ReMA：一种用于视频行为识别的无需训练的即插即用混合增强方法

Feng-Qi Cui[1,3], Jinyang Huang[2,†], Sirui Zhao[1],Jinglong Guo[2],Qifan Cai[2,3],Xin Yan[4],Zhi Liu[5]*1* 中国科学技术大学，中国合肥 *2* 合肥工业大学，中国合肥*3* 合肥小圣智能科技有限公司，中国合肥 *4* *Cylingo Group*，中国北京*5* 电气通信大学，日本东京

†通讯作者：hjy@hfut.edu.cn

arXiv:2601.00311v1 [cs.CV] 1 Jan 2026

摘要——**视频行为识别需要在复杂的时空变化下获得稳定且具有判别性的表征。然而，现有的视频数据增强策略大多仍由扰动驱动，往往引入不可控的变化，放大非判别性因素，最终削弱类内分布结构，并导致表征漂移在不同时间尺度上的效果不一致。为解决这些问题，我们提出表示感知混合增强（*ReMA*），这是一种即插即用的增强策略，将混合建模为受控的替换过程，以扩展表征，同时保持类别条件稳定性。*ReMA* 整合了两种互补机制。首先，表示对齐机制（*RAM*）在分布对齐约束下执行结构化类内混合，抑制无关的类内漂移，同时增强统计可靠性。随后，动态选择机制（*DSM*）生成运动感知的时空掩码，以定位扰动，引导其避开对判别敏感的区域，并促进时间连贯性。通过共同控制混合的方式与位置，*ReMA* 无需额外监督或可训练参数，即可提升表征鲁棒性。在多种视频行为基准上的大量实验表明，*ReMA*在不同的时空粒度下均能持续提升泛化能力和鲁棒性。**

索引词——**视频数据增强、表征感知混合、视频行为识别。**

一、引言

视频行为识别是计算机视觉中的一项基础问题，在人机交互、心理分析 [1], 和智能医疗保健等领域有着广泛的应用。随着研究不断推进，行为理解已从粗粒度的身体动作扩展到更精细的语义层面，例如面部表情 [2] 和细微的微动作 [3]。在这些任务中，模型需要学习具有判别性的表征，使其能够适应不同的时间尺度、空间分辨率和运动模式 [4]。然而，当前视频行为识别流程中的一项关键挑战，并非视频表征本身固有的不稳定性，而是缺乏对训练扰动如何塑造所学表征空间的显式控制。遗憾的是，在普遍采用的数据增强范式下，非判别性变化往往被不加区分地引入并逐步放大，这不可避免地会妨碍稳定且与任务相关的决策边界的形成 [5], [6]。具体而言，当增强效果未得到适当约束时，不受控的区域替换或插值可能会扭曲时间动态 [7],; 过度的执行风格变化可能导致类条件表征分散或漂移 [8], ; 随机或固定尺度的扰动则往往会放大特征空间中的高频冗余和局部噪声 [9]。这些因素共同导致非判别性变化在表征学习中占据主导地位，使决策边界在多样化场景和运动尺度下变得脆弱 [10]。

数据增强已被广泛用于提升泛化能力 [11]。除标准图像变换 [12], 外，混合增强（例如 Mixup [13] 和 CutMix [14] ）通过插值或区域替换 [15]扩展训练分布。然而，由于缺少结构感知约束 [16], 将这些策略直接扩展到视频往往会破坏时间连贯性和运动语义。尽管针对视频的混合策略已在动作识别 [17],中展现出有效性，但它们基本上侧重于扩展多样性，却忽视了类内统计一致性，因而生成的增强样本会偏离原始分布 [5]。此外，由于缺乏对相应运动结构的显式感知，扰动可能作用于关键动态区域，这对精细时间行为 [18]尤其有害。

因此，现有方法的局限与其说在于样本稀缺，不如说在于对扰动所带来的分布和时间副作用缺乏充分控制。许多方法仍以多样性为导向，依赖随机或固定扰动，而这些扰动对表征学习的影响在很大程度上未受约束 [19]。值得庆幸的是，自监督视频学习的最新证据进一步支持了控制导向的观点，即视频具有很强的时空冗余和相关性，即使在较高的掩码比例下也能学习到有效的表征 [20]。这一观察清楚地表明，视频中存在大量**可替换**或**可扰动**的空间，关键在于以受控的方式加以利用。因此，我们认为，混合增强应从不加区分的扰动转向受控替换，主要作用于可替换冗余，同时避免损害判别性动态和类内统计特性。从机制角度来看，当前

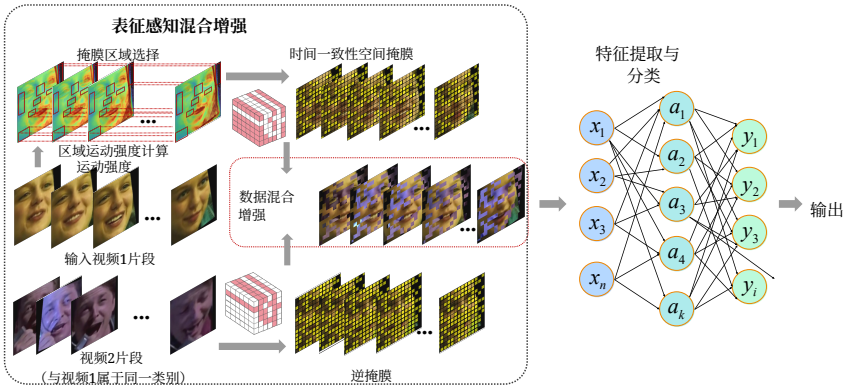

图1：所提出的 *ReMA*概述。*ReMA*执行运动引导、时间一致的空间掩蔽和类内混合。

算法1： *ReMA*的流程

要求： 视频数据集 $\mathcal{D}$，帧数 T，覆盖率 r，块大小 b_0
确保： 增强视频 $\tilde{x}$

采样 $(x_i,\ y_i) \sim \mathcal{D}$ 并进行均匀采样 T 帧 $x = \{x_t\}_{t=1}^{T}$
采样类内视频 $x_j \sim p(x \mid y = y_i)$ 并均匀采样 T 帧 x'
计算 运动 map $\mathcal{A} \leftarrow \frac{1}{3(T-1)} \sum_{t=1}^{T-1} \sum_{c=1}^{3} |x_{t+1}^{(c)} - x_t^{(c)}|$
将池化 $\mathcal{A}$ 合并到图块级运动图 $\mathcal{P}$ 中。块大小 b 0
根据 p_{ij} 采样 r 比例的图块 $\propto (1 - \mathcal{P}_{ij})$
构造管状一致性掩码 $\mathcal{M}$ 共享在所有帧中
通过 $\tilde{x} = (1 - \mathcal{M}) \odot x + \mathcal{M} \odot x'$ 混合视频
返回 $\tilde{x}$

方法仍有很大的改进空间，因为它们很少约束增强样本是否仍处于有效的类条件区域内，也很少施加结构引导、时间连贯且与运动敏感区域相匹配的扰动 [18]。为更好地释放混合增强的潜力，我们提出了表征感知混合增强 (***ReMA***)，这是一种受控的混合视频数据增强方法，可无缝集成到通用视频骨干网络中。*ReMA*不再将混合视为不加区分的扰动，而是明确将其定义为信息替换过程，并对其分布和结构影响进行分层控制。具体而言，表征对齐机制(***RAM***)在分布对齐约束下执行结构化类内混合，有效确保增强样本在统计上与原始类别分布保持一致，同时抑制无关的类内漂移。作为补充，动态选择机制(***DSM***) 根据帧间运动强度自适应生成时空掩码，引导扰动作用于判别敏感度相对较低的区域，并通过时间一致的掩码保持时间连续性 [20]。通过共同调控混合的应用方式和作用位置，*ReMA* 在数据层面提升了混合增强的可控性和稳定性，且无需引入额外监督或可训练参数。

总体而言，主要贡献总结如下：

- 我们从受控替换的视角重新审视视频混合增强，其动机在于视频的时空冗余以及判别性行为线索分布不均。基于这一视角，我们提出*ReMA*，**一种即插即用的受控混合策略，可在不引入额外监督或可训练参数的情况下，显式调节增强对表征层面的影响**。

- 我们提出了*RAM*，它在分布对齐约束下进行结构化的类内混合，以抑制类内非判别性变化，并提升增强样本的统计可靠性，从而获得更稳定的表征学习增益。
- 我们提出*DSM*，根据帧间运动强度自适应生成时空掩码，同时确保时间一致性，引导扰动避开运动敏感区域，从而在保留关键动态信息的同时，拓展有效的特征多样性。
- 我们在多个具有不同动作粒度的基准数据集上进行了广泛实验，结果表明*ReMA* 能够持续提升性能，并揭示了分布对齐与结构感知控制在提升混合增强效果上限方面所发挥的互补作用。

II. 一种新的数据增强方法

我们提出*ReMA*，一种面向表征的混合增强框架，用于视频行为识别，旨在通过受控的时空扰动扩展类内多样性。其核心思想是将基于混合的增强重新表述为一种类别条件且结构一致的替换过程，使增强样本能够稳定地促进判别性表示学习。为此，*ReMA* 由两种互补机制组成：表示对齐机制（*RAM*）用于调节类内混合的统计特性；动态选择机制（*DSM*）则根据视频运动特征调整替换的空间位置和粒度。两种机制完全在数据层面运行，不引入额外监督或可训练参数。*ReMA* 的整体增强流程如图1和算法1所示。

A. 表示对齐机制

RAM 旨在提升基于混合的增强的统计稳定性，使增强样本能够更可靠地促进视频表示学习。尽管类内混合在语义上有效，但其效果在很大程度上取决于是否对扰动源和

替换预算进行了明确约束。缺少此类约束时，不同尺度的混合操作可能会导致各样本间的扰动强度不一致，从而造成表征学习不稳定。

在*ReMA*中，混合增强被定义为一种类别条件且受预算控制的替换过程。给定来自同一类别的两个视频样本 x_i 和 x_j ，通过时空掩码$\mathcal{M}$构造增强样本，表示如下：

$$\tilde{x} = (1 - \mathcal{M}) \odot x_i + \mathcal{M} \odot x_j, \quad x_j \sim p(x \mid y = y_i). \quad (1)$$

类别条件采样约束确保替换内容来自同一类别，从而将扰动限制在类内变化的语义范围内。设 $\Phi(\cdot)$ 表示固定的视频特征映射。在此约束下，增强样本在统计上分布于相应类别的典型表示区域附近，因此混合不会引发系统性的类别级分布漂移：

$$\mathbb{E}[\Phi(\tilde{x}) \mid y_i] \approx \mathbb{E}[\Phi(x) \mid y_i]. \quad (2)$$

因此，混合增强主要扩展类内特征支持范围，而非移动类别中心。

除了约束扰动源之外，*RAM* 还通过时空掩码进一步控制替换预算。具体而言，我们将平均覆盖率定义为

$$r = \frac{1}{THW} \sum_{t,h,w} \mathcal{M}(t, h, w), \quad (3)$$

该指标表示视频中被替换内容所占的比例。通过固定 r，*RAM* 确保每个增强样本都经历相近程度的扰动，避免因掩码覆盖不平衡而导致替换过度或不足。这种预算控制并不要求不同样本之间的内容差异相似，而是使整个数据集中的类内混合整体强度保持一致。

通过上述设计，*RAM*无需引入额外的监督或可训练参数，却能为基于混合的增强提供明确的统计约束。通过将样本级混合转化为表征空间中受预算控制的类内扩展，*RAM* 使增强样本能够更一致地参与判别性决策边界的学习，从而为后续基于运动的动态选择机制奠定稳定基础。

B. 动态选择机制

DSM 负责调控混合增强的应用位置，从而确保*RAM* 定义的预算控制替换与视频数据的时空结构保持一致。*RAM* 限制扰动源和总体替换预算 r，而*DSM* 则侧重于根据视频运动特征自适应地选择替换位置。

给定视频序列 $x = \{x_t\}_{t=1}^{T}$，*DSM* 首先根据相邻帧差异计算运动强度图：

$$\mathcal{A}(h, w) = \frac{1}{3(T-1)} \sum_{t=1}^{T-1} \sum_{c=1}^{3} \left| x_{t+1}^{(c)}(h, w) - x_t^{(c)}(h, w) \right|, \quad (4)$$

其中， $\mathcal{A}(h, w)$ 表示空间位置（h, w）处的平均时间变化。该运动图提供了一种内容感知描述符，用于表征视频内时间动态的空间分布。

随后，使用大小为 b_0 的掩码块将运动图 $\mathcal{A}$ 池化为块级运动分布 $\mathcal{P}$ 。*DSM* 并非调整替换的空间尺度，而是利用运动统计来指导替换预算的分配。具体而言，替换位置根据逆运动概率进行采样：

$$p_{ij} \propto 1 - \mathcal{P}_{ij}, \quad (5)$$

以确保时间变化较低的区域更有可能被选中进行替换。该策略将扰动施加于相对稳定的区域，在这些区域进行替换不太可能破坏关键动态或引入时间不一致性。

最后，通过在所有帧间共享相同的空间掩码，*DSM* 构建了一个管状一致的时空掩码。具体而言，这种管状一致性掩码确保替换区域在整个视频中保持时间对齐，从而保留增强样本的结构连续性。通过上述过程，*DSM*在无需额外监督或可训练参数的情况下，引入了内容感知的空间与时间约束。

通过整合*RAM* 与*DSM*，*ReMA* 为受控混合增强提供了统一的形式化框架，其中统计对齐与时空自适应性得到联合约束。具体而言，*RAM* 调节替换内容的数量及其来源分布，而*DSM* 则根据视频运动特征，确定替换预算应用的位置和空间尺度。在不引入额外监督或可训练参数的情况下，*ReMA*通过受控扰动实现稳定的表征扩展，从而在时间和空间复杂度各异的视频行为识别任务中带来一致的性能提升。

III. 实验

A. 实验设置

1）数据集和指标：我们在四个具有代表性的视频基准数据集上评估*ReMA* ，这些数据集涵盖粗粒度、中粒度和细粒度行为识别，以评估其在不同运动尺度和结构复杂度下的泛化能力。UCF101 [21]用于粗粒度人体动作识别，具有较大的运动变化，并采用 Top-1 和 Top-5 准确率进行评估。DFEW [22] 和 FERV39k [23] 侧重于野外动态面部表情识别，其外观和强度多样性显著；我们报告加权平均召回率和非加权平均召回率（WAR/UAR）

方法	粗粒度		中粒度				细粒度		
	UCF101		DFEW		FERV39k		MA52：身体	MA52：动作	
	Top-1 ↑	Top-5 ↑	WAR ↑	UAR ↑	WAR ↑	UAR ↑	Top-1	Top-1	Top-5
基于二维卷积神经网络的方法									
ResNet	35.87	62.57	64.73	55.22	45.24	35.57	60.04	35.43	73.52
ResNet + *ReMA*	**36.45**$\uparrow_{0.58}$	**63.18**$\uparrow_{0.61}$	**66.48**$\uparrow_{1.75}$	**57.84**$\uparrow_{2.62}$	**47.05**$\uparrow_{1.81}$	**38.12** $\uparrow_{2.55}$	**60.45**$\uparrow_{0.41}$	**35.89**$\uparrow_{0.46}$	**74.10**$\uparrow_{0.58}$
ResNet 长短期记忆网络	41.18	64.74	67.08	54.63	45.88	37.14	60.59	36.87	73.88
ResNet 长短期记忆网络 + *ReMA*	**42.80**$\uparrow_{1.62}$	**65.05**$\uparrow_{0.31}$	**67.64**$\uparrow_{0.56}$	**55.62**$\uparrow_{0.99}$	**47.18**$\uparrow_{1.30}$	**38.28**$\uparrow_{1.14}$	**61.24**$\uparrow_{0.65}$	**37.63**$\uparrow_{0.76}$	**74.20**$\uparrow_{0.32}$
平均提升	**↑1.10**	**↑0.46**	**↑1.16**	**↑1.81**	**↑1.56**	**↑1.85**	**↑0.53**	**↑0.61**	**↑0.45**
基于三维卷积神经网络的方法									
R3D	62.69	84.06	69.25	56.10	46.00	37.95	72.68	50.20	83.93
R3D +*ReMA*	**63.91**$\uparrow_{1.22}$	**84.95**$\uparrow_{0.89}$	**70.31** $\uparrow_{1.06}$	**59.89** $\uparrow_{3.79}$	**48.09**$\uparrow_{2.06}$	**39.18**$\uparrow_{1.23}$	**74.10**$\uparrow_{1.42}$	**53.51**$\uparrow_{3.31}$	**85.79**$\uparrow_{1.86}$
X3D	64.26	85.67	67.21	57.76	46.16	38.40	76.62	51.52	84.64
X3D + *ReMA*	**64.55**$\uparrow_{0.29}$	**86.78**$\uparrow_{1.11}$	**69.09**$\uparrow_{1.88}$	**59.72**$\uparrow_{1.96}$	**48.67**$\uparrow_{2.51}$	**39.72**$\uparrow_{1.32}$	**77.87**$\uparrow_{1.25}$	**54.30**$\uparrow_{2.78}$	**86.54**$\uparrow_{1.90}$
平均提升	**↑0.76**	**↑1.00**	**↑1.47**	**↑2.88**	**↑2.29**	**↑1.28**	**↑1.34**	**↑3.05**	**↑1.88**
基于 Transformer 的方法									
TimeSformer	75.44	92.78	67.12	57.13	47.11	38.31	71.03	44.70	83.51
TimeSformer +*ReMA*	**76.71**$\uparrow_{1.27}$	**93.95**$\uparrow_{1.17}$	**67.94**$\uparrow_{0.82}$	**59.68**$\uparrow_{2.55}$	**48.21**$\uparrow_{1.10}$	**40.01**$\uparrow_{1.70}$	**72.18**$\uparrow_{1.15}$	**44.96**$\uparrow_{0.26}$	**84.04**$\uparrow_{0.53}$
ViedoMAE	72.11	92.02	69.35	59.50	47.37	38.62	76.07	55.67	83.60
ViedoMAE + *ReMA*	**73.91**$\uparrow_{1.80}$	**92.44**$\uparrow_{0.40}$	**71.15**$\uparrow_{1.80}$	**62.63**$\uparrow_{3.13}$	**47.99**$\uparrow_{0.62}$	**39.33**$\uparrow_{0.71}$	**76.82**$\uparrow_{0.75}$	**57.86**$\uparrow_{2.19}$	**86.00**$\uparrow_{2.40}$
平均提升	**↑1.54**	**↑0.79**	**↑1.31**	**↑2.84**	**↑0.86**	**↑1.21**	**↑0.95**	**↑1.23**	**↑1.47**

表 I：不同架构在 UCF101、DFEW、FERV39k 和 MA52 数据集上的测试比较（%）。（**粗体**：最佳；下划线：次佳。）

设置	方法		DFEW	
	RAM	*DSM*	WAR↑	UAR↑
a	×	×	67.21	57.76
a	✔	×	68.71	58.26
b	×	✔	68.79	57.40
d	✔	✔	**69.09**	**59.72**

表 II：*ReMA*在 DFEW 数据集上的消融研究（%）。

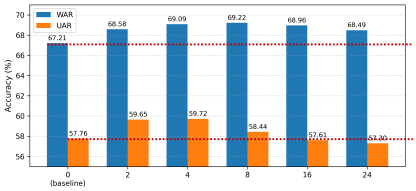

图2：不同基础块大小对 DFEW 上性能的影响。

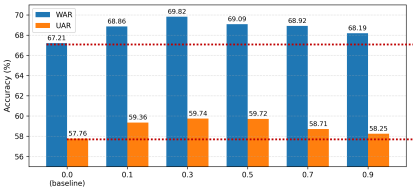

图3：不同覆盖率对 DFEW 性能的影响。

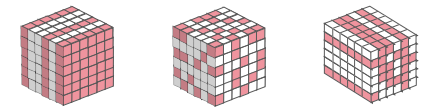

图4：三种时空掩码策略。

设置	方法	WAR↑	UAR↑
a	基线	67.21	57.76
b	A	67.29	58.03
c	B	68.24	57.12
d	C（*ReMA*）	**69.09**	**59.72**

表 III：DFEW 上不同时空策略的对比（%）。

设置	方法	WAR↑	UAR↑
a	基线	67.21	57.76
b	仅掩码	66.27	56.04
c	*ReMA*	**69.09**	**59.72**

表 IV：掩码-only 与*ReMA*在 DFEW 上的消融研究（%）。

为应对类别不平衡。MA-52 [3] 适用于具有细微运动和高类内相似性的细粒度微动作识别，性能通过 Top-1 和 Top-5 准确率衡量。

2）实现细节：所有实验均使用 PyTorch 实现，并在单张 NVIDIA RTX A6000 GPU 上进行。训练配置会根据不同的数据集和骨干架构进行调整。为确保公平比较，所有采用相同骨干架构的方法均在完全相同的配置下训练。*ReMA* 仅在训练期间作为即插即用的数据增强策略使用，不会引入额外的可训练参数，也不会修改骨干架构。

B. 不同数据集上的有效性

我们在多个基准数据集上评估*ReMA* ，涵盖粗粒度动作识别、野外面部表情识别和细粒度微动作识别。作为一种即插即用的增强策略，*ReMA*应用于多种骨干网络架构，从而能够在不同的时空特征和建模范式下进行全面评估。

1）总体性能分析：值得注意的是，在所有数据集上，*ReMA* 始终能够提升识别性能，优于相应的基线方法。这些提升在不同骨干网络系列中均有体现，包括二维卷积神经网络、三维卷积神经网络和基于Transformer的模型，这清楚地表明*ReMA*的有效性与骨干网络无关，其优势源于改进数据级表征学习，而非架构修改。通过在异质视频分布下稳定特征学习，*ReMA* 能够获得更鲁棒的表征。

值得注意的是，在类别不平衡和类内变异较为突出的面部表情数据集上，WAR 和 UAR 等平衡指标的提升尤为显著。这表明，*ReMA* 能够促进更一致的类内表示，并减轻对主导模式的过拟合，从而使各类别上的表现更加稳定、公平。

2）跨行为粒度的有效性：*ReMA*在不同粒度的行为识别任务中均能稳定提升性能，但其优势在不同任务中的表现方式各异。在粗粒度动作识别中，*ReMA* 通过受控替换扩大类内外观与运动多样性，同时不破坏全局时间语义。对于涉及细微动态且受试者间差异较大的面部表情识别，*ReMA*结合统计对齐的类内混合与运动感知掩码放置，在平衡指标上取得显著提升。在判别线索稀疏且高度局部化的细粒度微动作识别中，*ReMA*通过将扰动限制在低运动区域并确保时间一致性，在提升性能的同时保持了关键的细微运动，因此取得尤为稳定的改进。

3）跨骨干架构的一致性：*ReMA*在不同骨干架构上也展现出稳定的性能提升。它通过提供统计对齐且结构一致的训练样本，增强基于卷积的模型；同时，通过缓解无约束扰动所导致的表示噪声，使基于 Transformer 的模型受益。总体而言，在多个数据集和架构上的结果表明，*ReMA*能够提升异质时空条件下表征的稳定性，从而实现更加稳定且具有判别性的表征学习。

C. 消融实验

我们使用 X3D 在 DFEW 数据集上进行了一系列消融实验。DFEW 存在显著的类内差异和类别不平衡，旨在分析*ReMA*两个核心组件的作用。

1）*ReMA* 核心组件的消融实验：如表II所示，单独引入*RAM* 或*DSM* 均能提升基线性能，但仅采用单一机制时，性能提升仍然有限。

移除*RAM* 会导致 UAR 更明显地下降，表明分布感知对齐对于稳定类别级表征的重要性。相比之下，移除*DSM* 主要会降低 WAR，这表明运动感知调节对于保留判别性时间线索至关重要。当两个组件联合应用时，*ReMA* 在 WAR 和 UAR 上仍取得最佳结果，这有力地证实了*RAM* 与*DSM* 在异构视频条件下实现稳定且有效的增强方面所发挥的互补作用。

2）时空混合一致性的消融实验：我们进一步通过比较三种混合策略来考察时空一致性的作用，即空间一致混合、时空随机混合，以及*ReMA*采用的时间一致（顺序）混合。相应的消融结果见图4和表III。

只有时间一致混合才能在 WAR 和 UAR 上均带来明显且稳定的提升。空间一致混合仅带来边际增益，而时空随机混合则会导致 UAR 明显下降，这表明逐帧不一致扰动会破坏内在时间统计特性，尤其是在类别不平衡的情况下。

通过在各帧之间强制保持管级一致性，时间一致混合在保留时间结构的同时扩展了空间多样性，从而有效地使受控替换能够作用于时空冗余，而不会引入人为时间噪声。

3）基础块大小和覆盖率的影响：我们分析了*ReMA*中的两个关键超参数，即基础块大小和覆盖率对性能的影响；二者分别控制空间粒度和替换的整体强度，如图2和图3所示。

显然，适中的基础块大小始终能带来最佳性能。块过细会引入碎片化扰动，而块过粗则更可能干扰语义相关区域和主要运动结构。同样，适中的覆盖率可带来最稳定的增益，即覆盖率过低会导致变化不足，而覆盖率过高则会因替换过度而削弱判别性运动线索。

这些结果证实，*ReMA* 的有效性依赖于受控增强，其中必须联合调节空间粒度与扰动强度，以平衡多样性扩展和结构一致性。

4) 仅掩码增强的消融实验： 我们通过将*ReMA*与仅掩码增强策略进行比较，考察其性能提升是否源于时空掩码本身。该策略直接使用掩码后的样本进行训练，不进行类内混合 [24]。结果见表IV。与基线相比，采用仅掩码增强会导致性能明显下降。这清楚地表明，单独使用掩码主要会抑制信息性内容并造成不可逆的信息丢失，却无法提供有助于表征学习的互补类内变化；在类别不平衡的情况下，这一问题尤为不利。

相比之下，*ReMA* 的表现显著优于基线设置和仅使用掩码的设置。这证实了*ReMA*的有效性源于受时空结构调控的类内替换，而非将掩码作为独立操作。

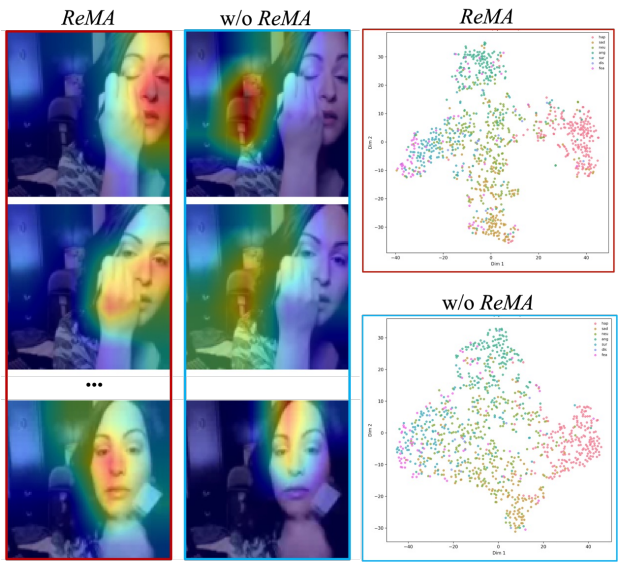

图 5：学习到的表征可视化。

在*ReMA*中，掩码充当结构约束，用于引导混合发生的位置，而不是直接丢弃信息，从而通过结构化的变化有效扩展表征。

D. 可视化

图 5 展示了使用和不使用 *ReMA* 时学习到的表征的定性可视化结果，包括 UCF101 上的 Grad-CAM 激活（左）和 DFEW 上的 t-SNE 嵌入（右）。不使用 *ReMA* 时，激活图在空间上分散且时间上不稳定，常对背景区域或无关纹理作出响应，而来自不同类别的特征嵌入则存在明显重叠。这些观察结果表明，不受约束的增强可能导致表征漂移，并削弱类别条件结构。

相比之下，使用 *ReMA* 训练的模型会产生更加紧凑且语义连贯的激活模式，并且这些模式在不同帧之间保持稳定；其特征嵌入也形成了更紧密的类内簇，类间分离更加清晰。这表明*ReMA* 能够有效抑制非判别性变化、增强时间一致性并规范表征空间，这与观察到的定量性能提升相一致。

IV. 结论

我们提出*ReMA*，一种面向视频行为识别的表征感知数据增强方法，用于解决时空异质性导致的表征不稳定性。通过将增强表述为受控的不变性过程，*ReMA*能够在保留分布结构的同时，稳定地扩展类内表征。*ReMA* 集成了两个轻量级、即插即用的组件。*RAM* 通过结构化类内混合增强判别多样性，而*DSM* 则根据运动线索自适应调节时空扰动，以保持时间连贯性。二者共同构成一种有效的数据级增强策略，无需额外监督或可训练参数。在多种数据集和骨干架构上开展的大量实验表明，该方法能够持续提升鲁棒性和泛化能力。未来的工作将探索把*ReMA* 扩展到更广泛的时空理解任务，并应用于效率更高的部署设置。

参考文献

[1] F.-Q. Cui、J. Huang、Z. Jia、X. Li、X. Yan、X. Zhou 和 M. Wang，《解耦情感基底与瞬态波动：一种用于视频情感分析的低秩稀疏分解方法》，2025年。

[2] [ ] 2511 11406 在线。可访问：https://arxiv.org/abs/。F.-Q. Cui、A. Tong、J. Huang、J. Zhang、D. Guo、Z. Liu 和 M. Wang，《从异质性中学习：通过分布鲁棒优化实现动态面部表情识别的泛化》，载于第*33*届*ACM*国际多媒体会议论文集。计算机协会，2025年。[3] D. Guo、K. Li、B. Hu、Y. Zhang 和 M. Wang，《微动作识别基准测试：数据集、方法与应用》，载于*IEEE*视频技术电路与系统汇刊，2024年。[4] H. Fan、B. Xiong、K. Mangalam、Y. Li、Z. Yan、J. Malik 和 C. Feichten-hofer，《多尺度视觉变换器》，载于*IEEE/CVF*国际计算机视觉会议论文集，2021年，第6824–6835页。[5] Y. Zou、J. Choi、Q. Wang 和 J.-B. Huang，《学习表征不变性以实现数据高效的动作识别》，载于《计算机视觉与图像理解》，2023年。[6] S. Yang、H. Yang、S. Guo、F. Shen 和 J. Zhao，《IPF-RDA：一种信息保持的鲁棒数据增强框架》，载于*IEEE*模式分析与机器智能汇刊，2025年。[7] R. Qian、T. Meng、B. Gong、M.-H. Yang、H. Wang、S. Belongie 和 Y. Cui，《时空对比视频表征学习》，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第6964–6974页。[8] J. Lin、C. Gan 和 S. Han，《TSM：用于高效视频理解的时间移位模块》，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年，第7083–7093页。[9] J. Yang、X. Li、J. Zhang 和 S. Li，《特征偏差校正：一种用于长尾识别的特征增强方法》，载于*2023*年*IEEE*多媒体与博览会国际会议（*ICME*），2023年。[10] D. Fan、J. Wang、S. Liao、Y. Zhu、V. Bhat、H. Santos-Villalobos、R. MV 和 X. Li，《用于时空表征学习的运动引导掩码》，载于*IEEE/CVF*国际计算机视觉会议论文集，2023年，第5619–5629页。[11] J. Shi、H. Ghazzai 和 Y. Massoud，《可微图像数据增强及其应用：综述》，载于*IEEE*模式分析与机器智能汇刊，2024年。[12] C. Shorten 和 T. M. Khoshgoftaar，《深度学习图像数据增强综述》，载于《大数据期刊》，2019年。[13] H. Zhang、M. Cisse、Y. N. Dauphin 和 D. Lopez-Paz，《Mixup：超越经验风险最小化》，载于*ICLR*，2018年。[14] S. Yun、D. Han、S. J. Oh、S. Chun、J. Choe 和 Y. Yoo，《CutMix：利用可定位特征训练强大分类器的正则化策略》，载于*IEEE/CVF*国际计算机视觉会议论文集，2019年。[15] X. Li、M. Xu 和 X. Zhou，《Twins-Mix：在潜在空间中进行自混合，以合理增强三维计算机辅助设计生成建模数据》，载于*2023*年*IEEE*多媒体与博览会国际会议（*ICME*），2023年。[16] C. Cao、F. Zhou、Y. Dai、J. Wang 和 K. Zhang，《基于混合的数据增强综述：分类体系、方法、应用与可解释性》，载于*ACM*计算综述，2024年。[17] S. N. Gowda、M. Rohrbach、F. Keller 和 L. Sevilla-Lara，《Learn2Augment：学习合成视频以增强动作识别数据》，载于欧洲计算机视觉会议，2022年。[18] T. Hong、Y. Wang、X. Sun、F. Lian、Z. Kang 和 J. Ma，《GradSalMix：基于梯度显著性的图像数据增强混合方法》，载于*2023*年*IEEE*多媒体与博览会国际会议（*ICME*），2023年。[19] L. Zhang 和 K. Ma，《良好的数据增强策略并非全部所需：多任务学习视角》，载于*IEEE*视频技术电路与系统汇刊，2023年。

[20] Z. Tong、Y. Song、J. Wang 和 L. Wang，《Videomae：掩码自编码器是适用于自监督视频预训练的数据高效学习器》，载于神经信息处理系统进展，2022年。[21] K. Soomro、A. R. Zamir 和 M. Shah，《UCF101：一个包含101类野外视频人类动作的数据集》，载于*arXiv*，2012年。[22] X. Jiang、Y. Zong、W. Zheng、C. Tang、W. Xia、C. Lu 和 J. Liu，《DFEW：一个用于野外动态面部表情识别的大规模数据库》，载于第*28*届*ACM*国际多媒体会议论文集，系列：MM '20，2020年。[23] Y. Wang、Y. Sun、Y. Huang、Z. Liu、S. Gao、W. Zhang、W. Ge 和 W. Zhang，《FERV39k：一个用于视频面部表情识别的大规模多场景数据集》，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），2022年。[24] Z. Zhong、L. Zheng、G. Kang、S. Li 和 Y. Yang，《随机擦除数据增强》，载于*AAAI*人工智能会议论文集（*AAAI*），2020年。