自监督流形结构的第一视角视频时序分割算法

张明明[1,2)] 闫小强[1)] 孙中川[1)] 胡世哲[1)] 叶阳东[1)]

[1)](郑州大学计算机与人工智能学院 郑州 450000)

[2)](河南工业大学人工智能与大数据学院 郑州 450000)

摘 要 随着可穿戴设备和智能存储技术的普及,第一视角视频的使用量高速增长。将这类视频划分成独立的视频片段以提取关键的内容信息,成为了视频理解领域的重要研究方向。这类视频数据规模大、维度高、内容多样,基于欧氏空间的特征学习方法难以有效地处理复杂高维的视频数据。现有时序分割算法在处理第一视角长视频时,很难应对因手部遮挡和运动模糊而导致的帧信息丢失问题。针对上述问题,本文提出了一种自监督流形结构的第一视角视频时序分割算法(Self-Supervised Manifold Structure, SSMS)。受高维视频数据在低维流形空间中具有相似语义聚集现象的启发,该算法将包含时序信息的帧特征进行低维嵌入,使得语义相似的帧特征映射到流形空间中相近位置。首先,本文提出了一种改进的局部流形结构特征学习策略,提取帧数据的局部流形结构。其次,SSMS算法构建了动态时序网络,基于最大相似关系来获得具有不变性的特征表示。然后,将帧数据的流形结构特征作为监督信号进行自监督学习。经过不断迭代优化,得到低维高质量的帧数据特征。最后,通过聚类过程实现第一视角视频的无监督时序分割,避免了标注数据的限制和成本。相比于现有的无监督时序分割算法,本文方法在五个第一视角数据集上平均提高了3.37%的准确度。

关键词 第一视角视频;流形结构;自监督学习;时序分割;特征表示

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.00266

Self-Supervised Manifold-Structured Algorithm for Egocentric Video Temporal Segmentation

ZHANG Ming-Ming[1,2)] YAN Xiao-Qiang[1)] SUN Zhong-Chuan[1)] HU Shi-Zhe[1)] YE Yang-Dong[1)]

[1)](School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450000)

[2)](School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450000)

Abstract With the increasing popularity of wearable devices and intelligent storage technology, the usage of egocentric view videos is experiencing rapid growth. Segmenting such videos into independent video segments to extract key content information has become an important research direction in the field of video understanding. These videos exhibit characteristics of large scale, high dimensionality, and diverse content, posing challenges for Euclidean-based feature learning methods to effectively handle complex high-dimensional video data. Existing temporal segmentation algorithms struggle to address the issue of frame information loss in long egocentric videos due to hand occlusion and motion blur. To address these challenges, this paper proposes a self-supervised Manifold Structure algorithm (SSMS) for egocentric temporal segmentation. Inspired by the similar semantic clustering phenomenon of high-dimensional video data in low-dimensional manifold space, the algorithm embeds frames containing temporal information into a

收稿日期:2023-12-13;在线发布日期:2024-09-30。本课题得到国家自然科学基金(62176239)资助。**张明明**,博士,中国计算机学会(CCF)会员,主要研究领域为机器学习、计算机视觉。E-mail:mmzhang@haut.edu.cn。**闫小强**,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、计算机视觉。**孙中川**,博士,主要研究领域为深度学习、推荐系统。**胡世哲**,博士,中国计算机学会(CCF)会员,主要研究领域为模式识别、信息理论。**叶阳东**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为机器学习、知识工程、智能系统。E-mail:yeyd@zzu.edu.cn。

low-dimensional manner, mapping semantically similar frames to nearby positions in the manifold space. Firstly, an improved local manifold structure feature learning method is introduced to extract local manifold structures from frames. Secondly, the SSMS algorithm constructs a dynamic temporal network to obtain invariant feature representation based on the maximum similarity relationship. Then, the manifold structure features of the frame data are used as supervised signals for self-supervised learning. Through iterative optimization, high-quality low-dimensional features of frame data are obtained. Finally, unsupervised temporal segmentation of egocentric videos is achieved through a clustering process, avoiding the limitations and costs associated with annotated data. Compared to existing unsupervised temporal segmentation algorithms, the accuracy of our method has improved by an average of 3.37% on five egocentric datasets.

Keywords　egocentric video; manifold structure; self-supervised learning; temporal segmentation; feature representation

1　引　言

近年来，随着新一代人工智能由感知智能转向认知智能，第一视角视频分析也进入到了智能化发展阶段[1]。第一视角视频由佩戴在人或机器上的摄像头记录产生，也可在虚拟现实环境下自动生成。以佩戴者或观测者的视角来记录当前所交互的场景，更贴近人类的感知并能够反映真实的行为意图。第一视角视频分析有着广泛的应用场景，如自动驾驶、元宇宙、智能养老等。在自动驾驶场景下，第一视角视频行驶记录能够为驾驶员提供实时的路况信息。自动驾驶系统则根据感知的第一视角视频信息处理、判断和决策，替代驾驶员做出行驶动作[2]。在元宇宙场景下，算法可以使用在虚拟环境中所发生的第一视角交互记录，实时向使用者提供个性化服务。在智能养老场景下，人们使用第一视角记录老人的日常活动，并通过视频分析来实现辅助建议、事件提醒和危险预警等功能。因此，第一视角视频分析具有极其重要的研究价值。

时序分割是解决第一视角视频分析的一项基础且充满挑战的任务[3]。视频时序分割任务指的是给定一段视频数据，通过挖掘视频画面的语义信息、时序关系以及运动状态将整段视频划分为包含特定事件的视频片段，从而实现对视频内容的理解和分析。如图1所示，人们通过获取视频帧的语义信息以及时序关系将整段视频划分为包含特定活动的视频片段。当前的第一视角视频时序分割问题主要存在以下难点：(1)视频画面的高维特征难以在低维空间有效表示。如果将视频画面中的像素点对应于空间上的一个维度，一个帧画面则可视为高维图像空间中的一个点，一种行为活动在不同时间上所有帧的集合就是图像空间上的一个连续流形。然而大多数基于深度学习的视频特征提取算法仅考虑帧的视觉以及时序信息，忽略了行为活动固有的结构信息，难以有效识别不同的行为。(2)第一视角视频画面不稳定，算法识别准确率低。由于佩戴者的身体不断活动，捕捉到的第一视角视频画面存在运动时差和运动模糊等问题，相邻帧之间画面易出现空间上的不连续。(3)标注数据集有限。当前很多监督算法需要大量的标注数据才能发挥作用。不同于第三视角视频数据集，第一视角视频主要以记录为目的，每个视频记录通常持续几十分钟甚至数小时，数据冗余大，逐帧的标注将花费大量的人力物力。

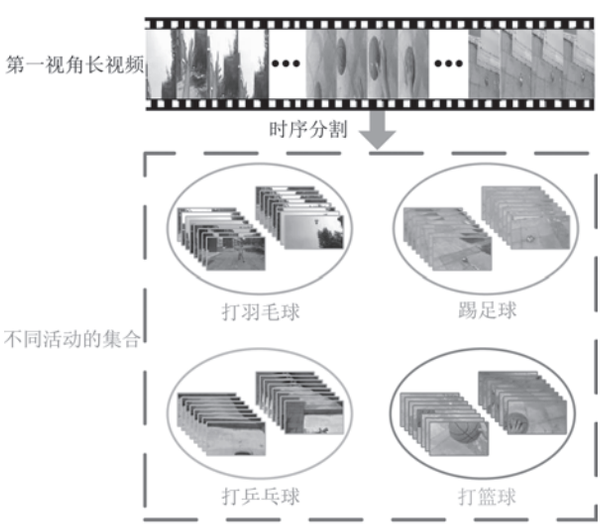

图1　时序分割的总流程

针对第一视角视频时序分割的难点，研究人员从多个方面进行探究：(1)从特征提取的方面解决时序分割问题：这类算法通过充分利用视频的时间连

续性，性能取得显著提升[4-5]。其中，自监督学习算法通常引入额外的辅助任务，促使模型自动学习有用的视觉特征。由于辅助任务的引入，复杂的学习过程增加了算法对计算资源的需求。李等人[6]考虑了视频的帧级和动作级结构，采用了自监督学习的方式来实现无监督时序分割问题。在自监督学习过程中，采用RNN(Recurrent Neural Network)网络识别正负样本，并将RNN的隐藏层作为动作级别特征嵌入。(2)通过高级语义信息解决时序分割问题：Sarfraz等人[7]提出了一种有效的时间加权分层聚类算法，该算法通过1阶最近邻图的方式来表示视频内容，形成语义和时间一致的帧簇。Bueno-Benitode等人[8]提出了一种深度度量学习方法，有效地对时间和语义建模，以在新的表示空间中解决动作分割问题。(3)采用检测活动边界的方式解决无监督动作分割问题[9-12]：王等人[12]提出了一种阶段级联的边感知网络，该网络能够自适应地调节感受野，并对模糊帧进行预测。杜等人[10]通过检测边界的方式来解决无监督动作分割问题，该方法估计平滑帧之间的相似性，通过非最大抑制方法选择最小的点作为候选边界。李等人[13]提出了一种全局和局部特征提取以及边界选择的方法，通过融合特征并检测显著边界进行动作分割。第一视角视频以日志记录为主，视频内容的相似度较大，仅依靠视频的自身属性很容易出现过度分割和边界模糊等问题。

以上方法虽在时序分割相关问题上有所探索，但仍存在不足之处。第一视角摄像头通常佩戴在受试者身体上，捕捉到的画面更容易出现运动时差和运动模糊等问题，使得采集的视频样本具有较大的类内离散性以及类间模糊性。通过增强内在时序结构的关联性可以减少无关信息对时序分割任务的影响。研究表明，视频图像数据往往分布在一个低维流形空间上，采用传统的欧氏表征并不能有效表示数据特征[14-15]。现有的流形学习方法很少考虑有关时序的流形映射，通常基于数据内部关系结构的通用学习方法[16-17]。相比于其他高维数据，第一视角视频数据具有时空上的相似性，即帧数据之间会受到时间距离以及时间顺序的影响。通过分析时序数据的局部流形结构，可以揭示出这些动态变化的底层模式。比如，连续的时间步对应着在流形结构上紧密相邻的点，而时间步间的突变可能表现为流形结构上的“跳跃”。单纯依赖时序特征的低维表示可能无法充分反映数据的复杂动态，而通过结合局部流形结构，可以更好地引导低维表示的生成，使得它们不仅能够保留数据的全局信息，还能捕捉到局部的细微变化和动态特性。本文提出了一种自监督流形结构的第一视角视频时序分割算法(Self-Supervised Manifold Structure，SSMS)。该算法通过时间窗口捕捉视频内的局部流形信息，拟合平滑的流形结构进行自监督学习，从而获得有效的视频特征表示。首先，SSMS算法使用了一种滑动窗口的策略来获取视频帧的局部时序流形结构。第一视角视频数据具有区域短时序依赖特性，即在一定的时空域内，佩戴者所进行的行为活动是连续且一致的，视频帧之间具有语义相似性。因此，相比于全局的流形特征，限定在一定窗口内获得的局部流形特征可以减少不同行为产生的语义干扰。接着，构建了一种动态时序网络来获取最大化的近邻时序信息。在短时空域内，第一视角视频的相邻帧之间具有较强语义相似性。但佩戴者的剧烈运动导致摄像头捕捉的部分画面中出现交互物体的模糊、变形、画面丢失等问题。这些异常画面为局部时序信息融合带来一定的干扰。在动态时序网络中通过对基于目标帧的近邻矩阵重排序，获得基于目标帧的最大近邻集。接着将最大近邻集输入到时序网络中，获得低维近邻时序特征。同时，SSMS算法设计了一种基于自监督学习的辅助任务，将流形特征作为自监督信号，使数据的流形信息以及近邻时序信息融合在同一目标函数中。最后，使用K-Means算法对得到的低维高级语义特征进行聚类分析来获得时序分割的结果。在五个第一视角视频数据集以及两个大规模数据集上的实验结果表明SSMS算法性能优于现有方法。本文的主要创新点总结如下：

(1)提出了一种自监督流形结构的第一视角视频时序分割算法。该算法采用自监督学习的方式，通过融入局部流形结构信息以及动态时序信息，以更全面地挖掘和利用第一视角视频的数据特征，从而实现了第一视角视频数据的无监督时序分割。

(2)提出了一种改进的局部流形结构信息获取方法。在一定的时域范围内，使用局部流形结构信息能够减少全局不同的行为活动所产生的语义干扰。

(3)构建了一种动态时序网络。该网络通过构建基于目标帧的近邻相似矩阵来度量获取时域内的近邻信息，从而有效避免异常画面对算法的干扰。

(4)提出了一种新颖的自监督学习策略。该方法将流形特征以及时序特征作为自监督信号，使得数据的流形信息以及近邻时序信息融合在同一目标函数中，从而获得更加有效的特征表达。

本文第2节回顾相关工作；第3节介绍本文所研究的第一视角视频时序分割方法的实现细节；第4节对本文提出的方法进行实验评估及分析；第5节介绍当前算法的局限性；第6节总结全文并对未来工作进行展望。

2　相关工作

2.1　第一视角视频分析

随着可穿戴设备的广泛使用，面向第一视角的视频数据快速、持续增长。第一视角视频的研究受到了越来越多的关注[1, 18]，例如行为识别[19-20]、行为预测[21-22]、视频总结[23-24]等。传统的算法主要基于手工特征来实现第一视角视频的活动识别，如物体识别[25]、手势识别[26]、眼动识别[27]等。目前基于深度学习的算法取得了不错的表现。张等人[28]提出了一种多任务聚类算法，用于解决第一视角视频中时序分割问题。宋等人[29]提出了一种多模态多流的深度学习框架来解决第一视角活动识别问题。Zatsarynna等人[30]提出了一种基于时间卷积的多模态算法，用来预测第一视角中人类行为。Perochon等人[31]提出了针对第一视角视频的无监督时序分割算法，该算法可以检测信息损失的帧，并使用核变化点检测法来估计动作边界。本文关注的是第一视角视频中无监督时序分割问题。

2.2　无监督时序分割

时序动作分割指的是一种对未修剪视频中的每一帧依据活动的种类进行分类的任务，它是理解复杂活动的基础与关键的任务。目前，很多深度学习算法在动作理解上取得了阶段性的进展。然而，在时序动作分割的任务上，依旧有很多问题值得进一步去探索。熊等人[4]采用卷积神经网络和双向LSTM来捕捉视频的局部时序关联性和全局时序信息。Lea等人[5]提出了一种时间卷积网络的方法对视频帧进行动作分割。该算法使用池化和上采样的方式捕捉长期的时间关系。雷等人[32]进一步扩展了时间卷积网络，提出时间可变形残差网络。该算法同时计算两个时间流以分析全时间的视频信息残差流。Farha等人[33]在Lea的基础上，提出了一种多阶段时间卷积网络的算法。该算法构建了多阶段的时间卷积来分层预测活动分割。黄等人[34]提出了一种基于图的时序推理模型。该模型在现有的动作分类模型的基础上，使用图网络建模不同时间跨度内多个动作分段之间的关系。以上算法依靠视频中时空层面的时序关系进行建模，而本文则引入了自监督学习策略，使用相邻帧之间的流形结构信息对视频中的时序关系进行建模。该策略能够进一步地捕捉到短时空域的时序关系，增强时序特征的表征能力。为了减少对视频中每个帧进行动作标注，人们通过弱监督或无监督的方式来提高性能并减少标注的依赖[35-36]。Sarfraz等人[7]提出了一种时间加权层次聚类算法，该算法编码了最近邻图上帧之间的时空相似性，从而辅助层次聚类过程。王等人[37]提出了一种扩展的时态推理图模型来捕捉和模拟不同时段内动作之间的时间依赖关系。Kruger等人[38]引入了一种基于自相似结构的聚类运动时序分割方法。本文通过学习局部流形结构信息以及相邻帧之间时序关系的自监督学习的方式来解决无监督的时序分割问题。

2.3　流形学习

经典的流形学习算法通过假定数据点嵌入到欧几里得空间中，并通过局部或全局的映射来计算数据点的相似度。如局部线性嵌入[39]、拉普拉斯特征映射[16]、Hessian局部线性映射[40]、增强的局部线性嵌入[41]等。这些算法假定K个最近邻点近似地描述局部邻居点的流形。还有一些算法通过全局的视角来构建流形空间，如等距特征映射[42]、统一流形逼近与投影[17]、t分布和随机近邻嵌入[43]等。然而对于视频数据，这些流形嵌入的算法仅仅考虑数据之间的特征空间信息，缺少相邻数据之间的时序关系。本文算法中的流形嵌入阶段借鉴了统一流形逼近与投影[17]。该方法建立在黎曼几何和代数拓扑理论框架上，获得了目标帧的局部流形结构。

3　研究框架

本文的目标主要是学习视频帧的低维特征，并使用聚类算法进行无监督的时序分割。将一段视频按一定的帧率进行采样，集合$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中$\boldsymbol{x}_t\in\mathbb{R}^{1\times d}$表示的是第$t$时刻采样帧的$d$维视觉特征，$n$是采样帧的数量。真实的标签$Y=\{y_1,y_2,\cdots,y_n\}$表示采样帧所对应的活动类别，以此作为聚类结果的评价依据。SSMS算法中特征学习过程如图2所示。给定一段第一视角视频帧集合，其中视频帧的局部流形结构以及时序特征都是从不同方面对当前帧的描述，两者之间的对应关系是自监督学习的关键。首先，算法将视频帧输入到

ImageNet数据集上预训练好的ResNet50网络中，得到视觉特征集$\boldsymbol{X}$。接着，SSMS算法引入了滑动窗口的策略获得视频帧的局部流形结构。算法将视频帧按照一定的长度分为n/L_W个子集合，L_W为滑动窗口大小。每个子集合对应着滑动窗口的一次流形嵌入操作，最终得到视频帧的流形表示矩阵$\boldsymbol{M}$。SSMS算法构建了最大相似近邻矩阵$\boldsymbol{N}$，得到每个帧对应的局部时域集合$\boldsymbol{N}_t$。将$\boldsymbol{N}_t$输入到时序编码器后，便得到包含动态时序信息的低维特征g_t。最后通过自监督学习方法，将局部流形结构信息与视频的动态时序信息映射到统一的自监督损失中。经过不断地迭代优化，输出最终的低维特征$\boldsymbol{G}$。

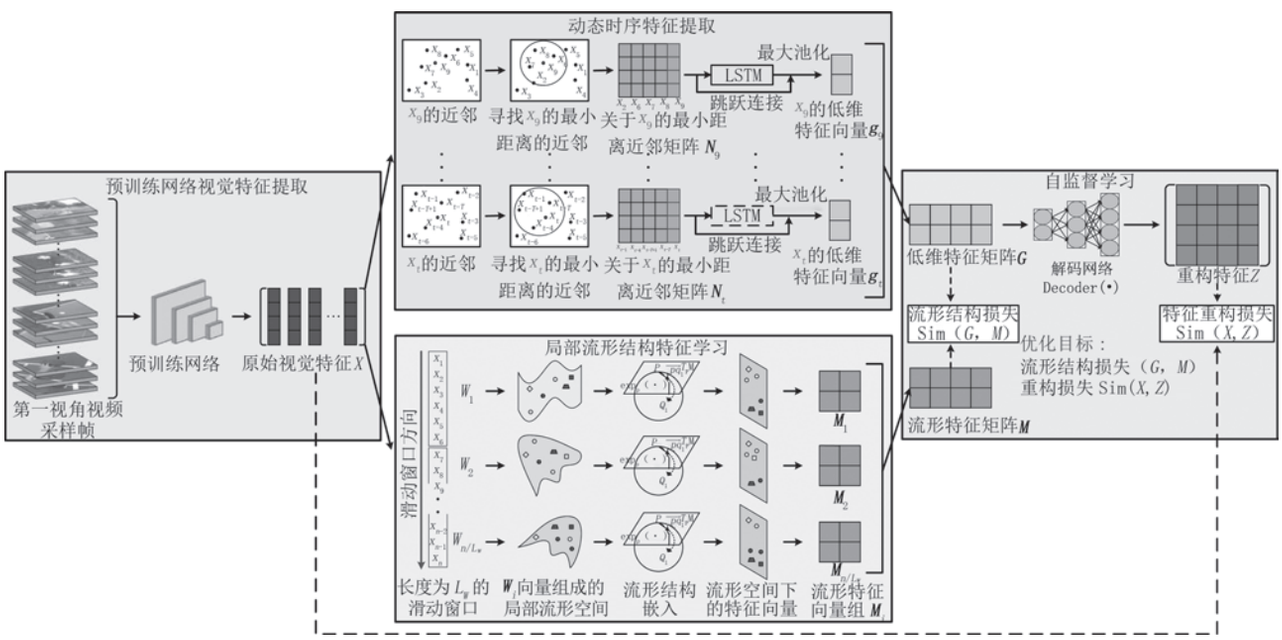

图2 SSMS算法中特征学习过程

算法1 局部流形嵌入算法

输入：第一视角视频的特征集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，滑动窗口$L_W$，嵌入维度$D$，近邻逼近数量$K$，流形嵌入的紧密程度$min_dist$

输出：视频帧$\boldsymbol{X}$的局部流形嵌入矩阵$\boldsymbol{M}$

1. **FOR ALL** $i<n/L_W$ **DO**}
2. 依次取L_W个帧构成$\boldsymbol{W}_i$，其中$\boldsymbol{W}_i=\{x_{i*L_W},\cdots,x_{i*L_W+L_W}\}$
3. $F\_set[L_i]\leftarrow LocalFuzzy(L_i,L_W,K)$；//将$N_i$中所有数据点构建统一的模糊单形集
4. $top\_Rep\leftarrow\cup_{t\in L_i}F\_set[Li]$；//使用t-余模求出最上面的表示$top_Rep$
5. $\boldsymbol{M}_i\leftarrow SpectralEmbedding(top\_Rep,D)$；//使用谱嵌入的方法，将$top_Rep$嵌入到$D$维数据$M_i$中
6. $\boldsymbol{M}_i\leftarrow OptimizeEmbedding(top\_Rep,i,min\_dist)$；//通过设定的紧密程度$min_dist$对$M_i$进行优化
7. **END FOR**
8. $\boldsymbol{M}=\{m_i|i=1,2,\cdots,n/L_W\}$。//输出流形特征

3.1 局部流形结构特征学习

本节提出了一种改进的方法提取视频帧的局部流形结构特征，首先，采用滑动窗口的策略对视频帧进行采样，生成多组无重叠的视频帧集。然后使用均匀流形近似映射的方式学习其流形结构，得到所有视频帧的局部流形结构特征集。

给定视频帧特征集$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，构建长度为$L_W$的滑动窗口，流形特征嵌入维度为$D$，$D<L_W$。第一视角的相邻帧序列具有内容致性，因此在局部范围内流形结构是统一且相近的。本文将视频分为n/L_W个视频片段$\boldsymbol{W}=\{\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_{n/L_W}\}$，每个片段包含长度为$L_W$视频帧。对于每个小片段$\boldsymbol{W}_i$，本文使用近似黎曼流形[44]的方式，生成均匀流形的低维表示$\boldsymbol{M}_i$。

由黎曼流形的相关定理[17]可知，假设数据在流形上是均匀分布的（相对于g），无论流形中心在哪里，在任何固定体积的球面上，都包含相同数量的$\boldsymbol{X}$点。因此，在一个以$\boldsymbol{X}_i$为中心的球近似包含固定的K个近邻。这样算法可以用$\boldsymbol{X}_i$到k个近邻的归一化距离来近似$\boldsymbol{X}_i$与其近邻的测地距离。因为流形假设下均匀分布的有效性，对每一个$\boldsymbol{X}_i$创建单独的自定义距离，将其合并为一致的全局结构，并通过度

量空间转换成可以用来表示全局流形的模糊单形集[45]。然后，算法找寻与源数据的拓扑结构紧密匹配的低维向量，以此得到数据的流形表示。具体流形嵌入的流程如算法1所示。其中滑动窗口L_w设置为250，流形嵌入维度D为200。由于近邻嵌入规则，滑动窗口长度L_w的设置应大于流形嵌入维度。若滑动窗口长度L_w设置过小，则窗口内采样帧数量也会变小，导致流形嵌入效果不明显。流形嵌入的超参数紧密程度*min_dist*设置为0.1，它定义了相邻嵌入点之间的期望分离，控制了嵌入后低维表示之间的距离。近邻逼近数量K设置为15，该参数控制了一个点直接连接到其第几个邻居的可能性。至此，算法得到了局部的流形特征$\boldsymbol{M}$。由于视频中包含帧的数量是庞大的，如果使用全局的流形特征得到的低维嵌入，流形特征会变得缺乏局部时序信息。

3.2　动态时序特征提取

在第一视角视频中，相邻视频帧具有较强的时序关系。SSMS算法使用了共享参数的LSTM结构获取每个帧的前向时序信息。佩戴者的运动会导致某些时序帧画面不受控制地发生改变，选取哪些稳定的相邻视频帧作为LSTM网络的输入，也是需要解决的一个关键问题。本节构建了一种基于目标帧的最大相似近邻矩阵，保证输入的近邻帧包含稳定的近邻信息。

给定近邻时域长度L_N，最大相似近邻数*ner*，构建关于视频帧集X的近邻距离集*pairwide_dist*，其中*pairwide_dist*=$\{x_1^{dist}, x_2^{dist}, \cdots, x_n^{dist}\}$，$x_1^{dist}$为在近邻长度$L_N$范围内$x_1$与其他近邻帧的距离。两个视频帧之间的距离公式如下计算：

$$Dist(\boldsymbol{X}_i, \boldsymbol{X}_j)=\sum_{k=1}^{d}\left(X_{ik}-X_{jk}\right)^2 \tag{1}$$

其中X_{ik}为$\boldsymbol{X}_i$第k维特征值，d为X_{ik}的特征维度。对*pairwide_dist*集进行排序，取出包含$\boldsymbol{X}_i$的前*ner*+1个最小距离的近邻矩阵$\boldsymbol{N}_i\in\mathbb{R}^{ner*M}$，并按照时序顺序对该近邻矩阵进行重排序。

给定输入时域大小T，最大相似近邻矩阵$\boldsymbol{N}$，SSMS评估每一个近邻x_{t-k}的隐藏状态h_{t-k}作为下一个帧x_{t-k+1}的隐藏状态。对于第$t-k-$th时刻，时序网络的隐藏状态如下计算：

$$\boldsymbol{f}_{t-k}=\sigma\left(W_{xf}\boldsymbol{x}_{t-k}+W_{hf}\boldsymbol{h}_{t-k-1}+\boldsymbol{b}_f\right) \tag{2}$$

$$\boldsymbol{i}_{t-k}=\sigma(W_{xi}\boldsymbol{x}_{t-k}+W_{hi}\boldsymbol{h}_{t-k-1}+\boldsymbol{b}_i) \tag{3}$$

$$\boldsymbol{o}_{t-k}=\sigma(W_{xo}\boldsymbol{x}_{t-k}+W_{ho}\boldsymbol{h}_{t-k-1}+\boldsymbol{b}_o) \tag{4}$$

$$\boldsymbol{c}_{t-k}=\boldsymbol{f}_{t-k}\odot\boldsymbol{c}_{t-k-1}+\boldsymbol{i}_{t-k}\odot\tanh(W_{xc}\boldsymbol{x}_{t-k}+W_{hc}\boldsymbol{h}_{t-k-1}+\boldsymbol{b}_c) \tag{5}$$

$$\boldsymbol{h}_{t-k}=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_{t-k}) \tag{6}$$

以上公式包含了输入门$\boldsymbol{i}_{t-k}$，遗忘门$\boldsymbol{f}_{t-k}$以及输出门$\boldsymbol{o}_{t-k}$。$\boldsymbol{c}_{t-k}$以及$\boldsymbol{h}_{t-k}$分别代表了记忆状态、隐藏状态的向量。σ是Sigmoid函数，$\odot$代表了两个元素的点乘。W_{**}是线性转换函数的参数，$\boldsymbol{b}_*$代表了偏差向量。最后，将所有LSTM网络输出的结果$\boldsymbol{h}_1,\cdots,\boldsymbol{h}_t$通过最大池化Pooling，并采用shortcut连接的方式，避免$\boldsymbol{h}_t$的信息遗失。对于g_t的第u维，可以由以下公式得到：

$$\boldsymbol{g}_t^u=\text{Pooling}\left(\boldsymbol{h}_{t-T}^u,\boldsymbol{h}_{t-T+1}^u,\cdots,\boldsymbol{h}_t^u\right)+\boldsymbol{h}_t^u \tag{7}$$

3.3　自监督学习过程

SSMS算法中自监督学习过程主要包括两个部分，其中一个是原特征的重构学习，另一个是局部流形结构特征的对比学习。对于原特征的重构学习，先将g_t重构为z_t：

$$\boldsymbol{z}_t=\text{Decoder}\left(\boldsymbol{g}_t\right)=\sigma\left(\omega t\left(\boldsymbol{g}_t\right)+\boldsymbol{b}\right) \tag{8}$$

其中Decoder(*)是解码过程，σ是Sigmoid函数。将D维的隐藏层特征g_t恢复到d维的重构特征向量$\boldsymbol{z}_t$。本文中采用了一层线性层和一个Sigmoid()激活函数作为解码网络，解码器的输入维度为D，输出维度为d。

重构特征向量$\boldsymbol{z}_t$并不能与原始$\boldsymbol{x}_t$完全相同，可以计算两者之间的相似度作为重构误差。对于成对帧的多维帧向量的相似度，可以使用交叉熵函数来度量：

$$\text{Sim}(\boldsymbol{z}_t,\boldsymbol{x}_t)=\boldsymbol{x}_t\cdot\log\boldsymbol{z}_t+(1-\boldsymbol{x}_t)\cdot\log(1-\boldsymbol{z}_t) \tag{9}$$

因此，重构损失为

$$L_R=\text{Sim}(\boldsymbol{x}_t,\boldsymbol{z}_t) \tag{10}$$

接下来，介绍局部流形结构特征的对比学习。这种方法的核心在于通过将局部流形结构与时序特征降维相结合，可以更好地保留和表达时序数据的动态特性，从而提升模型的性能和数据的可解释性。首先，设定动态时序网络输出的特征维度与流形结构特征维度相同。然后将时序特征与流形特征进行对比学习，构建流形结构对比损失：

$$L_M=\text{Sim}\left(\boldsymbol{g}_t,\boldsymbol{m}_t\right) \tag{11}$$

其中Sim()函数为公式(9)中交叉熵相似度度量函数，$\boldsymbol{g}_t$表示时序特征网络在第t时刻输出的D维隐藏层特征，$\boldsymbol{m}_t$为该时刻下的局部流形结构特征。通过最小化交叉熵，网络被优化为更接近局部流形结构特征的输出，从而使得网络的特征学习与流形结构的几何关系紧密结合。因此，时序特征网络能够更

好地学习到与局部流形结构一致的特征表示，捕捉到视频帧之间的局部几何特性和时空关系。

因此，最终的SSMS的目标函数为

$$L = L_R + L_M \tag{12}$$

3.4 聚类过程

给定一段第一视角视频，以滑动窗口的方式获取每个帧的局部流形特征，同时将每个帧的*ner*个近邻输入到动态时序网络中，以自监督学习的方式得到了包含时序关系以及局部流形关系的每个帧的低维特征。接下来，给定聚类数目，使用K-Means聚类算法对这些特征进行聚类划分，得到时序结果。

在K-Means算法中，使用欧氏距离来计算数据点与簇中心之间的距离：

$$dist_{Ed}(x, C) = \sqrt{\sum_{u=1}^{D} |x_u - C_u|^2} \tag{13}$$

其中，x为任一数据点，C为某个簇中心，D为数据点的维度。

3.5 算法优化

使用Adam优化器来优化模型的参数，设置一个较小的学习率0.001。为了更清楚地描述算法，整个算法流程如图2所示。

算法2. SSMS算法

输入：第一视角视频帧集$\boldsymbol{X}$，其$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$中流形学习相关参数$\theta$，近邻时域长度$L_N$，最大相似近邻数*ner*，最大迭代次数*maxIter*。

输出：聚类划分C

1. $\boldsymbol{W}\leftarrow\boldsymbol{X}$;//划分视频片段
2. $\boldsymbol{M}\leftarrow\theta, \boldsymbol{W}$;//生成视频片段$X$的低维流形表示$\boldsymbol{M}$
3. $\boldsymbol{N}\leftarrow\boldsymbol{X}, L_N, ner$;//生成近邻矩阵$\boldsymbol{N}$
4. WHILE $i<maxIter$ DO
5. $[\boldsymbol{h}_{t-T}^u,\boldsymbol{h}_{t-T+1}^u,\cdots,\boldsymbol{h}_t^u]\leftarrow$LSTM($\boldsymbol{N}$);//通过LSTM网络生成近邻时序特征
6. $\boldsymbol{g}_t\leftarrow$pooling($\boldsymbol{h}_{t-T},\boldsymbol{h}_{t-T+1},\cdots,\boldsymbol{h}_t$)$+\boldsymbol{h}_t$;//生成低维时序特征
7. $\boldsymbol{z}_t\leftarrow$Decoder($\boldsymbol{g}_t$);//重构原始帧特征
8. $L_R\leftarrow$Sim($\boldsymbol{x}_t,\boldsymbol{z}_t$);//计算重构损失
9. $L_M\leftarrow$Sim($\boldsymbol{g}_t,\boldsymbol{m}_t$);//计算流形结构损失
10. END WHILE
11. $C\leftarrow$KM($\boldsymbol{G}$);//通过K-Means算法得到时序分割的簇标签C。

4 实验与结果分析

4.1 第一视角视频数据集

本文在Office[46]、Outdoor[28]、GTEA[47]、Huji[48]、ADL[49]、EGTEA[50]以及EPIC-55数据集[51]上评估所提出的SSMS算法，并与现有的无监督时序分割算法、聚类算法、自监督学习算法进行对比。数据集的详细设置如表1所示。

表1 第一视角数据集详情

数据集	视频序号	活动类型数	采样帧数
Office	Office-1	6	8100
	Office-2	6	8100
	Office-4	6	8100
Outdoor	Outdoor-1	6	7500
	Outdoor-3	6	7500
	Outdoor-5	6	7500
GTEA	GTEA-1	7	9100
	GTEA-2	7	8140
	GTEA-3	7	7351
Huji	Huji-1	5	2000
	Huji-2	5	2000
	Huji-3	4	1600
ADL	ADL-1到ADL-20	32	64 857
EGTEA	OP01-R01到P26-R05	106	344 568
EPIC-55	P01-01到P31-14	351	662 258

Office数据集主要收集办公室环境下的活动，每个受试者的视频时长为25分钟～30分钟。每个视频的活动内容相同，包括：阅读、看电影、打字、写作和浏览网页。在每个活动之间设置了30秒的空白间隔。这些间隔主要是为了活动之间的切换，包括谈话，歌唱等一些放松活动，为了减少重复数据的处理以及加快实验验证，本文随机选取了其中三个视频，并以5 fps帧率对每个视频进行平均采样，共获得24 300个视频帧。

Outdoor数据集主要收集室外的活动，包含六个时长为25分钟，帧率为15 fps的第一视角视频。其中，视频的内容包括五项运动以及空白活动，例如，打篮球，踢足球，打乒乓球等。每个运动持续两分钟，运动之间会有30秒的空白活动，作为场景的切换以及休息。为了降低重复数据的计算本文随机选取了其中三个视频，并以5 fps帧率对每个视频帧进行平均采样，共获得22 500个视频帧。

GTEA数据集包含多个室内活动的第一视角视频，共包含七种不同类型的活动，例如做三明治、煮咖啡、泡茶等。每项活动由四个人执行，共记录了四个时长约十分钟，帧率为15 fps的视频。为了降低重复活动的计算成本，本文随机选取了三个人的活动视频，按照15 fps帧率进行平均采样，共获得了

24 591个采样帧。

Huji数据集通过GoPro设备记录了3个人的44段第一视角视频。为了保证数据集的多样性以及减少相似图片的数量，本文以5 fps的帧率对这3个人的4-5种活动视频进行平均采样。主要的活动包括走路、坐、骑行、静止、站立等。每段记录包括2000个采样帧，包含4-5个不同的聚类个数。

ADL数据集包含了20个人在不同的家庭环境中进行无剧本的日常活动，活动时长不固定。视频内容共包含32种不同的活动种类，每个视频内的活动类型与场景各不相同。本文对20个视频按照2 fps帧率进行平均采样，共获得了64 857个视频帧。

EGTEA数据集是一个大规模的第一视角时序分割数据集，共包含了32个受试者的86段，总时长为28小时的厨房烹饪活动的视频。为减少冗余帧的处理以及加快实验验证，本文以5 fps帧率对每个视频进行平均采样，共获得344 568个视频帧。

EPIC KITCHENS-55数据集是一个大型的第一视角视频数据集，包含了日常生活中的厨房活动，总时长为55小时。该数据集的拍摄环境为佩戴者的家庭自由环境，记录了多天内的厨房日常活动。为减少冗余帧的处理以及加快实验验证，本文以5 fps帧率对每个视频进行平均采样，共获得662 258个视频帧。

4.2 基准算法

(1)基础聚类算法

KM(K-Means)：本文选取KM算法作为基础的对比算法，KM算法可直观的对比出SSMS算法的提升。

(2)无监督特征选择算法

AE(Autoencoder clustering)：一个基础的自动编码器对单个视频帧进行特征表达。设置的隐藏特征维度为100，迭代次数为300。选取最佳的聚类结果作为实验结果。

N2D[52]：该算法使用了流形嵌入与自编码学习，获得了高质量的低维特征。在数据集上手动调节了参数设置以保证获得最佳的实验结果。自编码器中编码层与解码层的深度均为2层；流形嵌入的相关参数设置如下：umap算法最小距离*umap_mindist*设置为0.1，umap算法近邻数*umap_neighbors*设置为10，嵌入维度为10；迭代次数为200。

(3)时序聚类算法

TSC(Temporal Subspace Clustering)[53]：TSC算法是一种基于时序信息的子空间聚类算法。在数据集上手动调节所有的参数设置来获取最佳的聚类效果。

(4)自监督学习算法

DC(Deep Clustering)[54]：DC算法采用了生成伪标签方式进行深度无监督学习的算法。本文中DC算法采用了与SSMS算法所使用的特征提取网络(ResNet50)一致的网络结构，设置迭代次数为200。

CPC(Contrastive Predictive Coding)[55]：CPC算法是一种基于时间序列的无监督特征学习算法，该算法的核心是通过自回归模型来预测未来数据的隐变量。编码网络采用的是ResNet50，自回归网络采用的是基于像素的CNN网络。手动调节超参数来获取最佳的聚类效果。

SimCLR[56]：SimCLR是一种视觉表示的自监督学习框架，该算法通过对图像的随机变换来获得成对的正样本。使用预训练在ImageNet数据集上的ResNet50作为特征提取网络，输出特征维度为200，手动调节超参数来获取最佳的聚类效果。

CC(Contrastive Clustering)[57]：CC是基于聚类的自监督学习聚类算法。该算法通过最大化成对的正样例和最小化成对的负样例来实现样本的特征学习。其中，使用预训练在ImageNet数据集上的ResNet50作为特征提取网络，手动调整超参数来获取最佳聚类效果。

(5)无监督时序分割算法

TW-FINCH(Temporally Weighted First NN Clustering Hierarchy)[7]：TW-FINCH算法采用对时间加权的方式划分语义一致的视频帧。算法的输入来自于预训练ResNet50网络输出的视觉特征。

ASAL(Action Shuffle Alternating Learning)[6]：ASAL算法采用了自监督学习的方法，使用隐马尔可夫模型对动作长度进行建模。算法的输入来自预训练ResNet50网络输出的视觉特征，手动调整参数来获取最佳的聚类结果。

ABD(Action Boundary Detection)[10]：ABD是一种基于活动边界检测的时序分割算法。本文使用预训练ResNet50网络的输出特征作为算法的输入，手动调整参数来获取最佳的聚类结果。

TSA(Temporal-Semantic Aware)[8]：TSA算法是一种基于时间-语义相似度加权矩阵的三联体选择算法。本文中使用预训练ResNet50网络的输出特征作为算法的输入，经过多次迭代训练，获得最佳的聚类结果。

UASUEA（Unsupervised Action Segmentation of Untrimmed Egocentric Videos）[31]：UASUEA是一种针对第一视角视频的时序分割算法，该方法使用核变化点来检测估计动作边界。本文中使用预训练ResNet50网络的输出特征作为算法的输入，手动调整参数来获取最佳的聚类结果。

在多种类型的对比算法中，本文中使用预训练ResNet50网络的输出特征作为算法的输入，对未直接输出时序分割结果的算法，采用了KM算法对输出特征进行聚类，以此得到帧级别的划分结果。

4.3 度量指标

视频中的时序分割问题通常采用帧级别的度量指标，本文使用聚类中常用的三种指标来评估各算法的聚类性能。对于这三种指标来说，所得值越大说明聚类效果越好。

（1）聚类准确度（Clustering Accuracy，ACC）：

$$\mathrm{ACC}=\frac{\sum_{i=1}^{n}\delta\left(map(l_i),g_i\right)}{n} \tag{14}$$

其中，n为采样视频帧的总数；l_i和g_i为x_i的聚类划分以及真实标签；$\delta(x,y)$为狄拉克函数，当两个值相等时，该值为1，否则为0；$map(l_i)$为排列映射函数，将聚类结果映射到对应的真实标签上，保证尽可能多的聚类划分与真实标签对应，该过程可由Kuhn-Munkres算法优化。

（2）标准化互信息（Normalized Mutual Information，NMI）：

$$\mathrm{NMI}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{c}n_{ij}\log\left(\frac{n\cdot n_{ij}}{n_i\cdot n_j}\right)}{\sqrt{\left(\sum_{i=1}^{c}n_i\log\left(\frac{n_i}{n}\right)\right)\left(\sum_{j=1}^{c}n_j\log\left(\frac{n_j}{n}\right)\right)}} \tag{15}$$

其中，n_i表示在C_i簇中包含了多个视频帧，n_i代表了在类l_j中包含多少个视频帧，n_{ij}代表了有多少个视频帧既属于簇C_i，又属于类l_j。NMI值越高，聚类算法的表现越好。

（3）调整兰德指数（Adjusted Rand Index，ARI）：

$$\mathrm{ARI}=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_i}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_j}{2}\right]-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]/\binom{n}{2}} \tag{16}$$

其中，n_{ij}表示有多少帧既属于簇C_i，又属于类L_j；a_i表示在簇C_i中有多少视频帧被划分到类L_i中。ARI的取值范围为$[-1,1]$，当该值接近0的时候，聚类结果趋向于随机划分。

（4）F1值（F1 Score，F1）：

$$\mathrm{F1}=\frac{2\mathrm{P}\cdot\mathrm{R}}{(\mathrm{P}+\mathrm{R})} \tag{17}$$

F1是纯度P和召回率R的调和平均值。纯度P的定义如下：

$$\mathrm{P}=\frac{1}{N}\sum_i \underset{j}{max}\left|c_i\cap l_j\right| \tag{18}$$

其中N是视频片段的总数，c_i代表隶属于i-th簇的划分数量，l_j代表隶属于j-th类的划分数量。

召回率R的定义如下：

$$\mathrm{R}_c=\frac{L_{c,t}}{\sum_c L_{c,t}} \tag{19}$$

其中t是聚类c对应的真实类标签，它代表的是c中对应包含出现次数最多的那个类，$L_{c,t}$是c中包含t的视频片段数。

4.4 实验结果

本节分析了七个数据集在不同评价指标下的表现：

（1）Office数据集：如表2中Office数据集所示，SSMS取得了较好的结果。其他算法相对于基础的KM算法也有一定的提升。在办公室场景下，虽然视频中的活动较为丰富，但画面背景相对单一，如看书、写字、打字等活动。由于画面视觉区分度较低，极大增加了活动分割的难度。数据集中包含了活动之间的30秒自由时间。这些间隔既不属于上一个活动，也不属于下一个活动。虽然这些活动出现在不同的时间戳上，但是为了有效地分析人的具体行为活动模式，本文将其视为同一类。这种较长时间的过渡导致了基于边界检测算法无法对其进行有效的划分，这些算法大都将其视为前一活动的结束或者下一活动的开始。基于第一视角视频时序分割的UASUEA算法保持了帧的连续性，相较于其他边界检测算法有一定的优势。而基于帧的时序分割算法能够较好地适应这种情况的出现，如SimCLR算法取得了较好且稳定的结果。在Office数据集中，SSMS算法在ACC、ARI、F1等指标上高于其他算法0.4%、2.7%、0.6%。

（2）Outdoor数据集：如表2中Outdoor数据集所示，SSMS相比于其他算法取得了最佳的结果。N2D算法通过数据流形的嵌入以及自编码网络的学习，也取得了较为不错的结果。在Outdoor上取得了ACC 68.0%、NMI 57.7%、ARI 46.6%以及F1 54.0%。在户外场景下，不同的运动发生在特定的

表2　五个第一视角视频数据集上不同算法的性能对比

方法	Office				Outdoor				GTEA				Huji				ADL			
	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
KM	47.5	48.2	27.5	45.8	58.1	44.1	31.8	43.6	52.6	44.6	34.1	50.6	85.1	73.2	69.8	76.8	51.8	60.7	39.6	48.3
AE	68.7	61.1	46.8	60.6	62.5	48.5	38.8	48.3	61.9	52.2	40.2	51.9	88.6	77.2	75.4	80.5	56.3	62.0	43.7	50.6
DC[54]	49.7	45.6	27.7	45.1	60.8	45.8	35.5	46.4	52.3	43.1	30.1	43.7	79.0	71.9	65.7	43.8	40.2	42.8	24.0	36.6
N2D[52]	56.3	51.7	31.0	47.8	68.0	57.7	46.6	54.0	54.6	58.2	39.9	49.3	73.4	71.0	61.8	75.8	54.1	63.6	40.6	42.9
TSC[53]	53.4	53.5	36.8	49.6	67.9	63.4	53.9	52.1	66.6	66.5	53.1	61.2	81.4	67.0	63.1	74.1	50.8	55.8	39.1	44.1
CPC[55]	57.9	55.4	39.3	50.0	56.9	44.3	34.1	45.4	54.4	47.6	36.0	46.5	83.6	71.8	67.9	75.3	54.1	60.9	42.5	49.9
SimCLR[56]	69.1	60.7	49.6	49.2	66.8	52.0	43.7	44.3	71.5	71.6	61.6	67.9	64.9	42.2	36.8	43.8	50.5	51.3	35.0	45.4
CC[57]	56.6	55.5	40.4	49.6	68.6	53.3	47.3	55.8	58.3	53.8	40.8	48.1	87.8	76.6	74.5	78.9	53.6	57.7	40.3	44.2
TW-FINCH[7]	44.4	41.2	17.3	35.3	58.7	59.9	40.0	53.1	**88.8**	**91.9**	**88.1**	**87.5**	86.5	83.8	76.0	88.6	61.0	**70.4**	51.4	57.5
ASAL[6]	44.2	45.0	25.1	44.3	41.2	40.1	20.9	42.5	87.7	86.1	79.6	77.2	49.1	58.0	44.3	40.5	-	-	-	-
ABD[10]	40.9	38.1	18.2	34.8	66.5	65.1	51.7	61.0	71.6	76.1	59.2	67.4	85.2	82.6	76.7	82.3	58.8	68.2	46.1	54.4
TSA[8]	46.8	36.5	17.1	30.4	44.6	41.7	23.4	32.5	71.7	76.0	62.2	68.5	88.5	81.7	77.0	86.6	51.1	63.8	39.0	50.1
UASUEA[31]	57.9	**65.0**	44.7	48.2	59.6	52.8	41.5	53.0	59.3	57.0	44.6	52.7	72.7	67.2	58.4	69.3	57.9	65.0	44.7	53.2
SSMS	**69.5**	64.8	**52.3**	**61.2**	**80.8**	**69.6**	**63.4**	**67.4**	86.3	81.1	74.5	85.6	**91.6**	**84.9**	**81.9**	**89.1**	**62.7**	68.9	**52.4**	**58.4**

场地上。因此，在户外运动的视频画面上，不同的运动之间具有巨大的差异性。相对于其他自监督学习算法，SSMS算法能够从流形空间上为数据特征生成监督信号，提高算法准确率。SSMS算法在Outdoor数据集上高于其他算法ACC 12.2%、NMI 4.47%、ARI 9.5%、F1 6.4%。

（3）GTEA数据集：如表2中GTEA数据集所示，SSMS相比于其他算法取得了最佳的结果。图3展示了GTEA-2视频中时序分割的具体情况（GND表示视频中真实活动的划分）。虽然大部分活动均正确地分类，但某些活动中出现了分类错误的帧集合。如泡咖啡活动中出现了泡蜂蜜咖啡的错误聚类情况。这种情况源于两者的活动较为接近，泡蜂蜜咖啡中包含了部分泡咖啡的活动。然而，SSMS算法考虑相邻帧的时序关系，并没有增加错误划分的范围。同时，相比于其他算法，SSMS算法更注重时序帧之间的聚合性，也就是说相邻帧尽可能视为同一个活动类别。在GTEA数据集中，基于边界检测的时序分割算法表现出较高的正确率，TW-FINCH算法和ASAL算法均略高于SSMS算法。TW-FINCH算法同时遍历数据集中所有的长视频，而所提出方法在处理该问题时，仅对单个长视频进行时序分割，并没有获取其他视频的数据信息。对于镜头稳定，画面背景单一的GTEA视频来说，基于边界检测的算法有一定的优势。

（4）Huji数据集：如表2中Huji数据集所示，SSMS算法相比于其他算法获得了较高的正确率。Huji数据集相对于其他数据集采样规模小，活动数量少，分类难度大大降低。在Huji数据集中SSMS算法在ACC、NMI、ARI、F1等指标上高于其他算法3%、1%、5.1%、0.5%。

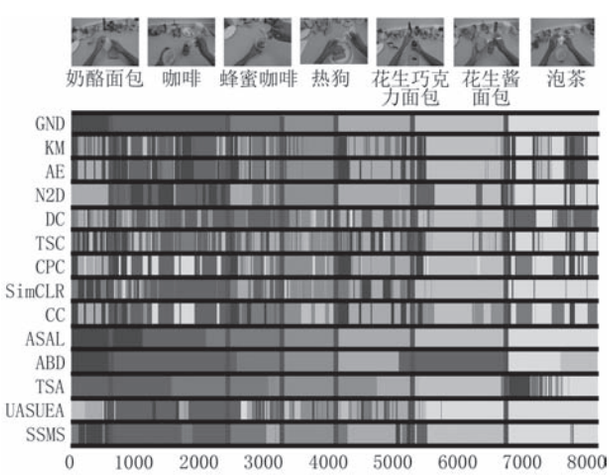

图3　GTEA-2视频中聚类结果的可视化展示

（5）ADL数据集：如表2中ADL数据集所示，SSMS算法相比于其他算法获得了较高的正确率。ADL数据集包含20个视频，活动种类丰富，数据规模大，分类难度大大提升。从整体算法的表现上看，基于边界检测的算法更占优势。

（6）EGTEA数据集：如表3中EGTEA数据集所示，SSMS算法在ACC、ARI、F1指标上均有不错的表现。EGTEA数据集是GTEA数据集的一个扩展，视频内容上更加丰富，活动种类有106类。视频活动均发生在厨房场景下，但EGTEA数据集的规模与活动数量明显高于GTEA。从整体算法的表现上看，SSMS算法略微优于其他算法。

表3 时序分割算法在大规模数据集上的表现

方法	EGTEA				EPIC			
	ACC	NMI	ARI	F1	ACC	NMI	ARI	F1
TW-TINCH[7]	43.3	64.0	31.8	39.0	43.0	43.4	23.6	37.3
ABD[10]	43.7	**65.2**	32.6	39.9	44.5	42.5	21.7	39.7
TSA[8]	35.2	61.1	24.6	30.4	43.7	43.2	23.2	37.0
UASUEA[31]	34.8	50.6	10.2	22.9	45.9	35.2	16.7	39.2
SSMS	**44.0**	64.8	**38.1**	**43.4**	**50.0**	**43.4**	**29.2**	**43.6**

(7) EPIC KITCHENS-55数据集：如表3中EPIC数据集所示，SSMS算法略高于其他算法。EPIC KITCHENS-55数据集中记录的视频活动更自由，数据量更大。在两个大规模数据集下的实验可以看出，SSMS在指标ACC、ARI、F1上均取得了最佳结果，而在NMI上略低于其他算法0.4%。原因在于这些视频内包含了大量的琐碎动作，区分难度较大。SSMS聚类算法在将数据分到正确的簇方面表现良好，而在区分类别间的差异方面需要进一步增强。

如图4所示，本文选取Office-1视频和Outdoor-1视频在不同的训练阶段(0，1，50，100)获得高质量数据特征，并使用T-SNE算法对其进行可视化展示并计算NMI值。如图4所示，在训练的开始阶段，NMI会先下降，然后缓慢增长。在获得的可视化图中，也能看到数据分布呈现出先分散后集中的特点。这是因为预训练Resnet50网络得到的数据特征本身包含大量的视觉特征。随着训练的进行，SSMS算法融入了时序特征，同时在自监督学习的过程中，辅助任务又考虑了流形结构信息。训练阶段包含了三者之间的信息融合过程，数据分布的可视化呈现出先分散后集中的特点。

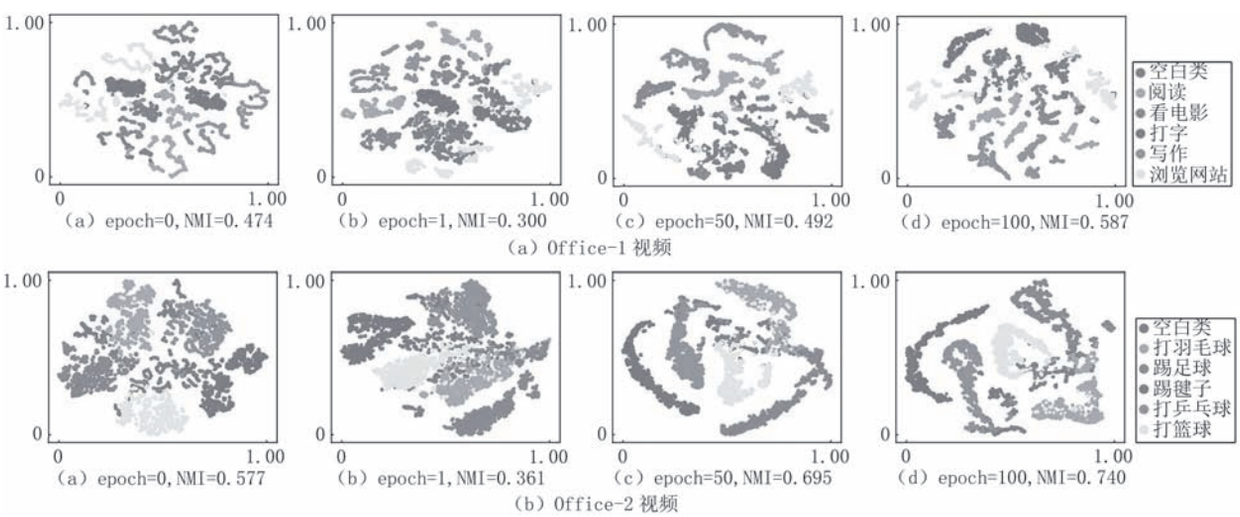

图4 训练过程中Office-1以及Outdoor-1视频数据特征的可视化展示

4.5 参数分析

4.5.1 近邻数量对SSMS算法的影响

本实验通过验证四个数据集中的首个视频，使用固定的时域大小(16)进行检验。实验中设定近邻数量的范围为2、4、8和16，以观察SSMS算法在时序分割上的准确性。如图5所示，近邻数量的变化会直接影响算法的准确性。值得注意的是，并非近邻数量越多，算法准确性就越高。较多冗余的近邻信息在某些场景下会造成一定的干扰。近邻数量的增加会影响算法的准确性，但增长到一定值时(8)，会出现下降的情况。因此，固定时域大小为16，设置近邻数量为8时，当前算法达到最优。

4.5.2 时域大小对SSMS算法的影响

本实验以五个数据集中的第一个视频作为验证对象，固定近邻数量等于时域大小，以降低近邻数量的变化对实验结果的影响。随后，设置时域大小的取值范围为2、4、8和16，观察SSMS算法在时序分割上的准确度。如图6所示，算法准确性会随着时域大小而缓慢上升，但上升的幅度较小。时域越大，算法的时序视野范围扩大，因此算法的准确性提升。在Office-1视频中，算法在时域大小为16的情况下低于时域大小为8的准确度。这种现象表明在不考虑其他因素的情况下，一味地增大时域视野范围，固定近邻数量不变，过多的冗余信息会对算法产生负面影响，导致准确度下降。因此，在大多数视频中，随着时域大小的增加，算法准确度呈平稳上升的趋势。当固定近邻数量与时域大小相等的情况，时域大小设置为16较为合适。

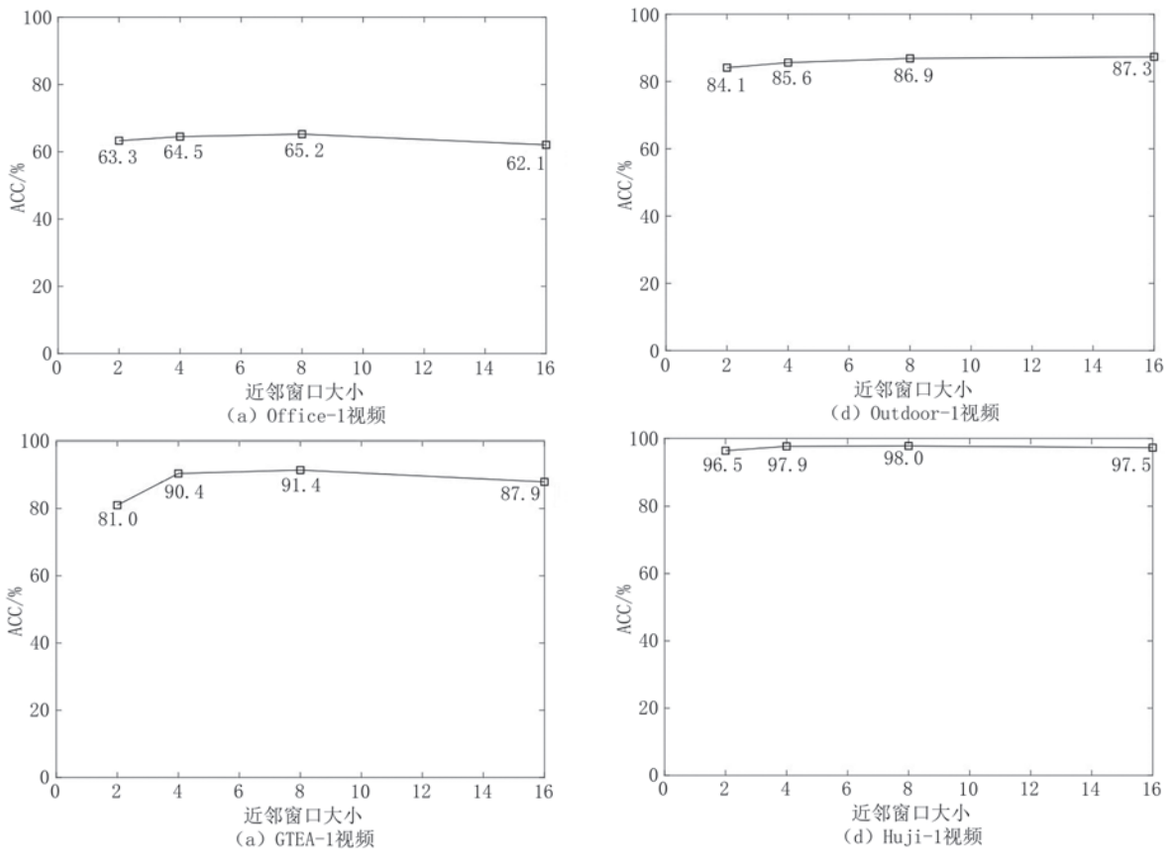

图5　近邻窗口大小对SSMS算法的影响

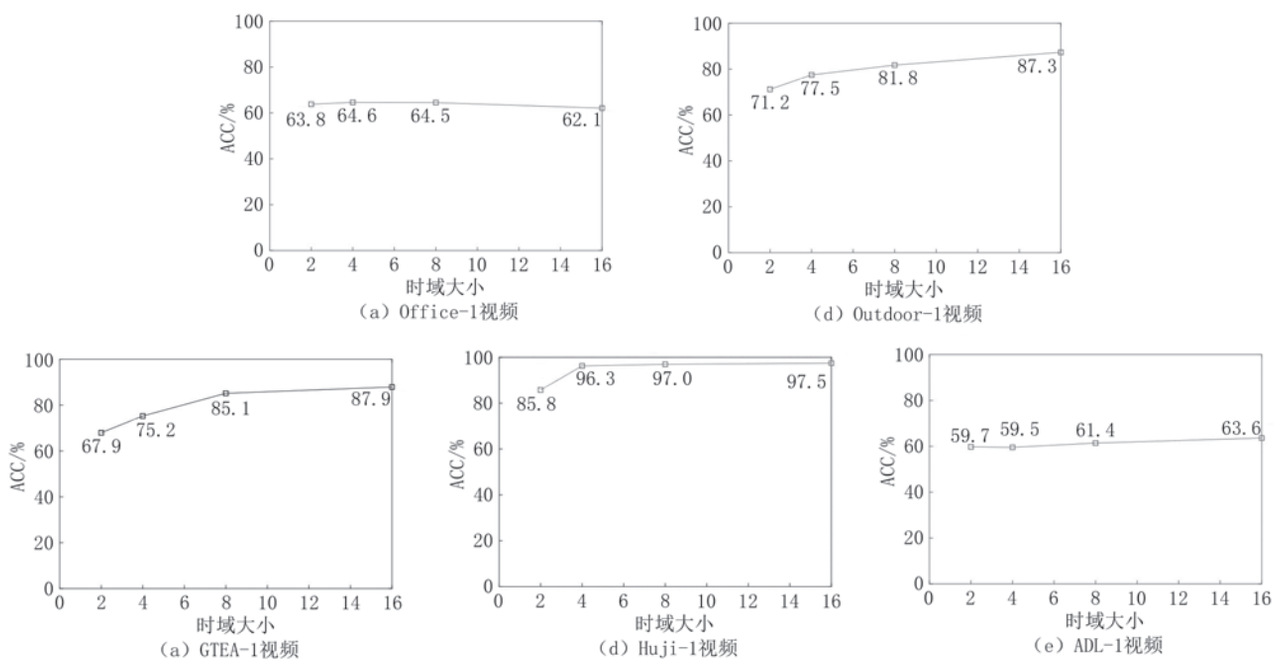

图6　时域大小对SSMS算法的影响

4.6　算法运行时间与存储效率

表4显示了在单个GPU上不同算法的运行时间与存储效率结果。实验包括FLOPs(浮点运算数)、参数量和算法每次迭代所需的时间等评价指标。SSMS算法取得了与其他方法相近的FLOPs值。该值越小，说明计算速度越快。SSMS算法直接对预训练模型提取的特征进行训练，而其他自监督学习算法仍需要训练ResNet50网络。因此

SSMS算法在训练过程中具有较小的参数量，对算法的计算资源需求较低。在单轮训练过程中，SSMS算法以较低的时间优于其他算法，具有较小的计算开销。不同算法的迭代次数对总时间花费也有影响，本节对比了算法的总时间花费。由表4可知，SSMS算法在总时间花费上也具有较小的值。

表4 算法运行时间与存储效率对比

方法	浮点运算数/G	参数量(M)	单轮迭代时间花费/s					总时间花费/min				
			Office-1	Outdoor-1	GTEA-1	Huji-1	ADL-1	Office-1	Outdoor-1	GTEA-1	Huji-1	ADL-1
DC[54]	4.11	24	64.1	58.4	76.5	10.89	15.5	283.3	281.2	360.6	76.2	134.2
CPC[55]	**3.83**	24	64.4	60.4	72.2	24.86	43.37	191.1	183.4	738.9	48.5	91.8
SimCLR[56]	4.11	24	101.4	93.4	113.8	18.67	30.51	317.3	293.1	237.1	79.6	99.5
CC[57]	8.24	32	87.9	81.3	100.7	16.88	27.55	248.6	305.2	287.7	59.6	104.5
SSMS	5	**2.5**	**20.4**	**19**	**22.7**	**2.24**	**3.86**	**48.8**	**44.5**	**55.4**	**16.9**	**24.6**

4.7 消融实验

本节评估了流形结构以及自监督学习算法对SSMS算法的影响，选取了五个数据集中的五个视频。以下算法代表了在不同条件下SSMS算法的消融实验：

(1)w/o TL(Temporal Learning)：该算法探究了缺少动态时序网络的情况下SSMS算法的性能。该算法将动态时序网络替换成简单的线性变化，其余算法过程与SSMS算法保持一致。

(2)w/o NL(Neighbour Learning)：该算法探究了不使用最大近邻相似矩阵情况下SSMS算法的性能。在SSMS算法中设置最大相似近邻数*ner*与近邻时域长度L_N相等，其余算法过程与SSMS算法保持一致。

(3)w/o MS(Manifold Structure)：该算法探究了流形结构对SSMS算法的影响。在SSMS算法中使用PCA算法降维后的特征作为监督信息，其余算法过程与SSMS算法保持一致。

(4)w/o SSL(Self-Supervised Learning)：该算法探究了SSMS算法中自监督学习的重要性。在SSMS算法中仅保留动态时序特征学习过程。

(5)MS(Manifold Structure)：该算法探究了直接使用流形结构特征进行时序分割任务的效果。将滑动窗口所获得的流形结构特征直接输入到KM聚类算法中，获得时序分割结果。

由表5观察发现，不使用流形学习的算法(w/o MS)相比于不使用时间关系算法(w/o TL)效果更好。这表明在第一视角视频数据中时序关系相较于流形特征更为关键，仅使用流形学习策略难以有效地描述第一视角视频数据的特征和规律。不使用最大近邻相似矩阵的算法(w/o NL)考虑了时序特征以及流形特征，但缺少了对近邻数量的约束，导致邻域内冗余信息增多，易出现过度分割的情况。SSMS算法略高于w/o NL算法，但在Outdoor-1视频数据中，w/o NL算法高于SSMS算法ACC 0.9%、NMI 1.9%、ARI 1.6%。这种情况可能发生在剧烈运动过程中视野不在观察物体上的某一时间段内，当近邻长度(相当于算法输入时的感受野)小于此现象的时间长度时，就会出现近邻矩阵效果不明显的情况。由于运动发生在同一场景下，这种情况很容易与静息状态时捕捉到的画面相似，导致误划分的结果。在其他相对稳定的数据集中，如ADL-1视频，最大近邻相似矩阵的引入使得算法提升了ACC 5.5%、NMI 5.1%、ARI 2.7%。在大多数情况下，使用最大近邻矩阵仍然能够给算法带来一定的提升。不使用自监督学习的算法(w/o SSL)要明显低于SSMS算法，该现象验证了自监督学习的有效性。通过多组消融实验，验证了流形结构的嵌入过程以及自监督学习方法在整个SSMS算法流程中发挥了至关重要的作用，二者相辅相成，缺一不可。最后，直接使用流形结构特征的算法(MS)效果很差。原因在于SSMS算法中所用到的监督信息

表5 SSMS算法的消融实验

方法	Office-1			Outdoor-1			GTEA-1			Huji-1			ADL-1		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MAE度量	62.4	54.3	41.8	86.2	74.7	70.9	78.0	78.5	71.8	**98.8**	**96.3**	**96.9**	**62.6**	**74.0**	57.9
MSE度量	57.0	49.5	35.0	85.5	73.7	69.6	84.4	77.9	69.6	98.1	95.2	95.4	62.5	71.7	**58.3**
本文的度量	**65.2**	**59.8**	**46.2**	**86.4**	**74.9**	**71.2**	**91.4**	**88.1**	**83.7**	98.0	94.5	95.0	61.6	71.9	55.9

包含了流形结构信息，动态时序信息以及自身特征的信息。单一的流形结构信息仅仅代表了滑动窗口内近邻数据之间的流形嵌入关系，并不能完全表示视频帧的语义特征。

4.8　不同度量对SSMS算法的影响

本节探究不同的度量对SSMS算法的影响，包括绝对误差损失（Mean Absolute Error，MAE），均方差损失（Mean Square Error，MSE）以及本文所选的交叉熵度量（Cross Entropy）。

由表6可知，在多数数据集中交叉熵度量表现最佳。交叉熵衡量了两个概率分布之间的差异，这种差异可以用来指导模型学习如何重构输入数据。交叉熵度量对网络结构、数据分布和梯度稳定性都具有更好的适应性。

表6　SSMS算法中不同度量函数的影响

方法	Office-1			Outdoor-1			GTEA-1			Huji-1			ADL-1		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
w/o TL	52.9	45.1	28.1	72.3	56.9	51.2	71.5	62.6	55.0	83.9	75.5	71.3	56.1	66.8	53.2
w/o NL	62.1	55.1	42.6	**87.3**	**76.8**	**72.8**	87.9	**88.5**	78.8	97.5	93.7	93.9	56.1	66.8	53.2
w/o MS	63.5	55.5	42.2	82.1	68.7	63.2	74.7	73.1	60.9	92.4	81.7	82.3	59.5	68.3	53.2
w/o SSL	64.0	57.9	44.2	82.1	69.4	64.2	84.0	80.5	77.1	84.1	76.9	71.5	52.3	57.5	47.0
MS	28.9	12.6	5.8	36.4	17.5	9.1	27.6	13.8	5.5	59.7	57.7	38.2	51.9	69.0	42.4
SSMS	**65.2**	**59.8**	**46.2**	86.4	74.9	71.2	**91.4**	88.1	**83.7**	**98.0**	**94.5**	**95.0**	**61.6**	**71.9**	**55.9**

注：w/o 表示“without”

4.9　不同流形对算法的影响

本节验证了不同流形对SSMS算法的影响，包括双曲流形、庞加莱流形、洛伦兹流形以及黎曼流形。由表7可以观察到，采用不同的流形对算法结果产生了一定的影响。值得注意的是，本文所采用的黎曼流形结构在效果上优于其他流形。这一优越性源于黎曼流形结构能够有效地建模时空关系，强化了局部的时序关系，减弱了因相机运动带来的画面模糊问题。

表7　不同流形对算法的影响

方法	Office-1			Outdoor-1			GTEA-1			Huji-1			ADL-1		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
双曲流形	60.8	52.1	32.7	84.7	73.3	68.6	75.4	75.8	65.1	95.4	90.3	89.0	60.0	68.5	53.6
庞加莱流形	**65.5**	56.4	38.0	77.5	73.6	67.0	73.4	68.8	63.2	95.6	90.8	89.6	61.4	69.6	56.4
洛伦兹流形	64.7	55.8	36.7	85.2	73.3	69.3	70.3	60.7	50.1	95.3	90.5	88.9	61.0	65.5	**58.4**
黎曼流形	65.2	**59.8**	**46.2**	**86.4**	**74.9**	**71.2**	**91.4**	**88.1**	**83.7**	**98.0**	**94.5**	**95.0**	**61.6**	**71.9**	55.9

5　局限性分析

本文提出的SSMS算法在第一视角时序分割任务上取得了不错效果的同时，本节也分析了所提模型存在的瓶颈：

（1）本文提出的SSMS算法对相同场景下不同活动的视频易出现误划分的情况。如图7所示，在SSMS算法并没有发现Office-1视频在开始阶段的空白类，甚至忽略了“看电影”的活动，KM算法的实验结果与SSMS算法类似。这种错误的划分来源于视频内环境的高度相似性，极大增加了时序分割任务的难度。如图8可以发现，“阅读”跟“看电影”周围所处的环境几乎相同，唯一不同的是实验者所注视的区域有所不同。在这种场景下的日常活动，时序分割算法很难有效地区分不同的活动。

图7　SSMS在Office-1视频上的表现

图8　Office-1视频中阅读和看电影的对比

（2）时序分割模型的泛化能力不足，本文仅使用自编码网络对特征进行自监督学习，后续可增加对抗学习来提升模型的泛化性。

6 结 论

本文提出了一种自监督流形结构的第一视角视频时序分割算法，命名为SSMS算法。该算法通过自监督学习方法同时考虑第一视角视频中的流形特征信息和时序信息，从而获取更为合理的特征表示。首先，提出了一种滑动窗口的方法来获取视频帧的局部流形信息。其次，构建了一种动态时序网络来获取最大化的近邻时序信息。然后，设计了自监督学习的辅助任务，将数据的流形结构和近邻时序信息统一融合在同一目标函数中。最后，采用K-Means算法对获得的低维特征进行聚类分析。

相比于目前的自监督学习算法，本文所提出的算法将局部流形结构信息作为自监督学习的监督信号，能够显著提高时序分割任务的正确率。在七个第一视角视频数据集上，SSMS算法相较于其他算法表现出更好的结果。这一研究的贡献在于兼顾视频数据中的流形特征和时序信息，通过自监督学习实现更有效的特征学习，为第一视角视频时序分割领域提供了一种有力的解决方案。

在未来的研究工作中，我们考虑第一视角中其他传感器信息融入时序分割算法研究中，如眼动，重力加速度等信息，以拓展算法对周围环境的感知。通过整合多源传感器信息，有望使算法获得更全面、更多样的信息，从而提高算法的准确性和适用性。此外，在元宇宙的背景下，第一视角数据不仅来源于日常生活记录，未来还将涌现在元宇宙等虚拟交互环境中。这样的大规模数据涌现为算法性能带来了巨大的挑战。我们将重点关注在这一背景下的研究，致力于应对由于数据规模庞大而引发的问题，探索更加高效和智能的算法设计。这包括但不限于数据处理、特征学习、模型优化等方面的研究，以满足未来元宇宙环境下第一视角视频理解的需求。

参 考 文 献

[1] Alejandro Betancourt, Pietro Morerio, Carlo S. Regazzoni, Matthias Rauterberg. The evolution of first person vision methods: A survey. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(5):744-760

[2] Yichen Ding, Ziming Zhang, Yanhua Li, et al. EgoSpeed-net: Forecasting speed-control in driver behavior from egocentric video data//Proceedings of the International Conference on Advances in Geographic Information Systems. Seattle, USA, 2022: 12:1-12:10

[3] Shen Qing, Ban Xiao-Juan, Chang Zheng, et al. On-line detection and temporal segmentation of actions in video based human computer interaction. Chinese Journal of Computers, 2015, 38(12):2477-2487 (in chinese)
(沈晴，班晓娟，常征等.基于视频的人机交互中动作在线发现与时域分割.计算机学报，2015，38(12)：2477-2487)

[4] Xiong Cheng-Xin, Guo Dan, Liu Xue-Liang. Temporal proposal optimization for temporal action detection. Journal of Image and Graphics, 2020, 25(7):1447-1458 (in chinese)
(熊成鑫，郭丹，刘学亮.时域候选优化的时序动作检测.中国图象图形学，2020，25(7)：1447-1458)

[5] Colin Lea, Michael D. Flynn, René Vidal, et al. Temporal convolutional networks for action segmentation and detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1003-1012

[6] Jun Li, Sinisa Todorovic. Action shuffle alternating learning for unsupervised action segmentation //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 2021: 12628-12636

[7] M. Saquib Sarfraz, Naila Murray, Vivek Sharma, et al. Temporally-weighted hierarchical clustering for unsupervised action segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 11225-11234

[8] Elena Belén Bueno-Benito, Biel Tura Vecino, Mariella Dimiccoli. Leveraging triplet loss for unsupervised action segmentation//IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver, Canada, 2023: 4922-4930

[9] Li Ding, Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA, 2018: 6508-6516

[10] Zexing Du, Xue Wang, Guoqing Zhou, et al. Fast and unsupervised action boundary detection for action segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual, 2022: 3313-3322

[11] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, et al. Alleviating over-segmentation errors by detecting action boundaries//IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2021: 2321-2330

[12] Zhenzhi Wang, Ziteng Gao, Limin Wang, et al. Boundary-aware cascade networks for temporal action segmentation//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 34-51

[13] Yuerong Li, Zhengrong Xue, Huazhe Xu. Unsupervised boundary detection for object-centric temporal action segmentation//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2023: 6423-6432

[14] Meng Yang, Pengfei Zhu, Luc Van Gool, et al. Face recognition based on regularized nearest points between image sets//Proceedings of the IEEE International Conference and

Workshops on Automatic Face and Gesture Recognition. Shanghai, China,2013: 1-7

[15] Enrique G. Ortiz, Alan Wright, Mubarak Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2013: 3531-3538

[16] Xiaofei He, Shuicheng Yan, Yuxiao Hu, et al. Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3): 328-340

[17] Leland McInnes, John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018

[18] Adrián Núñez-Marcos, Gorka Azkune, Ignacio Arganda-Carreras. Egocentric vision-based action recognition: A survey. Neurocomputing, 2022, 472:175-197

[19] Yan Yan, Elisa Ricci, Gaowen Liu, et al. Egocentric daily activity recognition via multitask clustering. IEEE Transactions on Image Processing, 2015, 24(10):2984-2995

[20] Haoxin Li, Wei-Shi Zheng, Jianguo Zhang, et al. Egocentric action recognition by automatic relation modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1):489-507

[21] Antonino Furnari, Giovanni Maria Farinella. Rolling-unrolling LSTMs for action anticipation from first-person video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(11): 4021-4036

[22] Esteve Valls Mascaro, AhnHyemin, LeeDongheui. Intention-conditioned long-term human egocentric action anticipation// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.2023:6037-6046

[23] V. Javier Traver, Dima Damen. Egocentric video summarisation via purpose-oriented frame scoring and selection. Expert Systems with Applications, 2022, 189:116079

[24] Abhimanyu Sahu, Ananda S. Chowdhury. Egocentric video co-summarization using transfer learning and refined random walk on a constrained graph. Pattern Recognition, 2023, 134:109128

[25] Minghuang Ma, Haoqi Fan, Kris M. Kitani. Going deeper into first-person activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1894-1903

[26] Heng Wang, Alexander Kläser, Cordelia Schmid, et al. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 2013, 103(1):60-79

[27] Shiro Kumano, Kazuhiro Otsuka, Ryo Ishii, et al. Collective first-person vision for automatic gaze analysis in multiparty conversations. IEEE Transactions on Multimedia, 2017, 19(1): 107-122

[28] Mingming Zhang, Xiaoqiang Yan, Shizhe Hu, Yangdong Ye. An information maximization multi-task clustering method for egocentric temporal segmentation. Applied Soft Computing, 2020, 94:106425

[29] Sibo Song, Vijay Chandrasekhar, Bappaditya Mandal, et al. Multimodal multi-stream deep learning for egocentric activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA, 2016: 378-385

[30] Olga Zatsarynna, Yazan Abu Farha, Juergen Gall. Multi-modal temporal convolutional network for anticipating actions in egocentric videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Tennessee, USA ,2021: 2249-2258

[31] Sam Perochon, Laurent Oudre. Unsupervised action segmentation of untrimmed egocentric videos//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island, Greece, 2023: 1-5

[32] Peng Lei, Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA,2018: 6742-6751

[33] Yazan Abu Farha, Jürgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Angeles, USA, 2019: 3575-3584

[34] Yifei Huang, Yusuke Sugano, Yoichi Sato. Improving action segmentation via graph-based temporal reasoning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA,2020: 14021-14031

[35] Hilde Kuehne, Alexander Richard, Juergen Gall. Weakly supervised learning of actions from transcripts. Computer Vision and Image Understanding, 2017, 163:78-89

[36] Fadime Sener, Angela Yao. Unsupervised learning and segmentation of complex activities from video//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, United States, 2018: 8368-8376

[37] Dong Wang, Di Hu, Xingjian Li, et al. Temporal relational modeling with self-supervision for action segmentation// Proceedings of the AAAI Conference on Artificial Intelligence. 2021: 2729-2737

[38] Björn Krüger, Anna Vögele, Tobias Willig, et al. Efficient unsupervised temporal segmentation of motion data. IEEE Transactions on Multimedia, 2016, 19(4):797-812

[39] Sam T. Roweis, Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290 (5500):2323-2326

[40] David Donoho, Carrie Grimes. Hessian eigenmaps: New tools for nonlinear dimensionality reduction//Proceedings of the National Academy of Sciences of the United States of America. USA, 2003, 100:5591-5596

[41] Zhenyue Zhang, Jing Wang. MLLE: Modified locally linear embedding using multiple weights//Proceedings of the Conference on Neural Information Processing Systems. Denver, USA, 2006: 1593-1600

[42] Joshua B. Tenenbaum, Vin De Silva, John C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500):2319-2323

[43] Laurens van der Maaten, HintonGeoffrey. Visualizing data

using t-sne. Journal of Machine Learning Research, 2008, 9: 2579-2605

[44] Chen Xing-Sheng, Chen Wei-Heng. Lectures on differential geometry.Beijing: Peking University Press, 1983 (in chinese)
(陈省身,陈维桓.微分几何讲义.北京:北京大学出版社,1983)

[45] David I. Spivak Metric realization of fuzzy simplicial sets. Self published notes, 2012

[46] Keisuke Ogaki, Kris Makoto Kitani, SuganoYusuke, et al. Coupling eye-motion and ego-motion features for first-person activity recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2012: 1-7

[47] Alireza Fathi, Ali Farhadi, James M. Rehg. Understanding egocentric activities//Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 407-414

[48] Yair Poleg, Ariel Ephrat, Shmuel Peleg, et al. Compact CNN for indexing egocentric videos//Winter Conference on Applications of Computer Vision. Waikoloa, America, 2016: 1-9

[49] Hamed Pirsiavash, Deva Ramanan. Detecting activities of daily living in first-person camera views//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA,2012: 2847-2854

[50] Yin Li, Miao Liu, James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video//Proceedings of the European Conference on Computer Vision, Munich, Germany: volume 11209. 2018: 639-655

[51] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. The EPIC-KITCHENS dataset: Collection, challenges and baselines. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(11):4125-4141

[52] Ryan McConville, Raúl Santos-Rodríguez, Robert J. Piechocki, et al. N2d: (not too) Deep clustering via clustering the local manifold of an autoencoded embedding//Proceedings of the International Conference on Pattern Recognition. Rome, Italy, 2021: 5145-5152

[53] Sheng Li, Kang Li, Yun Fu. Temporal subspace clustering for human motion segmentation//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4453-4461

[54] Mathilde Caron, Piotr Bojanowski, Armand Joulin, et al. Deep clustering for unsupervised learning of visual features// Proceedings of the European Conference on Computer Vision. Munich, Germany,2018:139-156

[55] Olivier J. Hénaff. Data-efficient image recognition with contrastive predictive coding//Proceedings of the International Conference on Machine Learning. Virtual, 2020: 4182-4192

[56] Ting Chen, KornblithSimon, NorouziMohammad, et al. A simple framework for contrastive learning of visual representations//Proceedings of the International Conference on Machine Learning. Austria, Vienna,2020: 1597-1607

[57] Yunfan Li, Peng Hu, Jerry Zitao Liu, et al. Contrastive clustering//Proceedings of the AAAI Conference on Artificial Intelligence, 2021(10):8547-8555

ZHANG Ming-Ming, Ph. D. His main research interests include machine learning and computer vision.

YAN Xiao-Qiang, Ph. D., associate professor. His main research interests include machine learning and computer vision.

SUN Zhong-Chuan, Ph. D. His main research interests include deep learning and recommender systems.

HU Shi-Zhe, Ph. D. His main research interests include pattern recognition and information theory.

YE Yang-Dong, Ph. D., professor. His main research interests include machine learning, knowledge engineering and intelligent system.

Background

Recently, with the new generation of artificial intelligence shifting from perceptual to cognitive intelligence, First Person View (FPV) video analysis has also entered the development stage of intelligence. FPV videos record the scene being interacted with from the viewpoint of the wearer/observer, which is closer to human perception and reflects the true behavioral intent. Without annotated data, temporal segmentation is a fundamental and challenging task in the FPV video analysis field. The existing FPV unsupervised temporal segmentation still has the problem of feature representation.

Previous unsupervised temporal segmentation methods fall into two groups: (1) based on detecting the boundaries and (2) in conjunction with some clustering methods based on frame level. However, there is not much work that combines temporal

information and manifold features. Temporal information acquisition is achieved by constructing a pretext task in self-supervised learning, while manifold features are acquired through a local manifold embedding algorithm. Temporal information is an important clue to further improving the performance of the algorithm. In addition, manifold learning is also a powerful approach to feature representation. In the end, this paper is based on manifold learning and self-supervised learning, which has achieved good performance. Firstly, a sliding window approach is proposed to obtain the local manifold information of the frame level. Second, a dynamic temporal auto-encoder network is constructed to obtain the maximized nearest neighbor temporal information. Then, a pretext task in self-supervised learning is designed to fuse the manifold information of data as well as the nearest neighbor temporal information in the same objective function. Finally, the obtained high-level features are clustered and analyzed using the K-Means algorithm.

This paper was supported by the National Natural Science Foundation of China(Grant No. 62176239)