基于扩散模型和交叉注意力机制的骨骼点动作识别方法

陈恩庆,李佳惠,郭　新

(郑州大学 电气与信息工程学院,河南 郑州 450001)

摘　要:针对人体动作识别中骨骼序列因遮挡或关节点缺失导致的动作信息不完整,以及在标注样本有限情况下模型泛化能力不足等问题,提出了一种结合扩散模型和交叉注意力机制的骨骼点动作识别方法(DCMAE)。在自监督学习框架下,采用时空掩蔽策略,通过扩散模型在去噪过程中学习动作序列的全局分布特性,提升模型在数据缺损情况下的分类准确率;在解码阶段通过交叉注意力机制引入编码器特征,实现时空维度的信息交互与引导,从而增强模型在少标签条件下的泛化能力。实验在 NTU RGB+D 60 和 NTU RGB+D 120 数据集上进行,结果表明,所提方法在数据缺损情况下和少标签条件下的识别准确率较 SkeletonMAE 模型最高分别提升 14.9 百分点和 3 百分点。研究结果表明:所提方法能够有效增强骨骼动作识别模型对缺损数据和少标签数据的鲁棒性,为自监督动作识别提供了新思路。

关键词:骨骼点动作识别;自监督学习;掩蔽重建;扩散模型;交叉注意力机制

中图分类号:TP391;TP181　　**文献标志码**:A　　**doi**:10.13705/j.issn.1671-6833.2026.04.011

基于骨架的动作识别因其数据轻量、计算高效的特点,成为计算机视觉领域的热门研究方向,对深度学习在资源受限环境下的应用与推广具有重要意义[1]。相较于基于 RGB 视频或深度图像的方法,骨架数据在隐私保护、环境鲁棒性以及计算复杂度等方面具有天然优势,在智能监控、人机交互和可穿戴设备等场景中展现出广阔的应用前景。

早期研究主要集中于监督学习下的骨架动作识别。该方法高度依赖大量标注数据,不仅标注成本高昂,而且过程耗时,极大限制了其在实际场景中的应用。为此,自监督学习(self-supervised learning, SSL)[2]因其在表示学习上的显著优势备受关注。现有基于骨架的自监督方法按预训练任务可分为基于上下文、对比学习和生成学习 3 大类[3]。基于上下文的方法[4]利用骨架固有属性构建伪标签以学习时空关系,但设计局限导致难以覆盖全部语义信息。对比学习方法[5]通过实例判别学习高级表示以增强一致性,但性能高度依赖正负样本构造与数据增强策略。生成学习方法[6]旨在通过数据重建与预测来学习潜在分布,从而提升模型泛化能力,通过近似真实数据分布形成更为通用的自监督目标,有助于迫使模型提取语义更丰富的表示。因此,生成式方法在处理复杂动作结构和提升少标签条件下的性能方面具有独特优势。

生成学习方法拥有 VAE[7]和 GAN[8]等实现框架。VAE 能高效构建潜在空间,但生成结果平滑,细节不足;GAN 能产生高质量样本,但训练不稳定且易出现模式崩塌。上述方法在处理复杂、高维且具备时空依赖的骨架数据时,常面临表达能力不足或生成多样性受限的挑战。扩散模型[9]通过逐步加噪与去噪的迭代过程,能精细刻画数据分布,生成图片质量高,有效规避了 VAE 的生成模糊与 GAN 的训练不稳定问题,为骨架自监督学习提供了一种更强大通用的生成途径。扩散模型中包含多种实现形式,如基于随机微分方程的 Score-based 模型[10]、基于潜空间构建的 Latent Diffusion 模型[11]和基于确定性采样过程的 DDIM 模型[12]等。其中,去噪扩散概率模型 DDPM[13]作为代表性框架,通过加噪与

收稿日期:2026-01-31;**修订日期**:2026-03-02

基金项目:国家自然科学基金资助项目(62101503);河南省科技攻关计划项目(242102211017)

作者简介:陈恩庆(1977—),男,河南郑州人,郑州大学教授,博士,主要从事计算机视觉、模式识别和多媒体信息处理研究,E-mail:ieeqchen@zzu.edu.cn。

通信作者:郭新(1988—),女,河南郑州人,郑州大学副教授,博士,主要从事机器学习与人工智能研究,E-mail:iexguo@zzu.edu.cn。

去噪过程精确构建数据分布,兼具理论可解释性和实践稳定性。本文采用 DDPM 作为核心框架,以挖掘扩散模型在骨架自监督学习中的潜力。

尽管扩散模型在生成能力上表现优异,它在编码骨架数据潜在结构、捕捉复杂时空关系方面仍存在一定局限。为弥补这一不足,掩码自编码器(masked autoencoder,MAE)[14]被引入到骨架数据自监督学习中,例如 SkeletonMAE[15]通过 STTFormer[16]构建骨架的时空相关性,能够有效学习骨架的局部和全局结构信息,从而增强潜在表示能力,但 MAE 的生成能力相对有限。

针对上述问题,本文提出了一种结合扩散模型和交叉注意力(cross-attention)[17]机制的 DCMAE 模型。本文贡献如下。

(1)本文设计了一个新型自监督预训练框架,将掩码自编码器的表示学习能力与去噪扩散概率模型的生成学习能力相结合,实现了二者的优势互补,既解决了扩散模型在捕捉复杂时空关系上的局限,又弥补了 MAE 生成能力有限的缺陷。

(2)本文对现有交叉注意力机制进行改进,将时序和空间构建显式解耦,提出时空交叉注意力模块。首先,在每个关节维度上计算时间维度上的交叉注意力;其次,在每个时间维度上对不同关节的空间依赖关系进行交叉注意力运算;最后,将时间交叉注意力与空间交叉注意力串联。这样,模型先在时间维度捕捉动作动态,再在空间维度构建关节协同,实现了对骨架动作的逐层解耦建模。

1 本文方法

1.1 模型架构

本文提出了结合扩散模型和时空交叉注意力的动作识别模型 DCMAE,采用预训练-微调两阶段框架,以充分利用自监督学习的优势,整体结构如图 1 所示。

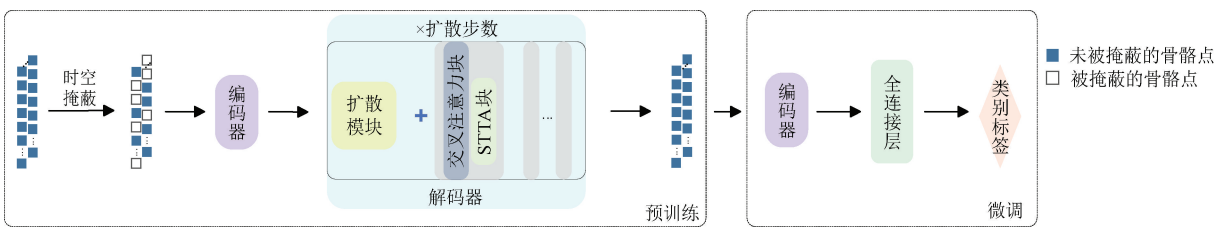

图 1 DCMAE 模型结构

Figure 1 Structure of the DCMAE model

模型以 SkeletonMAE 为基础,输入序列为随机变量 $\boldsymbol{x}_0 \in \mathbf{R}^{N\times C\times T\times V}$。对 $\boldsymbol{x}_0$ 采用时空掩蔽策略,将输入动作维度的掩蔽操作划分为时间与空间两个维度。根据预设的帧掩蔽比例 M_t,在时间掩蔽阶段首先会随机选取部分帧进行掩蔽处理;未被掩蔽的帧将进入下一阶段进行空间掩蔽处理。在空间掩蔽阶段,针对每帧未被掩蔽的关节点,按照设定的关节掩蔽比例 M_j 实施随机掩蔽操作,其中掩蔽的关节索引在每一帧中均为动态生成,即同一关节在不同帧中可能被掩蔽,也可能被保留。经过掩蔽后的序列 $\boldsymbol{x}_0 \in \mathbf{R}^{N\times C\times T\times(1-M_t)\times V\times(1-M_j)}$ 再输入到编码器。

在结构上,DCMAE 的编码器首先通过一个输入层接收数据,继而由一系列时空元组注意力 STTA[16]模块进行深层特征提取。编码器处理未掩蔽序列 $\boldsymbol{x}_0^{\mathrm{v}}$,提取有意义的时空特征。编码后的特征不仅用于骨架数据的高层表征学习,也作为解码器的输入,为后续的重建任务提供结构化信息。针对掩蔽序列 $\boldsymbol{x}_0^{\mathrm{m}}$,模型引入扩散机制进行噪声注入,通过模拟数据退化过程,迫使模型学习从缺损数据中恢复真实信息的能力,进而提升在数据缺失情况下的分类准确率。在解码器中,时空交叉注意力机制被用于聚合编码器特征,通过时空维度的信息交互与引导,增强模型在少标签条件下的泛化能力。

为评估 DCMAE 的性能,采用了端到端的微调策略。该策略将预训练所得编码器的参数权重迁移至所有训练数据上,并仅添加一个简单的全连接层来承接下游的识别分类任务。在整个微调过程中,优化目标为交叉熵损失函数,并最终保留验证集上精度最高的模型作为最佳模型。

1.2 扩散模块

本文是对条件概率分布 $p(\boldsymbol{x}_0^{\mathrm{m}} \mid \boldsymbol{x}_0^{\mathrm{v}})$ 进行建模,即在未掩蔽序列 $\boldsymbol{x}_0^{\mathrm{v}}$ 的条件下,推测并重建出掩蔽序列 $\boldsymbol{x}_0^{\mathrm{m}}$,使得重建序列在数据分布上保持连贯性和合理性。

在前向扩散阶段,仅对掩蔽序列 $\boldsymbol{x}_0^{\mathrm{m}}$ 施加噪声干扰,使其向纯噪声分布演化,而未掩蔽序列 $\boldsymbol{x}_0^{\mathrm{v}}$ 在整个过程中始终保持原始状态不变,以此作为稳定的条件信息传递至解码器,为后续重建过程提供关键参考依据。具体实施过程中,通过迭代方式逐步添加高斯噪声,经过 T 个时间步的连续变换,原始掩蔽

序列依次转变为 $\boldsymbol{x}_1^{\mathrm{m}},\boldsymbol{x}_2^{\mathrm{m}},\cdots,\boldsymbol{x}_T^{\mathrm{m}}$，最终形成完全符合标准正态分布的噪声 $\boldsymbol{x}_T^{\mathrm{m}}$。如文献[18]所示，该过程满足马尔可夫性质，其状态转移概率可表示为

$$p(\boldsymbol{x}_t^{\mathrm{m}} \mid \boldsymbol{x}_{t-1}^{\mathrm{m}}) = N(\boldsymbol{x}_t^{\mathrm{m}};\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}^{\mathrm{m}},\beta_t\boldsymbol{I})。 \quad (1)$$

式中：t 为时间步时间步，$t=1,2,\cdots,T$；β 为预先设定的噪声方差调度参数，用于控制在不同时间步中添加的噪声强度大小。基于高斯分布的可加性，直接对 $\boldsymbol{x}_t^{\mathrm{m}}$ 进行采样，而不需要逐步迭代执行扩散过程，具体采样公式如下所示：

$$p(\boldsymbol{x}_t^{\mathrm{m}} \mid \boldsymbol{x}_0^{\mathrm{m}}) = N(\boldsymbol{x}_t^{\mathrm{m}};\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{\mathrm{m}},(1-\bar{\alpha}_t)\boldsymbol{I})。 \quad (2)$$

式中：$\alpha_t = 1-\beta_t$；$\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$。进一步，将上述采样过程重参数化为

$$\boldsymbol{x}_t^{\mathrm{m}} = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0^{\mathrm{m}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\boldsymbol{\epsilon} \sim N(\boldsymbol{0},\boldsymbol{I})。 \quad (3)$$

通过合理设置噪声方差调度参数，当时间步 T 足够充分大时，$\bar{\alpha}_t$ 接近零，此时 $p(\boldsymbol{x}_t^{\mathrm{m}})$ 近似服从标准正态分布 $N(\boldsymbol{0},\boldsymbol{I})$，从而为后续反向去噪提供初始条件。

在反向去噪阶段，模型需要在已知未掩蔽序列 $\boldsymbol{x}_0^{\mathrm{v}}$ 的条件下，从加噪后的掩蔽序列 $\boldsymbol{x}_t^{\mathrm{m}}$ 出发，预测其对应的原始掩蔽序列 $\boldsymbol{x}_0^{\mathrm{m}}$，即对条件分布 $p(\boldsymbol{x}_0^{\mathrm{m}} \mid \boldsymbol{x}_t^{\mathrm{m}},\boldsymbol{x}_0^{\mathrm{v}})$ 进行建模。不同于标准扩散模型中逐步去除噪声的采样方式，本文直接采用单步重建策略，通过神经网络在某一随机时间步 t 上对加噪掩蔽序列 $\boldsymbol{x}_t^{\mathrm{m}}$ 进行去噪预测，重建原始掩蔽序列 $\boldsymbol{x}_0^{\mathrm{m}}$。该过程不需要从最终噪声序列 $\boldsymbol{x}_T^{\mathrm{m}} \sim N(\boldsymbol{0},\boldsymbol{I})$ 开始逐步采样，避免了多步反向扩散的高计算代价。与传统的多步重建相比，单步策略显著降低了推理计算量，推理速度理论上可提升至标准扩散模型的 T 倍，同时得益于未掩蔽序列 $\boldsymbol{x}_0^{\mathrm{v}}$ 所提供的强条件约束，仍能保持较高的重建质量。为提升重建质量，采用均方误差 *MSE* 作为训练损失函数，表达式为

$$L_{\mathrm{simple}} = E_{t,x_0,\boldsymbol{\epsilon}}\left\|\boldsymbol{x}_0^{\mathrm{m}} - D_\theta(\boldsymbol{x}_t^{\mathrm{m}},t,E_\varphi(\boldsymbol{x}_0^{\mathrm{v}}))\right\|^2, \quad (4)$$

式中：编码器 E_φ 用于将未掩蔽序列 $\boldsymbol{x}_0^{\mathrm{v}}$ 映射到潜在空间，提取高层语义特征并提供全局上下文信息；解码器 D_θ 负责根据噪声输入 $\boldsymbol{x}_t^{\mathrm{m}}$、时间步 t 以及未掩蔽序列的潜在特征表示 $E_\varphi(\boldsymbol{x}_0^{\mathrm{v}})$，完成对原始掩蔽序列 $\boldsymbol{x}_0^{\mathrm{m}}$ 的预测重建[18]。通过利用未掩蔽序列提供的条件信息，能够提升模型对掩蔽序列的重建精度，有效改善重建质量。

1.3 时空交叉注意力模块

SkeletonMAE 通过编码器-解码器框架重建缺失信息，编码器和解码器都由 8 个 STTA 模块组成，能够提取全局的时空特征，解码器负责将掩码信息还原。但原始 SkeletonMAE 的解码器缺乏与编码器输出之间的显式交互，即便编码器学习到了全局表示，解码器在还原过程中只能利用其内部自注意力传播上下文信息，导致在高掩码比例或复杂动作场景下，重建能力受限[19]。因此本文基于交叉注意力机制[17]引入时空交叉注意力模块，在解码器中的每个 STTA 模块之前加入时空交叉注意力模块，二者构成 CSTTA 模块，一共 8 个。

时空交叉注意力模块的核心思想是将时序建模和空间建模解耦，通过两阶段顺序执行，从而更符合骨架动作识别的先验结构。首先，在每个关节维度上，计算时间维度上的交叉注意力。假设第 v 个关节的特征为 $\boldsymbol{X}_v \in \mathbf{R}^{N\times T\times C}$，编码器对应特征为 $\boldsymbol{Z}_v \in \mathbf{R}^{N\times T\times C}$，则时间交叉注意力定义为

$$\mathrm{Attn_t}(\boldsymbol{X}_v,\boldsymbol{Z}_v) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_\mathrm{t}\boldsymbol{K}_\mathrm{t}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}_\mathrm{t}。 \quad (5)$$

式中：$\boldsymbol{Q}_\mathrm{t} = \boldsymbol{X}_v\boldsymbol{W}_Q^{\mathrm{t}}$；$\boldsymbol{K}_\mathrm{t} = \boldsymbol{Z}_v\boldsymbol{W}_K^{\mathrm{t}}$；$\boldsymbol{V}_\mathrm{t} = \boldsymbol{Z}_v\boldsymbol{W}_V^{\mathrm{t}}$。softmax 在时间维度上归一化，这一过程确保了模型在每个关节的时间维度上对齐编码器和解码器的动态特征。

其次，在每个时间步 t 上，对不同关节的空间依赖关系进行交叉注意力运算。设第 t 帧的解码器输入为 $\boldsymbol{X}_t \in \mathbf{R}^{N\times V\times C}$，编码器输出为 $\boldsymbol{Z}_t \in \mathbf{R}^{N\times V\times C}$，则空间交叉注意力定义为

$$\mathrm{Attn_s}(\boldsymbol{X}_t,\boldsymbol{Z}_t) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}_\mathrm{s}K_\mathrm{s}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}_\mathrm{s}。 \quad (6)$$

式中：$\boldsymbol{Q}_\mathrm{s} = \boldsymbol{X}_t\boldsymbol{W}_Q^{\mathrm{s}}$，$\boldsymbol{K}_\mathrm{s} = \boldsymbol{Z}_t\boldsymbol{W}_K^{\mathrm{s}}$，$\boldsymbol{V}_\mathrm{s} = \boldsymbol{Z}_t\boldsymbol{W}_V^{\mathrm{s}}$。softmax 在关节维度上归一化，从而刻画同一时间段的空间依赖关系，突出躯干、四肢等局部关节之间的交互作用。

最后，将时间交叉注意力与空间交叉注意力串联，得到最终输出：

$$\boldsymbol{Y} = \mathrm{Attn_s}(\mathrm{Attn_t}(\boldsymbol{X},\boldsymbol{Z}),\boldsymbol{Z})。 \quad (7)$$

式中：$\boldsymbol{X}$ 为解码器输入；$\boldsymbol{Z}$ 为编码器输出。至此，模型先在时间维度捕捉动作动态，再在空间维度构建关节协同，实现了对骨架动作的逐层解耦建模。

2 实验结果及分析

2.1 数据集

NTU-RGB+D 60[20]数据集共 56 880 个样本，每个样本含 4 种模态。骨架数据包括 25 个关节点的三维坐标；RGB 图像为 1 920×1 080 像素，深度与红外图像为 512×424 像素。动作类别共 60 种，分为 40 种单人日常动作、11 种交互动作和 9 种健康相关动作。数据集提供两种评估策略：跨个体(X-Sub)

和跨视角(X-View)。X-Sub 中,40 名参与者分成各 20 人用于训练和测试;X-View 是以不同视角的摄像机获取同一动作,用于评估模型在不同视角下的泛化能力。

NTU-RGB+D 120[21]数据集相较于 NTU RGB+D 60 数据集,动作类别从 60 种增加到了 120 种,设计了 32 种摄像机布置,使用 3 台相机同时采集不同角度数据,进一步提升了数据量与多样性。动作类别包括 82 种单人动作、26 种交互动作和 12 种健康相关动作,新增类别着重于小物体交互及物体相关的双人动作。数据集提供两种评估策略:跨个体(X-Sub)与跨配置(X-Set),其中在 X-Sub 中将 106 人分成各 53 人用于训练与测试;在 X-Set 中编号为偶数的摄像机布置用于训练,编号为奇数的用于测试。

2.2 实验设置

本文所有实验均在一台 Linux 服务器上进行,操作系统为 Ubuntu 20.04,搭载 2 块 RTX 2080Ti 显卡,采用 Python 3.8 语言编程,基于 PyTorch 1.12.1 框架实现,CUDA 版本为 11.3。在优化器方面,预训练与微调阶段均采用 SGD 优化器,其动量参数设定为 0.9。两阶段的基础学习率与权重衰减则有所不同:预训练阶段分别设为 0.001 和 0.000 2;微调阶段分别为 0.010 和 0.000 4。训练迭代次数均统一设定为 90 个周期,批次大小为 32,并遵循余弦退火策略对学习率进行衰减。

2.3 对比实验

2.3.1 全标签数据下的微调结果对比

本文将 DCMAE 与主流模型在 NTU RGB+D 60 及 NTU RGB+D 120 数据集上进行分类性能对比,结果如表 1 所示。

由表 1 可知,在 NTU RGB+D 60 数据集上,X-Sub 协议下 DCMAE 模型相对于 SkeletonMAE 在分类准确率上提升了 1.1 百分点,在 X-View 协议下提升了 1.2 百分点;在 NTU RGB+D 120 数据集上,DCMAE 在 X-Sub 协议下的分类准确率比 SkeletonMAE 提升了 5.5 百分点,在 X-Set 协议下提升了 5.3 百分点。实验结果表明,所提模型在小规模数据集上表现优于现有方法。

表 1 NTU RGB+D 60 与 NTU RGB+D 120 数据集全标签微调结果对比

Table 1 Comparison of fine-tuning performance on the fully labeled NTU RGB+D 60 and NTU RGB+D 120 datasets

方法	骨干网络	准确率/%			
		NTU RGB+D 60		NTU RGB+D 120	
		X-Sub	X-View	X-Sub	X-Set
CrossSCLR[22]	ST-GCN	82.2	88.9	73.6	75.3
AimCLR[23]	ST-GCN	83.0	89.2	76.4	76.7
CPM[24]	ST-GCN	84.8	91.1	78.4	78.9
AimCLR[23]	STTFormer	83.9	90.4	74.6	77.2
CrossSCLR[22]	STTFormer	84.6	90.5	75.0	77.9
Hi-TRS[25]	Transformer	86.0	93.0	80.6	81.6
SkeletonMAE[15]	STTFormer	86.6	92.9	76.8	79.1
SkeletonMVAE[26]	STTFormer	**88.4**	93.1	80.6	83.5
DCMAE	STTFormer	87.7	**94.1**	**82.3**	**84.4**

除了分类性能外,模型的计算效率和参数规模也是实际应用的重要指标。由表 2 可知,尽管 DCMAE 的参数量略高于 SkeletonMAE,但计算量大幅降低。这表明 DCMAE 在保持模型容量的同时,实现了更高的计算效率,使其更适合部署在实际场景。

表 2 性能比较

Table 2 Performance comparison

模型	参数量/M	计算量/GFLOPs
SkeletonMAE	11	42.8
DCMAE	14	18.2

2.3.2 少标签条件下的微调结果对比

为评估 DCMAE 模型在有限标注样本下的泛化能力,进行了针对性的实验设计,专门针对少标签数据下的情况进行测试。在 NTU RGB+D 60 和 NTU RGB+D 120 两种数据集上都随机抽取 5%和 10%的标记数据来进行微调,结果如表 3 所示。

表 3 NTU RGB+D 60 与 NTU RGB+D 120 数据集少标签微调结果对比

Table 3 Comparison of few-shot fine-tuning performance on the NTU RGB+D 60 and NTU RGB+D 120 datasets

方法	准确率/%							
	NTU RGB+D 60				NTU RGB+D 120			
	X-Sub		X-View		X-Sub		X-Set	
	5%	10%	5%	10%	5%	10%	5%	10%
CrossSCLR[22]	63.5	71.0	66.9	75.1	50.2	58.5	50.4	60.6
AimCLR[23]	63.9	70.2	67.5	76.2	49.0	58.6	51.8	60.5
SkeletonMAE[15]	64.4	73.0	68.8	76.9	50.4	61.8	52.0	62.5
SkeletonMVAE[26]	65.1	73.7	69.3	77.5	**53.9**	62.7	53.0	**64.6**
DCMAE	**65.5**	**76.0**	**70.4**	**77.6**	53.2	**63.3**	**54.3**	63.7

表 4　NTU RGB+D 60 与 NTU RGB+D 120 数据集缺损数据的微调结果对比

Table 4　Fine-tuning results on NTU RGB+D 60 and NTU RGB+D 120 datasets under missing data conditions

方法	准确率/%											
	NTU RGB+D 60						NTU RGB+D 120					
	X-Sub			X-View			X-Sub			X-Set		
	未掩蔽	5%	10%	未掩蔽	5%	10%	未掩蔽	5%	10%	未掩蔽	5%	10%
SkeletonMAE*	86.6	82.4	69.6	92.9	78.8	66.5	76.8	67.5	55.9	79.1	65.7	48.9
SkeletonMAE REC	86.6	82.8	70.9	92.9	79.2	70.1	76.8	68.8	58.3	79.1	67.4	52.5
DCMAE*	**87.7**	83.5	72.1	**94.1**	85.8	68.4	**82.3**	75.2	60.2	**84.4**	77.8	63.8
DCMAE REC	**87.7**	**84.3**	**74.2**	**94.1**	**86.4**	**74.4**	**82.3**	**79.0**	**65.6**	**84.4**	**80.2**	**70.4**

由表 3 可知,在 NTU RGB+D 60 和 NTU RGB+D 120 数据集上,DCMAE 分类准确率较 SkeletonMAE 有 0.7 百分点~3 百分点的性能提升。实验结果进一步验证了 DCMAE 模型在有限标注数据场景下的微调有效性及较好的泛化表现。

2.3.3　缺损数据下的分类结果对比

因为目前还没有骨骼点遮挡下的数据集,为了验证所提出的 DCMAE 模型对缺损数据的分类效果,设计了一组实验。在 NTU RGB+D 60 和 NTU RGB+D 120 数据集上按照 5%和 10%的比例对骨骼点进行随机掩蔽来模仿骨骼点的随机缺损或遮挡,结果如表 4 所示。

表 4 中,带 * 标记的为该模型经过编码器微调后再分类,SkeletonMAE REC 和 DCMAE REC 分别为 SkeletonMAE 和 DCMAE 模型先对缺损数据进行重建再分类。结果显示,当只使用编码器对遮挡数据进行分类时,分类准确率随着缺损比例的增加而降低。在 NTU RGB+D 60 和 NTU RGB+D 120 数据集上,DCMAE REC 的分类准确率比 DCMAE* 有最高 6.6 百分点的提升,而 SkeletonMAE REC 的分类准确率比 SkeletonMAE* 仅有最高 3.6 百分点的提升。这表明通过预训练的重建模型重建被遮蔽的部分,并将重建的人体骨骼再次输入编码器,可有效提升缺损骨骼点数据的分类准确率。此外,对比不同重建模型的提升幅度可知,本文提出的基于扩散的 DCMAE 重建方法性能提升优于 SkeletonMAE,证明了所提模型的有效性。

2.4　消融实验

对相关模块和相关超参数进行了消融实验,实验均在 NTU RGB+D 60 数据集上进行。

2.4.1　模块的消融实验

为了验证所加入的扩散模块和交叉注意力模块的有效性,分别在仅加入扩散模块和仅加入交叉注意力的情况下进行实验,实验结果如表 5 所示,可知在 DCMAE 中无论是去掉扩散模块还是去掉交叉注意力模块,分类准确率都会下降,由此证明了所加入的两个模块的有效性。

表 5　模块的消融实验

Table 5　Ablation experiment on module

方法	准确率/%	
	X-Sub	X-View
DCMAE	**87.7**	**94.1**
DCMAE 无交叉注意力模块	87.3	93.6
DCMAE 无扩散模块	87.0	93.2

2.4.2　解码器深度

本文进一步在解码器深度方面进行了消融实验,改变解码器深度为 5,7,9,11。在 X-Sub 上结果如表 6 所示,可知解码器深度为 9 层时分类准确率最高,而层数过多或过少均会导致性能下降。结合分类准确率与模型规模,最终确定解码器深度为 9 层。

表 6　不同解码器深度下的消融实验

Table 6　Ablation experiment under different decoder depths

解码器深度	准确率/%
5	87.3
7	86.9
9	**87.7**
11	87.5

2.4.3　帧和关节点的掩蔽率

在具有 X-Sub 的 NTU RGB+D 60 上使用随机掩蔽方法,测试当时间维度上的掩蔽率分别为 0.4,0.5,0.6 时,测试空间维度上的掩蔽率分别为 0.4,0.5,0.6 的分类准确率,结果如表 7 所示。最终结果表明,帧掩蔽率为 0.5 和关节点掩蔽率为 0.4 时实现了最佳结果。

2.4.4　噪声水平 t

本文进行了从较小噪声到较大噪声的实验,以观察噪声变化对模型性能的影响。在 X-Sub 上结果如表 8 所示,当噪声水平为 1 000 时表现最好。噪声过小时,模型容易只记住局部细节,无法捕捉更高

层的结构特征;适度增加噪声水平,可以迫使模型学习到更鲁棒的表示;但噪声太大时反而训练变得困难,模型很难有效学习,效果开始下降。考虑到兼顾模型的性能和可用资源,最终选择噪声水平为1 000。

表 7　不同帧和关节点掩蔽率下的消融实验

Table 7　Ablation experiment under different frame and joint masking ratios

帧掩蔽率	关节点掩蔽率	准确率/%
0.4	0.4	86.7
	0.5	86.5
	0.6	87.1
0.5	0.4	**87.7**
	0.5	86.6
	0.6	86.7
0.6	0.4	87.6
	0.5	86.5
	0.6	87.2

表 8　噪声水平的消融实验

Table 8　Ablation experiment on the noise level

噪声水平	准确率/%
500	86.8
750	87.2
1 000	**87.7**
1 250	87.5

2.4.5　解码器维度

本文对解码器嵌入维度进行了消融实验,改变在 DCMAE 解码器中不同的嵌入维数,X-Sub 上结果如表 9 所示。可知 256 维比大尺寸和小尺寸效果更好。同时,随着嵌入维数的增加,模型参数量也在增加,当设置维数为 512 时,模型参数为设置维数为 128 时模型参数量的 10 倍,训练时间更长。因此选择 256 作为默认嵌入维数。

表 9　不同解码器嵌入维度下的消融分析

Table 9　Ablation experiment under different decoder embedding dimensions

解码器嵌入维度	准确率/%	参数量/10^6
128	85.1	4
256	**87.7**	14
512	87.4	40

2.5　可视化

图 2 展示了 DCMAE 在 NTU RGB+D 60 数据集上采用时空掩码策略的可视化结果,选取“吃饭”动作的第 1,20,40,60,80 帧进行展示。

图 2 中第 1 行是原始骨架序列的可视化;第 2 行是掩蔽后骨架序列,红色关节和帧分别以 0.4,0.5 的概率被掩蔽;第 3 行是 DCMAE 重建后骨架序列,通过将原始骨架与重建后骨架进行比较,可以看出 DCMAE 重建的人体骨骼数据在细节上与原始数据略有不同,但整体框架的运动轨迹没有扭曲,也没

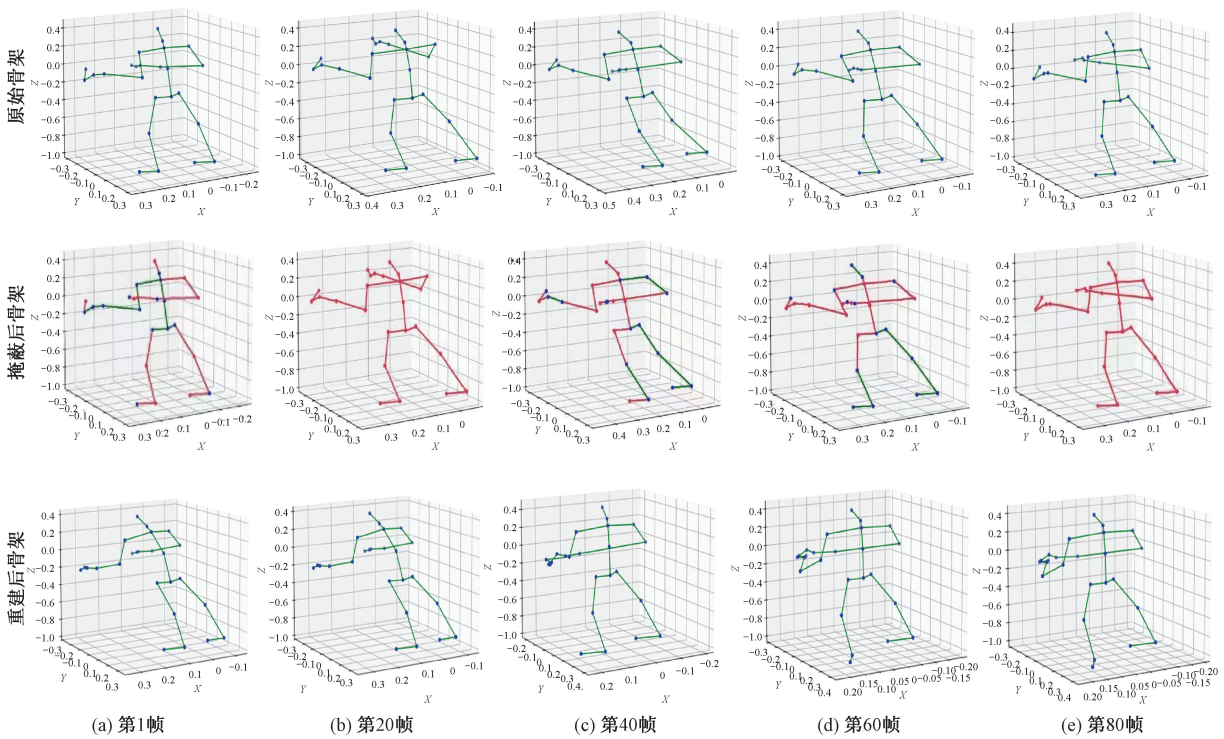

图 2　“吃饭”动作的重建可视化

Figure 2　Visualization of the reconstructed "eating" action

有明显的变形。可视化结果证明了 DCMAE 在重建缺损的人体骨骼序列方面的有效性。这表明 DCMAE 具有较好的生成能力,使骨架能够成功恢复被遮挡的关节并保持骨架序列的完整性。

3 结论

本文针对人体动作识别领域中数据缺损情况下分类准确率低、少标签数据泛化能力弱的问题,提出结合扩散模型和交叉注意力机制的 DCMAE 模型。通过时空掩蔽策略、扩散模型与时空交叉注意力机制的协同设计,构建完整的结构。在编码器阶段利用时空掩蔽策略对原始动作序列进行随机部分掩蔽处理,提取未掩蔽部分的时空特征;扩散模型在逆向去噪过程中学习动作序列全局分布,提升数据缺损情况下的分类性能;在解码器阶段借助时空交叉注意力机制实现时空信息交互,增强少标签数据条件下的泛化能力。在 NTU RGB+D 60 和 NTU RGB+D 120 数据集上的实验结果验证了该模型在复杂场景下的有效性。未来,尝试优化模型结构以降低计算复杂度,考虑改变掩蔽策略和采用其他的生成模型,以期得到更好的效果。

参考文献:

[1] Xin Wentian, Liu Ruyi, Liu Yi, et al. Transformer for Skeleton-based action recognition: a review of recent advances[J]. Neurocomputing, 2023, 537: 164-186.

[2] Gui Jie, Chen Tuo, Zhang Jing, et al. A survey on self-supervised learning: algorithms, applications, and future trends[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 9052-9071.

[3] Zhang Jiahang, Lin Lilang, Yang Shuai, et al. Self-supervised skeleton-based action representation learning: a benchmark and beyond[PP/OL]. V3. arXiv (2025-12-26)[2025-10-10]. https://arxiv.org/abs/2406.02978.

[4] Gao Lingling, Ji Yanli, Yang Yang, et al. Global-local cross-view fisher discrimination for view-invariant action recognition[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM, 2022: 5255-5264.

[5] Chen Zhan, Liu Hong, Guo, Tianyu et al. Contrastive learning from spatio-temporal mixed skeleton sequences for self-supervised skeleton-based action recognition[PP/OL]. V1. arXiv (2022-07-07)[2025-10-10]. https://arxiv.org/abs/2207.03065.

[6] Mao Yunyao, Deng Jiajun, Zhou Wengang, et al. Masked motion predictors are strong 3D action representation learners[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 10147-10157.

[7] Tomczak J M, Welling M. VAE with a VampPrior[PP/OL]. V5. arXiv (2018-02-26)[2025-10-10]. https://arxiv.org/abs/1705.07120.

[8] Liu Ziyu, Zhang Hongwen, Chen Zhenghao, et al. Disentangling and unifying graph convolutions for skeleton-based action recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 140-149.

[9] Fuest M, Ma P C, Gui Ming, et al. Diffusion models and representation learning: a survey[PP/OL]. V1. arXiv (2024-06-30)[2025-10-10]. https://arxiv.org/abs/2407.00783.

[10] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[PP/OL]. V2. arXiv (2021-02-10)[2025-10-10]. https://arxiv.org/abs/2011.13456.

[11] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 10674-10685.

[12] Lukoianov A, De Ocáriz Borde H S, Greenewald K, et al. Score distillation via reparametrized DDIM[PP/OL]. V3. arXiv (2024-10-10)[2025-10-10]. https://arxiv.org/abs/2405.15891.

[13] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[PP/OL]. V2. arXiv (2020-12-16)[2025-10-10]. https://arxiv.org/abs/2006.11239.

[14] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 15979-15988.

[15] Wu Wenhan, Hua Yilei, Zheng Ce, et al. SkeletonMAE: spatial-temporal masked autoencoders for self-supervised skeleton action recognition[C]//Proceedings of the 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Piscataway: IEEE, 2023: 224-229.

[16] Qiu Helei, Hou Biao, Ren Bo, et al. Spatio-temporal tuples transformer for skeleton-based action recognition[PP/OL]. V1. arXiv (2022-01-08)[2025-10-10]. https://doi.org/10.48550/arXiv.2201.02849.

[17] Chen C R, Fan Quanfu, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscat-

away: IEEE, 2021: 347-356.

[18] Wei Chen, Mangalam K, Huang Poyao, et al. Diffusion models as masked autoencoders[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2023: 16238-16248.

[19] Zhang Fuqiang, Bai Junyan, Mu Hui. Human-machine interaction oriented gesture recognition method based on improved GAN[J]. Journal of Zhengzhou University (Engineering Science), 2025, 46(2): 43-50. [张富强, 白筠妍, 穆慧. 基于改进 GAN 的人机交互手势行为识别方法[J]. 郑州大学学报(工学版), 2025, 46(2): 43-50.]

[20] Yue Rujing, Tian Zhiqiang, Du Shaoyi. Action recognition based on RGB and skeleton data sets: a survey[J]. Neurocomputing, 2022, 512: 287-306.

[21] Liu Jun, Shahroudy A, Perez M, et al. NTU RGB+D 120: a large-scale benchmark for 3D human activity understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(10): 2684-2701.

[22] Li Linguo, Wang Minsi, Ni Bingbing, et al. 3D human action representation learning via cross-view consistency pursuit[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 4739-4748.

[23] Guo Tianyu, Liu Hong, Chen Zhan, et al. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(1): 762-770.

[24] Hua Yilei, Wu Wenhan, Zheng Ce, et al. Part aware contrastive learning for self-supervised action recognition[PP/OL]. V2. arXiv (2023-05-11)[2025-10-10]. https://doi.org/10.48550/arXiv.2305.00666.

[25] Chen Yuxiao, Zhao Long, Yuan Jianbo, et al. Hierarchically self-supervised transformer for human skeleton representation learning[C]// Computer Vision-ECCV 2022. Cham: Springer, 2022: 185-202.

[26] Wang Xueting, Guo Xin, Wang Song, et al. Human skeleton action recognition method based on variational autoencoder masked reconstruction[J]. Journal of Graphics, 2025, 46(2): 270-278. [王雪婷, 郭新, 汪松, 等. 基于变分自编码器掩蔽重建的骨骼点动作识别方法[J]. 图学学报, 2025, 46(2): 270-278.]

Diffusion Method and Cross-attention Mechanisms for Skeleton-based Action Recognition Method

CHEN Enqing, LI Jiahui, GUO Xin

(School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: To address the problems of incomplete motion information caused by occlusion or missing joints in skeleton-based action recognition, as well as the limited generalization ability of models under few-label conditions, a skeleton-based action recognition method DCMAE was proposed, which integrated a diffusion model with a cross-attention mechanism. Within a self-supervised learning framework, a spatio-temporal masking strategy was adopted, where the diffusion model learned the global distribution characteristics of motion sequences during the denoising process to improve classification accuracy under data-missing conditions. In the decoding stage, the cross-attention mechanism introduced encoder features to achieve spatio-temporal information interaction and guidance, thereby enhancing the model's generalization ability in few-label conditions. Experiments conducted on the NTU RGB+D 60 and NTU RGB+D 120 datasets showed that the proposed method achieves accuracy improvements of up to 14.9 percentage points and 3 percentage points, respectively, over the SkeletonMAE models under data-missing conditions and few-label conditions. The results demonstrated that the proposed method effectively enhanced the robustness of skeleton-based action recognition models to data-missing and few-label data, providing a new perspective for self-supervised action recognition research.

Keywords: skeleton-based action recognition; self-supervised learning; masked reconstruction; diffusion model; cross-attention mechanism