融合时空切片和双注意力机制的视频摘要方法

张云佐，郭亚宁，李文博
（石家庄铁道大学信息科学与技术学院，050043，石家庄）

摘要：为解决现有视频摘要方法的视频帧特征信息提取不充分、摘要结果过分依赖单一特征的问题，提出了一种融合时空切片和双注意力机制的视频摘要方法。在原视频的精准分段阶段，提出了基于时空切片的核时序分割算法(STS-KTS)，将视频场景信息反映为时空切片纹理信息，采用水平映射法将预处理后的时空切片投影为一维数组，作为 KTS 的输入特征；以双注意力机制和分组卷积为基本组件，结合 BiLSTM 构建时空特征提取网络，以快速提取丰富的时空特征信息，从而配合纹理特征信息消除现有摘要模型对单一特征的过分依赖；采用帧参数预测模块获取最佳的视频帧贡献度分数、中心度分数以及帧序列位置；将帧分数转化为镜头分数，以选取内容丰富的片段，进而生成动态视频摘要。在 SumMe 和 TVSum 数据集上的实验表明：所提方法能提高生成摘要的准确性，比现有方法性能更高，尤其在 SumMe 数据集上的生成摘要准确性相比于现有方法提升了 0.58%。

关键词：视频摘要；时空切片；双注意力机制；时空特征提取；深度学习
中图分类号：TP391 **文献标志码**：A
DOI：10.7652/xjtuxb202212013 **文章编号**：0253-987X(2022)12-0127-09

Video Summarization Method Based on Spatiotemporal Slice and Dual Attention Mechanism

ZHANG Yunzuo, GUO Yaning, LI Wenbo
(School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China)

Abstract: In order to solve the problems of insufficient extraction of video frame feature information and excessive dependence on a single feature in the existing summarization methods, a video summarization method based on spatiotemporal slice and dual attention mechanism is proposed. Firstly, a kernel temporal segmentation algorithm based on the spatiotemporal slice (STS-KTS) is proposed for the accurate segmentation of the original video. The algorithm reflects the video scene information as spatiotemporal slice texture information, and uses the horizontal mapping method to project the preprocessed spatiotemporal slice into a one-dimensional array as the input feature of KTS. At the same time, taking the dual attention mechanism and grouping convolution as the basic components, combined with BiLSTM, the spatiotemporal feature extraction network is constructed to quickly extract rich spatiotemporal feature information, so as to eliminate the excessive dependence of the existing summarization model on a single feature with the texture feature information. Then, the frame parameter prediction module is used to obtain the best video frame contribution score, center score and frame sequence

收稿日期：2022-05-08。 作者简介：张云佐(1984—)，男，副教授，博士生导师。 基金项目：中央引导地方科技发展资金资助项目(226Z0501G)；河北省自然科学基金资助项目(F2022210007)；河北省高等学校科学技术研究项目(ZD2022100)。
网络出版时间：2022-08-25 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20220824.1634.002.html

position. Finally, the frame scores are converted into shot scores to select content-rich segments and generate dynamic video summarization. Experimental results on SumMe and TVSum datasets demonstrate that the proposed method can improve the accuracy of generating summarization and achieve better performance than the existing methods, in particular the accuracy of generating summarization on SumMe dataset is improved by 0.58% compared to the existing methods.

Keywords: video summarization; spatiotemporal slice; dual attention mechanism; spatiotemporal feature extraction; deep learning

随着移动电话、无人机、监控摄像头等具备视频拍摄、存储功能的设备的广泛使用,视频数据量呈现出爆炸式增长的趋势。如何以直观且高效的方式查阅视频数据,快速完成对视频内容的大致了解,成为计算机视觉领域的研究热点。视频摘要能以简洁的形式呈现长时视频中的有价值内容,为海量视频存储和检索提供了有效解决方案[1]。

早期的无监督视频摘要方法[1-3]大多基于聚类、图模型、稀疏编码等技术,以外观、运动特征等底层视觉信息为依据生成摘要。葛钊等[1]将聚类和图模型相结合并优化,提出了一种基于最短路径的视频摘要方法,将视频摘要问题转化为最短路径求解问题。Cong 等[2]采用字典学习对视频进行总结。近年来,Anuradha 等[4]提出了一种高压缩率的视频摘要方法,通过对初始帧的前景信息不断更新以生成视频摘要。Yaliniz 等[5]考虑生成摘要的一致性、多样性和代表性,设计了一个具备奖励功能的函数,提出了利用强化学习与独立递归神经网络完成无监督的视频摘要的方法。相比有监督方法,无监督学习方法摘要结果有待提高。

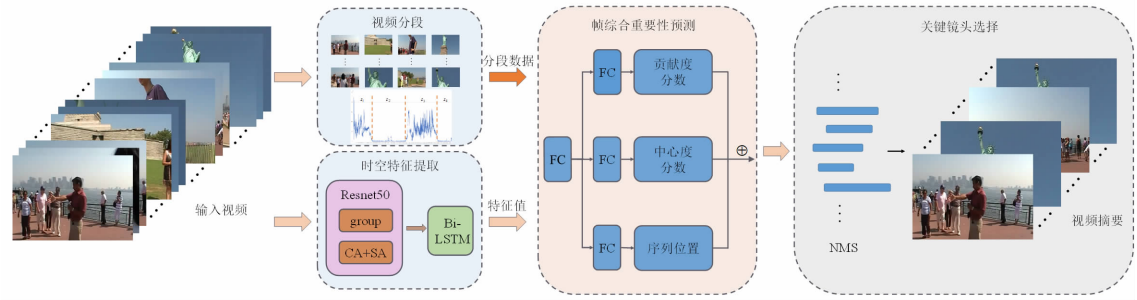
随着深度学习逐渐成为研究热点,基于深度学习的有监督视频摘要方法[6-13]由于能获得与人工标注数据较高相似度的摘要结果而受到青睐,成为当前视频摘要领域的主流方法。最具代表性的是 Zhang 等[7]将长短期记忆网络应用到视频摘要任务,并提出了 vsLSTM 模型,该方法通过双向长短期记忆网络预测视频帧的重要性得分,同时提出了 dppLSTM 模型以提高生成摘要的多样性,并选择出最终摘要。大批研究者在此基础上提出改进,例如:Zhao 等[8]提出分层 LSTM 框架以处理长时间视频序列;Huang 等[9]使用 2D 卷积神经网络、1D 卷积和 LSTM 建模视频帧的时空信息等。注意力机制也被广泛地应用在视频摘要任务中,并取得了重要突破。例如:Lin 等[10]使用 3D CNN 提取视频帧深度特征,并设计了一个基于注意力的分层 LSTM 模块来捕获视频帧之间的时序相关性;Ji 等[11]将 BiLSTM 作为编码器,将基于注意力机制的 LSTM 网络作为解码器构建视频摘要模型;Li 等[12]提出了一种高效的基于全局多元注意力的视频摘要卷积神经网络结构 SUM-GDA,从全局角度调整了注意力机制,考虑了视频帧间的成对时间关系。

目前,卷积神经网络(CNN)和长短期记忆网络(LSTM)仍是视频摘要中最常用的两种技术。其中,大多以预训练的 CNN(如 Googlenet 或 VGG-Net)提取视频帧深度特征,虽能有效捕捉视频帧空间信息,但缺少对视频帧局部信息及深浅层融合信息的关注,提取特征表达能力不足,网络模型选择关键帧的能力仍有待提高[10]。此外,视频片段分割多基于核时序分割(kernel temporal segmentation, KTS)算法[13],摘要模型以提取特征为唯一依赖项,导致模型集成度高、鲁棒性差[14]。

残差网络作为经典的深度学习模型,能充分利用视频帧的深层语义、浅层纹理信息,在图像分类、目标检测等任务中均表现优异。在残差网络中嵌入注意力机制能够对视频帧中运动目标区域进行有效监督,进一步提高关键帧选择精准度。时空切片作为一种高效的视频时空分析方法,可记录长时间尺度的视频场景信息,可以低复杂度快速完成视频片段分割。基于以上分析,本文提出基于时空切片的核时序分割算法(STS-KTS)以完成视频精准分段,并以双注意力机制和分组卷积为基本组件,结合 BiLSTM[15]构建时空特征提取网络,快速提取丰富的时空特征信息,采用帧参数预测模块获取视频帧的贡献度分数、中心度分数以及帧序列位置,进一步生成摘要视频。

1 本文方法

本文方法总体结构如图 1 所示,主要由视频分段模块、时空特征提取模块、帧综合重要性预测模块和关键镜头选择模块组成。视频分段模块采用时空切片完成视频片段分割;时空特征提取模块用来提取视频帧内和帧间深度特征;帧综合重要性预测模



FC—全连接层；group—分组卷积；CA—通道注意力模块；SA—空间注意力模块；NMS—非极大值抑制。

图 1　所提视频摘要方法总体结构

Fig. 1　Overall structure of the proposed video summarization method

块引入帧参数预测模块以对视频帧贡献度分数、中心度分数、帧序列位置等进行预测，并以此设计损失函数反向调节网络；关键镜头选择模块通过将帧分数转化为镜头分数以生成动态视频摘要。

1.1　视频分段模块

本文所提方法生成摘要为镜头集合，因此视频镜头分割为关键步骤。现有视频镜头分割方法大致分为两种：一种为传统的视频镜头分割方法，如基于 SURF 和 SIFT 特征的视频镜头分割方法[16]、基于 Hu 不变矩和三维颜色直方图的视频镜头分割方法[17]等，此类方法多针对图像全量像素点进行分析，计算复杂度较高；另一种为基于深度学习的视频镜头分割方法，其中，KTS 算法[13]通过对输入视频进行变点检测以完成视频分段，被广泛应用于视频镜头分割任务中。KTS 算法以视频帧特征值作为输入，分割效果直接取决于提取特征的表达能力，即由特征提取网络获得的特征值既作为视频摘要生成模型的输入，又作为参考镜头分割算法输入值，由此产生了整体模型对特征值的单一依赖，模型鲁棒性较差。时空切片由时间维 t 与空间维 x 或 y 组成，可记录长时间尺度的视频场景信息，目前主要应用于关键帧提取[18]、交通流分析[19]和目标检测与跟踪等任务中，大多与传统方法相结合以完成任务。综合时空切片对镜头转换十分敏感的特性，本文将其与深度学习结合，应用于视频镜头分割任务中，提出了基于时空切片的核时序分割算法（STS-KTS），以较小的计算量完成视频镜头分割，并消除模型对特征值的单一依赖。

本文所提的 STS-KTS 算法流程如图 2 所示。首先，以源视频为输入，按列采样构建时空切片；其次，对其做灰度化、Sobel 算子边缘检测、二值化等预处理操作；然后，采用水平投影法将二维图像映射为一维数据，并以视频帧号为横轴、映射值为纵轴绘制折线图；最后，将折线图中的点值作为 KTS 的输入，进而获得视频分段数据。

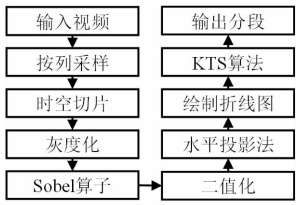


图 2　STS-KTS 算法流程

Fig. 2　Flow chart of STS-KTS algorithm

TVSum 数据集中的 EE-bNr36nyA 视频镜头转换频率较高，因此选用该段视频做结果展示。图 3 为对原视频的 1～1 200 帧按列采样形成的时空切片。可以看出，视频可粗分为 7 段，段与段之间有明显的镜头转换。

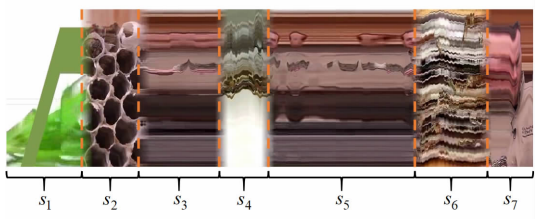


图 3　TVSum 数据集 EE-bNr36nyA 视频的时空切片

Fig. 3　Spatiotemporal slice (EE-bNr36nyA-TVSum)

采用 STS-KTS 算法进行分段的结果如图 4 所示。可以看出，所提算法将该段视频分为 8 段，后 6 段与图 3 中的人工分类结果一致，不同之处在于所提算法将 s_1 细分为 2 段，其分割结果更为细致。

将 STS-KTS 算法得到的视频分段 $C=(t_o^{\mathrm{s}}, t_o^{\mathrm{e}})$ 作为参考镜头数据，其中，t_o^{s}、t_o^{e} 分别为第 o 段的开始帧和结束帧，定义视频帧参数 $\delta_t^*=(\delta_{\mathrm{l}}^*, \delta_{\mathrm{r}}^*)$，$\delta_{\mathrm{l}}^*$ 和 δ_{r}^* 可表示为

$$\delta_{\mathrm{l}}^* = j - t_o^{\mathrm{s}} \tag{1}$$

$$\delta_{\mathrm{r}}^* = t_o^{\mathrm{e}} - j \tag{2}$$

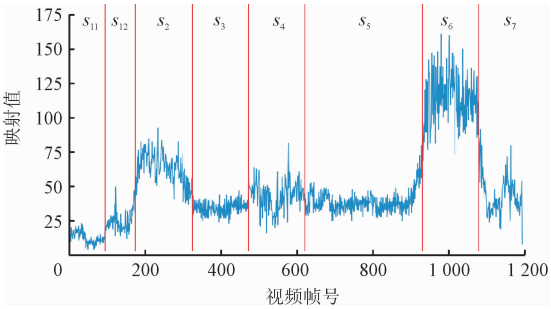


图 4 TVSum 数据集 EE-bNr36nyA 视频的分段结果

Fig. 4 Segmentation result (EE-bNr36nyA-TVSum)

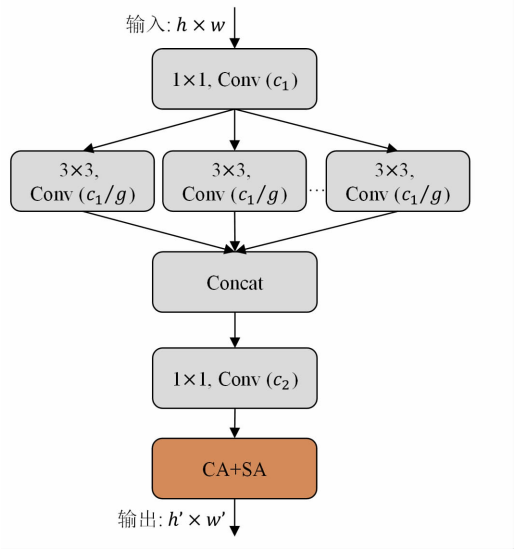
式中 j 为当前视频帧号。参考中心度分数 v_j^* 可表示为

$$v_j^* = \frac{\min(\delta_l^*, \delta_r^*)}{\max(\delta_l^*, \delta_r^*)} \tag{3}$$

此外，视频帧的参考重要性分数 s_j^* 由人为标注数据获得。

1.2 时空特征提取模块

本文方法使用 Resnet50[20] 提取视频帧深、浅层空间特征，使用 BiLSTM 提取视频帧间时序特征，两者融合以增强网络模型处理视频序列的能力。视频场景具有多样性，但卷积神经网络的通道权重通常固定且相等，这限制了网络对不同视频场景的表达能力[21]。为此，本文引入块内双注意力机制，使得网络聚焦视频帧的感兴趣区域，提高对目标区域的关注度。此外，综合考虑模型的体积、参数量、运算量和检测精度，结合分组卷积[22]搭建高效的 Resnet50 网络。改进的 Resnet50 残差模块结构如图 5 所示。



c_1，c_2—通道数。

图 5 改进的残差模块结构

Fig. 5 Improved residual module structure

为使网络聚焦于视频帧中的目标运动区域，进一步提高网络模型在复杂场景下选择关键帧的准确性，引入了图 6 的双注意力特征融合模块，将其运用到图 5 的 Resnet50 残差模块中最后一层卷积操作之后，以提取信息量丰富的帧特征。将来自 Resnet50 残差模块 3 层卷积操作之后的特征图 $\boldsymbol{F} \in \mathbb{R}^{c \times w \times h}$ 依次通过通道注意力模块 CA 和空间注意力模块 SA，以增大关注区域的特征权重，使网络关注到具有显著特征的局部位置信息，其中，c、w、h 分别为特征图的通道数、宽度和高度。

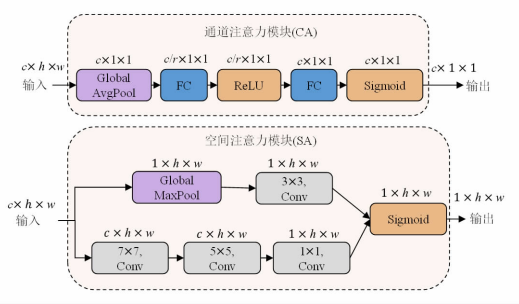


图 6 双注意力特征融合模块

Fig. 6 Dual attention feature fusion module

在通道注意力模块 CA 中，首先对特征图 $\boldsymbol{F}$ 做全局平均池化(global average pooling，GAP)操作以筛选显著通道特征构成通道描述符 $z_{\text{chn}}^{\max}$，然后通过两个 FC 层获得特征通道注意向量 $\boldsymbol{\Omega}_{\text{chn}} \in \mathbb{R}^{c \times 1 \times 1}$，即

$$\boldsymbol{\Omega}_{\text{chn}} = \sigma(\boldsymbol{W}_2(\delta(\boldsymbol{W}_1 z_{\text{chn}}^{\max}))) \tag{4}$$

式中：σ 为 sigmoid 函数；δ 为 ReLu 函数；$\boldsymbol{W}_1 \in \mathbb{R}^{c/r \times c}$ 和 $\boldsymbol{W}_2 \in \mathbb{R}^{c \times c/r}$ 分别为两个 FC 层参数；r 为降维比例。使用 $\boldsymbol{\Omega}_{\text{chn}}$ 对 $\boldsymbol{F}$ 逐通道加权即可得到输出特征图 $\boldsymbol{F}' \in \mathbb{R}^{c \times w \times h}$，其中 $\boldsymbol{F}' = \boldsymbol{\Omega}_{\text{chn}} \boldsymbol{F}$。

在空间注意力模块 SA 中：第一层注意力模块采用全局最大池化(global max pooling，GMP)获取特征通道全局信息，并引入 3×3 卷积以连接；第二层注意力模块对 $\boldsymbol{F}'$ 做 7×7、5×5 的空洞卷积操作以增大感受野，引入 1×1 卷积以降维；两层注意力模块同时作用得到最终特征图 $\boldsymbol{F}'' \in \mathbb{R}^{1 \times w \times h}$，即

$$\boldsymbol{F}'' = \sigma\{f_{\text{conv}}^{3\times3}[\text{GMP}(\boldsymbol{F}')] + f_{\text{conv}}^{1\times1}\{f_{\text{conv}}^{5\times5}[f_{\text{conv}}^{7\times7}(\boldsymbol{F}')]\}\} \tag{5}$$

式中 $f_{\text{conv}}^{3\times3}$、$f_{\text{conv}}^{1\times1}$、$f_{\text{conv}}^{5\times5}$、$f_{\text{conv}}^{7\times7}$ 分别为感受野大小为 3×3、1×1、5×5、7×7 的卷积操作。

Resnet50 中标准卷积的使用产生了冗余的特征图，造成了计算资源和存储资源的浪费。为在保证模型精度的同时提升所提模型的运行速度、减小模型尺寸、降低计算复杂度，引入分组卷积替换残差

块中的 3×3 标准卷积，如图 5 所示。考虑到模型性能会随参数量的减少而下降，同时更多特征图的处理会增加计算机内存消耗，因此将分组数 g 定为 4，以同时保证模型的精度和速度。

假设经改进的 Resnet50 提取的视频视觉特征为 $x_{\text{vis}}=\{x_{\text{vis}}^1,x_{\text{vis}}^2,\cdots,x_{\text{vis}}^t,\cdots,x_{\text{vis}}^k\}$，其中，$k$ 为测试段视频帧数，接下来需将 x_{vis}^t 输入到 BiLSTM 中。图 7 为 BiLSTM 结构。

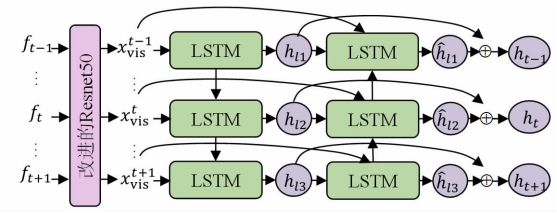


图 7　BiLSTM 结构

Fig. 7　Architecture of BiLSTM

BiLSTM 由前向传播 LSTM 和后向传播 LSTM 组成，双向传播结构以捕捉更完整的序列信息，保证网络更充分地学习。x_{vis}^t 通过前向传播 LSTM 的输出向量 $\boldsymbol{h}_{\text{fpt}}$ 和后向传播 LSTM 的输出向量 $\boldsymbol{h}_{\text{bpt}}$，更新当前状态，传播过程可表示为

$$\boldsymbol{h}_{\text{fpt}}=\text{LSTMFP}(\boldsymbol{h}_{t-1},x_t) \tag{6}$$

$$\boldsymbol{h}_{\text{bpt}}=\text{LSTMBP}(\boldsymbol{h}_{t-1},x_t) \tag{7}$$

$$\boldsymbol{h}_t=[\boldsymbol{h}_{\text{fpt}},\boldsymbol{h}_{\text{bpt}}] \tag{8}$$

1.3　帧综合重要性预测模块

将提取到的视频特征值输入至图 8 中的帧参数预测模块，以获取视频帧的贡献度分数 s_j、中心度分数 v_j 及帧序列位置 δ_t 的预测值。帧参数预测模块由共享的全连接层、ReLU 函数、Dropout 层、标准化层及 3 个并列输出支路组成，3 个分支依次输出 s_j、v_j 和 δ_t，其中，$\delta_t=(\delta_{\text{l}},\delta_{\text{r}})$，$\delta_{\text{l}}$、$\delta_{\text{r}}$ 分别代表当前帧与其所属的片段左边界、右边界的帧差。将 3 路输出相加和，以确定参考值与预测值间的损失以反向调节网络。损失函数 L 的定义如下

$$L=\frac{1}{N_{\text{pos}}}\sum_j L_{\text{cls}}(s_j,s_j^*)+$$

$$\frac{\lambda}{N_{\text{pos}}}\sum_j L_{\text{reg}}(\delta_{tj},\delta_{tj}^*)+\frac{\mu}{N_{\text{pos}}}\sum_j L_{\text{center}}(v_j,v_j^*) \tag{9}$$

式中：$L_{\text{cls}}(\cdot)$、$L_{\text{reg}}(\cdot)$、$L_{\text{center}}(\cdot)$分别为贡献度损失、序列位置损失和中心度损失；N_{pos}为视频摘要帧数；λ、μ 为权重系数，起平衡作用。

帧参数预测模块以视频帧为输入值，每一输入对应一个输出分段$(j-\delta_{\text{l}},j+\delta_{\text{r}})$，由此造成了分段的重复和冗余。因此，设置置信分数 $c_{\text{o}}=s_jv_j$，并引入 NMS 算法，以此保留置信度高的帧对应的视频段，去除低质量段。

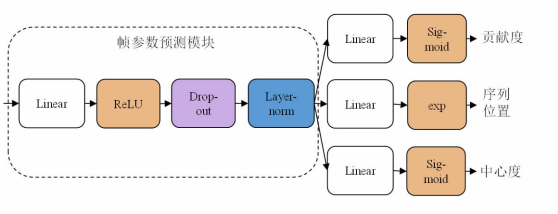


图 8　帧参数预测模块

Fig. 8　Frame parameter prediction module

1.4　关键镜头选择模块

动态摘要以镜头为单位，而非以帧为单位，因此本文需将帧贡献度分数 $s_{h,r}$ 转化为镜头贡献度分数 y_h，以形成基于片段的动态摘要，转换公式为

$$y_h=\frac{1}{n_h}\sum_{r=1}^{n_h}s_{h,r} \tag{10}$$

式中：$s_{h,r}$为第 h 个镜头第 r 帧的贡献度分数；n_h 为第 h 个镜头的总帧数。

本文选取原视频长度的 15%生成视频摘要[23]，采用 0/1 背包算法以在固定长度的摘要视频中保留重要性分数高的部分作为摘要片段，选取原则如下

$$\max\sum_{h=1}^{\zeta}u_hy_h\quad\text{s.t.}\quad\sum_{h=1}^{\zeta}u_hn_h\leqslant 0.15T \tag{11}$$

式中：$u_h\in\{0,1\}$表示第 h 个镜头是否被选为摘要帧；ζ 为镜头数；T 为视频总长度。

2　实验与分析

2.1　实验设置

本文实验的环境为一台处理器为 Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz、显卡为 Tesla V100S-PCIE-32 GB、内存为 188 GB、操作系统为 Centos7 的计算机。在标准数据集 SumMe 和 TVSum 上评估了所提方法的性能。SumMe 数据集包含 25 段视频，每段视频的时长在 1～6.5 min 之间，由 15～18 个人标注了帧重要性分数。TVSum 数据集包含 50 段视频，包含新闻、纪录片等主题，每段视频的时长在 2～10 min 之间，由 20 个人标注了帧重要性分数。这两个数据集是视频摘要领域的常用数据集。表 1 为两个标准数据集的详细信息。本文采用 F_1 分数作为衡量实验结果的指标，F_1 分数的定义为

$$F_1=2\frac{PR}{P+R} \tag{12}$$

$$P=\frac{O}{S} \tag{13}$$

$$R=\frac{O}{G} \tag{14}$$

式中：P 为精确率；R 为召回率；S 为所提方法生成的摘要；G 为人工选择摘要；O 为生成摘要与人工选择摘要重合的部分。F_1 分数越大，说明所提方法的效果越优。

表 1 两个标准数据集的详细信息

Table 1 Details of two standard datasets

参数	数值	
	SumMe	TVSum
视频数量	25	50
用户数	15～18	20
内容	用户生成视频	网络视频
标注类型	帧级分数	帧级分数
最短持续时间/s	32	83
最长持续时间/s	324	647
平均持续时间/s	146	235

2.2 消融实验

为验证所提模块的有效性，本文在 SumMe 数据集上进行消融实验，通过分步在模型中添加所提模块来展示每个模块对网络性能的贡献。定义不使用所提模块的模型为基础模型。SumMe 数据集上的消融实验结果如表 2 所示。可以看出：应用不同的模块后，F_1 分数逐渐增加；当 2 个所提模块同时应用时，达到实验的最优结果。所提模块的有效性得到了证明。

表 2 SumMe 数据集上的消融实验结果

Table 2 Ablation experiment results (SumMe)

实验编号	Resnet50	group+CA+SA	STS-KTS	F_1/%
1	√	×	×	39.19
2	√	√	×	43.54
3	√	×	√	40.72
4	√	√	√	45.02

2.3 定量结果与分析

表 3 为所提方法与近几年的有监督、无监督方法在标准数据集 SumMe 和 TVSum 上的对比结果。可以看出：所提方法在 SumMe 上的 F_1 分数为 45.02%，比 M-AVS 方法提高了 0.62%；在 TVSum 上的 F_1 分数为 59.4%，仅略低于 M-AVS 方法。所提方法在客观评估方面的有效性得到了证明。

表 3 不同视频摘要方法的性能对比

Table 3 Performance comparison of different video summarization algorithms

方法	F_1/%	
	SumMe	TVSum
A-AVS[11]	43.9	59.4
M-AVS[11]	44.4	**61.0**
dppLSTM[24]	38.6	54.7
MSDS-CC[25]	40.6	52.3
SUM-GAN$_{sup}$[26]	41.7	58.2
ERSUM[27]	43.1	59.4
DR-DSN$_{sup}$[28]	42.1	58.1
Cycle-LSTM[29]	41.9	57.6
PCDL$_{sup}$[30]	43.7	59.2
BLMSN[31]	38.9	56.1
ATPP-SUM$_{unsup}$[32]	43.2	58.5
本文方法	**45.02**	59.4

注：表中加粗数据为最优值。

2.4 定性结果与分析

本文选用 SumMe 中动作片段丰富的视频作为摘要示例展示，并挑选原始视频帧数的 15%生成摘要，图 9 为 Bearpark_climbing 视频的摘要示例。

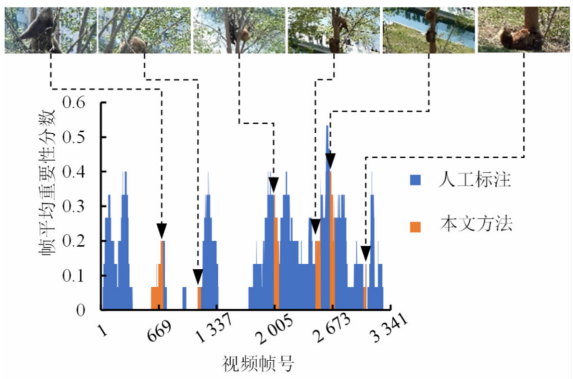


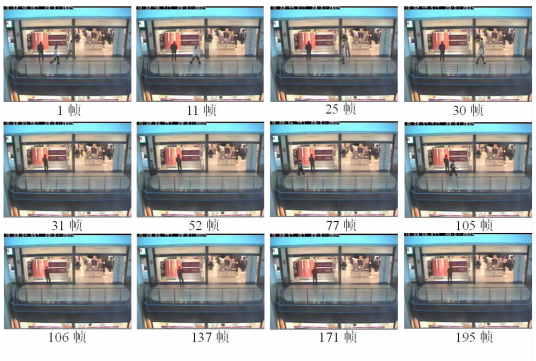
图 9 SumMe 数据集 Bearpark_climbing 视频摘要示例

Fig. 9 Example of video summarization (Bearpark_climbing-SumMe)

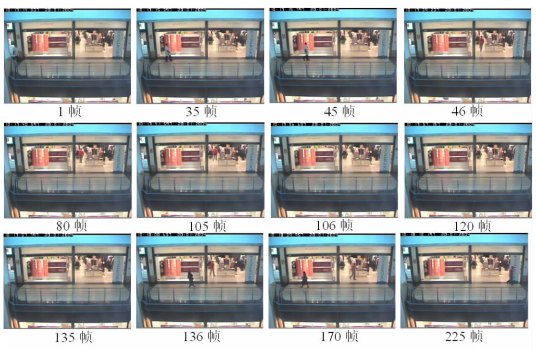
Bearpark_climbing 视频共 3 341 帧，包含两只熊吃树叶并爬下树的一系列过程。生成摘要共 495 帧，包含 6 个片段，可囊括熊吃树叶、爬下树、离开树等一系列事件。可以看出，所提方法选择的重要片段与人工标注的重要分数高的帧片段重合度较高，且摘要视频中涵盖的无用信息较少，说明所提方法能生成反映视频重要内容的简短视频。所提方法的有效性得到了验证。

为进一步验证所提方法的泛化能力，将所提方

法应用于监控视频环境中，所取得的效果如图 10 所示。监控视频数据采用公共数据集 CAVIAR[33] 中的 OneShopOneWait1front 视频和 OneStopEnter1front 视频。OneShopOneWait1front 视频有 1 375帧，所提方法生成的摘要视频为 195 帧，包含 3 个片段，分别为 1～30、31～105、106～195 帧，其中包括一人在商店门口等待、两人分别经过商店以及一人在商店内购物等的画面，可囊括原视频的重要事件。OneStopEnter1front 视频有 1 498 帧，所提方法生成的摘要视频为 225 帧，包含 4 个片段，分别为 1～45、46～105、106～135、136～225 帧，其中包括白衣男子经过商店、红衣男子进入并走出商店、黑衣女子经过商店等相关事件，基本可概括原视频关键事件。可以看出，将所提方法应用于监控视频环境中仍然适用，且能取得理想结果。所提方法的鲁棒性得到了证明。



(a)所提方法对 OneShopOneWait1front 的视频摘要结果



(b)所提方法对 OneStopEnter1front 的视频摘要结果

图 10　所提方法在监控视频公共数据集 CAVIAR 的部分摘要结果

Fig. 10　Partial summary results of the proposed method on the surveillance video public dataset CAVIAR

3　结　论

本文针对现有摘要方法存在的视频帧特征信息提取不充分、摘要结果过分依赖单一特征的问题，提出了一种融合时空切片和双注意力机制的视频摘要方法，本文结论如下。

(1)基于时空切片的核时序分割算法(STS-KTS)将原视频对应的时空切片映射得到的一维数组作为 KTS 的输入特征，以完成对原视频的精准分段，能够消除现有摘要方法对单一特征的依赖性。

(2)时空特征提取模块以双注意力机制和分组卷积为基本组件，结合 BiLSTM 能够快速提取丰富的时空特征信息。

(3)分别从客观评估、各模块效果验证和主观视觉效果多个角度对本文方法进行验证，结果均显示出所提方法的有效性和实用性。

(4)虽然所提方法取得了良好效果，但存在复杂度偏高的问题，后续将进行模型轻量化研究，进一步降低算法复杂度。

参考文献：

[1]　葛钊，赵烨. 一种基于最短路径的视频摘要方法 [J]. 合肥工业大学学报(自然科学版)，2021，44(2)：193-198.
GE Zhao，ZHAO Ye. A video summary method based on shortest path algorithm [J]. Journal of Hefei University of Technology (Natural Science)，2021，44(2)：193-198.

[2]　CONG Yang，YUAN Junsong，LUO Jiebo. Towards scalable summarization of consumer videos via sparse dictionary selection [J]. IEEE Transactions on Multimedia，2012，14(1)：66-75.

[3]　马明阳，梅少辉，万帅. 采用非线性块稀疏字典选择的视频总结 [J]. 西安交通大学学报，2019，53(5)：142-148.
MA Mingyang，MEI Shaohui，WAN Shuai. Nonlinear block sparse dictionary selection for video summarization [J]. Journal of Xi'an Jiaotong University，2019，53(5)：142-148.

[4]　ANURADHA K，ANAND V，RAAJAN N R. An effective technique for the creation of a video synopsis [J]. Journal of Ambient Intelligence and Humanized Computing，2020(7)：1-6.

[5]　YALINIZ G，IKIZLER-CINBIS N. Using independently recurrent networks for reinforcement learning based unsupervised video summarization [J]. Multimedia Tools and Applications，2021，80(12)：17827-17847.

[6]　LIU Tianrui，MENG Qingjie，HUANG Junjie，et al.

Video summarization through reinforcement learning with a 3D spatio-temporal U-net [J]. IEEE Transactions on Image Processing, 2022, 31: 1573-1586.

[7] ZHANG Ke, CHAO Weilun, SHA Fei, et al. Video summarization with long short-term memory [C]// Computer Vision — ECCV 2016. Cham, Germany: Springer International Publishing, 2016: 766-782.

[8] ZHAO Bin, LI Xuelong, LU Xiaoqiang. HSA-RNN: hierarchical structure-adaptive RNN for video summarization [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7405-7414.

[9] HUANG Siyu, LI Xi, ZHANG Zhongfei, et al. User-ranking video summarization with multi-stage spatio-temporal representation [J]. IEEE Transactions on Image Processing, 2019, 28(6): 2654-2664.

[10] LIN Jingxu, ZHONG Shenghua, FARES A. Deep hierarchical LSTM networks with attention for video summarization [J]. Computers & Electrical Engineering, 2022, 97: 107618.

[11] JI Zhong, XIONG Kailin, PANG Yanwei, et al. Video summarization with attention-based encoder-decoder networks [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(6): 1709-1717.

[12] LI Ping, YE Qinghao, ZHANG Luming, et al. Exploring global diverse attention via pairwise temporal relation for video summarization [J]. Pattern Recognition, 2021, 111: 107677.

[13] ZHU Wencheng, LU Jiwen, LI Jiahao, et al. DSNet: a flexible detect-to-summarize network for video summarization [J]. IEEE Transactions on Image Processing, 2021, 30: 948-962.

[14] 李依依, 王继龙. 自注意力机制的视频摘要模型 [J]. 计算机辅助设计与图形学学报, 2020, 32(4): 652-659.
LI Yiyi, WANG Jilong. Self-attention based video summarization [J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(4): 652-659.

[15] BROCKI Ł, MARASEK K. Deep belief neural networks and bidirectional long-short term memory hybrid for speech recognition [J]. Archives of Acoustics, 2015, 40(2): 191-195.

[16] 张昊骋, 朱晓龙, 胡新洲, 等. 基于 SURF 和 SIFT 特征的视频镜头分割算法 [J]. 液晶与显示, 2019, 34(5): 521-529.
ZHANG Haosu, ZHU Xiaolong, HU Xinzhou, et al. Shot segmentation technology based on SURF features and SIFT features [J]. Chinese Journal of Liquid Crystals and Displays, 2019, 34(5): 521-529.

[17] 杨瑞琴, 吕进来. 基于双重检测的视频镜头分割方法 [J]. 计算机工程与设计, 2018, 39(5): 1393-1398.
YANG Ruiqin, LV Jinlai. Video shot segmentation method based on dual-detection [J]. Computer Engineering and Design, 2018, 39(5): 1393-1398.

[18] ZHANG Yunzuo, TAO Ran, WANG Yue. Motion-state-adaptive video summarization via spatiotemporal analysis [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(6): 1340-1352.

[19] 张晓栋. 基于铁路货运大数据的重车车流推算关键技术研究 [D]. 北京: 北京交通大学, 2018.

[20] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 770-778.

[21] 陈勇, 刘曦, 刘焕淋. 基于特征通道和空间联合注意机制的遮挡行人检测方法 [J]. 电子与信息学报, 2020, 42(6): 1486-1493.
CHEN Yong, LIU Xi, LIU Huanlin. Occluded pedestrian detection based on joint attention mechanism of channel-wise and spatial information [J]. Journal of Electronics & Information Technology, 2020, 42(6): 1486-1493.

[22] CHOLLET F. Xception: deep learning with depthwise separable convolutions [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 1800-1807.

[23] FAJTL J, SOKEH H S, ARGYRIOU V, et al. Summarizing videos with attention [C]// Computer Vision — ACCV 2018 Workshops. Cham, Germany: Springer International Publishing, 2019: 39-54.

[24] ZHANG Ke, CHAO Weilun, SHA Fei, et al. Video summarization with long short-term memory [C]// Computer Vision — ECCV 2016. Cham, Germany: Springer International Publishing, 2016: 766-782.

[25] MENG Jingjing, WANG Suchen, WANG Hongxing, et al. Video summarization via multi-view representative selection [C]// 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Piscataway, NJ, USA: IEEE, 2017: 1189-1198.

[26] MAHASSENI B, LAM M, TODOROVIC S. Unsupervised video summarization with adversarial LSTM networks [C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 2982-2991.

[27] LI Xuelong, ZHAO Bin, LU Xiaoqiang. A general

framework for edited video and raw video summarization [J]. IEEE Transactions on Image Processing, 2017, 26(8): 3652-3664.

[28] ZHOU Kaiyang, QIAO Yu, XIANG Tao. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2018: 7582-7589.

[29] YUAN Li, TAY F E H, LI Ping, et al. Unsupervised video summarization with cycle-consistent adversarial LSTM networks [J]. IEEE Transactions on Multimedia, 2020, 22(10): 2711-2722.

[30] ZHAO Bin, LI Xuelong, LU Xiaoqiang. Property-constrained dual learning for video summarization [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(10): 3989-4000.

[31] 武光利，李雷霆，郭振洲，等. 基于改进的双向长短期记忆网络的视频摘要生成模型 [J]. 计算机应用，2021，41(7)：1908-1914.
WU Guangli, LI Leiting, GUO Zhenzhou, et al. Video summarization generation model based on improved bi-directional long short-term memory network [J]. Journal of Computer Applications, 2021, 41(7): 1908-1914.

[32] 王浩，彭力. 基于改进的全卷积网络的视频摘要算法 [J]. 激光与光电子学进展，2021，58(22)：407-415.
WANG Hao, PENG Li. Video summarization algorithm based on improved fully convolutional network [J]. Laser & Optoelectronics Progress, 2021, 58(22): 407-415.

[33] FISHER R. CAVIAR test case scenarios [EB/OL]. [2022-04-01]. https: // homepages. inf. ed. ac. uk/rbf/CAVIARDATA1/.

（编辑　陶晴　亢列梅）