跨视图时序对比学习的自监督视频表征算法

王露露[1,2]，徐增敏[1,2,3]，张雪莲[1,2]，蒙儒省[3]，卢 涛[4]

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 541004
3. 桂林安维科技有限公司, 广西 桂林 541010
4. 武汉工程大学 计算机科学与工程学院 智能机器人湖北省重点实验室，武汉 430205

摘 要：现有的自监督表征算法主要关注视频帧之间的短期运动特性，但是帧间动作序列的变化幅度较小，而且单视图数据因语义受限影响深度特征表达能力，视频动作中丰富的多视图信息未被充分利用。为此提出基于跨视图语义一致性的时序对比学习算法，自监督学习 RGB 帧和光流场两种数据中蕴含的动作时序变化特性，主要思路为：设计局部时序对比学习方法，采用不同正负样本划分策略，挖掘同一实例不重叠片段之间的时序相关性和判别可分性，增强细粒度特征表达能力；研究全局对比学习方法，通过跨视图语义协同训练来增加正样本，学习多实例不同视图的语义一致性，提高模型的泛化能力。通过两个下游任务对模型效果进行评估，在 UCF101 和 HMDB51 数据集的实验结果表明，所提方法在动作识别和视频检索任务上，较前沿主流方法平均提升了 2~3.5 个百分点。

关键词：自监督学习；视频表征学习；时序对比学习；局部对比学习；跨视图协同
文献标志码: A **中图分类号:**TP391.41; TP183

Cross-View Temporal Contrastive Learning for Self-Supervised Video Representation

WANG Lulu[1,2], XU Zengmin[1,2,3], ZHANG Xuelian[1,2], MENG Ruxing[3], LU Tao[4]

1. Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin,Guangxi 541004, China
2. Center for Applied Mathematics of Guangxi (GUET), Guilin,Guangxi 541004, China
3. Anview.ai, Guilin,Guangxi 541010, China
4. Hubei Key Laboratory of Intelligent Robot,School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China

Abstract： The existing self-supervised representation algorithms mainly focus on the short-term motion characteristics between video frames, but the variation range of the action sequence between frames is small, and the depth feature expression ability of single-view data is affected due to semantic limitations, so the rich multi-view information in video actions is not fully utilized. Therefore, a temporal contrast learning algorithm based on cross-view semantic consistency is proposed to self-supervised learn the action temporal variation characteristics

基金项目：国家自然科学基金(61862015,62072350);广西自然科学基金(2024GXNSFAA010493);广西科技基地和人才专项(AD23023002, AD21220114);广西重点研发计划项目(AB17195025)。
作者简介：王露露(1996—)，女，硕士，CCF 学生会员，研究方向为计算机视觉；徐增敏(1981—)，通信作者，男，博士，副教授，CCF 专业会员，研究方向为机器学习、人工智能，E-mail: xzm@guet.edu.cn。
收稿日期：2023-12-04；**修回日期：**2024-03-04

embedded in both RGB frames and optical flow field data. The main ideas are as follows: to design a local temporal contrast learning method, adopt different positive and negative sample division strategies to explore the temporal correlation and discriminative differentiability between non-overlapping segments of the same instance, and enhance the fine-grained feature expression capability; to study the global contrast learning method to increase the positive samples by cross-view semantic co-training, learn the semantic consistency of different views of multiple instances, and improve the generalization ability of the model. The model performance is evaluated through two downstream tasks, and the experimental results on UCF101 and HMDB51 datasets show that the proposed method improves on average 2~3.5percentage points over cutting-edge mainstream methods on action recognition and video retrieval tasks.

Key words：self-supervised learning; video representation learning; temporal contrastive learning; local contrastive learning; cross-view co-training

视频理解是计算机视觉中的重要任务，近年来随着互联网技术的飞速发展，多媒体信息数据源源不断产生，庞大的视频数据量给数据分析与理解带来了巨大的挑战。基于监督学习的视频表征方法在许多视觉识别任务中取得了优越的性能[1-2]，这有赖于大量手工标注的训练样本数据，如 YouTube8M[3]、Kinetics[4]等数据集，获取这些数据集需要极大的人力与时间成本，而大量未标注的视频数据很容易在互联网上获得。因此，利用自监督学习从未标注的原始数据出发，充分挖掘自身的监督信息，探索特征间的相关性。

由于自监督学习的不断发展，基于实例判别的对比学习方法[5-9]在图像表征领域取得了显著的成功，甚至优于监督算法的性能。因此，一些工作[10-15]考虑将实例对比学习应用到视频领域，通常将来自同一视频的片段视为正样本，来自不同视频的片段作为负样本，通过 InfoNCE[16]损失训练模型，使得模型能够区分不同类别的视频。虽然仅实例对比学习就可以在许多视频理解任务中取得显著效果，但该方法鼓励模型学习同一实例中相似的特征，忽略了视频的时序变化特性，且基于片段提取的特征较粗糙，在一些细粒度场景的任务中，有很大的局限性。

最近的一些工作验证了时序信息可以进一步提高模型学习表征的质量。例如，使用自适应丢弃帧的时序增强策略[11]或时序分割采样训练样本[14]来捕获视频的全局上下文信息，避免在时间距离较远的片段间强制特征不变性；设计前置任务来学习视频时序的变化，包括时间顺序识别[12,17]、视频帧采样速率识别[18-19]等。另外一些工作致力于增加表征的细粒度，通过视频序列帧级的特征相似度对比[20]，迫使模型区分同一实例在时序上的动态变化。尽管局部时序建模在一定程度上得到了改进，但仍存在两个问题：一方面帧级的对比学习旨在建模相邻帧之间的差异，然而连续帧之间的动作变化并不显著，很难学习到时序信息反而影响实例判别的效果；另一方面，在早期的对比学习方法中，跨视图的语义信息交互在很大程度上被忽视了，单视图由于语义受限不利于模型的聚类效果。

鉴于此，本文提出跨视图时序对比学习的自监督视频表征算法 CVTCL(cross-view temporal contrastive learning)，利用多视图语义信息挖掘正样本，并在对比学习中充分考虑视频序列的时序特性，增强模型学习表征的泛化性。首先，基于实例判别对比学习对每个视图的样本进行独立训练，提取多个视图的嵌入特征；其次，考虑到视频动作序列的时序连续性，提出局部时序对比学习方法，使用同一视频不重叠片段的聚合特征作为局部特征，通过不同的正负样本划分方式，促使模型学习视频序列的时序运动变化，增强表征的细粒度和时序性；最后，结合多视图语义一致性挖掘方法，将一个视图在嵌入空间中的相似样本映射到另一个视图作为正样本，引入视图间互补的伪监督约束，促进视图之间的信息共享，构建语义级的多实例对比学习网络。本文的主要贡献如下：

(1) 考虑到视频序列的连续性与时序性，提出局部时序对比学习方法，结合全局片段的长期语义依赖与局部片段间的时序运动特性，增强了仅实例判别的自监督视频表征。

(2) 引入光流来充分挖掘数据的语义信息，结合跨视图协同训练增加正样本多样性，构建语义级对比学习，实现视图内数据关联及视图间语义协同交互。

(3) 结合跨视图全局一致性挖掘网络与局部时序对比学习方法，提出新的自监督视频表征框架，并在视频分类和检索两个下游任务上对模型效果进行评估。

1 相关工作

1.1 对比自监督学习算法

目前的自监督学习算法中，基于实例判别的对比学习方法[5-15]为广大学者所研究。对比学习的核心思想是通过对两个事物的相似或不相似进行编码来构建表征[21]，其关键在于如何划分正负样本对，在实例判别对比学习中，将来自同一实例的样本作为正样本，不同实例的样本作为负样本。随着 MoCo[5]、SimCLR[6]等方法大量涌现，对比学习显示了在自监督图像表征领域的巨大潜力，近年来被广泛应用于视频领域。很多自监督视频表征算法[10-15]在实例判别对比学习基础上构建模型框架，并取得了显著的效果。本文受前人启发，在实例判别对比学习的基础上构建自监督模型框架。

1.2 多视图自监督学习算法

视频数据的多种视图 RGB 帧和光流都含有丰富的语义信息，因此多视图的自监督学习[10,22-25]也吸引了很多研究者的兴趣。该方法侧重于利用视图之间的重叠信息为模型提供自监督，以提高下游任务的性能。其中，CoCLR[10]提出一个自监督协同训练机制，利用 RGB 和光流两视图之间的语义相关性来互相补充正样本，通过 RGB 与光流网络协同训练来提升模型提取到的表征；MaCLR[22]通过光流网络分支来引导 RGB 特征更关注运动前景；MSCL[23]利用光流含有的运动信息，通过 RGB 和光流两视图的特征交叉对比学习，来提升自监督学习的运动感知力。受 CoCLR[10]启发，本文利用 RGB 和光流网络协同训练的方式，进一步提升模型表征的泛化性，但该方法只利用片段的长期语义依赖进行全局对比学习，缺少对视频时序运动信息的建模，本文通过提出的局部时序对比损失进一步改进。

1.3 细粒度的时序表征学习算法

视频数据具有时间连续性，也可以作为自监督学习的监督信号，为了充分利用视频的时序信息，最近的一些工作通过设计各种时序相关的前置任务，验证了时序信息可以进一步提高模型学习表征的质量。其中，TCGL[12]提出了时间对比图学习方法，将视频帧和片段顺序的先验知识集成到图结构中；VCLR[14]通过时序分割采样训练样本，设计片段帧顺序识别任务实现时序建模；SCVRL[17]引入打乱帧顺序的前置任务；TCLR[20]比较了不同时间跨度的特征。学习局部动作变化也是视频理解的一个重要课题，但在自监督学习领域，现有的方法对细粒度的时间特征关注并不够。在视频片段中连续帧的运动变化并不明显，很难学习到视频的时序信息，一些方法通过采用不同策略来筛选高运动片段。例如，TCGL[12]、MSCL[23]分别利用频域分析和运动差分来采样变化幅度大的片段，通过片段内的对比学习获得精细化特征。此外，在目标跟踪领域，一些工作[26-27]通过对粗粒度与细粒度的表征进行联合学习，实现了更好的性能。

尽管这些工作确实有所改进，但设置额外的前置任务和采样策略使得模型更复杂，且无法利用光流中的运动信息。受 TCLR[20]启发，使用不重叠短片段的特征作为局部特征，与其不同的是在局部对比学习中正负样本的划分，且利用视频多视图的相关性信息，提出基于跨视图时序对比学习的自监督视频表征算法，通过多视图全局语义互补增强正样本的多样性，结合局部时序动作变化，从而关联全局与局部语义信息，获得精细化的动作特征表达。

2 CVTCL 算法

基于对比学习的自监督方法需要构建正样本和负样本，通过度量正负样本之间的距离来实现自监督学习，使用损失函数优化模型，使得正样本对之间的距离远远大于负样本对之间的距离。因此，如何划分正负样本是本章的关键。本章利用多视图语义一致性挖掘和时序对比学习，结合协同训练机制，构建基于对比学习的跨视图自监督视频表征学习模型，整体框架如图 1 所示，包含两个层次的对比学习：(1) 跨视图全局对比学习(2.1 节、2.2 节)：在实例判别对比学习的基础上建模视频全局语义依赖，结合 RGB 与光流两分支网络协同训练，增加正样本的多样性；(2) 局部时序对比学习(2.3 节)：该模块为本文所提方法，显示在图 1 的虚线框中。为了解决连续帧运动变化不显著问题，使用同一视频不重叠片段的聚合特征作为局部特征，增强特征的细粒度和时序特性。不同于 CoCLR[10]算法，本文虽然同样使用了多视图协同的基本框架来构建网络模型，但在此基础上新增了局部时序对比学习模块，来建模视频的精细化时序特征。

2.1 实例对比学习

本节基于实例判别的对比损失 InfoNCE[16]来捕获视频的全局语义依赖。给定含有 N 个视频实例的数据集，设为 $V=\{v_1,v_2,\cdots,v_N\}$，从帧序列中提取对应光流图 $M=\{m_1,m_2,\cdots,m_N\}$。如图 1 所示，从一个视频

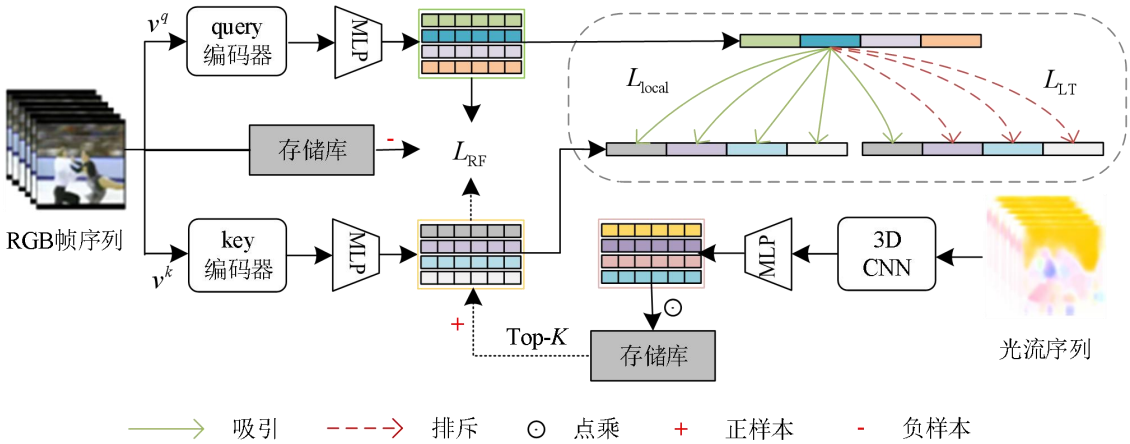

图 1 CVTCL 的算法框架

Fig.1 Architecture of CVTCL

序列 v_i 中随机采样一个片段 v_i^q，经过数据增强后得到视图 v_i^k，形成正样本对 $\{v_i^q, v_i^k\}$，其余视频片段为负样本 N^-，即 $v_j \in N^-, j \neq i$，且以队列的形式储存在存储库中，然后将样本分别输入到特征编码器 f_q 和 f_k，得到特征的可视化表示，通过 MLP 层 $g(\cdot)$ 将特征投影到低维嵌入空间进行相似度比较。

全局对比学习通过最大化同一视频不同视图的语义一致性来进行实例判别。为简单起见，将 v_i^q 的嵌入特征表示为 $z = g_q(f_q(v_i^q))$，正样本的嵌入特征为 $z^+ = g_k(f_k(v_i^k))$，负样本的嵌入特征为 $n_j = g_k(f_k(v_j^-))$，则实例对比损失为：

$$L_{\text{InfoNCE}} = L_{\text{NCE}}(z, z^+, N^-) = -\log \frac{\exp(z \cdot z^+ / \tau)}{\exp(z \cdot z^+ / \tau) + \sum_{j=1}^{N} \exp(z \cdot n_j / \tau)} \quad (1)$$

式中：$z \cdot z^+$ 是两向量间的点积；τ 是温度系数；N 为负样本个数；光流分支的实例对比损失可类似定义。通过实例对比学习，模型学习在嵌入空间中将相似的实例拉近，将不同的实例推远。

2.2 跨视图全局对比学习

单视图数据由于语义信息受限而影响深度特征表达能力，且视频动作中丰富的多视图信息未被充分利用。为了挖掘多视图间的语义一致性，本节在 CoCLR[10] 的启发下，使用 RGB 和光流两个分支网络协同训练，利用一种视图为另一种视图采集难正例，增加正样本的多样性。在 RGB 视图中，跨视图对比损失定义如下：

$$L_{\text{RF}} = L_{\text{NCE}}(z_{v_i}, z_{P_{1i}}, v_{N^-}) \quad (2)$$

式中：z_{v_i} 为样本 v_i 的特征；$z_{P_{1i}}$ 为正样本特征；v_{N^-} 为负样本；正样本集 P_{1i} 定义如下：

$$P_{1i} = \{v_i^a, v_k \mid k \in \text{topK}(z_{m_i} \cdot z_{m_j}), \forall j \in [1, N]\} \quad (3)$$

式中：v_i^a 为 RGB 样本 v_i 的数据增强视图；$z_{m_i} \cdot z_{m_j}$ 是光流视图中第 i 个和第 j 个视频之间的相似度；topK(·) 是指从 N 个样本中选择 K 个最相似的样本特征，并输出样本索引值，即 RGB 分支的正样本集为 v_i 的数据增强加上 v_i 在光流特征空间中的前 K 个最近邻。

同样的，在光流视图中，RGB 特征空间中相似的实例也可以作为正样本帮助光流网络进行更好地表征学习。类似地，光流视图中跨视图损失函数如下：

$$L_{\text{FR}} = L_{\text{NCE}}(z_{m_i}, z_{P_{2i}}, m_{N^-}) \quad (4)$$

$$P_{2i} = \{m_i^a, m_k \mid k \in \text{topK}(z_{v_i} \cdot z_{v_j}), \forall j \in [1, N]\} \quad (5)$$

式中：z_{m_i} 为光流样本 m_i 的特征；$z_{P_{2i}}$ 为正样本特征；m_{N^-} 为负样本；m_i^a 为 m_i 的数据增强视图；光流分支的正样本集 P_{2i}，包括 m_i 的数据增强加上 m_i 在 RGB 特征空间中的前 K 个最近邻，所涉及的参数含义与公式(3)相同。

2.3 局部时序对比学习

基于全局对比学习的特征很难建模视频动作的局部运动和时序性变化，在细粒度的场景中模型特征表达能力较差。为了充分捕捉同一视频动作之间的差异，使模型学习到帧之间的时序信息，设计局部时序对比学习模块(local temporal contrastive learning, LTCL)。主要包括两个对比学习任务：局部对比任务，学习同一视频不同局部片段之间的相似性，区分来自不同实例的特征，增加表征的细粒度；局部时序对比任务，学习同一视频不重叠局部片段之间的区别，增加表征的时序性。

为了解决连续帧运动信息不显著的问题，受TCLR[20]启发，本文使用不同时间戳片段的聚合特征作为局部特征，与其不同的是正负样本的设置，具体划分如图2所示。给定一个视频片段 $X=\{x_1,x_2,\cdots,x_n\}$，假设随机采样4个不重叠的局部片段 $\{s_1,s_2,s_3,s_4\}$，局部对比学习中 s_1 为样本，$\{s_1^a,s_2,s_3,s_4\}$ 视为该样本的正样本，其余视频片段为负样本 N^-，则片段间局部对比损失为：

$$L_{\text{local}}=\frac{1}{4}L_{\text{NCE}}(z_1,\{z_1^a,z_2,z_3,z_4\},z_{N^-}) \quad (6)$$

式中：z_i 为样本 s_i 的特征 $(i=1,\cdots,4)$；a 为数据增强。

通过局部对比学习增加了表征的细粒度，为了进一步学习到视频中的时序变化信息，设计局部时序对比学习，将 s_1 作为输入样本，s_1^a 视为该样本的正样本，$\{s_2,s_3,s_4\}$ 视为负样本，则片段间局部时序对比损失为：

$$L_{\text{LT}}=L_{\text{NCE}}(z_1,z_1^a,\{z_2,z_3,z_4\}) \quad (7)$$

因此，局部时序对比模块的损失为：

$$L_{\text{LTCL}}=L_{\text{local}}+L_{\text{LT}} \quad (8)$$

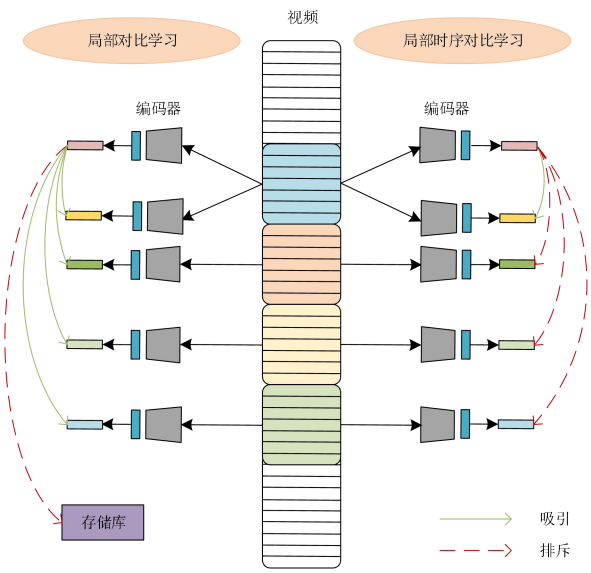

图2 局部时序对比学习模块

Fig.2 Local temporal contrastive learning

2.4 算法实现

在早期训练阶段，模型不够稳定，无法为另一视图提供可靠的正样本，因此，CVTCL算法采用两阶段训练策略，主要步骤如下所示：

1) 初始化阶段：使用损失函数 $L_1=\alpha L_{\text{InfoNCE}}+\beta L_{\text{LTCL}}$ 对RGB和光流两路径进行单独训练，形成初步的表征；

2) 协同训练阶段：使用损失函数 $L_2=\alpha L_{\text{RF}}+\beta L_{\text{LTCL}}$ 协同训练RGB和光流两个初始化后的模型，挖掘另一视图的 k 近邻，交替利用一个网络为另一个网络进行正样本的采样。其中，α 和 β 是平衡损失的系数，设置 α=β=1 使得算法更具有一般性。交替采样过程具体来说：首先将光流样本输入到编码器中提取特征，然后将该特征和存储库中其他样本的特征进行相似度比较，选取在光流嵌入空间中Top-K 个相近的实例（K 为超参数），最后将这 K 个在光流特征中相近的样本作为正样本用于RGB网络中对比损失的更新。算法1详细描述了局部时序对比学习在初始化阶段的实现过程。

算法1 LTCL模块的PyTorch实现

输入：视频数据矩阵 $\boldsymbol{X}$ [B, N, C, T, H, W]；
超参数 d，k，m，t
输出：N 维损失向量值，编码器参数更新

```
f_k.params = f_q.params # 初始化特征编码器 f_q, f_k
for s1, s2 in loader: # 加载 N 个样本数据
    随机数据增强得到 s1_q, s1_k, s2_q, s2_k
z1_q= f_q.forward(s1_q) # 编码得到 q 的特征
s1_k, s2_k = f_k.forward(s1_q, s2_q) # 编码得到 k 的特征
s1_k, s2_k = s1_k.detach( ), s2_k.detach( ) # k 不梯度更新
#正样本 logits: 2N×1
l_pos1=einsum('nc,nc->n',[s1_q,s1_k]).unsqueeze(-1)
l_pos2=einsum('nc,nc->n',[s1_q,s2_k]).unsqueeze(-1)
l_pos = torch.cat((l_pos1, l_pos2), dim=0)
#负样本 logits: N×K
l_neg = einsum('nc,ck->nk', [s1_q, queue])
# 计算局部对比 logits: 2N×(K+1)
logits = torch.cat((l_pos, l_neg.repeat(2, 1)), dim=1)
logits /= t # 应用温度系数
# 计算局部时序对比 logits1: N×2
l_pos = einsum('nc,nc->n',[s1_q, s1_k]).unsqueeze(-1)
  l_neg1 = einsum('nc,nc->n',[s1_q,.s2_k]).unsqueeze(-1)
logits1 = torch.cat((l_pos, l_neg1), dim=-1)
logits1 /= t # 应用温度系数
# 生成伪标签
labels = zeros(logits.shape[0], dtype=torch.long)
labels1=zeros(logits1.shape[0],dtype=torch.long)
# 计算对比损失
l_local=CrossEntropyLoss(logits,labels) # 公式(6)
l_LT=CrossEntropyLoss(logits1,labels1) # 公式(7)
loss = l_local + l_LT # 公式(8)
loss.backward( ) # 根据 loss 来计算网络参数的梯度
update(f_q.params) # f_q 使用优化器更新参数
# f_k 采用动量更新参数，m 为动量系数
f_k.params = m*f_k.params+(1-m)*f_q.params
self.dequeue_and_enqueue(k) # 更新队列
```

3 实验验证与分析

在本节中，首先描述 CVTCL 训练的数据集和实验设置。然后在 UCF101 数据集上对模型的各模块和参数进行了消融实验，以证明所提方法的有效性。最后，在动作识别和视频检索两个下游任务上对所提模型进行实验评估，并与其他模型进行比较。

3.1 实验数据集

本文在两个公开动作识别数据集上进行实验，分别为 UCF101[28]和 HMDB51[29]。其中 UCF101 数据集由用户上传的真实动作视频构成，该数据集包含 101 个动作类别，共 13320 个视频样本，具有很大的动作多样性，并且在摄像机运动、物体外观、姿势和杂乱背景等方面存在很大差异。HMDB51 数据集是从各种互联网资源和数字化电影中收集形成，该数据集的动作主要是人类日常行为，包含 6766 个视频样本，共 51 个动作类别，每种动作包含 101 个视频片段。

本文使用 UCF101 数据集进行模型自监督预训练，对于下游评估任务，在 UCF101 split1、HMDB51 split1 上进行基准测试。

3.2 实验设置

本文实验所用的硬件平台包括运行内存 128GB 的 8 块 TITAN XP 显卡，软件平台包括 Python3.8 和 PyTorch 1.7.1 框架。使用的参数配置与文献[10]保持一致，使用 S3D[30]网络作为所有实验的特征编码器，在 CVTCL 训练期间，参考 SimCLR[6]，在编码器后添加一个非线性投影层，并将其移除之后的网络用于下游任务。使用 32 帧的 RGB(或 flow)采样片段作为输入，视频片段的空间分辨率为 128×128 像素。在数据处理方面，应用了随机裁剪、水平翻转、高斯模糊和颜色抖动等空间数据增强策略，采用随机时间裁剪来模拟数据时间维度的自然变化，即从源视频的随机时间戳裁剪输入视频片段。光流计算采用 TV-L1 算法[31]，预处理过程与文献[32]相同。

在模型初始化阶段，运用基于 InfoNCE 的全局和局部时序对比学习训练 RGB 网络和光流网络 300 个 epochs，每个视频中随机采样一个片段进行训练，即输入实例总数为训练集中的视频数量。实验采用 MoCo[5]的动量更新队列来存储大量的特征。在协同训练阶段，对模型进行两个循环的训练，每个循环包括 200 个 epochs，即 RGB 和光流网络分别训练 100 个 epochs，并从另一个网络中进行难正例挖掘。为了优化模型，使用初始学习率 0.001，权重衰减为 0.0001，优化器为 Adam。

实验遵循文献[10]中的评估协议，包括两种类型的下游任务。(1)动作识别。增加一个分类器进行分类，然后用线性评估和微调训练整个模型，并评估 Top-1 准确性（Top-1 Acc）。(2)视频检索。直接使用预训练的骨干网络提取特征，进行最近邻（nearest-neighbour，NN）检索，实验使用来自测试集的视频表征来查询训练集中的 k-NN，并比较在 k 处的召回率(R@k)，即如果前 k 个最近邻中包含 1 个同类视频，则为正确的检索。

3.3 消融实验

本节的所有实验均在 UCF101 split1 数据集上进行，把动作识别和视频检索作为表征质量的主要度量，其中动作识别任务遵循线性评估协议。

3.3.1 验证初始化阶段各个对比损失的有效性

在本节中分析了在初始化阶段不同损失函数的有效性，对于所有的自监督预训练模型使用相同的参数设置，训练 300 个 epochs。验证结果如表 1 所示，在实例对比损失的基础上加入局部对比学习时，由于视频的 RGB 帧序列在时序上是变化的，而局部对比学习认为不同时间戳片段的特征是相似的，所以使得精度有所下降；在实例对比损失的基础上加入局部时序对比损失后，两个分支网络的性能都有提升；当三个损失函数同时使用时，局部时序对比损失能够学习视频帧序列在时序上的变化，弥补了前两个损失的不足，使得模型的精度达到最优。得到相应结论：相比于只使用实例判别损失进行模型初始化，加入局部时序对比学习模块后，RGB 和光流两分支网络都在一定程度上提升了学习到的表征质量。

表 1 初始化阶段关于 CVTCL 不同设计的消融实验

Table 1 Ablation study on different designs of CVTCL in initialization stage

对比损失				动作识别	视频检索	
L_{InfoNCE}	L_{local}	L_{LT}	输入	Top-1 Acc/%	R@1/%	R@5/%
✓	×	×	RGB	43.1	30.2	46.6
✓	×	×	flow	63.9	42.9	66.5
✓	✓	×	RGB	39.2	27.9	43.2
✓	✓	×	flow	**65.7**	41.7	64.1
✓	×	✓	RGB	45.1	30.4	48.5
✓	×	✓	flow	65.5	40.4	64.2
✓	✓	✓	RGB	**47.1**	**32.1**	**51.1**
✓	✓	✓	flow	65.6	**44.0**	**66.7**

3.3.2 验证所提局部时序对比学习的有效性

在本节中分析了不同训练阶段加入局部时序对比

学习(LTCL)模块时，模型的识别和检索效果的变化。上一节已验证在初始化训练阶段加入 LTCL 模型效果最好，因此本节选用加入 LTCL 的初始化模型进行协同训练 100 个 epochs，对比加入和去除 LTCL 的模型性能。如表 2 所示，分别在 RGB 和 flow 两个网络进行验证，得到结论：对于 RGB 网络在协同训练阶段加入 LTCL 模块在动作识别的效果优于未加 LTCL 的模型，但在最近邻视频检索上稍弱于不加 LTCL 的模型；而对于光流网络，协同训练加入 LTCL 后模型效果更差，这是由于同一实例的光流序列变化幅度小，将同一实例的不同局部片段拉远，会降低网络学习的能力。因此，在实际模型协同训练过程中，对于 RGB 分支网络使用 LTCL 方法，而光流网络不使用。

表 2 LTCL 在不同训练阶段的消融实验

Table 2 Ablation study of LTCL in different training stages

训练阶段			动作识别	视频检索		
初始	协同	输入	Top-1 Acc/%	R@1/%	R@5/%	R@10/%
✓	×	RGB	54.6	**43.9**	**63.1**	72.0
✓	×	flow	**69.1**	**54.7**	**74.1**	**81.6**
✓	✓	RGB	**57.2**	42.6	62.9	**72.5**
✓	✓	flow	67.3	50.6	72.4	78.0

3.3.3 验证协同训练阶段近邻数 *K* 的有效性

在协同训练阶段，公式(3)和公式(5)中的近邻数 K 决定着跨视图挖掘正样本的数量，对模型性能的影响至关重要。改变参数 K 的值，使用光流视图为 RGB 网络挖掘正样本训练 100 个 epochs，结果如表 3 所示，将 K 分别取值为 0、1、5、10。可以看出，跨视图挖掘近邻数 K 越大，模型的效果越好，表明利用另一个视图寻找的高置信度的样本可以有效增强同类样本的聚类效果，挖掘的正样本越多，模型基于伪标签信息的学习效果越好。但随着正样本的增多，计算复杂度更高，因此本文选择设置 K=5 进行协同训练。

3.3.4 验证协同训练挖掘正样本的有效性

为了更好证明采用 RGB 和光流网络协同训练的有效性，通过在 UCF101 数据集上进行视频检索任务，监测模型交替训练的过程。如图 3 所示，横轴表示训练阶段 epoch，前 300 个 epochs 是两个分支独立初始化阶段，然后交替训练两个循环，在交替训练过程中，同一个 epoch 区间固定一个网络为另一个网络提供正样本进行模型优化。随着协同训练，RGB 和光流网络视频检索性能 R@k 的增强表示，在嵌入空间中相同类别的视频被拉进，网络学习表征的质量得到了提高。

表 3 CVTCL 关于 Top-*K* 的消融实验

Table 3 Ablation study of CVTCL with Top-*K*

Top-*K*		动作识别	视频检索		
K	输入	Top-1 Acc/%	R@1/%	R@5/%	R@10/%
0	RGB	46.4	31.5	50.0	56.0
1	RGB	51.6	39.9	59.1	68.0
5	RGB	54.6	43.9	63.1	72.0
10	RGB	**55.8**	**45.0**	**64.9**	**72.9**

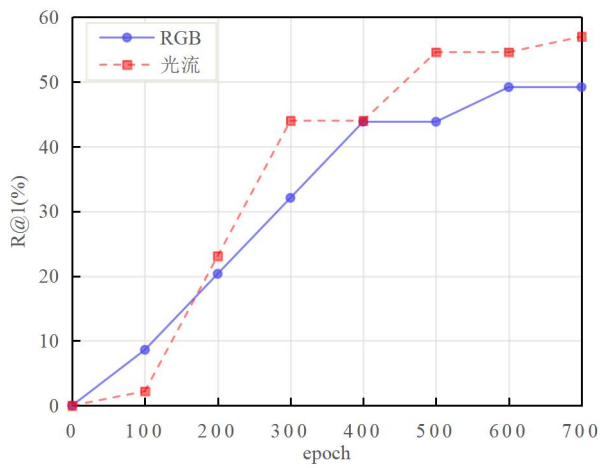

图 3 CVTCL 在 UCF101 数据集上的训练过程

Fig.3 CVTCL training process on UCF101

3.4 实验结果分析

本文的跨视图时序对比学习算法基于 CoCLR[10]，在该网络的基础上增加了局部时序对比学习模块，并使用 S3D[30]作为模型的主干网络。在 UCF101 和 HMDB51 两个数据集上使用动作识别和最近邻视频检索两个下游任务来评估自监督学习的表征。

3.4.1 定量结果分析

1) 动作识别

如表 4 所示，将本文的方法与其他基于自监督学习方面的工作进行了比较，主要对比了 MoCo[5]、VCOP[33]、CoCLR[10]、PacePred[18]、STS[34]、DSM[35]、TCLR[20]、TCGL[12]与 CACL[36]方法。所提方法 CVTCL 在 RGB 数据上动作识别的结果为 74.5%和 44.7%，两个分支网络微调后，在 UCF101 和 HMDB51 光流数据上结果为 81.8%和 52.1%，该方法在光流网络上取得了最优的识别效果，在 RGB 网络性能较 CoCLR 有所下降，由于训练设备的差异，复现 CoCLR 不能达到其所述精度。CVTCL 在运动差异明显的光流视图数据上更有效，表明该方法有利于模型对细粒度特征的建

模。

2) 最近邻视频检索

与动作识别任务类似，在 UCF101 数据集上进行自监督预训练，然后使用该网络提取的特征用于最近邻视频检索任务。本文使用 UCF101 和 HMDB51 数据集的测试集样本，分别从它们的训练集中查询 K 个最近邻视频样本，使用 R@k 作为该任务的评估指标。结果如表 5 所示。CVTCL 在 RGB 数据上取得了 48.3%和 19.6%的 R@1 检索精度，在光流数据上取得了 57.0%和 27.0%的 R@1 检索精度。结果表明增加局部时序对在光流数据上的提升效果更明显，比 CoCLR 方法分别提升了 5.1 和 3.1 个百分点的精度。相较于动作识别任务，CVTCL 在视频检索任务上性能更好，表明该方法更有助于聚合相同类别的嵌入特征。

3.4.2 定性结果分析

本文使用 CVTCL 预训练 500 个 epochs 后的表征，可视化模型在视频检索任务上的效果。如图 4 所示，左侧为从 UCF101 split1 测试集中选取 5 个视频裁剪样本，右侧为在训练集中查询到的对应类别 Top-3 最近邻视频序列。可以看出，本文所提方法可以很好地聚合相同类别的表征，分离不同类别的表征，模型具有检索相同语义类别视频序列的能力。

表 4 UCF101 和 HMDB51 数据集上的动作识别精度对比

Table 4 Comparison of accuracy for action classification on UCF101 and HMDB51

方法	年份	数据集	输入	网络	Top-1 Acc/%	
					UCF101	HMDB51
MoCo[5]	2020	UCF101	16×112	C3D	60.5	27.2
VCOP[33]	2020	UCF101	16×112	C3D	65.6	28.4
CoCLR[10]	2020	UCF101	32×128	S3D	81.4	**52.1**
PacePred[18]	2020	UCF101	112×112	R(2+1)D	75.9	35.9
STS[34]	2021	UCF101	16×112	C3D	69.3	34.2
DSM[35]	2021	UCF101	16×112	C3D	70.3	40.5
TCLR[20]	2022	UCF101	16×112	C3D	76.1	48.6
TCGL[12]	2022	UCF101	16×112	C3D	77.4	39.5
CACL[36]	2022	UCF101	16×112	R3D	77.5	43.8
Ours-RGB	-	UCF101	32×128	S3D	74.5	44.7
Ours-flow	-	UCF101	32×128	S3D	**81.8**	**52.1**

表 5 UCF101 和 HMDB51 数据集上的视频检索精度对比

Table 5 Comparison of accuracy for video retrieval on UCF101 and HMDB51

方法	年份	数据集	网络	UCF101			HMDB51		
				R@1/%	R@5/%	R@10/%	R@1/%	R@5/%	R@10/%
VCOP[33]	2020	UCF101	R18	14.1	30.30	40.4	7.6	22.9	34.4
CoCLR-RGB[10]	2020	UCF101	S3D	51.8	69.40	76.6	23.2	43.2	53.5
CoCLR-Flow[10]	2020	UCF101	S3D	51.9	68.50	75.0	23.9	47.3	58.3
PacePred[18]	2020	UCF101	R(2+1)D	31.9	49.70	59.2	12.9	32.2	45.4
BE[37]	2021	UCF101	R18	11.9	31.30	44.5	-	-	-
VCLR[14]	2021	UCF101	R50	46.8	61.80	70.4	17.6	38.6	51.1
TCLR[20]	2022	UCF101	R18	56.2	72.20	79.0	22.8	45.4	57.8
TCGL[12]	2022	UCF101	R18	23.4	42.20	51.9	11.7	28.9	40.5
CACL[36]	2022	UCF101	C3D	43.2	61.10	69.9	17.3	40.2	53.8
TransRank-ST[38]	2022	UCF101	R18	46.5	63.70	72.8	19.4	45.4	59.1
Ours-RGB	-	UCF101	S3D	48.3	68.10	76.8	19.6	43.3	57.0
Ours-flow	-	UCF101	S3D	**57.0**	**75.87**	**82.9**	**27.0**	**51.4**	**64.2**

图 4 使用 CVTCL 表征最近邻检索的结果

Fig.4 Nearest neighbour retrieval results with CVTCL representation

4 结论

本文提出了一种新的自监督视频表征学习方法 CVTCL，通过设计一个多任务的损失函数来学习更具表现力的特征表示。首先利用视图内与视图间的全局对比学习方法，捕获视频长期语义依赖和跨模态一致性，获得泛化能力较好的表征；其次利用同一实例不重叠片段间的时序相关性，提出局部时序对比学习方法，增加视频表征的时序性和细粒度；最后将自监督预训练模型迁移到下游任务，来提升视频检索和识别的性能，在 UCF101 和 HMDB51 两个数据集上对模型效果进行验证。经过实验证明，本文方法在针对细粒度场景的视频检索和动作识别任务上具有较好的可行性。在以后的工作中，将继续研究更有效地表征学习方法和不同层次的特征融合策略。

参考文献:

[1] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//Proceedings of the European Conference on Computer Vision, Amsterdam, Oct 10-16, 2016. Berlin: Springer, 2016: 20-36.

[2] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision, Santiago, Dec 7-13, 2015. New York: IEEE Press, 2015: 4489-4497.

[3] ABU-EL-HAIJA S, KOTHARI N, LEE J, et al. Youtube-8m: a large-scale video classification benchmark[EB/OL].(2016-09-27)[2022-11-16]. https://arxiv.org/abs/1609.08675.

[4] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[EB/OL].(2017-05-19)[2022-11-16]. https://arxiv.org/abs/1705.06950.

[5] HE K, FAN H, WU Y, et al. Momentum contrast for uns-upervised visual representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, Jun 14-19, 2020. New York: IEEE Press, 2020: 9729-9738.

[6] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the International Conference on Machine Learning, Jul 12-18, 2020. New York: PMLR, 2020: 1597-1607.

[7] GRILL J B, STRUB F, ALTCHÉ F, et al. Bootstrap your own latent a new approach to self-supervised learning[C]// Advances in Neural Information Processing Systems, 2020: 21271-21284.

[8] CARON M, MISRA I, MAIRAL J, et al. Unsupervised learning of visual features by contrasting cluster assignments[C]// Advances in Neural Information Processing Systems, 2020: 9912-9924.

[9] CHEN X, HE K. Exploring simple siamese representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 19-25, 2021. New York: IEEE Press, 2021: 15750-15758.

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

[11] PAN T, SONG Y, YANG T, et al. Videomoco: contrastive

video representation learning with temporally adversarial examples[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 19-25, 2021. New York: IEEE Press, 2021: 11205-11214.

[12] LIU Y, WANG K, LAN H, et al. Temporal contrastive graph learning for video action recognition and retrieval[EB/OL].(2021-03-17)[2022-11-05]. https://arxiv.org/abs/2101.00820.

[13] QIAN R, MENG T, GONG B, et al. Spatiotemporal contrastive video representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 19-25, 2021. New York: IEEE Press, 2021: 6964-6974.

[14] KUANG H, ZHU Y, ZHANG Z, et al. Video contrastive learning with global context[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Oct 10-17, 2021. New York: IEEE Press, 2021: 3195-3204.

[15] TAO L, WANG X, YAMASAKI T. Self-supervised video representation learning using inter-intra contrastive framework[C]//Proceedings of the 28th ACM International Conference on Multimedia, Seattle, Oct 12-16, 2020. New York: Association for Computing Machinery, 2020: 2193-2201.

[16] OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[EB/OL].(2019-01-22)[2022-11-05]. https://arxiv.org/abs/1807.03748.

[17] DORKENWALD M, XIAO F, BRATTOLI B, et al. SCVRL: shuffled contrastive video representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 19-21, 2022. New York: IEEE Press, 2022: 4132-4141.

[18] WANG J, JIAO J, LIU Y H. Self-supervised video representation learning by pace prediction[C]//Proceedings of the European Conference on Computer Vision, Aug 23-28, 2020. Cham: Springer, 2020: 504-521.

[19] SINGH A, CHAKRABORTY O, VARSHNEY A, et al. Semi-supervised action recognition with temporal contrastive learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 19-25, 2021. New York: IEEE Press, 2021: 10389-10399.

[20] DAVE I, GUPTA R, RIZVE M N, et al. TCLR: temporal contrastive learning for video representation[J]. Computer Vision and Image Understanding, 2022, 219: 103406-103414.

[21] JIAO J, CAI Y, ALSHARID M, et al. Self-supervised con-trastive video-speech representation learning for ultrasound[C]// Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, Lima, Oct 4-8, 2020. Cham: Springer, 2020: 534-543.

[22] XIAO F, TIGHE J, MODOLO D. MaCLR: motion-aware contrastive learning of representations for videos[C]//Proceedings of the European Conference on Computer Vision, Tel-Aviv, Oct 23-27, 2022. Cham: Springer, 2022: 353-370.

[23] NI J, ZHOU N, QIN J, et al. Motion sensitive contrastive learning for self-supervised video representation[C]//Proceedings of the European Conference on Computer Vision, Tel-Aviv, Oct 23-27, 2022. Cham: Springer, 2022: 457-474.

[24] YAO T, ZHANG Y, QIU Z, et al. Seco: exploring sequence supervision for unsupervised representation learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Feb 2-9, 2021. Menlo Park: AAAI Press, 2021: 10656-10664.

[25] SUN C, MYERS A, VONDRICK C, et al. Videobert: a joint model for video and language representation learning[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway, NJ: IEEE, 2019: 7464-7473.

[26] ZHANG D, ZHENG Z, LI M, et al. Reinforced similarity learning: siamese relation networks for robust object tracking[C]//Proceedings of the 28th ACM International Conference on Multimedia, Seattle, Oct 12-16, 2020. New York: ACM, 2020: 294-303.

[27] ZHANG D , ZHENG Z. Joint representation learning with deep quadruplet network for real-time visual tracking[C]//Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, Jul 19-24, 2020. Piscataway: IEEE, 2020: 1-8.

[28] SOOMRO K, ZAMIR A R, SHAH M. A dataset of 101 human action classes from videos in the wild[EB/OL].(2012-12-03)[2022-11-05]. https://arxiv.org/abs/1212.0402.

[29] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Nov 6-13, 2011. New York: IEEE Press, 2011: 2556-2563.

[30] XIE S, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning for video understanding[C]//Proceedings of the European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 305-321.

[31] ZACH C, POCK T, BISCHOF H. A duality based approach for realtime tv-l 1 optical flow[C]//Pattern Recog-nition: 29th DAGM Symposium, Heidelberg, Sep 12-14, 2007. Berlin: Springer, 2007: 214-223.

[32] CARREIRA J, ZISSERMAN A. Quo vadis, action

recognition? a new model and the kinetics dataset[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Jul 21-26, 2017. New York: IEEE Press, 2017: 6299-6308.

[33] XU D, XIAO J, ZHAO Z, et al. Self-supervised spatiotemporal learning via video clip order prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 16-20, 2019. New York: IEEE Press, 2019: 10334-10343.

[34] WANG J, JIAO J, BAO L, et al. Self-supervised video representation learning by uncovering spatio-temporal statistics[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3791-3806.

[35] WANG J, GAO Y, LI K, et al. Enhancing unsupervised video representation learning by decoupling the scene and the motion[C]//Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, Feb 2-9, 2021. Menlo Park, CA: AAAI Press, 2021: 10129-10137.

[36] GUO S, XIONG Z, ZHONG Y, et al. Cross-architecture self-supervised video representation learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 19-21, 2022. New York: IEEE Press, 2022: 19270-19279.

[37] WANG J, GAO Y, LI K, et al. Removing the background by adding the background: Towards background robust self-supervised video representation learning[C]//Pro-ceedings of the IEEE/CVF Conference on Computer V-ision and Pattern Recognition, Virtual, Jun 19-25, 2021. New York: IEEE Press, 2021: 11804-11813.

[38] DUAN H, ZHAO N, CHEN K, et al. Transrank: Self-supervised video representation learning via ranking-based transformation recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 19-21, 2022. New York: IEEE Press, 2022: 3000-3010.