研究与开发

面向机器视觉的自监督视频时域重采样方法

刘建然[1,2]，纪雯[1,3,4]，付哲[5]
（1. 中国科学院计算技术研究所处理器芯片全国重点实验室，北京 100190；
2. 中国科学院大学，北京 101408；
3. 中科工业人工智能研究院，江苏 南京 211135；
4. 龙眼国科（北京）智能信息技术有限公司，北京 100010；
5. 交控科技股份有限公司，北京 100071）

摘　要：针对视频时域重采样中帧间内容非线性变化导致的语义冗余问题，提出一种基于自监督特征嵌入和聚类的视频时域自适应重采样方法。所提方法通过预训练的ResNet-18提取视频帧特征并微调，利用自监督度量学习构建帧间相似性度量，采用余弦相似度衡量相邻帧相似性，并设计损失函数使同一视频序列内的帧在嵌入空间中流形分布光滑，同时抑制不同视频帧间相似度。之后，对嵌入后的帧特征进行基于流形等分点的时序数据聚类，并确保视频首尾完整。重采样后的视频序列经H.266/VVC编码，解码端结合插帧网络重建原始帧。实验表明，该方法在BDmAP和Pareto mAP指标上平均提高约2.3%和19.4%，且计算开销满足实时处理需求，有效平衡压缩效率、视觉任务精度与零样本兼容性，为机器视觉场景下的视频传输提供新思路。

关键词：机器视觉；时域重采样；视频特征嵌入；自监督聚类；视频压缩
中图分类号：TP393；TN911.73
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2026048

Self-supervised video temporal resampling method for machine vision

Liu Jianran[1,2], Ji Wen[1,3,4], Fu Zhe[5]
1. State Key Lab of Processors, Institute of Computing Technology, CAS，Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 101408, China
3. Institute of AI for Industries, Chinese Academy of Sciences, Nanjing 211135, China
4. LonganPi (Beijing) Intelligent Information Technology Co., Ltd., Beijing 100010, China
5. Traffic Control Technology Co., Ltd., Beijing 100071, China

Abstract: To address the semantic redundancy caused by nonlinear inter-frame content changes during video tempo-

收稿日期：2025−06−29；**修回日期：**2025−08−30
通信作者：纪雯，jiwen@ict.ac.cn
基金项目：国家重点研发计划项目（No.2024YFE0212000）；中国博士后科学基金资助项目（No.GZC20251091）；北京市自然科学基金资助项目（No.L221004）；江苏省工业人工智能重点实验室资助项目
Foundation Items: The National Key Research and Development Program of China (No.2024YFE0212000) , China Postdoctoral Science Foundation (No.GZC20251091), Beijing Municipal Natural Science Foundation (No.L221004), Jiangsu Provincial Key Laboratory of Industrial Artificial Intelligence

ral resampling, an adaptive resampling method based on self-supervised feature embedding and clustering was proposed. In this method, features from video frames were extracted using a pre-trained ResNet-18, which was subsequently fine-tuned. Self-supervised metric learning was employed to construct an inter-frame similarity measure, and cosine similarity was used to gauge the resemblance between adjacent frames. A loss function was designed to ensure a smooth manifold distribution for frames from the same video sequence in the embedding space, while similarity between frames from different videos was simultaneously suppressed. Subsequently, the embedded frame features were subjected to temporal data clustering based on manifold bisection points, and the integrity of the video's first and last frames was ensured. The sampled video sequence was then encoded using H.266/VVC, and the original frames were reconstructed at the decoding end by a frame interpolation network. Through experiments, it was demonstrated that average improvements of approximately 2.3% and 19.4% were achieved on the BDmAP and Pareto mAP metrics, respectively. Furthermore, the computational overhead was found to meet the demands of real-time processing. The proposed approach effectively balances compression efficiency, visual task accuracy, and zero-shot compatibility. It offers a novel solution for video transmission in machine vision scenarios.

Key words: machine vision, temporal resampling, video feature embedding, self-supervised clustering, video compression

0 引言

近年来，自动驾驶[1]、智慧城市[2]与实时监控[3]等应用的蓬勃发展，对大规模视频数据进行高效的机器视觉分析已成为一项关键需求。传统面向人眼的图像/视频编码标准主要面向人类视觉系统，旨在通过消除感知冗余来提升压缩率。然而，此类标准对于机器视觉任务并非最优，因为机器更关注图像中的高级语义信息而非像素级的视觉保真度。因此，在处理与目标相关任务时，传统编码方案常会在与任务无关的背景纹理上耗费大量码率[4]。动态图像专家组（Moving Picture Experts Group，MPEG）发起的机器视频编码（video coding for machine，VCM）项目也指出，为了在维持任务精度的同时显著降低码率，编码器有必要在早期处理阶段跳过非关键的视觉元素，例如，通过均匀时域重采样降低帧率[5-6]。

视频数据固有地包含大量时间冗余。例如，在基于视频的自监督预训练中，现有工作普遍采用步幅采样策略以大幅减少输入帧数，从而降低训练开销并提升效率[7]。而固定时序间隔均匀采样策略在实践中面临显著挑战。其一，当视频总帧数无法被采样间隔整除时，该策略会导致帧丢失或冗余。其二，对于内容动态变化（如包含静止与剧烈运动）的视频，采用统一的采样率会导致在静态场景中采样过剩，而在动态场景中遗漏关键信息。因此，设计一种能根据视频内容变化率快速、自适应地选取关键帧的重采样方案至关重要。与传统面向人眼的视频摘要方法不同，本文强调的是面向机器视觉任务性能的压缩策略，核心动机在于实现码率与下游任务精度的联合最优化。人眼导向方法与机器视觉导向的时域重采样方法对比见表1，人眼导向方法关注感知质量与视觉美学，而本文方法则更注重保留对任务有用的语义信息帧，跳过冗余的背景内容，从而提高检测等任务的效率与鲁棒性。

时域重采样技术常根据具体应用场景进行优化，其核心目标是在压缩视频体积的同时，生成一个能够代表原始视频核心内容的紧凑摘要，已广泛应用于视频摘要[8]和监控视频压缩[9]等领域。在机器视觉场景下，重采样效果则通过下游任务的性能指标来评估。当前，主流方法多为基于目标或语义的监督式时域重采样。例如，Tian等[10]提出了一种针对多场景视频进行压力检测的关键帧选择方法，通过特定算法从复杂的视频序列中筛选出最具代表性的关键帧。Zeng等[11]的研究则

表1　人眼导向方法与机器视觉导向的时域重采样方法对比

对比维度	人眼导向方法（视频摘要等）	机器视觉导向方法（本文）
驱动目标	主观观看体验，美学结构	下游视觉任务性能
优化指标	视频时长，人眼打分，PSNR/SSIM，…	任务精度与码率联合指标
帧选择依据	剧情变化，画面重复度，视觉风格	帧间语义特征流形的突变位置
压缩关注点	消除视觉冗余	跳过对任务精度影响不大的帧
方法特征	多基于视觉显著性或故事结构	强调编码效率与任务精度间的权衡
应用场景	视频编辑/展示	自动驾驶，监控，机器人感知

聚焦于敏感视频筛选检测领域，利用图像增强技术提升视频帧的视觉质量，再结合语义分析，精准定位到可能包含敏感内容的关键帧。Lee 等[12]提出了一种基于自适应帧控制的方法，通过多源预取器和准入控制器动态选择最新的帧进行目标检测。此类重采样方法常依赖下游任务标签，需要通过目标检测等先验获取任务相关帧。相比之下，本文提出的自监督方法完全摆脱了任务标签依赖，以帧间语义特征的结构变化为依据实现自适应采样。这种方式不仅提升了训练效率，更增强了对未知类别、新领域等零样本场景的泛化能力。监督式方法与自监督时域重采样方法对比见表2。

监督式方法存在一个共同的局限性：高度依赖于预先定义好的特定目标类别。这意味着在重采样前，视频帧需要经过目标检测或语义分析模型进行预处理，包含高置信度目标的帧会被赋予更高的重采样权重。尽管这类方法能根据特定目标精准决策，但其计算开销和零样本（zero-shot）场景下的泛化能力仍是限制其广泛应用的瓶颈。

为解决现有方法运算代价高昂及零样本适应性不足的问题，本文提出一种基于自监督特征嵌入和聚类的视频时域自适应重采样方法。主要贡献如下。首先，采用深度网络 ResNet-18 提取帧级特征，并通过一种基于余弦相似度的自监督度量学习损失函数进行微调。该设计促使在原始视频中时序连续的帧在嵌入空间中形成平滑的流形分布，从而有效建模帧间相似性。其次，对嵌入后的特征采用基于流形等分点的时序数据聚类算法，簇数量由目标采样帧数决定。重采样决策依据聚类结果动态生成：保留首帧，并仅当一帧与前一采样帧不属于同一簇时才进行采样，以确保对内容突变的快速响应。最后，重采样后的稀疏帧序列经由 H.266/VVC 编码，解码端则结合插值网络重构完整视频。该方法凸显了任务与数据驱动无关的特性，不需要特定任务的标注信息，即可依据帧间内容变化自适应重采样，从而更好地平衡压缩效率与下游视觉任务的性能。

1　方法论：基于深度度量学习的帧间相似性建模

传统视频处理通常采用与内容无关的均匀时域重采样，这种方式无法根据内容动态调整，易造成信息冗余或丢失。而基于特定语义（如预定义目标）的重采样方法，尽管在特定任务上表现优异，却面临泛化性差的瓶颈：当场景中不包含已知目标时，这类方法可能失效。为了克服这些

表2　监督式方法与自监督时域重采样方法对比

对比维度	监督式方法	自监督方法（本文）
是否需要标签	是（如目标检测标签）	否
对视频内容的依赖	画面中需要出现已知类别的目标	不限于画面中的目标
重采样依据	画面中对象位置	视频帧的特征流形结构变化
泛化能力	较差（对新目标不鲁棒）	适用于未标注/未知目标，甚至于无目标画面

局限，本文主张从数据自身出发，重新定义和度量视频帧间的“距离”或相似性。

本文引入了度量学习作为核心思想。度量学习的目标是学习一个映射函数，将原始数据投影到一个新的嵌入空间中，使得在该空间内，语义相似样本（如时序上相邻的视频帧）彼此靠近，而语义不相似样本（如来自不同场景的帧）则相互远离[13]。

对于度量学习，设$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]\in\mathbf{R}^{d\times N}$为训练样本，其中$\boldsymbol{x}_i\in\mathbf{R}^d$是第$i$个训练样本，$N$是训练样本总数，则样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间的距离被度量为：

$$d_{\boldsymbol{M}}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}_i-\boldsymbol{x}_j)} \tag{1}$$

其中，$d_{\boldsymbol{M}}$表示一个距离度量，矩阵$\boldsymbol{M}$对称且半正定，可以分解为$\boldsymbol{M}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}$。因此，对样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$距离的度量表示为：

$$d_{\boldsymbol{M}}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)=$$

$$\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{x}_i-\boldsymbol{x}_j)}=\left\|\boldsymbol{W}(\boldsymbol{x}_i-\boldsymbol{x}_j)\right\|_2 \tag{2}$$

其中，$\boldsymbol{W}$被视为一种线性变换，而$d_{\boldsymbol{M}}$这种度量被称为马哈拉诺比斯距离，即在线性变换空间中的欧几里得距离。事实上，对于被视为流形数据的视频序列而言，线性方法并不总是能够揭示数据的非线性特性。因此，对于本文面临的视频帧样本，$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间的距离度量需要经过非线性变换。而深度度量学习[14-16]基于样本之间的相似性原理，通过多层结构和非线性激活函数实现更高的抽象水平。

2 基于自监督学习的视频时域重采样框架

时域重采样方法旨在从一个完整的视频帧集合$\boldsymbol{V}=\{f_1,f_2,f_3,\cdots,f_n\}$中，自适应地选取一个最优子集$\boldsymbol{S}\subseteq\boldsymbol{V}$，使得该子集在下游机器视觉任务中的性能$Q(\boldsymbol{S})$最大化：

$$\underset{\boldsymbol{S}\subseteq\boldsymbol{V}}{\operatorname{argmax}}\ Q(\boldsymbol{S}) \tag{3}$$

为实现此目标，本文设计了一个端到端的处理流程，面向机器视觉的视频时域重采样技术路线如图1所示。整个框架包含两个主要阶段。

（1）视频重采样与编码，如图1（a）所示。此为本文的核心创新。输入的视频序列首先通过一个自监督的帧间度量与嵌入模块，将其投影到一个低维特征空间，形成能够反映内容变化的平滑流形。随后，在流形上进行聚类分析以识别内容相似的帧簇，并根据聚类结果进行自适应重采样，选取最具代表性的m帧。最后，这m帧及其位置元信息被送入标准视频编码器（如H.266/VVC）进行压缩。

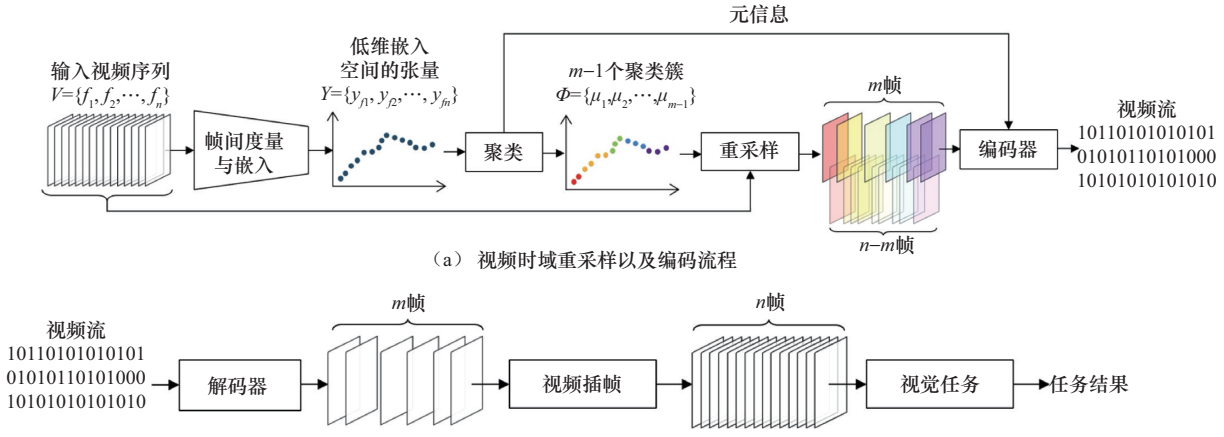


图1 面向机器视觉的视频时域重采样技术路线

（2）解码、重建与任务推理，如图1（b）所示。在接收端，码流被解码得到稀疏的采样帧。接着，一个深度视频插帧网络利用这些采样帧和元信息，重建出完整的n帧视频序列。最后，重建后的视频被送入下游视觉任务模型（如目标检测）进行推理。此阶段主要用于评估提出的重采样方法的有效性。

2.1　自监督嵌入与流形学习

本文方法的核心在于学习一个嵌入函数EBD(·)，该函数将高维的视频帧f_t映射到一个低维特征向量$\boldsymbol{y}_t=\mathrm{EBD}(f_t)$，且目标是使该嵌入空间具备两大特性。

（1）时序平滑：在时间上连续的视频帧，其内容变化通常是渐进的。因此期望它们的嵌入向量在特征空间中也彼此靠近，形成一条平滑的流形轨迹。这一特性可以直观地理解为，嵌入向量$\boldsymbol{y}_t$关于时间t的变化率应该尽可能小，即：

$$\frac{\partial \boldsymbol{y}_t}{\partial t}\approx 0 \tag{4}$$

（2）可区分性：仅追求平滑性可能导致一个平凡解，即所有帧都被映射到同一点。为避免这种情况，嵌入空间必须能够区分来自不同视频或不同场景的内容。

为同时实现这两个目标，本文设计了一个基于余弦相似度的自监督对比损失函数。对于一对在时序上相邻的帧f_t,f_{t+1}，它们的嵌入向量$\boldsymbol{y}_t$和$\boldsymbol{y}_{t+1}$被视为“正样本对”。而从其他不相关视频序列中随机采样的帧f_n所对应的嵌入向量$\boldsymbol{y}_n$则被视为“负样本”。损失函数定义为：

$$L(f_t,f_{t+1},\mathcal{N})=-E\left[\log\frac{\mathrm{e}^{C(\boldsymbol{y}_t,\boldsymbol{y}_{t+1})}}{\mathrm{e}^{C(\boldsymbol{y}_t,\boldsymbol{y}_{t+1})}+\sum_{\boldsymbol{y}_n\in\mathcal{N}}\mathrm{e}^{C(\boldsymbol{y}_t,\boldsymbol{y}_n)}}\right] \tag{5}$$

其中，视频帧f_t属于正样本集$\mathcal{T}$，而$\boldsymbol{y}_n$对应的视频帧表示负样本数据集$\mathcal{N}$，且满足$\mathcal{T}\cap\mathcal{N}=\varnothing$。$\sum_{\mathcal{N}}\mathrm{e}^{C(\boldsymbol{y}_t,\boldsymbol{y}_n)}$作为损失函数$\mathcal{L}$的惩罚项，当不属于同一视频场景的$f_t,f_n$在潜在空间中的距离过近，即$\sum_{\mathcal{N}}\mathrm{e}^{C(\boldsymbol{y}_t,\boldsymbol{y}_n)}$变小时，整项损失函数将增大。$C(a,b)$表示余弦相似度，即：

$$C(\boldsymbol{y}_t,\boldsymbol{y}_{t+1})=\frac{\sum_{i=1}^{n}\boldsymbol{y}_t^i\boldsymbol{y}_{t+1}^i}{\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{y}_t^i\right)^2}\sqrt{\sum_{i=1}^{n}\left(\boldsymbol{y}_{t+1}^i\right)^2}} \tag{6}$$

式（5）通过最大化正样本对的余弦相似度来鼓励时序平滑性，同时通过最小化与负样本集的相似度来确保可区分性。视频帧嵌入原理如图2所示，该过程在嵌入空间中将属于同一轨迹的加粗样本点拉近，同时推远不相关的灰色样本点。

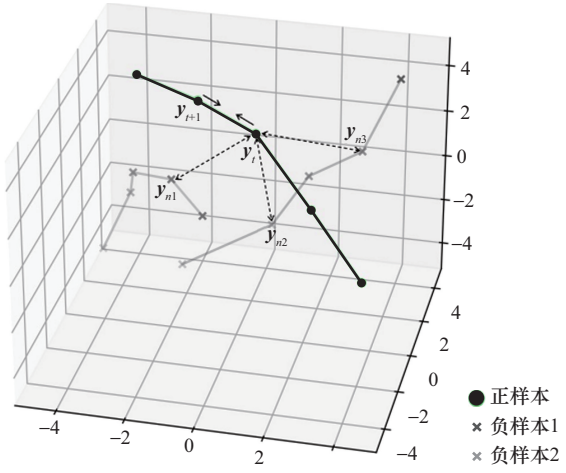


图2　视频帧嵌入原理

2.2　网络架构与训练策略

特征提取以及样本比对网络如图3所示，本文采用ResNet-18作为特征嵌入网络的骨干，其中stride控制卷积核的移动速度；padding用于调整输入数据的边界，可以保持特征图尺寸或避免边界信息丢失，dim表示当前特征图维数，⊕表示特征相加，即两个特征张量在通道维度上的逐元素相加操作，用于实现残差连接，在完成信息融合的同时缓解梯度消失问题。BN和ReLU分别表示批量归一化和激活函数。

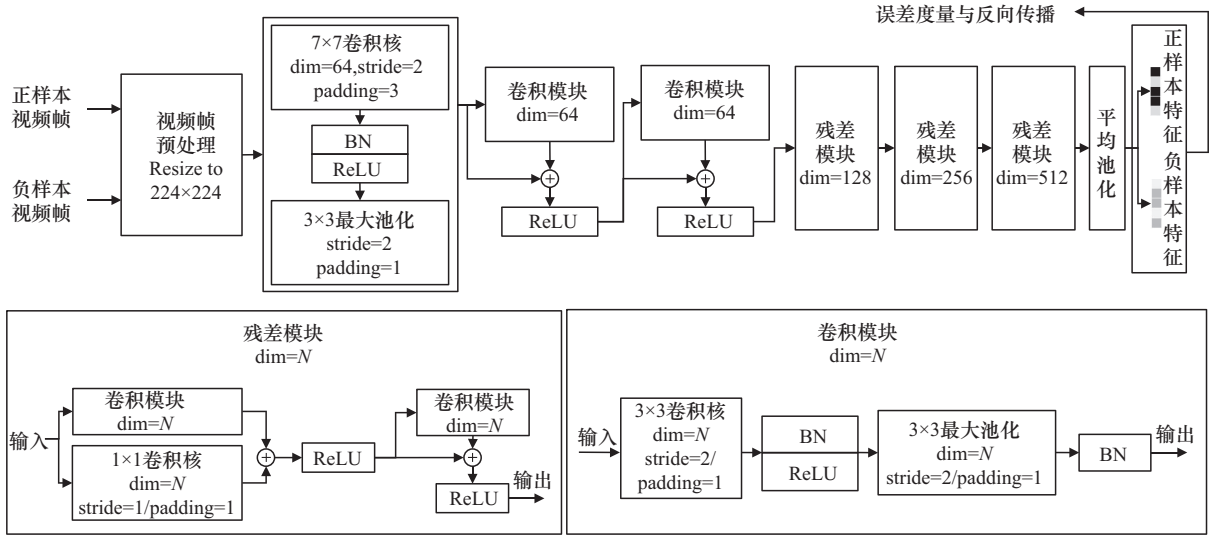


图3　特征提取以及样本比对网络

输入视频帧经卷积层提取空间特征后，将最终卷积层的输出展平并映射至128维嵌入空间。与ResNet-18原始网络的1 000维嵌入空间不同，128维常被用作特征比对[17-18]。为使网络学习到具有时序意义的表征，采用以下训练策略。

（1）对比损失优化：网络参数通过最小化式（5）中定义的对比损失进行端到端优化。使用随机梯度下降法进行反向传播。

（2）引入了基于动态半径的半硬负样本挖掘策略。该策略优先选择与当前锚点距离适中、具有一定挑战性的负样本，从而加速模型的收敛和性能提升。

（3）为防止嵌入空间坍缩至平凡解，额外引入了旋转预测的机制。要求模型在提取特征的同时，预测输入图像被施加的随机旋转角度，迫使网络学习图像的内在结构特征，从而起到正则化作用。

2.3　基于流形等分点的时序数据聚类

当视频帧序列被映射到平滑的特征流形后，下一步是对其进行聚类和重采样。假设用户的目标是得到一个包含m帧的视频，聚类中类数会被设定为$m-1$，因为原则上必须对原始视频的最后一帧进行采样以保证视频首尾完整。参数m的选择由用户自定义。

设原始高维时序数据$\boldsymbol{V}=\{f_1,f_2,f_3,\cdots,f_n\}$经非线性映射嵌入$d=128$维空间，形成流形轨迹$\mathcal{M}=\left\{R(f_t)\in\mathbf{R}^d|t=1,\cdots,n\right\}$。通过3次样条插值将离散点集$\mathcal{M}$拟合为光滑曲线$\gamma(s):[0,C]\to\mathbf{R}^d$，其中$l$为近似弧长，表示为：

$$l=\sum_{t=1}^{n-1}\left\|\boldsymbol{y}_{t-1}-\boldsymbol{y}_t\right\|_2 \tag{7}$$

之后，沿弧长方向均匀选取$m-2$个分割点$\{p_k\}_{k=1}^{m-2}$，将整个流形分割为m份，满足：

$$p_k=\gamma\left(\frac{k\cdot l}{m-1}\right),k=1,2,\cdots,m-2 \tag{8}$$

最后，将以等分点为边界将时序数据划分为$m-1$个子序列簇$B_1,\cdots,B_{m-1}$，其中：

$$B_j=\left\{f_t\left|\frac{(j-1)l}{m-1}\leqslant\sum_{i=1}^{t-1}\left\|\boldsymbol{y}_{t-1}-\boldsymbol{y}_t\right\|_2<\frac{jl}{m-1}\right.\right\} \tag{9}$$

该方法通过继承流形空间的局部拓扑结构，保持原始序列的时序连续性，确保轨迹数据的特征空间与时序规律得以完整保留。同时，自适应划分机制根据轨迹曲率动态调整等分点间距，避免了欧氏距离偏差。此外，每个簇均对应演化过程的特定阶段，赋予了聚类结果高度的可解

释性。

对于聚类结果，按照视频帧的时间顺序遍历类别标签，按照以下规则执行重采样。

（1）当前视频帧为第一帧时，对该帧进行采样。

（2）当前视频帧与相邻的前一帧不属于同一聚类类别时，对该帧进行采样。

以此，将原本包含 n 帧的视频序列重采样，得到一个包含 m 帧的视频序列，这些帧最大限度地捕捉了视频内容在特征层面的显著变化。最后将该视频序列进行视频编码，形成视频流。

2.4　码流元信息构建和视频插帧

为了使解码器能够恢复原始的完整视频，重采样过程中的决策信息必须作为元数据与压缩后的视频帧一同传输。本文中该元信息被高效地构建为相邻两个采样帧之间被丢弃的帧数集合。

在解码端，码流被解析后，一个支持任意数量插值的视频插帧网络启动工作。该网络以相邻的两个重采样帧为输入，并根据元信息中指定的应插帧数，精确地重建出中间被丢弃的帧序列。最终，重建后的全帧率视频被送入下游的机器视觉任务模型。需要说明的是，视频插帧算法仅作为验证本文重采样方法效果的工具，其具体实现并非本文的创新重点。

3　实验与分析

3.1　实验数据与模型

3.1.1　测试视频序列与实验动机

对于被重采样和深度模型推理的输入视频序列，本文使用公开的轨道交通场景数据集 OSDaR23[19]的测试序列，筛选出包含正样本目标的视频序列，每个场景包含不同角度的视频序列，每个序列包含 10 帧视频。需要特别说明的是，本文研究的核心目标并非针对轨道交通场景进行专门优化，而是旨在验证方法在跨领域、零样本（即未使用目标领域训练数据）条件下的通用性。轨道交通场景具有运动模糊、复杂背景及特定目标等挑战性因素，为检验方法的鲁棒性提供了良好的测试场景。方法的其他核心模块（特征提取、度量学习、插帧）均在其他通用数据集上预训练或学习，未使用任何轨道交通数据进行微调，此举正是为了强调方法本身学习到的时空表示与插帧能力的普适性。若旨在优化特定场景性能，可通过在目标领域数据上微调模型实现，但这并非本文关注的重点。

3.1.2　帧间度量与嵌入

特征提取与度量模块的预训练主要在大型视频动作识别数据集 Kinetics400[20]上进行。该数据集包含 306 245 个视频序列，覆盖 400 个动作类别。预训练阶段，从每帧视频中随机裁剪 224×224 区域作为网络输入，并施加随机裁剪、随机水平翻转、随机灰度化及色彩抖动等数据增强策略（动作类别标签不参与此阶段训练）。负样本从 Kinetics400 中随机选取 6 245 个视频序列。特征提取主干网络采用 ResNet-18，最终通过一个独立的全连接层输出 128 维特征向量。

3.1.3　编解码器

时域重采样效果的验证包含码率与视觉任务精度双维度指标。需要注意，即使编码帧数相同，不同重采样策略导致的帧内容差异亦会引起码率波动。此外，编码器可通过调整量化参数（quantization parameter，QP）精细控制被采样帧的画质：增大 QP 可降低码率，但可能导致画面模糊与锯齿感增强；减小 QP 可提升画质清晰度，但会增大码率。本文选用基于 H.266/VVC 标准的开源编解码器 VVenC[21]/VVdeC[22]。

3.1.4　视频解码后重建

在关键帧被提取后，为有效重建重采样过程中被舍弃的视频帧，本文方案采用阵列时间戳帧插值方法（EMA-VFI[23]）以适应变化的帧间采样间隔。该模块已在 Vimeo-90K[24]数据集上训练至收敛。其以相邻两帧重采样帧作为参考，根据元

信息中指定的插帧数量（一个大于0且小于256的整数，由8位二进制元信息约束），精确重建中间缺失的帧序列。这意味着，EMA-VFI网络本身支持非均匀间隔的帧间插值。具体来说，该模型以任意两帧视频图像为输入，并根据它们之间的跳帧数量（即待插帧数）执行多帧重建。本文编码端根据重采样结果构建的元信息为一组整数，表示每对相邻重采样帧之间丢弃的帧数。该跳帧数被直接传递给解码端的EMA-VFI网络，从而实现对任意长度间隔帧的插值操作。

需要指出的是，采样帧数m的增大虽然有助于压缩率的提升，但同时会导致相邻采样帧之间间隔变长，从而加大插帧网络的重建难度，进而可能对解码后视频的任务精度造成影响。根据EMA-VFI原文实验结果表明，帧间间隔越大，插帧精度越低，尤其在快速运动场景中误差更为明显，因此m的设定在压缩与精度之间存在显著权衡。

3.1.5 视觉任务

为充分评估时域重采样方法的性能优势，采用detectron2[25]框架进行机器视觉任务测试。该框架集成了目标检测与分割算法，可全面衡量重采样视频在高层视觉任务上的表现。具体地，选用框架内预训练的Faster R-CNN X101-FPN[26]模型进行目标检测验证，该模型已在COCO[27]数据集上预训练收敛。

3.2 实验描述

3.2.1 实验流程及细节

实验流程根据图1展示的过程完整进行。对于帧间度量与嵌入环节，模型训练的迭代次数为100次；视频插帧和视觉任务环节，均根据原始文献中的参数设置进行训练。视频重采样帧数约为原始视频帧数的1/3。精确地，假设原始视频有n帧，在本文实验中，对于OSDaR23数据集，被采样帧数m设置为$\left\lfloor \frac{n}{3} \right\rfloor +1$，其中$\lfloor \cdot \rfloor$表示向下取整符号，被采样帧数包含了每个被采样视频的最后一帧，因此对于聚类计算中的类别则是$\left\lfloor \frac{n}{3} \right\rfloor$。实际上，目标帧数的值可以取任意小于或等于$n$的正整数。本文在OSDaR23数据集时取$\left\lfloor \frac{n}{3} \right\rfloor +1$，是因为该数据集每个场景的视频帧为10帧，即$n$=10，此时$\left\lfloor \frac{n}{3} \right\rfloor +1=4$，对于均匀取值方法可取第1、4、7、10帧。如果不取$\left\lfloor \frac{n}{3} \right\rfloor +1$，就无法严格满足均匀取值的定义，造成算法比对上的歧义。

对照组分为两个方法，基线方法和基于目标跟踪的方法。基线方法即每个视频序列均匀地采样视频帧。基于目标跟踪的方法为标准方法[28]。标准方法使用DeepSORT对视频帧进行目标检测与跟踪，提取每帧目标轨迹并进行数据关联，识别出帧间对象运动的相似性。随后将连续相似的帧划为一组，仅保留每组首帧，其余帧丢弃，从而实现基于目标的自适应时序重采样。对于编码，本文选择全帧内（all intra，AI）编码模式和随机访问（random access，RA）编码模式。OSDaR23数据集的测试部分在每种编码模式下使用VVenC做6次不同QP的编码。本文提出的重采样方法和对照组重采样方法都执行以上步骤，并记录码流大小。

3.2.2 结果评价指标

作为目标检测任务主要的评价指标，全类平均精确率（mean average precision，mAP）综合了检测算法在不同召回率水平下的精确率表现。首先，计算每个类别的平均精确率（average precision，AP）。根据检测框的置信度对检测结果进行排序。之后计算每个置信度阈值下的查全率和查准率。并使用若干采样点查全率和查准率的对应关系计算每个类别下的AP：

$$\mathrm{AP}=\sum_a\left(\mathrm{Rec}_a-\mathrm{Rec}_{a-1}\right)\mathrm{Pre}_a \tag{10}$$

其中，Rec_a和Pre_a分别是第a个查全率和查准率的采样点。最终将所有类别的AP值进行平均得到最终的mAP指标：

$$\mathrm{mAP}=\frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i \tag{11}$$

在传统的目标检测推理任务中，重点关注推理精度，通常不考虑被推理数据的压缩情况。而在本文提出的方法中，需要综合考虑被推理数据的压缩率以及推理精度，因此需要更立体的评价指标。面向人眼视觉的视频编码中，常用评价指标为Bjøntegaard差分率（Bjøntegaard-delta rate，BDrate）。而BDrate中采用的质量度量为峰值信噪比（peak signal-to-noise ratio，PSNR）、结构相似性衡量指标（structure similarity index measure，SSIM）等差异性度量，而本文将使用mAP取代PSNR和SSIM，凸显机器视觉任务的精度，构建成BDmAP。本质上，BDmAP是计算对照视频的码率–检测精度曲线与实验视频码率–检测精度曲线之间面积的平均积分差，表示为：

$$\mathrm{BDmAP}\approx 10^{\frac{1}{D_u-D_b}\int_{D_b}^{D_u}\left[g_t(D)-g_a(D)\right]\mathrm{d}D}-1 \tag{12}$$

对于积分上下限：

$$\begin{aligned}D_u&=\min\left(\max\left(D_{a,i}\right),\max\left(D_{t,i}\right),\right.\\&\left.g_a\left(\max\left(R_{a,i}\right)\right),g_t\left(\max\left(R_{t,i}\right)\right)\right)\\D_b&=\max\left(\min\left(D_{a,i}\right),\min\left(D_{t,i}\right),\right.\\&\left.g_a\left(\min\left(R_{a,i}\right)\right),g_t\left(\min\left(R_{t,i}\right)\right)\right)\end{aligned} \tag{13}$$

其中，$D_{a,i}$和$D_{t,i}$分别为参考值和实验值的畸变测量值。$g_a()$和$g_t()$分别为参考值和实验值的拟合曲线。这些值是通过多次不同量化参数的编码得到的，取得其中的最大和最小值参与积分计算。BDmAP值为正时，表明实验视频比对照视频有更好的表现。

本文另一个需要使用的评价指标是帕雷托全类平均精确率（Pareto mAP），在实验表格中简写为P-mAP。该评价指标表示对照视频与实验视频之间的码率–检测精度曲线与坐标轴围成的面积的百分比大小。值为负时，表明实验视频比对照视频有更好的精度表现。此外，为了更全面地使用Pareto mAP指标，本文在选取码率–检测精度曲线的重采样点时，分别抽取6个重采样点中码率最低的4个点（low 4）、码率居中的4个点（mid 4）和码率最高的4个点（high 4）参与计算，以此体现不同码率层面的优势。

3.3　实验结果

3.3.1　比对试验

随机访问编码模式下的码率–检测精度曲线如图4所示，全帧内编码模式下的码率–检测精度曲线如图5所示。其中，本文方法在绝大多数测试序列上都展现出显著优势。相较于均匀采样方法，本文方法在所有码率点上几乎都能以更低的码率达到相同或更高的mAP。与标准方法相比，本文方法在多数情况下也表现出相当的性能。

随机访问编码模式下的Pareto mAP和BDmAP表现见表3，全帧内编码模式下的Pareto mAP和BDmAP表现见表4。随机访问（RA）模式：本文方法取得了–30.34%的总平均P-mAP，并在高、中、低码率段均表现稳定。总平均BDmAP提升了2.30%。唯一的例外是“3_fire_site”，其正值表明在该特定场景下表现欠佳，是因为画面中随机高频扰动与本文方法基于内容变化重采样的假设略有冲突。此外，全帧内编码模式下同样表现出色，总平均P-mAP为–27.63%，BDmAP提升2.95%。尤其在“1_calibration”和“13_station_ohlsdorf”视频上，P-mAP分别达到-49.73%和–52.98%。

OSDar23数据集中不同重采样方法重建帧的目标检测结果定性对比如图6所示，直观地展示了在相似码率下不同重采样方法重建后某一帧的

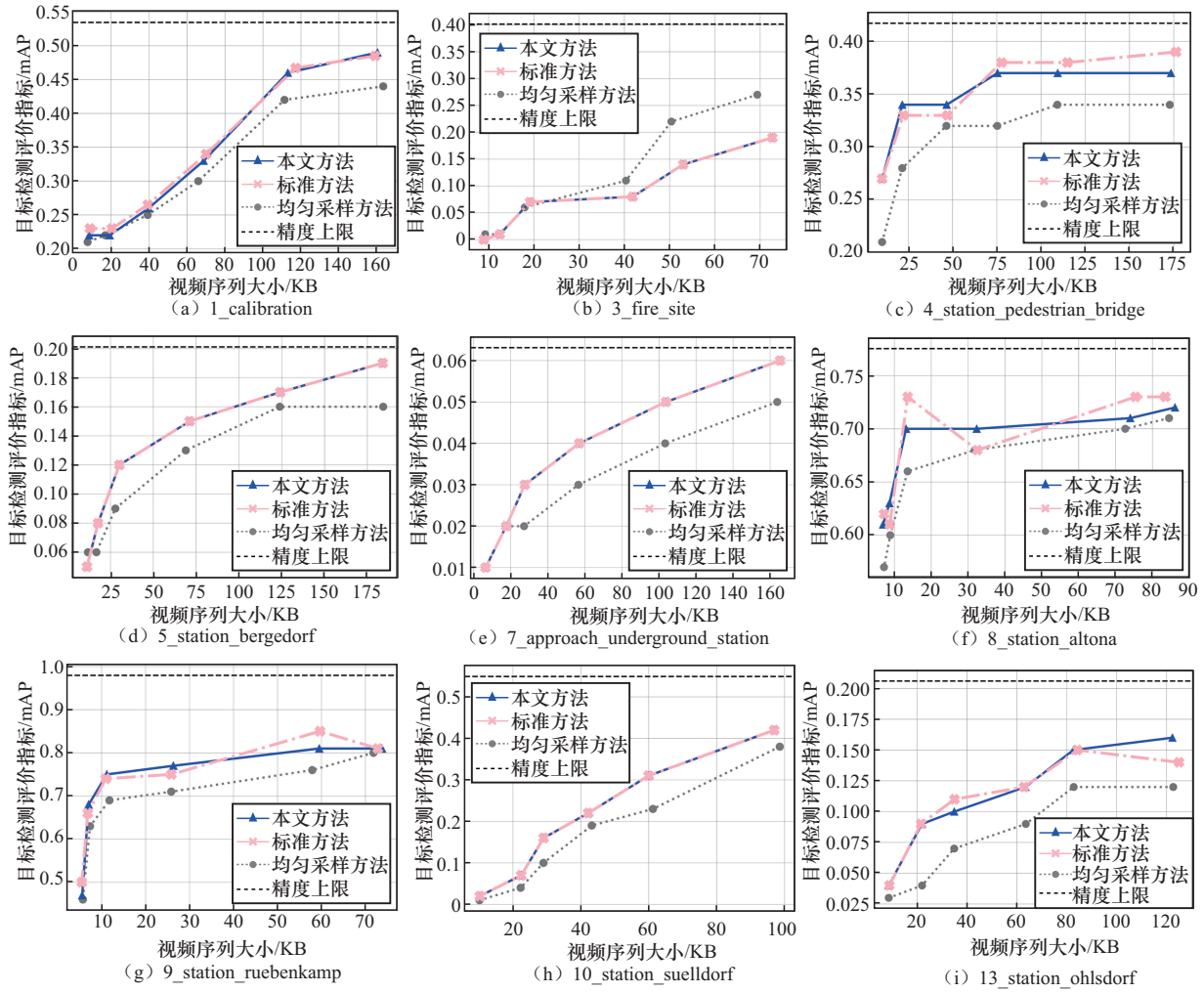


图4　随机访问编码模式下的码率-检测精度曲线

目标检测结果。结果表明，本文方法通过自监督学习，能够精准地保留视频中发生语义变化的关键时刻，即使这些变化并非由预定义的目标运动引起，从而为下游视觉任务提供信息量更丰富的输入。

3.3.2　其他指标和任务

为进一步验证本文所提重采样方法的有效性与泛化能力，本节将评估其在更广泛条件下的性能。首先，在不同视频序列大小（即不同码率）范围内，详细分析本文方法对目标检测任务精度（AP）及重建视频保真度（PSNR，SSIM）的影响。其次，为检验方法的零样本泛化能力，在一个全新的、更具挑战性的多目标视频分割任务上进行评估，并与标准方法进行对比。

本节在6个不同的视频序列大小（码率）区间内，对目标检测任务的平均精度（AP）以及重建视频的PSNR和SSIM进行了详细的性能评估。不同视频序列大小范围下目标检测与重建质量的详细性能对比见表5。从表5的平均结果可以看出，本文方法在3个主要目标类别（Bicycle，Person，Car）上的AP与标准方法基本持平或略有优势。这一结果表明，本文提出的自监督重采样方法在保留对目标检测任务至关重要的语义信息方面，至少与基于目标跟踪的标准方法同样有效。

另一个发现是，任务性能的提升或保持与传

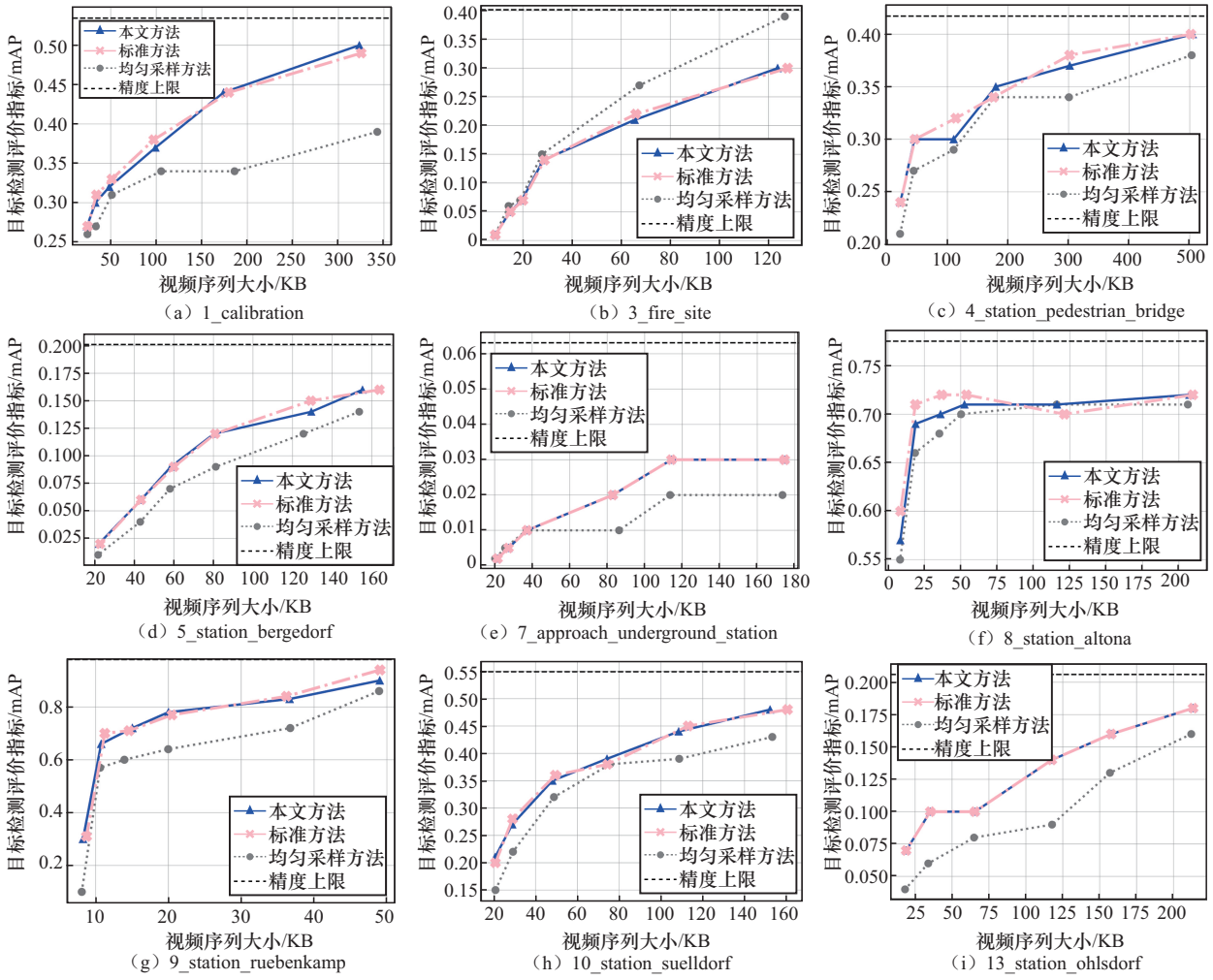


图 5　全帧内编码模式下的码率–检测精度曲线

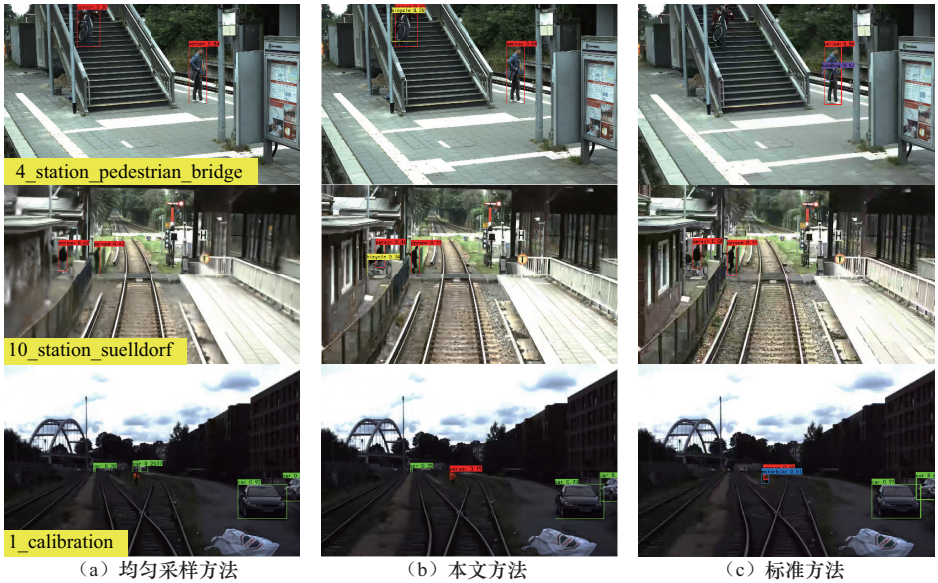
表 3　随机访问编码模式下的 Pareto mAP 和 BDmAP 表现

视频名称	P-mAP↓	P-mAP(high 4)↓	P-mAP(mid 4)↓	P-mAP(low 4)↓	BDmAP↑
1_calibration	−10.72%	−11.34%	−13.01%	−11.91%	1.51%
3_fire_site	15.79%	26.55%	8.14%	−0.04%	−2.33%
4_station_pedestrian_bridge	−62.41%	−65.53%	−69.29%	−57.78%	4.05%
5_station_bergedorf	−39.41%	−41.34%	−43.55%	−38.63%	2.18%
7_approach_underground_station	−26.61%	−25.45%	−2.30%	−27.44%	0.37%
8_station_altona	−39.03%	−50.49%	−45.80%	−26.55%	2.15%
9_station_ruebenkamp	−33.86%	−55.09%	−67.81%	−16.09%	5.46%
10_station_suelldorf	−23.22%	−25.36%	−19.34%	−19.90%	4.23%
13_station_ohlsdorf	−53.58%	−33.21%	−45.44%	−61.55%	3.09%
总平均	−30.34%	−31.25%	−33.16%	−28.88%	2.30%

统面向人类视觉的保真度指标（PSNR 和 SSIM）之间存在解耦现象。如表 5 所示，尽管本文方法在 AP 上表现更优，但其平均 PSNR 和 SSIM 值与标准方法相比稍弱。这种策略性地“牺牲”部分对机器任务无关紧要的像素保真度，换取了关键语义信息的完整性，从而在相似的码率预算下，

表4　全帧内编码模式下的Pareto mAP和BDmAP表现

视频名称	P-mAP↓	P-mAP(high 4)↓	P-mAP(mid 4)↓	P-mAP(low 4)↓	BDmAP↑
1_calibration	−49.73%	−64.34%	−24.63%	−28.75%	5.45%
3_fire_site	25.85%	30.53%	15.37%	6.90%	−3.15%
4_station_pedestrian_bridge	−33.49%	−29.63%	−12.05%	−31.37%	1.92%
5_station_bergedorf	−24.45%	−29.71%	−21.96%	−20.25%	1.90%
7_approach_underground_station	−29.46%	−37.33%	−37.98%	−12.34%	0.52%
8_station_altona	−30.58%	−68.11%	−47.47%	−28.79%	1.71%
9_station_ruebenkamp	−28.69%	−44.28%	−55.56%	−12.02%	10.85%
10_station_suelldorf	−25.15%	−27.61%	−20.67%	−22.91%	4.13%
13_station_ohlsdorf	−52.98%	−41.92%	−48.43%	−71.76%	3.23%
总平均	−27.63%	−34.71%	−28.15%	−24.59%	2.95%



（a）均匀采样方法　（b）本文方法　（c）标准方法

图6　OSDar23数据集中不同重采样方法重建帧的目标检测结果定性对比

表5　不同视频序列大小范围下目标检测与重建质量的详细性能对比

视频序列大小范围/KB	AP↑						PSNR↑		SSIM↑	
	Bicycle		Person		Car					
	标准方法	本文方法	标准方法	本文方法	标准方法	本文方法	标准方法	本文方法	标准方法	本文方法
5.49~11.83	0.105 3	**0.118 6**	**0.177 8**	0.164 3	0.280 3	**0.297 0**	**25.54**	25.32	**0.786 7**	0.785 8
6.86~22.43	0.164 3	**0.165 5**	0.221 5	**0.232 3**	**0.327 6**	0.302 1	**25.67**	25.64	**0.798 1**	0.793 2
10.94~47.23	0.213 5	**0.216 6**	0.267 4	**0.271 3**	0.362 4	**0.363 8**	26.09	**26.10**	0.809 5	**0.810 5**
25.83~77.69	**0.273 5**	0.271 4	**0.270 2**	0.263 3	0.383 0	**0.385 4**	26.78	**26.88**	0.820 9	**0.829 6**
50.38~124.11	0.295 1	**0.305 0**	0.351 9	**0.361 8**	0.409 6	**0.415 6**	27.01	**27.18**	0.832 3	**0.842 9**
69.36~184.25	0.314 8	**0.327 7**	**0.368 7**	0.365 6	**0.453 2**	0.444 9	27.22	**27.30**	**0.853 2**	0.850 4
平均	0.227 7	**0.234 1**	0.276 2	**0.276 4**	**0.369 3**	0.368 1	**26.39**	26.38	0.816 7	**0.818 7**

实现了更优的任务性能。这一特性对于带宽受限的实际应用场景（如自动驾驶、远程监控）尤为重要，因为它允许系统在不显著增加数据传输成本的前提下，维持或提升AI模型的分析精度。

为了评估本文方法在未知任务和数据集上的零样本泛化能力，本文将其应用于一个完全不同的视觉任务——多目标视频分割。该实验在极具挑战性的公开基准数据集上进行。DAVIS2017[29]专为多实例视频对象分割设计，包含复杂的场景，如物体遮挡、快速运动和多目标交互，是检验模型鲁棒性的理想选择。使用 YOLACT++[30]原文的预训练模型作为验证算法，插帧设置与目标检测任务相同。由于本文方法在训练阶段完全不依赖任何分割标签，其在该数据集上的表现能够直接反映其与任务无关的自适应时域重采样能力。在评估分割任务时，本文采用两个核心指标：Jaccard 相似系数和 F_1 分数。其中 Jaccard 相似系数（Jaccard similarity coefficient），也常被称为交并比，用于衡量预测分割掩码 A 与真实掩码 B 之间的重叠程度。该指标是分割任务评估的黄金标准。计算式为：

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{14}$$

而 F_1 分数，是精确率（Precision）和召回率（Recall）的调和平均值。它提供了一个平衡的度量，能够同时惩罚假正例（false positives，FP）和假反例（false negatives，FN），即错误分割出的背景像素和未能分割出的目标像素，这对于处理前景与背景区域不平衡的分割场景尤为重要。计算式为：

$$F_1=\frac{2\cdot TP}{2\cdot TP+FP+FN} \tag{15}$$

在超参数设置上，DAVIS2017 数据集的被采样帧数 m 设置为 $\left\lfloor\frac{n}{4}\right\rfloor+1$，编码方法为随机访问模式。

DAVIS2017 数据集下多目标分割任务精度表现见表 6，从总体平均结果来看，本文方法在 Jaccard 相似系数上优于标准方法，同时 F_1 分数与标准方法基本持平，证明了本文方法所学习到的特征表示具有很强的任务泛化性。更值得注意的是，在较低码率范围（25.83~77.69 KB）内，本文方法的 Jaccard 相似系数（0.369 4）显著优于标准方法（0.335 6）。这表明在高压缩率场景下，本文方法的优势更为突出。

这种现象的根本原因在于，视频压缩对分割任务的影响尤为严重，它会模糊物体的精细边界和纹理细节，从而干扰分割模型的判断 。标准方法可能因均匀采样或简单的运动判据，在压缩视频中丢失了对精确分割至关重要的少数关键帧。相比之下，本文方法通过对特征流形的分析来选择采样点，能够更可靠地捕捉到内容发生本质变化的时刻——例如，一个物体从被遮挡到重新出现，或者一个物体的姿态发生剧烈变化。这些时刻的帧对于下游分割模型重建时序一致的、准确的掩码至关重要。通过保留这些信息最丰富的帧，本文方法为分割模型提供了更高质量的输入，从而在高压缩条件下表现出更强的鲁棒性。

DAVIS2017 数据集上多目标分割结果定性对比如图 7 所示，更直观地展示本文方法在多目标分割任务上的优势。图 7 展示了在相似码率条件

表 6　DAVIS2017 数据集下多目标分割任务精度表现

视频序列大小范围/KB	Jaccard 相似系数↑		F_1↑		PSNR↑		SSIM↑	
	标准方法	本文方法	标准方法	本文方法	标准方法	本文方法	标准方法	本文方法
25.83~77.69	0.335 6	**0.369 4**	**0.534 1**	0.495 8	**25.45**	25.21	**0.784 5**	0.783 3
65.96~184.25	**0.393 1**	0.372 2	0.537 3	**0.557 9**	**25.61**	25.50	**0.796 0**	0.791 7
159.09~288.00	**0.397 5**	0.397 1	0.562 0	**0.563 1**	26.02	**26.03**	0.807 8	**0.808 9**
290.23~512.00	0.402 3	**0.413 7**	0.567 9	**0.579 6**	26.69	**26.81**	0.819 1	**0.826 4**
平均	0.382 1	**0.388 1**	**0.550 3**	0.549 1	**25.94**	25.89	0.801 9	**0.802 6**

下，不同方法重建视频后某一帧的分割效果；各列分别展示了均匀采样方法、标准方法以及本文方法的分割结果。通过对比可以发现，本文方法在处理复杂场景时表现出明显的优势。特别是在第二行展示的快速运动场景中，标准方法可能由于采样帧之间的时间间隔较大，对高速运动物体的跟踪失败，而本文方法能够自适应地在内容剧烈变化的区域增加采样密度，保留了定义运动轨迹的关键帧，因此其分割结果更为完整，这解释了F_1分数能够与标准方法保持相当水平的原因。

（a）均匀采样方法　（b）本文方法　（c）标准方法

图7　DAVIS2017数据集上多目标分割结果定性对比

通过对视频内容在特征空间的流形结构进行建模，本文方法能够智能识别并保留那些对下游任务（无论是检测还是分割）信息量最大的帧，而不是简单地依赖于像素相似度或预定义目标的运动。这种基于语义变化的采样策略，使其在面对压缩、遮挡和复杂运动等挑战时，能够为下游视觉模型提供更鲁棒、信息更丰富的输入。

3.3.3　时域重采样方法运行速度分析

为了验证本文方法在实时处理场景中的可行性，分析了其与标准方法在特征提取和重采样决策阶段的计算复杂度，本文方法与标准方法在计算开销与性能增益上的量化对比见表7。考虑计算设备的运行状态、电气环境等差异，本文通过算法核心部分复杂度间接反映算法的运行速度。本文方法的核心计算在于ResNet-18的特征提取和聚类，其模型参数量约为1 170万，处理单帧的浮点运算次数约为1.82 GFLOPS。相比之下，基于目标跟踪的标准方法采用更深、更复杂的网络参数量为3 690万，浮点运算次数为104.70 GFLOPS，表明本文方法的计算开销比标准方法低了一个数量级。尽管计算量大幅降低，但从BDmAP和其他两项BDrate指标看，两者性能在同一水平。这凸显了本文方法在效率上的巨大优势，实现了性能与效率的卓越平衡。

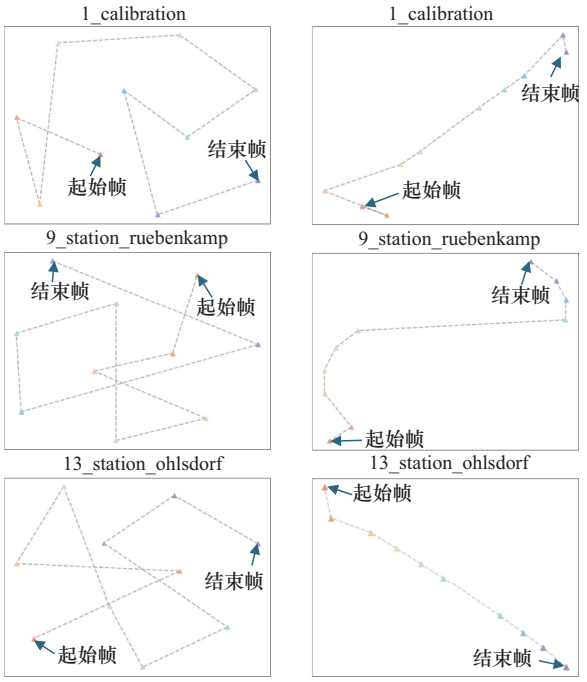
3.3.4　消融实验

本文比较了直接使用在ImageNet上预训练的ResNet-18模型与使用本文自监督方法训练后的模型在视频帧嵌入空间中的表现。自监督训练消融实验对比如图8所示，展示了3个视频序列的帧特征经过t-SNE降维后的二维可视化结果。在图8左列（使用预训练模型）中可以看到，未经自监督训练的通用模型所提取的特征，其在嵌入空间中的轨迹是混乱无序的。即使视频数据在像素层面是一种流形数据，其在特征空间中的分布也不能呈现流形。时间上相邻帧（由虚线连接）在空间中可能相距很远，整个序列的路径杂乱无章，无法体现视频连续性。经过本文方法训练后（图8右列），所有视频序列的帧嵌入点都形成了一条光滑、连续的流形轨迹。从起始帧到结束帧的虚线连接代表了视频序列在嵌入流形上的时序演化路径，可以看到轨迹的演进方向非常清晰。这直观地证明了提出的损失函数成功地将时间上

表7　本文方法与标准方法在计算开销与性能增益上的量化对比

算法	参数量$_{\downarrow}$/$\times10^6$	浮点运算次数$_{\downarrow}$/GFLOPS	BDmAP$_{\uparrow}$	BDrate(PSNR)$_{\uparrow}$	BDrate(SSIM)$_{\uparrow}$
本文方法	**~11.7**	**~1.82**	2.627 6%	−0.168 8%	**−0.047 1%**
标准方法	~36.9	~104.70	**2.668 1%**	**−0.132 9%**	−0.049 6%

连续的帧在嵌入空间中拉近，构建了一个能够准确反映内容演变的平滑流形结构。这个消融实验强有力地证明了自监督度量学习是本文方法成功的关键。正是这种对视频时序结构的有效建模，才使得后续基于流形等分点的聚类和重采样能够准确地找到内容变化的关键节点。



（a）使用预训练模型　（b）使用本文自监督训练

注：左列为使用通用预训练模型提取的特征，其在嵌入空间分布混乱；右列为使用本文方法训练后的特征，形成了光滑、连续的时序流形

图 8　自监督训练消融实验对比

4　结束语

本文提出了一种面向机器视觉的自监督视频时域重采样方法，旨在解决传统均匀采样在处理内容非线性变化的视频时面临的语义冗余与关键信息丢失问题。该方法通过自监督度量学习将视频帧序列映射至一个光滑的低维特征流形，并创新性地采用基于流形等分点的聚类算法进行自适应重采样，从而有效识别并保留了内容发生显著变化的语义关键帧。在公开数据集 OSDaR23 上的综合评估结果表明，本文方法在不降低下游任务精度的前提下，显著优化了码率–精度权衡性能，在 BDmAP 和 Pareto mAP 两项指标上均取得了明显提升。此外，实验也验证了该方法在多目标分割任务中的可行性与参考价值。特别地，本文方法在特征提取与聚类阶段的计算开销极低，赋予了其在实时处理场景中的巨大应用潜力，为未来机器视觉编码技术的落地提供了可行的技术路径。

参考文献：

[1] Guan Y C, Liao H C, Li Z N, et al. World models for autonomous driving: an initial survey[J]. IEEE Transactions on Intelligent Vehicles, 2024, 99: 1-17.

[2] Ji W, Xu J C, Qiao H X, et al. Visual IoT: enabling Internet of things visualization in smart cities[J]. IEEE Network, 2019, 33(2): 102-110.

[3] Dui H Y, Zhang S R, Liu M, et al. IoT-enabled real-time traffic monitoring and control management for intelligent transportation systems[J]. IEEE Internet of Things Journal, 2024, 11(9): 15842-15854.

[4] Kim A, Woo S T, Park M, et al. Deep learning-guided video compression for machine vision tasks[J]. EURASIP Journal on Image and Video Processing, 2024(1): 32.

[5] 熊皓萱, 徐媛媛, 朱琨. 面向机器视觉的 VVC 帧内编码算法[J]. 信号处理, 2025, 41(2): 350-358.
Xiong H X, Xu Y Y, Zhu K. VVC intra-coding scheme for machines[J]. Journal of Signal Processing, 2025, 41(2): 350-358.

[6] Ge X T, Luo J X, Zhang X J, et al. Task-aware encoder control for deep video compression[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 26036-26045.

[7] Schiappa M C, Rawat Y S, Shah M. Self-supervised learning for videos: a survey[J]. ACM Computing Surveys, 2023, 55 (13s): 1-37.

[8] Hua H, Tang Y L, Xu C L, et al. V2Xum-LLM: cross-modal video summarization with temporal prompt instruction tuning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2025, 39(4): 3599-3607.

[9] Zhao Y, Ye M, Ji L P, et al. Temporal adaptive learned surveillance video compression[J]. IEEE Transactions on Broadcasting, 2025, 71(1): 142-153.

[10] Tian J R, Lin Z X, Dai Y, et al. Keyframes selection from multi-scene videos for stress detection[J]. Information Processing & Management, 2025, 62(5): 104215.

[11] Zeng J H, Liang G, Ma Y X, et al. Pornographic video detection based on semantic and image enhancement[J]. The Computer Journal, 2024, 67(10): 3009-3019.

[12] Lee J, Hwang K I. YOLO with adaptive frame control for real-time object detection applications[J]. Multimedia Tools and Applications, 2022, 81(25): 36375-36396.

[13] Duan Y Q, Lu J W, Feng J J, et al. Deep localized metric learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(10): 2644-2656.

[14] Roth K, Milbich T, Sinha S, et al. Revisiting training strategies and generalization performance in deep metric learning[EB]. 2020.

[15] Zhang D Y, Li Y M, Zhang Z F. Deep metric learning with spherical embedding[EB]. 2020.

[16] Fu Z R, Li Y, Mao Z D, et al. Deep metric learning with self-supervised ranking[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2021, 35(2): 1370-1378.

[17] Jia X H, Han K, Zhu Y K, et al. Joint representation learning and novel category discovery on single- and multi-modal data[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 590-599.

[18] Wu Z R, Xiong Y J, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3733-3742.

[19] Tagiew R, Klasek P, Tilly R, et al. OSDaR23: open sensor data for rail 2023[C]//Proceedings of the 2023 8th International Conference on Robotics and Automation Engineering (ICRAE). Piscataway: IEEE Press, 2024: 270-276.

[20] Kay W, Carreira J, Simonyan K, et al. The Kinetics human action video dataset[EB]. 2017.

[21] Wieckowski A, Brandenburg J, Hinz T, et al. Vvenc: an open and optimized VVC encoder implementation[C]//Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). Piscataway: IEEE Press, 2021: 1-2.

[22] Wieckowski A, Hege G, Bartnik C, et al. Towards a live software decoder implementation for the upcoming versatile video coding (VVC) codec[C]//Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2020: 3124-3128.

[23] Zhang G Z, Zhu Y H, Wang H N, et al. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 5682-5692.

[24] Xue T F, Chen B A, Wu J J, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127(8): 1106-1125.

[25] Merz G, Liu Y C, Burke C J, et al. Detection, instance segmentation, and classification for astronomical surveys with deep learning (deepdisc): detectron2 implementation and demonstration with Hyper Suprime-Cam data[J]. Monthly Notices of the Royal Astronomical Society, 2023, 526(1): 1122-1137.

[26] Girshick R. Fast R-CNN[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2016: 1440-1448.

[27] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 2014 European Conference on Computer Vision (ECCV). Cham: Springer, 2014: 740-755.

[28] Perera A, Adzic V, Kalva H, et al. Comparative analysis of VCM and AhG8 for machine vision applications[EB]. 2025.

[29] Pont-Tuset J, Perazzi F, CAELLES S, et al. The 2017 Davis challenge on video object segmentation[EB]. 2017.

[30] Zhou C. Yolact++ better real-time instance segmentation[M]. Davis: University of California, Davis, 2020.

[作者简介]

刘建然（1994– ），男，中国科学院计算技术研究所博士生，中国计算机学会（CCF）会员，主要研究方向为机器视觉编码、图像处理、特征编码等。

纪雯（1976– ），女，博士，中国科学院计算技术研究所博士生导师，中国计算机学会（CCF）高级会员，主要研究方向为视觉处理器、多媒体系统、工业人工智能。

付哲（1992– ），男，交控科技股份有限公司高级工程师，主要研究方向为基于多传感器融合的轨道交通智能化平台及列车自主运行系统。