计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500158

骨架数据增强和双重最近邻检索自监督动作识别

吴雨珊[1,2]　徐增敏[1,2]　张雪莲[1,2]　王　涛[3]
1 桂林电子科技大学数学与计算科学学院广西高校数据分析与计算重点实验室　广西 桂林 541004
2 广西应用数学中心(桂林电子科技大学)　广西 桂林 541002
3 桂林电子科技大学建筑与交通工程学院广西智慧交通重点实验室　广西 桂林 541004
(wuyushan2929@163.com)

摘　要　传统基于骨架数据的自监督方法常将某一样本的不同增强作为正例,将其余样本均视为负例,这使得正负样本的比例严重失衡,限制了相同语义信息的样本发挥作用。针对上述问题,提出了一种正样本不受数据增强限制的双重最近邻检索动作识别算法 DNNCLR。首先,基于人体关节的物理连接设计了一个新的关节级空间数据增强,即 Bodypart 增强,对输入的骨架序列用正态分布数组随机替换,以获得高级语义嵌入;其次,为避免正样本受数据增强的限制,提出了一种更合理的双重最近邻检索(DNN)正样本扩充策略,进一步提出了双重最近邻检索对比损失 DNN Loss。具体为利用支撑集进行全局检索,将正样本集的寻找范围扩展到普通数据增强无法覆盖的新数据点;而负样本集中存在被误判的正样本,其是来自不同视频但语义信息相同的骨架样本。为此,再一次利用最近邻检索,从负样本集中寻找这种潜在的正例,二次扩展正样本集,并进一步提出双重最近邻检索对比损失,迫使模型学习更多的一般特征表示,使得模型优化更加合理。最后,将 DNNCLR 算法应用在 AimCLR 模型上,得到 AimDNNCLR 模型,并在 NTU-RGB+D 数据集上对该模型进行了线性评估,与前沿模型相比,所提方法在精度上平均提升了 3.6%。

关键词:对比学习;最近邻检索;数据增强;动作识别;人体骨架
中图法分类号　TP391.41;TP183

Self-supervised Action Recognition Based on Skeleton Data Augmentation and Double Nearest Neighbor Retrieval

WU Yushan[1,2],XU Zengmin[1,2],ZHANG Xuelian[1,2] and WANG Tao[3]
1 School of Mathematics and Computing Science,Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation,Guilin University of Electronic Technology,Guilin,Guangxi 541004,China
2 Center for Applied Mathematics of Guangxi(Guilin University of Electronic Technology),Guilin,Guangxi 541002,China
3 School of Architecture and Transportation Engineering,Guangxi Key Laboratory of ITS,Guilin University of Electronic Technology,Guilin,Guangxi 541004,China

Abstract　Traditional self-supervised methods based on skeleton data often take different data augmentation of a sample as positive examples,and the rest of the samples are regarded as negative examples,which makes the ratio of positive and negative samples seriously unbalanced,and limits the usefulness of samples with the same semantic information. In order to solve the above problems,this paper proposes a double nearest neighbor retrieval action recognition algorithm named DNNCLR,in which positive samples are not limited by data augmentation. First,a new joint level spatial data augmentation,namely Bodypart augmentation,is designed based on the physical connection of human joints. The input skeleton sequence is randomly replaced with a normal distribution array to obtain high-level semantic embedding. Secondly,in order to avoid the limitation of positive samples by data augmentation,a more reasonable double nearest neighbor retrieval(DNN) positive sample augmentation strategy is proposed,and further,a double nearest neighbor retrieval contrastive loss(DNN Loss) is proposed. Specifically,by using support sets for global retrieval,the search range of the positive sample set is expanded to new data points that cannot be covered by ordinary data augmentation. In the negative sample set,there are positive samples that have been misjudged,which are skeleton samples with the same semantic information but from different videos. Therefore,by using nearest neighbor retrieval again,these potential positive

到稿日期:2023-05-23　**返修日期**:2023-08-28
基金项目:国家自然科学基金(61862015,52262047);广西科技基地和人才专项(AD23023002,AD21220114,AD20159035);广西重点研发计划项目(AB17195025)
This work was supported by the National Natural Science Foundation of China(61862015,52262047),Science and Technology Project of Guangxi(AD23023002,AD21220114,AD20159035) and Guangxi Key Research and Development Program(AB17195025).
通信作者:徐增敏(xzm@guet.edu.cn)

examples are searched from the negative sample set to further expand the positive sample set, and the double nearest neighbor retrieval contrastive loss is further proposed, forcing the model to learn more general feature representations, making the model optimization more reasonable. Finally, the DNNCLR algorithm is applied to the AimCLR model to obtain the AimDNNCLR model, and the model is evaluated linearly on the NTU-RGB+D dataset. Compared with the first line model, the proposed method has an average improvement of 3.6% in accuracy.

Keywords Contrastive learning, Nearest neighbor retrieval, Data augmentation, Action recognition, Human skeleton

1 引言

人体动作识别基于视频中完整的动作执行来识别人类动作，是视频理解的核心任务，在生活中已经有着广泛的运用。例如，由动作识别算法驱动的视觉监控系统可以帮助人们通过视频捕捉罪犯，降低犯罪行为造成的风险[1]；视频检索使人们可以通过文本数据（如标题、关键字等）在互联网上搜索到与文本内容相符的视频[2]；基于深度传感器数据[3]的虚拟现实技术[4]在游戏领域吸引了大批各年龄段的人群。现有研究已经探索出视频特征表示的各种模式，如 RGB 帧、光流和人体骨架。在这些模式中，由于人体姿态估计算法的进步[5-7]，以及骨架数据只捕获人体动作信息而不受杂乱背景和明暗变化等上下文干扰[8]的特性，基于骨架数据的动作识别在近年来也越来越受青睐。

在已有的工作中，基于骨架数据的动作识别方法大多以有监督学习框架为主。无论是基于 CNN[9-10]，RNN[11-12]，还是基于 GCN[13-16]的方法，都不可避免地使用了大量标记数据来学习动作表示，但大规模标记良好的骨架数据比 RGB 视频数据更难获得。本文将重点放在自监督设置上，旨在避免 3D 动作表示学习中人工标注的繁重工作量。

自监督表示学习旨在在不使用昂贵的标签或注释的情况下，也能从原始数据中获得样本的稳健表示。基于对比度损失的自监督学习方法已被广泛用于计算机视觉领域[17-19]。一些对比学习方法[20-21]侧重于设计各种新颖的代理任务，以发现隐藏在未标记数据中的模式信息。还有一些对比学习方法[22-23]着重于正负样本对的设计，尽可能地扩展正样本集。然而，在骨架上应用对比学习仍然存在着未解决的问题。一方面，对比学习的成功很大程度上依赖于数据增强[23]。SkeletonCLR[24]仅使用了 Shear 与 Temporal Crop 两个简单的数据增强，获得的运动模式非常有限，无法为模型提供丰富的语义信息，AimCLR[25]则使用了大量的数据增强，但数据增强应该看重质量而非数量。另一方面，来自不同视频的骨架数据被一致认为是负样本[24]，这是不合理的，因为它们与正样本可能属于同一类别，即使在 CrosSCLR[24]中利用最近邻挖掘扩展了正样本集，但是也没有改变将来自不同视频的骨架样本嵌入视为负样本的本质，且在单视图 3D 动作识别上并不适用。

由于普通数据增强所获特征无法保留尽可能多的动作信息或突显重要语义信息，难以为模型训练提供更多指导，且同一样本经不同增强所得嵌入被刻板地设定为正样本，使得损失函数受到不合适甚至错误正样本的指导，从而导致动作识别产生错误。鉴于此，本文提出了基于高质量数据增强的双重最近邻检索视觉表示对比学习（Double Nearest Neighbor Retrieval Contrastive Learning of Visual Representations, DNNCLR）。首先根据多种数据增强理论，提出了一个新的高质量的 Bodypart 数据增强，通过对身体部位进行划分并用正态分布随机数组进行局部填充，增强骨架局部与整体的空间关系，以获得无标签骨架序列的不同动作表示，并对其编码建立基于骨架的对比学习网络。其次，受 RGB 工作中视觉表征的最近邻对比学习方法[26]的启发，提出双重最近邻检索对比损失（DNN Loss），不再对数据增强后的样本直接进行正负例的划分，而是通过最近邻检索（Nearest Neighbor Retrieval, NNR，在一组代表完整数据分布的嵌入上，检索出克服类内与类间差异的嵌入并将其作为模型的正样本。获得全局正样本后，再次利用最近邻检索，挖掘出潜在的正样本以进一步扩充正样本集，将正样本集扩展到普通增强无法覆盖的新数据点，丰富正样本的多样性，构建基于丰富正样本的对比学习网络。

图 1 给出了本文方法的有效性。结合了 Bodypart 增强与双重最近邻检索对比损失的 AimDNNCLR 模型在 NTU-RGB+D 数据集上优于许多其他方法，如 LongT GAN[27]，MS2L[28]，P&C[29]，AS-CAL[30]，SkeletonCLR[24]，CrosSCLR[24]，AimCLR[25]。本文的贡献可以总结为：

1）为给编码器提供带有高级语义信息的新运动模式，关注骨架部分与整体之间的联系，提出了一个高质量的 Bodypart 数据增强，使模型能够学习到更丰富的动作表示。

2）引入最近邻检索，使用一个更合理的正样本扩充策略，进一步提出了一个高效的对比损失函数，即双重最近邻检索对比损失。通过双重最近邻检索，首先得到由一组代表完整数据分布嵌入支撑的正样本，而不是刻板地将同一样本的不同增强视为正样本对，接着再次利用最近邻检索以二次扩充正例，使得最后得到的正样本集覆盖了不同视频的同类样本和同一视频的困难样本。

3）结合所提数据增强和对比损失函数，进一步提出了 DNNCLR 算法，并基于 NTU-RGB+D 数据集，将其应用在主流模型 AimCLR 上进行实验，两者的结合被称为 AimDNNCLR 模型。接着基于线性评估协议对模型效果进行评估，评估结果如图 1 所示，超过了许多前沿模型，验证了所提方法的有效性。

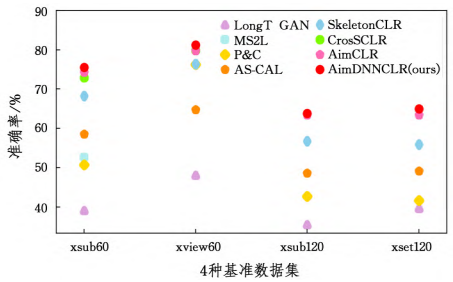


图 1 AimDNNCLR 与其他模型识别精度对比图

Fig. 1 Recognition accuracy comparison of AimDNNCLR and other models

2 相关工作

2.1 基于对比损失的自监督学习方法

自监督学习方法分为生成式和对比式两大类。生成式方法通过编码器将输入特征 u 编码为向量 v，训练解码器从向量 v 重建特征 u；对比式方法也称为对比学习，不同于生成式需要重建特征 u，对比学习是在得到向量 v 后用于计算相似度[31]。相比模型参数容易震荡的生成式方法，对比学习方法训练更为简单。对比学习的目的是将相似样本聚集到一起，将不相似样本推离。在早期工作中，这一目的通常借助代理任务实现[27-34]。而近年来，He 等[22]和 Chen 等[35]受字典查找的启发，提出了一个主流对比学习框架 MoCo，通过构建队列形式的存储库和动量更新编码器来促进自监督对比学习。

2.2 基于骨架数据的自监督动作识别

SkeletonCLR 是遵循 MoCov2[35]框架来实现基于骨骼的单流 3D 动作识别。首先使用两个不同的普通增强，将给定的骨架序列 S 转换为不同的增强序列 x 和 $\hat{x}$，且定义 $x,\hat{x}\in R^{C\times T\times V}$，其中 C,T 和 V 分别是通道、帧和节点的数量。SkeletonCLR使用的增强是 Shear 和 Temporal Crop(见 2.3 节)。接着，两个增强序列分别进入编码器 φ_{Query} 和 φ_{Key}，得到特征向量 $\boldsymbol{h}$ 和 $\hat{\boldsymbol{h}}$，其中 φ_{Key} 基于动量更新，$\boldsymbol{h},\hat{\boldsymbol{h}}\in R^{C_h}$。之后将特征向量 $\boldsymbol{h}$ 和 $\hat{\boldsymbol{h}}$ 通过 $g(\cdot)$ 和 $\hat{g}(\cdot)$ 投影至低维空间，得到 $z=g(h)$ 和 $\hat{z}=\hat{g}(\hat{\boldsymbol{h}})$，其中 $z,\hat{z}\in R^{C_z}$。然后将 Key 分支的嵌入 $\hat{z}$ 存储在先进先出的队列 $Queue=\{\boldsymbol{m}_i\}_{i=1}^{K}$ 中，以消除冗余计算，其中 K 表示队列大小，整个队列作为下一个训练步骤的负样本。按照 MoCov2 中构建样本对的标准，嵌入 z 和 $\hat{z}$ 形成正对，而 $\boldsymbol{u}$ 和 $Queue$ 中的嵌入形成负对。那么 InfoNCE 损失[36]可以写成式(1)的形式，被用于训练网络。

$$L_{\text{Info}}=-\log\frac{\exp(z\cdot\hat{z}/\tau)}{\exp(z\cdot\hat{z}/\tau)+\sum_{i=1}^{K}\exp(z\cdot\boldsymbol{m}_i/\tau)} \tag{1}$$

其中，τ 是温度超参数，“·”为点积运算，计算两个 L2 归一化嵌入 $z,\hat{z}$ 之间的相似性。

计算 InfoNCE 损失之后，φ_{Query} 的参数 θ_z 通过梯度更新，而 φ_{Key} 的参数 $\theta_{\hat{z}}$ 则通过 θ_z 的移动平均值进行更新，这个过程可以表示为式(2)。

$$\theta_{\hat{z}}\leftarrow m\theta_{\hat{z}}+(1-m)\theta_z. \tag{2}$$

其中，$m\in[0,1)$是动量系数，通常趋近于 1，以保持 $Queue$ 队列中嵌入的一致性。

现有基于骨架数据的对比自监督学习比较热门的有 CrosSCLR 模型和 AimCLR 模型，两者均是以 Li 等提出的 SkeletonCLR 模型为基础的进一步探索，本文同样是在此基础上进行研究。Li 等[24]利用多视图互补监督信号，提出了一个跨视图对比学习框架 CrosSCLR，仅在负样本中进行交叉视图一致性知识挖掘，但是这样忽略了当前视图中潜在的相似语义正样本。Guo 等[25]在提出的 AimCLR 模型中使用了大量的数据增强，以获取样本特征的新运动模式，并提出了最近邻挖掘(Nearest Neighbors Mining，NNM)扩展正样本。但过多的数据增强在获得新运动模式的同时也会因反复多次的增强而使样本丢失一些重要的特征，其最近邻挖掘也仅是在负样本队列中进行。本文强调数据增强的设计应该在获得新运动模式的同时保留本身的高级语义特征，正样本的设计也不应受限于数据增强和负样本队列。

2.3 骨架数据增强

处理骨架数据常用的数据增强为 Shear[30]和 Temporal Crop[24]。Shear 将人体关节点的三维坐标向任意角度倾斜，该过程通过式(3)所示的线性映射矩阵实现。

$$\boldsymbol{D}=\begin{bmatrix}1 & d_X^Y & d_X^Z\\ d_Y^X & 1 & d_Y^Z\\ d_Z^X & d_Z^Y & 1\end{bmatrix} \tag{3}$$

其中，$d_X^Y,d_X^Z,d_Y^X,d_Y^Z,d_Z^X,d_Z^Y$ 被称为剪切因子，例如 d_X^Y 对应从 X 轴到 Y 轴的剪切因子，它们的取值范围为$[-\beta,\beta]$，β 通常取值为 0.5[24]。

在图像数据中，Crop[37]是对图像的大小进行裁剪，而应用于骨架数据的 Temporal Crop 是在时间维度上对骨架序列进行裁剪。首先在时间维度上对骨架序列进行对称填充，然后将序列的长度随机裁剪到原始长度。这里定义填充长度为 T/γ，T 表示帧数，γ 是取值为正整数的填充比，γ 通常取 6[24]。用 aug 或 aug' 表示这两个增强，aug 和 aug' 随参数的设置不同而表示两个不同的普通增强。

2.4 计算机视觉中的最近邻检索

从图像检索到无监督特征学习，最近邻检索已成为计算机视觉应用[38-43]的重要工具。与本文工作相关，Han 等[44]证明在 InfoNCE 损失中添加正类语义有利于训练，利用最近邻检索将来自同一数据源不同视图(RGB 和光流)下的相似样本作为互补信息来改善 InfoNCE 损失。AimCLR 和 CrosSCLR 分别在单模态和多模态上，首先计算负样本同正样本的相似性，然后使用最近邻检索从负样本集中检索最近邻嵌入，该嵌入被重新划分为正样本以扩展正样本集。但是该方法得到的正例无法最大程度地覆盖不同视频的同类样本和同一视频的困难样本。本文使用的最近邻检索是在一个特定的支撑集上进行，正样本集由支撑集上的最近邻检索结果决定，既不受限于数据增强方法，又能尽可能地检索到普通增强无法覆盖的新数据点。

3 双重最近邻检索视觉表征对比学习

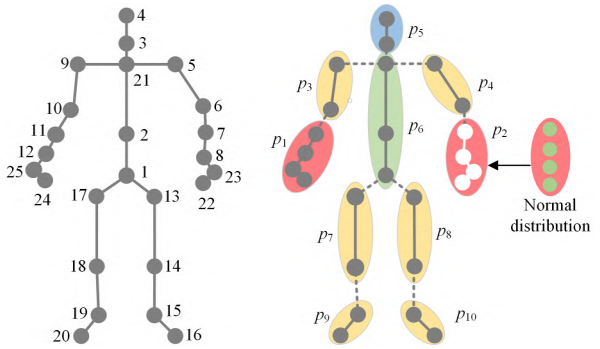
3.1 节详细阐述了 Bodypart 空间数据增强。3.2 节按照最近邻检索的顺序来解释所提正样本扩充策略，第一重最近邻检索依托于近邻正样本集，即支撑集，第二重最近邻检索后可进一步得到双重最近邻检索对比损失函数。

3.1 Bodypart 数据增强

Bodypart 增强是一种关节级空间增强，利用正态分布对骨架数据进行填充，能在一定程度上保持动作信息的一致性，正态分布数据又增强了模型的鲁棒性。以 NTU-RGB+D 数据集的骨架拓扑图为例。如图 2 所示，先将图 2(a)的人体关节划分为 10 个具有物理关系的部分，然后从中随机选取几个部分，以一组标准正态分布数据对原坐标信息进行替换，替换后的人体骨架即为增强后的骨架数据。

如图 2(b)所示，首先根据骨架数据的拓扑信息，按照

身体部位关系把人体关节划分为 10 个部分的不同集合 $P=\{p_1,p_2,p_3,\cdots,p_{10}\}$，其中，$p_1$ 和 p_2 含有 4 个身体关节点，如图 2(b)中的红色部分；而 p_6 含有 3 个身体关节点，如图 2(b)中的绿色部分，由于这 3 个关节在人体运动过程中幅度变化较其他关节缓和，因此将其归为一个部分进行处理；其余关节皆为两两相邻关节，如图 2(b)中的蓝色及黄色部分；注意，除 p_5 和 p_6 外，其余划分的身体部分实际上为 4 对人体几何对称关节，如图中黄色及红色部分。具体而言，使用两个离散均匀分布 $B_p-U(b_1,b_2)$ 和 $B_t-U(b_3,b_4)$ 来确定要进行替换的身体部位，并在保证裁剪骨架片段完整性的有效范围内随机采样一个起始帧 t_s，然后将从 t_s 帧开始的 N_t 帧进行 N_p 个部位的替换，最后得到替换完成的 T 帧骨架序列 S。通过 4.3.1 节实验设定 $N_p=1$，即 b_1 和 b_2 设置为 1，又参考文献[45]将 b_3 和 b_4 分别设为 7 和 11。



(a) NTU-RGB+D 骨架拓扑图　　(b) Bodypart 数据增强

注：骨架拓扑结构来自 NTU-RGB+D 数据集。

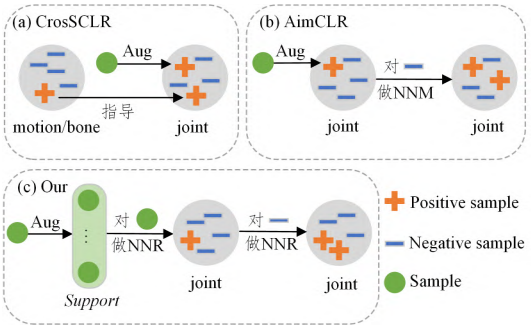
图 2　Bodypart 数据增强示意图(电子版为彩图)

Fig. 2　Bodypart data augmentation diagram

实验中将对原始骨架序列的所有关节坐标进行 2.3 节和 3.1 节介绍的 3 种数据增强，用 aug'' 表示，因此 aug'' 应包含 Shear，Temporal Crop 和 Bodypart 增强。只包含 Shear 和 Temporal Crop 增强时称为普通增强，aug 和 aug' 是两个不同的普通增强，因为 Shear 中的剪切因子是从参数 β 中随机取样得到的，Temporal Crop 进行裁剪时的起始帧受参数 T/γ 的影响[24](见 2.3 节)。数据增强的效果将在 4.3.1 节进行验证。

3.2　双重最近邻检索

在基于骨架数据的自监督对比学习领域中，通常将同一样本序列 S 进行不同增强得到的 z 与 $\hat{z}$ 嵌入视为正样本对，将存储在队列 $Queue$ 中的历史 $\hat{z}$ 嵌入全都作为负样本，使得负样本的数量远远多于正样本，这是不合理的[24]。由式(1)可知，正负样本的比例达到了 $1:K$。最新的关于正样本扩展的方法[24-25]中，CrosSCLR 根据嵌入相似性挖掘跨视图正样本对，如图 3(a)所示，CrosSCLR 将其他视图(motion，bone)中的高嵌入相似性样本所对应的当前视图样本设置为正样本，即利用一个视图中的正样本来指导另一个视图学习。但是多视图网络意味着需要比单视图网络更多的参数并以降低速度为代价来提高准确率，且有极大可能忽略了本视图中的潜在正样本。如图 3(b)所示，AimCLR 中同样指出队列 $Queue$ 中的样本并不一定都是负样本，因此提出最近邻挖掘(NNM)，在负样本中寻找与正样本 $\hat{z}$ 相似性最高的样本来扩充正样本集。但是，以上方法都是将同一样本的不同增强视图作为正样本对，以及在将存储库 $Queue$ 作为负样本的前提下，再寻找负样本中的正样本来弥补正样本稀少的问题。



注：CrosSCLR 通过跨模态查找正样本，低效复杂；AimCLR 只从负样本中查找正样本；本文方法则从全局支撑集与负样本集两个方面入手。

图 3　不同正样本扩充策略对比图

Fig. 3　Comparison of different positive sample expansion strategies

鉴于此，本文受最近邻检索与 NNCLR 模型[26]的启发，提出了双重最近邻检索对比损失，设置了新的正样本扩充策略。该策略在不增加模型参数量的情况下，不提前规定正样本，而是如图 3(c)所示，从一组代表完整数据分布的嵌入上检索出最近邻，得出包含丰富新数据点的正样本，然后再次对负样本进行最近邻检索。

3.2.1　近邻正样本集

与以往的 SkeletonCLR 和 AimCLR 方法[24-25]不同，为了获得更有力的正样本，本文不再从一开始就将同一样本的不同视图 z 与 $\hat{z}$ 视为正样本对，而是不局限于当前批次样本，在多个批次内，以样本视图的最近邻作为整个正样本集。SkeletonCLR 方法的概述见 2.2 节，模型架构如图 4(a)与图4(b)所示，故此处不再过多阐述。

如图 4(c)所示，首先设置一个支撑集($Support$)。支撑集在形式上与存储库 $Queue$ 相同，是一个存储了历史嵌入 $\hat{z}$ 的先进先出队列，其中提供了一组代表完整数据分布的嵌入，作为正样本选取的支撑。然后在支撑集上筛选嵌入 $\hat{z}$ 的最近邻，因此整个正样本集全由样本视图的最近邻充当。

形式上，如图 4(a)和图 4(c)所示，给定一组骨架序列 S，通过 aug 和 aug'' 增强生成两个不同的随机增强视图 x 和 $\hat{x}$，并经过编码器 φ_{Query} 与 φ_{Key} 分别得到特征 $\boldsymbol{h}=\varphi_{Query}(aug(\boldsymbol{x}))$ 与 $\hat{\boldsymbol{h}}=\varphi_{Key}(aug''(\hat{\boldsymbol{x}}))$，再应用多层感知机头 $g(\cdot)$ 和 $\hat{g}(\cdot)$ 得到嵌入 $z=g(\boldsymbol{h})$ 与 $\hat{z}=\hat{g}(\hat{\boldsymbol{h}})$。此时 z 与 $\hat{z}$ 不再是正样本对。随着训练迭代的进行，将 $\hat{z}$ 的历史嵌入存储在支撑集 $Support=\{\tilde{z}_j\}_{j=1}^{L}$ 中，L 为支撑集的大小，在支撑集 $Support$ 上检索 $\hat{z}$ 的最近邻，使用检索得到的 $\hat{z}$ 的最近邻 $NN(\hat{z},sup)$ 作为正样本。

基于 InfoNCE 损失，该过程的损失 L_{NN} 可以定义为式(4)，其中 $NN(\hat{z},sup)$ 由式(5)得到。注意，每个嵌入都经过了 L2 标准化。

$$L_{NN}=-\log\frac{\exp(z\cdot NN(\hat{z},sup)/\tau)}{\exp(z\cdot NN(\hat{z},sup)/\tau)+\sum_{i=1}^{M}\exp(z\cdot m_i/\tau)} \tag{4}$$

$$NN(\hat{z},sup)=\arg\min_{z\in sup}\|\hat{z}-\tilde{z}\|_2 \tag{5}$$

支撑集 $Support$ 与存储库 $Queue$ 都是将样本嵌入保存在内存中，但前者是为了提供数据增强无法覆盖的正样本，而后者被用来充当负样本。为了能够提取到在语义上更近似 $\hat{z}$ 的嵌入，支撑集应尽可能地近似嵌入空间中的完整数据分布，因此需要保持支撑集的大小足够大。支撑集 $Support$ 实现为队列的形式，在每次迭代结束后，当前训练步骤的 N 个嵌入从队列末尾入队，同时队列开头最早的 N 个嵌入出队。

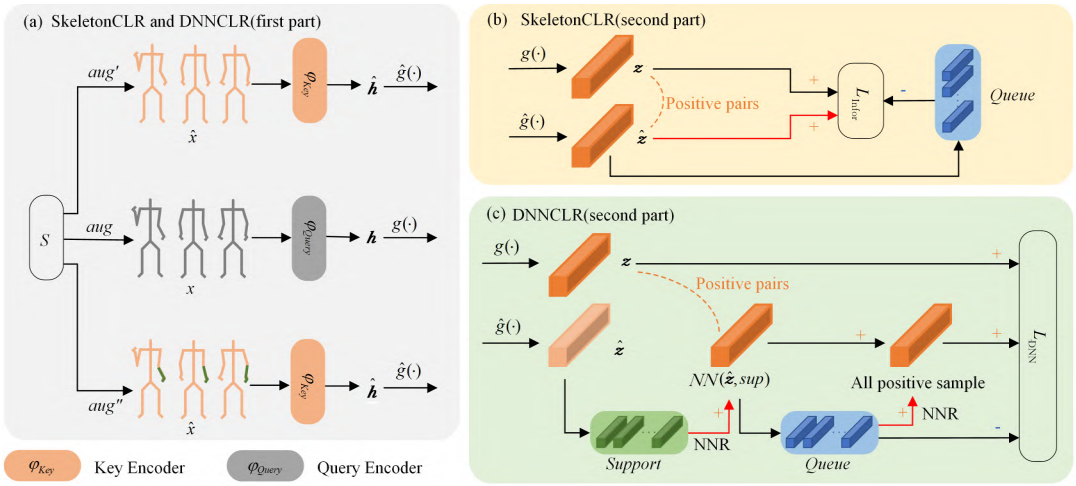


图 4 DNNCLR 算法结构图

Fig. 4 Diagram of DNNCLR algorithm structure

3.2.2 双重最近邻检索对比损失

得到全新的正样本后，通常应继续正负样本在特征空间中的比较，但经过式(4)的处理后得到的正样本仍然有限，故继续从负样本队列 $Queue$ 中挖掘式(5)所获正嵌入的最近邻嵌入，将其填充进正样本集中。如图 4(c)所示，计算嵌入 $NN(\hat{z},sup)$ 与 $Queue$ 之间的相似度，将相似度最高的 k 个嵌入设置为正样本，以二次扩充正样本集。将所获正样本对应的索引集设为 N_+。例如当 $k=1$ 时，取嵌入 $NN(\hat{z},sup)$ 与 $Queue$ 之间的相似度最大值所对应的 $Queue$ 中的样本为正样本，N_+ 中则存储该样本对应索引。这个过程可以表示为式(6)。

$$L_{NN'}=-\log\frac{\sum_{i\in N_+}\exp(z\cdot m_i/\tau)}{\exp(z\cdot NN(\hat{z},sup)/\tau)+\sum_{i=1}^{M}\exp(z\cdot m_i/\tau)} \tag{6}$$

结合式(4)和式(6)可以得到双重最近邻检索对比损失 L_{DNN}。

$$L_{DNN}=-\log\frac{\exp(z\cdot NN(\hat{z},sup)/\tau)+\sum_{i\in N_+}\exp(z\cdot m_i/\tau)}{\exp(z\cdot \mathrm{NN}(\hat{z},sup)/\tau)+\sum_{i=1}^{M}\exp(z\cdot m_i/\tau)} \tag{7}$$

如图 4(c)所示，对于嵌入 $\hat{z}$，先生成一个 $Support$ 队列，再根据嵌入 $NN(\hat{z},sup)$ 生成 $Queue$ 队列，前者作为用于提供正样本集的支撑集，后者作为负样本嵌入的载体，并从这些负样本队列 $Queue$ 中检索出与嵌入 $NN(\hat{z},sup)$ 相似的语义样本，并将其作为正例的补充。观察图 4(c)中正样本的流向，从 $Support$ 中检索到的 $NN(\hat{z},sup)$ 嵌入作为正样本的同时，其迭代的历史嵌入存入 $Queue$ 队列中作为负样本。$Queue$ 中检索出的潜在正样本与 $NN(\hat{z},sup)$ 合并后得到检索出的所有正例。最后这些正例和 $Query$ 分支的 z 嵌入、负样本 $Queue$ 队列中的嵌入一起计算 L_{DNN} 损失。

4 实验结果

本章将所提 Bodypart 数据增强与双重最近邻检索对比损失在主流的 AimCLR 上进行了实验，并按照常用的评估协议，将所提方法与 NTU-RGB+D 数据集上关于线性评估的其他自监督方法进行比较，最后对实验进行了总结。

4.1 实验数据集

将所提方法与 NTU-RGB+D 数据集上关于线性评估的其他自监督方法进行比较，最后对实验进行了总结。

NTU-RGB+D 60(NTU-60)[8] 是一个用于人体动作识别的大规模数据集，包含 4 个模态数据：RGB 视频、深度图序列、3D 骨架数据、红外视频。本文只使用其中的骨架模态数据。该数据集包含 60 个动作类别，56 880 个动作样本，每帧包含 25 个身体主要关节点，以三维坐标的形式存储。数据集中含有两种协议：1) Cross-Subject(xsub)，训练数据集由 40 名受试者中的一半人员提供，测试数据集由另一半受试者提供，训练集和测试集分别有 40 320 和 16 560 个样本；2) Cross-View(xview)，从 ID 为 1 的摄像机收集测试数据集，从 ID 为 2 和 3 的摄像机收集训练数据，训练集和测试集分别有 37 920 和 18 960 个样本。本文遵循推荐的评估方案来评估所提方法。

NTU-RGB+D 120(NTU-120)[46] 数据集是在受试者和动作类别数量上的扩展。动作类别新增到 120 个，动作样本扩大到 114 480 个。该数据集与 NTU-60 有一个相类似的 xsub 基准，该基准的训练集和测试集各由 53 名受试者提供，此外还有一个 Cross-Setup(xset)基准，该基准将偶数

ID摄像机采集的样本作为训练数据，将奇数ID摄像机采集的样本作为测试数据。同样使用推荐评估方案来评估本文方法。

4.2 实验设置

本文实验所用的硬件平台包括128 GB内存和4块TITAN XP显卡，软件平台包括Python 3.7和Pytorch 1.6.0框架。使用的参数配置与文献[25]保持一致，对于数据预处理，本文遵循SkeletonCLR的方法，以进行公平比较。编码器φ_{Query}与φ_{Key}使用ST-GCN网络。对于优化器，使用带有动量(0.999)和重量衰减(0.0001)的SGD。该模型以0.1的学习率训练300个epoch，并在250个epoch处降低到0.01。为了公平比较，此处还生成了3个骨架序列流，即joint，bone和motion。对于3个骨架序列流的所有报告结果，与其他多流GCN方法[25]一样，使用权重[0.6，0.6，0.4]进行加权融合。线性评估运行100个epoch，其学习率初值设置为3，在评估了80个epoch后学习率降为0.3。

本文在进行实验时，将所提方法DNNCLR应用在AimCLR框架上。那么Bodypart增强应用在AimCLR的极端增强分支，实验中的总对比损失可以表示为式(8)。

$$L=\alpha L_{\mathrm{DNN}}+\beta L_{\mathrm{D}} \tag{8}$$

注意，此处的L_{D}损失(D^3M Loss)也根据L_{DNN}损失做出了相应变换，α与β是控制着两部分损失权重的超参数，若无特殊说明两者取值均为1[25]。实验使用两阶段训练策略，在训练的前150个epoch中，先不使用负样本上的最近邻挖掘策略，即此时的损失为$\alpha L_{\mathrm{NN}}+\beta L_{\mathrm{D}}$，$L_{\mathrm{NN}}$如式(4)所示；在150个epoch之后使用完整的总对比损失L，即式(8)。为方便与其他一线方法进行比较，将更改后的模型称为AimDNNCLR。

4.3 消融实验结果分析

本节在NTU-60与NTU-120数据集上进行消融研究，以验证所提数据增强和对比损失产生的影响，并给出支撑集大小等参数不同取值的实验结果。

4.3.1 Bodypart增强与DNN Loss的有效性

为了进一步验证本文方法的有效性，分别对Bodypart(BP)增强和DNN Loss进行了测试。如表1所列，NA表示普通增强(Shear，Temporal Crop)，EA表示极端增强(Gaussian Noise等多种增强)，使用3s-SkeletonCLR(仅勾选w/NA)和3s-AimCLR(仅勾选w/NA和w/EA)作为实验的第一基线和第二基线，3s表示joint，motion，bone融合的结果。第一基线使用普通增强，在xsub和xview上分别达到了75.0%和79.8%的准确率，第二基线使用极端增强分别达到了78.9%和83.8%的准确率，在此基础上添加BP增强，使得xview基准下的模型精度相比第一基线与第二基线分别提升了4%和1%，这表明有效的数据增强在模型中起着重要的作用。DNN Loss使得xsub基准上相对于第一基线准确率提高了3.6%，xview基准上相对于两个基线分别提高了4.1%和0.1%，证明了双重最近邻检索理论的有效性。结合Bodypart增强与DNN Loss，AimDNNCLR模型的准确率超过两个基线，分别达到了79.2%(xsub)和84.6%(xview)的效果。

表1 NTU-60数据集上的消融实验结果

Table 1 Ablation study on NTU-60

w/NA	w/EA	w/BP	w/DNN	NTU-60/%	
				xsub	xview
√				75.0	79.8
√	√			78.9	83.8
√	√	√		78.9	84.8
√	√		√	78.6	83.9
√	√	√	√	**79.2**	**84.6**

表2列出了Bodypart数据增强中随机选取的身体部分数量N_p对AimDNNCLR模型效果的影响。由3.1节可知，N_p的值由离散均匀分布$B_p-U(b_1,b_2)$在每次训练迭代中采样确定，N_p为整数。当b_1和b_2取值为1时，N_p也为1；当$b_1=1,b_2=2$时，N_p可为1或2；当b_1和b_2取值为2时，N_p为2。从表2中可以看出，当身体部分数量N_p取值为1时效果最好。而随着N_p取值的增大，xsub基准的准确率在下降，说明过度的随机化骨骼节点会丢失骨架的运动信息。

表2 参数N_p在模型AimDNNCLR上的影响

Table 2 Influence of parameters N_p on AimDNNCLR

N_p	NTU-60/%	
	xsub	xview
1	**75.5**	81.2
1 or 2	74.9	80.3
2	73.3	**81.5**

4.3.2 支撑集的大小

支撑集*Support*的作用是存储一个近似原始嵌入空间的数据分布，以便提供数据增强无法覆盖的正样本。如表3所列，对于xsub基准，支撑集的大小取值为49152个样本时，达到了表3所列实验中的精度最大值，xview和xset基准的支撑集适合值分别为36864和32768个样本。通过表3的实验可知，支撑集过小或过大都不能为模型提供最佳的性能。若支撑集过小，提供的数据分布达不到近似原始嵌入空间的效果；若支撑集过大，历史嵌入的数量增加不会使模型性能进一步提升，反而会影响模型速度。因此综合考虑模型速度、准确率以及在3个流(joint，bone，motion)上的效果，对于xsub基准选择40960个样本作为支撑集的大小，对于其他基准使用32768个样本作为支撑集大小。

表3 *Support*大小对模型AimDNNCLR的影响

Table 3 Influence of *Support* size on AimDNNCLR

Support size/个	NTU-60		NTU-120	
	xsub	xview	xsub	xview
49152	**79.3**	84.4	**68.9**	70.1
40960	79.2	83.6	68.7	70.2
36864	78.1	**84.7**	68.7	70.2
32768	77.4	84.6	68.7	**70.8**
16384	78.7	84.6	67.9	69.0

4.3.3 AimDNNCLR模型的有效性

本文在NTU-60和NTU-120上进行了实验，表明了结合Bodypart增强和DNN Loss的AimDNNCLR性能。从表4可以看出，对于两种数据集的3个不同输入流，joint和bone流的性能明显增加，而motion流实际上在其他实验配置下，其性能也可以得到很好的提升，只不过本文选择的配置并不是最利于motion流。另外，Bodypart数据增强是关节级空间

增强，而 motion 流侧重于时间维度上的运动信息，过强的空间增强会使特征丢失一些时间信息。对于三流融合的结果，3s-AimDNNCLR 的结果在两个数据集上均超过了 3s-SkeletonCLR 和 3s-AimCLR。

表 4 NTU-RGB+D 数据集上的线性评估结果

Table 4 Linear evaluation results on NTU-RGB+D

（单位：%）

Method	Stream	NTU-60		NTU-120	
		xsub	xview	xsub	xset
SkeletonCLR[24]	J	68.3	76.4	56.8	55.9
AimCLR[25]	J	74.3	79.7	63.4	63.4
AimDNNCLR	J	**75.5**	**81.2**	**64.0**	**66.1**
SkeletonCLR[24]	B	69.4	67.4	48.4	52.0
AimCLR[25]	B	73.2	77.0	62.9	63.4
AimDNNCLR	B	**73.9**	**78.2**	**63.8**	**64.0**
SkeletonCLR[24]	M	53.3	50.8	39.6	40.2
AimCLR[25]	M	66.8	70.6	**57.3**	54.4
AimDNNCLR	M	**63.5**	**72.1**	54.0	**54.5**
SkeletonCLR†[24]	all	75.0	79.8	60.7	62.6
AimCLR†[25]	all	78.9	83.8	68.2	68.8
AimDNNCLR†	all	**79.2**	**84.6**	**68.7**	**70.8**

注：† 表示三流融合结果。

本文方法有效的原因是，Bodypart 增强提供了普通增强没有的骨架局部与整体之间的特征信息，而后 DNNCLR 以强劲的双重最近邻检索得到了经过真正意义上合理划分的正样本。以往方法（见图 4(b)）将增强后的嵌入直接归为正样本，历史嵌入不进行甄别便全归为负样本，但实际上同一视频序列之间不一定有最高的相似性，而不同视频也可能对应着相同动作类别，故通过第一重最近邻检索找出真正高相似性的嵌入归为正样本，再利用第二重最近邻检索寻找可能被归入了负样本中的潜在正样本，此时得到的正样本更加合理化。合理正样本促进了 DNN Loss 对网络的进一步优化，有效识别了动作类别。

图 5 给出了被 AimDNNCLR 模型所纠正样本的可视化结果，前两幅为类别 A54（Point Finger at the other Person）的可视化，后两幅为类别 A57（Touch other Person's Pocket）的可视化。在实验中 AimCLR 将类别 A57 识别为了 A54，而结合 DNNCLR 后大范围纠正了这一错误识别。

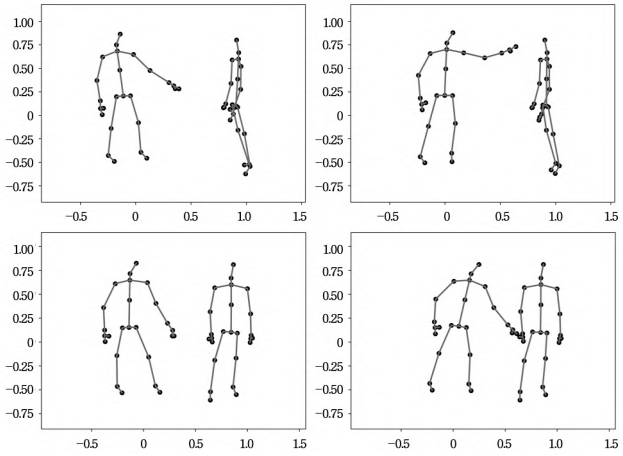


图 5 被 AimDNNCLR 模型所纠正样本的可视化

Fig. 5 Visualization of samples corrected by AimDNNCLR

表 5 列出了本文方法与 AimCLR、SkeletonCLR 的参数量和运算量的对比结果。同 AimCLR 和 SkeletonCLR 相比，本文方法在不增加参数量和运算量的基础上进一步提高了模型精度。AimDNNCLR 整个模型的运算量为 2.2×10^2 GFLOPs，介于 SkeletonCLR 和 CrosSCLR（J+M）模型运算量之间，但 AimDNNCLR 模型精度明显优于两者，体现了本文方法的有效性。

表 5 参数量（Params）和运算量（FLOPs）的对比

Table 5 Comparison of parameters（Params） and computational complexity（FLOPs）

NTU-60	xsub	xview	Params/($\times10^6$)	FLOPs/G
SkeletonCLR[24]	75.0	79.8	1.8	**1.5×10^2**
CrosSCLR(J+M)[24]	74.5	82.1	3.7	2.9×10^2
AimCLR[25]	78.9	83.8	1.8	2.2×10^2
AimDNNCLR(ours)	**79.2**	**84.6**	**1.8**	2.2×10^2

注：J+M 表示同时使用 Joint 和 Motion 视图。

4.4 实验结果分析

4.4.1 定量结果分析

表 6 列出了 AimDNNCLR 在 NTU-60 上的线性评估结果。对于单个流，本文的 AimDNNCLR 方法超过了 AimCLR 方法 1.2%（xsub）和 1.5%（xview），对于 3s-AimDNNCLR，在 xsub 和 xview 基准下也明显优于以往方法。

表 6 NTU-60 数据集上的线性评估结果

Table 6 Linear evaluation results on NTU-60

Method	NTU-60/%	
Single-stream:	xsub	xview
LongT GAN(AAAI 18)[27]	39.1	48.1
MS^{2}L(ACM MM 20)[28]	52.6	—
AS-CAL(Information Sciences 21)[30]	58.5	64.8
P&C(CVPR 20)[29]	50.7	76.3
SeBiReNet(ECCV 20)[33]	—	79.7
SkeletonCLR(CVPR 21)[24]	68.3	76.4
AimCLR(AAAI 22)[25]	74.3	79.7
AimDNNCLR(ours)	**75.5**	**81.2**
Three-stream:		
3s-SkeletonCLR(CVPR 21)[24]	75.0	79.8
3s-Colorization(ICCV 21)[48]	75.2	83.1
3s-CrosSCLR(CVPR 21)[24]	77.8	83.4
3s-AimCLR(AAAI 22)[25]	78.9	83.8
3s-AimDNNCLR(ours)	**79.2**	**84.6**

表 7 列出了 AimDNNCLR 在 NTU-120 上的线性评估结果与其他基于骨架数据的自监督学习方法的比较结果。本文的 AimDNNCLR 超过了表中 NTU-120 上的其他自监督方法，如 P&C[29]，AS-CAL[30]，CrosSCLR[24]，ISC[47]，AimCLR[25]。三流融合结果在 xsub 和 xset 基准上的准确率分别达到了 68.7%和 70.8%。实验结果表明，不需要跨模态嵌入的帮助，仅在单个模态下通过双重最近邻检索提供正样本也能学到更好的特征表示。

为验证 Bodypart 增强使用正态分布随机数作为填充获得的增强信息更为合理，将正态分布随机数填充（Normal distribution）与零填充方法（Zeros）进行比较。结果如表 8 所列，零填充方法对 xview 基准略有改善，对 xsub 基准则起了反作用，而使用正态分布填充对模型的改善效果更显著。

表 7 NTU-120 数据集上的线性评估结果

Table 7 Linear evaluation results on NTU-120

Method	NTU-120/%	
	xsub	xset
P&C(CVPR 20)	42.7	41.7
AS-CAL(Information Sciences 21)	48.6	49.2
3s-CrosSCLR(CVPR 21)	67.9	66.7
ISC(ACM MM 21)[47]	67.9	67.1
3s-AimCLR(AAAI 22)	68.2	68.8
3s-AimDNNCLR(ours)	**68.7**	**70.8**

表 8 正态分布填充与零填充的比较

Table 8 Comparison between normal distribution padding and zero padding

NTU-60-J	xsub	xview
SkeletonCLR[24]	68.3	76.4
AimCLR[25]	74.3	79.7
Zeros	74.1	80.2
Normal distribution(ours)	**75.5**	**81.2**

AimDNNCLR 使用的正样本一部分是通过支撑集获得的全局正样本，另一部分是从全局正样本更新的队列中检索得出，比直接由数据增强后的嵌入充当的正样本更加合理。为说明这一点，将 AimDNNCLR 与 AimCLR($top\_K=2$)进行比较，后者是最近邻挖掘数为 2 的 AimCLR。实验结果如表 9 所列，虽然 AimCLR($top\_K=2$)对精度略有提升，但本文方法的效果明显优于 AimCLR($top\_K=2$)策略，这说明了本文方法能捕获更优质的正样本。

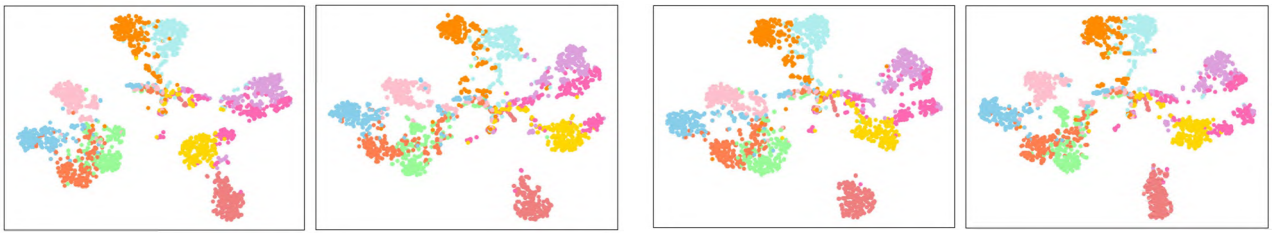
表 9 AimDNNCLR 与 AimCLR 的比较

Table 9 Comparison between AimDNNCLR and AimCLR

NTU-60-J	xsub	xview
AimCLR[25]	74.3	79.7
AimCLR($top\_K=2$)	74.5	81.0
AimDNNCLR (ours)	**75.5**	**81.2**

4.4.2 定性结果分析

为了更直观地展示本文所提 Bodypart 增强和 DNN Loss 函数的效果，利用 t-SNE[49] 降维算法，将预训练了 300 个 epoch 后的 AimDNNCLR，以及分别单独使用了 Bodypart 数据增强以及 DNN Loss 的 AimDNNCLR(BP)和 AimDNNCLR(DNN)的可视化嵌入分布，与 AimCLR 的可视化嵌入分布进行了比较。从 NTU-60 的 xview 基准上的 60 个类中选取 10 个类别的特征嵌入进行比较，得到图 6 所示的结果。可以看出，相比 AimCLR，随着 Bodypart 和 DNN Loss 的加入，同类嵌入聚集更为紧凑，不同类嵌入彼此更加远离。



(a) AimCLR (b) AimDNNCLR(BP) (c) AimDNNCLR(DNN) (d) AimDNNCLR

图 6 NTU-60(xview)数据集上对嵌入特征进行 t-SNE 可视化

Fig. 6 t-SNE visualization of embedded features on NTU-60(xview)

结束语 本文中，首先提出了一个关节级空间增强 Bodypart，以正态分布随机数对具有物理连接意义的 10 个身体部分随机进行正态分布填充，带来了新的语义信息，提高了特征表示的泛化性。其次，摒弃将同一样本的不同增强作为正对的传统方法，进行双重正样本的选取，对经过数据增强的原始嵌入利用最近邻检索筛选出最相似的正样本，减少了真正正样本被疏漏的可能性，接着从由历史特征更新的负样本队列中检索出与第一重正样本相似性最高的嵌入并入正样本集中。双重检索使得正样本集的设计更加合理，进一步提出了双重最近邻检索对比损失，迫使模型学习更多的一般特征表示，使得模型优化更加合理。最后，结合 Bodypart 增强与 DNN Loss 构建 DNNCLR 模型，并在 NTU-RGB+D 数据集上进行实验，在各个线性评估协议中模型准确率超过了文中涉及的其他前沿模型，说明了本文方法的有效性。

参 考 文 献

[1] SHOITAN R, MOUSSA M M, EL NEMR H A. Attribute based spatio-temporal person retrieval in video surveillance[J]. Alexandria Engineering Journal, 2023, 63: 441-454.

[2] TRAN M T, HOANG-XUAN N, TRANG-TRUNG H P, et al. V-FIRST: A Flexible Interactive Retrieval System for Video at VBS 2022[C]// MultiMedia Modeling: 28th International Conference. Cham: Springer International Publishing, 2022: 562-568.

[3] LIU W, BAO Q, SUN Y, et al. Recent advances of monocular 2d and 3d human pose estimation: a deep learning perspective[J]. ACM Computing Surveys, 2022, 55(4): 1-41.

[4] RAUTER M, ABSEHER C, SAFAR M. Augmenting virtual reality with near real world objects[C]// 2019 IEEE Conference on Virtual Reality and 3D User Interfaces(VR). USA: IEEE, 2019: 1134-1135.

[5] CAO Z, HIDALGO G, SIMON T, et al. Openpose: Realtime multi-person 2d pose estimation using part affinity fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 172-186.

[6] FANG H S, XIE S Q, TAI Y W, et al. Rmpe: Regional multi-person pose estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2334-2343.

[7] XU J W,YU Z B,NI B B,et al. Deep kinematics analysis for monocular 3d human pose estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA:IEEE,2020:899-908.

[8] SHAHROUDY A,LIU J,NG T T,et al. NTU RGB+D:A large scale dataset for 3d human activity analysis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas,USA:IEEE,2016:1010-1019.

[9] KE Q H,BENNAMOUN M,AN S J,et al. A new representation of skeleton sequences for 3d action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,USA:IEEE,2017:3288-3297.

[10] LIU M Y,LIU H,CHEN C. Enhanced skeleton visualization for view invariant human action recognition[J]. Pattern Recognition,2017,68:346-362.

[11] SONG S J,LAN C L,XING J L,et al. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection[J]. IEEE Transactions on Image Processing,2018,27(7):3459-3471.

[12] ZHANG P F,LAN C L,XING J L,et al. View adaptive neural networks for high performance skeleton-based human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,41(8):1963-1978.

[13] YAN S J,XIONG Y J,LIN D H. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Thirty-second AAAI Conference on Artificial Intelligence. New Orleans,USA:AAAI Press,2018:7444-7452.

[14] SHI L,ZHANG Y F,CHENG J,et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE,2019:12026-12035.

[15] SI C Y,CHEN W T,WANG W,et al. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE,2019:1227-1236.

[16] CHEN Z,LI S C,YANG B,et al. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver,Canada:AAAI Press,2021,35(2):1113-1122.

[17] ISLAM A,LUNDELL B,SAWHNEY H,et al. Self-supervised Learning with Local Contrastive Loss for Detection and Semantic Segmentation[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa,HI,USA:IEEE,2023:5624-5633.

[18] JIAO Y,YANG K,SONG D J,et al. Timeautoad:Autonomous anomaly detection with self-supervised contrastive loss for multivariate time series[J]. IEEE Transactions on Network Science and Engineering,2022,9(3):1604-1619.

[19] WICKSTRØM K,KAMPFFMEYER M,MIKALSEN K Ø,et al. Mixing up contrastive learning:Self-supervised representation learning for time series[J]. Pattern Recognition Letters,2022,155:54-61.

[20] ALBELWI S. Survey on self-supervised learning:auxiliary pretext tasks and contrastive learning methods in imaging[J]. Entropy,2022,24(4):551.

[21] KOMODAKIS N,GIDARIS S. Unsupervised representation learning by predicting image rotations[C]// International Conference on Learning Representations. Canada:ICLR,2018.

[22] HE K M,FAN H Q,WU Y X,et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA:IEEE,2020:9729-9738.

[23] CHEN T,KORNBLITH S,NOROUZI M,et al. A simple framework for contrastive learning of visual representations[C]// International Conference on Machine Learning. Virtual Event:PMLR,2020:1597-1607.

[24] LI L G,WANG M S,NI B B,et al. 3d human action representation learning via cross-view consistency pursuit[C]// Proceedings of the IEEE FConference on Computer Vision and Pattern Recognition. Nashville,USA:IEEE,2021:4741-4750.

[25] GUO T Y,LIU H,CHEN Z,et al. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Event:AAAI Press,2022,36(1):762-770.

[26] DWIBEDI D,AYTAR Y,TOMPSON J,et al. With a little help from my friends:Nearest-neighbor contrastive learning of visual representations[C]// Proceedings of the IEEE International Conference on Computer Vision. Montreal,Canada:IEEE,2021:9588-9597.

[27] ZHENG N G,WEN J,LIU R S,et al. Unsupervised representation learning with long-term dynamics for skeleton based action recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New Orleans,USA:AAAI Press,2018,32(1):2644-2651.

[28] LIN L L,SONG S J,YANG W H,et al. Ms2l:Multi-task self-supervised learning for skeleton based action recognition[C]// Proceedings of the 28th ACM International Conference on Multimedia. Seattle,USA:ACM,2020:2490-2498.

[29] SU K,LIU X L,SHLIZERMAN E. Predict & cluster:Unsupervised skeleton based action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle,USA:IEEE,2020:9631-9640.

[30] RAO H C,XU S H,HU X P,et al. Augmented skeleton based contrastive action learning with momentum LSTM for unsupervised action recognition[J]. Information Sciences,2021,569:90-109.

[31] LIU X,ZHANG F J,HOU Z Y,et al. Self-supervised learning:Generative or contrastive[J]. IEEE Transactions on Knowledge and Data Engineering,2023,35(1):857-876.

[32] MISRA I,ZITNICK C L,HEBERT M. Shuffle and learn:unsu-

pervised learning using temporal order verification[C]// European Conference on Computer Vision. Amsterdam, Netherlands: Springer, Cham, 2016: 527-544.

[33] NIE Q, LIU Z W, LIU Y H. Unsupervised 3d human pose representation with viewpoint and pose disentanglement[C]// European Conference on Computer Vision. Glasgow, UK: Springer, Cham, 2020: 102-118.

[34] NOROOZI M, FAVARO P. Unsupervised learning of visual representations by solving jigsaw puzzles[C]// European Conference on Computer Vision. Amsterdam, Netherlands: Springer, Cham, 2016: 69-84.

[35] CHEN X L, FAN H Q, GIRSHICK R, et al. Improved baselines with momentum contrastive learning[J]. arXiv: 2003. 04297, 2020.

[36] OORD A, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[J]. arXiv: 1807. 03748, 2018.

[37] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of Big Data, 2019, 6(1): 1-48.

[38] MEMMESHEIMER R, HÄRING S, THEISEN N, et al. Skeleton-DML: deep metric learning for skeleton-based one-shot action recognition[C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. USA: IEEE, 2022: 3702-3710.

[39] LIN C C, LIN K, WANG L J, et al. Cross-modal representation learning for zero-shot action recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2022: 19978-19988.

[40] WU C R, PENG Q L, LEE J, et al. Effective hierarchical clustering based on structural similarities in nearest neighbor graphs [J]. Knowledge-Based Systems, 2021, 228: 107295.

[41] DANG Z Y, DENG C, YANG X, et al. Nearest neighbor matching for deep clustering[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2021: 13693-13702.

[42] CARON M, TOUVRON H, MISRA I, et al. Emerging properties in self-supervised vision transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2021: 9650-9660.

[43] WU Z R, EFROS A A, YU S X. Improving generalization via scalable neighborhood component analysis[C]// Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018: 685-701.

[44] HAN T, XIE W, ZISSERMAN A. Self-supervised co-training for video representation learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 5679-5690.

[45] CHEN Z, LIU H, GUO T Y, et al. Contrastive Learning from Spatio-Temporal Mixed Skeleton Sequences for Self-Supervised Skeleton-Based Action Recognition [J]. arXiv: 2207. 03065, 2022.

[46] LIU J, 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, 2019, 42(10): 2684-2701.

[47] THOKER F M, DOUGHTY H, SNOEK C G M. Skeleton-contrastive 3D action representation learning[C]// Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM, 2021: 1655-1663.

[48] YANG S Y, LIU J, LU S J, et al. Skeleton cloud colorization for unsupervised 3d action representation learning[C]// Proceedings of the IEEE International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 13423-13433.

[49] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

WU Yushan, born in 1998, postgraduate, is a member of China Computer Federation. Her main research interests include action recognition, self-supervised learning and applied mathematics, etc.

XU Zengmin, born in 1981, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include human action recognition, multimodal semantic understanding, computer vision and pattern recognition, etc.

（责任编辑：喻藜）