Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(2), 300-312
Published Online May 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.132032

Hans汉斯

基于时空采样的视频行为识别

王　冠，彭梦昊，陶应诚，徐　浩，景圣恩

合肥工业大学计算机与信息学院，安徽　合肥

收稿日期：2024年4月11日；录用日期：2024年5月17日；发布日期：2024年5月27日

摘　要

视频特征包含了行为执行时的时间、空间冗余信息。该信息和行为类别无关，会干扰行为识别，造成行为类别的错误判断。本文提出了一种基于时空采样的视频行为识别模型。模型包括关键帧采样和Token采样的视频Transformer。关键帧采样过程，通过量化相邻帧间的像素差异，识别出包含显著变化的关键帧，累积多个连续帧的更新概率处理两个关键帧间的可能存在的长时间间隔，引入一个可训练的采样概率阈值从而将更新概率二值化，增强对于关键帧的建模能力。因此该过程保证了视频关键信息的获取。本文认为不同的Token对识别任务的重要性会有所不同，因此在时空Transformer块中，本文采用一种数据依赖的Token采样策略，通过分层减少Token的数量有效降低空间冗余信息，同时也减少了模型计算量。最终通过全连接层完成视频行为识别。实验在ActivityNet-v1.3、Mini-Kinetics数据集上进行验证。实验表明，本文基于时空采样的视频行为识别方法，具有较小计算量的同时，能够达到现有行为识别方法的准确率。

关键词

视频行为识别，时空采样，视频Transformer

Video Action Recognition Based on Spatiotemporal Sampling

Guan Wang, Menghao Peng, Yingcheng Tao, Hao Xu, Sheng'en Jing

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

Received: Apr. 11th, 2024; accepted: May 17th, 2024; published: May 27th, 2024

Abstract

Video features contain the time and space redundancy information when the action is executed. This information has nothing to do with the action category, which will interfere with the action

文章引用: 王冠, 彭梦昊, 陶应诚, 徐浩, 景圣恩. 基于时空采样的视频行为识别[J]. 人工智能与机器人研究, 2024, 13(2): 300-312. DOI: 10.12677/airr.2024.132032

identification and cause the wrong judgment of the action category. This thesis proposes a video action recognition model based on spatiotemporal sampling. The model includes key frame sampling and Token sampling video Transformer. Key frame sampling, by quantifying the pixel difference between adjacent frames, identifies key frames with significant changes, accumulates the update probability of multiple consecutive frames, processes the possible long time interval between two key frames, introduces a trained sampling probability threshold to binarize the update probability, enhances the modeling ability of key frames, and ensures the acquisition of video key information. This thesis believes that different tokens have different importance to recognition tasks. Therefore, in the Transformer block, this thesis adopts a data-dependent Token sampling strategy to reduce the number of tokens by layers to effectively reduce spatial redundancy information and reduce the amount of computation. Finally, the video action recognition is completed through the fully-connected layer. The experiments are validated on ActivityNet-v1.3, Mini-Kinetics dataset. The experiments show that in this thesis, the action recognition method based on spatiotemporal sampling, can achieve the accuracy of existing action recognition methods with less computation.

Keywords

Video Action Recognition, Saptio-Temporal Sampling, Video Transformer

1. 引言

行为识别的目的是从视频帧序列中识别出目标正在执行的行为类别[1]，应用于人物交互、视频监控和视频检索等方面[2] [3]。

近年来，许多工作提出了有效的行为识别方法，基于 Transformer 的视频识别方法的一般做法是，将数据集中的每个视频均匀采样固定数量的视频帧，将视频帧划分为不重叠的 Token，通过 Transformer 块计算 Token 之间的自注意力学习特征信息，最后将学习所有特征信息的类别 Token 输入到多层感知机完成分类。但是均匀采样可能会错失关键信息同时引入干扰信息，而且视频 Transformer 的计算成本随着 Token 数量的增加呈平方级增长，甚至无法完成一些高空间分辨率或长视频的行为识别。因此，本文基于上述视频识别方法存在的问题，致力于探索新的方法，在提升视频分析性能的同时，有效降低计算成本和内存占用。

均匀采样作为一种常见的策略，能够满足视频行为识别的基本要求，然而当面对行为变化迅速的场景时，均匀采样的方法无法捕捉到行为的高速变化，因此会错失关键信息，影响到行为识别的准确性。均匀采样还可能引入与行为识别无关的干扰信息，进一步增加了分析的难度。因此，为了避免错失视频关键信息，同时降低干扰信息，需要更为精细和灵活的采样策略。本文针对视频片段设计了一种关键帧采样策略，通过量化相邻帧间像素的差异，识别出包含显著变化的关键帧。同时为了处理两个关键帧间的时间间隔，本文积累多个连续帧的更新概率。为了增强对于关键帧的建模能力，本文引入可训练的采样概率阈值，将更新概率二值化，保证了本方法对于视频关键信息的获取。

视频 Transformer 有很高的计算成本，在空间和时间上都是二阶复杂度。视频 Transformer 的输入尺寸非常大，即使使用图像块分辨率为 16 × 16 的 Token 化的采样方式，一张分辨率 224 × 224 的 RGB 图片，就会产生多达 196 个视觉 Token。为了缓解高计算成本问题，Timesformer [4]和 ViViT [5]分别沿时

间维度和时空维度进行全时空自注意力分解，以实现视频识别的准确性和效率之间的平衡。虽然已经优化了视频 Transformer 中的注意力计算，但还缺乏基于数据特征自适应分配计算资源的方法，在保证识别性能的同时降低整体计算量，提高效率。本文认为不同的 Token 对检测任务的重要性可能会有所不同，有的 Token 在网络的前期阶段是有用的，但在后期阶段包含的有效信息较少。例如，在空的道路上检测车辆可能需要来自车辆本身的一些后期 Token，但只需要来自周围道路的几个前期 Token。因此可以通过关键 Token 采样达到降低计算成本的目的。本文在不引入额外网络的情况下，利用一种依赖于数据的 Token 采样策略来分层减少 Token 的数量，显著地降低识别所需的计算量。

综上，本文的主要内容如下：

1) 本文设计了基于时空采样的视频行为识别方法，主要包括帧间像素差异引导的关键帧采样和自适应 Token 采样的视频 Transformer。在关键帧采样部分，通过帧更新概率估计、累积更新概率和更新概率二值化，完成关键帧的采样。

2) 在自适应 Token 采样的视频 Transformer 部分，将 Token 采样模块集成到视频 Transformer 的时空注意力块中，完成分层 Token 采样，有效降低需要处理的 Token 数量，使用联合时空注意力学习特征信息，最终通过全连接层完成视频分类。

3) 在数据集 ActivityNet-v1.3 [6]、Mini-Kinetics [7]上的实验证明了，本文提出方法的有效性。

2. 相关工作

2.1. 时间采样

为了降低视频行为识别的计算成本，一些方法将帧选择问题定义为一种顺序决策任务。其核心思想在于，利用顺序决策机制来平衡计算效率与识别性能。在这种思想指导下，决策过程需要依赖先前累积的信息来指导后续步骤，确定下一个应观察的帧或判断是否终止选择过程。这种方法通过智能地选择关键帧来减少冗余计算，从而在保持识别准确率的同时降低整体计算负担。

FrameGlimpses [8]将视频帧编码为带有时序信息和特征信息的向量，作为循环网络的输入，循环网络在每个时间步确定是否停止检测，若停止检测则输出预测结果，若继续检测则输出需要检测的下一帧的位置。AdaFrame [9]认为简单行为使用一到两帧即可正确预测，而复杂行为需要更多帧才能正确预测，因此提出了一个记忆增强 LSTM (长短期记忆网络，Long Short-Term Memory)，它提供了上下文信息，用于自适应确定何时检测下一帧。同时使用强化学习计算每个时间点检测更多帧的预期奖励。ListenToLook [10]认为视频帧保存了视频片段的外观信息，音频保存了视频片段的动态信息，因此使用视频帧和对应的音频组成图像－音频对，代替直接对视频片段分析。同时设计了 LSTM 网络与视频帧、音频索引序列交互，确定下一个时间步要处理的视频帧与音频。FrameExit [11]基于提前退出机制，对于每个视频，遵循一个确定的时间维度采样策略，提取采样得到的帧的特征并使用累计特征池化模块聚合，训练门控模块根据当前帧特征和聚合特征，决定是否抛出一个信号来退出当前视频片段，门控模块主要由多层感知机组成，可以避免大量的计算成本。虽然这个策略可以避免复杂的计算，但是确定性采样对于前后变化不确定的视频并非是最优选择。

另有一些方法并行完成对于视频帧、片段的采样。该策略的核心思想在于，同时并独立地确定每一帧或片段应采取的行为，从而并行地获取最终的选择结果。这种方法通过并行化处理，显著提高了采样效率，使得视频识别过程中的计算资源得到更加有效的利用。同时，并行采样策略也有助于减少由于顺序决策导致的潜在延迟，进一步提升了视频识别的实时性能。

SCSampler [12]将视频划分为之指定长度的片段，使用轻量级的权重网络估计每个指定长度视频片段的显著性得分，用得分前 K 的片段训练识别模型。DSN [13]将视频分为同等长度的片段，在每个视频片

段内并行地动态采样一个鉴别帧，同时保证权值共享，使用识别模块处理每个采样出的视频帧完成识别。AR-Net [14]对于给定的一个视频帧，使用一个策略网络来决定行为识别模型应该使用什么输入分辨率来处理，然后使用主干网络处理重新缩放后的帧，以生成预测。VideoIQ [15]训练一个与识别网络并行的轻量级网络，对给定的视频片段，将每一帧依次输入网络产生一个动态策略，表明每帧在识别视频时使用的数值精度。

2.2. 空间采样

K 中心图像块采样方法[16]将一个视频片段预处理成一组图像块，并将每个图像块映射到一个具有固定位置编码的向量中。然后通过最远点采样算法采样 K 个图像块，将采样出的图像块 Token 化，作为 Transformer 的输入。Xie [17]等提出一种随机采样插值网络，基于输入特征映射生成稀疏采样掩模，然后只计算在采样点上的特征，形成稀疏特征映射。通过插值模块对未采样点的特征进行插值，形成输出特征映射。STTS [18]将动态 Token 选择视为排序问题：将视频嵌入为 Token 序列后，采用一个轻量化的评分网络估计出每个输入 Token 的分数，然后选择前 K 个得分最高的 Token 计算自注意力进而完成识别。在时间维度上，文章保留了与动作类别最相关的帧，而在空间维度上，文章识别特征图中最具区分性的区域，保证了不会影响视频 Transformer 中以分层方式使用的空间上下文。

2.3. 高效注意力机制

一些基于 Transformer 的视频行为识别方法设计了更加高效的注意力机制。TubeViT [19]认为稀疏性对视频是有效的，对图像 Token 化的过程中，将二维卷积的时间步幅设置为 16 帧，同时为了防止处理短视频时可能出现的信息丢失，文章设计了不同形状的稀疏 tube 用于采样，文章设计了图像-视频联合训练方法，加入位置编码后同时使用 ViT 计算自注意力。X-ViT [20]将时间注意力限制于局部的时间窗口，利用 Transformer 的深度来获取视频序列的全时间覆盖，使用有效的时空混合来联合建模空间和时间信息，与纯空间模型相比没有增加任何额外计算成本。Timesformer [4]将帧分割为不重叠的图像块，通过线性嵌入层转化为向量，先使用时间注意力编码，进行残差连接后使用空间注意力编码，再次进行残差连接后，使用多层感知机完成行为识别。WLiT [21]将特征映射沿着空间维度划分为多个窗口，并分别计算窗口内部的注意力，同时沿通道维度计算线性注意，使模型能够捕获完整的时空信息。

3. 基于时空采样的视频行为识别

3.1. 关键帧采样

本章方法如图 1 所示，其中关键帧采样包括帧更新概率估计，更新概率累积，更新概率二值化。本章模型的输入是视频片段 $Clip=\{v_t\}$，$Clip\in\mathbb{R}^{T\times H\times W\times C}$，其中 T 是视频片段帧数，t 是帧索引，$t\in[1,T]$，$H\times W$ 是空间分辨率，C 是通道数。

对于视频帧 v_t，预处理将其转化为像素矩阵 $pixel_t$，$pixel_t\in\mathbb{R}^{H\times W\times C}$。像素矩阵是一个数组，其中每个元素代表图像中对应位置的像素值。计算相邻两个视频帧像素矩阵 $pixel_t$ 和 $pixel_{t+1}$ 的帧差 $d_{t,t+1}=\left\|pixel_{t+1}-pixel_t\right\|_2$，使用高斯分布估计帧差的分布 $d_{t,t+1}\sim\mathrm{N}\left(\mu,\sigma^2\right)$，其中，分布的均值 μ 和方差 σ^2 分别为：

$$\mu=\frac{1}{T-1}\sum_{i}^{T-1}d_{t,t+1} \tag{1}$$

$$\sigma^2=\frac{1}{T-1}\sum_{i}^{T-1}\left(d_{t,t+1}-\mu\right)^2 \tag{2}$$

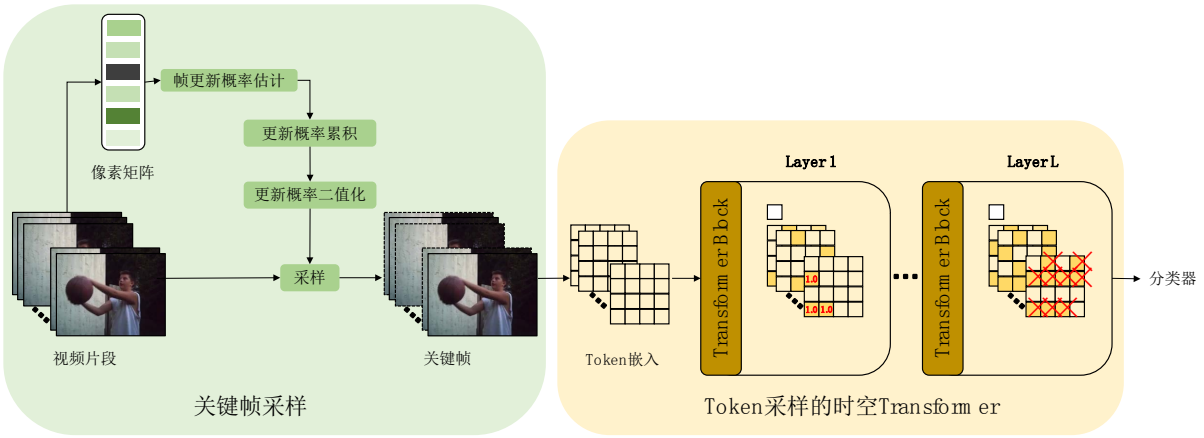

Figure 1. Video action recognition model based on spatiotemporal sampling
图 1. 基于时空采样的视频行为识别模型

帧差概率 p_{t+1}^{diff} 表示已知第 t 帧的条件下，根据第 t 帧与第 $t+1$ 帧间的像素矩阵的差值得到的第 $t+1$ 帧采样的概率：

$$p_{t+1}^{diff} = \frac{1}{1+\exp\left(\mu+\sigma-d_{t,t+1}\right)} \tag{3}$$

本文认为当选中一个视频帧作为关键帧之后，短期内选中下一个关键帧的概率较小。相应的，当长期未选中关键帧，接下来的视频帧作为关键帧的概率逐渐增大。因此本文通过沿帧序列的更新概率累积机制来估计多个帧的更新概率。当一个视频帧被选中为关键帧时，下一帧的累积概率用帧差概率表示。当一个视频帧被认为是非关键帧时，下一帧的累积概率用当前帧的更新概率与下一帧的帧差概率相加之和表示。累积概率 p_{t+1}^{sum} 考虑了两个关键帧之间的时间间隔内的所有帧。累积概率公式为：

$$p_{t+1}^{sum}\begin{cases} p_{t+1}^{diff} & \text{if } b_t = 1 \\ \min\left(p_t^{sum}+p_{t+1}^{diff},1\right) & \text{if } b_t = 0 \end{cases} \tag{4}$$

其中，$b_t=1$ 表示采样第 t 帧作为关键帧，$b_t=0$ 表示不采样第 t 帧作为关键帧。

不同目标执行相同行为时，存在个体差异性。这种差异性不仅体现在行为执行的细节上，更增加了关键帧选取的难度。因此即使是相同的行为，其关键帧也可能存在很大的变化。为了应对这种变化，本文引入了可训练的采样阈值 $threshold \in (0,1)$，当帧的更新概率累积大于采样阈值，则将其视为关键帧。为了解决存在变化较大的关键帧造成的选择不确定问题，本文通过确定采样阈值，将更新概率二值化：

$$b_t = \begin{cases} 1 & \text{if } p_t^{sum} \geq threshold \\ 0 & \text{if } p_t^{sum} \leq threshold \end{cases} \tag{5}$$

处理视频片段 *Clip* 内的全部视频帧获得关键帧集合 $frame = \{frame_t\}$，$frame \in \mathbb{R}^{T_{key}\times H\times W\times C}$，$T_{key}$ 是关键帧帧数。关键帧采样从原始视频片段中筛选出最具代表性的帧，确保在降低计算复杂度的同时，尽可能保留了行为的关键信息。

3.2. Token 采样的时空 Transformer

对于 3.2 获得的关键帧集合 *frame* 中的视频帧 $frame_t \in \mathbb{R}^{H\times W\times C}$，本文将其嵌入为二维 Token 序列

$X=\{x_k\}$，$X\in\mathbb{R}^{K\times D}$，$(P,P)$是图像块的空间分辨率，$K=HW/P^2$是获得的 Token 数量，即 Transformer 的输入序列长度，k 表示 Token 索引，$k\in[1,K]$，D 是嵌入维度。

本文使用可学习的矩阵 $E\in\mathbb{R}^{D\times 3P^2}$ 将每个 Token 线性嵌入为向量，使用可学习的位置嵌入 $e_k^{pos}\in\mathbb{R}^D$ 表示每个 Token 的时空位置，结合了视频帧的局部特征和全局空间关系，能够保证后续对视频帧内容的理解，可以得到：

$$x_k^{(l)}=Ex_k+e_k^{pos} \tag{6}$$

$x_k^{(l)}\in\mathbb{R}^D$ 其中 $x_k^{(l)}$ 的上标 $l=0,1,\cdots,L$ 表示经过第 l 层后的 Token。本文在序列的前面添加一个可学习向量 $x_c^{(l)}$ 作为类别 Token，经过所有 Transformer 层之后，使用线性层处理此类别 Token 得到视频类别。至此，获得序列作为 Transformer 的输入。

输入视频帧经过嵌入编码得到 Token 序列后，送入图 2 所示的 Transformer 块进行自注意力的计算。一个完整的 Transformer 网络由一系列 Transformer 块构成，输入的序列长度在网络整个阶段不会发生改变，并且 Token 的特征维度也保持不变。

本文的 Transformer 结构包括个 L 编码块，将处理好的 $Z^{(0)}$ 输入到 Transformer 编码块中，在每个编码块 l 计算 $query_k^{(l,a)}$，$key_k^{(l,a)}$，$value_k^{(l,a)}$，公式如下：

$$query_k^{(l,a)}=W_Q^{(l,a)}\mathrm{LN}\left(x_k^{(l-1)}\right)\in\mathbb{R}^{D_h} \tag{7}$$

$$key_k^{(l,a)}=W_K^{(l,a)}\mathrm{LN}\left(x_k^{(l-1)}\right)\in\mathbb{R}^{D_h} \tag{8}$$

$$value_k^{(l,a)}=W_V^{(l,a)}\mathrm{LN}\left(x_k^{(l-1)}\right)\in\mathbb{R}^{D_h} \tag{9}$$

其中，$query_k^{(l,a)}$ 表示在第 a 个注意力头经过第 l 层 Transformer 块的 Token $x_k^{(l-1)}$ 的查询(*query*)表示，$key_k^{(l,a)}$ 是其键(*key*)表示，$value_k^{(l,a)}$ 是其值(*value*)表示。$a\in[1,A]$ 是注意力头的索引，A 表示注意力头的总数，$W_Q^{(l,a)}$ 表示第 a 个注意力头在第 l 层线性变换的参数矩阵，LN 表示层归一化[22]，$D_h=D/A$ 表示每个注意力头的维度。

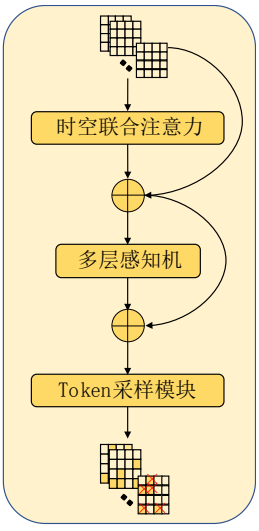

Figure 2. Transformer block structure with Token sampling
图 2. 集成 Token 采样的 Transformer 块结构

本文在每个注意力头，使用点积计算时空联合自注意力，经过 Softmax 激活函数，获得每个 query 的自注意力权重，每个层的自注意力计算如下：

$$\alpha_k^{(l,a)} = \text{Softmax}\left(\frac{query_k^{(l,a)^\top} key_k^{(l,a)}}{\sqrt{D_h}}\right) \tag{10}$$

将得到的权重值与 *value* 的值相乘求和，得到每个头部的注意力向量：

$$s_k^{(l,a)} = \sum_{k=0}^{K} \alpha_k^{(l,a)} value_k^{(l,a)} \tag{11}$$

将所有注意力头获得的关联信息结果拼接，通过线性层 W_O 后，与第 $l-1$ 个编码器的输出相加：

$$x_k'^{(l)} = W_O \begin{bmatrix} s_k^{(l,1)} \\ \vdots \\ s_k^{(l,A)} \end{bmatrix} + x_k^{(l-1)} \tag{12}$$

通过 MLP (多层感知机，Multilayer Perceptron)处理层归一化 LN 计算得到的值，再和上述得到的结果使用残差连接，得到最后的输出 Token：

$$x_k^{(l)} = \text{MLP}\left(\text{LN}\left(x_k'^{(l)}\right)\right) + x_k'^{(l)} \tag{13}$$

本文在上述 MLP 操作后引入一个依赖于输入的 Token 采样模块，计算获得在第 l 层的 Token 不再向后续 Transformer 层传播的丢弃概率：

$$h_k^{(l)} = \sigma\left(\gamma \cdot x_{k,d}^{(l)} + \beta\right) \tag{14}$$

其中，$x_{k,d}^{(l)}$ 表示第 l 层 Token $x_k^{(l)}$ 的第 d 个维度，嵌入维度 D 只有单个维度用于丢弃概率计算，本文使用 $d=0$ (第一维度)完成计算，$\sigma(\cdot)$ 表示 Sigmoid 函数，γ 和 β 是用于调整嵌入的移动和缩放参数。调整分数值的大小，从而影响 Token 减少的程度和视频识别的性能。这两个标量参数在所有 Token 的所有层中共享。丢弃概率用于指导 Token 传播操作，从而逐步减少 Token 的数量。

在第 N_k 层，当 Token 的累积丢弃概率超过 $1-\epsilon$ 时，表示此 Token 的计算在此结束，并且它将不会传播到后续的层，ϵ 是一个正常数用于允许 Token 在第一层之后被丢弃：

$$N_k = \arg\min_{n \le L} \sum_{l=1}^{n} h_k^{(l)} \ge 1-\epsilon \tag{15}$$

本文的 Token 采样在推理时，只需从计算中删除已丢弃的 Token，以衡量本文的 Token 采样机制所获得的实际加速。

为了跟踪跨层的丢弃概率的进展，本文计算每个 Token 的剩余丢弃分数 r_k 为：

$$r_k = 1 - \sum_{l=1}^{N_k-1} h_k^{(l)} \tag{16}$$

为了使模型学习一个合理的 Token 采样机制，本文设计一个衡量损失：

$$L_p = \frac{1}{K}\sum_{k=1}^{K}\left(N_k + r_k\right) \tag{17}$$

至此，本文完成对于每一层 Token 的采样。本文首先基于所有层中的丢弃概率分数，对类别 Token $x_c^{(l)}$ 的加权求和得到 x_o，然后将 x_o 输入后续的视频识别分类器 C 输出得到视频行为识别结果，其中分类器 C 是全连接层。

行为识别的损失函数为 L_{task}：

$$L_{task} = L_{CE}\left(C\left(x_o\right)\right) \tag{18}$$

$$x_o = \sum_{l=1}^{N_c-1} h_c^{(l)} x_c^{(l)} + r_c x_c^{(N_c)} \tag{19}$$

$L_{CE}(\cdot)$ 是分类的交叉熵损失，衡量模型对于视频行为类别预测与真实标签之间的不一致程度。通过最小化这一损失，模型能够提升将输入数据正确映射到相应行为类别的能力。模型的总损失函数为：

$$L_{All} = L_{task} + \alpha_p L_p \tag{20}$$

其中，α_p 为控制衡量损失的超参数。

4. 实验结果与分析

4.1. 数据集与评价指标

本章的实验 ActivityNet-v1.3 [6]和 Mini-Kinetics [7]两个数据集上进行。

ActivityNet-v1.3 该数据集是一个用于视频活动识别和检测的大规模数据集，涵盖了丰富多样的活动类别，如运动、日常生活、社交互动等，每个视频片段都被标注了相应的活动类别和时间段。ActivityNet-v1.3 数据集是由带有 200 种类别标签的 10,024 个训练视频和 4926 个测试视频组成的，视频平均时长 117 秒。

Mini-Kinetics 该数据集包含从 Kinetics 数据集中随机选择的 200 个类，131,082 个视频。视频平均时长为 10 秒。本文使用 121,215 个视频进行训练，使用 9867 个视频进行测试。

为了全面评估模型的准确性，本文针对不同数据集采取了不同的评估指标。针对 ActivityNet-v1.3 数据集，本文使用 mAP [23] (全类平均正确率)作为识别准确性评估指标。这一指标是通过对所有类别的检测平均正确率进行加权平均得出的，它能够有效地反映出模型在识别各类活动时的整体性能。对于 Mini-Kinetics 数据集，本文使用 Top1 准确率作为评估标准。Top1 准确率是指模型预测结果中，概率最高的预测标签与实际标签相符的比例。此外，为了评估模型在视频识别任务中的计算效率，本文还计算了视频级 GFLOPs 表示处理单个视频所需的平均计算成本。

4.2. 实验细节

本章提出了基于帧间差异来估计帧更新概率选择关键帧的方法，所以对于两个数据集的处理使用关键帧采样从视频片段中采样 10 帧。本文所有实验使用的硬件配置为 Intel Core i7-5960X、CPU 3GHz 8cores RAM 8 GB、图像显卡为 2 张 NVIDIA GeForce GTX 2080Ti、Linux18.04 操作系统。软件框架使用 Pytorch 深度学习框架。在训练过程中，本文随机裁剪图像为 224 × 224，随机翻转图像进行增强。在推理过程中，将所有帧的大小调整为 256 × 256，中心裁剪为 224 × 224。在 Token 采样模块，本文将 γ 设置为 10，β 设置为 10，衡量损失函数的超参数 α_p 设置为 5×10^{-4}。本文使用 Adam (Adaptive Moment Estimation)优化器，其中使用余弦学习速率，初始学习率设置为 1×10^{-5}。

4.3. 对比实验

在本节中，将本文的方法与不同数据集上现有的先进方法进行比较。

AvtivityNet-v1.3 数据集。在表 1 将本文提出的方法与 AvtivityNet-v1.3 数据集上的其他现有方法进行比较，由于 AvtivityNet-v1.3 数据集的视频持续时间很长，存在大量冗余信息，因此本文采取的关键帧采样的方法通过时间去冗余，与数据集特性契合。同时本文集成 Token 采样的时空 Transformer 方法，能够

在利用注意力的同时，去除空间冗余信息。实验结果表明，本文所提出的方法在平均精度均值(mAP)分数优于表 1 中其他先进方法。具体而言，在 mAP 这一评估指标上，本文方法与 TSQNet [25]相比，取得了 1.2% (76.6%→77.8%)的精度提升，同时视频级 GFLOPs 降低了 0.2 (26.1→25.9)，体现了本文方法在提升行为识别准确性方面的有效性，在保持高效计算性能方面的优越性。

Table 1. Comparison of different methods on ActivityNet-v1.3 dataset
表 1. 不同方法在 ActivityNet-v1.3 数据集上的对比实验

方法	Backbones	mAP (%)	GFLOPs
FrameGlimpses [8]	VGG	60.2	32.9
AdaFrame [9]	ResNet101	71.5	79.0
LiteEval [24]	MobileNetV2+ResNet101	72.7	95.1
ListenToLook [10]	MobileNetV2+ResNet50	72.3	81.4
SCSampler [12]	MobileNetV2+ResNet50	72.9	42.0
AR-Net [14]	MobileNetV2+ResNet50	73.8	33.5
FrameExit [11]	ResNet50	76.1	26.1
TSQNet [25]	ResNet50	76.6	26.1
本文方法	TimeSformer	77.8	25.9

Mini-Kinetics 数据集。在表 2 将本文提出的方法与 Mini-Kinetics 数据集上的其他现有方法进行比较。从数据集特点上来看，Mini-Kinetics 数据集的行为类型与场景高度相关，网络模型可能仅从视频帧空间背景的外观特征就可以推断出行为类型，视频帧存在空间冗余，本文的 Token 采样方法可以去除空间冗余。与另一种先进的方法 D-STEP [26]相比，本文的方法利用时间和空间采样，取得了 6.5% (65.4%→73.9%)的 Top1 分数提升。与表 2 中其他方法相比，本文方法在 Top1 分数和视频级 GFLOPs 两个关键指标都取得领先。作为基于 Transformer 的行为识别模型，本文方法在保证高效的情况下，取得了最佳识别结果。

Table 2. Comparison of different methods on Mini-Kinetics dataset
表 2. 不同方法在 Mini-Kinetics 数据集上的对比实验

方法	Backbones	Top1 (%)	GFLOPs
LiteEval [24]	MobileNetV2 + ResNet101	61.0	99.0
SCSampler [12]	MobileNetV2 + ResNet50	70.8	42.0
AR-Net [14]	MobileNetV2 + ResNet50	71.7	32.0
D-STEP [26]	ResNet50	65.4	12.4
本文方法	TimeSformer	73.9	25.9

4.4. 消融实验

在本节，本文在 ActivityNet-v1.3 数据集上进行消融实验分析，从而验证本文方法的有效性。首先本文验证所提出的关键帧采样方法对模型性能的影响。表 3 给出了，采用均匀采样和关键帧采样两种方式对模型识别准确率以及计算量的影响，为确保对比的公正性，本文将均匀采样与关键帧采样数量均设置为 10 帧。通过分析表 3 中的数据，可以观察到关键帧采样方式在仅增加了 0.8 GFLOPs 的情况下，模型的平均精均值(mAP)增加了 2.4%。这一提升体现了关键帧采样方法相较于均匀采样的优越性。

Table 3. The effect of key frames sampling on model performance
表 3. 关键帧采样对模型性能的影响

采样方式	mAP (%)	Δ (差值)	GFLOPs
均匀采样	75.4	0	24.5
关键帧采样	77.8	+2.4	25.9

其次本文验证关键帧采样数量对模型性能的影响。为此，本文在表 4 中对比了采样 4 帧、8 帧、10 帧、16 帧的情况下，模型的识别准确率以及计算量。通过分析表 4 中的数据，可以观察到随着采样帧数的增加，模型的识别准确率也呈现上升趋势，但准确率会达到上限。具体来说，本文将采样帧数从 4 帧增加到 8 帧，mAP 上涨了 6.2%；采样帧数从 8 帧增加到 10 帧，mAP 上涨了 2.4%；但是采样帧数从 10 帧增加到 16 帧，mAP 保持不变，同时计算量增加了 8.9 GFLOPs，模型处理更多帧数据却并未获得性能提升，这说明模型性能达到饱和，因此为了达到模型识别准确率和效率间的平衡，本文将重要性采样帧数设置为 10。

Table 4. The effect of the quantity of key frames sampling on model performance
表 4. 关键帧采样数量对模型性能的影响

关键帧采样数量	4	8	10	16
mAP (%)	69.2	75.4	77.8	77.8
GFLOPs	15.0	21.6	25.9	34.8

最后本文验证在嵌入向量中使用单个维度来计算和表示丢弃概率分数对模型性能的影响。如表 5 所示，与使用完整向量相比，使用向量的第一个维度计算丢弃概率，模型的识别 mAP 分数仅下降了 0.1%。因此在本文的模型中，选择嵌入向量中的第一个元素，并使用它来进行丢弃概率计算。

Table 5. The effect of calculating discard probabilities using a single dimension of the vector on model performance
表 5. 使用向量单个维度计算丢弃概率对模型性能的影响

计算方式	mAP (%)	GFLOPs
使用向量第一个维度	77.8	25.9
使用完整向量	77.9	29.6

4.5. 可视化结果

图 3 将本文提出的方法时空采样 Token 的结果可视化。这里展示了本文方法对 Mini-Kinetics 数据集视频片段的处理。可以观察到，图 3 是一个滑雪样本的视频序列，第一行是本文对原始视频帧使用关键帧采样获得的关键帧序列，获得了视频片段中最具信息量的视频帧。在第二、第三行，经过时空联合注意力块中的 Token 采样模块，冗余 Token 被逐层丢弃，白色区域表示被丢弃的 Token。

同时可以观察到，在不同的帧之间，被移除的 Token 数量会自适应地变化。例如，第一帧相对于其他帧有更多的冗余 Token 被丢弃，主要归因于第一帧包含更为丰富的背景信息，这体现了本文方法处理不同帧的灵活性。

从第二行可以观察到，在本文方法中，浅层注意力已经可以快速定位关键信息区域，过滤掉复杂的背景干扰。结合第二行、第三行可以观察到本文方法在丢弃冗余 Token 的同时，能够有效保留运动目标的语义信息，确保目标的空间结构不被破坏。

Figure 3. Visualization of spatial-temporal sampling
图 3. 时空采样可视化

5. 总结与展望

本文首先分析了现有的基于 Transformer 的视频行为识别方法的缺陷，均匀采样视频帧可能会错失关键信息同时引入干扰信息，而且视频 Transformer 的计算成本随着 Token 数量的增加呈平方级增长，甚至无法完成一些高空间分辨率或长视频的行为识别。本文认为不同的 Token 对检测任务的重要性可能会有所不同，有的 Token 在网络的早期阶段是有用的，但在后期阶段包含的有效信息较少。针对上述问题，本文提出一种基于时空采样的行为识别模型。本文针对视频片段设计了一种关键帧采样策略，通过对相邻帧间像素的差异量化，识别出包含显著变化的关键帧。同时为了处理两个关键帧间的时间间隔，本文积累多个连续帧的更新概率。为了增强对于关键帧的建模能力，利用可训练的采样阈值，将更新概率二值化，保证了本方法对于视频关键信息的获取。在时空联合 Transformer 块中，本文在不引入额外网络的情况下，利用一种依赖于数据的 Token 采样策略来分层减少 Token 的数量，显著地降低识别所需的计算量。最终本文使用全连接层作为分类器完成视频级的行为识别。在 ActivityNet-v1.3 数据集和 Mini-Kinetics 数据集上进行实验分析，本文的方法能够获取视频关键信息，具有较小计算量的同时，能够达到现有行为识别方法的准确率。

参考文献

[1] Karpathy, A., Toderici, G., Shetty, S., *et al*. (2014) Large-Scale Video Classification with Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1725-1732. https://doi.org/10.1109/CVPR.2014.223

[2] Goyal, R., Ebrahimi Kahou, S., Michalski, V., *et al*. (2017) The “Something Something” Video Database for Learning and Evaluating Visual Common Sense. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 5842-5850. https://doi.org/10.1109/ICCV.2017.622

[3] Chen, J., Li, K., Deng, Q., *et al*. (2019) Distributed Deep Learning Model for Intelligent Video Surveillance Systems with Edge Computing. *IEEE Transactions on Industrial Informatics*. https://doi.org/10.1109/TII.2019.2909473

[4] Bertasius, G., Wang, H. and Torresani, L. (2021) Is Space-Time Attention All You Need for Video Understanding? *The* 38*th International Conference on Machine Learning* (*ICML* 2021), 18-24 July 2021, 1-12.

[5] Arnab, A., Dehghani, M., Heigold, G., *et al*. (2021) Vivit: A Video Vision Transformer. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 11-17 October 2021, 6836-6846. https://doi.org/10.1109/ICCV48922.2021.00676

[6] Caba Heilbron, F., Escorcia, V., Ghanem, B. and Carlos Niebles, J. (2015) Activitynet: A Large-Scale Video Benchmark for Human Activity Understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 961-970. https://doi.org/10.1109/CVPR.2015.7298698

[7] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S. and Zisserman, A. (2017) The Kinetics Human Action Video Dataset.

[8] Yeung, S., Russakovsky, O., Mori, G. and Fei-Fei, L. (2016) End-to-End Learning of Action Detection from Frame Glimpses in Videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2678-2687. https://doi.org/10.1109/CVPR.2016.293

[9] Wu, Z., Xiong, C., Ma, C.Y., Socher, R. and Davis, L.S. (2019) Adaframe: Adaptive Frame Selection for Fast Video Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 1278-1287. https://doi.org/10.1109/CVPR.2019.00137

[10] Gao, R., Oh, T.H., Grauman, K. and Torresani, L. (2020) Listen to Look: Action Recognition by Previewing Audio. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 14-19 June 2020, 10457-10467. https://doi.org/10.1109/CVPR42600.2020.01047

[11] Ghodrati, A., Bejnordi, B.E. and Habibian, A. (2021) Frameexit: Conditional Early Exiting for Efficient Video Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 15608-15618. https://doi.org/10.1109/CVPR46437.2021.01535

[12] Korbar, B., Tran, D. and Torresani, L. (2019) Scsampler: Sampling Salient Clips from Video for Efficient Action Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, 27-28 October 2019, 6232-6242. https://doi.org/10.1109/ICCV.2019.00633

[13] Zheng, Y.D., Liu, Z., Lu, T. and Wang, L. (2020) Dynamic Sampling Networks for Efficient Action Recognition in Videos. *IEEE Transactions on Image Processing*, **29**, 7970-7983. https://doi.org/10.1109/TIP.2020.3007826

[14] Meng, Y., Lin, C.C., Panda, R., Sattigeri, P., Karlinsky, L., Oliva, A., Feris, R., *et al*. (2020) Ar-Net: Adaptive Frame Resolution for Efficient Action Recognition. *Computer Vision-ECCV* 2020: 16*th European Conference*, Glasgow, 23-28 August 2020, 86-104. https://doi.org/10.1007/978-3-030-58571-6_6

[15] Sun, X., Panda, R., Chen, C.F.R., Oliva, A., Feris, R. and Saenko, K. (2021) Dynamic Network Quantization for Efficient Video Inference. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11-17 October 2021, 7375-7385. https://doi.org/10.1109/ICCV48922.2021.00728

[16] Park, S.H., Tack, J., Heo, B., Ha, J.W. and Shin, J. (2022) K-Centered Patch Sampling for Efficient Video Recognition. In: *European Conference on Computer Vision*, Springer, Cham, 160-176. https://doi.org/10.1007/978-3-031-19833-5_10

[17] Xie, Z., Zhang, Z., Zhu, X., Huang, G. and Lin, S. (2020) Spatially Adaptive Inference with Stochastic Feature Sampling and Interpolation. *Computer Vision-ECCV* 2020: 16*th European Conference*, Glasgow, 23-28 August 2020, 531-548. https://doi.org/10.1007/978-3-030-58452-8_31

[18] Wang, J., Yang, X., Li, H., Liu, L., Wu, Z. and Jiang, Y.G. (2022) Efficient Video Transformers with Spatial-Temporal Token Selection. In: *European Conference on Computer Vision*, Springer, Cham, 69-86. https://doi.org/10.1007/978-3-031-19833-5_5

[19] Piergiovanni, A.J., Kuo, W. and Angelova, A. (2023) Rethinking Video Vits: Sparse Video Tubes for Joint Image and Video Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 2214-2224. https://doi.org/10.1109/CVPR52729.2023.00220

[20] Bulat, A., Perez Rua, J.M., Sudhakaran, S., Martinez, B. and Tzimiropoulos, G. (2021) Space-Time Mixing Attention for Video Transformer. *Advances in Neural Information Processing Systems*, **34**, 19594-19607.

[21] Sun, R., Zhang, T., Wan, Y., Zhang, F. and Wei, J. (2023) Wlit: Windows and Linear Transformer for Video Action Recognition. *Sensors*, **23**, Article No. 1616. https://doi.org/10.3390/s23031616

[22] Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016) Layer Normalization.

[23] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J. and Tian, Q. (2015) Scalable Person Re-Identification: A Benchmark. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1116-1124. https://doi.org/10.1109/ICCV.2015.133

[24] Wu, Z., Xiong, C., Jiang, Y.G. and Davis, L.S. (2019) Liteeval: A Coarse-to-Fine Framework for Resource Efficient Video Recognition. 33*rd Conference on Neural Information Processing Systems* (*NeurIPS* 2019), Vancouver, 8-14 December 2019, 1-10.

[25] Xia, B., Wang, Z., Wu, W., Wang, H. and Han, J. (2022) Temporal Saliency Query Network for Efficient Video Recognition. In: *European Conference on Computer Vision*, Springer, Cham, 741-759. https://doi.org/10.1007/978-3-031-19830-4_42

[26] Raviv, A., Dinai, Y., Drozdov, I., Zehngut, N., Goldin, I. and Center, S.I.R.D. (2022) D-Step: Dynamic Spatio-Temporal Pruning. *Proceedings of the British Machine Vision Conference*, London, 21-24 November 2022, 1-13.