

视频表征自监督对比学习方法和装置

相关申请的交叉引用

本申请是以 CN 申请号为 202111085396.0，申请日为 2021 年 09 月 16 的申请为基
5 础，并主张其优先权，该 CN 申请的公开内容在此作为整体引入本申请中。

技术领域

本公开涉及视频学习领域，特别涉及一种视频表征自监督对比学习方法和装置。

10 背景技术

视频表征自监督学习的目标是通过探索未经标注的视频中存在的内在属性来学习
视频的特征表达。

一种视频表征自监督对比学习方法，其基于对比学习技术实现高效的自监督视频
表征学习。然而，目前的视频表征自监督对比学习技术通常关注如何根据图像对比学
15 习的研究成果提升对比学习性能。

发明内容

本公开一些实施例提出一种视频表征自监督对比学习方法，包括：

根据视频片段的每个视频帧相应的光流信息，计算得到所述视频片段的每个视频
20 帧相应的运动幅度图；

根据所述视频片段的各视频帧相应的运动幅度图，确定所述视频片段相应的运动
信息；

根据视频片段序列和每个视频片段相应的运动信息，进行视频表征自监督对比学
习。

25 在一些实施例中，所述根据视频片段的每个视频帧相应的光流信息，计算得到所
述视频片段的每个视频帧相应的运动幅度图，包括：

提取视频片段中每一对相邻视频帧之间的光流场，以确定所述视频片段的每个视
频帧相应的光流场；

计算每个视频帧相应的光流场在第一方向和第二方向的梯度场；

将第一方向和第二方向的梯度场的幅值进行聚合得到每个视频帧相应的运动幅度图。

在一些实施例中，第一方向和第二方向相互垂直。

5 在一些实施例中，所述计算每个视频帧相应的光流场在第一方向和第二方向的梯度场包括：

计算每个视频帧相应的光流场的水平分量在第一方向和第二方向的梯度；

计算每个视频帧相应的光流场的垂直分量在第一方向和第二方向的梯度；

每个视频帧相应的光流场的水平分量和垂直分量在第一方向和第二方向的梯度，构成所述光流场在第一方向和第二方向的梯度场。

10 在一些实施例中，所述视频片段相应的运动信息包括：所述视频片段相应的时空运动图、空间运动图、时间运动图中的一项或多项；其中，

确定所述视频片段相应的时空运动图包括：在时间维度将视频片段的各视频帧的运动幅度图叠加构成所述视频片段的时空运动图；

15 确定所述视频片段相应的空间运动图包括：对所述视频片段的时空运动图沿着时间维度进行池化，得到所述视频片段的空间运动图；

确定所述视频片段相应的时间运动图包括：对所述视频片段的时空运动图沿着空间维度行池化，得到所述视频片段的时间运动图。

在一些实施例中，所述根据视频片段序列和每个视频片段相应的运动信息，进行视频表征自监督对比学习，包括：

20 根据每个视频片段相应的运动信息对视频片段进行数据增强，根据增强后的视频片段序列并结合对比损失进行视频表征自监督对比学习；或者，

根据视频片段序列并结合运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习；或者，

25 根据每个视频片段相应的运动信息对视频片段进行数据增强，根据增强后的视频片段序列并结合运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习；

其中，所述运动对齐损失通过对齐进行学习的主干网络的最后卷积层的输出与视频片段相应的运动信息来确定。

在一些实施例中，所述根据每个视频片段相应的运动信息对视频片段进行数据增强包括：

在视频片段相应的运动信息包括视频片段相应的时空运动图的情况下，根据时空运动图中各像素的运动速度的大小确定第一阈值，根据第一阈值确定视频片段中具备较大运动幅度的三维时空区域；或者，

5 在视频片段相应的运动信息包括视频片段相应的时间运动图的情况下，根据视频片段相应的时间运动图计算所述视频片段的运动幅度，对序列中的各视频片段进行时域采样，采样得到的视频片段的运动幅度不小于第二阈值，第二阈值根据各视频片段的运动幅度确定；或者，

10 在视频片段相应的运动信息包括视频片段相应的空间运动图的情况下，根据视频片段相应的空间运动图中各像素的运动速度的大小确定第三阈值，根据第三阈值对各像素进行划分，对空间运动图反复执行随机多尺度空间裁剪、并确保裁剪得到的矩形空间区域至少覆盖了超过预设比例的空间运动图中大于第三阈值的像素，对视频片段中的每一视频帧都裁剪与矩形空间区域相同的区域。

在一些实施例中，所述根据视频片段相应的时间运动图计算所述视频片段的运动幅度包括：

15 将视频片段相应的时间运动图作为视频帧级别的运动图，计算视频片段内所有帧的视频帧级别的运动图的均值，作为所述视频片段的运动幅度。

在一些实施例中，第一阈值、第二阈值、第三阈值分别采用中位数的方法确定。

在一些实施例中，所述对视频片段进行数据增强还包括：对视频片段中的视频帧进行图像数据增强操作。

20 在一些实施例中，所述运动对齐损失相应的损失函数表示为以下的一项或多项的累加：

主干网络的最后卷积层输出的特征图在所有通道的累加与视频片段相应的时空运动图之间的距离，

25 所述累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离，
所述累加沿着空间维度的池化结果与视频片段相应的时间运动图之间的距离。

在一些实施例中，所述运动对齐损失相应的损失函数表示为以下的一项或多项的累加：

主干网络的最后卷积层输出的特征图按照各通道的权重在所有通道的第一加权累加与视频片段相应的时空运动图之间的距离，

所述第一加权累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离，

所述第一加权累加沿着空间维度的池化结果与视频片段相应的时间运动图之间的距离。

5 在一些实施例中，所述运动对齐损失相应的损失函数表示为以下的一项或多项的累加：

主干网络的最后卷积层输出的特征图的各通道的梯度按照各通道的权重在所有通道的第二加权累加与视频片段相应的时空运动图之间的距离，

10 所述第二加权累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离，

所述第二加权累加沿着空间维度的池化结果与视频片段相应的时间运动图之间的距离。

15 在一些实施例中，通道的权重的确定方法包括：计算视频片段相应的查询样本和正例样本之间的相似度相对于卷积层输出的特征图的某个通道的梯度，计算该通道的梯度的均值，作为该通道的权重。

在一些实施例中，所述对比损失根据对比学习的损失函数确定。

在一些实施例中，对比学习的损失函数包括 InfoNCE 损失函数。

在一些实施例中，所述主干网络包括三维卷积神经网络。

20 在一些实施例中，还包括：根据学习得到的视频表征模型，对待处理视频进行处理得到相应的视频特征。

本公开一些实施例提出一种视频表征自监督对比学习装置，包括：存储器；以及耦接至所述存储器的处理器，所述处理器被配置为基于存储在所述存储器中的指令，执行所述视频表征自监督对比学习方法。

25 本公开一些实施例提出一种非瞬时性计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现所述视频表征自监督对比学习方法的步骤。

附图说明

下面将对实施例或相关技术描述中所需要使用的附图作简单地介绍。根据下面参照附图的详细描述，可以更加清楚地理解本公开。

显而易见地，下面描述中的附图仅仅是本公开的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图 1 示出本公开一些实施例的运动聚焦的视频表征自监督对比学习方法的流程示意图。

图 2a、2b、2c、2d 示出本公开一些实施例的提取视频片段的运动信息以及基于运动信息
5 的视频数据增强的示意图。

图 3 示出了本公开通过运动聚焦的视频数据增强和运动聚焦的特征学习两方面的结合同时进行视频表征自监督对比学习的示意图。

图 4 示出了本公开一些实施例的运动对齐损失函数的对齐示意图。

图 5 为本公开一些实施例的运动聚焦的视频表征自监督对比学习装置的结构示意图。

10

具体实施方式

下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述。

除非特别说明，否则，本公开中的“第一”“第二”等描述用来区分不同的对象，并不
15 用来表示大小或时序等含义。

经研究发现，目前的视频表征自监督对比学习技术通常关注如何根据图像对比学习的研究成果提升对比学习性能，往往会忽视视频和图像之间存在的最关键的时间维度的差异，进而导致视频中广泛存在的运动信息没有得到充分的重视与根据，但是在实际场景中，视频的语义信息和运动信息是高度相关的。

20 本公开提出一种用于视频表征自监督学习的运动聚焦对比学习方案，使得视频中广泛存在且十分重要的运动信息在学习过程中被充分利用，从而提升视频表征自监督对比学习性能。

图 1 示出本公开一些实施例的运动聚焦的视频表征自监督对比学习方法的流程示意图。

25 如图 1 所示，该实施例的方法包括：步骤 110-130。

在步骤 110，根据视频片段的每个视频帧相应的光流信息，计算得到所述视频片段的每个视频帧相应的运动幅度图。

在视频中，不同区域的运动本质上是不同的。采用运动速度大小（即运动幅度）来衡量视频帧中各区域相对参考帧的位置变化率。一般来说，速度较大的区域具有更

丰富的信息，并且更有利于对比学习。

在一些实施例中，本步骤 110 例如包括：步骤 111-113。

在步骤 111，提取视频片段中每一对相邻视频帧之间的光流场，以确定所述视频片段的每个视频帧相应的光流场。

5 针对拥有 N 帧的分辨率为 $H \times W$ 的视频片段（图 2a，其中的视频图像仅是一种示例，本申请并不保护视频图像内容），根据无监督的 TV-L1 算法提取每一对相邻视频帧之间的光流场，以确定所述视频片段的每个视频帧相应的光流场，记作 $\{(u_1, v_1), (u_2, v_2), \dots, (u_N, v_N)\}$ 。其中 u_i, v_i 分别是光流场在水平方向和竖直方向的分量，用于表示每个像素在 i 帧和 $i + 1$ 帧之间的位移，其分辨率均为 $H \times W$ 。

10 光流场，是指图像中所有像素点构成的一种二维瞬时速度场，其中的二维速度矢量是景物中可见点的三维速度矢量在成像表面的投影。

在步骤 112，计算每个视频帧相应的光流场在第一方向和第二方向的梯度场。

在根据光流计算运动幅度的过程中，受到相机运动的影响，直接根据光流进行计算运动幅度很可能遭遇稳定性问题。例如，当相机快速运动时，原本静止的物体或者背景像素会在光流中呈现出很高的运动速度，这对于获得高质量的视频内容的运动信息不利。为了消除相机镜头抖动带来的不稳定问题，进一步计算光流场在第一方向和第二方向的梯度场作为运动边界。

15

在一些实施例中，计算每个视频帧相应的光流场在第一方向和第二方向的梯度场包括：计算每个视频帧相应的光流场的水平分量在第一方向和第二方向的梯度；计算每个视频帧相应的光流场的垂直分量在第一方向和第二方向的梯度；每个视频帧相应的光流场的水平分量和垂直分量在第一方向和第二方向的梯度，构成所述光流场在第一方向和第二方向的梯度场。在一些实施例中，第一方向和第二方向可相互垂直。例如，将坐标系中的相互垂直的 x 方向和 y 方向作为第一方向和第二方向。

20

计算每个视频帧相应的光流场在 x 方向和 y 方向的梯度信息作为运动边界。例如，对于第 i 帧的光流场 (u_i, v_i) ，可以计算其在 x 方向和 y 方向的梯度场： $(\frac{\partial u_i}{\partial x}, \frac{\partial u_i}{\partial y}, \frac{\partial v_i}{\partial x}, \frac{\partial v_i}{\partial y})$ 。

25

在步骤 113，将第一方向和第二方向的梯度场的幅值进行聚合得到每个视频帧相应的运动幅度图。

基于上述梯度场可以进一步将各个方向的梯度场的幅值进行聚合得到第 i 帧的运动幅度图 m_i （图 2b）：

$$m_i = \sqrt{\left(\frac{\partial u_i}{\partial x}\right)^2 + \left(\frac{\partial u_i}{\partial y}\right)^2 + \left(\frac{\partial v_i}{\partial x}\right)^2 + \left(\frac{\partial v_i}{\partial y}\right)^2}$$

其中 $m_i \in \mathbb{R}^{H \times W}$ 用于表征第 i 帧中每个像素的运动速度大小（即运动幅度），略去了运动的方向信息。正如图 2b 所示，本公开定义的运动幅度图没有受到相机运动的影响，对时频片段中事实上的运动物体显示出高响应，高亮部分对应运动物体。

5 在步骤 120，根据所述视频片段的各视频帧相应的运动幅度图，确定所述视频片段相应的运动信息。

所述视频片段相应的运动信息包括：所述视频片段相应的时空运动图（ $m^{ST} \in \mathbb{R}^{N \times H \times W}$ ，ST-motion）、空间运动图（ $m^S \in \mathbb{R}^{H \times W}$ ，S-motion）、时间运动图（ $m^T \in \mathbb{R}^N$ ，T-motion）中的一项或多项。

10 确定所述视频片段相应的时空运动图包括：在时间维度将视频片段的各视频帧的运动幅度图叠加构成所述视频片段的时空运动图。例如，对于长度为 N 帧的视频片段，在时间维度将视频片段的各视频帧的运动幅度图 m_i 叠加构成时空运动图 m^{ST} 。

确定所述视频片段相应的空间运动图包括：对所述视频片段的时空运动图沿着时间维度进行池化，得到所述视频片段的空间运动图。例如，对 $m^{ST} \in \mathbb{R}^{N \times H \times W}$ 沿着时间
15 维度进行池化，得到所述视频片段的空间运动图 $m^S \in \mathbb{R}^{H \times W}$ 。

确定所述视频片段相应的时间运动图包括：对所述视频片段的时空运动图沿着空间维度行池化，得到所述视频片段的时间运动图。例如，对 $m^{ST} \in \mathbb{R}^{N \times H \times W}$ 沿着空间维度行池化，得到所述视频片段的时间运动图 $m^T \in \mathbb{R}^N$ 。

20 在步骤 130，根据视频片段序列和每个视频片段相应的运动信息，通过运动聚焦的视频数据增强、运动聚焦的特征学习这两方面的任一方面或者两方面的结合，进行视频表征自监督对比学习。提升在视频表征自监督对比学习任务中的表现。

图 3 示出了本公开通过运动聚焦的视频数据增强和运动聚焦的特征学习两方面的结合同时进行视频表征自监督对比学习的示意图。

其中，运动聚焦的视频数据增强（Motion-Focused Video Augmentation）能够根据预先计算得到的视频运动图（motion map）来产生具有丰富运动信息的三维管道作为骨干网络的输入。三维管道是指由采样自一系列连续视频帧的图像块在时间维度拼接在一起构成的视频样本。运动聚焦的视频数据增强可以分为两个部分：1）用于滤除画面相对静止的视频时间片段的时域采样（Temporal Sampling），和 2）用于选取视频中具有较大运动速度的空间区域的空间域裁剪（Spatial Cropping）。由于视频语
25

义与视频中运动信息的相关性，通过运动聚焦的视频数据增强，生成包含丰富运动信息的语义更相关的视频样本。

其中，运动聚焦的特征学习（Motion-Focused Feature Learning），通过本公开提出的新的运动对齐损失（Motion Alignment Loss）实现，通过在随机梯度下降的优化过程中对齐输入视频样本（三维管道）中每个位置对应的梯度幅度和运动图来促使骨干网络在特征学习过程中更多地关注视频中具备更高动态信息的区域。在对比学习损失（如 InfoNCE 损失）的基础上，运动对齐损失以额外约束条件的形式被集成入对比学习框架中。最终，整个运动聚焦对比学习框架以端到端的方式联合优化。所述主干网络包括三维卷积神经网络，例如三维 resnet 等，但不限于所举示例。主干网络后面还可以级联多层感知机（Multilayer Perceptron, MLP）等。通过运动聚焦的特征学习，使得学习过程中更多地聚焦于视频中的运动区域，进而使得学习到的视频特征包含充足的运动信息，更好地描述视频中的内容。

也即，本步骤 130 包括如下三种实现方式。

第一种，进行运动聚焦的视频数据增强：根据每个视频片段相应的运动信息对视频片段进行数据增强，根据增强后的视频片段序列并结合对比损失进行视频表征自监督对比学习，也即，针对增强后的视频片段序列，利用对比损失进行视频表征自监督对比学习。

第二种，进行运动聚焦的特征学习：根据视频片段序列并结合运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习，也即，针对视频片段序列，利用运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习。

第三种，同时进行运动聚焦的视频数据增强和运动聚焦的特征学习：根据每个视频片段相应的运动信息对视频片段进行数据增强，根据增强后的视频片段序列并结合运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习，也即，针对增强后的视频片段序列，利用运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习。

其中，所述运动对齐损失通过对齐进行学习的主干网络的最后卷积层的输出与视频片段相应的运动信息来确定。其中，所述对比损失根据对比学习的损失函数确定。对比学习的损失函数例如包括 InfoNCE 损失函数等，但不限于所举示例。后面会具体描述运动对齐损失和对比损失。

下面描述运动聚焦的视频数据增强。

基于前述的视频片段的各种运动图，运动聚焦的视频数据增强可以更好地关注视频中运动较大的区域。通过为对比学习算法选取更好的数据视图，提升模型学习到的视频表征的泛化能力。这是因为基于对比学习的自监督学习方法往往能够较好地收益于数据视图之间的互信息（MI，Mutual Information），而为了提高模型针对下游任务的泛化能力，“好”视图应该包含尽可能多的与下游任务相关信息，同时尽可能多地丢弃输入中的不相关信息。鉴于绝大多数视频相关的下游任务中都需要视频中的运动信息，例如，图 2c 中矩形框标示出了包含了较大运动幅度的两个视频区域样本，运动中的马和骑手包含了更有价值的互信息，图 2d 中矩形框标示出了从视频中静态区域采样得到的两个样本，包含灌木丛与地面等相对来说不太重要的背景信息，图 2c 中的样本更有助于提升模型对比学习的效果。本公开根据无需人工标注即可获得的运动图来寻找包含更多运动信息的视频时空区域。

在一些实施例中，所述根据每个视频片段相应的运动信息对视频片段进行数据增强包括以下至少三种实现方式。

15 第一种，在视频片段相应的运动信息包括视频片段相应的时空运动图的情况下，根据时空运动图中各像素的运动速度的大小确定第一阈值，第一阈值可采用中位数的方法确定，例如，将时空运动图中各像素的运动速度大小的中位数确定为第一阈值，然后，根据第一阈值确定视频片段中具备较大运动幅度的三维时空区域，例如，三维时空区域至少覆盖了超过预设比例（如 80%）的时空运动图中大于第一阈值的像素。

20 从而，通过时空运动图直接获得视频中具备较大运动的三维时空区域。

 第二种，在视频片段相应的运动信息包括视频片段相应的时间运动图的情况下，根据视频片段相应的时间运动图计算所述视频片段的运动幅度，例如，将视频片段相应的时间运动图作为视频帧级别的运动图，计算视频片段内所有帧的视频帧级别的运动图的均值，作为所述视频片段的运动幅度，然后对视频片段序列中的各视频片段进行时域采样，采样得到的视频片段的运动幅度不小于第二阈值，运动幅度小于第二阈值的视频片段可不被采样。第二阈值根据各视频片段的运动幅度确定，例如，将各视频片段的运动幅度的中位数作为第二阈值。

 从而，通过基于时间运动图的时域采样，可提取视频片段序列中具备较大运动的视频片段。

第三种，在视频片段相应的运动信息包括视频片段相应的空间运动图的情况下，根据视频片段相应的空间运动图中各像素的运动速度的大小确定第三阈值，根据第三阈值对各像素进行划分，对空间运动图反复执行随机多尺度空间裁剪、并确保裁剪得到的矩形空间区域至少覆盖了超过预设比例的空间运动图中大于第三阈值的像素，对
5 视频片段中的每一视频帧都裁剪与矩形空间区域相同的区域。

从而，通过基于空间运动图的空间裁剪，可获得视频片段中具备较大运动的三维时空区域。

上述的第二种和第三种还可以结合起来使用。也即，在运动图的引导下，运动聚焦的视频增强依次通过时域采样和空间裁剪两个步骤对原始视频数据进行采样。由于
10 在时域采样可过滤到一半的候选视频片段，减少空间裁剪的处理对象，提升视频数据增强的效率。

在运动聚焦的视频数据增强之后，对视频片段中的视频帧进行图像数据增强操作，如颜色抖动、随机灰度、随机模糊和随机镜像等。从而，保持传统视频数据增强方法中存在的随机性。

15 下面描述运动聚焦的特征学习。

利用从视频中提取得到的运动图作为模型特征学习的监督信号，进一步地引导模型的对比学习过程，如前所述，结合运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习。也即，运动聚焦的视频表征自监督对比学习的损失函数 $\mathcal{L} = \mathcal{L}_{MAL} + \mathcal{L}_{NCE}$ ，其中， $\mathcal{L}_{MAL}$ 表示运动对齐损失函数，例如为候选描述的 $\mathcal{L}_{MAL-v1}$ 、 $\mathcal{L}_{MAL-v2}$ 或
20 $\mathcal{L}_{MAL-v3}$ ， $\mathcal{L}_{NCE}$ 表示对比损失函数，例如为 InfoNCE。

常规的对比学习，给定一个经过编码器编码的查询样本 $\mathbf{q} \in \mathbb{R}^d$ ，一组包含了一个正例样本键值 $\mathbf{k}^+ \in \mathbb{R}^d$ 和 K 个负例样本键值 $\mathcal{K}^- = \{\mathbf{k}_i^-\}$ 的经过编码器编码的键值向量。其中查询样本和正例样本通常是对同一个数据实例（图像、视频等）进行不同数据增强后得到的样本，而负例样本则是采样自其它不同数据实例的样本。在对比学习中的
25 实例判别任务的目标是引导查询样本 $\mathbf{q}$ 与正例样本 $\mathbf{k}^+$ 之间在特征空间中更加相似，同时保证查询样本 $\mathbf{q}$ 与其它负例样本 $\mathcal{K}^-$ 之间存在足够的区分度。通常对比学习会采用 InfoNCE 作为其损失函数：

$$\mathcal{L}_{NCE} = -\log \frac{\exp(\mathbf{q}^T \mathbf{k}^+ / \tau)}{\exp(\mathbf{q}^T \mathbf{k}^+ / \tau) + \sum_{i=1}^K \exp(\mathbf{q}^T \mathbf{k}_i^- / \tau)}$$

其中， τ 是预设的超参数。

对比学习的损失函数是在编码后的视频样本（三维管道）级别上执行对比学习，在这个过程中三维管道中的每个时间-空间位置都被平等地看待。鉴于视频中的语义信息更多地集中在运动较为剧烈的区域，为了帮助模型在训练过程中更多地聚焦于视频中运动区域，更好地发掘视频中的运动信息，本公开提出了一种新的运动对齐损失（MAL, Motion Alignment Loss）来对齐主干网络卷积层的输出和视频样本的运动图中的运动幅度，并作为 InfoNCE 之外的监督信号作用与模型的优化过程，进而使学习到的视频特征表达能够更好地描述视频中的运动信息。

下面描述三种运动对齐损失相应的损失函数，简称运动对齐损失函数。

第一种运动对齐损失函数，对齐特征图，也即对齐主干网络最后的卷积层输出的特征图的幅度与运动图，从而使得主干网络输出的卷积层的特征图在运动较大的区域有着更大的响应。

第一种运动对齐损失函数表示为以下的一项或多项的累加：主干网络的最后卷积层输出的特征图在所有通道的累加与视频片段相应的时空运动图之间的距离，所述累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离，所述累加沿着空间维度的池化结果与视频片段相应的时间运动图之间的距离。

当包括以上三项时，第一种运动对齐损失函数表示为：

$$\mathcal{L}_{MAL-v1} = \|h^{ST} - \langle m^{ST} \rangle\|_2^2 + \|h^S - \langle m^S \rangle\|_2^2 + \|h^T - \langle m^T \rangle\|_2^2$$

其中， $h^{ST} = \langle \sum_c h_c \rangle$ ， h_c 表示卷积层输出的特征图的第 c 个通道的响应幅度， $\sum_c h_c$ 表示卷积层输出的特征图在所有通道的响应幅度的累加， h^{ST} 沿着时间维度的池化结果表示为 h^S ， h^{ST} 沿着空间维度的池化结果表示为 h^T ， m^{ST} 表示时空运动图， m^S 表示空间运动图， m^T 表示时间运动图。

第二种运动对齐损失函数，对齐加权特征图，即对齐主干网络的最后卷积层输出的特征图按照各通道的权重在所有通道的加权累加与运动图。

考虑到特征图相应的梯度幅值能够更好地衡量特征图中每个位置的特征对模型推理结果，即对比学习损失函数 InfoNCE 的贡献，故可以采用梯度幅值对特征图响应进行加权。各通道的权重的确定方法包括：计算视频片段相应的查询样本和正例样本之间的相似度相对于卷积层输出的特征图的某个通道的梯度，计算该通道的梯度的均值，作为该通道的权重。具体而言，根据 InfoNCE 损失函数的形式，首先需要计算查询样本和正例样本之间相似度 $q^T k^+$ 相对于卷积层输出的特征图的某个通道的梯度

$g_c = \frac{\partial q^T k^+}{\partial h_c}$, 然后对于每个通道 c , 计算梯度 g_c 的均值 w_c 用于表示通道 c 的权重, 最后利用各通道的权重对特征图进行通道维度的加权。

第二种运动对齐损失函数表示为以下的一项或多项的累加: 主干网络的最后卷积层输出的特征图按照各通道的权重在所有通道的第一加权累加与视频片段相应的时空运动图之间的距离, 所述第一加权累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离, 所述第一加权累加沿着空间维度的池化结果与视频片段相应的时间运动图之间的距离。

当包括以上三项时, 第二种运动对齐损失函数表示为:

$$\mathcal{L}_{MAL-v2} = \|\bar{h}^{ST} - \langle m^{ST} \rangle\|_2^2 + \|\bar{h}^S - \langle m^S \rangle\|_2^2 + \|\bar{h}^T - \langle m^T \rangle\|_2^2$$

其中, $\bar{h}^{ST} = \langle \text{ReLU}(\sum_c w_c h_c) \rangle$, h_c 表示卷积层输出的特征图的第 c 个通道的响应幅度, w_c 表示第 c 个通道的权重, ReLU表示线性整流函数(Rectified Linear Unit), $\bar{h}^{ST}$ 沿着时间维度的池化结果表示为 $\bar{h}^S$, $\bar{h}^{ST}$ 沿着空间维度的池化结果表示为 $\bar{h}^T$, m^{ST} 表示时空运动图, m^S 表示空间运动图, m^T 表示时间运动图。

第三种运动对齐损失函数, 对齐加权梯度图, 即对齐主干网络的最后卷积层输出的特征图的各通道的梯度按照各通道的权重在所有通道的加权累加与运动图, 如图4所示。

第三种运动对齐损失函数表示为以下的一项或多项的累加: 主干网络的最后卷积层输出的特征图的各通道的梯度按照各通道的权重在所有通道的第二加权累加与视频片段相应的时空运动图之间的距离, 所述第二加权累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离, 所述第二加权累加沿着空间维度的池化结果与视频片段相应的时间运动图之间的距离。其中, 各通道的权重的计算方法参见前述。

当包括以上三项时, 第三种运动对齐损失函数表示为:

$$\mathcal{L}_{MAL-v3} = \|\bar{g}^{ST} - \langle m^{ST} \rangle\|_2^2 + \|\bar{g}^S - \langle m^S \rangle\|_2^2 + \|\bar{g}^T - \langle m^T \rangle\|_2^2$$

其中, $\bar{g}^{ST} = \langle \text{ReLU}(\sum_c w_c g_c) \rangle$, $\bar{g}^{ST}$ 沿着时间维度的池化结果表示为 $\bar{g}^S$, $\bar{g}^{ST}$ 沿着空间维度的池化结果表示为 $\bar{g}^T$, m^{ST} 表示时空运动图, m^S 表示空间运动图, m^T 表示时间运动图, w_c 、 g_c 的含义参考前述。

通过上述各实施例, 学习得到视频表征模型, 根据学习得到的视频表征模型, 对待处理视频进行处理得到相应的视频特征。

图5为本公开一些实施例的运动聚焦的视频表征自监督对比学习装置的结构示意图。

如图 5 所示, 该实施例的装置 500 包括: 存储器 510 以及耦接至该存储器 510 的处理器 520, 处理器 520 被配置为基于存储在存储器 510 中的指令, 执行前述任意一些实施例中的运动聚焦的视频表征自监督对比学习方法。

其中, 存储器 510 例如可以包括系统存储器、固定非易失性存储介质等。系统存储器例如存储有操作系统、应用程序、引导装载程序 (Boot Loader) 以及其他程序等。

其中, 处理器 520 可以用通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、应用专用集成电路 (Application Specific Integrated Circuit, ASIC)、现场可编程门阵列 (Field Programmable Gate Array, FPGA) 或其它可编程逻辑设备、分立门或晶体管等分立硬件组件方式来实现。

10 装置 500 还可以包括输入输出接口 530、网络接口 540、存储接口 550 等。这些接口 530, 540, 550 以及存储器 510 和处理器 520 之间例如可以通过总线 560 连接。其中, 输入输出接口 530 为显示器、鼠标、键盘、触摸屏等输入输出设备提供连接接口。网络接口 540 为各种联网设备提供连接接口。存储接口 550 为 SD 卡、U 盘等外置存储设备提供连接接口。总线 560 可以使用多种总线结构中的任意总线结构。例如, 总线结构包括但不限于工业标准体系结构 (Industry Standard Architecture, ISA) 总线、微通道体系结构 (Micro Channel Architecture, MCA) 总线、外围组件互连 (Peripheral Component Interconnect, PCI) 总线。

本领域内的技术人员应当明白, 本公开的实施例可提供为方法、系统、或计算机程序产品。因此, 本公开可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且, 本公开可采用在一个或多个其中包含有计算机程序代码的非瞬时性计算机可读存储介质 (包括但不限于磁盘存储器、CD-ROM、光学存储器等) 上实施的计算机程序产品的形式。

本公开是参照根据本公开实施例的方法、设备 (系统)、和计算机程序产品的流程图和 / 或方框图来描述的。应理解为可由计算机程序指令实现流程图和 / 或方框图中的每一流程和 / 或方框、以及流程图和 / 或方框图中的流程和 / 或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器, 使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

- 5 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

- 10 以上所述仅为本公开的较佳实施例，并不用以限制本公开，凡在本公开的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本公开的保护范围之内。

权 利 要 求

1. 一种视频表征自监督对比学习方法，包括：

根据视频片段的每个视频帧相应的光流信息，计算得到所述视频片段的每个视频帧相应的运动幅度图；

根据所述视频片段的各视频帧相应的运动幅度图，确定所述视频片段相应的运动信息；

根据视频片段序列和每个视频片段相应的运动信息，进行视频表征自监督对比学习。

2. 根据权利要求1所述的方法，其中，所述根据视频片段的每个视频帧相应的光流信息，计算得到所述视频片段的每个视频帧相应的运动幅度图，包括：

提取视频片段中每一对相邻视频帧之间的光流场，以确定所述视频片段的每个视频帧相应的光流场；

计算每个视频帧相应的光流场在第一方向和第二方向的梯度场；

将第一方向和第二方向的梯度场的幅值进行聚合得到每个视频帧相应的运动幅度图。

3. 根据权利要求2所述的方法，其中，第一方向和第二方向相互垂直。

4. 根据权利要求2所述的方法，其中，所述计算每个视频帧相应的光流场在第一方向和第二方向的梯度场包括：

计算每个视频帧相应的光流场的水平分量在第一方向和第二方向的梯度；

计算每个视频帧相应的光流场的垂直分量在第一方向和第二方向的梯度；

每个视频帧相应的光流场的水平分量和垂直分量在第一方向和第二方向的梯度，构成所述光流场在第一方向和第二方向的梯度场。

5. 根据权利要求1所述的方法，其中，所述视频片段相应的运动信息包括：所述视频片段相应的时空运动图、空间运动图、时间运动图中的一项或多项；其中，

确定所述视频片段相应的时空运动图包括：在时间维度将视频片段的各视频帧的运动幅度图叠加构成所述视频片段的时空运动图；

确定所述视频片段相应的空间运动图包括：对所述视频片段的时空运动图沿着时间维度进行池化，得到所述视频片段的空间运动图；

确定所述视频片段相应的时间运动图包括：对所述视频片段的时空运动图沿着空间维度进行池化，得到所述视频片段的时间运动图。

6. 根据权利要求1所述的方法，其中，所述根据视频片段序列和每个视频片段相应的运动信息，进行视频表征自监督对比学习，包括：

根据每个视频片段相应的运动信息对视频片段进行数据增强，根据增强后的视频片段序列并结合对比损失进行视频表征自监督对比学习；或者，

根据视频片段序列并结合运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习；或者，

根据每个视频片段相应的运动信息对视频片段进行数据增强，根据增强后的视频片段序列并结合运动对齐损失和对比损失进行运动聚焦的视频表征自监督对比学习；

其中，所述运动对齐损失通过对齐进行学习的主干网络的最后卷积层的输出与视频片段相应的运动信息来确定。

7. 根据权利要求6所述的方法，其中，所述根据每个视频片段相应的运动信息对视频片段进行数据增强包括：

在视频片段相应的运动信息包括视频片段相应的时空运动图的情况下，根据时空运动图中各像素的运动速度的大小确定第一阈值，根据第一阈值确定视频片段中具备较大运动幅度的三维时空区域；或者，

在视频片段相应的运动信息包括视频片段相应的时间运动图的情况下，根据视频片段相应的时间运动图计算所述视频片段的运动幅度，对序列中的各视频片段进行时域采样，采样得到的视频片段的运动幅度不小于第二阈值，第二阈值根据各视频片段的运动幅度确定；或者，

在视频片段相应的运动信息包括视频片段相应的空间运动图的情况下，根据视频片段相应的空间运动图中各像素的运动速度的大小确定第三阈值，根据第三阈值对各

像素进行划分，对空间运动图反复执行随机多尺度空间裁剪、并确保裁剪得到的矩形空间区域至少覆盖了超过预设比例的空间运动图中大于第三阈值的像素，对视频片段中的每一视频帧都裁剪与矩形空间区域相同的区域。

8. 根据权利要求7所述的方法，其中，所述根据视频片段相应的时间运动图计算所述视频片段的运动幅度包括：

将视频片段相应的时间运动图作为视频帧级别的运动图，计算视频片段内所有帧的视频帧级别的运动图的均值，作为所述视频片段的运动幅度。

9. 根据权利要求7所述的方法，其中，第一阈值、第二阈值、第三阈值分别采用中位数的方法确定。

10. 根据权利要求7所述的方法，其中，所述对视频片段进行数据增强还包括：对视频片段中的视频帧进行图像数据增强操作。

11. 根据权利要求6所述的方法，其中，所述运动对齐损失相应的损失函数表示为以下的一项或多项的累加：

主干网络的最后卷积层输出的特征图在所有通道的累加与视频片段相应的时空运动图之间的距离，

所述累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离，

所述累加沿着空间维度的池化结果与视频片段相应的时间运动图之间的距离。

12. 根据权利要求6所述的方法，其中，所述运动对齐损失相应的损失函数表示为以下的一项或多项的累加：

主干网络的最后卷积层输出的特征图按照各通道的权重在所有通道的第一加权累加与视频片段相应的时空运动图之间的距离，

所述第一加权累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离，

所述第一加权累加沿着空间维度的池化结果与视频片段相应的时间运动图之间

的距离。

13. 根据权利要求 6 所述的方法，其中，所述运动对齐损失相应的损失函数表示为以下的一项或多项的累加：

主干网络的最后卷积层输出的特征图的各通道的梯度按照各通道的权重在所有通道的第二加权累加与视频片段相应的时空运动图之间的距离，

所述第二加权累加沿着时间维度的池化结果与视频片段相应的空间运动图之间的距离，

所述第二加权累加沿着空间维度的池化结果与视频片段相应的空间运动图之间的距离。

14. 根据权利要求 12 或 13 所述的方法，其中，通道的权重的确定方法包括：

计算视频片段相应的查询样本和正例样本之间的相似度相对于卷积层输出的特征图的某个通道的梯度，计算该通道的梯度的均值，作为该通道的权重。

15. 根据权利要求 6 所述的方法，其中，所述对比损失根据对比学习的损失函数确定。

16. 根据权利要求 15 所述的方法，其中，对比学习的损失函数包括 InfoNCE 损失函数。

17. 根据权利要求 6 所述的方法，其中，所述主干网络包括三维卷积神经网络。

18. 根据权利要求 1 所述的方法，还包括：

根据学习得到的视频表征模型，对待处理视频进行处理得到相应的视频特征。

19. 一种视频表征自监督对比学习装置，包括：

存储器；以及

耦接至所述存储器的处理器，所述处理器被配置为基于存储在所述存储器中的指

令，执行权利要求 1-18 中任一项所述的视频表征自监督对比学习方法。

20. 一种非瞬时性计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现权利要求 1-18 中任一项所述的视频表征自监督对比学习方法的步骤。