arXiv:2304.04023v2 [cs.CV] 2 Oct 2024

攻击增强混合对比骨骼表征学习

宾倩 Xu、Xiangbo Shu、*IEEE* 高级会员、Jiachao Zhang、Rui Yan 和 Guo-Sen Xi e

摘要—对比学习依赖有效的正负样本对，有助于在无监督骨架动作识别中学习具有信息量的骨架表征。为了获得这些正负样本对，现有的弱/强数据增强方法必须随机改变骨架的外观，以间接追求语义扰动。然而，这些方法存在两个局限：i）仅对外观进行扰动无法很好地捕捉骨架的内在语义信息；ii）随机扰动可能会将原始的正/负样本对改变为软正/负样本对。为解决上述困境，我们首次尝试探索一种基于攻击的增强方案，该方案额外引入直接的语义扰动，用于构建困难正样本对，并进一步辅助构建困难负样本对。具体而言，我们提出了一种新颖的攻击增强混合对比骨骼表征学习（A²MC），通过对比困难正样本特征和困难负样本特征，学习更加鲁棒的骨架表征。在A²MC 中，攻击增强（Att-Aug）被设计为分别通过攻击和增强协同执行骨架的定向扰动与无目标扰动，以生成高质量的困难正样本特征。同时，提出正负混合器（PNM），将困难正样本特征与负样本特征进行混合，以生成困难负样本特征，并利用这些特征更新混合记忆库。在三个公开数据集上的大量实验表明，A²MC 具有与最先进方法相竞争的性能。代码将在A²MC上提供。

索引词—骨骼表征、动作识别、攻击、对比学习。

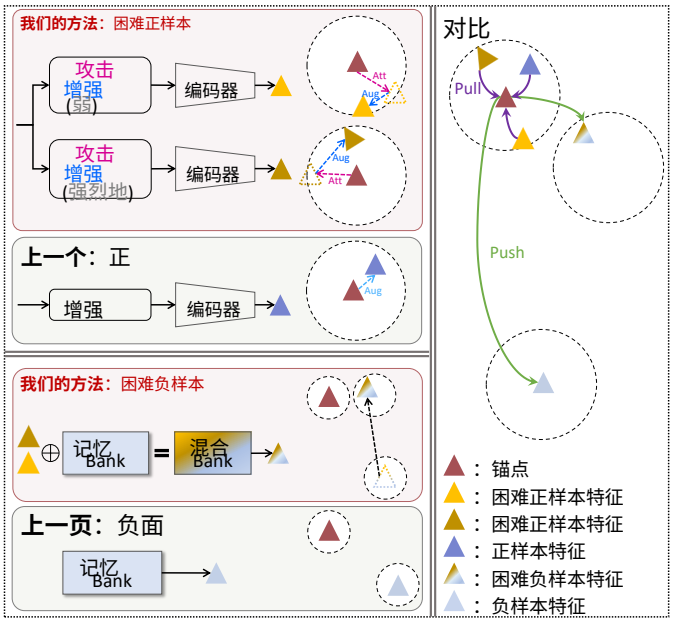


图 1. 本文的主要思想。以往的弱/强增强会随机为锚点构造正样本特征，同时直接将其作为记忆库中的负样本特征。本文引入攻击增强（**“Att”：攻击特征，使其向语义边界移动；“Aug”：随机增强特征，以提高多样性**）来生成困难正样本特征，并进一步构造正负混合特征作为困难负样本特征。与正/负样本特征相比，我们的困难正/负特征更有利于对比学习。

一、引言

HUMAN 动作识别是计算机视觉领域的基础课题，在视频监控、人机交互、娱乐等多种应用场景中发挥着重要作用。[1], [2], [3], [4], [5]。由于骨架具有轻量和鲁棒等特点，基于骨架的动作识别受到了广泛关注 [6], [7], [8], [9]。近年来，监督方法取得了令人满意的性能，但对充足标注数据的需求仍然是一个障碍 [10], [11], [12]。此外，该领域涌现出许多无监督方法，主要基于以下三类模型，即编码器-解码器 [13], [14]，对比学习 [15], [16]，和 Transformer [17], [18]。其中，基于对比学习的方法天然适用于无监督任务，并持续受到学界关注 [19], [15], [18]。

通常，对比学习的过程是拉近正样本对并推远负样本对，以学习数据表征 [20], [21]。因此，对比学习的关键在于构建正（相似）样本对和负（不相似）样本对 [22], [23]。在无监督的基于骨架的动作识别任务中，弱/强增强是构建正样本特征和负样本特征 [15], [16], [24], 的最通用方式，它总是进行随机的外观扰动，以间接实现语义扰动。然而，仅对骨架的外观进行扰动无法很好地捕捉骨架的内在语义信息，这是因为外观扰动所导致的间接语义扰动与直接语义扰动并不一致。此外，随机扰动外观可能会将原始的正/负样本对改变为软正/负样本对。与软正/负

B. Xu、X. Shu 和 *G.-S. Xie* 就职于南京理工大学计算机科学与工程学院，中国南京 *210094*。电子邮箱：{*xubinq11*、*gsxiehm*}*@gmail.com*、{*shuxb*、*jinhui- tang*}*@njust.edu.cn*。通讯作者：*Xiangbo Shu*。*J. Zhang* 就职于人工智能产业技术研究院、南京工程学院，中国南京 *211167*。电子邮箱：*zhangjc07@foxmail.com*。*R. Yan* 就职于南京大学计算机科学学院，中国南京 *210023*。电子邮箱：*ruiyan@nju.edu.cn*。

特征、困难正/负特征对对比学习更有益。直观地说，理想的困难正样本特征应接近锚点的语义内边界，也就是说，它应尽可能远离锚点。同样，理想的困难负样本特征应接近锚点的语义外边界 [25][23]。

因此，如何分别将正样本特征和负样本特征有效地推向锚点的语义内边界和外边界，是构造理想困难正样本对和困难负样本对的关键。语义边界是指动作类别之间的边界。由于原始类别未知，在一定程度上压平所有类别中特征的置信度，会降低对原始类别的置信度，使特征更接近语义边界。上述思路受到白盒攻击相关方法的启发 [26], [27]。基于此，我们首次尝试设计一种攻击增强（Att-Aug）机制，通过弱/强攻击增强实现直接的语义扰动和外观扰动，从而有效生成困难正样本特征，并进一步辅助构造困难负样本特征。如1所示，弱/强攻击增强通过攻击骨架特征，使其接近语义边界，从而对骨架进行语义扰动；同时还进行外观扰动，以生成更多样化的困难正样本特征。随后，将这些困难正样本特征与来自更新中的记忆库的负样本特征进一步混合，生成困难负样本特征，并用于更新混合记忆库。

基于 Att-Aug，我们提出了一种新颖的攻击-增强混合对比骨骼表征学习（A^{2}MC）框架，用于无监督骨架动作识别，如 2 所示。具体而言，攻击-增强（Att-Aug）机制用于构造困难正样本特征：首先通过攻击损失更新骨架，使其接近语义边界；随后将其输入弱/强增强变换和查询编码器，以获得丰富的困难正样本特征。接着，正负混合器（PNM）将少量困难正样本特征与大量对抗负样本特征（位于可学习记忆库中）进行混合，从而生成困难负样本特征（位于混合记忆库中）。最后，混合对比损失（MC）通过拉近相似度分布与独热分布，或与根据基础增强特征计算得到的相似度分布，来训练整个网络。在 NTU RGB+D 60、NTU RGB+D 120 和 PKU-MMD 上的实验验证了所提出的 A^{2}MC 在无监督骨架动作识别任务上的有效性。总体而言，本工作的主要贡献可概括如下：

- 我们提出了一种新颖的攻击增强混合对比骨骼表征学习（A^{2}MC）框架，能够有效构造困难正样本对和负样本对，使模型学习到更鲁棒的骨架表征，用于无监督骨架动作识别。
- 为探索更丰富的困难正样本特征，我们设计了一种新的攻击增强（Att-Aug）机制，分别针对目标语义扰动和非目标外观扰动，对骨架序列进行攻击和增强。
- 为构造和更新困难负样本特征，我们提出了一种新的正负混合器（PNM），用于混合困难正样本特征和对抗负样本特征，随后更新混合记忆库。

II. 相关工作

无监督骨架动作识别。 近年来出现了许多用于基于骨架的动作识别的无监督方法，旨在从无标签数据中学习表征 [19], [15], [18], [28], [16], [24], [29], [30], [13], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]。现有方法几乎可分为三类：编码器-解码器、对比学习和 Transformer。自 MoCo [41], [21]于 2020年出现以来，基于对比学习的方法得到了广泛研究。例如，骨架间对比（ISC） [33] 被提出用于以骨架间和骨架内对比的方式学习多个不同的骨架表征。除了单视图对比学习外，CrosSCLR [16]还进一步探索了关节视图与运动视图之间的跨视图对比。在 ActCLR [42], 中，运动和静态部分以不同程度进行增强，以构造更加多样的正样本对用于对比学习。近期，跨模态互蒸馏（CMD） [29]利用记忆库中锚点的 k 近邻来构造更多正样本对，并在对比框架中探索邻域相似度分布。在本文中，所提出的 A^{2}MC 同样由 MoCo 演化而来，通过使用混合记忆库替代传统记忆库，以解决无监督骨架动作识别问题。

以生成困难正样本对为主要目标，HaLP [43] 通过最小化所生成正样本对与真实正样本对之间的相似度，同时确保二者属于同一原型。A^{2}MC不仅通过基于梯度的白盒攻击生成困难正样本对，还通过将困难正样本与记忆库中的负样本特征混合，生成困难负样本对。

对比学习。 对比学习旨在通过衡量所构造的正样本对/负样本对的相似度/不相似度来学习数据表征 [20], [22], [21], [23]。许多方法考虑不同的增强策略，为锚点特征构造正样本特征 [20], [15][24], [44], [33], [45]。其中，AS-CAL [44] 利用七种不同的常规增强来生成正样本特征，例如旋转、剪切、反转等。此外，除了常规增强之外，AimCLR [15] 结合八种增强策略生成极端增强，以丰富多样化特征。为缓解极端增强造成的不一致性，Zhang 等人 [45] 引入了逐步的层次化增强，以构造更加一致的正样本对。此外，一些研究证明，困难负样本特征对于在对比学习中学习鲁棒表征同样十分重要 [25], [23]。同样，困难正样本

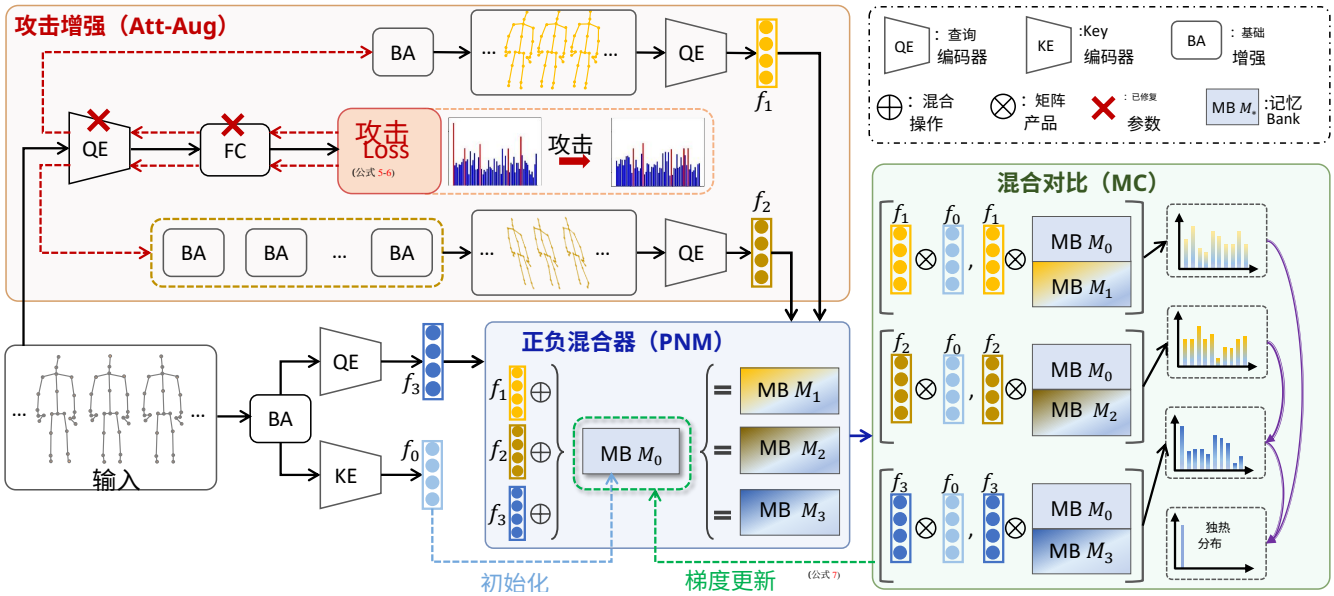


图 2. 所提出的攻击增强混合对比骨架表征学习框架²MC。给定骨架输入，Att-Aug 会生成两种攻击特征 f_1/f_2（弱版本和强版本），作为困难正样本特征（§III-B）；而特征 f_0/f_3 则通过基础增强和键查询编码器获得（§III-A）。PNM 将特征 $f_1/f_2/f_3$ 和记忆库 M_0 混合，以生成混合记忆库 $M_1/M_2/M_3$，其中记忆库 M_0 是由梯度更新的对抗负样本特征集成而成，并以特征 f_0 初始化（§III-C）。在 MC 中，由 f_1 和 f_3 计算得到的相似度分布分别拉向独热分布，而由 f_2 计算得到的相似度分布则拉向由 f_3 计算得到的相似度分布（§III-D）。

配对也可能很重要，但这一点在该领域几乎被忽视。在这项工作中，我们探索困难正样本特征的特性，并旨在同时构造困难正样本特征和负样本特征。

基于梯度的白盒攻击。尽管深度学习模型取得了卓越的性能，但仍然容易受到对抗扰动和对抗攻击的影响 [46], [47], [48], 。因此，人们探索了各种对抗攻击 [49], [50], [51]。其中，白盒攻击是最具代表性的攻击之一，在这种攻击中，目标模型的所有信息都是已知的 [49]。在白盒攻击系列中，基于梯度的攻击是主流技术之一，通过沿着模型损失曲面的梯度上升来欺骗模型 [50]。近期，一些工作 [26],[27]尝试利用基于梯度的攻击来测试学习模型的鲁棒性。在本文中，我们探索基于梯度的白盒攻击，以生成困难正样本特征，从而进一步提升对比表征的鲁棒性。

III. 方法

A. 预备知识

给定骨架序列 x，我们首先对其进行基本的数据增强 [33]（包括姿态增强或关节抖动，以及时间裁剪-缩放），以获得两个增强版本 x_0 和 x_3，遵循 MoCo v2 [41]。随后，x_0 和 x_3 分别输入键编码器和查询编码器，以生成特征 f_0 和 f_3。记忆库$M = \{m_i\}_{i=1}^K$ 以先进先出的方式存储来自 的大量负样本特征。我们采用 InfoNCE 损失 [52] 来训练整个对比网络，即，

$$\mathcal{L}_{in} = -\log \frac{\exp(f_3^\top f_0/\tau)}{\exp(f_3^\top f_0/\tau) + \sum_{i=1}^K \exp(f_3^\top m_i/\tau)} \quad (1)$$

其中，τ 是温度超参数。类似于 [15],，我们也可以将上述 InfoNCE损失转换为另一个版本，以便后续理解。第一步是计算 u 和 v 的相似度分布，如下所示，

$$\psi(u, v) = \frac{\exp(u^\top v/\tau)}{\exp(u^\top v/\tau) + \sum_{i=1}^K \exp(u^\top m_i/\tau)} \quad (2)$$

因此，InfoNCE 损失的另一种形式表示为

$$\mathcal{L}_{in} = -\mathbf{1} \cdot \log \psi(f_3, f_0) - \sum_{i=1}^K \mathbf{0} \cdot \log \psi(f_3, m_i) \quad (3)$$

其中，$\mathbf{1}$ 和$\mathbf{0}$ 是独热分布中的元素指示符。为保持记忆库 M 中存储的特征一致性，键编码器是查询编码器的动量更新版本，如下所示，

$$\theta_k \leftarrow \alpha\theta_k + (1-\alpha)\theta_q \quad (4)$$

其中，α 是动量系数，θ_k/θ_q 表示键编码器/查询编码器的参数集。

[{"id": 0, "output": "B. 攻击增强（Att-Aug）"}]

许多基于对比学习的方法采用弱/强 Aug-增强来生成正样本对 [15], [16], [24], [29],[44], [45], ，但很少探索如何生成真正的困难正样本对。直观地说，困难正样本对应主要更接近语义内边界，这可以提升对比能力，从而学习更鲁棒的表征。因此，为了

构造困难正样本特征，我们设计了一种基于梯度的白盒攻击的新型攻击-增强（Att-Aug）机制，协同执行定向语义攻击和无目标外观增强。具体而言，前者旨在将骨架特征移至锚点的语义边界，后者则随机移动骨架特征，使其更加多样化。

定向语义攻击。接近锚点语义边界的一种方法是平滑其语义分布，在一定程度上使其远离原始类别语义，并接近其他类别的语义。需要注意的是，这里的"定向"指的是语义的定向变化，与下文"无目标外观增强"中的随机语义变化形成对比。具体而言，首先将骨架序列 x 依次输入固定的查询编码器和全连接层，得到其类别特征 $f_a \in \mathcal{R}^{C\times 1}$。其次，我们构造攻击损失来平滑 f_a，使 f_a 的分布趋向均匀分布，从而使受攻击的 $\hat{x}$ 相比其原始形式（锚点）更接近语义边界。值得注意的是，将 f_a 的分布拉向均匀分布，必然会降低其所属类别的得分，同时在一定程度上提高其余类别的得分。因此，f_a 会更接近其所属类别与其余类别之间的边界。形式上，攻击损失定义为最大化 f_a 的熵，如下所示，

$$\begin{aligned}\mathcal{L}_{at} &= -\text{Entropy}(f_a) \\ &\triangleq \text{MSE}(f_a - \frac{1}{C}\mathbb{1})\end{aligned} \tag{5}$$

其中，MSE( ) 表示均方误差函数，用于实现最大化熵的功能以便于推导，而 $\mathbb{1}$ 是全 1 向量。计算出攻击损失后，x 的梯度为 $\frac{\partial \mathcal{L}_{at}}{\partial x}$，用于更新骨架序列。攻击后的骨架序列 $\hat{x}$ 计算如下:

$$\begin{aligned}\rho &= \epsilon\Phi(\frac{\partial \mathcal{L}_{at}}{\partial x}, x) \\ \hat{x} &= x + \rho, \quad \|\rho\|_2 < \eta\end{aligned} \tag{6}$$

其中 ϵ 是学习率，Φ 是用于 [53] 计算更新的 Adam优化器，ρ 表示受 ℓ_2-范数限制的扰动，并且 η 是一个标量值，用于确保骨架具有一定程度的不可感知性。到目前为止，在定向语义攻击之后，骨架序列在语义空间中更加接近锚点的语义边界。

无目标外观增强。给定攻击后的骨架序列$\hat{x}$，分别对其进行弱增强和强增强，以改变其外观。具体而言，弱增强与对比学习的基础增强一致，包括姿态增强或关节抖动，以及时间裁剪-缩放。强增强由更多基础增强组成 [15]: 姿态增强或关节抖动、时间裁剪-缩放、空间翻转、旋转、高斯噪声、高斯滤波器和轴向掩码。随后，弱/强增强的骨架序列通过查询编码器转换为困难正样本特征 f_1/f_2。

C. 正负混合器（PNM）

一些研究表明，困难负样本特征也能提升对比能力 [25], [23]。在本文中，正负混合器（PNM）旨在根据不同的权重，将来自在线更新记忆库的负样本特征与正样本特征（$f_1/f_2/f_3$）进行混合，并进一步生成参与混合记忆库更新的困难负样本特征，具体如下所述。

梯度更新记忆库。与基础对比学习中采用先进先出式部分更新的传统记忆库不同，PNM 的记忆库 M_0 整体上沿着梯度 [54], [55]的方向更新，使其更接近锚点。首先，通过存储键编码器的特征来初始化 M_0。假设前向传播后计算出的总损失为$\mathcal{L}$，则相对于 M_0的梯度可表示为 $\frac{\partial \mathcal{L}}{\partial M_0}$。随后，$M_0$ 的更新过程表示如下，

$$M_0 \leftarrow M_0 + \beta\frac{\partial \mathcal{L}}{\partial M_0} \tag{7}$$

其中 β 为学习率。

正负样本混合。为了进一步基于 M_0生成困难负样本特征，我们采用在线混合正负样本特征的策略。具体而言，给定正样本特征（f_1、f_2和 f_3）以及来自M_0的负样本特征，生成记忆库 M_1、M_2和 M_3 的混合过程计算如下:

$$M_* = \{m_{*i}\}_{i=1}^{K} = \{\lambda f_* + (1-\lambda)m_{0i}\}_{i=1}^{K} \tag{8}$$

其中，下标 $*\in\{1,2,3\}$、K 表示记忆库 M_*的长度，λ 为混合系数。为确保混合特征更接近负样本特征，λ 通常小于0.5。在此，由于引入了困难正样本特征的语义，生成的困难负样本特征相对接近锚点的语义外边界。

D. 混合对比（MC）

本文采用混合对比损失（MC）对所有正负样本特征进行对比，从而将正负样本特征之间的相似度分布拉近至对应的分布，如2所示。形式上，给定所有类型的正样本特征 $\{f_0, f_1, f_2, f_3\}$和记忆库 $\{M_0, M_1, M_2, M_3\}$，MC损失 $\mathcal{L}$ 定义为最终的对比损失，即:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \tag{9}$$

在 $\mathcal{L}_1$ 中，f_0、f_1、M_0 和 M_1 的相似度分布被拉向独热分布。将 M_0 和 M_1 拼接以生成最终记忆库 M_{01}，然后按如下方式计算 $\mathcal{L}_1$，

$$\begin{aligned}M_{01} &= M_0 \cup M_1 = \{m_i\}_{i=1}^{2K} \\ \mathcal{L}_1 &= -\mathbf{1}\cdot\log\psi(f_1, f_0) - \sum_{i=1}^{2K}\mathbf{0}\cdot\log\psi(f_1, m_i)\end{aligned} \tag{10}$$

$\mathcal{L}_3$ 和 $\mathcal{L}_1$ 的损失公式相同，$\mathcal{L}_1$（$\mathcal{L}_3$）最大化 f_1（f_3）与 f_0（f_0）的一致性，同时最小化 f_1（f_3）与 M_{01}（M_{03}）的一致性。不同于

上述独热分布的拉取，$\mathcal{L}_2$ 需要拉取 f_0，f_2，M_0，以及 M_2 由于强增强而朝向基础增强的相似度分布 [15], 如下所示，

$$\begin{aligned} M_{02} =& M_0 \cup M_2 = \{m_i\}_{i=1}^{2K} \\ \mathcal{L}_2 =& -\psi(f_3, f_0) \cdot \log \psi(f_2, f_0) \\ & - \sum_{i=1}^{2K} \psi(f_3, m_i) \cdot \log \psi(f_2, m_i) \end{aligned} \tag{11}$$

最后，使用 MC损失 $\mathcal{L}$ （式（9））来训练 A^{2}MC 的整个框架。

IV. 实验

A. 数据集

NTU RGB+D 60 (NTU-60) [56]: 包含 60 个类别的 56,880 个骨架序列，这些序列由 3 台摄像机采集自 40 名受试者。在跨主体基准中，训练集来自 20 名受试者，测试集来自另外 20 名受试者。在跨视角基准中，训练集来自摄像机 2 和 3，测试集来自摄像机 1。

NTU RGB+D120 (NTU-120) [57]: 包含 120 个类别的 114,480 个骨架序列，这些序列由 32 种设置采集自 106 名受试者，是在 NTU RGB+D 60 的基础上扩展而来。在跨主体基准中，训练集来自 53 名受试者，测试集来自另外 53 名受试者。在跨设置基准中，训练集来自偶数编号的设置，测试集来自奇数编号的设置。

PKU-MMD [58]: 包含近 20,000 个动作片段，涵盖 51 个动作类别，由使用 Kinect摄像机采集的第一阶段和第二阶段（PKU-II）组成。由于视角变化较大，PKU-II 更具挑战性，因此建议用于验证模型。PKU-II 包含 6,945 个骨架序列，涵盖 41 个动作类别，由 13 名受试者完成。在推荐的跨受试者基准中，训练集和测试集分别包含 5,332 个和 1,613 个骨架序列。

B. 实现细节与评估

遵循此前的相关工作 [29], [33],，为进行公平比较，采用双向门控循环单元作为编码器。除表 I 中使用 Transformer [59], 图卷积网络 [12], 以及图卷积网络与 Transformer [60],等不同编码器的实验外，其他所有表格均使用双向门控循环单元作为编码器。所有序列长度均通过时间裁剪-缩放调整为固定的 64 帧 [33]。我们采用动量为 0.9 、权重衰减为 10^{-4} 的随机梯度下降优化所提出的 A^{2}MC。初始学习率和小批量大小分别设为 0.01 和 64。所有实验均在 PyTorch 框架 [61]上进行。

自监督预训练。在基于 MoCo v2 [41]，的对比学习中，记忆库大小 K、温度超参数 τ 和动量系数 α 分别设为 16,384、0.07 和 0.999。在 NTU- 60 和 NTU-120 上，A^{2}MC 预训练总计 450 个训练周期，并在 350 个训练周期后将学习率降至 0.001。在 PKU-II 上，A^{2}MC 预训练总计 1,000 个训练周期，并在 800 个训练周期后将学习率降至 0.001。在 Att-Aug 中，对于单步基于梯度的白盒攻击，学习率 ϵ 和标量值 η 分别设为 0.1 和 0.5。对于 the PNM 的可学习记忆库 M_0 的学习率 β 为 3.0，温度较低，为0.03。为获得更多困难负样本特征，在预训练中设置了不同的混合系数（$\lambda = 0.4$，0。3，0。2，0。1）。

线性评估。 在冻结的查询编码器上附加一个可学习的线性分类器，以预测动作类别。网络训练 80 个训练周期，初始学习率为 0.1（在第 50 和 70 个训练周期时分别降至 0.01 和 0.001）。

KNN评估。 一个K近邻（K近邻）分类器无需训练即可直接对测试样本进行分类，其依据是查询编码器学习到的训练样本特征的K近邻，遵循 [13]。

迁移学习评估。查询编码器首先在源数据集上进行预训练，然后与线性分类器一起在目标数据集上微调 80 个训练周期，初始学习率为 0.01（在第 50 和第 70 个训练周期分别降至 0.001 和 0.0001）。源数据集为 NTU-60/NTU-120（跨主体）的训练集，目标数据集为 PKU-II 的测试集。

C. 与最先进方法的比较

线性评估结果。如 I 所示，与其他相关工作相比，所提出的 A^{2}MC 性能最佳。值得注意的是，采用双向门控循环单元作为编码器的 A^{2}MC 在大多数基准测试集（NTU-60 跨视角、NTU-120 跨主体/跨设置）上表现最佳。具体而言，在双向门控循环单元配置一致的情况下，A^{2}MC 在 NTU-120（跨设置）和 PKU-II（跨主体）上分别比相应的最先进方法（即 CMD [29]）高出 1.7% 和 5.4%。CMD 基于邻近相似度分布，在现有负样本对中挖掘正样本对；而 A^{2}MC 则通过攻击和增强骨架来构造新的困难正样本特征。总体而言，A^{2}MC 优于方法 HaLP [43] 。二者都会生成困难正样本对，但 A^{2}MC 还会基于困难正样本对生成困难负样本对。在采用相同的图卷积网络（即 ST-GCN [12]）配置时，A^{2}MC 的表现与 CPM [39]和 ActCLR [42]相当。A^{2}MC 将语义攻击本身视为困难正样本特征；CPM [39] 则将上下文队列中的其他样本识别为困难正样本；ActCLR [42] 则将运动和静态部分均采用不同增强变换得到的骨架视为更多样化的样本。以 Transformer（即 DSTA-Net [59]）为编码器时，与基于 Transformer 的方法 GL-Transformer [18]相比，A^{2}MC 使用更少的参数，也取得了更好的性能。与同时采用图卷积网络和 Transformer 的最先进方法 SCD-Net [60]，相比，在使用相同骨干网络时，我们的方法性能相当，且在 NTU-120 和 PKU-I 上表现更优。

我们还进行了对比实验，以评估 A^{2}MC 在 NTU-60（跨主体）上以三模态数据（包括关节、运动和骨骼）作为输入时的性能。对于包含关节、运动和骨骼的多模态骨架序列，我们仅将 A^{2}MC 扩展为三流版本，即 3s-A^{2}MC，没有进行任何复杂设计。如II所示，3s-A^{2}MC 的性能与 3s-CMD [29]和 3s-ActCLR [42],相当；后者是专为处理多模态而设计的最先进方法。

表 I NTU-60、NTU-120 和 PKU-II 上不同方法的线性评估结果。最佳和次佳结果分别以**粗体** 和下划线标出。

方法	编码器	浮点运算次数	参数	NTU-60		NTU-120		PKU-I
				跨主体	跨视角	跨主体	跨设置	跨主体
MS^{2}L [24](ACM MM 20)	GRU	-	2.28M	52.6	-	-	-	64.9
PCRP [35](IEEE TMM 21)	GRU	-	-	54.9	63.4	43.0	44.6	-
LongT GAN [14](AAAI 18)	GRU	-	40.2M	39.1	48.1	-	-	-
EnGAN [62](WACV 19)	LSTM	-	-	68.6	77.8	-	-	85.9
P&C [13](CVPR 20)	GRU	-	-	50.7	76.3	42.7	41.7	-
SeBiReNet [63](ECCV 20)	GRU	-	0.27M	-	79.7	-	-	-
TS着色 [37](ICCV 21)	GCN	-	-	71.6	79.9	-	-	-
H-Transformer [17](ICME 21)	Transformer		>100M	69.3	72.8	-	-	-
GL-Transformer [18](ECCV 22)	Transformer	118.62G	214M	76.3	83.8	66.0	68.7	-
AS-CAL [44](INS 21)	LSTM	-	0.43M	58.5	64.8	48.6	49.2	-
ISC [33](ACM MM 21)	门控循环单元&卷积神经网络&图卷积网络	5.76G	10.0M	76.3	85.2	67.1	67.9	80.9
CrosSCLR [16](CVPR 21)	GCN	5.76G	0.85M	72.9	79.9	-	-	-
CrosSCLR-B [16](CVPR 21)	双向门控循环单元	5.76G	10.0M	77.3	85.1	67.1	68.6	-
CRRL [64](IEEE TIP 22)	LSTM	-	-	67.6	73.8	56.2	57.0	82.1
AimCLR [15](AAAI 22)	GCN	1.15G	0.85M	74.3	79.7	63.4	63.4	83.4
RVTCLR+ [65](ICCV 23)	GCN	-	0.85M	74.7	79.1	-	-	-
PSTL [66](AAAI 23)	GCN	1.15G	0.85M	77.3	81.8	66.2	67.7	88.4
ViA [67](IJCV 24)	GCN	-	0.85M	78.1	85.8	69.2	66.9	-
HYSP [68](ICLR 23)	GCN	-	0.85M	78.2	82.6	61.8	64.6	83.8
CPM [39](ECCV 22)	GCN	2.22G	0.85M	78.7	84.9	68.7	69.6	88.8
CMD [29](ECCV 22)	双向门控循环单元	5.76G	10.0M	79.8	86.9	70.3	71.5	-
SkeAttnCLR [69](IJCAI 23)	GCN	1.15G	0.85M	80.3	86.1	66.3	74.5	87.3
HiCLR [70](AAAI 23)	GCN	1.15G	0.85M	80.4	85.5	70.0	70.4	-
HaLP [43](CVPR 23)	双向门控循环单元	-	10.0M	79.7	86.8	71.1	72.2	-
ActCLR [42](CVPR 23)	GCN	1.15G	0.84M	80.9	86.7	69.0	70.5	-
UmURL [71](ACM MM 23)	Transformer	1.74G	20.3M	82.3	89.8	73.5	74.3	-
Eq-Contrast [72](IEEE TIP 24)	双向门控循环单元	2.88G	10.0M	83.9	90.3	75.7	77.2	89.7
PCM^3 [73](ACM MM 23)	双向门控循环单元	2.88G	10.0M	83.9	90.4	76.5	77.5	-
MAMP [40](ICCV 23)	Transformer	-	-	84.9	89.1	78.6	79.1	92.2
SCD-Net [60](AAAI 24)	图卷积网络&Transformer	7.14G	83.26M	86.6	91.7	76.9	80.1	91.9
A^{3}MC（我们的方法）	Transformer	4.12G	3.12M	77.4	84.7	69.9	71.7	86.2
	GCN	1.21G	0.85M	78.8	85.1	69.5	71.3	86.1
	双向门控循环单元	2.93G	10.0M	80.0	87.9	71.4	73.2	87.4
	图卷积网络& Transformer	7.33G	83.26M	**86.7**	**91.8**	**79.6**	**80.4**	**92.5**

K近邻评估结果。在一些先前的工作中，K近邻评估也被视为动作检索任务 [74], [33]。从III中的K近邻评估结果可见，A^{2}MC在NTU-120上比最先进方法（即CMD [29]）高出1.7%，在NTU-60上则与CMD表现相当。这表明，A^{2}MC以无监督方式在判别空间中学习到了高质量特征。

迁移学习评估结果。我们进行了实验，以检验从源数据集学到的知识是否有助于在目标数据集上进行学习，即所学特征是否具有泛化能力。迁移学习评估结果如IV所示。与其他方法相比，A^{2}MC 学到的特征具有更好的泛化能力，尤其是在 [29] 从 NTU-60（跨主体）迁移到 PKU-II 时，性能高出 2.1%。

D. 消融研究

不同组件的有效性。 为验证 A^{2}MC 框架中各组件（包括 Att-Aug、PNM 和 MC损失）的有效性，我们在 NTU-60 上对不同组件进行了消融研究，如 V 所示。以 B1（采用基础增强和 InfoNCE 损失的基础对比学习）为基准，B3（加入 Att-Aug）在跨主体和跨视角设置下的性能分别提升了 2.9% 和 2.8%，证明了 Att-Aug 的优越性。与 B1 相比，B2（仅采用 PNM）分别提升了 1.3% 和 1.4%，证实了 [54]适用于骨架。将 B4 中的 InfoNCE 损失替换为 MC损失后，性能进一步提升。InfoNCE 损失与 MC 损失的区别在于，后者将特征分布拉向根据正样本特征计算得出的相似度分布，而不是独热分布。最后，在 B4 的基础上加入 PNM 后，A^{2}MC 达到了最佳性能，这表明

表 II NTU-60 上不同模态的线性评估结果。J、M 和 B 分别指三种模态：关节、运动和骨骼。

方法	模态	跨主体
A^{2}MC	J	80.0
A^{2}MC	M	77.1
A^{2}MC	B	78.3
3s-CrosSCLR [16]	J+M+B	77.8
3s-AimCLR [15]	J+M+B	78.9
3s-RVTCLR+ [65]	J+M+B	79.7
3s-HiCLR [45]	J+M+B	80.4
3s-CrosSCLR-B [16]	J+M+B	82.1
3秒-CPM [39]	J+M+B	83.2
3s-HiCo [74]	J+M+B	83.8
3s-CMD [29]	J+M+B	84.1
3s-ActCLR [42]	J+M+B	84.3
3s-A^{2}MC	J+M+B	**84.6**

表 III NTU-60 和 NTU-120 上的 K近邻评估结果。

方法	NTU-60		NTU-120	
	跨主体	跨视角	跨主体	跨设置
LongT GAN [14]	39.1	48.1	31.5	35.5
P&C [13]	50.7	76.3	39.5	41.8
ISC [33]	62.5	82.6	50.6	52.3
CrosSCLR-B [16]	66.1	81.3	52.5	54.9
HiCLR [45]	67.3	75.3	-	-
HaLP [43]	65.8	83.6	55.8	59.0
HiCo [74]	68.3	84.8	56.6	59.1
CMD [29]	70.6	85.4	58.3	60.9
A^{2}MC（我们的方法）	**70.8**	**85.4**	**59.1**	**62.6**

PNM 的有效性。

定向与无目标增强的有效性。我们测试了 Att-Aug 模块中定向语义攻击（Att）和无目标外观增强（Aug）的效果。VI 显示，分别单独加入 Att 或 Aug 均能显著提升性能，其中 Att 的影响更大。这是因为它能生成更具挑战性、且更接近语义边界的样本。在 NTU-60（跨视角）上，同时使用 Att 和 Aug 相较于单独使用 Att 和 Aug，性能分别提升了 1.2% 和 0.8%。因此，

表 IV 迁移学习评估结果。

方法	迁移至 PKU-II	
	NTU-60（跨主体）	NTU-120（跨主体）
LongT GAN [14]	44.8	-
MS^{2}L [24]	45.8	-
ISC [33]	51.1	52.3
CrosSCLR-B [16]	54.0	52.8
CMD [29]	56.0	57.0
A^{2}MC（我们的方法）	**58.1**	**58.9**

表 V 在 NTU-60 上进行线性评估的消融研究。

基线	Att-Aug	PNM	MC	NTU-60	
				跨主体	跨视角
B1	✗	✗	✗	75.6	82.8
B2	✗	✓	✗	76.9	84.2
B3	✓	✗	✗	78.5	85.6
B4	✓	✗	✓	79.0	86.7
A^{2}MC	✓	✓	✓	**80.0**	**87.9**

表 VI NTU-60 上定向攻击与非定向增强的线性评估消融研究。

基线	Att	Aug	NTU-60	
			跨主体	跨视角
A1	✗	✗	75.6	82.8
A2	✗	✓	79.0	86.7
A3	✓	✗	79.9	87.1
A^{2}MC	✓	✓	**80.0**	**87.9**

它们作为攻击-增强模块集成到最终模型中。**攻击增强的可视化**。为了直观地说明攻击增强（Att-Aug）在语义层面和外观层面的不同影响，我们将攻击、弱攻击增强和强攻击增强前后的特征分布与骨架进行可视化，如3所示。从语义层面的角度来看，攻击后的特征分布比攻击前更加平滑（参见前五条形图），而弱攻击增强和强攻击增强所产生的结果具有轻微的不确定性。从外观层面的角度来看，攻击后的骨架变化可以忽略不计，而强攻击增强后的变化相比弱攻击增强后的变化更加明显，这可以通过多样化的外观变化进一步丰富表征。

特征的 t-SNE 可视化。我们采用 t-SNE算法来可视化A^{2}MC 学习到的骨架表征，并将其与三个基线进行比较，包括 A^{2}MC 仅使用弱增强、A^{2}MC 仅使用弱增强和强增强、A^{2}MC 仅使用攻击。如4所示，与这三个基线相比，A^{2}MC学习到的特征在同一类别中更加聚合，在不同类别中更具判别性，证明了其表征学习能力。

Att-Aug 中学习率的影响 ϵ 和标量值 η。学习率 ϵ 是影响 Att-Aug 中扰动的重要参数。为探究 ϵ的影响，我们对 ϵ的不同取值进行了实验。如图5a所示，当 $\epsilon = 0.1$时，NTU-60（x-view）上的性能最佳。因此， ϵ 在最终实现中设为 0.1。标量值 η 是另一个关键参数，用于限制攻击产生的扰动幅度。图5b 显示，取值过大或过小都会损害性能：较大的取值可能破坏语义，而较小的取值则可能削弱困难正样本特征。在所有实验中，我们选择 $\eta = 0.5$ 作为默认设置。

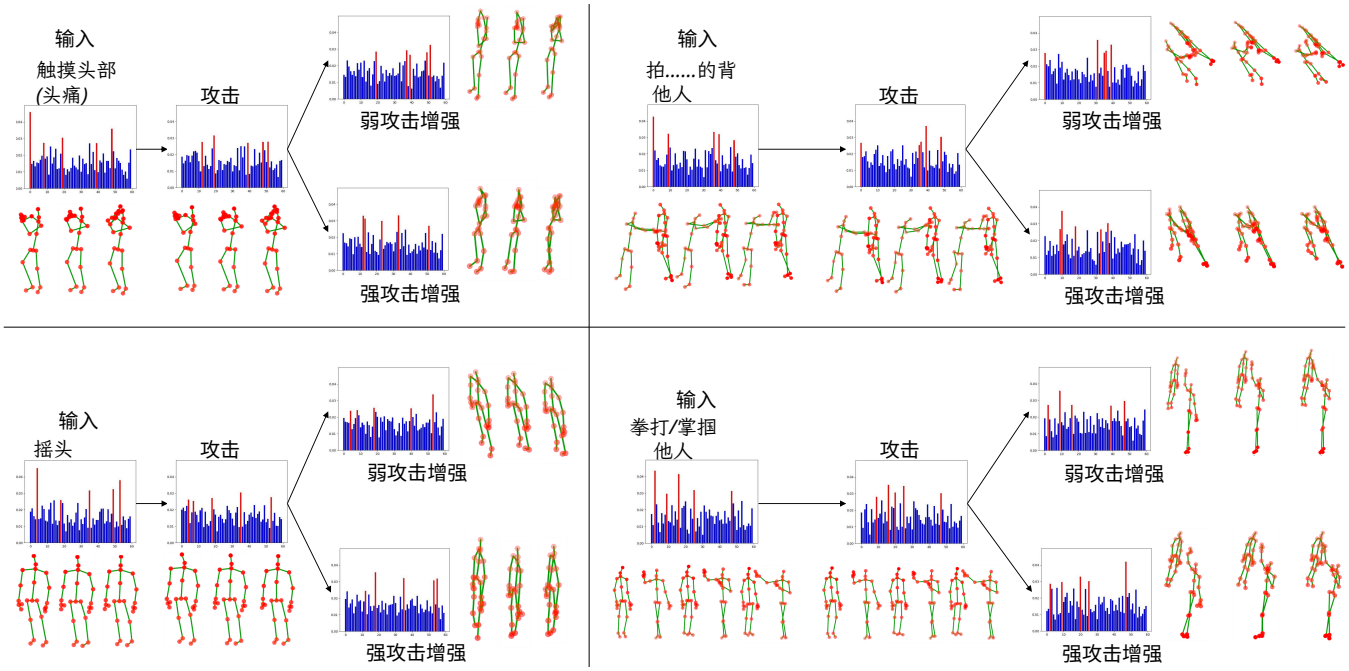


图3. NTU-60 上攻击增强的可视化。对于每组，给定原始骨架序列，分别将其经过攻击、弱攻击增强和强攻击增强后的骨架及特征分布进行可视化。在这里，我们利用 FC 层获取特征分布，其中 Top-5 条形图被突出显示。

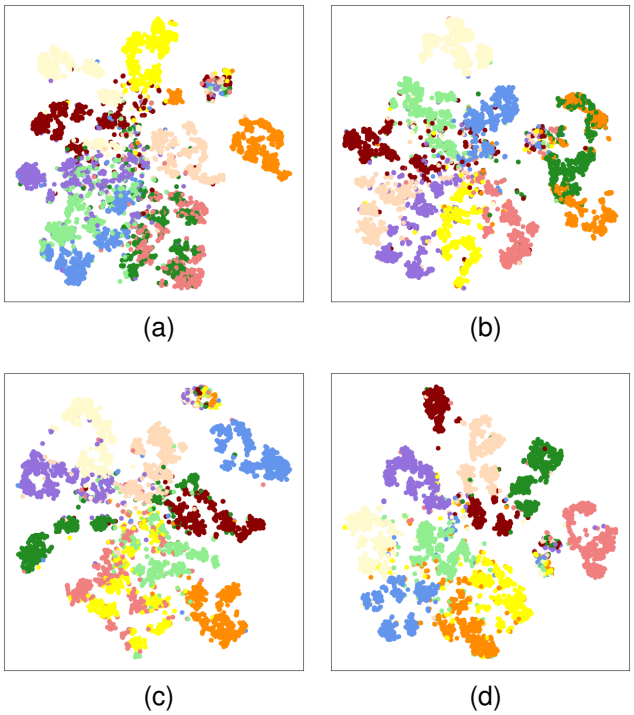


图 4。NTU-60上的 t-SNE。随机选取并报告十个动作类别。（a）使用弱。（b）使用弱和强。（c）使用攻击。（d）A^2MC。

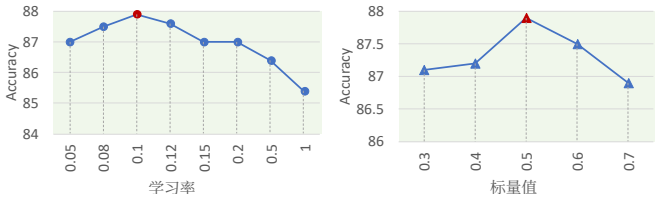


图 5. Att-Aug 在 NTU-60（跨视角）上不同学习率 ϵ 和标量值 η 的线性评估结果。（a）学习率 ϵ。（b）标量值 η。

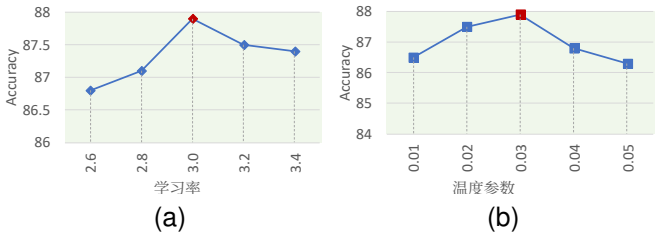


图6.PNM在 NTU-60（跨视角）上的不同学习率 β 和温度的线性评估结果。（a）学习率 β。（b）温度。

PNM 中学习率的影响 β 和温度。在 PNM 中，学习率 β 与温度共同用于控制混合记忆库 [54]的更新幅度。如图6所示，它在下游任务上取得了最佳性能——

表现，当 $\beta = 3.0$ 采用温度为0.03时。

混合系数的影响 λ （PNM 中）。为探究混合系数 λ的影响，我们通过经验性地设置四种组合混合系数来测试性能，即 $\lambda \in$ {(0.4)、(0.4，0.3)、(0.4，0.3，0.2)、(0.4，0.3，0.2，0.1)}。这里，λ = (0.4)表示采用该权重混合

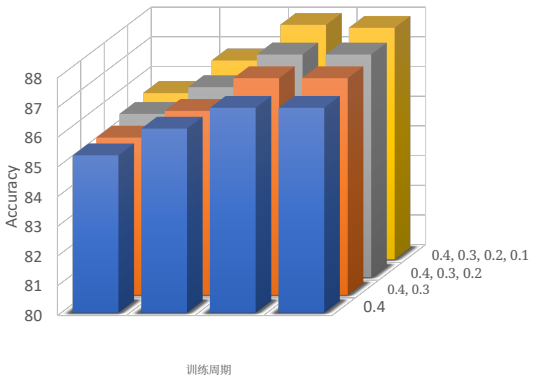


图 7. NTU-60（跨视角）上不同混合系数 λ 和训练周期的线性评估结果。

表 VII NTU-60 上的微调评估结果。

方法	NTU-60	
	跨主体	跨视角
ST-GCN [12]	81.5	88.3
SkeletonCLR [16]	82.2	88.9
AimCLR [15]	83.0	89.2
SkeleMixCLR [75]	84.5	91.1
CPM [39]	84.8	91.1
A²MC（我们的方法）	**84.8**	**92.3**

正负特征以生成 K 困难负样本特征，而 $\lambda = (0.4, 0.3, 0.2, 0.1)$表示采用这四个权重混合正负特征，以生成 $4K$ 困难负样本特征。如图 7 所示，通过 $\lambda = (0.4, 0.3, 0.2, 0.1)$混合更多困难负样本特征，有助于 A²MC 在不同训练周期中持续取得更好的性能。因此，最终实现采用 $\lambda = (0.4, 0.3, 0.2, 0.1)$。

不同训练周期下的性能。 为展示 A²MC 在困难负样本特征数量不同时的收敛情况，我们报告了预训练阶段总训练周期不同时的性能。如图7所示，我们观察到，性能从第 250 个训练周期到第 750 个训练周期逐渐提升，最终在第 750 个训练周期趋于稳定。与相关工作中 CMD [29]/CrosSCLR [16] 的总训练周期 450/300 相比，更多的困难负样本特征可能需要更多迭代。

NTU-60 上的微调评估结果。 在微调评估中，预训练的查询编码器和线性分类器以全监督方式在所有标注数据上进行训练。与最先进方法（即 CPM [39]）相比，微调后的 A²MC 达到了可比的性能，如表 VII所示。这体现了 A² MC 的泛化能力。**NTU-60 上的半监督评估结果。**为验证 A²MC 在半监督场景下的性能，我们在 NTU-60 上使用少量标注数据进行了实验。如表 VIII所示，A²MC 在跨主体 [29] 使用 1% 标注数据和跨视角使用 10% 标注数据时，分别比 CMD 高出 1.4% 和 1.3%。此外，我们还进一步提供了随机

表 VIII NTU-60 半监督评估结果。

方法	跨主体		跨视角	
	1%	10%	1%	10%
LongT GAN [14]	35.2	62.0	-	-
MS²L [24]	33.1	65.2	-	-
ASSL [30]	-	64.3	-	69.8
ISC [33]	35.7	65.9	38.1	72.5
CrosSCLR-B [16]	48.6	72.4	49.8	77.0
CMD [29]	50.6	75.4	53.0	80.2
A²MC（我们的方法）	**52.0**	**76.4**	**53.5**	**81.5**

表 IX NTU-60 上更多半监督评估结果。

方法	跨主体		跨视角	
	5%	20%	5%	20%
ASSL [30]	57.3	68.0	63.6	74.7
ISC [33]	59.6	70.8	65.7	78.2
CrosSCLR-B [16]	67.7	76.1	70.6	81.9
CMD [29]	71.0	78.7	75.3	84.3
A²MC（我们的方法）	**72.2**	**79.4**	**75.9**	**85.4**

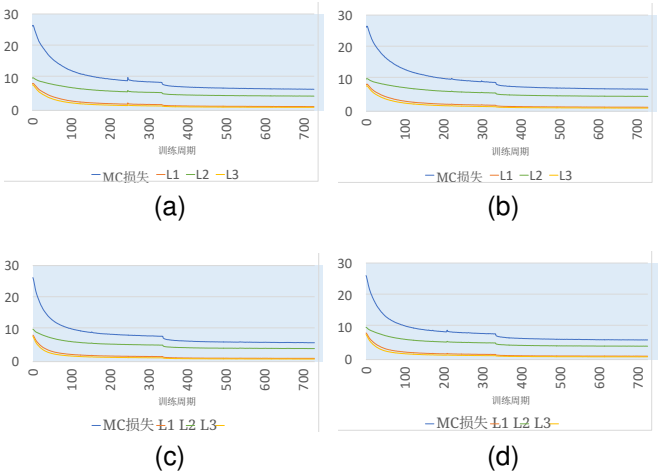


图8.预训练阶段NTU-60和NTU-120上MC损失、 $\mathcal{L}_1$ 损失、 $\mathcal{L}_2$ 损失和 $\mathcal{L}_3$ 损失的变化曲线。(a)NTU-60（跨主体）上。(b) NTU-60（跨视角）上。(c)NTU-120（跨主体）上。(d) NTU-120（跨设置）上。

采样了 5% 和 20% 的标注数据。如表 IX 所示，A²MC 在不同比例的标注数据下均表现良好，证明了 A²MC 的有效性。

三个模态在其余基准测试集上的线性评估结果。 除了三个模态在 NTU-60 xsub 上的线性评估比较结果外，我们还进一步提供了其余基准测试集上的结果，包括 NTU-60xview 和 NTU-120 xsub/xset，如表X所示。与最先进方法（即 3s-HiCo [74] 和 3s-CMD [29]）相比，所提出的 3s-A²MC 具有相当的性能，证明了其有效性。

损失和准确率的变化曲线。图 8展示了MC损失、 $\mathcal{L}_1$ （来自弱攻击增强）的损失、$\mathcal{L}_2$ （来自强攻击增强）的损失，以及 $\mathcal{L}_3$ （来自

表 X：NTU-60 和 NTU-120 上三种模态（关节、运动和骨骼）的线性评估结果。

方法	NTU-60 xview	NTU-120 xsub	NTU-120 xset
3s-CrosSCLR [16]	83.4	67.9	66.7
3s-AimCLR [15]	83.8	68.2	68.8
3s-RVTCLR+ [65]	84.6	60.0	68.9
3s-HiCLR [45]	85.5	-	-
3s-CrosSCLR-B [16]	89.2	71.6	73.4
3s-CPM [39]	87.0	73.0	74.0
3s-HiCo [74]	90.8	**75.9**	**77.3**
3s-CMD [29]	90.9	74.7	76.1
3s-ActCLR [42]	88.8	74.3	75.7
3s-A^{2}MC	**91.2**	75.5	77.1

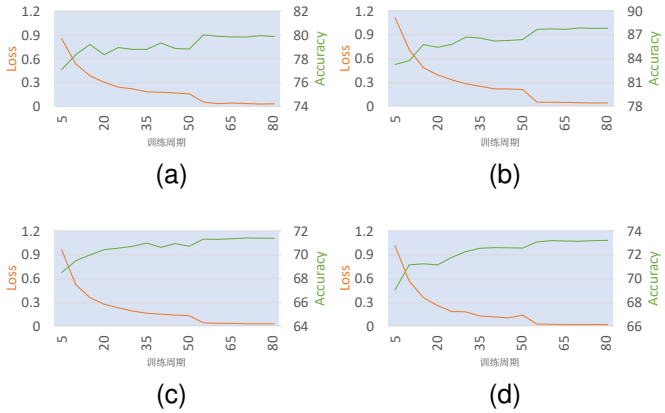


图9.线性评估中 NTU-60和 NTU-120上识别损失与准确率的变化曲线。(a) NTU-60（跨主体）。(b)NTU-60（跨视角）。(c) NTU-120（跨主体）。(d) NTU-120（跨设置）。

基础增强）在预训练对比学习中。我们可以观察到：1）$\mathcal{L}_2$ 的收敛速度比 $\mathcal{L}_1$ 和 $\mathcal{L}_3$的更慢；2）由于学习率下降，所有损失在第 350个 epoch进一步降低，并最终在 NTU-60 和NTU-120 上于第 750 个 epoch 稳定下来。此外，图9展示了线性评估过程中识别损失和准确率的变化。我们可以发现：1）识别损失和准确率的数值在第 80 个 epoch 变得稳定；2）在第 50个 epoch降低学习率对于提升 NTU-60 和 NTU- 120 上的识别性能非常重要。

A^{2}MC获得的混淆矩阵。如图 10所示，在线性评估中，NTU-60（跨主体/跨视角）和NTU-120（跨主体/跨设置）上有四组混淆矩阵。具体而言，每组的左侧混淆矩阵来自第 11 至 20 类。从图 10中可以看出：1）NTU-60/120（跨主体）上的分类准确率低于 NTU-60/120（跨视角/跨设置）上的分类准确率，这表明不同主体执行相同动作之间的差异对 A^{2}MC来说更具挑战性；2）NTU-60 和 NTU-120中细粒度相似动作（例如“阅读”与“写作”，以及“穿鞋”与“脱鞋”）的分类准确率相对较低。

K近邻评估结果的可视化。K近邻评估

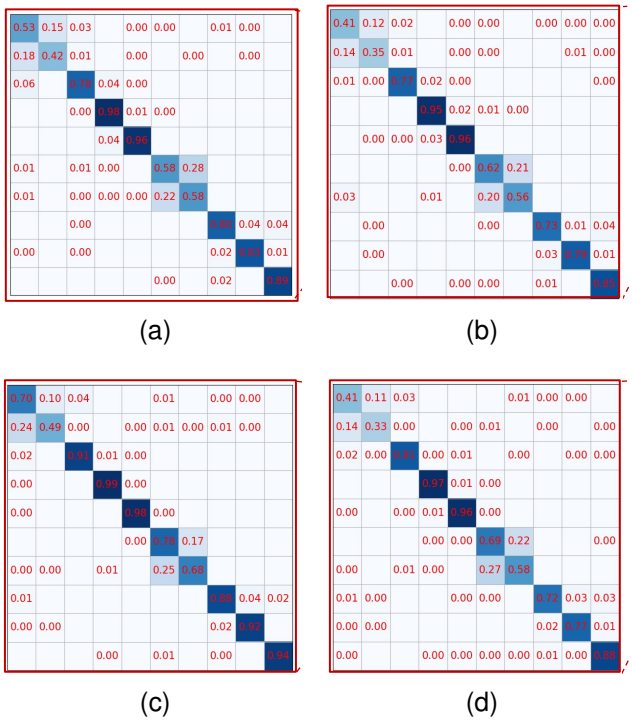


图 10. 在线性评估中，A^{2}MC 在 NTU-60 和 NTU-120 上得到的混淆矩阵。每组中左侧的混淆矩阵对应类别 11 至 20。(a) NTU-60（跨主体）。(b) NTU-120（跨主体）。(c) NTU-60（跨视角）。(d) NTU-120（跨设置）。

第4.3节的结果证明，A^{2}MC 学到了高质量的特征空间。遵循 ISC [33],，我们还对给定查询的前4个最近邻进行了可视化。如图 11所示，每组都包含查询，以及 ISC [33] 和 A^{2}MC 分别得到的前4个最近邻。所有实验结果最初都来自骨架数据，但为了便于理解，我们展示了对应的RGB视频，而非骨架数据。与 ISC 的结果相比，A^{2}MC 学到的前4个邻居中，包含了更多与查询类别一致的动作。这表明，A^{2}MC 学到的表征对相似动作具有更强的判别能力。

V. 讨论

“恶人即天使”。攻击正是我们在这项工作中所需要的天使。我们设计的攻击增强（Att-Aug）首次尝试结合攻击与增强的优势来构造困难正样本特征。前者主要在语义层面对骨架序列进行扰动，旨在获得接近语义边界的特征；后者则主要在外观层面对骨架序列进行扰动，以随机获得多样化的特征。具体而言，为了实现这种攻击效果，我们以无监督的方式采用攻击损失来平滑特征分布，在一定程度上抑制 top-k 条形图之间的差异。这种方法简单而有效，尽管还有各种攻击策略也能实现这一目的。



图 11. NTU-60 上的 K近邻评估结果。每组包括查询样本，以及 ISC 生成的前4个最近邻 [33] 和 A²MC 生成的前4个最近邻。所有实验结果最初均来自骨架数据，但为了便于理解，我们展示了对应的 RGB视频，而非骨架数据。

VI. 结论

我们提出了一种新颖的攻击增强混合对比骨骼表征学习（A^{2}MC）框架，用于无监督的基于骨架的动作识别。该框架整合了攻击增强（Att-Aug）机制和正负混合器（PNM），为混合对比（MC）生成困难正/负特征。具体而言，Att-Aug 协同实现了对骨架的定向语义攻击和无目标外观增强，以促进模型学习更加鲁棒的表征。大量实验表明，A^{2}MC具有优越性。作为首次尝试，我们在 A^2 MC 中采用的攻击相对直接，不同于各种弱/强版本中的增强系列。这启发我们未来进一步探索更多攻击组合策略。

参考文献

[1] J. Carreira 和 A. Zisserman，“Quo vadis，动作识别？一种新模型与 Kinetics数据集”，收录于 *CVPR*，2017年。[2] X. Shu、J. Tang、G. Qi、W. Liu 和 J. Yang，“用于人类交互识别的分层长短期并发记忆”，*IEEE TPAMI*，第40卷，第3期，第1110–1118页，2021年。[3] X. Shu、L. Zhang、Y. Sun 和 J. Tang，“宿主–寄生体：用于群体活动识别的图 LSTM-in-LSTM”，*IEEE TNNLS*，第32卷，第2期，第663–674页，2021年。[4] S. Kumawat、T. Okawara、M. Yoshida、H. Nagahara 和 Y. Yagi，“从单张编码图像进行动作识别”，*IEEE TPAMI*，第45卷，第4期，第4109–4121页，2022年。[5] H. Li、W.-S. Zheng、J. Zhang、H. Hu、J. Lu 和 J.-H. Lai，“通过自动关系建模进行自我中心动作识别”，*IEEE TPAMI*，第45卷，第1期，第489–507页，2022年。[6] T. S. Kim 和 A. Reiter，“利用时间卷积网络进行可解释的三维人体动作分析”，收录于 *CVPR*研讨会，2017年。[7] S. Li、W. Li、C. Cook、C. Zhu 和 Y. Gao，“独立循环神经网络（IndRNN）：构建更长、更深的循环神经网络”，收录于 *CVPR*，2018年。[8] X. Shu、L. Zhang、G.-J. Qi、W. Liu 和 J. Tang，“用于人体骨架运动预测的时空协同注意力循环神经网络”，*IEEE TPAMI*，第44卷，第6期，第3300–3315页，2021年。[9] X. Shu、B. Xu、L. Zhang 和 J. Tang，“用于半监督基于骨架的动作识别的多粒度锚点对比表征学习”，*IEEE TPAMI*，第45卷，第6期，第7559–7576页，2022年。[10] M. Li、S. Chen、X. Chen、Y. Zhang、Y. Wang 和 Q. Tian，“用于基于骨架的动作识别的动作-结构图卷积网络”，收录于 *CVPR*，2019年。[11] Q. Wang、J. Peng、S. Shi、T. Liu、J. He 和 R. Weng，“IIP-Transformer：用于基于骨架的动作识别的部位内外交互变换器”，*arXiv*，2021年。[12] S. Yan、Y. Xiong 和 D. Lin，“用于基于骨架的动作识别的时空图卷积网络”，收录于 *AAAI*，2018年。[13] K. Su、X. Liu 和 E. Shlizerman，“预测与聚类：无监督基于骨架的动作识别”，收录于 *CVPR*，2020年。[14] N. Zheng、J. Wen、R. Liu、L. Long、J. Dai 和 Z. Gong，“利用长期动态进行无监督表征学习，以实现基于骨架的动作识别”，收录于 *AAAI*，2018年。[15] T. Guo、H. Liu、Z. Chen、M. Liu、T. Wang 和 R. Ding，“从极度增强的骨架序列中进行对比学习，以实现动作自监督识别”，收录于 *AAAI*，2022年。[16] L. Li、M. Wang、B. Ni、H. Wang、J. Yang 和 W. Zhang，“通过跨视角一致性追求进行三维人体动作表征学习”，收录于 *CVPR*，2021年。[17] Y.-B. Cheng、X. Chen、J. Chen、P. Wei、D. Zhang 和 L. Lin，“分层 Transformer：用于基于骨架的人体动作识别的无监督表征学习”，收录于 *ICME*，2021年。[18] B. Kim、H. J. Chang、J. Kim 和 J. Y. Choi，“用于无监督基于骨架的动作学习的全局-局部运动变换器”，收录于 *ECCV*，2022年。[19] Y. Chen、L. Zhao、J. Yuan、Y. Tian、Z. Xia、S. Geng、L. Han 和 D. N. Metaxas，“用于人体骨架表征学习的分层自监督 Transformer”，收录于 ECCV，2022年。[20] T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton，“用于视觉表征对比学习的简单框架”，收录于国际机器学习会议，2020年。[21] K. He、H. Fan、Y. Wu、S. Xie 和 R. Girshick，“用于无监督视觉表征学习的动量对比”，收录于 CVPR，2020年。[22] M. Gutmann 和 A. Hyv¨arinen，“噪声对比估计：未归一化统计模型的一种新估计原理”，收录于人工智能与统计学国际会议，2010年。[23] J. Robinson、C.-Y. Chuang、S. Sra 和 S. Jegelka，“利用困难负样本进行对比学习”，arXiv，2020年。[24] L. Lin、S. Song、W. Yang 和 J. Liu，“MS2L：用于基于骨架的动作识别的多任务自监督学习”，收录于 ACM MM，2020年。[25] Y. Kalantidis、M. B. Sariyildiz、N. Pion、P. Weinzaepfel 和 D. Larlus，“用于对比学习的困难负样本混合”，收录于 NeurIPS，2020年。[26] J. Liu、N. Akhtar 和 A. Mian，“针对基于骨架的人体动作识别的对抗攻击”，IEEE TNNLS，第33卷，1622，2020 4，1609号页码。[27] H. Wang、F. He、Z. Peng、T. Shao、Y.-L. Yang、K. Zhou 和 D. Hogg，“理解对抗攻击下基于骨架的动作识别的鲁棒性”，收录于 CVPR，2021年。[28] J. Li 和 E. Shlizerman，“利用主动学习进行稀疏半监督动作识别”，arXiv，2020年。[29] Y. Mao、W. Zhou、Z. Lu、J. Deng 和 H. Li，“CMD：通过跨模态互蒸馏进行自监督三维动作表征学习”，收录于 ECCV，2022年。[30] C. Si、X. Nie、W. Wang、L. Wang、T. Tan 和 J. Feng，“用于半监督三维动作识别的对抗自监督学习”，收录于 ECCV，2020年。[31] Y. Su、G. Lin、R. Sun、Y. Hao 和 Q. Wu，“为自监督三维骨架动作表征学习建模不确定性”，收录于 ACM MM，2021年。[32] Y. Su、G. Lin 和 Q. Wu，“利用运动一致性与连续性进行自监督三维骨架动作表征学习”，收录于 ICCV，2021年。[33] F. M. Thoker、H. Doughty 和 C. G. Snoek，“骨架对比式三维动作表征学习”，收录于 ACM MM，2021年。[34] Z. Tu、J. Zhang、H. Li、Y. Chen 和 J. Yuan，“用于半监督骨架动作识别的关节-骨骼融合图卷积网络”，IEEE多媒体汇刊，2022年。[35] S. Xu、H. Rao、X. Hu、J. Cheng 和 B. Hu，“原型对比与反向预测：无监督基于骨架的动作识别”，IEEE多媒体汇刊，第25卷，第624–634页，2021年。[36] Z. Xu、X. Shen、Y. Wong 和 M. S. Kankanhalli，“利用胶囊自编码器进行无监督运动表征学习”，收录于 NeurIPS，2021年。[37] S. Yang、J. Liu、S. Lu、M. H. Er 和 A. C. Kot，“用于无监督三维动作表征学习的骨架云着色”，收录于 ICCV，2021年。[38] H. Yao、S.-J. Zhao、C. Xie、K. Ye 和 S. Liang，“用于无监督基于骨架的动作识别的循环图卷积自编码器”，收录于 ICME，2021年。[39] H. Zhang、Y. Hou、W. Zhang 和 W. Li，“用于无监督三维动作表征学习的对比正样本挖掘”，收录于 ECCV，2022年。[40] Y. Mao、J. Deng、W. Zhou、Y. Fang、W. Ouyang 和 H. Li，“掩码运动预测器是强大的三维动作表征学习器”，收录于 ICCV，2023年，第10 181–10 191页。[41] X. Chen、H. Fan、R. Girshick 和 K. He，“基于动量对比学习的改进基线方法”，arXiv，2020年。[42] L. Lin、J. Zhang 和 J. Liu，“用于无监督基于骨架的动作识别的动作元依赖对比学习”，收录于 CVPR，2023年，第2363–2372页。[43] A. Shah、A. Roy、K. Shah、S. Mishra、D. Jacobs、A. Cherian 和 R. Chellappa，“HaLP：为基于骨架的动作自监督学习幻构潜在正样本”，收录于 CVPR，2023年，第18 846–18 856页。[44] H. Rao、S. Xu、X. Hu、J. Cheng 和 B. Hu，“利用动量LSTM进行增强骨架对比动作学习，以实现无监督动作识别”，《信息科学》，第569卷，第90–109页，2021年。[45] J. Zhang、L. Lin 和 J. Liu，“通过渐增式增强进行基于骨架的动作识别的分层一致性对比学习”，arXiv，2022年。[46] C. Szegedy、W. Zaremba、I. Sutskever、J. Bruna、D. Erhan、I. Goodfellow 和 R. Fergus，“神经网络的有趣特性”，arXiv，2013年。[47] N. Tanaka、H. Kera 和 K. Kawamoto，“针对动作识别的对抗性骨长攻击

，” 在*AAAI* ，2022，第 2335–2343 页。

[48] L. Zhang、Y. Zhou、Y. Yang 和 X. Gao，《面向未知对抗攻击的可泛化鲁棒性元不变性防御》，IEEE TPAMI，2024年。[49] N. Akhtar、A. Mian、N. Kardan 和 M. Shah，《计算机视觉中对抗攻击与防御的进展：综述》，IEEE Access，第 9 卷，第 155 161–155 196 页，2021年。[50] I. J. Goodfellow、J. Shlens 和 C. Szegedy，《解释并利用对抗样本》，arXiv，2014年。[51] A. Madry、A. Makelov、L. Schmidt、D. Tsipras 和 A. Vladu，《迈向能够抵御对抗攻击的深度学习模型》，arXiv，2017年。[52] A. v. d. Oord、Y. Li 和 O. Vinyals，《基于对比预测编码的表征学习》，arXiv，2018年。[53] D. P. Kingma 和 J. Ba，《Adam：一种随机优化方法》，arXiv，2014年。[54] Q. Hu、X. Wang、W. Hu 和 G.-J. Qi，《AdCo：利用自训练负对抗样本高效学习无监督表征的对抗对比》，发表于 CVPR，2021年。[55] G.-J. Qi 和 M. Shah，《自监督深度网络的对抗预训练：过去、现在与未来》，arXiv，2022年。[56] A. Shahroudy、J. Liu、T.-T. Ng 和 G. Wang，《NTU RGB+ D：一个用于三维人体活动分析的大规模数据集》，发表于 CVPR，2016年。[57] J. Liu、A. Shahroudy、M. Perez、G. Wang、L.-Y. Duan 和 A. C. Kot，《NTU RGB+ D 120：一个用于三维人体活动理解的大规模基准》，IEEE TPAMI，第 42 卷，第 10 期，第 2684–2701 页，2019年。[58] C. Liu、Y. Hu、Y. Li、S. Song 和 J. Liu，《PKU-MMD：一个用于连续多模态人体动作理解的大规模基准》，arXiv，2017年。[59] L. Shi、Y. Zhang、J. Cheng 和 H. Lu，《用于基于骨架的动作与手势识别的解耦时空注意力网络》，发表于 ACCV，2020年。[60] C. Wu、X.-J. Wu、J. Kittler、T. Xu、S. Ahmed、M. Awais 和 Z. Feng，《SCD-Net：用于自监督基于骨架的动作识别的时空线索解耦网络》，发表于 AAAI，2024年，第 5949–5957 页。[61] A. Paszke、S. Gross、S. Chintala、G. Chanan、E. Yang、Z. DeVito、Z. Lin、A. Desmaison、L. Antiga 和 A. Lerer，《PyTorch 中的自动微分》，发表于 NeurIPS研讨会，2017年。[62] J. N. Kundu、M. Gor、P. K. Uppala 和 V. B. Radhakrishnan，《将人体动作作为姿态嵌入流形中的轨迹进行无监督特征学习》，发表于 WACV，2019年。[63] Q. Nie、Z. Liu 和 Y. Liu，《通过视角与姿态解耦进行无监督三维人体姿态表征》，发表于 ECCV，2020年。[64] P. Wang、J. Wen、C. Si、Y. Qian 和 L. Wang，《用于自监督基于骨架的动作识别的对比重构表征学习》，IEEE TIP，第 31 卷，第 6224–6238 页，2022年。[65] Y. Zhu、H. Han、Z. Yu 和 G. Liu，《为自监督基于骨架的动作识别建模相对视觉节奏》，发表于 ICCV，2023年，第 13 913–13 922 页。[66] Y. Zhou、H. Duan、A. Rao、B. Su 和 J. Wang，《从部分时空骨架序列进行自监督动作表征学习》，发表于 AAAI，2023年，第 3825–3833 页。[67] D. Yang、Y. Wang、A. Dantcheva、L. Garattoni、G. Francesca 和 F. Br´emond，《通过动作重定向进行视角不变的骨架动作表征学习》，IJCV，第 132 卷，第 2351–2366 页，2024年。[68] L. Franco、P. Mandica、B. Munjal 和 F. Galasso，《用于自监督骨架动作表征的双曲自步学习》，发表于 ICLR，2023年。[69] Y. Hua、W. Wu、C. Zheng、A. Lu、M. Liu、C. Chen 和 S. Wu，《面向自监督动作识别的部件感知对比学习》，发表于 IJCAI，2023年，第 855–863 页。[70] J. Zhang、L. Lin 和 J. Liu，《结合渐增式增强的骨架动作识别分层一致性对比学习》，发表于 AAAI，2023年，第 3427–3435 页。[71] S. Sun、D. Liu、J. Dong、X. Qu、J. Gao、X. Yang、X. Wang 和 M. Wang，《用于基于骨架的动作理解的统一多模态无监督表征学习》，发表于 ACM MM，2023年，第 2973–2984 页。[72] L. Lin、J. Zhang 和 J. Liu，《用于三维动作表征学习的互信息驱动等变对比学习》，IEEE TIP，第 33 卷，第 1883–1897 页，2024年。[73] J. Zhang、L. Lin 和 J. Liu，《结合掩码动作建模的提示对比：迈向通用三维动作表征学习》，发表于 ACM MM，2023年，第 7175–7183 页。[74] J. Dong、S. Sun、Z. Liu、S. Chen、B. Liu 和 X. Wang，《用于无监督基于骨架的动作表征学习的层次对比》，arXiv，2022年。

[75] Z. Chen、H. Liu、T. Guo、Z. Chen、P. Song 和 H. Tang，“通过时空混合骨架序列的对比学习，实现自监督的基于骨架的动作识别”，*arXiv*，2022年。



Binqian Xu 目前是中国南京理工大学计算机科学与工程学院的博士学位候选人。她目前还在新加坡国立大学担任访问学者。她的研究兴趣包括计算机视觉和深度学习。她在这些领域发表了多篇期刊/会议论文，例如 IEEE TPAMI、IEEE TIP、NeurIPS 等。



Xiangbo Shu 目前是中国南京理工大学计算机科学与工程学院的教授。在此之前，他曾在新加坡新加坡国立大学担任访问学者。他目前的研究兴趣包括计算机视觉和多媒体。他在这些领域发表了 100 多篇期刊和会议论文，包括 IEEE TPAMI、IEEE TNNLS、IEEE TIP、CVPR、ICCV、ECCV、ACM MM 等。他曾获得 MMM 2016 最佳学生论文奖，以及最佳 ACM MM 2015 年论文亚军。他曾担任 IEEE TNNLs、IEEE TCSVT、Pattern Recognition 等期刊的编委会成员。他还是 ACM 会员、CCF 高级会员和 IEEE 高级会员。



Jiachao Zhang 目前是中国南京工程学院人工智能产业技术研究院副教授。她于2011年获得南京理工大学学士学位，2015年获得代顿大学硕士学位，并于2018年获得南京理工大学博士学位。她于2016年至2017年以访问学者身份访问代顿大学。她目前的研究兴趣是图像处理和计算机视觉。她是CCF会员，以及 IEEE 会员。



Rui Yan 在中国南京理工大学智能媒体分析组（IMAG）获得博士学位。他目前是中国南京大学计算机科学与技术系的助理研究员。他于2022年1月至2022年8月在字节跳动担任研究实习生（兼职）。他于2021年9月至2021年12月在腾讯担任研究实习生（兼职）。他于2021年8月至2022年8月在新加坡国立大学（NUS）担任访问研究员。他曾在华为担任研究实习生 NOAH’S ARK 实验室，任职时间为2018年12月至2019年12月。他的研究主要聚焦于复杂人体行为理解和视频-语言理解。他在这些领域发表了20余篇期刊和会议论文，包括 IEEE TPAMI、IEEE TNNLS、IEEE TCSVT、CVPR、NeurIPS、ECCV、ACM MM 等。



Guo-Sen Xie 于2016年在中国北京的中国科学院自动化研究所模式识别国家重点实验室获得模式识别与智能系统博士学位。2014年至2015年，他在新加坡新加坡国立大学担任访问学者。他现任中国南京理工大学计算机科学与工程学院教授。他的研究成果已发表于30余篇顶级期刊和

在 IEEE TPAMI、IJCV、IEEE TIP、IEEE TNNLS、NeurIPS、CVPR、ICCV 和 ECCV 等知名会议上发表论文。他曾获 MMM 2016 最佳学生论文奖。研究兴趣包括计算机视觉和机器学习。