第 58 卷 第 22 期/2021 年 11 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于改进的全卷积网络的视频摘要算法

王浩，彭力*

江南大学物联网工程学院，江苏 无锡 214122

摘要 面对海量的视频数据，视频摘要技术在视频检索、视频浏览等领域发挥着越来越重要的作用，其旨在通过生成简短的视频片段或选择关键帧集合来获取输入视频中的重要信息。现有的方法大多集中在研究视频摘要的代表性和多样性上，没有考虑到视频结构等多尺度上下文信息。针对上述问题，提出了一种基于全卷积序列网络的视频摘要模型，模型中利用时间金字塔池化对视频中的多尺度上下文信息进行提取，并利用全连接的条件随机场对视频帧序列进行标注。在 SumMe 和 TVSum 数据集上的实验表明，所提模型取得了比全卷积序列网络更好的性能，在这两个数据集上 F 分指标分别提高了 1.6%和 3.0%。

关键词 机器视觉；视频摘要；深度学习；全卷积序列网络；卷积神经网络

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2215008

Video Summarization Algorithm Based on Improved Fully Convolutional Network

Wang Hao, Peng Li*

School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu 214122, *China*

Abstract In the face of massive video data, video summarization technique plays an increasingly important role in video retrieval, video browsing and other fields. It aims to obtain important information in input videos by generating short video clips or selecting a set of key frames. Most of the existing methods focus on the representativeness and diversity of video summarization, without considering the multi-scale contextual information such as the structure of the video. To solve the above problems, a video summarization model based on improved fully convolutional network is proposed, in which time pyramid pooling is used to extract multi-scale contextual information, and the fully connected conditional random field is used to label the video frame sequence. Experiments on SumMe and TVSum datasets show that the proposed model achieves better performance than fully convolutional sequence networks, and the F-score indexes on these two data sets are improved by 1.6% and 3.0%, respectively.

Key words machine vision; video summarization; deep learning; fully convolutional sequence networks; convolutional neural networks

OCIS codes 150.1135; 110.4155; 100.4996

1 引 言

随着视频采集设备的不断普及和成本的不断降低，在过去十几年中视频数据量急剧增加，视频已经逐渐成为可视化数据的重要形式之一。由于人们完整地观看数据量巨大的视频并获取其中有用的信息是不现实的，因此开发能够实现高效浏览庞大视频数据的计算机视觉技术变得越来越重要。为了使这些海量的视频数据易于浏览和访问，视频摘要技术应运而生。

收稿日期：2021-03-15；**修回日期**：2021-05-26；**录用日期**：2021-07-15

基金项目：国家自然科学基金(61873112)

通信作者：*penglimail2002@163.com

当给定一个输入视频时，视频摘要的目标是生成一个较短的视频，同时能够保留输入视频中的重要信息。视频摘要技术在现实世界中得到了广泛应用。例如，在视频监控工作中，工作人员需要浏览冗长的监控摄像头拍摄的视频，这是一件很繁琐且耗时的事情。如果能够提供一个简短的视频摘要，将大大减少视频监控工作中所需要的人力。由于短视频较容易存储和传输，因此还为很多后续的视频分析任务提供了帮助。例如，在短视频基础上运行行为识别算法，可以加快识别速度。

早期的视频摘要算法主要是基于无监督的。一种最常用的方法就是聚类[1-2]，大多数的聚类方法是根据视频帧的颜色特征和一些运动特征将视频帧分成几簇，每一簇中都是内容相似的视频帧。另一种常见的方法是视频大纲[3-4]，通过优化一些能量函数，将不同时间和空间的运动目标压缩在同一个时空域中，以生成大纲视频，这样最大程度地消除了时空域中的冗余。

近年来，深度学习逐渐被用于视频摘要任务中。基于深度学习的视频摘要算法通常将视频摘要视为一种序列标注问题，并使用基于循环神经网络(RNN)的变种长短期记忆(LSTM)单元来解决该问题[5]。LSTM 模型中的每一个时间步，对应于输入视频中的一帧。在一个时间步中，LSTM 模型输出一个二进制值，用以表示该视频帧是否是关键帧。LSTM 的优势在于它可以发现长期的结构依赖关系。然而，这些基于 LSTM 的模型也具有一定的局限性，即 LSTM 中的计算顺序通常是从左到右，这意味着 LSTM 一次只能处理一个视频帧，而且每一帧必须等待前一帧的处理结果。因此，前面的视频帧是否被选入视频摘要中是无法依赖于后面的视频帧的。Li 等[6]在双向 LSTM[7]的基础上，将视频分类网络 NeXtVlad[8]作为视频语义的生成器，通过最小化原始视频和视频摘要的语义相似度获得了较好的效果。虽然有双向 LSTM 的存在，但是文献[6]中仍然会遇到与 LSTM 相同的问题。

为了解决上述问题，Rochan 等[9]使用全卷积神经网络(FCN)[10]代替 RNN。FCN 在语义分割领域已经得到了广泛的应用。视频摘要和语义分割常被认为是计算机视觉中两个完全不同的问题，但其实两者具有很大的相似性。语义分割本质是对图片中的每一个像素进行分类，而视频摘要可以被看作对视频中的每一帧进行分类。在将两个任务建立起联系后，FCN 可转化为全卷积序列网络(FCSN)，较适用于视频摘要任务，并取得了极佳的效果。

与 FCN 在语义分割中的缺陷类似，FCSN 没有考虑到视频帧之间的时间一致性关系。此外，FCSN 也没有考虑到视频中的结构。针对上述问题，本文提出一种了基于时间金字塔池化的上下文模块，用于提取时域上的多尺度上下文信息。同时，利用全连接的条件随机场(DenseCRF)，并考虑视频帧之间的时间一致性关系，对分类结果进行优化。最终，所提模型在两个公开的数据集 TVSum 和 SumMe 中，相比于 FCSN 模型，性能提升了 1.6% 和 3.0%。

2 相关工作

FCN 是一种广泛用于语义分割的模型。Rochan 等[9]对 FCN 进行了一些修改，使它适用于视频摘要任务，并将其命名为 FCSN。在 FCN 中，输入是尺寸为 $m\times n\times 3$ 的 RGB 图像，m 和 n 是图像的高度和宽度。输出图像的尺寸为 $m\times n\times C$，C 为分类的数量。在 FCSN 中，输入图像的尺寸是 $1\times T\times D$，其中 T 为视频中的帧数、D 为每一帧的特征向量的维度，输出图像尺寸为 $1\times T\times D$。在视频摘要任务中设定 C 为 2，因为需要把视频中的每一帧分为两类(关键帧或非关键帧)。FCSN 中将 FCN 中所有空间维度的卷积都转换为时间卷积，如图 1 所示。

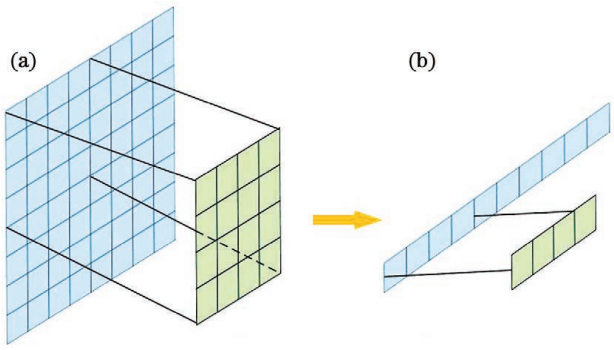

图 1 空间卷积和时间卷积。(a) 空间卷积；(b) 时间卷积

Fig. 1 Spatial convolution and temporal convolution. (a) Spatial convolution; (b) temporal convolution

Long 等[10]通过对 FCSN 的研究在语义分割和视频摘要间建立起了联系。在语义分割领域中，Yu 等[11]为了提取上下文信息设计了基于空洞卷积的前端模块和上下文模块。在前端模块中，将 VGG-16 网络中的最后两个池化层移除，并用扩张卷积代替随后的卷积层。4.4 节的实验表明，仅仅使用前

端模块就可取得比 FCN 更好的分割效果。Chen 等[12-13]提出了基于空洞卷积的空间金字塔池化(ASPP),以提取多尺度的上下文信息。Krhenbühl 等[14]考虑了像素与像素间的关系,提出了利用 DenseCRF 对分割结果进行优化。相比于视频摘要,语义分割是一个得到更广泛研究的领域[15],并且很多针对语义分割的方法都可以应用到视频摘要领域中。

3　改进的全卷积序列网络

3.1　时间金字塔池化

FCSN 中采用最大池化层在减小视频帧序列长度的同时,扩大了感受野,但是大大降低了帧序列的分辨率。帧序列分辨率的降低导致在最后的上采样操作中,很多细节信息无法被还原,比如一些时间长度很短的镜头,此时短镜头中的视频帧可能无法成为视频摘要中的关键帧。另外,在 FCSN 中使用了跳层结构来融合不同尺度的上下文信息,但是融合的尺度较少。由于 FCSN 使用 FCN-16 网络作为原型,因此只融合了两个尺度的特征信息。然而,视频作为一种结构清晰的可视化数据,需要提取较多尺度的上下文信息。

受 ASPP[12-13]启发,提出了基于空洞卷积的时间金字塔池化(ATPP)模块,如图 2 所示。输入为一维的特征图,四个空洞卷积核的尺寸为 1×3,它们的扩张率分别为 1、6、12、18。其中,扩张率为 1 表示正常卷积核,扩张率为 6 表示每个输入都跳过 5 个视频帧。输入的特征图经由 4 个扩张率逐渐增大的空洞卷积核提取特征后,再依次相加,相加后的特征图与输入的特征图在长度上保持一致,从而在不改变分辨率的情况下,获得了时域的多尺度上下文信息。

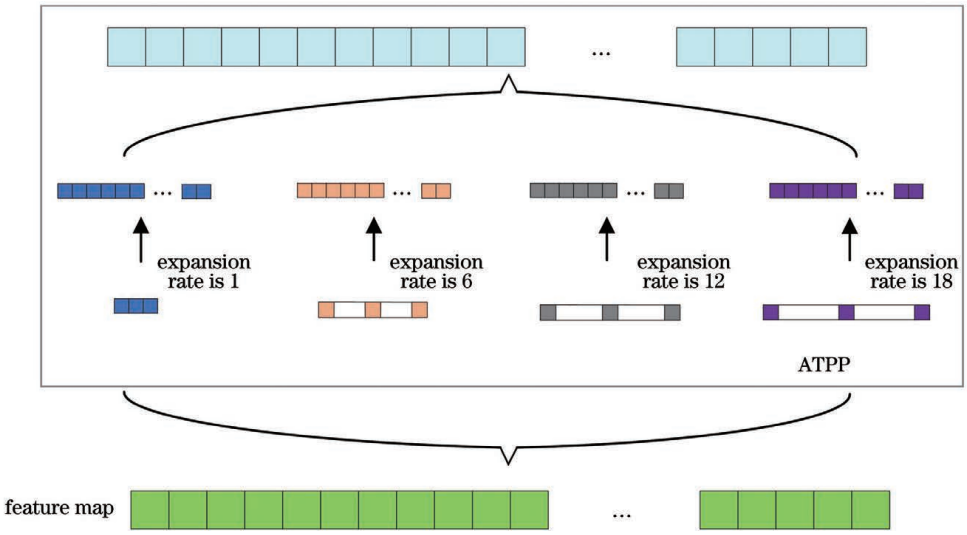

图 2　ATPP 示意图

Fig. 2　Schematic diagram of ATPP

3.2　全连接的条件随机场

FCSN 中没有考虑视频帧之间的关系,而在视频摘要的任务中,视频中的重要内容往往会持续数帧。视频中关键镜头所包含的连续数帧,需要被统一分类为关键帧。受文献[14,16]的启发,提出了针对视频摘要的 DenseCRF 模型,该模型的能量函数表达式为

$$E(s)=\sum_i \psi_u(s_i)+\sum_{i<j}\psi_p(s_i,s_j), \tag{1}$$

式中:s 为视频帧的标签;s_i 是视频中的第 i 帧;s_j 是视频中的第 j 帧;将 $\psi_u(s_i)=-\log[P(s_i)]$ 作为能量函数的一元势函数,其中 $P(s_i)$ 是每一帧成为关键帧或非关键帧的概率,也就是改进的全卷积序列网络(ATPP-SUM)的输出。$\psi_p(s_i,s_j)$ 为能量函数的二元势函数,其表达式为

$$\psi_p(s_i,s_j)=\mu(s_i,s_j)\left[w_1\exp\left(-\frac{|t_i-t_j|}{2\sigma_\alpha^2}-\frac{\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2}{2\sigma_\beta^2}\right)+w_2\exp\left(-\frac{|t_i-t_j|}{2\sigma_\gamma^2}\right)\right], \tag{2}$$

式中:$\mu(s_i,s_j)$ 是标签兼容项,当 $s_i\neq s_j$ 时,$\mu(s_i,s_j)=1$,其他情况下为 0;w_1 为表现核的权重

值；w_2 为光滑核的权重值；$\exp\left(-\frac{|t_i-t_j|}{2\sigma_\alpha^2}-\frac{\|\boldsymbol{f}_i-\boldsymbol{f}_j\|_2}{2\sigma_\beta^2}\right)$为表现核，其中 t_i 是第 i 帧的时刻、t_j 是第 j 帧的时刻；$\boldsymbol{f}_i$ 是第 i 帧的特征、$\boldsymbol{f}_j$ 是第 j 帧的特征，每帧的特征是由 GoogleNet[17] 提取的 1024 维的特征向量；$\exp\left(-\frac{|t_i-t_j|}{2\sigma_\gamma^2}\right)$为光滑核，光滑核只取决于视频帧之间的间隔；$\sigma_\alpha$、$\sigma_\beta$ 和 σ_γ 为控制着上述两个高斯核的尺度。因为视频帧序列是一维的，所以视频帧的时刻信息也是一维的。可以看出，如果两个视频帧的时间间隔相近且特征值相近，那么 DenseCRF 更可能将它们划分为同一个标签。

3.3　改进的全卷积序列网络

所提模型的特性：1）视频摘要的 LSTM 模型[2]是按顺序处理视频帧的，与 LSTM 模型相比，所提模型能够同时处理所有帧，并且能够通过 ATPP 模块提取多尺度的上下文信息；2）使用 DenseCRF 考虑视频帧间的关系，对模型的分类结果进行优化；3）语义分割模型通常采用编码器-解码器的架构。首先，由编码器对图像进行处理以提取特征。然后，解码器使用编码器提取的特征来产生分类结果。同样地，所提模型也可以被认为是编码器-解码器架构。编码器用于处理视频帧，以提取多尺度的上下文信息。解码器用于产生一个 0/1 标签的序列。将改进后的模型称为 ATPP-SUM，如图 3 所示。

在 ATPP-SUM 中，前 4 个卷积层（Conv1～Conv4）由多个时间卷积组成，并且在每一个时间卷积后都添加批量归一化层（BN）和线性修正单元（ReLU）。此外，在前三个卷积层中，还会加一个池化层，来对特征图进行下采样。在卷积层 Conv5 中，使用扩张率为 2 的空洞卷积代替 FCSN 中的池化层，在不减小分辨率的情况下，扩大了感受野。将 Conv5 的输出送入 ATPP 模块中，在 ATPP 模块中共有 4 种不同扩张率的空洞卷积核，从不同的尺度上提取上下文信息。然后，利用元素加法将提取到的多尺度上下文信息融合，送入卷积层 Conv6 和 Conv7 中，可获得通道数为 2 的输出。接着，将可以学习的反卷积层作为上采样以得到与输入相同长度的预测输出。最后，使用 DenseCRF 对分类结果进行优化，从而得到 0/1 标签的视频帧序列。

3.4　有监督学习的损失函数

在基于关键帧的监督学习中，由于在视频摘要

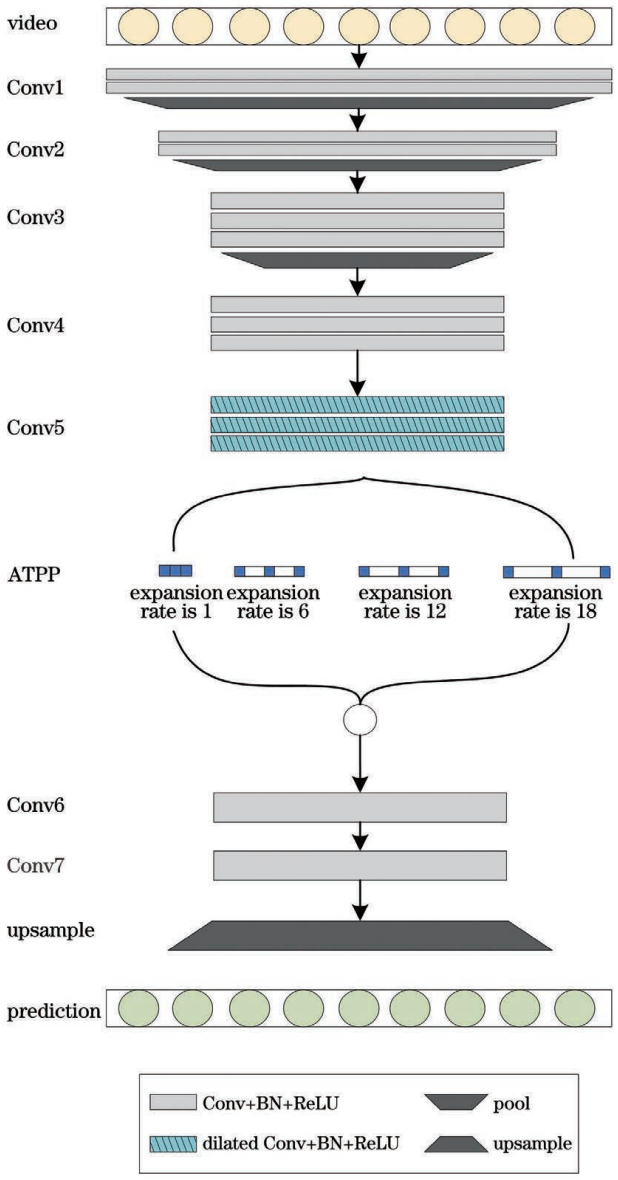

图 3　ATPP-SUM 网络结构

Fig. 3　Network structure of ATPP-SUM

中只选择了输入视频中的一小部分视频帧，即与非关键帧相比，关键帧非常少，因此类之间是极不平衡的。一个常见的处理策略是给每个类设定不同的权重值来进行学习，这种类平衡的策略经常被用于序列标注任务中[18]。对第 c 类来说，将权重定义为

$$w_c=\frac{f_{\text{median}}}{f_c},\tag{3}$$

式中：f_c 是第 c 类的频率，即视频中为 c 的视频帧的数量除以视频帧的总数；f_{median} 是计算出的频率的中位数。

假设训练的视频帧数为 T，定义损失函数为

$$L_{\text{sup}}=-\frac{1}{T}\sum_{t=1}^{T}w_{c_t}\lg\left[\frac{\exp(\varphi_{t,c_t})}{\sum_{c=1}^{C}\exp(\varphi_{t,c})}\right],\tag{4}$$

式中：c_t 是第 t 帧的真实标签；$\varphi_{t,c}$ 为视频中第 t 帧被分类为第 c 类的预测值；C 为总类数。

3.5 无监督学习的损失函数

一方面，需要视频摘要中的帧具有多样性和代表性。另一方面，视频摘要的数据集非常稀少。因此，有必要对模型进行无监督学习。为了提高视频摘要的多样性和代表性，提出了两种损失函数，对模型进行约束。同时，对 ATPP-SUM 进行了修改。首先，根据模型的预测分数选择 Y 个关键帧。然后，对这些关键帧解码后得到的特征向量应用尺寸为 1×1 的卷积来重建它们的原始特征。接着，使用跳跃连接合并选定 Y 个关键帧的输入帧特征向量。最后，使用尺寸为 1×1 的卷积来获得 Y 个关键帧的最终重构特征，使得每个关键帧的特征向量维数与输入帧特征向量维数保持相同。

重建损失的表达式为

$$L_{\text{reconst}}(\boldsymbol{E}_v,\boldsymbol{v})=\frac{1}{k}\sum_{t=1}^{k}\|\boldsymbol{E}_v-\boldsymbol{v}\|_2^2,\tag{5}$$

式中：$\boldsymbol{E}_v$ 和 $\boldsymbol{v}$ 是分别是 ATPP-SUM 生成的视频摘要中的第 t 帧的特征向量和输入视频中相对应的第 t 帧的特征向量；k 是 ATPP-SUM 生成的视频摘要的总帧数。这个损失函数的目标是使关键帧重构后的特征向量尽可能地接近于输入视频中对应关键帧的特征向量。

多样性损失的表达式为

$$L_{\text{div}}=\frac{1}{|Y|(|Y|-1)}\sum_{t\in Y}\sum_{t'\in Y,t'\neq t}d(\boldsymbol{f}_t,\boldsymbol{f}_{t'}),\ d(\boldsymbol{f}_t,\boldsymbol{f}_{t'})=\frac{\boldsymbol{f}_t^{\mathrm{T}}\boldsymbol{f}_{t'}}{\|\boldsymbol{f}_t\|_2\|\boldsymbol{f}_{t'}\|_2},\tag{6}$$

式中：$\boldsymbol{f}_t$ 为第 t 帧的重构特征；$\boldsymbol{f}_t$ 为第 t' 帧的重构特征。该损失函数的值越小，则视频摘要种类越丰富。

4 实验结果与分析

4.1 数据集

在 SumMe[19] 和 TVSum[20] 两个数据集上对模型进行了评估。SumMe 数据集中一共有 25 个视频，视频中包含着各种各样的内容，例如体育运动、节日、烹饪等。数据集中视频的长度为 1.5～6.5 min。TVSum 数据集包含了 50 个视频，视频可分为 10 种类型，包括新闻、纪录片等主题，每个视频长度为 2～10 min。

4.2 实验细节

首先，按照文献[5]的要求，将视频统一下采样到 2 frame/s。然后，将 GoogleNet[17] 中 pool 层的输出作为每个视频帧的特征，维度为 1024。将 GoogleNet 作为提取视频特征的神经网络，可以与之前的先进方法进行公平比较。

由于不同的数据集中人工标注方式不同，故本文按照文献[5]的方法为数据集中的每个视频生成关键帧集合。这个关键帧集合将被用来训练 ATPP-SUM 模型。为了与一些先进的方法进行公平比较，需要将模型生成的预测分数转换成基于镜头的视频摘要。同时，需要将数据集中基于关键帧的标注转换为基于镜头的标注。在 SumMe 数据集中，人工标注是基于镜头的，所以不需要对数据集进行处理。然而，在 TVSum 数据集中，只有针对关键帧的标注。为了将针对关键帧的标注转换为针对镜头的标注，本文将按照文献[5]中的步骤对数据集进行处理：1）使用 KTS(Kernel Temporal Segmentation) 算法[2]将完整的视频分割为不相交的视频段；2）计算出每个视频段中视频帧的分数，并将整个视频段的平均分数分配给视频段中的每一帧；3）根据分数对视频中的每一帧进行排序；4）通过阈值的设置，利用背包算法[20]选择一定数量的视频帧形成基于镜头的视频摘要。

在对模型进行测试时，使用均匀采样将测试视频采样为 $T=320$ 的视频后作为模型的输入，模型的输出同样是 T 为 320 的视频。最终，使用邻近算法将模型的输出恢复为原始视频长度。本文按照文献[5]中的方法将预测的关键帧转化为关键镜头。使用 KTS 算法[2]分割测试视频生成不相交的视频段。如果视频段中包含关键帧，则将视频段中所有的视频帧标记为 1，反之则标记为 0。这些被标记为 1 的视频段便组成了基于镜头的视频摘要。

4.3 评价指标

按照文献[5]中的方法，本文使用基于镜头的评价指标。对于给定视频 V，假定 A_O 为视频摘要的总时长，A_G 为人工标注视频的总时长，B 是视频摘要和人工标注视频重叠的时长。通过时间上的重叠来计算准确率（P）和召回率（R）以得到 F 分数，$F=(2P\times R)/(P+R)$，并将 F 分数作为评价指标。P 和 R 的表达式为

$$\begin{cases} P=\dfrac{B}{A_O} \\ R=\dfrac{B}{A_G} \end{cases}。\tag{7}$$

最后，随机选择 20%（视频时间长度占比）视频时长作为测试集，剩下的 80%视频时长作为训练集和验证集。因为数据是随机划分的，所以需要重复划分 5 次以获取平均的 F 分数。

4.4 实验结果与分析

选择了 5 种视频摘要模型与所提模型进行比较。vsLSTM[5]是利用 LSTM 的视频摘要生成模型，该模型将深度学习引用到视频摘要任务中，并获得了非常好的效果。ddpLSTM[5]同样是利用 LSTM 的模型，它通过使用行列式点过程（DPP）算法将选择的关键帧进行多样化，获得了比 vsLSTM 更好的效果。SUM-FCN[9]将经典的语义分割网络 FCN 进行修改，使 FCN 能够适用于视频摘要任务且获得了极佳的效果。SUM-GAN[21]将生成式对抗网络引入到视频摘要任务中，将 LSTM 作为模型的生成器和判别器。Cycle-SUM[22]借鉴了 Cycle-GAN[23]的结构，将判别器进行了修改，提高了原始视频和视频摘要之间的一致性。

表 1 和表 2 中比较了 ATPP-SUM 与之前的先进模型在 SumMe 和 TVSum 数据集中的表现。可以看出，所提模型的性能较好。此外，在 vsLSTM 中，使用了 F 分数对模型进行训练。在 ddpLSTM 中，使用了 F 分数和基于镜头的标注对模型进行训练。然而，在 ATPP-SUM 中，仅仅使用了基于关键帧的二进制标签对模型进行了训练却展示出了更好的性能，同时也反映出了卷积神经网络在序列标注任务上的能力。

表 1　不同模型在 SumMe 数据集上的性能比较

Table 1　Performance comparison of different models on SumMe dataset

Model	F /%
vsLSTM	37.6
ddpLSTM	38.6
SUM-FCN	47.5
SUM-GAN	41.7
Cycle-SUM	41.9
ATPP-SUM	48.6
ATPP-SUM$_{unsup}$	43.2

表 2　不同模型在数据集上的性能比较

Table 2　Performance comparison of different models on TVSum dataset

Model	F /%
vsLSTM	54.2
ddpLSTM	54.7
SUM-FCN	56.8
SUM-GAN	56.3
Cycle-SUM	57.6
ATPP-SUM	58.5
ATPP-SUM$_{unsup}$	55.5

本文借鉴了文献[9]中的两个损失函数 L_{rec} 和 L_{div}，对 ATPP-SUM 进行无监督学习，将此时的模型称为 ATPP-SUM$_{unsup}$。实验结果表明，仅仅使用无监督学习对模型进行训练就可以获得很好的性能，其结果可以与其他有监督学习的模型媲美。

采用有监督的方法对本文提出的模块进行了消融实验，以验证其有效性，如表 3 所示。Dilation-SUM 模块的设计是受文献[12]的启发，将 ATPP 模块去除后，使用空洞卷积代替 FCSN 中的跳层连接。DenseCRF 则对模型的输出进行了优化。

表 3　消融实验

Table 3　Ablation experiment

Module	F on SumMe /%	F on TVSum /%
Dilation-SUM	46.1	55.2
Dilation-SUM+DenseCRF	46.8	56.4
ATPP-SUM	48.6	58.5
ATPP-SUM+DenseCRF	49.1	59.8

从表 3 中可以看出，去掉跳跃连接后，如果只使用空洞卷积来提取上下文信息，则与原模型相比，FCSN 性能会变差。在加入 ATPP 模块后，在 SumMe 数据集上性能提高了 2.5%，而在 TVSum 数据集上性能则提高了 3.3%，这是因为在 SumMe 数据集中镜头较少且镜头内容变化缓慢，而在 TVSum 数据集中镜头很多，且场景变化频繁。ATPP 模块通过提取多尺度的上下文信息可以很好地解决视频结构复杂、镜头多、场景变化频繁的问题。在加入 DenseCRF 后，Dilation-SUM 在两个数据集上的性能分别提高了 0.7%和 1.2%，ATPP-SUM 的性能则分别提高了 0.5%和 1.3%。因为在

SumMe 数据集中的视频都是第一视角的，所以镜头往往是“一镜到底”，镜头较少。而本文中的评价方法是基于镜头的，在将关键帧集合转换为关键镜头集合时，会把含有关键帧的镜头中的视频帧都标记为关键帧。因此，DenseCRF 在对视频帧间进行优化时，无法将优化后的结果完全反映在评价指标中，即评价指标对细节“不敏感”。当视频中的镜头少，且每个镜头时间很长时，则更加“不敏感”。所以 DenseCRF 对 SumMe 数据集中的视频提升较小。

表 4 反映了 DenseCRF 中参数对模型性能产生的影响。实验中使用的数据集是 TVSum。为了便于研究，将(2)式中的 w_1 设置为常数，w_2 设置为 0。可以看出，随着 σ_α 和 σ_β 增大，模型的性能也随之提升。最终在 $\sigma_\alpha=15$，$\sigma_\beta=20$ 附近达到最大值。

表 4　DenseCRF 中参数对模型性能的影响

Table 4　Influence of parameters in DenseCRF on model performance　　unit：%

Parameter	$\sigma_\beta=5$	$\sigma_\beta=15$	$\sigma_\beta=25$
$\sigma_\alpha=10$	56.8	57.4	58.9
$\sigma_\alpha=15$	57.7	58.3	59.4
$\sigma_\alpha=20$	57.3	59.8	59.1

表 5 反映了 ATPP 中不同扩张率对模型性能的影响，其中 ATPP-SUM(1，2，4，8)表示在 ATPP 中四个卷积核的扩张率分别为 1、2、4、8，ATPP-SUM(1，6，12，18) 表示在 ATPP 中四个卷积核的扩张率分别为 1、16、12、18，ATPP-SUM(1，12，24，36) 表示在 ATPP 中四个卷积核的扩张率分别为 1、12、24、36。可以发现，随着扩张率的增大，模型可以获得更丰富的多尺度上下文信息，进而模型性能得到了提高。但是，当扩张率过大时，模型的性能会下降，如表 5 中 ATPP-SUM(1，12，24，36)所示，这是因为扩张率过大时，为了保持图片尺寸不变，填充会增加，此时会产生很多无意义的上下文信息。最终实验结果显示，当卷积核的扩张率在 1、16、12、18 时，模型性能最佳。

表 5　ATPP 中扩张率对模型性能的影响

Table 5　Influence of expansion rates in ATPP on model performance

Model	F on SumMe /%	F on TVSum /%
ATPP-SUM(1，2，4，8)	47.5	57.6
ATPP-SUM(1，6，12，18)	49.1	59.8
ATPP-SUM(1，12，24，36)	48.6	59.5

图 4 中比较了不同方法生成的视频摘要结果。其中，带有向右倾斜线段的图形表示通过模型计算后选择的关键镜头，带有向左倾斜线段的图形表示没有被选中镜头。可以看到，相比于 SUM-FCN，所提两个方法均提高了视频摘要的质量。ATPP-

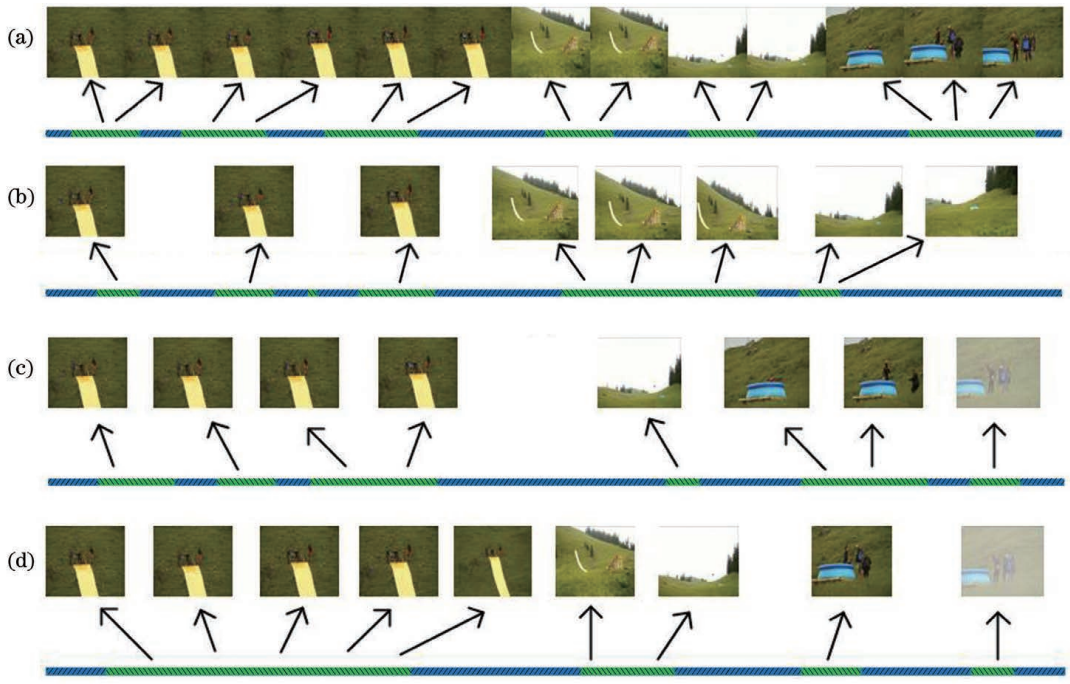

图 4　不同方法生成的视频摘要结果比较。(a)人工标注；(b) SUM-FCN；(c) ATPP-SUM；(d) ATPP-SUM＋DenseCRF

Fig. 4　Comparison of video summarization results by different methods. (a) Manual labeling; (b) SUM-FCN; (c) ATPP-SUM; (d) ATPP-SUM＋DenseCRF

SUM 中增加了 ATPP 模块后，获取到了更多的原始视频内容，进而视频摘要更加完整和紧凑。在 ATPP-SUM+DenseCRF 中，利用 DenseCRF 模块对分类结果进行了优化，使得关键帧能够聚集在一起，从而生成了更高质量的视频摘要。

5 结 论

针对现有方法在视频结构等多尺度上下文信息提取不足的问题，在 FCSN 的基础中引入了 ATPP 和 DenseCRF 两个模块，同时也对语义分割和视频摘要之间的联系做了深入的研究。所提模型在加入 ATPP 模块后，通过提取多尺度的上下文信息，在视频结构复杂、镜头多、场景变化频繁的视频中取得了较好的效果。在引入 DenseCRF 模块后，提升了模型在细节上的分类能力，使时间间隔相近且特征值相近的视频帧，能够较容易地被划分为同一个标签。实验结果表明，所提模型相比于传统的基于 LSTM 的模型在 TVSum 和 SumMe 数据集上展现出了更好的性能。然而，所提方法以及其他基于深度学习的视频摘要模型，在对输入视频进行预处理时，只提取了视频的视觉特征，并没有考虑视频中的其他特征，如音频特征等，这损失了视频中的大量信息。下一步将考虑发展基于多模态融合的视频摘要算法。

参 考 文 献

[1] Potapov D, Douze M, Harchaoui Z, et al. Category-specific video summarization[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8694: 540-555.

[2] de Avila S E F, Lopes A P B, da Luz A, Jr, et al. VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method[J]. Pattern Recognition Letters, 2011, 32(1): 56-68.

[3] Pritch Y, Rav-Acha A, Peleg S. Nonchronological video synopsis and indexing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1971-1984.

[4] Pritch Y, Rav-Acha A, Gutman A, et al. Webcam synopsis: peeking around the world[C]//2007 IEEE 11th International Conference on Computer Vision, October 14-21, 2007, Rio de Janeiro, Brazil. New York: IEEE Press, 2007: 9848979.

[5] Zhang K, Chao W L, Sha F, et al. Video summarization with long short-term memory[M]// Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9911: 766-782.

[6] Li Z T, Yang L. Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward[C]// 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), January 3-8, 2021, Waikoloa, HI, USA. New York: IEEE Press, 2021: 3238-3246.

[7] Lea C, Flynn M D, Vidal R, et al. Temporal convolutional networks for action segmentation and detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1003-1012.

[8] Lin R C, Xiao J, Fan J P. NeXtVLAD: an efficient neural network to aggregate frame-level features for large-scale video classification[M]// Leal-Taixé L, Roth S. Computer vision-ECCV 2018 Workshops. Lecture notes in computer science. Cham: Springer, 2019, 11132: 206-218.

[9] Rochan M, Ye L W, Wang Y. Video summarization using fully convolutional sequence networks[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11216: 358-374.

[10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.

[11] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]//4th International Conference on Learning Representations, ICLR 2016, May 2-4, 2016, San Juan, Puerto Rico. [S.l.: s.n.], 2016.

[12] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[13] He K M, Zhang X Y, Ren S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[14] Krhenbühl P, Koltun V. Efficient inference in fully connected CRFs with gaussian edge potentials[EB/OL]. (2012-10-20)[2021-03-10]. https://arxiv.org/abs/1210.5644.

[15] Zhang X F, Liu J, Shi Z S, et al. Review of deep

learning-based semantic segmentation[J]. Laser & Optoelectronics Progress, 2019, 56(15): 150003.
张祥甫, 刘健, 石章松, 等. 基于深度学习的语义分割问题研究综述[J]. 激光与光电子学进展, 2019, 56(15): 150003.

[16] Dong Y F, Yang Y X, Wang L Q. Image semantic segmentation based on multi-scale feature extraction and fully connected conditional random fields[J]. Laser & Optoelectronics Progress, 2019, 56(13): 131007.
董永峰, 杨雨沂, 王利琴. 基于多尺度特征提取和全连接条件随机场的图像语义分割方法[J]. 激光与光电子学进展, 2019, 56(13): 131007.

[17] Szegedy C, Liu W, Jia Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 15523970.

[18] Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 2650-2658.

[19] Gygli M, Grabner H, Riemenschneider H, et al. Creating summaries from user videos[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8695: 505-520.

[20] Song Y L, Vallmitjana J, Stent A, et al. TVSum: Summarizing web videos using titles[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 5179-5187.

[21] Mahasseni B, Lam M, Todorovic S. Unsupervised video summarization with adversarial LSTM networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2982-2991.

[22] Yuan L, Tay F E, Li P, et al. Cycle-SUM: cycle-consistent adversarial LSTM networks for unsupervised video summarization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 9143-9150.

[23] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2242-2251.