Computer Science and Application 计算机科学与应用, 2023, 13(3), 433-443
Published Online March 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.133041

Hans汉斯

多任务对比学习的自监督视频表达

单东风，于　磊，骆文杰，熊思璇，刘家仁，吴克伟

合肥工业大学计算机与信息学院，安徽　合肥

收稿日期：2023年2月16日；录用日期：2023年3月17日；发布日期：2023年3月28日

摘　要

现有的自监督学习使用单一的空间或时间代理任务。单一的代理任务，从未标记的数据中提供单一的监督信号，不足以描述视频表示学习的空间特征和时间特征之间的差异。在本文中，我们提出了一个多任务对比学习方法，它通过对多个时空代理任务的对比学习，在时空自注意力的情况下学习有区别的时空特征。不同的空间代理任务学习不同的空间特征，包括空间旋转和空间拼图。不同的时间代理任务学习不同的时间特征，包括时间顺序和时间节奏。我们将视频表示为每个代理任务的多个不同特征，并设计基于代理任务的对比损失来分离一个视频中学习的空间特征和时间特征。基于代理任务的对比损失鼓励不同代理任务学习不同的特征，同一代理任务学习相似的特征，可以学习到同一视频中每个代理任务的判别特征。实验表明，在UCF-101数据集和HMDB-51数据集的行为识别上优于现有的自监督学习方法。

关键词

自监督，空间特征，时间特征，多任务对比学习方法，时空自注意力

Multitask Contrastive Learning for Self-Supervised Video Representation

Dongfeng Shan, Lei Yu, Wenjie Luo, Sixuan Xiong, Jiaren Liu, Kewei Wu

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei Anhui

Received: Feb. 16th, 2023; accepted: Mar. 17th, 2023; published: Mar. 28th, 2023

Abstract

Most existing self-supervised works use a single spatial or temporal pretext task. A single pretext task, providing single supervision from unlabeled data, is insufficient to describe the difference between spatial features and temporal features for video representation learning. In this paper, we propose an attentive spatiotemporal contrastive learning network, which learns discrimina-

文章引用: 单东风, 于磊, 骆文杰, 熊思璇, 刘家仁, 吴克伟. 多任务对比学习的自监督视频表达[J]. 计算机科学与应用, 2023, 13(3): 433-443. DOI: 10.12677/csa.2023.133041

tive spatial-temporal features with self-attention by contrastive learning between multiple spatial and temporal pretext tasks. Different spatial features are learned by multiple spatial pretext tasks, including spatial rotation, and spatial jigsaw. Different temporal features are learned by multiple temporal pretext tasks, including temporal order, and temporal pace. We represent video as multiple different features for each pretext task, and design pretext task-based contrastive loss to separate the spatial feature and the temporal feature learned in one video. The pretext task-based contrastive loss encourages the different pretext tasks to learn dissimilar features and the same pretext task to learn similar features, which can learn the discriminative features for each pretext task in one video. Experiments show that it outperforms existing self-supervised learning methods for behavior recognition on the UCF-101 dataset and the HMDB-51 dataset.

Keywords

Self-Supervised, Spatial Feature, Temporal Feature, Multitask Contrastive Learning Method, Spatiotemporal Self-Attention

1. 引言

视频表达的自监督学习可以从未标记的数据中提取时空特征，不仅减少了昂贵的手工标注需求，而且可以为下游的有监督任务学习到泛化能力很强的视频表达，因此在视频分析应用中具有很大的必要性。

自监督方法生成无注释标签作为代理任务。在现有的代理任务中，有的方法通过单一的空间变换学习空间特征[1] [2]，有的方法通过单一的时间变换来学习时间特征，其中有时间顺序预测[3]、时间速度预测[4] [5]、时间连续性预测[6]，有的方法通过时空变换变换[7]来学习时空特征，有的方法通过多个代理任务[8]学习一些时间特征。然而上述方法忽略了描述不同视频之间的特征差异，这可能会在动作相似的视频中造成混淆。针对这一问题，一些自监督方法通过基于实例的对比学习对视频特征进行改进，这种方法可以为每个代理任务学习不同视频的不同特征[9] [10]。有些方法通过设计多个代理任务进行对比学习来学习视频特征[11] [12]。然而，上述方法学习到的特征忽略了描述在多个代理任务中学习到的空间特征和时间特征之间的差异。因此设计一个具有对比学习的自监督视频表示模型仍然是一个挑战。

图 1 显示了我们的多任务包含 2 个时间代理任务(时间速度代理任务、时间排序代理任务)和 2 个空间代理任务(空间旋转代理任务、空间拼图代理任务)。在没有对比学习的情况下，多任务学习方法无法学习时间代理任务和空间代理任务的差异性，从而导致视频特征表达精度损失。因此本文引入对比学习可以对多任务学习到的时空特征进行分离，避免信息混淆，捕捉更详细的时空信息。

本文中提出了多任务对比学习方法，该方法通过学习不同时空代理任务中不同的特征来解决基于代理任务的对比学习问题。然后，为了增强时空特征，该方法引入注意力特征提取模块来捕捉视频帧空间通道关系和视频帧之间的时间关系。

首先，本文利用多任务学习特征，让网络可以同时学习到时间特征和空间特征，包括时间顺序预测、时间速度预测、空间旋转预测和空间拼图预测。本文设计任务级的对比学习来区别时间代理任务和空间代理任务学习的特征。本文设置在同一类代理任务(2 个时间代理任务或者 2 个空间代理任务)中学习到的特征作为正例，在不同类代理任务(1 个时间代理任务和 1 个空间代理任务)学习到的特征作为负例。与基于实例的对比损失相比，本文的代理任务的对比损失使网络能够描述一个视频中更多的时空变化。

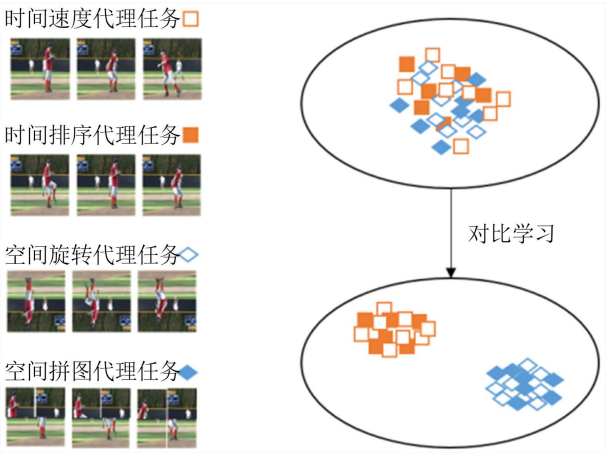

Figure 1. Multitask comparison learning example diagram
图 1. 多任务对比学习示例图

其次，本文在网络中设置注意力特征提取模块来增强网络学习到的时空信息，其中注意力特征提取模块包含空间通道注意力模块和时间注意力模块。空间通道自注意模块学习空间之间的关系，指出空间中感兴趣的对象。时间自注意力模块学习帧之间的时间关系，从而表明视频中的关键帧。空间关系和时间关系分别提供了空间权重和时间权重，以集成所有帧的特征作为视频表示，充分描述了视频表示的空间和时间变化。

本文主要贡献如下：

1) 本文提出了一个多任务对比学习方法，通过设计基于代理任务的对比学习损失，学习空间代理任务的特征与时间代理任务的特征的差异性。

2) 多任务对比学习方法引入了注意力特征提取模块，帮助网络获取关键空间信息和关键视频帧信息来增强空间和时间特征。

3) 多任务对比学习方法在自监督上进行预训练，将训练好的模型进行微调应用于监督行为识别任务。实验结果表明，本文的多任务对比学习方法在 UCF-101 数据集和 HDMB-51 数据集上实现了最先进的监督行为识别。

2. 多任务对比学习方法

本文如图 2 所示，本文提出的多任务对比学习方法由 4 个模块组成：任务数据扩充模块、注意力特征提取模块、代理任务损失模块和对比学习损失模块，后续章节会详细介绍上述模块。

2.1. 多任务数据扩充模块

如图 2(a)所示，本文使用 4 个代理任务训练网络，分别通过 2 个时间代理任务和 2 个空间代理任务来进行数据扩充。

时间速度代理任务(Temporal Pace Pretext Task: TP)。通过以 4 个速率(1×, 2×, 3×, 4×)进行采样，得到不同播放速度的片段作为输入，使得本文的网络模型预测正确的速率。

时间顺序代理任务(Temporal Order Pretext Task: TO)。通过打乱三个片段(每个片段包含 3 帧)，得到无序片段作为输入，使得本文的网络模型预测正确的视频片段顺序。

空间旋转代理任务(Spatial Rotation Pretext Task: SR)。通过旋转视频剪辑的每一帧作为输入，旋转有四个角度(0˚，90˚，180˚，270˚)，使得本文的网络模型预测正确的视频旋转角度。

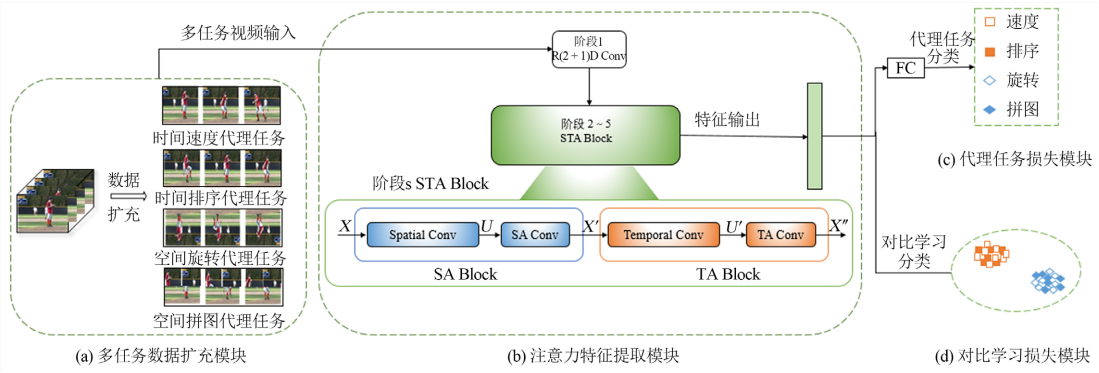

Figure 2. The multi-task comparison learning method consists of the following four modules. (a) Multi-task data expansion module. (b) Attentional feature extraction module. (c) Agent task loss module. (d) Contrast learning loss module.
图 2. 多任务对比学习方法包含以下 4 个模块。(a) 多任务数据扩充模块。(b) 注意力特征提取模块。(c) 代理任务损失模块。(d) 对比学习损失模块

空间拼图代理任务(Spatial Jigsaw Pretext Task: SJ)。将视频剪辑的每一帧划分为 2 × 2 网格块。然后，该任务通过对视频片段的网格快进行随机打乱，使得本文的网络模型预测正确的空间顺序。

我们将多任务数据扩充视频表示为 $V_n^{task_i}$，n 为视频下标，其中 $task_i=\{\mathrm{TP},\mathrm{TO},\mathrm{SR},\mathrm{SJ}\}$，$i=1,\cdots,M$，$M$ 为代理任务数，因此 $M=4$。

2.2. 注意力特征提取模块

多任务对比学习方法的注意力特征模块包含 5 个阶段，阶段 1 是 R(2+1)D Conv 模块，阶段 s 是时空注意力模块(STA Block)，$s=2$，3，4，5。

如图 2(b)所示的注意力特征提取模块的 STA Block 所示，STA Block 包含空间通道注意力模块(SA Block)和时间注意力模块(TA Block)。本文将阶段 s 的 STA Block 的输入特征简化定义为 $X\in\mathrm{R}^{T\times C\times H\times W}$，其中 T 为视频时间帧数，C 为通道数，$H\times W$ 为图像的高和宽。空间通道注意力模块的输出特征为 $X'\in\mathrm{R}^{T\times C'\times H'\times W'}$，其中 X' 学习到的是视频帧不同通道之间的空间信息。时间注意力的输出特征为 $X''\in\mathrm{R}^{T\times C''\times H''\times W''}$，其中 X'' 学习到的是视频帧序列时间维度的信息。

2.2.1. 空间通道注意力模块

空间特征描述了是视频中的目标与其他目标空间位置的关系。两个位置之间的关系描述了一个目标的上下文，然而不同通道在空间同一位置与其他空间位置的关系是不同的，因此可以通过将整个帧不同通道的同一空间位置特征进行聚合，来描述当前空间位置的特征。本文将当前空间位置与其他空间位置上的特征相比，上下文提供了补充特征。因此，本文运用空间通道关系来增强空间特征。

本文对 X 进行时间分割来得到 t 时刻的特征 x_t。本文使用 R(2 + 1)D 的 Spatial Conv 对 x_t 使用二维空间卷积得到特征 $u_t\in\mathrm{R}^{C\times H\times W}$，其中 u_t 是 t 时刻的特征，u_t 的集合表示为 $U=\{u_t\},t=1,\cdots,T$。为了得到每一层通道的权重关系，本文需要获取 u_t 第 c 通道的特征 $u_{t,c}$，$u_{t,c}$ 的集合表示为 $u_t=\{u_{t,c}\},c=1,\cdots,C'$。

如图 3 所示，本文使用空间通道注意力卷积解决 u_t 无法利用到每一层通道即 u_t 的上下文信息的问题。本文通过池化操作将全局空间信息压缩到一个通道描述符中。这是通过使用全局平均池化来生成通道级统计信息来实现的。形式上一个统计量 z_t 是对空间维数 $H'\times W'$ 进行压缩生成的。这样 z_t 的第 c 通道 $z_{t,c}$ 可以通过以下公式得到：

$$z_{t,c}=\frac{1}{H'\times W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'}u_{t,c} \tag{1}$$

式(1)中，$z_{t,c}$ 集合表示为 $z_t=\{z_{t,c}\},c=1,\cdots,C'$。为了完全捕获通道依赖性，函数必须满足两个标准：首先，它必须是灵活的(特别是，它必须能够学习通道之间的非线性相互作用)。其次，它必须学习一种非互斥关系。为了满足这些条件，本文选择采用一个简单的带有 Sigmoid 激活的门控机制，公式如下：

$$s_t=\text{Sigmoid}\left(\text{FC}\left(\text{ReLU}\left(\text{FC}\left(z_t\right)\right)\right)\right) \tag{2}$$

式(2)中，第一个 FC 层通过一个降维参数 r 在空间通道维度上降维，第二个 FC 层用来返回原来的空间通道维度，2 个 FC 层整体形成了一个空间通道维度的非线性层。空间通道注意力模块第 t 时刻的权重输出 x_t' 表示为：

$$x_t'=s_t\odot u_t \tag{3}$$

式(3)中，$\odot$ 为乘法，空间通道注意力模块的最终输出为 $X'=\{x_t'\},t=1,\cdots,T$。

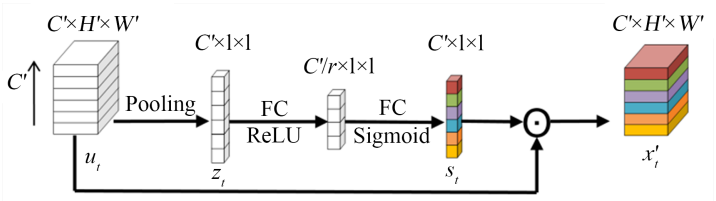

Figure 3. SA Conv
图 3. 空间通道注意力卷积

2.2.2. 时间注意力模块

不同于空间通道注意力模块，时间注意力模块关注的是视频帧之间的时间相关性。局部时间特征通常采用二维卷积或三维卷积来捕捉，但忽略了描述视频帧序列中的长期时间关系。两个时间框架之间的长期时间关系可以为具体的动作找到相关的框架。时间关系被认为是时间权重，它可以用来聚合相关帧的时间特征，从而学习视频的上下文特征。与局部时间特征相比，上下文提供了补充特征。因此，我们应用时间关系来增强时间特征。

本文使用 R(2 + 1)D 的 Temporal Conv 对 X' 使用一维时间卷积分解得到特征 $U'\in \mathrm{R}^{T\times C''\times H''\times W''}$，其中 $U'=\{u_t'\},t=1,\cdots,T$。

如图 4 所示，本文使用时间注意力卷积来解决 U' 无法获取长期时间关系的问题。时间注意力卷积通过池化操作压缩通道特征和空间特征 $C''\times H''\times W''$ 来获取每帧的时间描述符特征表示为 $z_t'\in \mathrm{R}^{1\times 1}$。$z_t'$ 计算公式如下：

$$z_t'=\frac{1}{C''\times H''\times W''}\sum_{i=1}^{C''}\sum_{j=1}^{H''}\sum_{k=1}^{W''}u_t' \tag{4}$$

式(4)中，z_t' 的集合表示为 $Z'=\{z_t'\},t=1,\cdots,T$ 为了完全捕捉时间依赖性，与空间通道注意力模块的操作类似，时间注意力模块的权重输出 S' 表示为：

$$S'=\text{Sigmoid}\left(\text{FC}\left(\text{ReLU}\left(\text{FC}\left(Z'\right)\right)\right)\right) \tag{5}$$

式(5)中，第一个 FC 层通过一个降维参数 r' 在时间维度上降维，第二个 FC 层用来返回原来的时间维度，2 个 FC 层整体形成了一个时间维度的非线性层。时间注意力模块的输出 X'' 表示为：

$$X''=S'\odot U' \tag{6}$$

式(6)中，$\odot$ 为乘法，X'' 为 STA Block 阶段 s 的注意力特征提取模块的最终输出。

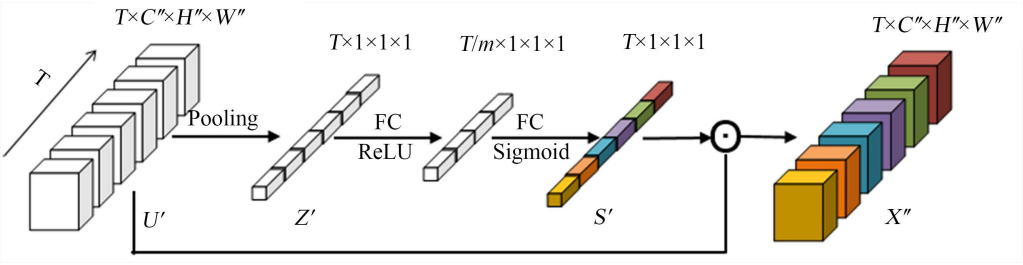

Figure 4. TA conv
图 4. 时间注意力卷积

2.3. 总体损失函数

多任务对比学习方法的总体损失函数包括代理任务损失和基于代理任务的对比学习损失，因此本文的总体损失函数表示如下：

$$L_{overall} = L_{con} + \lambda L_{task} \tag{7}$$

式(7)中，L_{con} 为对比学习损失，L_{task} 为代理任务损失，其中正负对的对比损失总和通常比代理任务损失大很多，因此 λ 被用于平衡损失项。本文会在 2.3.1 和 2.3.2 章节介绍代理任务损失和对比学习损失。

2.3.1. 代理任务损失模块

如图 2(c)所示，本文多任务对应的分类标签分别是速度、排序、旋转、拼图，为了学习到与分类标签对应的代理任务的视频特征信息，本文使用如下方法学习对应的视频特征信息。

首先，本文将注意力特征提取模块阶段 5 的输出特征表示为 $h_n^{task_i}$。其次，本文通过一个 FC 层得到代理任务概率分布 $p_n^{task_i} = \text{Sigmoid}\left(\text{FC}\left(h_n^{task_i}\right)\right)$。则代理任务损失 L_{task} 计算公式如下：

$$L_{task} = \frac{1}{N}\sum_{n=1}^{N}\left(-\sum_{i=1}^{M} y_i^{gt} \log p_n^{task_i}\right) \tag{8}$$

式(8)中，N 表示视频数，M 表示为代理任务数，y_i^{gt} 表示对应的真实分类标签。

2.3.2. 对比学习损失模块

现有的自监督方法通常采用基于实例的对比学习来分离两个视频的特征。本文认为空间特征和时间特征是视频的两个不同方面。基于实例的对比学习没有充分分析时间特征和空间特征之间的差异，可能会阻碍发现空间变化和时间变化。为此，本文利用自监督的空间标签和时间标签，在本文的对比学习方法中增加不同的代理任务，允许分别学习空间特征和时间特征。其中空间特征通过空间旋转代理任务(SR)和空间拼图代理任务(SJ)两个空间代理任务来学习。时间特征通过时间顺序代理任务(TO)和时间速度代理任务(TP)来学习。

注意力特征提取模块阶段 5 的输出特征表示为 $h_n^{task_i}$，$task_i = \{\text{TP},\text{TO},\text{SR},\text{SJ}\}$。本文的代理任务损失考虑来自同一视频两个时间代理任务的特征对 $\left(h_n^{\text{TP}}, h_n^{\text{TO}}\right)$ 表示为正样本对，来自同一视频两个空间代理任务的特征对 $\left(h_n^{\text{SR}}, h_n^{\text{SJ}}\right)$ 表示为正样本对，来自同一视频一个时间代理任务和一个空间代理任务的特征对表示为负样本对，有 $\left(h_n^{\text{TP}}, h_n^{\text{SR}}\right)$，$\left(h_n^{\text{TP}}, h_n^{\text{SJ}}\right)$，$\left(h_n^{\text{TO}}, h_n^{\text{SR}}\right)$，$\left(h_n^{\text{TO}}, h_n^{\text{SJ}}\right)$。特征对的点积相似度公式表示为 $sim(u,v) = uv^{\text{T}} / \|u\|\|v\|$。本文设置对比学习损失的分子是正样本对和负样本对的集合，分母为正样本对，因此本文将特征 $h_n^{task_i}$ 基于代理任务的对比损失 L_{con} 表示为：

$$L_{con}=\frac{1}{N}\sum_{n=1}^{N}-\log\frac{\exp\left(sim\left(h_n^{\mathrm{TP}},h_n^{\mathrm{TO}}\right)/\tau\right)+\exp\left(sim\left(h_n^{\mathrm{SR}},h_n^{\mathrm{SJ}}\right)/\tau\right)}{\sum_{i\neq j}\exp\left(sim\left(h_n^{task_i},h_n^{task_j}\right)/\tau\right)} \tag{9}$$

式(9)中，τ 是一个对比学习损失的平衡参数，N 表示视频数。

3. 实验

3.1. 数据集与评价标准

本文使用两阶段实验[12]，分别是自监督预训练和监督行为识别微调。本文采用了数据集 Kinetics-40 [13]进行自监督预训练。对于监督行为识别微调，本文将在 Kinetics-400 数据集进行自监督预训练的模型微调应用于广泛使用的两个数据集 UCF-101 [14]以及 HDMB-51 [15]上进行了监督行为识别。

Kinetics-400 该数据集包含 400 个不同的动作类别，共计有超过 400,000 个视频，其中每个视频的长度为 10 秒。这些视频的场景涵盖了人们日常生活中的许多活动。本文在该数据集上进行预训练，其中80%作为训练集，20%作为测试集。

UCF-101 该数据集包含了 101 类不同的行为视频，其中每类包含 13,320 个视频，每个视频的时长为 5 秒。UCF-101 数据集涵盖了各种人类动作，如游泳，跳舞，拍拍手，跳绳，等等。它收集了超过 13,000 个视频，其中 70%作为训练集，30%作为测试集。

HDMB-51 该数据集是一个用于室内行人识别的数据集，由三个实验室收集的 51 个室内视频序列组成，视频序列中的每一帧都有一个标签，以指示行人的位置。该数据集包括 2790 个视频帧，其中 70%作为训练集，30%作为测试集。

3.2. 实验细节

自监督预训练。在多任务数据扩充模块中，提取视频图像帧数 $T=9$ 帧，每帧随机裁剪为 224 × 224。每个视频提取 $M=4$ 个代理任务的特征，包括时间速度代理任务(TP)、时间顺序代理任务(TO)、空间旋转代理任务(SR)和空间拼图代理任务(SJ)。在特征提取模块中，空间通道注意力卷积设置 $r=16$，时间注意力卷积设置 $r'=3$。在对比学习模块中，对比学习损失参数 $\tau=0.07$。由于正负对的对比损失下降梯度大约是代理任务损失的 10 倍，因此平衡参数 $\lambda=10$。

训练期间，网络超参数设置如下：训练周期 epoch = 18，批次大小 batch = 32。每个批次由代理任务损失和对比学习损失组成的总体损失进行学习。模型的优化器动量为 0.9，初始学习率为 0.06，每 3 个周期以线性余弦的方式下降[16]。

监督微调。在自监督模型预训练完之后，对自监督预训练模型进行微调，即针对下游不同的分类任务对最后一层进行微调。为了进行公平的比较，本文在 UCF-101 和 HDMB-51 数据集上微调分类器进行行为识别。与 SlowFast [17]一样，本文对整个视频的 10 个剪辑进行采样，并对所有剪辑的 Softmax 概率取平均值作为最终预测。训练期间，网络超参数设置如下：训练周期 epoch = 200，批次大小 batch = 128，该模型的模型优化器动量为 0.9，初始学习率为 0.2，分别在 50，100 和 150 批次降低了 10 倍。

本文实验所采用的 PC 机配置为 Intel Core i7-5960X、CPU 3GHz×8 cores RAM 8GB、图像显卡为 1 张 NVIDIA GeForce GTX 1080Ti、Linux16.04 操作系统。深度学习框架为 pytorch。

3.3. 对比实验

在本文的多任务对比学习方法不丧失通用性的情况下，以下的实验都是基于 R(2 + 1)D 基础网络来说明我们的多任务对比学习方法的有效性。

直接应用多任务对性能的影响。在参文[2]中实验证实了时间代理任务的数据扩充不会改善甚至会降低性能。本文类似使用了多代理任务数据扩充进行实验。结果证明，在表 1 中直接应用多代理任务数据扩充与单一代理任务相比结果并没有改善。相反，多任务组合的性能都有所下降。

Table 1. Comparison of multitask and contrastive learning
表 1. 多任务与对比学习的比较

预训练 代理任务	L_{task}	L_{con}	UCF-101 (%)	HDMB-51 (%)
SJ	√	--	78.6	39.2
SR	√	--	79.0	**42.0**
TP	√	--	79.5	41.3
TO	√	--	**80.1**	40.7
SJ + TP	√	--	75.2	36.6
SJ + TO	√	--	74.2	37.2
SR + TP	√	--	75.4	38.2
SR + TO	√	--	76.3	37.9
SJ + TP	√	√	80.4	43.5
SJ + TO	√	√	80.7	41.1
SR + TP	√	√	82.2	44.6
SR + TO	√	√	81.7	42.6
TO + TP + SR + SJ	√	--	74.9	39.3
TO + TP + SR + SJ	√	√	**86.9**	**54.3**

将多任务转化为自监督信号对性能的影响。从表 1 可以看出，将多任务作为额外的对比自监督信号，比直接应用多任务的性能得到明显改善，并且优于最好的单一任务的性能。这表明，本文提出的多任务对比学习提供了额外的对比自监督信号，从而实现了时空信息的辨别，因此指导我们的模型学习到更好的视频表达。最后该模型使用 TO + TP + SR + SJ 四种代理任务和整体损失来获得最佳性能(UCF-101 为 86.9%，HDMB-51 为 54.3%)。

Table 2. Comparison of attentive STA Block's contrastive learning
表 2. 时空注意力模块的对比学习比较

Self-attention	UCF-101 (%)	HDMB-51 (%)
Without	86.9	54.3
SA	87.4	55.4
TA	87.0	56.3
SA + TA	**88.4**	**57.6**

时空注意力模块对性能的影响。表 2 给出了不同注意模块下的行为识别精度。具有空间注意的模型学习视频帧通道之间的空间关系，即目标之间的位置关系。具有时间注意的模型学习帧间的时间关系，可以增强具有时间背景的特征，即长时间帧中的物体运动。视频表示的空间关系和时间关系是互补的，最后实验得到了最佳的效果(UCF-101 为 88.4%，HDMB-51 为 57.6%)。

3.4. 消融实验

表 3 显示了最先进的自监督方法与微调行为识别精度的比较。在自监督方法中，将 3D CNN 主干用于学习时空特征。3D RotNet 考虑空间旋转的代理任务。3D ST-puzzle 应用了时空变换的代理任务。SpeedNet 使用速度预测的代理任务。TBE 运用了时序的代理任务和背景擦除的特点。LSFD 使用基于实例的长视视频和短视视频特征之间的对比损耗。在具有运动保留特征的视频之间使用基于实例的对比损失。本文的方法相比于最先进的自监督方法在 UCF-101 提升了 0.8%，HDMB-51 提升了 0.5%。

Table 3. Comparison of state-of-the-art self-supervised methods
表 3. 最先进的自监督方法的比较

方法	Backbone	预训练数据集	UCF-101 (%)	HDMB-51 (%)
3D RotNet [1]	3D ResNet-18	Kinetics-400	76.6	47.0
3D ST-puzzle [7]	3D ResNet-18	Kinetics-400	65.8	33.7
LSFD [18]	C3D	Kinetics-400	79.8	52.1
SpeedNet [5]	S3D-G	Kinetics-400	81.1	48.8
TBE [19]	I3D	Kinetics-400	87.1	56.2
CoCLR [20]	S3D	Kinetics-400	87.9	54.6
Video Jigsaw [2]	CaffeNet	Kinetics-400	55.4	27.0
ERUV [21]	R(2 + 1)D-18	Thumos14	68.4	31.9
PRP [4]	R(2 + 1)D-18	UCF-101	72.1	35.0
COP [3]	R(2 + 1)D-18	GTEA	72.4	30.9
PacePred [11]	R(2 + 1)D-18	Kinetics-400	77.1	36.6
TCGL [22]	R(2 + 1)D-18	Kinetics-400	81.2	50.1
TaCo [12]	R(2 + 1)D-18	Kinetics-400	81.8	46.0
CPNet [6]	R(2 + 1)D-18	Kinetics-400	83.8	**57.1**
Elo [15]	R(2 + 1)D-26	Youtube	84.2	53.7
TCLR [10]	R(2 + 1)D-18	UCF-101	84.3	54.2
CMD [9]	R(2 + 1)D-26	Kinetics-400	85.7	54.0
STOR [23]	R(2 + 1)D-18	Kinetics-400	**87.6**	56.4
Ours	R(2 + 1)D-18	Kinetics-400	**88.4**	**57.6**

3.5. 可视化结果分析

为了找到不同代理任务下的微调行为识别模型的特征响应，图 5 可视化了 Stage5 中输出特征的显著性热图。如图 5(a)对于动作“铅球”，TO 和 TP 倾向于捕捉人推铅球的运动轨迹。SR 和 SJ 倾向于捕获人体站立状态和身体位置。本文的模型(TP + TO + SR + SJ)利用多任务的对比损失同时捕捉了推铅球的运动轨迹和人体的位置，效果更加显著。

如图 5(b)对于动作“棒球”，TO 和 TP 倾向于捕捉人打棒球的运动轨迹。SR 和 SJ 倾向于捕获人体站立状态和身体位置。本文的模型(TP + TO + SR + SJ)利用多任务的对比损失同时捕捉了打棒球的运动轨迹和人体的位置，效果更加显著。如图 5(c)对于动作“篮球”，TO 和 TP 倾向于捕捉人打篮球的运动轨迹。SR 和 SJ 倾向于捕获人体站立状态、身体位置和篮球框的位置。本文的模型(TP + TO + SR + SJ)利用

多任务的对比损失同时捕捉了打篮球的运动轨迹、人体的位置和篮球框的位置，效果更加显著。因此本文的模型利用多任务的对比损失来捕捉上述时空特征更显著。

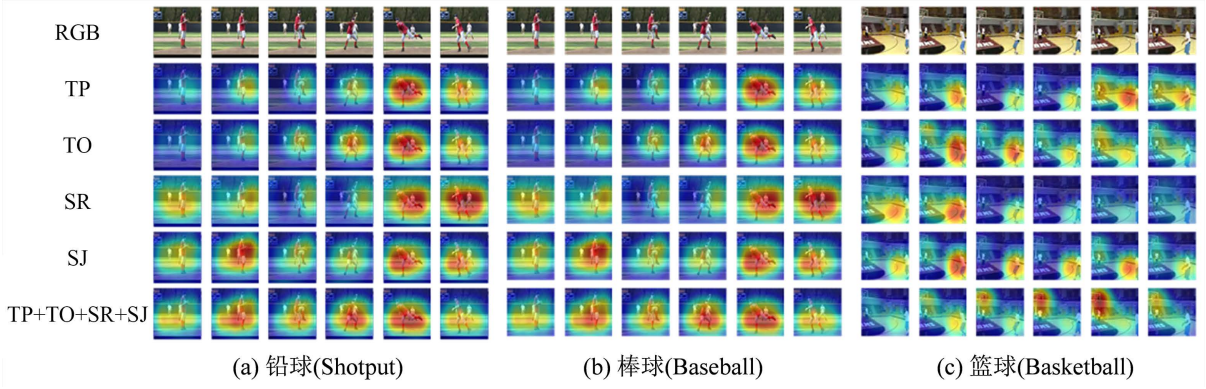

Figure 5. The output heat map of the 5th stage of attention feature extraction module
图 5. 注意力特征提取模块第 5 阶段的输出热图

4. 结论

本文提出了一个多任务对比学习方法，解决了单一代理任务监督信号单一，不足以详细描述视频学习空间特征和时间特征之间差异的问题。

多任务对比学习方法的多任务数据扩充模块，使本文的模型可以同时学到时间特征和空间特征。其中注意力特征提取模块进一步增强了多任务学习到的时间特征和空间特征。如果只应用多任务数据扩充模块，本文的模型无法区分时间特征和空间特征的差异，因此提出了对比学习损失模块。在自监督学习中，基于代理任务的对比损失使得空间代理任务学习到的特征与时间代理任务学习到的特征不同。本文将自监督模型微调应用于监督行为识别任务。在 UCF-101 数据集和 HMDB-51 数据集上的实验结果验证了该方法的有效性。

基金项目

安徽省重点研究与开发计划(202004d07020004)，安徽省自然科学基金项目(2108085MF203)，中央高校基本科研业务费专项资金(PA2021GDSK0072, JZ2021HGQA0219)。

参考文献

[1] Jing, L. and Tian, Y. (2018) Self-Supervised Spatiotemporal Feature Learning by Video Geometric Transformations.

[2] Ahsan, U., Madhok, R. and Essa, I. (2019) Video Jigsaw: Unsupervised Learning of Spatiotemporal Context for Video Action Recognition. 2019 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 7-11 January 2019, 179-189. https://doi.org/10.1109/WACV.2019.00025

[3] Xu, D., Xiao, J., Zhao, Z., *et al*. (2019) Self-Supervised Spatiotemporal Learning via Video Clip Order Prediction. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 10326-10335. https://doi.org/10.1109/CVPR.2019.01058

[4] Yao, Y., Liu, C., Luo, D., *et al*. (2020) Video Playback Rate Perception for Self-Supervised Spatio-Temporal Representation Learning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6547-6556. https://doi.org/10.1109/CVPR42600.2020.00658

[5] Benaim, S., Ephrat, A., Lang, O., *et al*. (2020) SpeedNet: Learning the Speediness in Videos. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 9919-9928. https://doi.org/10.1109/CVPR42600.2020.00994

[6] Liang, H., Quader, N., Chi, Z., *et al*. (2021) Self-Supervised Spatiotemporal Representation Learning by Exploiting Video

Continuity. *The* 36*th AAAI Conference on Artificial Intelligence* (*AAAI*-22), 22 February-1 March 2022, 1564-1573.

[7] Kim, D., Cho, D. and Kweon, I.S. (2018) Self-Supervised Video Representation Learning with Space-Time Cubic Puzzles. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8545-8552. https://doi.org/10.1609/aaai.v33i01.33018545

[8] Piergiovanni, A.J., Angelova, A. and Ryoo, M.S. (2020) Evolving Losses for Unsupervised Video Representation Learning. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 130-139. https://doi.org/10.1109/CVPR42600.2020.00021

[9] Huang, L., Liu, Y., Wang, B., *et al*. (2021) Self-Supervised Video Representation Learning by Context and Motion Decoupling. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 13881-13890. https://doi.org/10.1109/CVPR46437.2021.01367

[10] Dave, I., Gupta, R., Rizve, M.N., *et al*. (2021) TCLR: Temporal Contrastive Learning for Video Representation. *Computer Vision and Image Understanding*, **219**, Article ID: 103406. https://doi.org/10.1016/j.cviu.2022.103406

[11] Wang, J., Jiao, J. and Liu, Y.H. (2020) Self-Supervised Video Representation Learning by Pace Prediction. *Computer Vision—ECCV* 2020 16*th European Conference*, Glasgow, 23-28 August 2020, 504-521.

[12] Bai, Y., Fan, H., Misra, I., *et al*. (2020) Can Temporal Information Help with Contrastive Self-Supervised Learning?

[13] Kay, W., Carreira, J., Simonyan, K., *et al*. (2017) The Kinetics Human Action Video Dataset.

[14] Soomro, K., Zamir, A.R. and Shah, M. (2012) UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild.

[15] Kuehne, H., Jhuang, H., Garrote, E., *et al*. (2011) HMDB: A Large Video Database for Human Motion Recognition. *IEEE International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 2556-2563. https://doi.org/10.1109/ICCV.2011.6126543

[16] Chen, X., Xie, S. and He, K. (2021) An Empirical Study of Training Self-Supervised Vision Transformers. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9620-9629. https://doi.org/10.1109/ICCV48922.2021.00950

[17] Feichtenhofer, C., Fan, H., Malik, J., *et al*. (2019) SlowFast Networks for Video Recognition. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 6201-6210. https://doi.org/10.1109/ICCV.2019.00630

[18] Behrmann, N., Fayyaz, M., Gall, J., *et al*. (2021) Long Short View Feature Decomposition via Contrastive Video Representation Learning. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9224-9233. https://doi.org/10.1109/ICCV48922.2021.00911

[19] Wang, J., Gao, Y., Li, K., *et al*. (2021) Removing the Background by Adding the Background: Towards Background Robust Self-Supervised Video Representation Learning. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 11799-11808. https://doi.org/10.1109/CVPR46437.2021.01163

[20] Han, T., Xie, W. and Zisserman, A. (2020) Self-Supervised Co-Training for Video Representation Learning.

[21] Luo, D., Fang, B., Zhou, Y., *et al*. (2020) Exploring Relations in Untrimmed Videos for Self-Supervised Learning. *ACM Transactions on Multimedia Computing*, *Communications*, *and Applications*, **18**, Article No. 35.

[22] Liu, Y., Wang, K., Lan, H., *et al*. (2021) Temporal Contrastive Graph Learning for Video Action Recognition and Retrieval.

[23] Zhang, Y., Po, L.M., Xu, X., *et al*. (2021) Contrastive Spatio-Temporal Pretext Learning for Self-Supervised Video Representation. *The* 36*th AAAI Conference on Artificial Intelligence* (*AAAI*-22), 22 February-1 March 2022, 3380-3389.