面向行为边界框生成的端到端时间全局相关网络

马百腾[1],张士伟[2],高常鑫[1],桑 农[1]
(1. 华中科技大学人工智能与自动化学院图像信息处理与智能控制教育部重点实验室,湖北武汉 430000;
2. 阿里巴巴达摩院科技有限公司,浙江杭州 310000)

摘 要: 时序行为边界框生成任务的目的是定位未剪辑视频中行为的开始和结束时间. 现有的生成行为边界框的方法存在两个缺点:所使用的特征不具有足够的时间全局信息,导致了边界框的不准确;特征提取和边界框生成的过程是分开的,导致生成的特征不完全适合边界框生成任务. 为了解决上述问题,本文提出了时间全局相关网络(Temporal Global Correlation Network,TGCNet),利用时间全局相关(Temporal Global Correlation,TGC)模块获取全局信息. TGC模块主要包含动态相关结构和静态相关结构,分别编码动态和静态全局信息. TGCNet网络可以以端到端的方式训练,使得所学习到的特征更适合时序行为边界框生成任务. 本文在两个具有挑战性的数据集THUMOS14和ActivityNet1.3上进行了实验,结果表明,所提出的TGCNet网络在这两个数据集上均达到了最好的时序行为边界框生成性能.

关键词: 时间全局信息;时间全局相关模块;时间全局相关网络;时序行为边界框生成;时序行为检测
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)10-2452-10
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20201302

Temporal Global Correlation Network for End-to-End Action Proposal Generation

MA Bai-teng[1], ZHANG Shi-wei[2], GAO Chang-xin[1], SANG Nong[1]
(1. *Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, Hubei* 430000, *China*;
2. *Alibaba DAMO Academy, Hangzhou, Zhejiang* 310000, *China*)

Abstract: The purpose of the temporal action proposal generation task is to locate the start and end time of the action in the untrimmed video. The existing methods of temporal action proposal generation are suboptimal because of two reasons: the applied features cannot encode sufficient temporal global information, which may result in imprecise proposals; the procedures of feature extracting and proposal generating are separate, hence the features may be not completely suitable for the proposal generation task. To solve this problem, we propose the temporal global correlation network (TGCNet) by repeatedly embedding well designed temporal global correlation (TGC) module to encode temporal global information. Specifically, the TGC module mainly contains a dynamic correlation structure and a static correlation structure, which target to encode dynamic and static global information, respectively. Most importantly, TGCNet can be trained in an end-to-end manner, which makes the features leaned by TGCNet are more suitable for action proposal generation. We perform experiments on two challenging datasets: THUMOS14 and ActivityNet1.3, and the results show that the proposed TGCNet achieves state-of-the-art temporal action proposal generation performance on the both datasets.

Key words: temporal global information; Temporal Global Correlation module; Temporal Global Correlation Network; temporal action proposal generation; temporal action detection

1 引言

时序行为边界框生成是时序行为检测的重要组成部分. 该任务的目的是在未修剪的视频中定位行为的开始和结束位置,以及生成具有精确的时间边界和可靠的置信度的边界框. 现有的方法可以分为两类:基于锚的方法(anchor-based methods)[1~3]和基于边界的方法

收稿日期:2020-11-18;修回日期:2021-03-01;责任编辑:王天慧
基金项目:国家自然科学基金(No. 61871435)

(boundary-based methods)[4~6]. 基于锚的方法使用一系列手动设置的锚来定位行为的时间边界,因此不适用于动作持续时间变化很大的情况. 与之相反,基于时间边界的方法能够评估视频每个时间位置的开始和结束的概率,因此可以生成具有精确时间边界的边界框且适用于动作持续时间变化很大的情况. 例如,BSN[4]预测每个时间位置为行为开始和结束位置的概率以生成边界框,并使用这些边界框的全局信息来生成置信度;BMN[5]使用边界匹配机制(boundary-matching mechanism),利用丰富的上下文信息将边界框生成和置信度生成结合在一起;DBG[6]利用边界框的全局信息来生成行为的开始和结束位置的概率.

上述基于时间边界的网络中,BSN 和 BMN 没有使用边界框的全局信息和视频的全局信息,而 DBG 虽然使用了边界框的全局信息,但并未使用视频的全局信息,这导致它们对整个视频均没有很好的理解. 另外,上述网络的边界框生成和特征提取不是端到端训练的. 图 1(*a*) 表示了上述网络的训练方式. 首先,由骨干网络提取特征,然后将特征送入"边界框生成"部分以生成边界框,最后计算损失(Loss). 在这种训练方式中,梯度仅反向传播到特征图,而没有到达骨干网络,因此无法根据损失调整骨干网络. 特别是这些方法的骨干网络的训练和微调是基于视频的分类信息来进行的,因此通过骨干网络获得的特征更倾向于分类任务,而不是时序行为边界框生成任务. 为了解决上述问题,本文提出时间全局相关网络(Temporal Global Correlation Network,TGCNet)来探索全局信息和端到端训练方式的优势. 为了使网络能够使用视频的全局信息以生成精确的时间边界,本文提出了时间全局相关(Temporal Global Correlation module,TGC)模块,该模块由使用注意力机制的动态相关结构和使用大卷积核的静态相关结构组成,它们可以分别生成动态和静态的全局信息以促使精确的时间边界的生成. 为了获得更好的用于时序行为边界框生成的特征图,本文采用了端到端的训练方式. 图 1(*b*) 显示了本文的端到端的训练方式. 该训练方式使得梯度反向传播到骨干网络,因此整个网络都可以同时被训练并且生成的特征图更有利于精确的时间边界框的生成. 特别是因为端到端的训练方式和 TGC 模块存在相互作用的关系,相较于仅引入 TGC 模块而不使用端到端训练方式的情况,可以进一步提升时序行为边界框生成的性能.

总而言之,本文的工作有 3 个主要贡献.

(1)提出了一种时间全局相关网络(TGCNet),可以根据使用注意力机制的动态相关结构和使用大卷积核的静态相关结构来分别生成动态和静态全局信息,从而为行为边界框提供精确的行为边界和可靠的置信度得分.

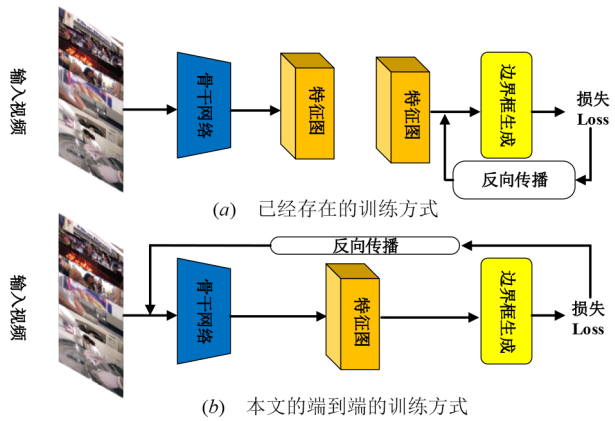

图 1　已经存在的训练方式和本文的端到端训练方式的比较

(2)采用端到端的训练方式有利于发挥时间全局相关(TGC)模块的作用.

(3)在两个具有挑战性的数据集 THUMOS14 和 ActivityNet1.3 上实现了最先进的性能,这证明了时间全局相关网络(TGCNet)在时序行为边界框生成任务上比其他方法更好.

2　相关工作

2.1　行为识别

行为识别是时序行为提议生成任务的基础. 较早的方法[7]使用手工制作的特征,例如 HOG,HOF 和 MBH. 近年来,深度学习方法极大地提高了行为识别任务的准确性. 行为识别网络可分为两类:双流网络[8]主要从 RGB 图像和堆叠的光流中探究视频的表现(appearance)和运动线索(motion clues);3D 网络[9~12]利用时空卷积探索视频的表现(appearance)和运动线索(motion clues).

2.2　全局相关性

全局信息被广泛用于许多计算机视觉任务,例如语义分割、行为识别. 虽然近年的工作[13~16]证明了全局信息的有效性,但其尚未在时序行动边界框生成任务中得到应用. 其中,文献[16]使用注意力机制来构建局部特征的丰富的上下文依赖性,从而提高语义分割效果;文献[13]使用全卷积来建立局部与全局之间的关系.

2.3　时序行为边界框生成

时序行为边界框生成任务的目标是为边界框生成精确的行为边界和可靠的置信度分数. 基于锚的方法可以使用预设的锚来生成边界框. SCNN[3]使用 C3D[11]网络来选择包含行为的边界框. TURN[2]使用单元回归生成边界框. 基于边界的方法评估视频中的每个时间位置以生成边界框. BSN[4]生成开始和结束位置的概率以生成边界框. BMN[5]提出了一种边界匹配机制来为密集分布的边界框生成置信度. DBG[6]使用边界框的

全局信息生成开始和结束位置的概率.

3 方法

3.1 定义

假设存在未修剪的视频 $V=\{f_n\}_{n=1}^{l_v}$,其中 f_n 是第 n 个 RGB 帧,l_v 是帧数. 视频 V 的标签由一组行为实例 $U_g=\{u_g=(t_{s,n},t_{e,n})\}_{n=1}^{N_g}$ 组成,其中 N_g 是视频 V 中的真实的行为实例的数量,$t_{s,n}$ 是行为实例 u_g 的开始位置,$t_{e,n}$ 是行为实例 u_g 的结束位置. 时序行为边界框生成任务的目的是预测可以精确覆盖 U_g 的行为实例的时间边界 $U_p=\{u_p=(t_{s,n},t_{e,n})\}_{n=1}^{N_p}$,其中 N_p 是预测的行为实例 u_p 的数量.

3.2 框架概述

图 2 显示了网络框架的概述. 在视频编码部分,视频被送到骨干网络并被编码成形状为 $T/16\times H/32\times W/32\times C$ 的视频特征,其中 T 是视频 V 的帧数,H 和 W 分别是高度和宽度,C 是通道数. 时间全局相关特征生成部分由 n 个时间全局相关(TGC)模块和平均池化层组成. TGC 模块的功能是对时间全局信息进行编码,使得特征图中的每个位置都具有全局信息. 通过堆叠 n 个 TGC 模块,该模型能够以更复杂的方式对全局信息进行编码. 形状为 $T/16\times C$ 的一维特征(1D Feature)是通过平均池化生成的. 边界框生成部分类似于 BMN[5],由 3 个模块组成:基本模块(Module)、时序评估模块(Temporal Evaluation Module,TEM)和边界框评估模块(Proposal Evaluation Module,PEM). 基本模块处理 1D Feature,输出被 TEM 和 PEM 共享的特征;TEM 为视频的每个时间位置生成开始和结束的概率;PEM 为密集分布的边界框提供置信度分数. 后处理指的是利用开始和结束的概率和密集分布的边界框的置信度得分以及 Soft-NMS[17]生成边界框.

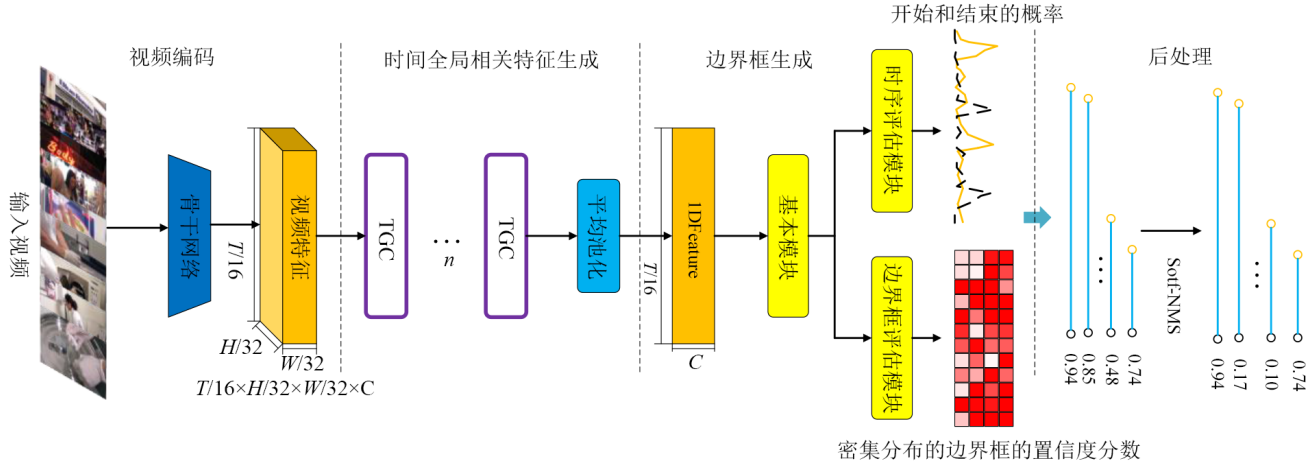

图 2 时间全局相关网络的框架

3.3 视频编码

本文使用 3D-ResNet[18]编码视频的时间和空间信息. 由于未剪辑的视频很长,本文将视频 $V=\{f_n\}_{n=1}^{l_v}$ 划分为一系列的片段 $S=\{s_m\}^{l_s}$,其中 l_v 表示视频的帧数,s_m 包含固定数量的帧 L,l_s 表示视频能被分成的片段数,分割间隔为 $\sigma=l_v/l_s$. 在 THUMOS14 数据集中,分割间隔 $\sigma=512$,固定数量的帧 $L=512$;在 ActivityNet1.3 数据集中,片段数 $l_s=1$,固定数量的帧 $L=1600$. 采用上述设置是为了保证输入的视频片段的时间长度大于视频中最长行为的时间长度.

3.4 时间全局相关特征生成

时间全局相关特征生成部分的功能是获取全局信息,它由 n 个 TGC 模块和一个平均池化层组成. TGC 模块通过注意力机制和卷积运算分别获得动态和静态的全局信息.

时间全局相关(TGC)模块. 全局信息可以有效地在语义分割任务中生成具有判别性的特征,这充分表明了全局信息的有效性. 受此启发,本文提出了一种可生成动态和静态全局信息的时间全局相关(TGC)模块. 如图 3 所示,TGC 模块主要由两部分组成:动态相关结构和静态相关结构.

动态相关结构. 场景分割方法[16]使用注意力机制生成全局上下文信息. 该方法会根据输入动态地在局部特征上构建丰富的全局上下文信息,从而生成动态全局信息. 本文将其引入到时序行为边界框生成中,提出了基于注意力机制的动态相关结构,以获得关于行为的时序动态全局信息.

如图 3 所示,给定局部特征 $\boldsymbol{A}\in\mathbb{R}^{C'\times T'\times H'\times W'}$,将其输入 3 个 $1\times1\times1$ 的卷积层,得到特征 $\{\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}\}$,其中 $\{\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}\}\in\mathbb{R}^{C'\times T'\times H'\times W'}$. 然后,将 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 变形为 $\mathbb{R}^{C'\times N}$,其中 $N=T'\times H'\times W'$. 同时,在 $\boldsymbol{B}$ 和 $\boldsymbol{C}$ 的转置之间执行矩阵乘法,并使用 softmax 层计算时间空间注意力矩阵

$\boldsymbol{X}\in\mathbb{R}^{N\times N}$,即

$$x_{(t,n),(t',n')}=\frac{\exp B_{(t',n')}\cdot C_{(t,n)}}{\sum_{n'=1}^{N_{n'}}\sum_{t'=1}^{N_{t'}}\exp B_{(t',n')}\cdot C_{(t,n)}} \tag{1}$$

其中 $x_{(t,n),(t',n')}$ 表示第 (t',n') 个位置对第 (t,n) 个位置的影响,t 和 t' 是特征图在时间维度上的位置,n 和 n' 是特征图在空间维度上的位置,$N_{n'}=W'\times H'$, $N_{t'}=T'$. 同时,将 $\boldsymbol{D}$ 变形为 $\mathbb{R}^{C'\times N}$. 然后在 $\boldsymbol{D}$ 和 $\boldsymbol{X}$ 之间执行矩阵乘法,并对结果进行变形以获得包含动态全局信息的特征图 $\boldsymbol{E}\in\mathbb{R}^{C'\times T'\times H'\times W'}$,即

$$E_{(t,n)}=\sum_{n'=1}^{N_{n'}}\sum_{t'=1}^{N_{t'}}x_{(t,n),(t',n')}\cdot D_{(t',n')} \tag{2}$$

根据式(2)可知,特征图 $\boldsymbol{E}$ 的每个时间和空间位置是所有时间和空间位置的加权和,因此特征图 $\boldsymbol{E}$ 具有全局信息. 此外,由于时间空间注意力矩阵 $\boldsymbol{X}$ 根据输入而变化,因此特征图 $\boldsymbol{E}$ 中的每个位置与所有位置之间的关系都随着输入而变化. 所以,特征图 $\boldsymbol{E}$ 具有动态的全局信息.

静态相关结构. 动态全局信息与行为类别有关,不同的行为可能具有不同的动态全局信息. 另外,所有的行为可能具有某种类似的、与行为类别无关的静态全局信息,这些信息在动态全局信息中无法得到体现,但这些信息对于区分行为与非行为是有价值的. 为此,本文提出了一种静态相关结构以获取可以表示该关系的静态全局信息.

如图 3 所示,本文将局部特征图 $\boldsymbol{A}$ 输入 $K\times 1\times 1$ 的卷积层中,以获得包含静态全局信息的特征图 $\boldsymbol{F}\in\mathbb{R}^{C'\times T'\times H'\times W'}$,即

$$F_{(t,n)}=\sum_{t'=1}^{N_{t'}}w_{(t,n),(t',n)}\cdot A_{(t',n)} \tag{3}$$

其中,t 和 t' 是时间位置;n 是空间位置;$w_{(t,n),(t',n')}$ 表示第 (t',n) 个位置对第 (t,n) 个位置的影响,$N_{t'}=T'$. 但是,如何获取静态全局信息? 实际上,本文将卷积核的参数 K 设置为大于或等于 T',此时卷积核的感受野是全局的,那么特征图 $\boldsymbol{F}$ 包含全局信息. 又因为卷积核参数在推理过程中是固定的,所以特征图 $\boldsymbol{F}$ 具有静态的全局信息.

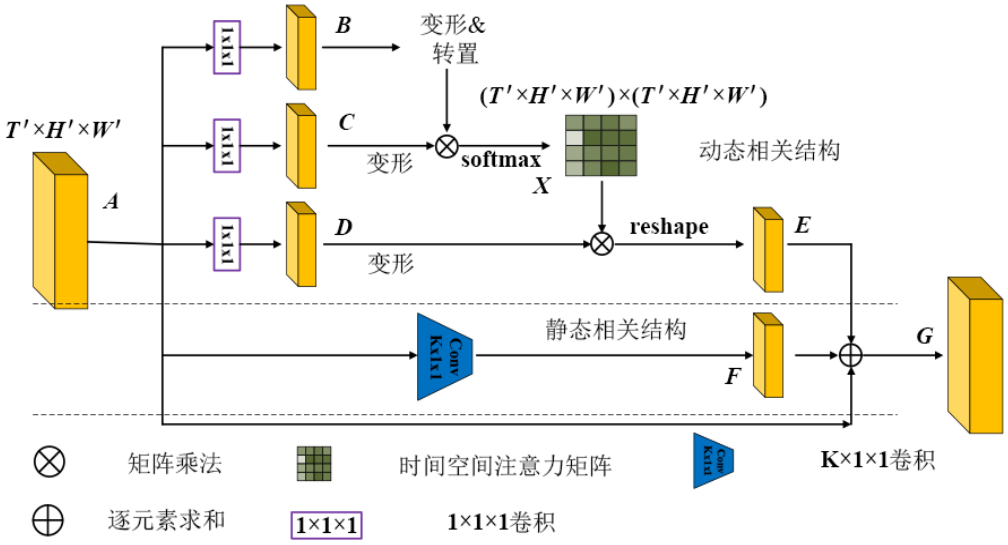

图 3　时间全局相关模块的详细信息

特征融合. 网络基于动态相关结构和静态相关结构获得特征图 $\boldsymbol{E}$ 和 $\boldsymbol{F}$,并且分别提取视频的动态全局信息和静态全局信息. 这些信息对于检测持续时间较短的行为可能不利,或者说检测持续时间较短的行为时,局部信息更有效. 因此,本文通过 $\boldsymbol{A}$,$\boldsymbol{E}$,$\boldsymbol{F}$ 的融合,将 $\boldsymbol{E}$ 和 $\boldsymbol{F}$ 包含的全局信息与 $\boldsymbol{A}$ 包含的局部信息相结合,实现持续时间较长和持续时间较短的行为都能有效地检测出来. 为了融合这 3 个特征图,本文将包含动态全局信息的特征图 $\boldsymbol{E}$ 乘以比例参数 α,将包含静态全局信息的特征图 $\boldsymbol{F}$ 乘以比例参数 β,并执行逐元素求和运算,以获得包含 2 种全局信息的特征图 $\boldsymbol{G}\in\mathbb{R}^{C'\times T'\times H'\times W'}$,即

$$G_{(t,n)}=\alpha\cdot\boldsymbol{E}_{(t,n)}+\beta\cdot\boldsymbol{F}_{(t,n)}+\boldsymbol{A}_{(t,n)} \tag{4}$$

其中,α 和 β 是可学习的参数,并被初始化为 0. 通过组合静态和动态全局信息,可以生成更具有判别能力的特征. 为了充分发挥时间全局相关模块的作用,本文级联了 n 个 TGC 模块.

3.5　边界框生成

如图 2 所示,通过时间全局相关(TGC)模块,得到 1D Feature $\in\mathbb{R}^{C'\times T'}$. 将 1D Feature 输入基本模块(Base Module),随后将输出的特征图输入时序评估模块(TEM)和边界框评估模块(PEM)以分别生成密集分布的边界框的开始和结束概率以及置信度分数. Base

Module，TEM 和 PEM 的结构和原理与 BMN[5] 相同．为了更好地了解整个网络，本文简要介绍这 3 个模块的结构．

基本模块．Base Module 处理 1D Feature，输出被 TEM 和 PEM 共享的时序特征图 $T_F \in \mathbb{R}^{C\times T'}$．如表 1 所示，基本模块由 Conv1D_1 和 Conv1D_2 两个卷积层组成．

时序评估模块．时序评估模块的目的是获取视频中所有时间位置的开始和结束概率，然后根据开始和结束概率生成边界框．如表 1 所示，时序评估模块主要由 Conv1D_3 和 Conv1D_4 两个卷积层组成，并且使用 sigmoid 函数激活输出特征以生成开始点的概率 $P_s=\{P_{tn}^s\}_{n=1}^{lv}$ 和结束点的概率 $P_e=\{P_{tn}^e\}_{n=1}^{lv}$，其中 lv 是时间位置的数量，P_{tn}^s 是第 n 个时间位置是行为开始点的概率，P_{tn}^e 是第 n 个时间位置是行为结束点的概率．

边界框评估模块．边界框评估模块生成边界匹配置信度图（Boundary-Matching confidence map），为密集分布的边界框提供置信度分数．如表 1 所示，边界框评估模块主要由边界匹配层（BM layer）和多个 3d，2d 卷积层组成．通过边界匹配层（BM layer），时序特征图 $T_F \in \mathbb{R}^{C\times T'}$ 能够转化成边界匹配特征图 $\text{BM}_F \in \mathbb{R}^{C\times N\times D\times T'}$，其中 D 是边界框的最大时间长度并且 $D=T'$．边界匹配层的目的是在每个边界框 $\rho_{(i,j)}$ 的开始点 t_s 和结束点 t_e 之间的特征中均匀采样 N 个点，以生成边界框的特征 $F_{(i,j)}^{\rho} \in \mathbb{R}^{C\times N}$，其中 $N=32$．然后，对所有的边界框执行相同的采样操作以获得边界匹配特征图 BM_F 以及通过 Conv3D_1 卷积将边界匹配特征图 BM_F 的采样维度从 N 变为 1，同时将通道数 C 从 256 变为 512．最终，通过 Conv2D_1，Conv2D_2，Conv2D_3 以及 sigmoid 函数得到边界匹配置信度图：$\text{BM}_C \in \mathbb{R}^{D\times T'}$ 和 $\text{BM}_R \in \mathbb{R}^{D\times T'}$，其中 BM_C 将用于二分类损失函数（binary classification loss function），而 BM_R 将用于回归损失函数（regression loss function）．

表 1 基本模块、时序评估模块、边界框评估模块的详细结构

模块	层名称	卷积核	激活函数	输出
基本模块	Conv1D_1	3	relu	$256\times T'$
	Conv1D_2	3	relu	$256\times T'$
时序评估模块	Conv1D_3	3	relu	$256\times T'$
	Conv1D_4	3	sigmoid	$2\times T'$
边界框评估模块	BM layer			$256\times 32\times D\times T'$
	Conv3D_1	32×1×1	relu	$512\times 1\times D\times T'$
	squeeze			$512\times D\times T'$
	Conv2D_1	1×1	relu	$128\times D\times T'$
	Conv2D_2	3×3	relu	$128\times D\times T'$
	Conv2D_3	1×1	sigmoid	$2\times D\times T'$

注：T' 是特征序列的时间长度，D 是边界框的最大时间长度．

3.6 后处理

后处理的功能是利用开始和结束的概率生成边界框，并使用密集分布的边界框的置信度分数为生成的边界框提供置信度分数．生成边界框后，网络利用 Soft-NMS[17] 降低置信度得分消除冗余的结果．后处理的参数与 BMN[5] 相同．

4 实验

4.1 评估数据集

为了评估时间全局相关网络（TGCNet）的有效性，本文在 THUMOS14 和 ActivityNet1.3 两个数据集上进行了实验．THUMOS14[19] 数据集包含 200 个验证和 213 个测试视频，包含 20 个动作类别．ActivityNet1.3[20] 是一个大型视频数据集，包含 19 994 个未剪辑的视频，其中包含 200 个动作类别．

4.2 实验细节

对于视频编码，本文选择在 Kinetics-400[21] 上预训练的 3D-ResNet[18] 作为骨干网络．在 ActivityNet1.3 中，为了方便与其他方法比较，包括 TCN[22]，MSRA[23]，Prop-SSAD[24]，CTAP[25]，BSN[4]，BMN[5]，DBG[6]，本文将所有视频的帧数归一化为 1 600．在 THUMOS14 中，本文不对视频帧数进行归一化，而是参考 BMN[5] 设置了固定长度的滑动窗口 Lw=512，该参数可以涵盖 98% 的行为实例的长度．在训练过程中，此研究使用 momentum SGD 进行优化，batch 设置为 4．在前 7 个 epoch 中将学习率设置为 5×10^{-4}，接下来的 3 个 epoch 中学习率衰减为 5×10^{-5}．与 BMN 相似，本文使用时间边界标签（temporal boundary label）和边界匹配标签（boundary-matching label）．对于损失函数，本文使用二分类（binary classification function）和回归损失函数（regression loss function）．

4.3 时序行为边界框生成

为了评估边界框的质量，本文使用在不同 IoU（Intersection over Union）阈值下的平均召回率和平均提议数量曲线（Average Recall（AR）with Average Number of proposals（AN））．其中，IoU 阈值 [0.5∶0.05∶0.95] 和 [0.5∶0.05∶1.0] 分别用于 ActivityNet1.3 和 THUMOS14．对于 ActivityNet1.3，本文计算 AR 与 AN 曲线下的面积（Area Under the AR vs. AN Curve，AUC）以验证结果，AN 的值为 0~100．

最新方法的比较．本文将 TGCNet 与在 ActivityNet1.3 验证集上测试的其他最新方法进行了比较．如表 2 所示，TGCNet 达到了最先进的性能，并将 AUC 从 68.23% 提高到 68.72%，表明本文的网络可以生成涵盖所有行为实例的高质量的边界框．

表 3 显示了 TGCNet 在 THUMOS14 数据集上的结果．和 BMN 一样，本文分别在使用 Greedy-NMS 和 Soft-

表2　本文的TGCNet与其他最新的边界框生成网络在AR@AN和AUC方面在ActivityNet1.3数据集的验证集上的比较

方法	TCN	MSRA	Prop-SSAD	CTAP	BSN	MGG	BMN	DBG	TGCNet
AR@100(val)	-	-	73.01	73.17	74.16	74.54	75.01	**76.65**	76.52
AUC(val)	59.58	63.12	64.40	65.72	66.17	66.43	67.10	68.23	**68.72**

NMS情况下获得结果. 当边界框的平均数量设置为[50,100,200,500,1 000]时,本文的结果要优于其他网络.

表3　本文的TGCNet在THUMOS14数据集上与其他最好的边界框生成网络之间的比较

方法	@50	@100	@200	@500	@1000
BSN+NMS	35.41	43.55	52.23	61.35	65.10
BSN+SNMS	37.46	46.06	53.21	60.64	64.52
BMN+NMS	37.15	46.75	54.84	62.19	65.22
BMN+SNMS	39.36	47.72	54.70	62.07	65.49
DBG+NMS	40.89	49.24	55.76	61.43	61.95
DBG+SNMS	37.32	46.67	54.50	62.21	66.40
TGCNet+NMS	40.72	48.22	56.06	**62.37**	**67.25**
TGCNet+SNMS	**41.26**	**50.33**	**56.21**	62.34	66.08

注:指标是AR@AN.其中NMS代表Greedy-NMS,SNMS代表Soft-NMS.

骨干网络的消融实验. 为了比较具有不同深度和结构的3D-ResNet[18]对结果的影响,本文使用不同的骨干网络进行实验. 表4中的结果表明,即便使用深度为18的3D-ResNext-18网络,也可以通过端到端的训练方式使得TGCNet获得良好的性能. 随着骨干网络深度的增加,网络的性能越来越好. 最终,当骨干网络为3D-ResNext-101时,获得最佳性能,并且AUC为67.33%. 性能最好的原因是该网络比其他网络更深,能够提取更好的用于时序行为边界框生成的特征.

表4　根据AR@AN和AUC在ActivityNet1.3数据集的验证集上的结果,研究TGCNet中骨干网络的影响

骨干网络	@1	@5	@10	@100	AUC
3D-ResNet-18	32.46	45.26	52.31	72.96	64.23
3D-Resnet-34	33.64	47.29	54.43	74.79	66.21
3D-ResNet-50	33.33	48.30	55.51	74.58	66.55
3D-ResNext-101	**34.21**	**49.84**	**56.80**	**75.33**	**67.33**

时间全局相关(TGC)模块的消融实验. 表5显示了TGC模块对结果的影响,结果都是在骨干网络为3D-ResNet-101时获得的. 添加一层TGC模块后,结果显著改善,AUC从67.33%提高到68.37%. 为了探究一层TGC模块是否可以充分地编码全局信息,本文将多个TGC模块进行级联. 当有5个TGC模块时,效果最好,此时AUC为68.72%. 当数量超过5时,由于级联过多的TGC模块而引入了过多的参数,使得网络过拟合,导致结果降低.

动态相关结构和静态相关结构的消融实验. 为了探讨动态相关和静态相关结构对结果的影响,本文分别进行了2个实验. 实验以3D-ResNext-101为骨干网络,并且只利用一个TGC模块. 如表6所示,仅添加静态相关结构时AUC为67.92%,仅添加动态相关结构时AUC为67.58%,这表明两种结构都可以分别提高网络的性能. 然后,本文将动态和静态相关结构同时添加到网络中,得到AUC为68.37%. 网络性能得到提升的原因是融合了存在互补关系的动态全局信息和静态全局信息.

表5　根据AR@AN和AUC在ActivityNet1.3数据集的验证集上的结果,研究TGCNet中TGC模块数量的影响

NoM	@1	@5	@10	@100	AUC
0	34.21	49.84	56.80	75.33	67.33
1	34.76	50.82	58.07	76.24	68.37
2	35.07	51.05	58.13	76.28	68.54
3	35.10	50.97	57.80	**76.59**	68.57
4	35.12	50.69	57.83	76.55	68.63
5	35.15	50.95	**58.28**	76.52	**68.72**
6	**35.26**	**51.08**	57.90	76.43	68.49

注:NoM:TGC模块的数量.

表6　根据AR@AN和AUC在ActivityNet1.3数据集的验证集上的结果,研究TGCNet中静态相关结构和动态相关结构的影响

SC	DC	@1	@5	@10	@100	AUC
×	×	34.21	49.84	56.80	75.33	67.33
√	×	34.67	50.00	57.43	75.79	67.92
×	√	34.76	49.92	56.66	75.78	67.58
√	√	**34.76**	**50.82**	**58.07**	**76.24**	**68.37**

注:SC:静态相关结构,DC:动态相关结构.符号√表示已添加相应的结构,符号×表示未添加任何结构.

端到端的训练方式的消融实验. 为了验证端到端训练方式的有效性,本文将端到端与现有的训练方式进行了比较. 实验以3D-ResNext-101为骨干网络,并且仅添加了一个TGC模块. 如表7所示,不添加TGC模块,端到端训练方式将AUC从64.04%提高到67.33%,原因是该训练方式能够提取更加适用于行为边界生成的特征图. 当仅仅添加TGC模块时,AUC为63.34%,出现这种情况的原因是TGC模块参数量大,没有端到端训练方式的配合导致过拟合. 添加TGC模块和使用端到端训练方式时,AUC为68.37%,结果提升明显的原因是端到端训练和TGC模块相互配合能够生成更有利于

行为边界框生成的特征. 为了进一步说明端到端训练方式的优势,本文还向BMN添加了TGC模块. 如表8所示,由于BMN没有使用端到端训练方式,添加TGC模块后性能降低,这进一步表明端到端训练方式更有效.

数据处理方式的消融实验. 在ActivityNet1.3中,由于视频较长,为了实现端到端的训练方式,本文将帧数标准化为1 600. 在R-C3D[24]中,为了实现端到端的训练,所有视频的帧率均固定为3. 如表9所示,本文比较了这两种数据处理方法,可以看出本文的数据处理方法更好. 因为固定的帧率会使短动作缩短,更加不利于动作定位. 该实验以3D-ResNext-101为骨干网络进行.

表7 根据AR@AN和AUC在ActivityNet1.3数据集的验证集上的结果,研究端到端训练方式和TGC模块的影响

E2E	TGC	@1	@5	@10	@100	AUC
×	×	31.99	43.91	50.94	73.48	64.04
√	×	34.21	49.84	56.80	75.33	67.33
×	√	32.66	43.34	49.87	72.90	63.34
√	√	**34.76**	**50.82**	**58.07**	**76.24**	**68.37**

注:E2E为端到端的训练方法;TGC为时间全局相关模块.E2E列中的符号√表示使用了端到端训练方式,而符号×则相反.TGC列中的符号√表示已使用TGC模块,而符号×则相反.

表8 根据AR@AN和AUC在ActivityNet1.3数据集的验证集上的结果,研究BMN中的TGC模块的影响

TGC	@1	@5	@10	@100	AUC
×	33.28	**48.78**	**56.40**	**75.22**	**67.10**
√	**33.55**	48.74	55.66	74.43	66.24

注:TGC为时间全局相关模块. TGC列中的符号√表示已使用TGC模块,而符号×则相反.

表9 根据AR@AN和AUC在ActivityNet1.3数据集的验证集上的结果,研究TGCNet中数据处理的影响

数据处理	@1	@5	@10	@100	AUC
FR	0.00	3.04	11.01	56.75	52.80
NF	**34.21**	**49.84**	**56.80**	**75.33**	**67.33**

注:FR:帧率为3. NF帧数为1 600.

边界框的通用性. 一个好的边界框生成网络的重要特性是能够为未曾见过的动作生成高质量的边界框. 和BMN一样,本文使用ActivityNet1.3的两个动作子集,即“运动,锻炼和娱乐”和“社交,放松和休闲”,分别作为可见子集(seen)和不可见子集(unseen). 可见子集和不可见子集分别有87和38个行为类别,4 455和1 903个训练视频,2 198和896个验证视频. 在实验中,本文选择3D-ResNext-101作为骨干网络并级联5个TGC模块. 本文分别用可见(seen)和可见与不可见(seen+unseen)训练数据视频训练网络,并分别在可见(seen)和不可见(unseen)验证数据视频中评估网络. 本实验的目的是说明是否在训练数据中添加不可见子集的数据,其对网络在验证集上的结果影响很小. 如表10所示,当在训练数据的可见子集的视频上训练网络并在验证数据的不可见子集的视频上测试网络时,AUC只会略有下降,这表明本文的网络可以为未曾见过的动作生成高质量的边界框.

表10 根据AR@AN和AUC在ActivityNet1.3数据集的验证集上的结果,验证TGCNet的通用性

训练数据	验证数据			
	可见		不可见	
	AR@100	AUC	AR@100	AUC
可见+不可见	75.43	67.26	75.11	67.07
可见	75.08	67.20	**72.78**	**65.02**

推理速度. 为了探究动态相关结构和静态相关结构对TGCNet推理速度的影响,本文进行了消融实验. 实验以3D-ResNext-101为骨干网络,并且只利用一个TGC模块. 如表11所示,不添加动态相关结构和静态相关结构时,网络处理一个视频平均花费0.815 8 s,每秒钟可以处理1 961帧. 在分别添加动态相关结构和静态相关结构后,网络处理一个视频平均花费0.817 6 s和0.817 4 s,每秒钟可以处理1 956帧和1 957帧. 将动态和静态相关结构同时添加到网络中时,网络处理一个视频平均花费0.822 7 s,每秒钟可以处理1 944帧. 虽然添加一个TGC模块会将网络处理一个视频的平均时间从0.815 8 s增加到0.822 7 s,但增加的时间非常少,这表明TGC模块对网络推理速度的影响很微小.

表11 根据AR@AN,AUC和推理速度(inference speed)在ActivityNet1.3数据集的验证集上的结果,研究TGCNet中静态相关结构和动态相关结构的影响

SC	DC	AR@100	AUC	T_{inf}	FPS
×	×	75.33	67.33	0.815 8	1 961
√	×	75.79	67.92	0.817 4	1 957
×	√	75.78	67.58	0.817 6	1 956
√	√	**76.24**	**68.37**	**0.822 7**	**1 944**

注:SC为静态相关结构,DC为动态相关结构. 符号√表示已添加相应的结构,符号×表示未添加任何结构. 推理速度T_{inf}(s/视频)指的是处理完验证集中所有视频的平均时间; FPS指的是网络每秒钟可以处理的帧数.

本文将TGCNet的推理速度与同样使用端到端训练方式的R-C3D[26]进行了比较. TGCNet的推理速度是在以3D-ResNext-101为骨干网络并添加5个TGC模块的情况下得到的. 如表12所示,TGCNet的FPS为1 824,R-C3D的FPS为1 030,这表明TGCNet能够以更高的速度处理视频.

检测误差. 为了探究TGCNet的检测误差,本文在ActivityNet1.3数据集的验证集上进行了实验. 实验以

3D-ResNext-101为骨干网络,并且仅添加了一个TGC模块. ActivityNet1.3数据集的验证集有7 654个行为实例,本文用在不同IoU(Intersection over Union)阈值和不同数量的边界框的情况下将网络所能检测出的行为实例占验证集中行为实例的比例作为结果. 表13中表示的是在分别选取得分在前1,5,10,100的边界框的情况下,并在IoU阈值为0.5,0.75,0.95的条件下网络检测出的行为实例占验证集中行为实例的比例(%). 如表13所示,在选取得分为前1,5,10,100的边界框的情况下和在IoU阈值为0.5,0.75,0.95的条件下,添加了TGC模块后的网络都能检测出更多的行为实例. 同时,在选取得分在前100的边界框和IoU为0.5时,网络能够检测出91.45%的行为实例,说明在该条件下,网络能够检测出绝大部分的行为实例,同时也说明TGCNet的检测误差很小.

表12　TGCNet与其它网络在推理速度上的比较

方法	T_{inf}	FPS
R-C3D	-	1 030
TGCNet	0.877 0	1 824

注:推理速度T_{inf}(秒/视频)指的是处理完验证集中所有视频的平均时间; FPS指的是网络每秒钟可以处理的帧数.

表13　根据检测出的行为实例占ActivityNet1.3数据集的验证集中行为实例的比例(%),研究TGCNet的检测误差

IoU	TGC	@1	@5	@10	@100
0.50	×	46.90	63.78	71.71	91.12
	√	47.31	64.09	71.97	91.45
0.75	×	35.27	52.39	60.15	82.30
	√	35.57	52.70	60.46	82.59
0.95	×	15.46	22.28	25.21	34.12
	√	15.78	22.60	25.52	34.42

注:IoU为网络生成的边界框与数据集中行为实例的交并比;TGC为时间全局相关模块.TGC列中的符号√表示已使用TGC模块,而符号×则相反.

特征图的可视化. 为了进一步了解时间全局相关(TGC)模块,本文将时间空间注意力图$\boldsymbol{X}\in\mathbb{R}^{N\times N}$,包含动态全局信息的特征图$\boldsymbol{E}\in\mathbb{R}^{C'\times T'\times H'\times W'}$和包含静态全局信息的特征图$\boldsymbol{F}\in\mathbb{R}^{C'\times T'\times H'\times W'}$可视化. 对于注意力图$\boldsymbol{X}$,本文直接对其进行可视化;对于特征图$\boldsymbol{E}$,本文将其变形为特征图$\boldsymbol{E}'\in\mathbb{R}^{N\times 1}$,然后对特征图$\boldsymbol{E}'$进行转置得到特征图$\boldsymbol{E}''\in\mathbb{R}^{1\times N}$,随后特征图$\boldsymbol{E}'$和特征图$\boldsymbol{E}''$进行矩阵乘法生成特征图$\in\mathbb{R}^{N\times N}$进行可视化. 本文在特征图$\boldsymbol{F}$上执行相同的操作.

如图4所示,本文在图中用白框绘制了真实标签(ground truth). 其中,(3)和(4)视频中包含的是同一种行为,特征图$\boldsymbol{X}$和$\boldsymbol{E}$的可视化表明白框所包围的特征在视频(3)和(4)是很相似的; 而视频(1)(2)和(3)中包含的是不同的行为,所以特征图$\boldsymbol{X}$和$\boldsymbol{E}$的可视化结果在视频(1)(2)和(3)差距很大. 这表明,特征图$\boldsymbol{E}$包含的动态全局信息与行为类别有关,不同的行为可能具有不同的动态全局信息. 对于特征图$\boldsymbol{F}$,需要从对角线进行观察,可以明显看到白框的角很接近特征图的高亮区域. 并且包含不同行为的视频(1)(2)(3)和(4)的特征图$\boldsymbol{F}$的可视化结果很相似,这表明特征图$\boldsymbol{F}$可以提取不同行为所具有的某种类似的、与行为类别无关的静态全局信息. 时间空间注意力图$\boldsymbol{X}$,包含动态全局信息的特征图$\boldsymbol{E}$和包含静态全局信息的特征图$\boldsymbol{F}$的可视化表明本文的动态相关结构和静态相关结构可以很好地提取动态全局信息和静态全局信息以提升网络对行为的定位能力.

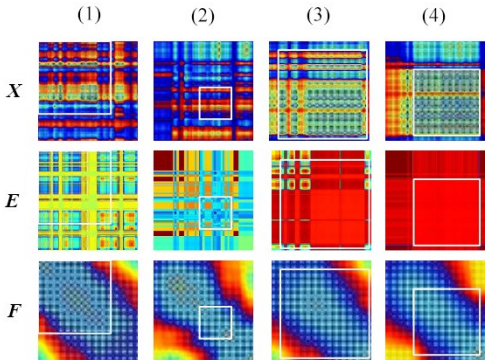

图4　注意力矩阵$\boldsymbol{X}$及特征图$\boldsymbol{E}$,$\boldsymbol{F}$的可视化

4.4　时序行为检测结果

评估边界框质量的另一种方法是将边界框输入到时序行为检测框架中并计算检测结果. 时序行为检测任务的评估指标是平均精度(mean Average Precision, mAP). 本文在不同IoU阈值[0.3,0.4,0.5,0.6,0.7]下计算在THUMOS14数据集中的mAP.

为了获得检测结果,本文遵循两阶段的"按边界框进行检测"框架(two-stage "detection by classifying proposals" framework),将TGCNet的边界框与最新的分类网络结合起来. 和BMN一样,本文针对每个边界框使用UntrimmedNet[27]在THUMOS14上生成的前2个视频级别分类结果. 表14显示了本文的TGCNet和其他方法的时序行为检测结果. 特别是当IoU为0.7时,TGCNet的检测性能比DBG高1.1%,这也表明我们的网络可以生成高质量的时序行为边界框.

表14　以mAP@IoU形式表示的在THUMOS14的测试集上的时序行为检测结果

方法	分类网络	0.7	0.6	0.5	0.4	0.3
BSN	UNet	20.0	28.4	36.9	45.0	53.5
BMN	UNet	20.5	29.7	38.8	47.4	56.0
DBG	UNet	21.7	30.2	39.8	49.4	57.8
TGCNet	UNet	**22.8**	**30.4**	**40.2**	**50.9**	**58.4**

5 结论

本文提出了一个由动态相关和静态相关结构组成的时间全局相关网络(TGCNet)以捕获动态和静态全局信息,从而为边界框提供精确的边界和可靠的置信度. 特别地,本文通过端到端的训练方式来训练TGCNet,能够生成更适合时序行为边界框生成任务的特征. 在两个具有挑战性的数据集THUMOS14和ActivityNet1.3上进行的实验证明了TGCNet优于其他最新方法.

参考文献

[1] BUCH S, ESCORCIA V, SHEN C, et al. Sst: Single-stream temporal action proposals[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2911-2920.

[2] GAO J, YANG Z, CHEN K, et al. Turn tap: Temporal unit regression network for temporal action proposals[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3628-3636.

[3] SHOU Z, WANG D, CHANG S F. Temporal action localization in untrimmed videos via multi-stage cnns[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 1049-1058.

[4] 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. Munich: Springer, 2018: 3-19.

[5] LIN T, LIU X, LI X, et al. Bmn: boundary-matching network for temporal action proposal generation[C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 3889-3898.

[6] LIN C, LI J, WANG Y, et al. Fast learning of temporal action proposal via dense boundary generator[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 11499-11506.

[7] WANG H, SCHMID C. Action recognition with improved trajectories[C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 3551-3558.

[8] WANG L, XIONG Y, WANG Z. Temporal segment networks: towards good practices for deep action recognition [C]//Proceedings of the European Conference on Computer Vision. Amsterdam: Springer, 2016: 20-36.

[9] 丰艳, 张甜甜, 王传旭. 基于伪3D残差网络与交互关系建模的群组行为识别方法[J]. 电子学报, 2020, 48(7): 1269-1275.
FENG Y, ZHANG T T, WANG C X. Group activity recognition method based on pseudo 3d residual network and interaction modeling[J]. Acta Electronica Sinica, 2020, 48 (7): 1269-1275. (in Chinese)

[10] 胡正平, 刁鹏成, 张瑞雪, 等. 3D多支路聚合轻量网络视频行为识别算法研究[J]. 电子学报, 2020, 48(7): 1261-1268.
HU Z P, DIAO P C, ZHANG R X, et al. Research on 3d muli-branch aggregated lightweight network video action recognition algorithm[J]. Acta Electronica Sinica, 2020, 48(7): 1261-1268. (in Chinese)

[11] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 4489-4497.

[12] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5534-5542

[13] PENG C, ZHANG X, YU G, et al. Large kernel matters--improve semantic segmentation by global convolutional network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4353-4361.

[14] YU C, WANG J, PENG C, et al. Learning a discriminative feature network for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1857-1866.

[15] ZHANG H, DANA K, SHI J, et al. Context encoding for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7151-7160.

[16] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3146-3154.

[17] BODLA N, SINGH B, CHELLAPPA R. Soft-nms improving object detection with one line of code[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5561-5569.

[18] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE,

2018: 6546-6555.
[19] GORBAN A, IDREES H, JIANG Y G, et al. THUMOS Challenge: Action Recognition with a Large Number of Classes[EB/OL]. (2015)[2020]. http://www.thumos.info.
[20] CABA HEILBRON F, 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. Boston: IEEE, 2015: 961-970.
[21] KAY W, CARREIRA J, SIMONYAN K, et al. The Kinetics Human Action Video Dataset[EB/OL]. (2017-05-19)[2020-11-18]. https://arxiv.org/abs/1705.06950.
[22] DAI X, SINGH B, ZHANG G, et al. Temporal context network for activity localization in videos[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5793-5802.
[23] GHANEM B, NIEBLES J C, SNOEK C, et al. Activitynet Challenge 2017 Summary[EB/OL]. (2017-10-22)[2020-11-18]. https://arxiv.org/abs/1710.08011.
[24] LIN T W, ZHAO X, SHOU Z. Temporal Convolution Based Action Proposal: Submission to Activitynet 2017[EB/OL]. (2017-06-21)[2020-11-18]. https://arxiv.org/abs/1707.06750.
[25] GAO J, CHEN K, NEVATIA R. Ctap: Complementary temporal action proposal generation[C]//Proceedings of the European Conference on Computer Vision. Munich: IEEE, 2018: 68-83.
[26] XU H, DAS A, SAENKO K. R-c3d: region convolutional 3d network for temporal activity detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5783-5792.
[27] WANG L, XIONG Y, LIN D, et al. Untrimmednets for weakly supervised action recognition and detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4325-4334.

作者简介

马百腾　男,1995年生. 华中科技大学人工智能与自动化学院硕士研究生. 主要研究方向为视频处理、行为检测、计算机视觉与模式识别.
E-mail: btm@hust.edu.cn

桑　农(通讯作者)　男,华中科技大学人工智能与自动化学院教授. 主要研究方向为计算机视觉、模式识别.
E-mail: nsang@hust.edu.cn

张士伟　男,阿里巴巴达摩院(杭州)科技有限公司高级算法工程师. 主要研究方向为行为检测、计算机视觉与模式识别.
E-mail: zhangjin.zsw@alibaba-inc.com

高常鑫　男,华中科技大学人工智能与自动化学院副教授. 主要研究方向为计算机视觉、模式识别和智能视频分析.
E-mail: cgao@hust.edu.cn