的 MoE 架构，显示出在视频理解领域的潜力。

为进一步优化计算与通信开销，研究者提出了多种创新策略。例如，Expert-Choice MoE[45] 颠覆传统由门控控制的路径选择逻辑，允许专家主动决定是否处理特定输入，提升了路由的表达能力与效率。Brainformer[46] 设计了分层式专家激活机制，减少了不同专家之间的通信依赖，优化了大规模部署时的系统开销。MoE-LLaVA[47] 则聚焦于轻量级视觉-语言模型中的专家集成问题，通过引入时空适配器提高了跨模态微调能力。最新的 DeepSeek-V2[48] 进一步在稀疏激活过程中引入细粒度的负载均衡机制，提升了专家路由的稳定性与使用效率。

尽管 MoE 架构已在 NLP 及部分图像任务中取得突破性成果，其在视频理解，尤其是暴力行为识别中的应用仍属新兴。考虑到视频数据本身具有显著的时间和空间异质性，MoE 能够通过动态激活不同专家，按需建模视频中关键时间片段或局部区域，从而有效捕捉长时序结构与复杂动态变化。本文进一步结合 LoRA 参数高效适配策略，提出了一种任务驱动的 2 阶段专家初始化方法，使得 MoE 架构在数据稀缺、标注困难的暴力识别任务中能够发挥更强泛化能力与迁移潜力。

2 方 法

暴力行为在实际监控中呈现“长尾-低先验”分布，且与日常动作存在高度时空重叠；同时，背景冗余占比高、视角光照变化大，导致现有模型在稀缺数据条件下既难捕捉关键边缘特征，又面临计算与部署瓶颈。为此，本文提出“先剪枝、再分工”的 RDT-MoE 框架：先由 RDT 模块对冗余词元进行稀疏化过滤，再通过 LoRA 初始化的 MoE 网络实现样本自适应的专家激活与组合。实验结果表明，在参数量减半的前提下，所提方法在 RWF-2000 与 DVAD 数据集上的准确率分别提升了 2.00 个百分点和 2.34 个百分点，验证了其在低资源实时暴力识别任务中的效率优势与泛化潜力。

2.1 DVAD 数据集构建与预处理

如图 2 所示，本文的自建数据集 DVAD 构建流程包括多个步骤，旨在为暴力行为识别任务提供高质量的标注数据。

1）原始数据采集。本文收集原始视频数据有 2

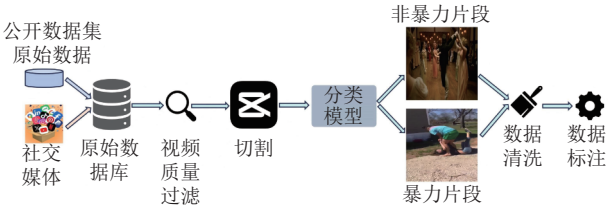


Fig. 2 Workflow of the self-constructed dataset

图 2 自建数据集流程

个主要来源：一方面使用了公开的现有数据集，如 SMFI[49] 和 HMDB[50]，这些数据集包含了各种暴力和非暴力行为的未经剪辑的原始视频；另一方面，从互联网上收集了大量相关的暴力行为视频，包括社交媒体、新闻报道及其他公开的内容。

2）视频质量过滤及切割。在大多数视频中，暴力行为通常只占据整个视频的一小部分，如图 3 所示，本文截取一个暴力行为视频片段的相邻几帧，可以看出连续帧是高度冗余的，同时相邻帧之间也存在着内在的对应关系。本文需要对这些视频进行有效的剪辑和处理，以便提取出具有代表性的暴力行为片段。为了确保数据的一致性和可比性，参考现有暴力行为数据集 RWF-2000 的处理方法，本文将每个视频片段统一切割为 5 s 的长度。这种剪辑方法的选择是基于以下考虑：5 s 的长度足够捕捉到暴力行为的瞬间，但又不会因为过长的时间跨度而混淆非暴力行为；5 s 的长度是处理视频时的一种比较标准化的选择，可以使得模型在训练时能处理统一时长的样本。



Fig. 3 The temporal relationships in video data exhibit a high degree of redundancy across consecutive frames

图 3 视频数据的时间关系展示了连续帧高度冗余

3）自动分类。为了有效地对这些剪辑后的片段进行分类，本文借助已经在 RWF-2000 数据集上微调过的 VideoMAE 模型进行自动分类，该模型能够识别暴力和非暴力行为的视频片段，并将每个片段分为相应类别。

4）半自动数据清洗及标注。由于视频中的一些内容可能被模型误判，为了保证标注质量，本文采用了“人机协同+双人复核”的标注策略。首先利用预训练模型进行初筛，随后由 3 名研究人员进行人工清洗与标注。每段视频至少经过 2 人的交叉验证，当 2 人标注意见不一致时，引入第 3 人进行仲裁，以确保标签的准确性。通过人工筛选和修正，剔除了视频模糊、质量较差及重复的片段，并且对误标注的片段进行了手动标注纠正。

如图 4 所示，通过上述过程，本文构建了一个场景更为复杂、质量更高、数据量更庞大的 DVAD。



(a) 多场景暴力行为帧 (b) 易与暴力混淆的非暴力行为帧

Fig. 4 Sample frames of DVAD dataset

图 4 DVAD 数据集示例帧

2.2 RDT-MoE 模型框架

图 5 展示了本文提出的 RDT-MoE 整体框架。为应对暴力行为识别中普遍存在的时空特征冗余与场景差异带来的建模困难，模型采用“冗余削减+专家分工”的协同设计。

具体而言，在输入端，首先对视频帧进行下采样，并通过立方体嵌入（tubelet embedding）将视频转换为时空词元序列。随后，词元序列进入 RDT 模块，对冗余信息进行稀疏化筛选，保留更具判别性的关键词元，从而在降低计算开销的同时强化对暴力行为关键线索的关注。经过 RDT 处理后的词元被送入编码器进行深层时空特征提取，得到的高层表示进一步输入 MoE 模块。MoE 通过多个专家子网络对特征进行分工建模与动态聚合，利用不同专家在特征模式建模上的互补性，提升模型对复杂暴力场景的表达能力与适应能力，最终由分类头完成暴力行为识别。

RDT-MoE 的训练采用 2 阶段策略，以兼顾表征学习稳定性与专家建模能力。第 1 阶段旨在学习 RDT 模块与编码器参数，并获得后续专家初始化所需的 LoRA 参数。具体做法是在 ViT 骨干基础上引入 RDT 模块并调整池化结构，使用 VideoMAE 在 Kinetics-400 数据集上的预训练权重进行初始化。随后，在编码器各层 MLP 中嵌入 LoRA 结构，并在目标任务数据集上进行微调训练，从而得到与暴力行为识别任务相关的轻量化适配参数。第 2 阶段在第 1 阶段模型的基础上引入 MoE 模块，保留并冻结已训练好的 RDT 模块与编码器，以维持稳定的底层时空表征。同时，将第 1 阶段学习得到的 LoRA 参数迁移

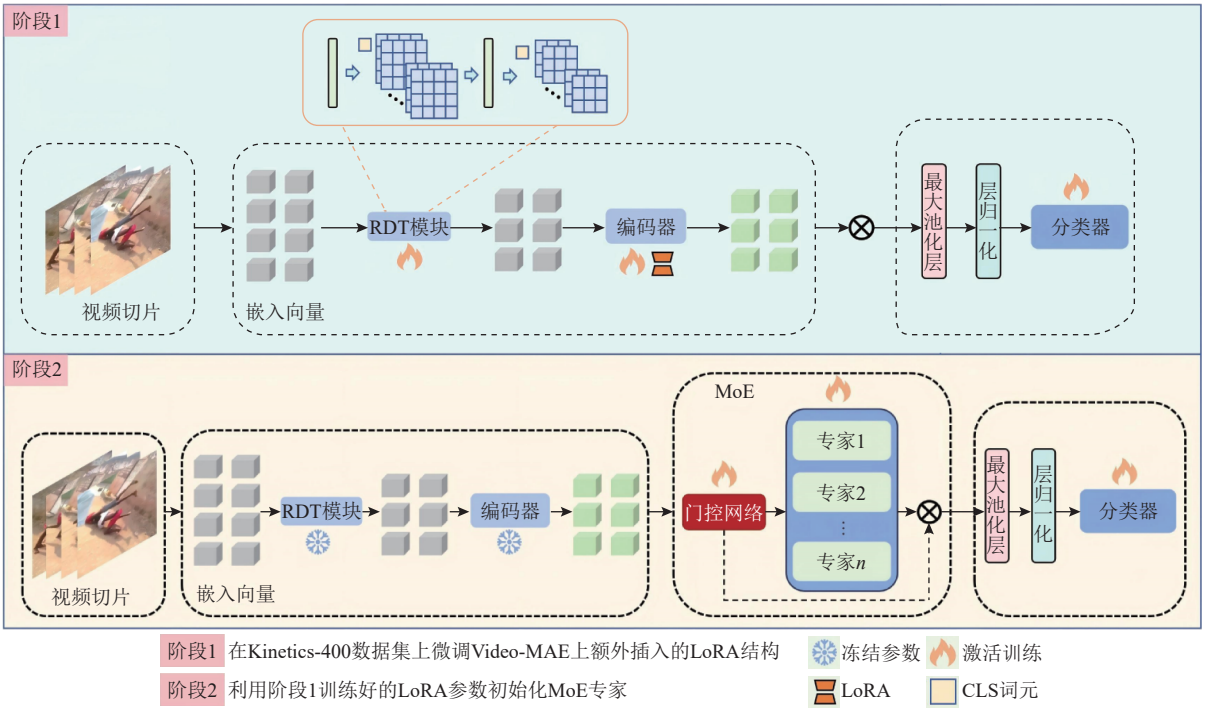


Fig. 5 Framework of RDT-MoE model

图 5 RDT-MoE 模型框架

用于各专家的初始化，以提供任务相关的先验能力。在此基础上，对 MoE 模块进行微调训练，使专家分工与门控机制逐步适配暴力行为识别任务，最终实现更高效且更具泛化性的分类性能。

2.3 冗余削减 RDT 模块

RDT 模块的每一层由 DST-Transformer(divide-space-time Transformer)[17] 块和词元选择策略组成，如图 6 所示。

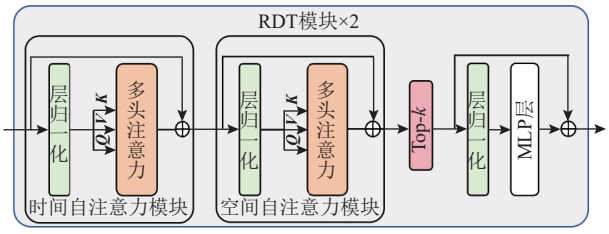


Fig. 6 Structure of RDT module

图 6 RDT 模块结构

DST-Transformer 是基于 Transformer 架构的一种变体，通过分解时空注意力来高效处理视频时空数据，其核心思想是将空间和时间维度的注意力计算分离来降低计算量，同时保持视频数据中时序信息的捕捉能力。其中时间注意力和空间注意力的区别是时间注意力处理时间维度上的词元依赖，而空间注意力是通过对每一帧都引入 CLS 词元来捕获空间维度上词元之间的联系并引导 Top-k 选择。其时间注意力的计算公式为：

$$Attention_{\text{tmp}} = Softmax\left(\frac{\boldsymbol{Q}_{\text{tmp}} \cdot \boldsymbol{K}_{\text{tmp}}^{\text{T}}}{\sqrt{d_{\text{key}}}}\right)\boldsymbol{V}_{\text{tmp}} 。 \quad (1)$$

假设含一个全局 CLS 词元的原始输入视频块序列为 $\boldsymbol{X}_{\text{tmp}} \in \mathbb{R}^{B\times(1+T\times H\times W)\times C}$，$B$ 是批大小，T 是帧数，H 是空间高度方向的块词元数，W 是空间宽度方向的块词元数，C 是通道数(也就是每个词元的特征长度)。为在时间维度上建模局部区域的动态变化，时间注意力并非直接作用于全部时空词元，而是在固定空间位置的跨帧词元序列上进行计算。具体地，将去除 CLS 词元后的特征 $\boldsymbol{X}_{\text{tmp}} \in \mathbb{R}^{B\times(T\times H\times W)\times C}$ 重排为 $\boldsymbol{X}'_{\text{tmp}} \in \mathbb{R}^{(B\times H\times W)\times T\times C}$，从而使每个空间位置对应一条长度为 T 的时序特征序列，并在该序列上施加时间注意力，以捕获其跨帧时序依赖关系。式(1)中的 $\boldsymbol{Q}_{\text{tmp}} = \boldsymbol{X}'_{\text{tmp}} \times \boldsymbol{W}_{\boldsymbol{Q}}^{\text{tmp}}$，$\boldsymbol{K}_{\text{tmp}} = \boldsymbol{X}'_{\text{tmp}}\boldsymbol{W}_{\boldsymbol{K}}^{\text{tmp}}$，$\boldsymbol{V}_{\text{tmp}} = \boldsymbol{X}'_{\text{tmp}}\boldsymbol{W}_{\boldsymbol{V}}^{\text{tmp}}$，其中 $\boldsymbol{W}_{\boldsymbol{Q}}^{\text{tmp}}$，$\boldsymbol{W}_{\boldsymbol{K}}^{\text{tmp}}$，$\boldsymbol{W}_{\boldsymbol{V}}^{\text{tmp}} \in \mathbb{R}^{C\times d_k}$，$d_{\text{key}}$ 为映射后每个注意力头内部参与相似度计算的向量长度，$d_{\text{key}} = C/h$，h 是注意力头数。之后需要将时间注意力的结果恢复形状，然后经过残差连接后再进行空间注意力的计算。其空间注意力计算公式为：

$$Attention_{\text{spc}} = Softmax\left(\frac{\boldsymbol{Q}_{\text{spc}} \cdot \boldsymbol{K}_{\text{spc}}^{\text{T}}}{\sqrt{d_{\text{key}}}}\right)\boldsymbol{V}_{\text{spc}} 。 \quad (2)$$

空间注意力是在每一帧内部进行的，所以重排后每一帧是一个样本，共 $B\times T$ 帧。每帧有 $H\times W$ 个视

频块序列，再插入 1 个可学习的 CLS 词元，因此长度变为 $H\times W+1$，得到 $\boldsymbol{X}'_{\text{spc}}\in\mathbb{R}^{(B\times T)\times(H\times W+1)\times C}$，其中 $\boldsymbol{Q}_{\text{spc}}=\boldsymbol{X}'_{\text{spc}}\boldsymbol{W}_{\boldsymbol{Q}}^{\text{spc}}$，$\boldsymbol{K}_{\text{spc}}=\boldsymbol{X}'_{\text{spc}}\boldsymbol{W}_{\boldsymbol{K}}^{\text{spc}}$，$\boldsymbol{V}_{\text{spc}}=\boldsymbol{X}'_{\text{spc}}\boldsymbol{W}_{\boldsymbol{V}}^{\text{spc}}$，其中 $\boldsymbol{W}_{\boldsymbol{Q}}^{\text{spc}}$，$\boldsymbol{W}_{\boldsymbol{K}}^{\text{spc}}$，$\boldsymbol{W}_{\boldsymbol{V}}^{\text{spc}}\in\mathbb{R}^{C\times d_k}$。

RDT 模块的冗余词元去除机制通过 4 个阶段流程实现高效特征选择：

1)输入的视频词元序列在时空维度解耦，被划分为空间主导组和时间主导组，这种划分使得模型能够独立分析时空维度的关联性。

2)引入 CLS 词元作为全局查询向量，通过跨时空自注意力机制计算每个空间-时间词元与 CLS 词元的注意力权重，此过程利用 CLS 词元的全局语义感知能力，生成反映各词元重要性的注意力分布。

3)在动态选择阶段，采用 Top-k 阈值筛选机制，仅保留注意力权重最高的前 k 个词元。这种硬选择策略直接去除低权重词元，显著降低后续计算复杂度。

4)为进一步优化特征质量，RDT 采用双级联处理：第 1 级进行粗粒度筛选，第 2 级则对剩余词元进行细粒度校准，通过局部注意力交互剔除次要信息。最终，筛选后的词元与类令牌融合，形成精简的时空特征表示。

2.4 MoE 网络

门控网络（gating network）在 RDT-MoE 网络中起到决定性的作用，它负责根据输入特征判断和选择最合适的专家来进行任务处理。在暴力行为识别任务中，视频数据的复杂性和多样性要求模型能够灵活地选择合适的专家来处理不同场景下的特征，门控网络的核心功能是路由选择，它通过学习从输入视频帧中提取的特征来动态决定哪些专家应该被激活。

本文采用 Top-k 路由策略来决定哪些专家被激活，虽然会节省计算资源和提高效率，但是由于每次只选择 Top-k 个专家进行激活，可能会导致某些专家频繁地被激活，而有些几乎从不被激活，从而引发负载不均衡的问题，未被选中的专家可能会失去更新的机会，导致它们无法在模型中发挥作用。为了尽可能使得所有的专家都得到训练，本文引入了随机噪声。给定输入特征 $\boldsymbol{x}$，门控网络会输出 1 组概率 $\boldsymbol{p}$ 表示每个专家被选择的概率，计算公式为：

$$\boldsymbol{p}=TopK(Softmax(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}+\boldsymbol{\varepsilon}))\text{。}\tag{3}$$

假设有 E 个专家，其中 $\boldsymbol{p}=(p_1,p_2,\cdots,p_E)$ 是长度为 E 的概率向量，其中 p_i 表示第 i 个专家的选择概率；$\boldsymbol{W}$ 是门控网络的权重矩阵，维度为 $d\times E$，其中 d 是输入特征的维度；$\boldsymbol{b}$ 为偏置项，维度为 E；$\boldsymbol{\varepsilon}$ 是加入的噪声项，函数 *Softmax* 确保加了噪声之后进行归一化使得所有选择概率之和为 1，*TopK*操作选择 *Softmax* 后最大的 k 个专家，并将其他的值置 0。

RDT-MoE 模型中的每个专家通过对输入数据进行非线性变换和非线性激活以捕捉数据的高维特征，在多任务学习或多样化的数据下具备灵活性，本文采用并行的处理方式，使得多个专家能够并行处理不同部分的输入数据，提高了计算效率。结合路由机制，模型能够很好地选择合适的专家用于处理特定的任务或特征。

最终专家层的输出也是和路由机制相关联的，本文采用 Top-k 的方式进行路由的选择，最终可以将输出合并为 y，其计算公式为：

$$y=\sum_{j=1}^{k}p(i)_j\cdot y(i)_j\text{，}\tag{4}$$

其中 $p(i)$ 为对应专家的路由概率，$y(i)$ 为第 i 个专家的输出。这种结构可以使得模型能够根据输入的特征选择合适的专家，并且由于本文添加了噪声又使得其不会导致某些专家几乎不被激活，通过加权的方式将各个被选择的专家的输出融合在一起，从而实现更丰富的特征学习和更灵活的任务处理能力。

2.5 2 阶段 LoRA 训练策略

在第 1 阶段，该模型在引入 RDT 模块后，其编码器部分由 10 层 Transformer 块构成，每一层的前馈神经网络（MLP 层）作为构建专家的基础。其他操作都和之前一样，只是在使用预训练模型初始化之后需要对编码器中每一层的 MLP 模块进行 LoRA 封装，从而获得低秩参数。具体来说，对于 MLP 层中的权重矩阵 $\boldsymbol{W}$，本文引入 2 个低秩矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$，并将更新后的权重表示为 $\boldsymbol{W}'$，其计算公式为：

$$\boldsymbol{W}'=\boldsymbol{W}+\boldsymbol{B}\boldsymbol{A}\text{，}\tag{5}$$

其中矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的秩 r 远小于原矩阵的维度。通过这种方法，只需更新这 2 个低秩矩阵即可模拟权重的变化。在这一阶段，预训练模型的原始参数仍保持可更新状态，整个模型与 LoRA 模块共同训练，从而捕获与特定任务相关的微调信息。经过训练获得了一系列不同维度的 LoRA 参数，这些参数反映了每一层 MLP 在特定任务下的调整需求。

在完成第 1 阶段的训练后，模型中的编码器和 RDT 模块已经学到了丰富的特征和任务相关知识。为了进一步聚焦于下游任务的优化，在第 2 阶段，本文将冻结编码器和 RDT 模块的参数，只对门控网络和专家层进行微调。具体操作如下：

1)冻结模块。在第2阶段,本文将预训练模型中的编码器和RDT模块参数冻结,以保持其在预训练阶段所学到的知识。

2)微调门控网络和专家模型。利用第1阶段获得的LoRA参数作为专家模型的初始值,对MoE网络中的各专家进行微调,同时训练门控网络,优化专家选择策略,使模型能够更准确地分配计算资源,并提高最终的任务性能。

相比于端到端的联合训练,这种2阶段策略充分利用了第1阶段预训练模型在大规模数据上学习到的通用特征,通过收敛的LoRA低秩参数高效捕捉特定任务所需的调整信息。作为第2阶段专家的热启动权重,有效避免了MoE冷启动时的路由震荡问题,并确保了不同专家在初始状态即具备一定的任务适配性,从而提升了最终模型的泛化能力。

2.6 总体损失函数

在RDT-MoE网络中,损失函数的设计至关重要,它直接影响到模型的训练效果和专家的利用效率。RDT-MoE在计算专家的选择概率的时候引入了噪声来使得专家的选择尽可能的均衡,这里进一步引入负载均衡损失[41]来缓解MoE模型中专家使用不均衡的问题。

负载均衡损失的基本思想是,强制要求每个专家接受到的训练样本数量相对均衡,从而避免某些专家被过度训练而某些专家被忽视从而不被训练。定义上,辅助损失通常与专家的激活频率相关,如果第 i 个专家被选中的概率为 p_i,则其损失的计算公式为:

$$L_{\text{aux}} = \lambda \sum_{i=1}^{E} \left(p_i - \frac{1}{E} \right)^2 , \tag{6}$$

其中 λ 是一个超参数,用于平衡辅助损失和分类损失的权重;p_i 是第 i 个专家被选择的概率;E 为专家的总数。该损失鼓励每个专家的激活概率接近 $1/E$ 来促使路由的均衡,从而提高所有专家的利用率。

交叉熵损失函数被广泛用于分类任务,尤其适用于处理概率输出和标注数据间的差异,在本文的任务中有效地衡量了模型在暴力与非暴力行为识别方面的性能,其计算公式为:

$$L_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i) \right] 。 \tag{7}$$

总损失是模型训练过程中优化的目标,它是多个损失的加权,在RDT-MoE网络中,总损失由交叉熵损失和负载均衡损失共2个部分组成,具体如下:

$$L_{\text{total}} = L_{\text{CE}} + L_{\text{aux}} , \tag{8}$$

其中 L_{CE} 为交叉熵损失,用于衡量模型分类任务的表现,L_{aux} 为负载均衡损失,旨在促进专家的均衡激活。

通过结合这2部分损失,模型不仅能够提高暴力行为识别的分类性能,还能在训练过程中保持专家的均衡使用,防止出现过度饱和或专家忽视的情况,确保了模型能够在复杂视频数据的暴力行为识别任务中取得更好的效果。

3 实 验

3.1 实验设置

为兼顾可重复性与工程落地需求,本文在自建数据集DVAD与公开数据集RWF-2000[38]上采用完全一致的“2阶段-3划分”实验流程,具体如下:

1)数据划分

① 按照7∶1∶2比例将原始片段随机划分为训练集、验证集与测试集,划分前对类别标签进行分层采样,以确保暴力和非暴力样本比例在3个子集中保持一致;

② 对训练集进一步执行“场景级”随机打乱,避免连续采集的同一摄像头片段集中进入网络,从而降低过拟合风险。

2)预处理与增强

① 统一采样16帧,间隔帧率自适应(25 fps视频按 Δt=4抽帧,30 fps视频按 Δt=5抽帧),以约5 s片段覆盖主要动作区间;

② 空间尺寸缩放至224×224,采用“中心裁剪+随机左右翻转”作为基础增强,并在训练阶段以50%概率启用随机亮度与对比度扰动(亮度系数 $L \in [0.8, 1.2]$,对比度系数 $C \in [0.9, 1.1]$),以提升模型对光照变化的鲁棒性;

③ 验证与测试阶段仅执行中心裁剪,不使用颜色扰动。

3)第1阶段训练(RDT+编码器)

① 使用VideoMAE预训练权重初始化骨干模型,冻结前2层ViT参数5轮,并在Kinetics-400上进行微调热身,随后解冻全部参数;

② 在编码器的10层MLP块中嵌入LoRA结构(LoRA秩 $rank$ = 32,缩放系数 α = 64),与主干网络联合训练;

③ 采用AdamW优化器(初始学习率 1×10^{-4},经余弦退火至 1×10^{-6}),批次大小设为8,训练150轮;

④ 每10轮在验证集上评估1次,以 $F1$ 分数最高

的权重保存 LoRA 参数，用于第 2 阶段专家初始化。

4）第 2 阶段训练（MoE 微调）

① 载入第 1 阶段最优 LoRA 参数，将其复制至 10 个专家子网络作为初始化；

② 冻结 RDT 与编码器全部参数，仅训练门控网络及专家内 LoRA-B 矩阵；

③ 继续使用 AdamW，学习率分别设为 5×10^{-5}（门控层）与 1×10^{-4}（专家层），训练 50 轮，早停窗口设置为 7；

④ 设置负载均衡损失系数 $\lambda=1\times10^{-2}$，与交叉熵损失联合优化，确保专家利用率大于 85%。

5）测试/推理阶段

① 使用单帧中心裁剪、无增强方式进行前向推理；

② 每次仅激活 4 个专家，单次推理耗时小于 5 ms（RTX-3090，FP16 模式）；

③ 若片段级概率大于 0.5 则判定为暴力行为，输出类别标签与置信度，并记录每帧专家分配情况用于后续可解释性分析。

通过上述标准化流程，RDT-MoE 在 DVAD 与 RWF-2000 上分别取得 94.67% 与 91.50% 的准确率，跨数据集性能差异小于 3%，验证了该流程的鲁棒性与跨域可迁移性。

3.2 参数配置

为便于复现与对比，本文将实验配置按“基础参数—训练—测试”3 阶段分别汇总。以下表格统一描述了 DVAD 与 RWF-2000 数据集在相同硬件环境下的完整超参数设置，后续所有实验均在表 2 所列条件下执行，表 3 为训练阶段模型设置参数，表 4 为推理阶段模型设置参数，所有的实验结果都进行 5 次取平均值，以确保结果的可重复性与公平性。

Table 2 Experimental Environment Settings

表 2 实验环境设置

类别	参数	取值
软件环境	PyTorch	1.11.0
	CUDA	11.3
	Python	3.8
硬件环境	GPU	RTX 3090（24 GB）
输入	采样帧数	16（均匀间隔）
	帧尺寸	224×224，RGB
优化器	类型	AdamW (β_1=0.9，β_2=0.999，ε=1E–8)
	批次大小	8
	权重衰减	1E–4

Table 3 Model Parameters in Training Stage

表 3 训练阶段模型参数

参数	取值	说明
总训练批次	150	含 5 个 epoch 的 warm-up
初始学习率	1E–4	余弦退火至 1E–6
LoRA rank	32	α=64，*dropout*=0.1
RDT 保留率 R	0.75	双层叠加后约为 0.56
数据增强	随机翻转+亮度/对比度扰动	仅训练阶段启用

Table 4 Model Parameters in Inference Stage

表 4 推理阶段模型参数

参数	取值	说明
推理帧数	16	与训练一致
增强方式	中心裁剪	无颜色扰动
专家激活	4	稀疏路由
精度模式	FP16	单次推理小于 5 ms

3.3 对比方法

为验证本文方法有效性，选取以下代表性模型进行对比，涵盖 CNN，Transformer 与混合架构：

1）Flow Gated Network[38]。通过引入光流门控机制融合运动信息，有效增强了模型对暴力行为的动态特征感知能力。

2）SepConvLSTM[24]。提出一种高效的双流网络结构，利用可分离卷积与长短时记忆单元（LSTM）联合建模外观与运动特征，用于暴力行为识别。

3）TimeSFormer[26]。基于纯 Transformer 架构，联合建模空间与时间注意力机制，验证了注意力机制在视频理解任务中的有效性。

4）VideoMAE[3]。本文选择的基线方法，采用掩码自监督预训练，具备强时空表征能力。

5）ST-GCN[27]。开创性地将时空图卷积网络应用于骨架动作识别，通过图结构建模人体关键点间的空间依赖与时间动态。

6）SPIL[51]。基于 3D 骨架点云的人体交互学习方法，通过空间几何关系建模实现更稳健的暴力行为识别。

7）HSCD[52]。结合人体骨架建模与变化检测策略，在监控视频中高效识别暴力行为，兼顾识别性能与计算效率。

8）Prompt-Guided Zero-Shot[53]。基于预训练深度骨架特征的提示引导零样本学习方法，无需额外训练即可实现异常与暴力行为识别。

9）BrutNet[54]。结合深度卷积神经网络（DCNN）

与门控循环单元(GRU)的混合模型，通过联合学习空间与时间特征，实现暴力行为的精确检测与分类。

10)RDTNet。指在 VideoMAE 基线模型基础上集成 RDT 模块，并采用标准 MLP 作为分类头的网络变体。该模型未引入 MoE 架构，旨在作为消融基线以量化 MoE 模块的性能增益。

所有方法均在 RWF-2000[38] 与 DVAD 上统一复现，输入帧数、分辨率与训练策略保持一致，确保公平对比。

3.4 主要结果

表 5 和表 6 分别在公开数据集 RWF-2000 与自建数据集 DVAD 上验证了 RDT-MoE 模型的优越性能。为确保比较的公平性，本文选取了暴力行为识别及通用行为识别领域中具有代表性的模型，包括双流网络架构、3D 卷积神经网络架构、骨架模型以及 Transformer 架构。所有方法均采用最常见的 16 帧均匀采样作为输入。光流特征用于补充运动信息，骨骼特征虽然能够刻画人体结构，但在遮挡严重或场景复杂的情况下表现不稳定，因此在暴力行为识别任务中存在一定局限。

Table 5 Compared with Existing Methods on RWF-2000

表 5 在 RWF-2000 上与现有方法的对比

模型名称	是否使用额外数据	输入形式	采样帧数	参数量	准确率/%
Flow Gated Network[38]		R+F		**0.20×10⁶**	87.25
SPIL[51]		S			89.30
ST-GCN[27]		S			83.30
SepConvLSTM[24]		R+F	32	0.30×10⁶	89.75
TimeSFormer[26]	Kinetics-400	R		121.00×10⁶	89.30
HSCD[52]		R+S	50		90.25
VideoMAE[3]	Kinetics-400	R	16	86.23×10⁶	89.50
Prompt-Guided Zero-Shot[53]	Kinetics-400	R			90.30
BrutNet[54]		R	24	3.42×10⁶	86.43
RDTNet（本文）	Kinetics-400	R	16	92.14×10⁶	90.50
RDT-MoE（本文）	Kinetics-400	R	16	48.42×10⁶	**91.50**

注：R 表示图像帧，F 表示光流信息，S 表示骨骼信息。黑体数值表示最优值。

Table 6 Compared with Existing Methods on DVAD

表 6 在 DVAD 上与现有方法的对比

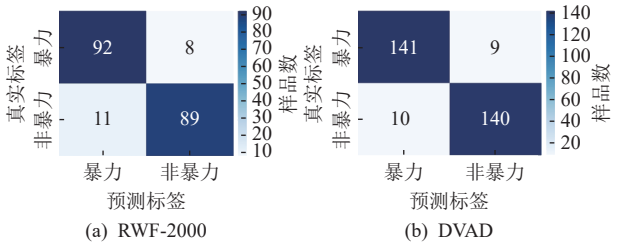
模型	数据集名称	输入形式	采样帧数	参数量	准确率/%
VideoMAE[3]	Kinetics-400	R	16	86.23×10⁶	92.33
RDTNet（本文）	Kinetics-400	R	16	92.14×10⁶	93.67
RDT-MoE（本文）	Kinetics-400	R	16	**48.42×10⁶**	**94.67**

注：R 表示图像帧。黑体数值表示最优值。

实验结果表明，RDT-MoE 在仅使用 RGB 单模态输入的条件下的准确率仍优于多数多模态方法。其性能优势主要得益于多专家动态协作机制：通过 Top-k 路由策略激活最相关的专家，使模型能够在单模态条件下覆盖暴力行为的多样化特征，从而避免多模态融合带来的结构复杂度与计算开销。如表 5 所示，RDT-MoE 相较于 RDTNet 在 RWF-2000 数据集上准确率进一步提升，验证了 MoE 架构在处理复杂暴力特征时的优势。轻量化设计使模型参数量较 RDTNet 减少约 50%，并低于多数主流模型。这一优势源于参数冻结策略与 LoRA 初始化的协同优化机制：前者抑制冗余参数更新，后者提升专家间的多样性。跨数据集的稳定性能提升进一步验证了模型的强泛化能力，表明 RDT-MoE 能够在复杂场景下保持良好的鲁棒性与适应性。

综上，RDT-MoE 为暴力行为检测任务提供了一种兼具高精度、轻量化与易部署特性的可扩展方案，验证了在单模态模型中通过特征解耦替代多模态融合的可行性。

如图 7 所示，通过对比 RDT-MoE 模型在公开数据集 RWF-2000 与自建数据集 DVAD 上的混淆矩阵，揭示了模型改进与数据特性的关联。在 RWF-2000 上，RDT-MoE 对暴力样本的误判率显著降低，表明 MoE 网络的专家分工机制强化了对边界样本的细粒度特征提取能力；然而，非暴力样本的误判率未显著改善，主要受限于该数据集的样本分布不均导致模型难以充分学习其判别性特征。相比之下，在自建数据集 DVAD 上，RDT-MoE 的暴力与非暴力误判率同步下降，且分类均衡性显著提升，这得益于 DVAD 的多样性设计迫使模型通过动态路由激活差异化专家，从而增强对挑战性样本的鲁棒性。结果表明，数据质量与模型架构的协同优化是暴力检测任务的关键，自建数据集的丰富性为模型提供更全面的训练环境，而公开数据集的局限性需通过数据增强或损失函数调整进一步突破。



(a) RWF-2000 (b) DVAD

Fig. 7 Confusion matrices of RDT-MoE on datasets

图 7 RDT-MoE 在数据集上的混淆矩阵

3.5 消融实验

为验证各模块对整体性能的真实增益，并厘清超参数的影响边界，本文围绕“预训练来源-池化方式-词元保留率-专家配置”展开系统消融。每项实验仅改动单一变量，其余设定保持与表 2~4 相同，以回答以下关键问题：

1）不同预训练源对跨域迁移的贡献差异如何；

2）最大池化为何更适合暴力边缘特征；

3）RDT 压缩比如何权衡速度与精度；

4）MoE 中“专家结构/数量/路由 Top-k”对多样类内行为的敏感度。

迁移预训练模型可以让模型学习更多的通用特征表示能力，表 7 对比了分别在 UCF101[55]，SSV2[56]，Kinetics-400[12] 数据集上进行预训练，然后向暴力检测任务的迁移性能。UCF101 因含暴力片段，迁移效果优于 SSV2；而 Kinetics-400 凭借多场景覆盖和 400 类细粒度行为分类，使模型学习到更普适的时空特征，达到最优性能。这表明预训练数据与目标任务的语义适配性至关重要：Kinetics-400 通过细粒度分类隐式强化暴力相关特征的提取能力，而 SSV2 的时序组合逻辑与暴力行为的突发性特征存在本质差异。因此，暴力检测模型需优先选择数据分布相似性高且能促进特征解耦的预训练策略，而非单纯追求数据规模。

Table 7 Performance Comparison of Different Pretraining Datasets

表 7 不同预训练数据集性能对比

预训练数据集	数据量	迁移数据集准确率/%	
		RWF-2000	DVAD
UCF101	9 500	88.75	92.00
SSV2	169 000	87.25	91.67
Kinetics-400	240 000	89.50	92.33

表 8 分析了最大池化（max pooling）与平均池化（average pooling）在暴力行为识别任务中的性能差异。最大池化在 RWF-2000 和 DVAD 数据集上均取得更高准确率，其优势源于暴力行为的关键局部特征敏感性：最大池化通过保留特征图局部最大值，能有效捕捉突发性剧烈动作的显著信号，同时抑制背景噪声干扰；而平均池化因对局部特征取均值，易弱化短时高强度运动的判别性细节。暴力行为的识别更依赖稀疏强响应特征的提取，而非全局平缓特征的表征。因此，最大池化因其对突变信号的强化能力，成为暴力检测任务的优化选择。

Table 8 Accuracy of Different Pooling Strategies

表 8 不同池化策略的准确率 %

池化策略	RWF-2000	DVAD
最大池化	90.00	92.67
平均池化	89.50	92.33

表 9 分析了词元保留率（R）对暴力检测模型性能的影响。当 R=0.75 时，模型在 RWF-2000 和 DVAD 数据集上达到峰值准确率，而 R=0.9 时性能下降，表明适度稀疏化可抑制冗余词元干扰。由于模型采用双层 RDT 模块，实际词元保留率为 R^2（如 R=0.75 时仅保留 56.25%），而 R=0.5 时过度丢弃关键动作特征导致了性能显著降低。实验揭示计算效率与特征完整性需动态平衡：较低 R 值通过剪枝冗余信息加速训练，同时保留判别性局部特征；而过高 R 值因保留过多非显著词元，引入噪声导致性能衰退。当前最优 R 值尚未饱和，未来可通过任务自适应调整进一步提升。

Table 9 Accuracy of Different Token Retention Ratios

表 9 不同词元保留比率的准确率

R	准确率/%	
	RWF-2000	DVAD
0.50	88.50	91.17
0.75	**90.50**	**93.67**
0.90	89.25	92.83

注：黑体数值表示最优值。

表 10 分析验证了 RDT 模块对暴力行为识别模型的优化效果。引入 RDT 模块后，模型在 RWF-2000 数据集上准确率提升 0.5 个百分点，训练时间缩短 40%，计算量降低 24.5%。该模块通过动态词元剪枝（保留率 R=0.75）去除 50% 冗余特征，使模型聚焦关键动作片段，同时减少背景噪声干扰。尽管特征量大幅缩减，但性能仍优于基线，证明暴力识别任务中稀疏强响应特征的主导性。实验还表明，词元保留率需平衡特征完整性与计算效率，过高或过低均会导致性能下降。未来可通过自适应调整 R 值进一步优化，为实时暴力检测部署提供更高效的解决方案。

Table 10 Effectiveness Analysis of RDT Module

表 10 RDT 模块的有效性分析

RDT 模块	训练批次	准确率/%	训练时间/h	训练速度/GFLOPs
×	150	90.00	5	102
√	150	90.50	3	77

注：GFLOPs 在相同输入设置下按每个视频片段进行统计。

表 11 实验通过对比线性(linear)层与 MLP 作为 MoE 专家模型的性能,揭示了模型结构与训练策略的协同优化机制。MLP 因非线性特征提取能力,准确率显著优于线性层,其可建模暴力行为中复杂的时空交互模式;冻结训练策略通过固定部分参数,在压缩 MLP 参数量 65.3% 的同时提升准确率至 94.17%,表明其能强化路由网络的特征分配效率;LoRA 参数初始化进一步差异化专家能力,使冻结后的 MLP 模型准确率达 94.67%。三者联合优化表明,暴力行为识别需兼顾非线性建模、高效特征筛选与专家多样性。

Table 11 Impact Analysis of the Expert Models and Training Strategies

表 11 专家模型及训练策略的影响分析

专家模型	LoRA 参数初始化	训练技巧	参数量	准确率/%
Linear	×		98.05×10^6	92.17
	×	冻结	7.10×10^6	92.67
MLP	×		139.37×10^6	94.00
	×	冻结	48.42×10^6	94.17
	√		139.37×10^6	94.33
	√	冻结	48.42×10^6	94.67

表 12 对比了 Softmax 与 Top-k 这 2 种路由策略对暴力检测性能的影响。Top-k 策略的准确率显著优于 Softmax,其核心优势在于稀疏专家激活机制:仅动态筛选与输入特征最相关的 k 个专家,抑制次要专家的干扰噪声。暴力行为的判别依赖关键特征子集的精准提取,而 Softmax 的全专家加权融合易受弱相关特征干扰,导致判别性下降。Top-k 策略通过硬性筛选,迫使模型强化对高响应专家输出的依赖,从而提升特征解耦能力。

Table 12 Impact Analysis of Routing Strategies

表 12 路由策略的影响分析

路由策略	准确率/%
Softmax	93.50
Top-k	94.67

表 13 验证了 Top-k 路由策略中不同激活专家数(k)对暴力行为识别性能的影响。当 k=4 时模型达到峰值准确率,而 k 值过小或过大均导致性能下降。k=2 时,激活的专家数量不足,难以覆盖暴力行为的多维度特征,造成关键信息缺失;k=6 并逐渐增大至 k=8 时,冗余专家的引入导致特征冲突,弱化暴力判别边界。k=4 的设定平衡了特征多样性与噪声抑制:既激活足够专家以捕捉暴力核心线索,又避免过度稀释模型对关键特征的聚焦。

表 14 探究了 MoE 网络中专家数量对暴力行为识别性能的影响。结果显示,专家数量为 10 时模型准确率最高,而数量过少或过多均导致性能衰退。专家不足时,模型难以覆盖暴力行为的多维度特征,导致关键判别信息缺失;专家过量时,则因训练样本分配稀疏化和负载不均衡,引发冗余噪声干扰。最佳专家数量的设定平衡了模型容量与训练稳定性:既保证专家专业化,又避免资源分散导致的特征稀释。

Table 13 Analysis of k Values in Top-k

表 13 Top-k 中 k 的取值分析

k	准确率/%
2	93.67
4	94.67
6	94.50
8	94.17

Table 14 Performance Analysis with Different Numbers of Experts

表 14 不同专家数量下的性能分析

专家数量	准确率/%
5	94.00
10	94.67
20	94.33

图 8 通过对比负载均衡策略优化前后各专家的词元处理占比,揭示了路由策略改进的关键机制。改进前,门控网络存在显著专家激活偏向(如 10 号专家处理占比接近 0),导致部分专家长期闲置,限制模型特征多样性。改进策略通过引入随机噪声打破门控分数固化,并施加负载均衡损失约束激活分布,使

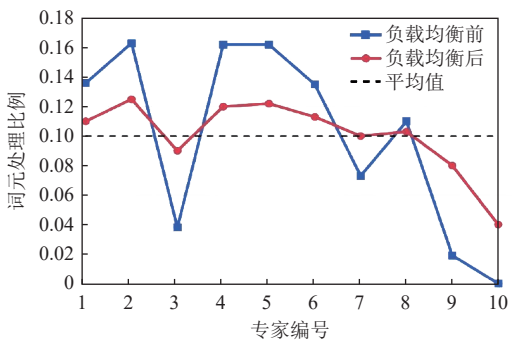


Fig. 8 Token distribution before and after expert load balancing

图 8 专家负载均衡前后的词元分布

各专家处理占比趋于均衡。此优化促使所有专家参与特征解析，覆盖暴力行为的细粒度模式，同时避免高频专家过载引发的训练不稳定问题。均衡化策略显著提升了模型在复杂场景下的鲁棒性，支撑了整体识别准确率的提升。

3.6 可视化分析

为对实验结果进行更直观、全面的评估，本节从测试集中选取若干具有代表性的错误样例进行可视化展示，重点覆盖 2 类典型误差：非暴力误判为暴力（FP）与暴力漏检（FN）。通过逐帧对比观察这些样例在不同场景与成像条件下的表现特征，并与前述误差因素统计结果进行对照，定性验证错误归因的合理性。

如图 9 所示，从 FP 样例可以观察到，误报通常出现在“动作相似”与“背景干扰”更为突出的片段中：例如打闹、摔跤、庆祝、快速挥臂/投掷等高动态行为在局部运动模式上与攻击动作高度相似，容易触发模型的暴力判别；当场景中存在复杂背景、遮挡或多人同时运动时，关键判别线索更易被非关键运动所掩盖，从而进一步加剧误判。



Fig. 9 Error examples of non-violent samples misclassified as violent samples

图 9 非暴力样本误判为暴力样本的错误示例

相对地，FN 样例更多集中在“光照不足”“小目标/远距离”等低质量视频条件下：夜间监控、远景拍摄往往导致人物尺度较小、动作细节模糊且伴随压缩噪声，使得关键肢体接触与攻击动作难以被稳定捕获，如图 10 所示。同时，多人拥挤带来的遮挡与交互混杂也会削弱可辨识线索，造成漏检风险上升。

总体而言，非暴力误判为暴力更集中于相似动作与背景干扰，暴力漏检则更受低照度与小目标信息不足的影响，而“多人拥挤/遮挡”对这 2 类错误均具有显著作用。

3.7 局限性分析

尽管 RDT-MoE 在多个数据集上取得了较好的识别性能，但结合前述统计与可视化结果可以看出，



Fig. 10 Sample example of missing detection of violence action

图 10 暴力行为漏检样本示例

该方法在真实复杂场景中仍存在一定局限性。1）在动作语义边界模糊的情况下，模型对“暴力相似动作”的区分仍不够稳健，例如打闹、摔跤、庆祝、快速挥臂等高动态行为在运动模式上与攻击动作高度接近，容易触发误报，说明当前表征对细粒度意图差异（攻击性/非攻击性）与交互语义的刻画仍有限。2）模型对低质量视频的鲁棒性有待提升：夜间监控、光照不足、模糊与压缩噪声，以及远距离拍摄导致的小目标会显著削弱关键动作细节，使暴力行为更易被漏检，表明在信息缺失与域偏移（成像条件变化）下模型的稳定性仍受挑战。3）多人拥挤与遮挡对 2 类错误均具有显著影响，当多人交互同时发生、遮挡频繁或目标间关联复杂时，模型难以持续锁定冲突中心并准确判断交互关系，从而增加混淆与漏检风险。4）从数据与评测角度看，暴力识别任务本身存在一定长尾与标注边界问题：不同数据集对“暴力”的定义与标注口径可能存在差异，且部分边界样本具有主观性，这会在一定程度上影响模型泛化评估与跨场景迁移效果。

这些局限性表明，后续工作需要进一步关注相似动作下的细粒度判别、低照度与小目标条件下的鲁棒建模，以及拥挤交互场景中的关系理解与稳定识别能力。通过引入更强的时空细粒度建模以提升相似动作区分能力、结合多尺度与小目标增强策略提升远距离场景召回、通过光照鲁棒的数据增强与域适配提升低质量视频稳定性，以及面向拥挤场景引入交互关系建模与更稳健的特征选择机制，以推动模型在真实复杂环境中的泛化表现。

4 结　　论

本文结合 MoE 网络设计了高效的 RDT-MoE 模型，不仅显著减少了冗余特征对任务性能的干扰，还

有效缓解了单路径模型在捕获判别性特征方面的局限。为进一步提升模型的训练稳定性与专家多样性，本文设计了2阶段的训练策略：第1阶段训练得到LoRA参数，第2阶段利用该参数差异化初始化专家模型，使不同专家具备互补的特征处理能力，并通过冻结主干网络，使优化过程聚焦于门控机制与专家组合策略，从而在保持通用表征能力的同时强化模型对判别性特征的动态选择能力。门控网络采用Top-k稀疏激活策略，根据输入场景特征自适应选择最相关的专家组合并融合输出。综合实验结果表明，所提出的RDT-MoE在暴力行为识别任务中取得了优于现有方法的性能，并展现出良好的泛化性与扩展潜力，可推广至多视角行为分析、公共安全监控等实际应用场景。

作者贡献声明：杨建锋提出研究构思并指导研究开展，审阅和修改论文；于大为负责方法实现、模型搭建、实验与数据分析，并撰写论文；曹瑞审阅和修改论文，统筹审稿意见回复并负责最终定稿；任佳航参与论文审阅与修改；王平参与论文审阅与修改；胡琦瑶提出改进思路，并提供方法优化建议与分析支持。

参 考 文 献

[1] Mumtaz N, Ejaz N, Habib S, et al. An overview of violence detection techniques: Current challenges and future directions[J]. Artificial Intelligence Review, 2023, 56(5): 4641−4666

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2017: 5998–6008

[3] Tong Zhan, Song Yibing, Wang Jue, et al. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training[C]//Proc of the 36th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 10078−10093

[4] Sudhakaran S, Lanz O. Learning to detect violent videos using convolutional long short-term memory[C]//Proc of the 14th IEEE Int Conf on Advanced Video and Signal Based Surveillance. Piscataway, NJ: IEEE, 2017: 1−6

[5] Wu Qian, Cui Ruoxuan, Li Yuke, et al. Haltingvt: Adaptive token halting Transformer for efficient video recognition[C]//Proc of the 49th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2024: 4305−4309

[6] Climent-Perez P, Florez-Revuelta F. Improved action recognition with separable spatio-temporal attention using alternative skeletal and video pre-processing[J]. Sensors, 2021, 21(3): 1005−1019

[7] Korbar B, Tran D, Torresani L. Scsampler: Sampling salient clips from video for efficient action recognition[C]//Proc of the 17th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6232−6242

[8] Liu Mingyu, Huang Xun, Yu Jiahui, et al. Generative adversarial networks for image and video synthesis: Algorithms and applications[J]. Proceedings of the IEEE, 2021, 109(5): 839−862

[9] Wang Yu, Zhao Shengjie. Weakly-supervised action localization by hierarchical attention mechanism with multi-scale fusion strategies[C]//Proc of the 25th IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2024: 1−6

[10] Chen Ting, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//Proc of the 37th Int Conf on Machine Learning. Cambridge, MA: MIT, 2020: 1597−1607

[11] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proc of the 15th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2015: 4489−4497

[12] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. arXiv preprint, arXiv: 1705.06950, 2017

[13] Lepikhin D, Lee H J, Xu Y, et al. GShard: Scaling giant models with conditional computation and automatic sharding[J]. arXiv preprint, arXiv: 2006.16668, 2020

[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[C/OL]// Proc of the 9th Int Conf on Learning Representations. 2021[2026-02-10]. https://openreview.net/forum?id=YicbFdNTTy

[15] Hu E, Shen Yelong, Wallis P, et al. Lora: Low-rank adaptation of large language models[C/OL]//Proc of the 10th Int Conf on Learning Representations. 2022[2026-02-10]. https://openreview.net/forum?id=nZeVKeeFYf9

[16] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Proc of the 27th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1725−1732

[17] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Proc of the 28th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 568–576

[18] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// Proc of the 14th European Conf on Computer Vision. Cham: Springer, 2016: 20−36

[19] Feichtenhofer C, Fan Haoqi, Malik J, et al. SlowFast networks for video recognition[C]//Proc of the 17th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6202−6211

[20] Carreira J, Zisserman A. Quo vadis, Action recognition? a new model and the kinetics dataset[C]//Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6299−6308

[21] Tran D, Wang Heng, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6450−6459

[22] Hu Kai, Zheng Fei, Weng Liguo, et al. Action recognition algorithm of spatio-temporal differential LSTM based on feature enhance-

ment[J]. Applied Sciences, 2021, 11(17): 7876

[23] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proc of the 28th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 2625−2634

[24] Islam Z, Rukonuzzaman M, Ahmed R, et al. Efficient two-stream network for violence detection using separable convolutional LSTM[C]//Proc of the 31st Int Joint Conf on Neural Networks. Piscataway, NJ: IEEE, 2021: 1−8

[25] Arnab A, Dehghani M, Heigold G, et al. VIVIT: A video vision transformer[C]//Proc of the 18th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 6836−6846

[26] Bertasius G, Wang Heng, Torresani L. Is space-time attention all you need for video understanding?[C]//Proc of the 38th Int Conf on Machine Learning. Cambridge, MA: MIT, 2021: 1−4

[27] Yan Sijie, Xiong Yuanjun, Lin Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Proc of the 32nd AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI. https://ojs.aaai.org/index.php/AAAI/article/riew/12328

[28] Chen Yuxin, Zhang Ziqi, Yuan Chunfeng, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]//Proc of the 29th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 13359−13368

[29] Li Maosen, Chen Siheng, Chen Xu, et al. Actional-structural graph convolutional networks for skeleton-based action recognition[C]// Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3595−3603

[30] Plizzari C, Cannici M, Matteucci M. Skeleton-based action recognition via spatial and temporal transformer networks[J]. Computer Vision and Image Understanding, 2021, 208: 103219

[31] Lee P, Uh Y, Byun H. Background suppression network for weakly-supervised temporal action localization[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 11320−11327

[32] Wu Peng, Zhou Xuerong, Pang Guansong, et al. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 6074−6082

[33] Fan Haoqi, Xiong Bo, Mangalam K, et al. Multiscale vision transformers[C]//Proc of the 18th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 6824−6835

[34] Faure G, Chen M, Lai S. Holistic interaction transformer network for action detection[C]//Proc of the 23rd IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2023: 3340−3350

[35] Pan Junting, Lin Ziyi, Zhu Xiatian, et al. ST-Adapter: Parameter-efficient image-to-video transfer learning[C]//Proc of the 36th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 26462−26477

[36] Bermejo N, Deniz S, Bueno G, et al. Violence detection in video using computer vision techniques[C]//Proc of the 14th Int Conf on Computer Analysis of Images and Patterns. Berlin: Springer, 2011: 332−339

[37] Hassner T, Itcher Y, Kliper-Gross O. Violent flows: Real-time detection of ViolentCrowd behavior[C]//Proc of 25th IEEE Computer Society Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2012: 1−6

[38] Cheng Ming, Cai Kunjing, Li Ming. RWF-2000: An open large scale video database for violence detection[C]//Proc of the 25th Int Conf on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2021: 4183−4190

[39] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts[J]. Neural Computation, 1991, 3(1): 79−87

[40] Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer[J]. arXiv preprint, arXiv: 1701.06538, 2017

[41] Fedus W, Zoph B, Shazeer N. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(1): 5232−5270

[42] Riquelme C, Puigcerver J, Mustafa B, et al. Scaling vision with sparse mixture of experts[C]//Proc of the 35th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 8583−8595

[43] Hwang Changho, Cui Wei, Xiong Yifan, et al. Tutel: Adaptive mixture-of-experts at scale[J]. Proceedings of Machine Learning and Systems, 2023, 5: 269−287

[44] Zoph B, Bello I, Kumar S, et al. ST-MoE: Designing stable and transferable sparse expert models[J]. arXiv preprint, arXiv: 2202.08906, 2022

[45] Zhou Yanqi, Lei Tao, Liu Hanxiao, et al. Mixture-of-experts with expert choice routing[C]//Proc of the 36th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 7103−7114

[46] Zhou Yanqi, Du Nan, Huang Yanping, et al. Brainformers: Trading simplicity for efficiency[C]//Proc of the 40th Int Conf on Machine Learning. Cambridge, MA: MIT, 2023: 42531−42542

[47] Lin Bin, Tang Zhenyu, Ye Yang, et al. MoE-LLaVA: Mixture of experts for large vision-language models[J]. arXiv preprint, arXiv: 2401.15947, 2024

[48] Liu Aixin, Feng Bei, Wang Bin, et al. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model[J]. arXiv preprint, arXiv: 2405.04434, 2024

[49] Aktı Ş, Ofli F, Imran M, et al. Fight detection from still images in the wild[C]//Proc of the 22nd IEEE Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2022: 550−559

[50] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition[C]//Proc of the 18th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2011: 2556-2563.

[51] Su Yukun, Lin Guosheng, Zhu Jinhui, et al. Human interaction learning on 3D skeleton point clouds for video violence recognition[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 74−90

[52] Garcia-Cobo G, SanMiguel J C. Human skeletons and change

detection for efficient violence detection in surveillance videos[J]. Computer Vision and Image Understanding, 2023, 233: 103739

[53] Sato F, Hachiuma R, Sekii T. Prompt-Guided Zero-Shot anomaly action recognition using pretrained deep skeleton features[C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 6471−6480

[54] Haque M, Nyeem H, Afsha S. BrutNet: A novel approach for violence detection and classification using DCNN with GRU[J]. The Journal of Engineering, 2024, 2024(4): e12375

[55] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint, arXiv: 1212. 0402, 2012

[56] Goyal R, Ebrahimi K S, Michalski V, et al. The "something something" video database for learning and evaluating visual common sense[C]//Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 5842−5850



Yang Jianfeng, born in 1981. Master, senior engineer, master supervisor. His main research interests include software development methodologies, and data science and its applications in higher education.

杨建锋, 1981 年生。硕士，高级工程师，硕士生导师。主要研究方向为软件开发方法、数据科学及其在高校教育中的应用。



Yu Dawei, born in 1999. Master, engineer. His main research interest includes action recognition.

于大为, 1999 年生。硕士，工程师。主要研究方向为行为识别。



Cao Rui, born in 1991. PhD, lecturer, master supervisor. Member of CCF. His research interests include text representation learning, application optimization of generative large language models, and diffusion-based model artistic generation.

曹　瑞, 1991 年生。博士，讲师，硕士生导师。CCF 会员。主要研究方向为文本表示学习、生成式大语言模型应用与优化、扩散模型艺术生成。



Ren Jiahang, born in 2002. Master candidate. His main research interests include large model development and application, and network security.

任佳航, 2002 年生。硕士研究生。主要研究方向为大模型开发应用、网络安全。



Wang Ping, born in 2002. Master candidate. His main research interest includes large model visual reasoning.

王　平, 2002 年生。硕士研究生。主要研究方向为大模型视觉推理。



Hu Qiyao, born in 1992. PhD, engineer, master supervisor. Member of IEEE. Her main research interests include AI art and multimodal generation.

胡琦瑶, 1992 年生。博士，工程师，硕士生导师。IEEE 会员。主要研究方向为人工智能艺术、多模态生成。