第 10 卷 第 6 期
2021 年 11 月
集 成 技 术
JOURNAL OF INTEGRATION TECHNOLOGY
Vol. 10 No. 6
Nov. 2021

引文格式：
李煜堃, 刘熠, 周林, 等. 短视频场景在线起始检测任务及方法研究 [J]. 集成技术, 2021, 10(6): 86-96.
Li YK, Liu Y, Zhou L, et al. On the online highlight start detection in short video scene [J]. Journal of Integration Technology, 2021, 10(6): 86-96.

短视频场景在线起始检测任务及方法研究

李煜堃[1,2] 刘 熠[1,2] 周 林[1,2] 何艾莲[1,2] 王亚立[1] 彭小江[1] 乔 宇[1*]

[1]（中国科学院深圳先进技术研究院 深圳 518055）
[2]（中国科学院大学深圳先进技术研究院 深圳 518055）

摘 要 现有视频在线检测研究所用数据集主要集中于长视频，且类别范畴相对单一，同时，需要设计符合在线形式需求的检测评价体系以满足日益增长的手机端短视频应用需求。该文提出一项短视频场景下在线精彩时刻起始检测的新任务，以辅助引导手机在拍摄过程中智能捕捉精彩时刻或实现其他短视频应用。具体实验包括：(1) 构建经过仔细时序标注的基于手机端拍摄的短视频数据集 Highlight45，用于填补新任务的训练和评估数据空缺；(2) 设定在线评价指标——首个检出的平均查准率，可视化结果显示该指标更加契合起始检测任务需求；(3) 设计带有序列对比损失函数的混合双流网络作为该任务的基线方法。实验结果显示，相比传统方法，该研究所提出的方法在已有起始检测指标和首个检出的平均查准率指标中分别取得了 6.98% 和 4.11% 的性能提升。

关键词 短视频数据集；在线起始检测；混合双流网络；序列对比损失函数
中图分类号 TP 399 **文献标志码** A doi: 10.12146/j.issn.2095-3135.20210318001

On the Online Highlight Start Detection in Short Video Scene

LI Yukun[1,2] LIU Yi[1,2] ZHOU Lin[1,2] HE Ailian[1,2] WANG Yali[1] PENG Xiaojiang[1] QIAO Yu[1*]

[1](*Shenzhen Institute of Advanced Technology*, *Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[2](*Shenzhen College of Advanced Technology*, *University of Chinese Academy of Sciences*, *Shenzhen* 518055, *China*)
[*]*Corresponding Author*: *yu.qiao@siat.ac.cn*

Abstract The existing datasets used in video online detection research are mainly concentrated on long videos and the category is relatively simple. At the same time, the detection and evaluation system that meets the needs of online setting is needed to meet the growing demand for short video applications on mobile phones. This paper proposes a new task called online highlight start detection (OHSD) in short video scenarios to assist in guiding the mobile phone to automatically capture highlights or other short video applications

收稿日期：2021-03-18 **修回日期**：2021-04-25
作者简介：李煜堃，硕士研究生，研究方向为视频行为分析；刘熠，博士研究生，研究方向为视频行为分析；周林，硕士研究生，研究方向为视频行为分析；何艾莲，硕士研究生，研究方向为视频行为分析；王亚立，副研究员，研究方向为计算机视觉；彭小江，副研究员，研究方向为计算机视觉；乔宇（通讯作者），研究员，研究方向为计算机视觉，E-mail：yu.qiao@siat.ac.cn。

during the shooting process. The specific experiment is as below: (1) construct a short video dataset called Highlight45, which is carefully temporally labeled based on mobile phone shooting, to fill the gaps in training and evaluation data for new tasks; (2) set the online evaluation metric——the average precision of the first detection, and the visualization results show that this metric is more suitable for the online start detection task requirements; (3) design a hybrid dual-stream network with sequence contrastive loss function as the baseline method for this task. The experiments show that, compared with the traditional method, the proposed method has achieved 6.98% and 4.11% performance improvements in the existing start detection metric and the average precision of the first detection respectively.

Keywords　short video dataset; online start detection; hybrid dual-stream network; sequence contrastive loss

1　引　言

视频行为理解因其在视频内容分析、智能监控、人机交互等方面的广阔应用前景而在人工智能和计算机视觉领域得到了广泛的研究。在学术界，关于视频理解已存在许多相关主题，例如修剪视频的行为分类[1]、未修剪视频中的行为识别[2]、时序行为检测[3]、时空行为定位[4]及视频高光时刻检测[5]。但是，这些任务所专注的研究都是离线设置下的，即需要以完整视频作为输入，待获取全部视频信息后输出结果。

自 2016 年以来，先后出现了一些在线设置下的视频动作研究工作，即在仅获取过去及当前视频帧信息的条件下输出当前结果，如在线动作检测[6]、在线动作预测[7]和在线动作起始检测[8-9]。在线任务除了需要解决离线设置下视频理解任务的所有难点以外，还需要解决视频帧下文信息不足的问题。因此，与离线任务相比，在线任务更具挑战性。目前，上述任务的研究大多是在原本作为离线设置下动作检测的数据集上完成的，如 THUMOS'14[10]、ActivityNet[2]或一些从电视剧集中获取的长视频数据集（如 TVseries[11]）。这些数据集的视频通常平均时长为数分钟甚至数十分钟，并且类别基本集中在人的动作上。尽管短视频在诸多移动端应用中无处不在，但丰富场景下的短视频在线检测的相关研究依然有所欠缺。

基于此，本文着眼于手机端短视频的视频理解提出在线精彩时刻起始检测（Online Highlight Start Detection，OHSD）任务。OHSD 任务的研究有助于在手机端 AI 相机应用中实现智能启用慢动作录制或触发其他预设特定效果。为了适应此任务的研究，本研究首先采集和构建了一个名为 Highlight45 的大规模手机短视频数据集。该数据集包含来自日常生活中 45 个不同类别的 9 751 个高分辨率手机拍摄视频。这些类别的设定主要是通过调研手机用户在拍摄过程中的偏好确定，定义为精彩时刻（Highlight）：一方面，因为本数据集中的类别不仅局限于人类动作，还包括自然场景、动物、人物交互等大类，因此需要与之前的动作检测进行区分；另一方面，这一定义也契合本研究所关注的应用场景。数据集中所有视频均是未经裁剪的原始手机视频，并对每一个视频进行了精彩时刻起止点的标注。针对 OHSD 任务，本研究设计了两个评测指标以评估在线起始检测的效果：首次检测时的平均查准率（Average Precision@First，AP@1）和平均次数的平均召回率（Average Recall@Average Number，AR@AN）。具体来说，前者侧重于在线评估，仅考虑网络输出的首个检测结果；而后者则对完整视频

处理完后的所有检测结果进行整体评价。

实验部分给出了在线检测任务中常用的基于递归神经网络系列的几种网络的基准结果，并设计了一种基于带孔时序卷积的网络结构(Highlight-Net)以更好地利用图像色彩(RGB)信息和光流(Flow)信息。为了更好地解决起始检测任务中背景帧和前景帧之间难以区分的问题，本研究进一步设计了序列对比损失函数。实验结果表明，新的网络结构及损失函数显著地提升了检测效果，可以作为 OHSD 任务很强的一个基线方法。在最后，本研究通过具体类别的实例分析，阐明了以往在线评价指标存在的问题和本研究所提出的评价指标的合理性。

2 在线起始检测数据集介绍

2.1 数据采集

鉴于手机短视频场景下视频在线检测及在线起始检测任务数据集较少，本文构建了一个名为 Highlight45 的大规模手机短视频数据集：首先，通过调研日常生活中手机拍摄精彩时刻视频内容确定了涵盖动物、人类行为、人物交互和场景 4 大类型共计 45 个类别的设定；然后，以众包的方式收集视频以确保每个类别内容的多样性，并经过人工逐个检查视频质量，剔除了分辨率低、摄像机运动剧烈等低质量视频。最终形成的数据集中每个类别均有约 200 个视频，共计 9 751 个视频，以保证样本平衡。为了获得尽可能准确的精彩时刻起始标注，首先对每个类别提供了起始判定的参考依据并给出参考实例。考虑到手机短视频的特性及本任务应用的侧重点，每个视频仅标注 1 个实例。经过统计，本数据集中大部分视频帧数少于 200，同时有很大比例视频的精彩时刻持续帧数少于 20，这意味着对本数据集精细化的时序起始检测将更具难度。从结果分析来看，表现不好的类别也确实是这些帧数少的类别。图 1 展示了本数据集的统计特性。

2.2 任务定义及评价指标

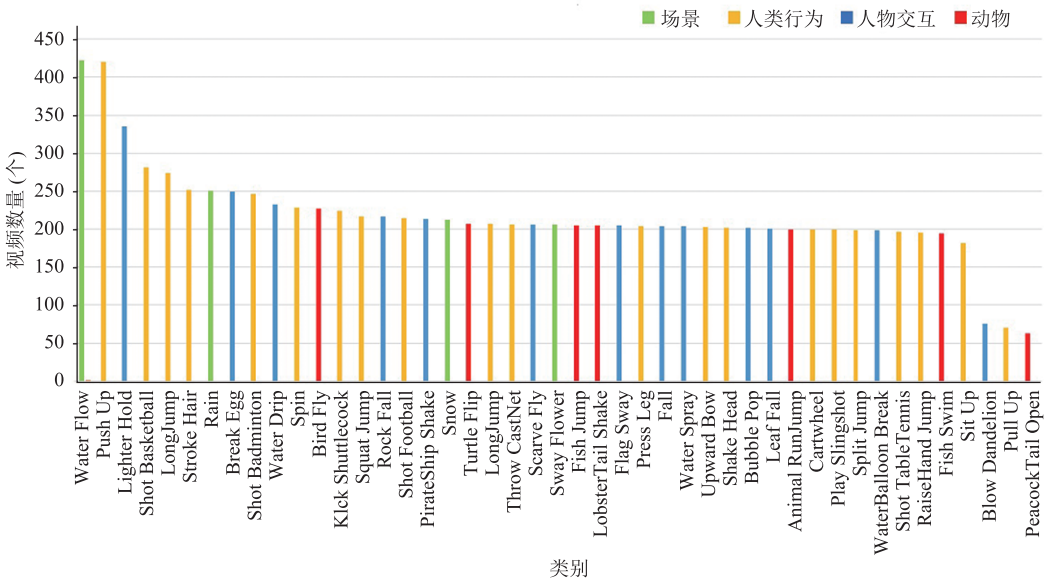
2.2.1 在线起始检测任务

对于 OHSD 任务而言，网络需要在仅获取过去和当前帧信息的情况下，输出当前帧的类别以及 Highlight 分数。整体而言，首先通过特征提取网络提取帧级别特征，然后使用时序建模模块集成历史信息以帮助当前帧的分类，最后使用分类器来判断 Highlight 事件的起始。具体来说，可以划分为类别相关和类别无关两个子任务。在类别相关的设置中，网络除了需要输出判定为事件起始的分数以外还需要作出正确的分类，而类别无关的设置下仅需要给出起始的分数而对类别正确与否没有要求。考虑到实际应用中具体到帧级别的起始点判定方式过于严苛，同时不同类别实际上有不同程度的检测敏感度要求，因此本文提出自适应时间容差窗口(Time Tolerance Window)的概念，即网络判定的起始帧只要落在实际标注的起始帧前后若干帧内即算正确，窗口大小与实例时长相关。

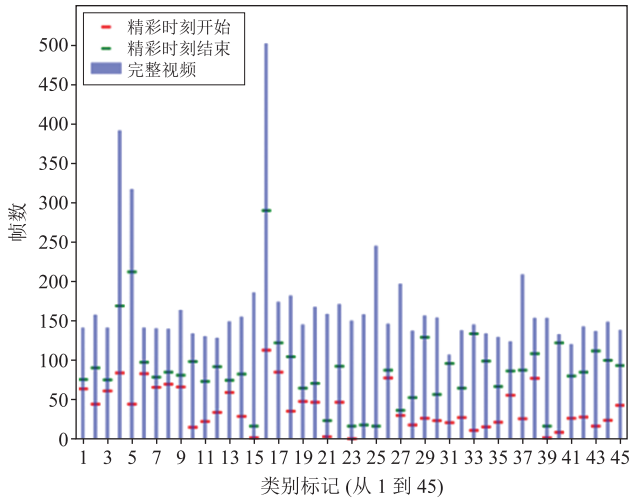
2.2.2 评价指标

参考以往在线检测任务中，在线动作检测(Online Action Detection)[11]通常使用帧级别平均查准率(frame Average Precision，frame-AP)以及考虑了平衡背景影响的帧级别校准平均查准率(calibrated Average Precision，cAP)两个指标。鉴于这两个指标主要适用于在线帧分类问题而不适用于起始检测，有学者在在线起始检测任务中[9]提出了点级别平均查准率(point-level Average Precition，p-AP)——更多地适应于多实例视频下的起始点评价(如 THUMOS'14 数据集)。然而，该指标统计评价整个视频所有的检测结果，并不能适配 OHSD 任务中面向实际应用的情形。

对于短视频场景下的 OHSD 任务，往往关注算法能否及时正确地输出首个起始点的检测结果。因此，本研究引入了两个新颖的视频级评价



(a) 每个类别的视频数量



(b) 每个类别的平均帧数

图 1　Highlight45 数据集统计特性

Fig. 1　Statics of Highlight45

标准——AP@1 和 AR@AN。为了更加公平地评估不同难度下的起始查准率，使用实例自适应时间容差窗口来判定正确的预测。对于每个视频，实例自适应时间容差窗口的定义为：

$$\left[S_{\mathrm{gt}}-\alpha\left(E_{\mathrm{gt}}-S_{\mathrm{gt}}\right),S_{\mathrm{gt}}+\alpha\left(E_{\mathrm{gt}}-S_{\mathrm{gt}}\right)\right] \tag{1}$$

其中，S_{gt} 和 E_{gt} 分别为视频中标注的起始和结束时间点；α 为偏移容差系数，其大小决定了评价指标的严格程度，在实验中，该系数分别设置为 0.1、0.2、0.3 以进行比较。

AP@1 是完全在线评估的指标，算法不能在处理完全部视频后进行后处理（如按照分数进行排序筛选），仅提供输出的首个检测结果作为评判。形式上 AP@1 可以表示为：

$$AP@1=\frac{\sum_{i=1}^{N}I\left[S_0(i)\right]}{N} \tag{2}$$

其中，N 为参与评价的视频总数；$S_0(i)$为第 i 个视频中首个起始检测时间点(帧号)；$I(\cdot)$为指示函数，如果检测到的帧号落入上述时间窗口内，则判定为 1，否则为 0。AP@1 反映了所有视频在线输出起始检测的正确比例。

AR@AN 作为离线评价指标进行辅助评价，允许算法在处理完全部视频后进行后处理。对于每个视频，系统首先将所有检测出的起始结果按其置信度分数排序，然后将前 N 个预测结果用于召回率评估。若前 N 个预测结果中有任何一个落入上述时间窗口，则正确值加 1。在本数据集的设置下，每个视频只有一个实例需要判断，因此设定 AN(Average Number)的值为 1 和 2 进行评价。

3 网络设计

3.1 混合双流网络

在线动作分析的方法流程通常是先通过双流网络分别提取 RGB 和 Flow 特征，然后将两种模态的特征拼接起来作为后续网络框架的输入。这种简单且直接的先融合策略虽然同时利用了两种模态的信息，但可能会使后续时序建模网络对外观和运动特征产生混淆，导致对时间维度信息更敏感的在线设置下任务产生更为显著的混淆。经过实验可以验证，在线起始检测任务中，在帧级别的 RGB 特征上添加时序建模模块对整体性能有负面影响。在线时序检测任务中时序建模网络通常使用循环神经网络(如 LSTM[12]和 GRU[13])，但 Wang 等[14]指出，使用带孔因果卷积的效果会优于循环神经网络。因此，本研究针对 OHSD 任务设计了一种基于带孔因果卷积的混合双流网络结构(Highlight-Net)从而更有效地利用两种不同模态的特征。

图 2 展示了该网络结构的整体流程图。整个网络划分为 RGB 分支和 Flow 分支。对于 RGB 分支，为了最大限度保留帧本身的信息，采用图像领域常用的卷积神经网络 ResNet50[15]进行帧级别的特征提取和分类。对于 Flow分支，首先采取 BN-Inception[16]对过去 $L-1$ 帧及当前帧的光流输入进行特征提取，这些特征通过全连接层和 ReLU 激活函数变形后组成片段级特征序列；然后将此特征序列输入名为“带孔因果卷积”(Dilated Casual Convolution，DCC)[17]的时序建模模块中，用来替代之前序列任务一般采用的循环神经网络。因果卷积保证了网络的“在

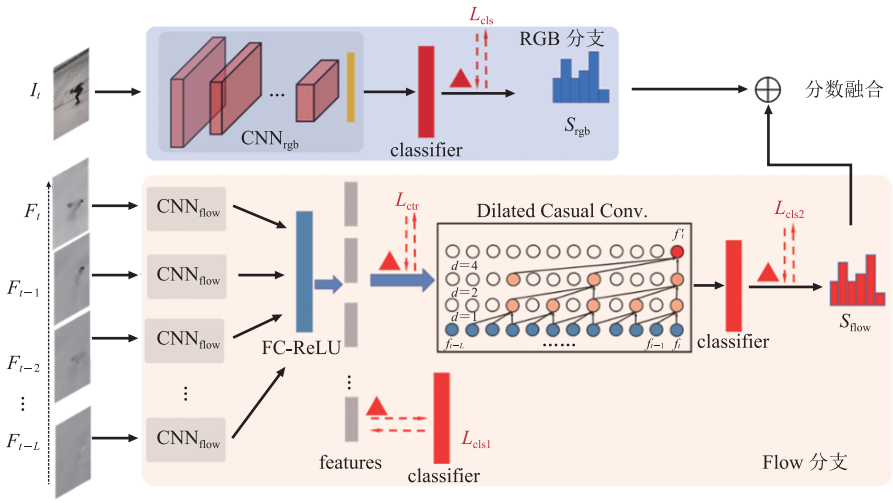


图 2 Highlight-Net 网络结构

Fig. 2 Network architecture of Highlight-Net

线”性质，同时带孔卷积保证了对长时历史信息的保留。

形式上来说，给定时序长度为 L，特征维度为 d 的输入 $F=\{f_1,f_2,...,f_L\}\in\mathbf{R}^{d\times L}$，DCC 模块将产生与输入时序同样长度的输出 $F_o=\{f_{o1},f_{o2},...,f_{oL}\}\in\mathbf{R}^{d_o\times L}$，每个时刻 t 的输出为：

$$f_{ot}^{(j)}=\mathrm{ReLU}\left(\sum_{i=1}^{s}f_{[t+r\cdot i]}\times W_{[i]}\right) \tag{3}$$

其中，r 为带孔卷积比例，反应时序卷积在采样帧的间隔帧数；W 表示尺寸为 s、通道数为 d_o 的 1D 卷积核。RGB 和 Flow 分支的分数最终通过加权求和的方式进行后融合以作为当前帧最终的起始判定分数。

3.2　序列对比损失

在 Flow 分支中，特征以时序序列的形式进行输入。由于视频任务中帧的连续性特点，起始点前后若干帧的特征十分相近，从而造成网络难以准确检测起始点。基于此，本研究分析数据特性，设计了序列对比损失函数(Sequential Contrastive Loss)以监督网络增大背景帧和前景帧光流特征建模的区分度，从而使最终输出的起始点更加准确。

对于长度为 L 的特征序列，以相邻两帧作为一对计算对比损失。形式上可以表述为：

$$\begin{aligned}L_{\mathrm{ctr}}=&\frac{1}{L-1}\sum_{i=1}^{L-1}y_{(i,i+1)}D^2\left(f_i,f_{i+1}\right)+\\&\left(1-y_{(i,i+1)}\right)\max\left[0,m-D^2\left(f_i,f_{i+1}\right)\right]\end{aligned} \tag{4}$$

其中，y_i 为 L 序列中第 i 帧的类别，当相邻两帧类别相同时(都是前景或背景)，$y_{(i,i+1)}$的值为 1，否则为 0；$D^2\left(f_i,f_{i+1}\right)$为相邻两帧特征之间的欧氏距离；$m$ 为调整损失函数边界的参数，实验中设置为 2。

3.3　实验设置细节

在训练阶段，RGB 分支和 Flow 分支独立进行。对 RGB 分支，先采用在 ImageNet[18]上预训练的 ResNet50 模型进行初始化，然后基于标准交叉熵损失(图 2 中的 L_{cls})在 Highlight45 数据集上进行精调。对 Flow 分支，先采用在 Kinetics400[19]数据集上预训练的 BN-Inception 模型进行初始化，然后进行图像级的精调以作为单帧特征抽取器。对于 Flow 分支的时序建模部分，在一个 FC-ReLU 整合特征模块之后，采用两层带孔因果卷积层(卷积核尺寸为 3，带孔尺寸分别为 1、2，通道数为 3 072)作为时序建模模块。每一层卷积层之后均使用 ReLU 和 dropout 来控制过拟合。光流部分的损失函数可形式化表示为：

$$L_{\mathrm{flow}}=L_{\mathrm{cls2}}+\lambda\left(L_{\mathrm{cls1}}+L_{\mathrm{ctr}}\right) \tag{5}$$

其中，L_{cls1} 和 L_{cls2} 均为交叉熵损失函数，L_{cls1} 作为一项提前监督的策略来更加有效地训练深层网络；λ 为平衡权重因子，在实验中设置为 0.5。

数据处理阶段，首先将短边为 256 的图片按长短边比例调整尺寸，然后进行中心裁剪为 224×224 大小的图片作为输入。网络训练过程中，使用动量为 0.9 的 SGD 优化器，正则化权重为 0.000 5，批处理规模为 64。初始学习率为 0.001，在第 3 和第 5 个迭代周期学习率衰减 10 倍，共训练 10 个训练周期。所有代码基于 Pytorch 框架进行实现，使用 8 张 NVIDIA RTX 2080Ti GPU 显卡进行所有实验。

在测试阶段，Highlight-Net 以间距为 1 的滑动窗口对输入视频流进行逐帧在线检测。RGB 分支逐帧处理视频流，Flow 分支处理长度为 L 的帧序列。为了对齐两分支的当前帧位置，在测试开始阶段对每个视频的开头添加空帧以填补 Flow 分支的空缺。通过加和两个分支的全连接层输出分数来融合二者的特征信息后，使用 Softmax 函数来获得用于 OHSD 任务的多分类或二分类概率。进一步，通过计算当前帧相对前一帧预测为前景概率值的差作为判断当前帧是否为起始帧的依据。在类别相关的设定下，还需要判

定该帧的前景帧对应分类是否与前一帧一致。

4 实验与讨论

在本节中，首先对本研究提出的 Highlight-Net 和一系列经典方法进行比较，然后通过消融实验评估新提出的混合双流结构及序列对比损失函数的提升效果。

4.1 实验结果

OHSD 任务最直观的方法是进行逐帧分类。因此，本研究最基本的对比方法是直接拼接 RGB 和光流的双流特征(Two Stream feature，TS)作为后续网络输入的逐帧分类器。实验中分别采用直接使用全连接层分类器以及递归神经网络(LSTM 和 GRU)对 TS 特征进行时序建模后分类的方式作为基线方法。为了保证对比公平，所有的特征提取网络均保持一致，时序建模层的通道数也与 Highlight-Net 中 DCC 模块的通道数相同。

对于 OHSD 任务的性能评估，除了使用本文提出的更符合任务设定的指标(AP@1 和 AR@AN)进行测评以外，还采用了在线动作起始检测工作[9]中提出的 p-AP 指标。由于 p-AP 指标中起始评测范围(1～10 s)是针对长视频数据集设定的，与本文构建的短视频数据集的数据特性不相符，因此在评价过程中需对该指标进行修正(0.5～5 s)。所有实验均在两套判定体系(类别相关和类别无关)下进行评估。偏移容差系数 α 对应任务的难度，分别按 0.1、0.2、0.3 进行评估。

从表 1 中 Highlight-Net 和其他经典方法之间的实验结果可知：(1)从本文提出的新指标及在线动作起始检测中采取的 p-AP 评价标准来看，混合双流网络的方法始终表现更好，尤其是在类别无关的设定下，提升效果非常显著，这一定程度上反映了 RGB 特征的拼接会影响后续效果；(2)从 AP@1 和 AR@AN 的角度来看，当容差系数变小时，所有方法的性能都会显著降低，这说明精确检测起始点非常困难；(3)相对于只使用当前帧进行分类的方式，LSTM 和 GRU 在这两套判定体系下均有一定提升，说明历史时序信息的融合有助于起始检测。

4.2 消融实验

本节对双流模态融合方式、时序建模方法、序列对比损失以及提前监督策略进行消融实验研究，旨在证明本文提出方法的有效性。所有实验均在类别相关、偏移容差系数为 0.2 的设定下以 AP@1 和 AR@1 两大指标进行对比实验。

4.2.1混合策略及时序模型

表 2 展示了使用不同混合策略及时序模型的组合进行实验，共计 13 个模型的性能比较。Flow 分支均使用本文设计的提前监督策略和序

表 1 Highlight45 上各种指标下 OHSD 任务的实验结果

Table 1 Results of online highlight start detection with varied metrics on Highlight45

设定	方法	AP@1(%)			AR@1(%)			AR@2(%)			p-AP(%) 0.5~5 s
		α=0.1	α=0.2	α=0.3	α=0.1	α=0.2	α=0.3	α=0.1	α=0.2	α=0.3	
类别相关	Baseline(TS)	62.61	74.38	80.95	58.63	71.12	78.64	64.68	78.12	85.47	84.36
	TS w/LSTM	65.15	75.33	81.49	66.28	77.40	84.17	69.34	80.97	87.59	86.01
	TS w/GRU	64.01	75.49	82.21	64.27	75.85	82.68	67.58	79.68	86.61	87.05
	Highlight-Net	68.30	78.49	83.45	69.90	80.19	85.26	74.32	84.17	89.29	91.14
类别无关	Baseline(TS)	40.93	57.98	68.00	25.88	36.82	45.19	33.28	48.74	59.27	66.56
	TS w/LSTM	52.67	64.85	72.11	43.31	50.44	61.15	50.17	62.19	78.25	68.13
	TS w/GRU	51.03	63.77	72.79	41.98	48.94	60.09	48.88	60.36	77.68	67.08
	Highlight-Net	66.92	73.60	77.09	69.79	78.12	81.77	79.82	88.69	92.03	70.62

列对比损失函数进行优化从而保证对比的公平性。结果表明，(1)在两个指标中，由于是类别相关的设定，在不使用时序建模的情形下，仅使用 RGB 特征的性能略优于 Flow 特征。(2)时序建模可以显著提升 Flow 分支的效果，但对 RGB 特征却会有所损害。具体来说，AP@1 指标中，Flow 分支提升 4.75%，而 RGB 分支却下降 2.21%，这一现象说明了混合双流结构的必要性，时序建模对空间特征的融合并不友好。(3)对比不同时序模型，不论是哪种特征输入方式，DCC 均略优于 LSTM 和 GRU。(4)相较于特征拼接的输入方式，使用混合模型可以显著改善效果，带有 DCC 的 Highlight-Net 可获得最佳结果。

表 2　特征不同混合方式和时序模型的对比实验

Table 2　Evaluation of hybrid strategy and temporal modeling methods

模型	方法	AP@1/α=0.2 (%)	AR@1/α=0.2 (%)
0	Baseline(TS)	74.38	71.12
1	RGB	69.30	70.63
2	Flow	68.90	69.12
3	RGB w/DCC	67.09	68.42
4	Flow w/DCC	73.65	73.75
5	Flow w/LSTM	72.85	71.92
6	Flow w/GRU	72.59	71.78
7	TS w/DCC	75.81	78.04
8	TS w/LSTM	75.33	77.40
9	TS w/GRU	75.49	75.85
10	Highlight-Net w/LSTM	78.17	79.71
11	Highlight-Net w/GRU	77.95	79.29
12	Highlight-Net w/DCC	78.49	80.19

4.2.2 损失函数

表 3 中评估了 Highlight-Net 采用的序列对比损失函数和提前监督策略，由于这两个损失函数仅作用在 Flow 分支，因此表中仅对比使用光流模态的结果，时序建模网络使用 DCC 模块。从表 3 可以看出，二者一致地提高了性能。这说明序列对比损失可以监督时序建模网络更有效地将起始点前后前景帧、背景帧特征进行区分，从而更好地服务于后续起始检测任务。另外，提前监督的策略辅助了整个模型的优化。这二者共同将首个检出的平均查准率 AP@1 提高 2.82%。

表 3　损失函数的对比实验

Table 3　Evaluation of the loss functions

L_{ctr}	L_{cls1}	AP@1/α=0.2 (%)	AR@1/α=0.2 (%)
		70.83	70.25
	√	71.92	72.64
√		72.13	72.38
√	√	73.65	73.75

4.2.3 评价指标

本节挑选 3 个典型类别(Animal RunJump、Squart Jump、Throw Castnet)及全部数据(Whole data)进行多个指标的评估对比并对具体例子进行可视化，用以说明本研究设计指标的合理性。如图 3 所示，frame-AP 是在线动作检测的评价指标，用以统计所有被判定为前景帧的查准率，不能反映起始区域帧的准确程度，从结果上反映出来每一类下该指标的数值都非常高。p-AP 指标对网络所有检测出的起始提名按置信度进行排序，若起始提名位于统一的固定时间偏移(如 0.5 s)中，则认为该提名是正确的，所有符合要求的检测都会纳入计算。本研究提出的指标与 p-AP 之间的主要区别在于：(1)AP@1 和 AR@AN 使用实例自适应时间偏移窗口，更契合视频长短不同情形的不同检测需求；(2)AP@1 仅评判首个检出的结果，而 p-AP 指标需要全部视频作出输出才进行计算，本指标更符合在线要求和实际需求；(3)AP@1 和 AR@AN 在视频数量级别进行平均。

综上所述，p-AP 的评价会受到时间偏移量和当前视频中非首个检测产生的误报的影响，以 Animal RunJump 类为例可以发现，该类别的 p-AP 极低，但 AP@1 高出两倍以上。从可视化

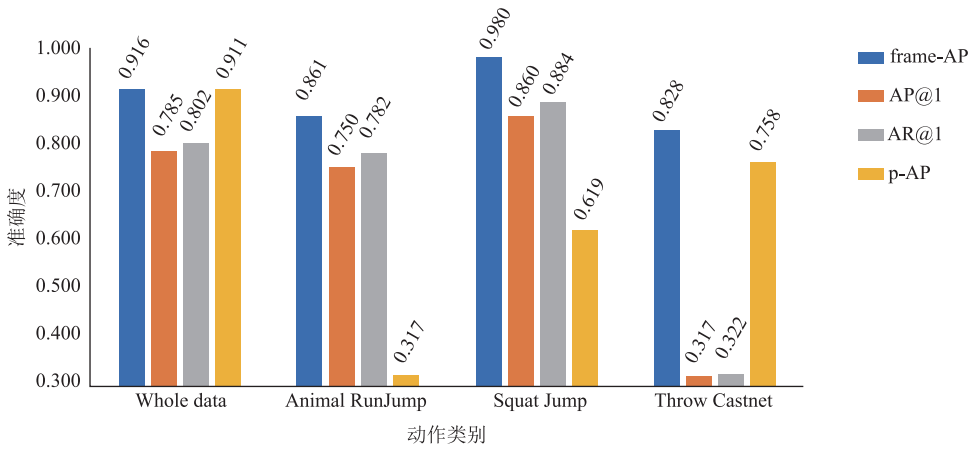
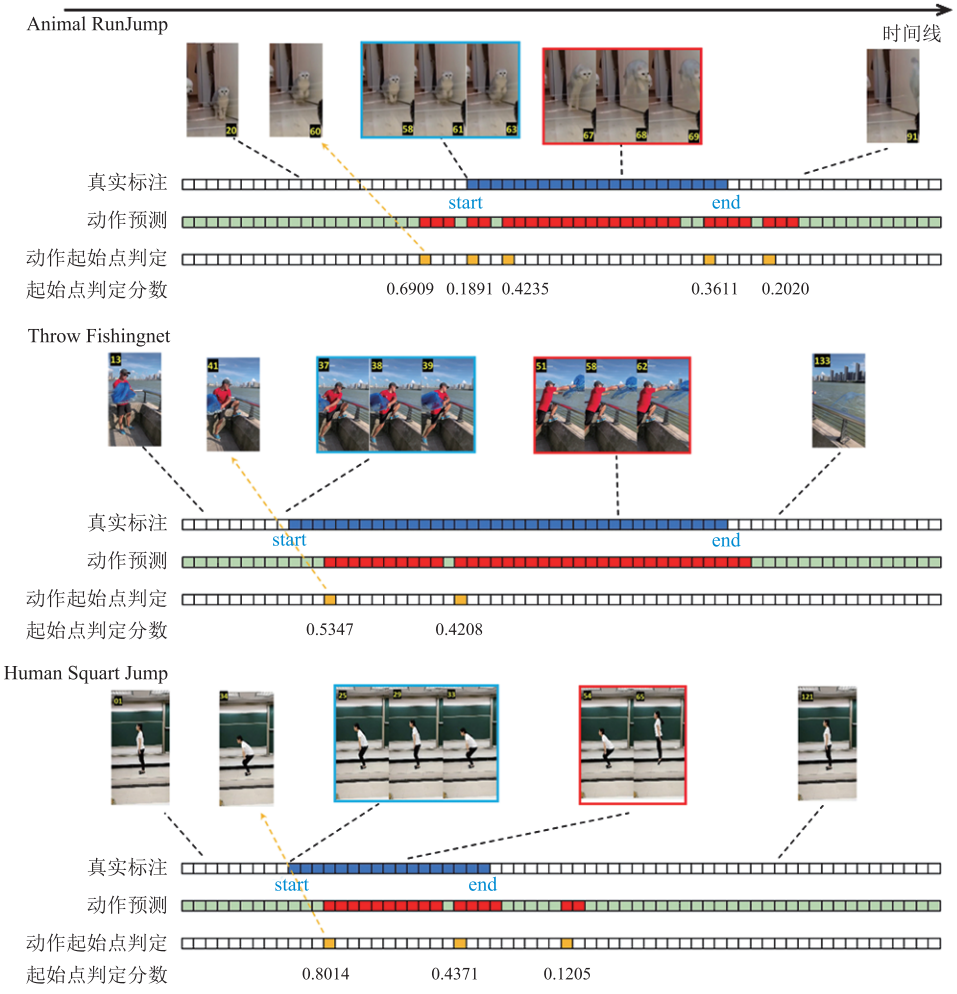


图 3 不同评价指标详细对比

Fig. 3 Comparison between different evaluation metrics



注：每一方格简化代表每一帧

图 4 可视化实例示意图

Fig. 4 Visualization of online highlight start detection

情况来看，这一类别的首个实际预测往往是正确的，但 p-AP 指标由于会考虑超过阈值的全部起始预测并且按照置信度排序，从而拉低了整体结果。而另一类 Throw Castnet 则正好相反，这是因为该类 Highlight 持续时间普遍较短，与 p-AP 固定时间窗口模式相比，本研究指标中自适应窗口模式会判定更多的正确预测。实际需求下，时长短的类别往往需要更灵敏的起始检测，所以 AP@1 可以更好地反映类别难度，与该类别相似的几个类别的起始检测效果均不理想。图 4 中给出了 3 个类别典型例子的可视化结果图来帮助解释上述情形。

5 结　论

本研究基于手机短视频场景提出在线起始检测任务(OHSD)并配套构建了 Highlight45 数据集和契合 OHSD 任务需求的两个新的评估指标，即 AP@1 和 AR@AN。类别相关和类别无关设定下的大量实验表明，与传统评估指标相比，本研究的度量标准更合理实用。针对 OHSD 任务，本研究设计了 Highlight-Net 网络结构，通过探索全新的双流融合策略和使用新的损失函数监督取得了较好的检测效果，以作为强有力的基线方法。短视频研究的应用前景十分广阔，本研究仅初步进行了数据、评价标准和方法上的探索，未来可以在数据的扩充和方法的优化上进行更多的研究。

参 考 文 献

[1] Soomro K, Zamir AR, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild [Z/OL]. arXivPreprint arXiv: 1212.0402, 2012.

[2] Heilbron FC, Escorcia V, Ghanem B, et al. ActivityNet: a large-scale video benchmark for human activity understanding [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 961-970.

[3] Lin T, Zhao X, Su H, et al. BSN: boundary sensitive network for temporal action proposal generation [C] // Proceedings of the European Conference on Computer Vision, 2018: 3-19.

[4] Gu C, Sun C, Ross DA, et al. AVA: a video dataset of spatio-temporally localized atomic visual actions [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6047-6056.

[5] Sun M, Farhadi A, Seitz S. Ranking domain-specific highlights by analyzing edited videos [C] // European Conference on Computer Vision, 2014: 787-802.

[6] Gao J, Yang Z, Nevatia R. RED: reinforced encoder-decoder networks for action anticipation [Z/OL]. arXivPreprint arXiv: 1707.04818, 2017.

[7] Gao M, Xu M, Davis LS, et al. StartNet: online detection of action start in untrimmed videos [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 5542-5551.

[8] Xu M, Gao M, Chen YT, et al. Temporal recurrent networks for online action detection [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 5532-5541.

[9] Shou Z, Pan J, Chan J, et al. Online detection of action start in untrimmed, streaming videos [C] // Proceedings of the European Conference on Computer Vision, 2018: 534-551.

[10] Idrees H, Zamir AR, Jiang YG, et al. The THUMOS challenge on action recognition for videos “in the wild” [J]. Computer Vision and Image Understanding, 2017, 155: 1-23.

[11] Geest RD, Gavves E, Ghodrati A, et al. Online action detection [C] // European Conference on

Computer Vision, 2016: 269-284.

[12] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[13] Cho K, Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN Encoder-Decoder for statistical machine translation [J/OL]. arXiv preprint arXiv: 1406.1078, 2014.

[14] Wang W, Peng X, Qiao Y, et al. A comprehensive study on temporal modeling for online action detection [Z/OL]. arXivPreprint arXiv: 2001.07501, 2020.

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[16] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C] // International Conference on Machine Learning, 2015: 448-456.

[17] Oord A, Dieleman S, Zen H, et al. WaveNet: a generative model for raw audio [Z/OL]. arXivPreprint arXiv: 1609.03499, 2016.

[18] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database [C] // 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.

[19] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the Kinetics dataset [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.