Video Grounding：视频时序定位

王 鑫 朱文武
清华大学

关键词：视频理解 多模态 视觉与语言 跨模态 视频检索 时序定位

引言

随着手机等终端设备上的多媒体技术不断发展，人们可以更容易地访问来自世界各地的视频。与文本、图像等其他信息传播和交换媒介相比，视频包含了更多的动态信息。一般来说，一个视频由伴有音频和字幕的连续帧图像序列组成，因此，视频在多媒体智能的探索和研究中具有天然的优势。比如，快速从一段长视频中检索特定视频片段可以让用户方便地找到感兴趣的高亮时刻，目前已有众多研究试图自动捕捉视频中的关键信息，例如，视频总结[42, 74, 81]、视频亮点检测[28, 70]。更为基础地，一些工作能够检测包含特定动作的视频片段，一般称为动作检测（action detection）或视频中的时序动作定位（Temporal Action Grounding in Videos，TAGV）。然而，TAGV受限于预先定义的动作类别集合，不能完全覆盖所有的活动。因此，引入自然语言描述复杂多样的活动更为合理，视频中的时序定位（Temporal Sentence Grounding in Videos，TSGV）就是这样一项任务：使一个句子查询与视频中具有相同语义的一个片段（也被称作时刻）相匹配。如图1所示，给定查询“一个小女孩从一个小男孩身边走过并继续吹树叶”作为输入，TSGV的目标是预测目标片段在原视频中的起点和终点（即从第7.11 s到第12.7 s）。作为典型的视频－文本多模态问题，TSGV任务往往需要首先采用预训练模型（如C3D和GloVe）提取视频/文本特征，再设计模态间/模态内的细粒度交互模块来学习跨模态的语义匹配关系。TSGV可以作为各种下游视觉－语言任务的中间任务，例如视频问答和视频内容检索，因此，TSGV值得深入探索，它连接了计算机视觉和自然语言处理社区，进一步促进了各种下游应用。然而，由于以下原因，TSGV更具挑战性：

- 视频和句子查询都是具有丰富的语义和时序性的。因此，视频和句子之间的匹配关系相当复杂，需要以更精细的方式建模，以实现准确的时间定位。
- 与查询相对应的目标片段在空间和时间尺度上是相当灵活的。如果通过滑动窗口获取候选视频片

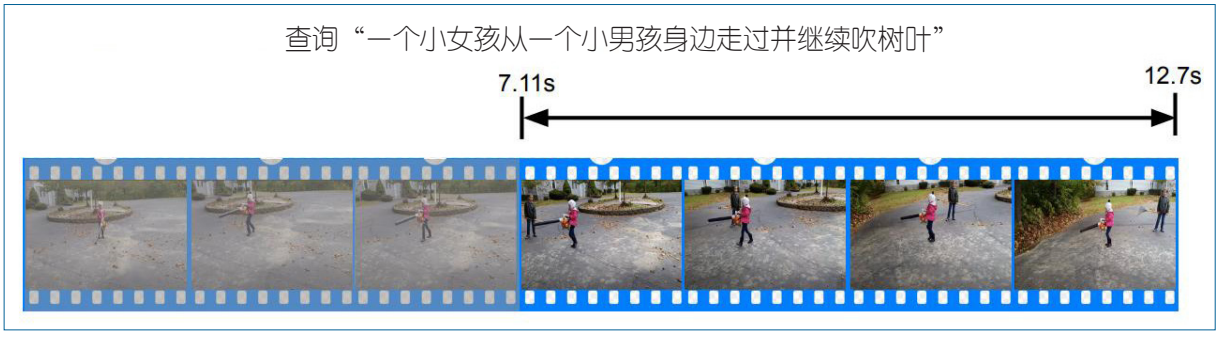


图1 TSGV的一个具体例子

段，计算成本会很高。因此，如何有效地全面覆盖目标片段，也是TSGV面临的挑战。

• 视频中的活动通常不是独立出现的，它们往往有内部的语义关联和时间上的相互依赖。因此，对视频上下文信息在句子语义引导下的内在逻辑关系进行建模也十分重要。

尽管存在上述挑战，近几年来，仍然出现了许多有前景的研究工作，从早期的基于两阶段匹配的方法[16, 18, 23, 37, 63]、端到端方法[8, 73, 75, 78]、基于强化学习的方法[21, 22, 65]，到最近的弱监督方法[14, 43]。本文总结了现有方法的分类、评价准则以及当前基准设计的潜在问题，并进一步明确了有前景的研究方向。

方法总述

如图2所示，根据是否生成候选片段和监督方式的不同，可以将TSGV模型分为四大类。早期的工作采用了两阶段的架构，即首先扫描整个视频，并通过滑动窗口或提案生成网络（proposal generation network）预先生成候选片段，然后根据跨模态匹配模块对候选片段进行排名。然而，候选片段的重叠导致了太多的冗余计算，而且单独的成对的片段查询匹配也可能忽略了上下文的视频信息。

考虑到上述问题，一些研究人员开始尝试以端到端方式解决TSGV问题。这种端到端模型没有预先切割出候选片段作为模型的输入。有的方法采用长短期记忆（LSTM）或卷积神经网络（CNN）依次维护在每一时间步结束的多尺度候选片段，它们被称为基于锚点（anchor-based）的方法。其他一些端到端方法预测每个视频单元（即帧级或片段级）是目标片段起点和终点的概率，或者根据整个视频和句子查询的多模态特征直接回归目标起点和终点坐标。这些方法不依赖任何生成候选片段的过程，被称为无锚点（anchor-free）的方法。

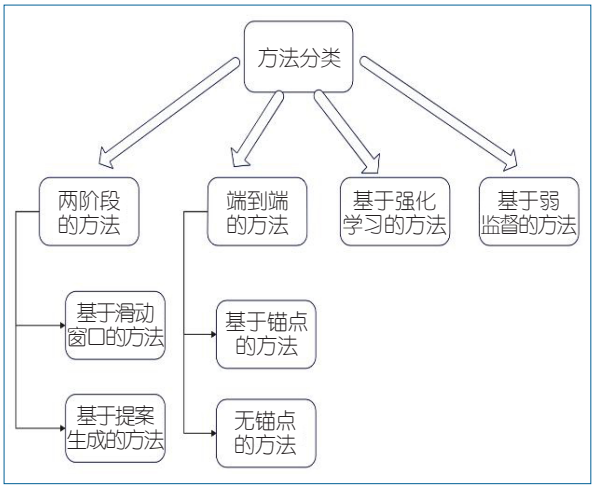


图2 TSGV方法分类

值得注意的是，有些工作借助深度强化学习技术解决TSGV问题，将这个任务视为一个顺序决策过程，这也是无锚点的。除了上述三类全监督方法，为了减少标注真实标签的时刻边界所需的大量人力，也有人提出了只用视频级标注的弱监督方法。

两阶段方法

对于两阶段方法，预分割候选片段与模型计算是分开进行的。两阶段方法可以分为两大类：基于滑动窗口的方法和基于提案生成的方法。

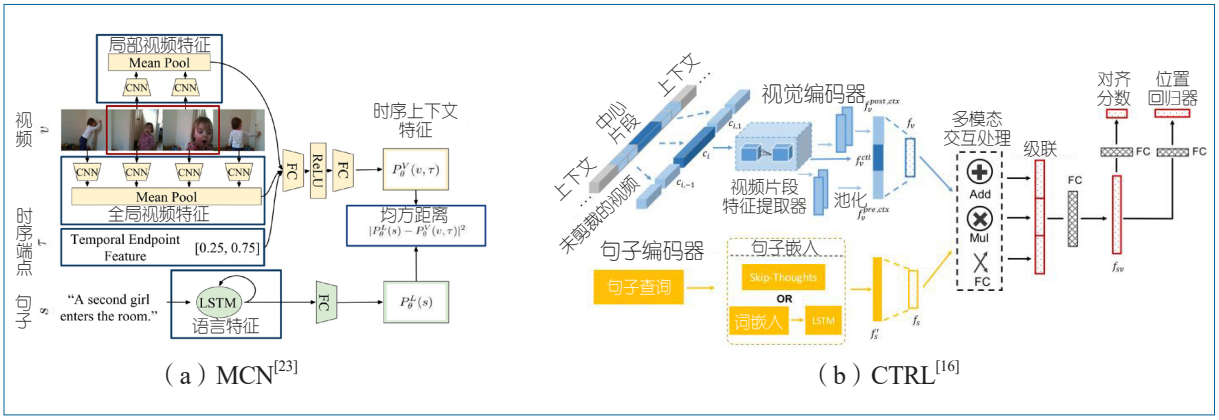
基于滑动窗口的方法

MCN[23]和CTRL[16]是开创性的工作，它们定义了TSGV任务并构建了基准数据集。Hendricks等人[23]提出MCN，它通过滑动窗口机制采样得到候选片段，然后将视频片段表示和查询表示嵌入到同一个向量空间。在这个空间中，句子查询和相应的目标视频片段之间的L2距离被最小化，以监督模型的训练（参见图3（a））。

Gao等人[16]提出了CTRL，这是第一个将R-CNN[20]从物体检测适应到TSGV的方法。CTRL利用滑动窗口获得不同长度的候选片段。如图3（b）所示，它利用多模态处理模块将候选片段的表征与句子表征相融合，然后将融合后的表征送入另一个全连接层，以预测候选片段的对齐分数以及候选段和目标段之间的位置偏移。

与上述将查询视作一个整体的CTRL相比，Liu等人[38]对查询进行分解，并根据视频的时间背景自适应地获取重要的文本成分，做了进一步的改进。

由于CTRL忽略了视频片段和句子查询里的时



（a）MCN[23]　　（b）CTRL[16]

图3　MCN和CTRL是首次提出TSGV任务的两项先驱性工作

空信息，Liu 等人 [37] 进一步提出了一个注意力跨模态检索网络 ACRN。通过由句子查询引导的记忆注意力网络自适应地将权重分配给用于记忆的上下文片段表征。SLTA[27] 也设计了一个注意力模型，根据查询信息自适应地识别相关的对象和互动。

Wu 和 Han[63] 提出了一种多模态环形融合模型（Multi-modal Circulant Fusion，MCF），将视觉 / 文本向量扩展到环状矩阵，可以充分地利用视觉和文本表征的相互作用。

然而，CTRL、ACRN 和 MCF 等都是直接计算视觉 – 语义的相关性，没有明确地对模态内活动信息进行建模，而且通过滑动窗口采样得到的候选片段可能包含各种无意义的噪音。因此，Ge 等人 [18] 明确地将视频和文本挖掘活动信息作为先验知识，为每个候选片段计算包含活动的信心程度，从而提高定位准确性。

基于提案生成的方法

考虑到基于滑动窗口方法的缺点，一些研究致力于减少候选片段的数量，被称为提案生成法。这种方法仍然采用两阶段方案，但通过不同种类的提案网络来避免密集的滑动窗口采样。QSPN[67] 通过将句子查询作为条件限制候选片段的选取，以此减少候选片段的数量，从而减轻了计算负载。类似地，Chen 和 Jiang[9] 提出的 SAP 将句子查询的语义信息整合到提案生成过程中。

尽管两阶段方法取得了一定的成功，但也有一些缺点。为了达到较高的定位精度（即候选片段中至少应该有一个接近真实标注），候选片段的长度和位置分布应该是多样化的，从而不可避免地增加了候选片段的数量，导致后续匹配过程的计算效率低下。

端到端方法

端到端模型遵循单流模式，可以被分为两类：基于锚点的模型和无锚点的模型。

基于锚点的模型

TGN[5] 是一个典型的端到端深度神经网络结构，它可以单程内定位目标时刻，而不用处理大量重叠的预分割候选片段。TGN 通过细粒度逐词帧交互动态匹配句子和视频单元。在每个时间步，定位器会同时对结束于该时间步的一组不同时长的候选片段进行评分。

CMIN[84] 与 TGN 一样进行序列化评分，并通过边界回归细化候选时刻。为了进一步加强跨模态匹配，它设计了一个新的跨模态交互网络。同样，CBP[60] 建立了一个单流模型，它在每个时间步上联合预测时间锚点及边界，以获得精确的定位。CSMGAN[36] 也采用了单流方案。它建立了一个联合图，通过迭代的消息传递对跨 / 自模态的关系建模，以有效捕捉两种模态之间的高阶相互作用。

Qu 等人 [46] 提出了一个细粒度的迭代注意力网络（FIAN），不同于上述使用 RNN 的方法，它设计了一个以内容为导向的策略生成候选片段。

TGN 通过 LSTM 网络建立时序定位结构，而 Yuan 等人 [73] 提出了 SCDM，利用分层的时间卷

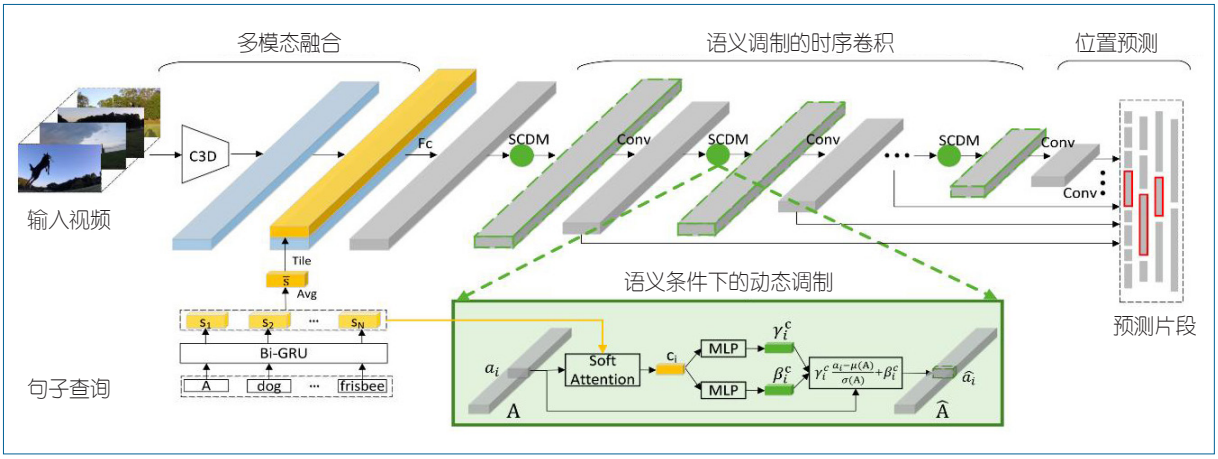


图4　SCDM的结构，将时间卷积网络与语义条件下的动态调制相结合[73]

积网络进行目标片段定位，如图 4 所示，这个多模态融合模块以细粒度的方式融合整个句子和每个视频片段。将融合的表示作为输入，语义调制的时间卷积模块在时间卷积过程中进一步关联与句子相关的视频内容，动态调制与句子相关的时间特征图。

MAN[78] 也采用时序卷积网络，其中句子查询作为动态过滤器被整合到卷积过程中。SCDM 和 MAN 都只考虑一维时间特征，而 2D-TAN[83] 通过一个二维时间特征图对视频片段的时间关系进行建模。2D-TAN 首先将句子表征与每个候选片段的特征融合，然后利用卷积神经网络嵌入视频上下文信息，最后预测每个候选片段作为目标片段的置信度得分。

Wang 等人 [59] 提出了一个结构化多级互动网络（SMIN），该网络在二维时间特征图上做了进一步的修改，作为其提案生成模块。Zhang 等人 [82] 也采用了与 2D-TAN 相同的提案生成方法，设计了一个视觉语言转换器主干，然后接上一个多阶段的整合模块，以获得有区别性的视频片段特征。

尽管基于锚点的方法取得了卓越的性能，但其性能对人工设计的启发式规则（即锚点的数量和尺度）很敏感。因此，这种基于锚点的方法不适用于视频长度可变的情况。同时，虽然不需要像两阶段方法那样进行预分割，但它的结果仍取决于被提案出的候选片段的排名，这也会影响其效率。

无锚点的模型

无锚点的方法没有对大量的候选方案进行排名，而是着眼于更精细的视频单元，如帧或片段，旨在预测每一帧 / 片段是目标片段的起点和终点的概率，或者直接从全局角度回归起点和终点。

Yuan 等人提出了 ABLR[75]。为了保留上下文信息，ABLR 首先通过双向 LSTM 网络对视频和句子进行编码。然后，引入多模态协同注意力机制，既生成能反映全局视频结构的视频注意力，还生成能突出时间定位关键细节的句子注意力。最后，设计了一个基于注意力的坐标预测模块，对时刻坐标进行回归。LGI[44] 也采用了和 ABLR 一样的基于注意力的位置回归。它进一步提出了一个更有效的局部 – 全局、视频 – 文本交互模块。VSLNet[80] 采用了一个标准问答框架，进一步区分了视频序列和文本段落之间的差异，以便更好地适应 TSGV 任务。L-Net[6] 引入了一个边界模型预测开始和结束的边界，以及包含交叉门控的循环神经网络强调相关的视频部分。Rodriguez 等人 [48] 也对每个视频单元预测了其是目标片段的起点或终点的概率，还进一步对边界标签的不确定性进行建模。

Lu 等人 [39] 提出了一个密集的自下而上的定位框架 DEBUG。一个典型的密集无锚点模型通常包含一个用于多模态特征编码的骨干框架和一个用于帧级预测的头部网络。DEBUG 设计了三个分支作为头部网络，分别预测每一帧的分类分数、边界距

离和置信分数。

类似地，DRN[76] 和 GDP[7] 也采用了这样一个密集无锚点的框架。DRN 的骨干网络使用一个视频和查询交互模块来获得融合的分层特征图。DRN 的头部网络密集地预测边界的距离、匹配分数和在真实标签片段内每一帧的近似 IoU（Intersection over Union，交并比，目标检测任务中衡量预测结果位置信息准确程度的指标）。同时，GDP 的骨干网络利用 Graph-FPN 层增强整合的帧特征。GDP 的头部网络预测每一帧到目标片段的边界的距离和置信分数。

与基于锚点的方法相比，无锚点的方法具有更高的计算效率和对可变时长视频的鲁棒性。虽然无锚点的方法具有这些显著的优势，但它很难捕捉到多模态交互的片段级特征。

上述端到端的方法要么利用多尺度锚点取样，要么直接回归最终坐标。也有一些方法跳出了这个模式。BPNet[66] 通过无锚点方法生成提案，然后以基于锚点的方式将它们与句子查询进行匹配。Wang 等人 [58] 提出了一个包含两个分支的双路径交互网络（DPIN）补充定位目标时刻。受自然语言处理领域中依存树分析任务的启发，有研究者还提出了一个基于双仿射的架构 CBLN [35]。

基于强化学习的方法

作为另一种无锚点方法，基于强化学习的框架将这样的任务视为一个连续的决策过程。每一步的行动空间是一组人为设计的基本操作（如移位、缩放）。

He 等人 [22] 首先引入深度强化学习技术解决 TSGV 任务，将 TSGV 形式化为一个顺序决策问题，在每个时间步骤中，观察网络输出环境的当前状态，供演员 – 评论员（actor-critic）模块生成行动策略，在此基础上，智能体执行行动来调整时间边界。

Wang 等人 [62] 提出了一种模型，依次观察一组有选择的视频帧，最后获得给定查询的时间边界。Cao 等人 [2] 首先利用空间场景跟踪任务，利用空间级强化学习过滤掉与文本查询不相关的信息。TripNet[21] 使用门控注意力调整文本和视觉特征从而提高准确性。

TSP-PRL[65] 采用了一种树状结构的策略，与传统的基于强化学习的方法不同，其灵感来自人类从粗略到精细的决策模式。同时，AVMR[3] 在对抗性学习范式下解决 TSGV 问题，它设计了一个基于强化学习的提案生成器生成提案候选。

弱监督方法

之后，TSGV 被扩展到训练阶段无法获得基准事实片段位置的弱监督场景下，即弱监督 TSGV。弱监督方法大致可分为基于多实例学习（Multi-Instance Learning，MIL）和基于重建两类。

一些工作 [12, 17, 43, 55] 采用多实例学习，整个视频被视为具有袋级标注的实例袋，对实例（视频段提案）的预测被聚合为袋级预测。

TGA[43] 是一种典型的基于 MIL 的方法，它通过将视频和其对应描述的匹配分数最大化，同时将视频和其他描述的匹配分数最小化来学习视频层面的视觉 – 文本对齐。它提出了文本引导的注意力（Text-Guided Attention，TGA）来获得特定文本的全局视频表征、学习视频和视频级描述的联合表征。

WSLLN[17] 是另一个基于 MIL 的端到端弱监督语言定位网络，同时进行片段语句对齐和片段选择。Huang 等人 [26] 提出了一种跨语句关系挖掘（Cross-sentence Relations Mining，CRM）方法，探索段落级范围内的跨语句关系，以提高单句定位精度。Ma 等人 [40] 提出了一个视频语言对齐网络，利用代理提案模块修剪不相关的时刻候选。Wu 等人 [64] 试图将基于强化学习的模型应用于弱监督 TSGV，提出了一个边界自适应细化框架，以实现边界灵活和内容感知的定位结果。Chen 等人 [12] 提出了一种由粗到精的模型，通过对帧进行分组来细化粗略段的边界。Tan 等人 [55] 提出了潜在图协同注意网络，在整个视频中进行细粒度的语义推理。

基于 MIL 的方法通常以基于正负样本对的三元组损失来学习视觉 – 文本对齐，因此严重依赖随机选择的负样本的质量，这些负样本通常很容易与正样本区分开，因而不能提供很强的监督信号。

基于重建的方法 [11, 14, 34, 54] 根据选定的视频片段重建给定的语句查询，并使用中间结果进行语句定位。Lin 等人 [34] 提出了一个语义补全网络（Semantic Completion Network，SCN），根据生成和选择的视频提案的视觉背景预测查询中重要的掩码词汇。

Song 等人 [54] 利用注意力重建思想提出了一种多级注意力重建网络（Multi-level Attentional Reconstruction Network，MARN），使用提案级注意力对片段候选进行排序。

Duan 等人 [14] 解决了视频中的弱监督密集事件描述问题，这是弱监督 TSGV 的一个对偶问题，他们提出了一个可以同时解决对偶问题的循环系统，弱监督 TSGV 可以被看作这种循环系统中的一个中间任务。与 Duan 等人 [14] 类似，Chen 和 Jiang[11] 也采用了一个循环系统处理密集事件描述任务。

此外，Zhang 等人 [86] 没有基于重建或 MIL，而是设计了一个反事实的对比学习范式改进视觉和语言的基础任务。

数据集与评估

数据集

DiDeMo[23] 由 Flickr 收集，包含个人用户上传的各种人类活动。Hendricks 等人 [23] 通过聚合 5 秒钟的片段单位，从原始的未剪辑的视频中分割并标记视频片段，基准片段的长度是 5 秒的倍数。DiDeMo 中有 33008、4180 和 4022 个视频 – 句子对，分别用于训练、验证和测试。

TACoS[47] 是基于 MPII-Compositive 数据集建立的 [49]。它包含了 127 个以烹饪活动为主题的复杂视频，每个视频都有几个片段，用句子描述烹饪活动。视频的平均长度约为 300 秒，比其他基准数据集的视频长度长很多。该数据集的句子 – 片段对总数为 17344 个，其中 50%、25%、25% 分别用于训练、验证和测试。

Charades-STA[16] 基于 Charades[52] 建立，Charades 最初是为视频活动识别而收集的。该数据集包含 13898 个用于训练的句子 – 片段对，4233 个简单句子 – 片段对（每句 6.3 个单词），以及 1378 个用于测试的复杂句子 – 片段对（每句 12.4 个单词）。

ActivityNet Captions[31] 最初是为视频密集事件描述生成而提出的。ActivityNet Captions 包含的视频数量最多，它将视频与一系列有时间标记的句子描述相匹配。每个句子的平均长度为 13.48 个单词，句子长度呈正态分布。训练集有 10009 个视频和 37421 个句子 – 片段对，测试集有 4917 个视频和 34536 个句子 – 片段对。

指标

TSGV 有两类指标，即 mIoU（即平均 IoU）和 R@n, IoU=m。IoU 在物体检测中被广泛用于评估两个边界框之间的相似性，TSGV 也类似，采用时序 IoU 衡量片段相似性。指标 mIoU 通过平均所有样本的时序 IoU 来评估结果。另一个常用的指标是 R@n, IoU=m[25]。对于样本 i，如果当前 n 个被检索的片段中存在一个与基准片段的时间 IoU 超过 m 的片段时，则视为检索成功。R@n, IoU=m 是检索成功的样本占所有样本的百分比。研究者习惯设置 $n \in \{1, 5, 10\}$ 和 $m \in \{0.3, 0.5, 0.7\}$。通常，当方法采用无提案方式（即属于无锚点或基于强化学习的框架）时，n=1。

性能比较

我们基于四个基准数据集对上述方法的性能进行了全面比较。

两阶段方法 如表 1 所示，两阶段方法的总体性能似乎比其他方法差。可能的原因有三个方面。（1）大多数两阶段方法只粗略地结合了视频和句子的特征，而忽略了细粒度的视觉和文本交互。（2）将候选片段生成和句段匹配过程分开，使模型无法进行全局优化。（3）在句子查询和单个片段之间建立匹配关系将使局部视频内容与全局视频上下文分离。

具体来说，MCN 在 Charades-STA 数据集上得到的结果最差，这表明其简单的候选片段多模态匹配和排序策略不能很好地处理各种灵活位置的片段。而 CTRL、ACRN、ROLE、SLTA 和 ACL-K 可以根

表1 两阶段方法的性能比较（SW：基于滑动窗口，PG：基于提案生成）

类型	方法	DiDeMo			TACoS			Charades-STA			ActivityNet Captions		
		0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
SW	MCN[23]	—	—	—	—	—	—	13.57	4.05	—	—	—	—
	CTRL[16]	—	—	—	18.32	13.3	—	—	23.63	8.89	—	—	—
	MCF[63]	—	—	—	18.64	12.53	—	—	—	—	—	—	—
	ROLE[38]	29.4	15.68	—	—	—	—	25.26	12.12	—	—	—	—
	ACRN[37]	—	—	—	19.52	14.62	—	—	—	—	—	—	—
	SLTA[27]	—	30.92	17.16	17.07	11.92	—	38.96	22.81	8.25	—	—	—
	ACL-K[18]	—	—	—	24.17	20.01	—	—	30.48	12.2	—	—	—
PG	QSPN[67]	—	—	—	—	—	—	54.7	35.6	15.8	45.3	27.7	13.6
	SAP[9]	—	—	—	—	18.24	—	—	27.42	13.36	—	—	—

据模型的位置偏移预测调整候选片段的边界，从而提高了方法的性能。

与基于滑动窗口的方法相比，提案生成方法在提案候选数量减少的情况下取得了更好的性能。QSPN 在 Charades 数据集上的性能明显优于其他两阶段方法，验证了所提出的提案生成网络能够以更细的时间粒度提供更有效的候选时刻。

端到端方法 如表 2 所示，对于基于锚点的方法，TGN 在 TACoS 和 ActivityNet Captions 数据集上取得了最低的性能。CMIN 在 TACoS 上的表现也很差。TGN、CMIN 和 CBP 的准确性普遍较差可能归因于它们的单流锚点定位框架，无法通过连续的 RNN 推理出复杂的跨模态关系。SCDM 和 MAN 都没有采用基于 RNN 的框架，而是使用卷积神经网络更好地捕捉细粒度的互动和不同时间粒度的不同视频内容，从而持续取得更好的性能。此外，CSMGAN、FIAN、SMIN 和 Zhang 等人 [82] 在 ActivityNet Captions 数据集上都取得了优异的结果。值得注意的是，虽然 CSMGAN 采用了类似 TGN 的顺序 RNN，但它建立了一个联合图对跨模态 / 自模态关系进行建模，可以有效地捕捉两种模态之间的高阶互动，而 FIAN 采用对称的迭代注意力获得更鲁棒的跨模态特征，以实现更准确的定位。

对于无锚点方法，包括 ExCL、VSLNet 和 Rodriguez 等人 [48] 在内的受机器阅读理解任务启发的方法的表现大幅优于其他方法。包括 DRN、GDP 和 DEBUG 在内的密集无锚点方法的表现优于早期的稀疏回归网络 ABLR，证明了增加正训练样本数量的重要性。而包括 PMI、HVTG 和 LGI 在内的其他基于回归的方法在 ActivityNet Captions 数据集上取得了优异的表现。

基于强化学习的方法 表 3 的上半部分展示了基于强化学习的方法的表现。可以看到，TSP-PRL 在 ActivityNet Captions 上取得了很好的性能，证明了借用人类决策过程从粗略到精细的想法的有效性。虽然基于强化学习的方法不能达到端到端的最先进方法的性能，但它们为解决 TSGV 任务和提高可解释性提供了全新的思路。

弱监督方法 表 3 的下半部分展示了弱监督方法的实验结果。我们无法根据整体性能判断哪个框架（即基于 MIL 或基于重构）有绝对的进步。具体而言，在所有弱监督方法中，CRM 在 Charades-STA 和 ActivityNet Captions 数据集上取得了最好的表现。与其他全监督方法相比，其结果也很有竞争力。

讨论

当前评价基准的局限

尽管 TSGV 取得了令人鼓舞的成果，但最近也有一些工作对当前评价指标提出了质疑。一些研

表2 端到端方法的性能比较（AB：基于锚点的，AF：无锚点的，OT：其他）

类型	方法	TACoS			Charades-STA			ActivityNet Captions		
		0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
AB	TGN[5]	21.77	18.9	—	—	—	—	45.51	28.47	
	MAN[78]	—	—	—	—	46.53	22.72	—	—	—
	CMIN[84]	24.64	18.05	—	—	—	—	63.61	43.4	23.88
	SCDM[73]	26.11	21.17	—	—	54.44	33.43	54.8	36.75	19.86
	CBP[60]	27.31	24.79	19.1	—	36.8	18.87	54.3	35.76	17.8
	2D-TAN[83]	37.29	25.32	—	—	39.7	23.31	59.45	44.51	26.54
	FIAN[46]	33.87	28.58	—	—	58.55	37.72	64.1	47.9	29.81
	CSMGAN[36]	33.9	27.09	—	—	—	—	68.52	49.11	29.15
	SMIN[59]	48.01	35.24	—	—	64.06	40.75	—	48.46	30.34
	Zhang等人[82]	48.79	37.57	—	—	—	—	—	48.02	31.78
AF	ABLR[75]	18.9	9.3	—	—	—	—	55.67	36.79	—
	DEBUG[39]	23.45	—	—	54.95	37.39	17.69	55.91	39.72	—
	GDP[7]	24.14	—	—	54.54	39.47	18.49	56.17	39.27	—
	PMI[8]	—	—	—	55.48	39.73	19.27	59.69	38.28	17.83
	ExCL[19]	44.4	27.8	14.6	61.4	41.2	21.3	62.1	41.6	23.9
	DRN[76]	—	23.17	—	—	45.4	26.4	—	42.49	22.25
	HVTG[10]	—	—	—	61.37	47.27	23.3	57.6	40.15	18.27
	Rodriguez等人[48]	24.54	21.65	16.46	67.53	52.02	33.74	51.28	33.04	19.26
	LGI[44]	—	—	—	72.96	59.46	35.48	58.52	41.51	23.07
	VSLNet[80]	29.61	24.27	20.03	70.46	54.19	35.22	63.16	43.22	26.16
OT	BPNet[66]	25.96	20.96	14.08	55.46	38.25	20.51	58.98	42.07	24.69
	DPIN[58]	46.74	32.92	—	—	47.98	26.96	62.4	47.27	28.31
	CBLN[35]	38.98	27.65	—	—	61.13	38.22	66.34	48.12	27.6

究[45, 72]将基准片段的时间位置分布可视化，发现基准片段的起始和结束时间戳在训练和测试集中的联合分布相同且有明显的分布偏差。他们设计了一些简单基线方法，不需要任何有效的视觉和文本输入，性能就能超过一些精心设计的深度模型。此外，Yuan等人[72]重新组织了两个基准数据集，创建了两个不同的测试集：一个测试集遵循与训练集相同的时间位置分布，即test-iid；另一个测试集与训练集的分布完全不同，即test-ood。在比较了各种基线方法在这两个测试集上的实验结果后，他们发现几乎所有的方法在test-ood上的表现都明显下降，这表明现有的方法受时间标注偏差的影响很大，并没有真正模拟视频和文本之间的语义匹配关系。因此，对于未来的工作，构建去偏差的数据集和建立不受偏差影响的鲁棒模型是至关重要的。

未来研究方向

大规模视频语料库时刻检索 大规模视频语料库时刻检索（Video Corpus Moment Retrieval, VCMR）是在TSGV基础上不断探索的一个研究方向[15, 32, 77, 79]。它具有更高的应用价值，因为可以从大规模视频语料库（即未修剪和未分割的视频集合）而不是从单个视频中检索与给定文本语义对应的目标片段。与TSGV相比，VCMR有更高的效率要求，因为它不仅需要从单个视频中检索特定片段，而且还要从视频语料库中定位目标视频。

时空定位 视频中的时空语句定位是TSGV的另一个扩展，它主要从视频中通过自然语言描述将

表3 基于强化学习的方法和弱监督方法的性能比较（RL：基于强化学习的，WS：弱监督的）

类型	方法	TACoS			Charades-STA			ActivityNet Captions		
		0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
RL	R-W-M[22]	—	—	—	—	36.7	—	—	36.9	—
	SM-RL[62]	20.25	15.95	—	—	24.36	11.17	—	—	—
	TripNet[21]	—	—	—	51.33	36.61	14.5	48.42	32.19	13.93
	TSP-PRL[65]	—	—	—	—	45.45	24.75	56.02	38.82	—
	STRONG[2]	72.14	49.73	18.29	78.1	50.14	19.3	—	—	—
	AVMR[3]	72.16	49.13	—	77.72	54.59	—	—	—	—
WS	WSDEC[14]	—	—	—	—	—	—	41.98	23.34	—
	TGA[43]	—	—	—	32.14	19.94	8.84	—	—	—
	WSLLN[17]	—	—	—	—	—	—	42.8	22.7	—
	EC-SL[11]	—	—	—	—	—	—	44.29	24.16	—
	SCN[34]	—	—	—	42.96	23.58	9.97	47.23	29.22	—
	Chen等人[12]	—	—	—	39.8	27.3	12.9	44.3	23.6	—
	VLANet[40]	—	—	—	45.24	31.83	14.17	—	—	—
	MARN[54]	—	—	—	48.55	31.94	14.81	47.01	29.95	—
	RTBPN[85]	—	—	—	60.04	32.36	13.24	49.77	29.63	—
	BAR[64]	—	—	—	44.97	27.04	12.23	49.03	30.73	—
	CCL[86]	—	—	—	—	33.21	15.68	50.12	31.07	—
	LoGAN[55]	—	—	—	51.67	34.68	14.54	—	—	—
	CRM[26]	—	—	—	53.66	34.76	16.37	55.26	32.19	—

指定对象或实例定位为连续的时空管道（即边界框序列）。由于对 STSGV 定位管道的细粒度标记过程（即为视频中的每一帧标注一个空间区域）费力且复杂，Chen 等人 [13] 提出用新构建的 VID-sentence 数据集以弱监督的方式解决这一问题，只需要视频级别的描述即可。

音频增强定位 目前 TSGV 的输入只包含给定的语句和未剪辑的视频，然而音频信号并没有得到有效的利用，这些信号可能为视频定位提供额外的指导，例如，在厨房使用电子产品时发出的巨大噪音，或足球运动员踢球时观众的欢呼声。这些不同形式的声音为更精确地定位目标时刻提供了辅助性且必不可少的线索，而这一点还没有被探索。如今，在基于音频增强辅助的视觉和语言领域已经有很多工作 [24, 68]，证明了其对性能改进的有效性。因此，在 TSGV 任务中结合音频线索是一个很有前景的未来方向。

广义视频定位

进一步，我们提出了广义视频定位（generalized video grounding）的概念，包括基于文本的视频定位、基于事件的视频定位以及基于音频的视频定位。广义视频定位涉及多种媒体形式，可以应用于各类多模态场景。

基于文本的视频定位 基于自然语言文本的视频时序定位是连接计算机视觉和自然语言处理的一项基本的、具有挑战性的任务，可以被看作一些下游视频理解应用的中间任务，例如视频问题回答、视频总结和视频内容检索，因此具有研究价值。

基于事件的视频定位 不同于使用文本进行的视频定位，基于事件的视频定位只需要根据给定的事件进行时间和空间上的定位（主要是时间上），而不需要输入完整的句子。相比整个句子，用词语或者短语描述的事件不需要给定主语和宾语，通常

代表更加抽象的概念，从而能更好地表达视频中的动态信息。比如“人在跑”和“狗在跑”，基于文本判断两者可能是不同的动作，但是在事件上两者都表示了“跑”这一视频中的动态，在某些对动作而不是对目标更加敏感的任务（例如目标跟踪）中，基于事件的视频定位能够比基于文本发挥更大的作用。

在具体的研究中，基于事件的视频定位通常以时序动作定位和时序动作检测等方式展开，主要针对人类的活动进行定位，这一任务通常与动作识别任务共同进行。Zhao 等人 [87] 提出了一种结构化的分段网络 SSN，通过时间金字塔对每个动作实例的时间结构建模。Chao 等人 [89] 提出的 TAL-Net 进一步改进了动作检测的框架，从而能够更好地应对时间尺度的变化和聚合上下文信息。

基于声音的视频定位 同样地，基于声音的视频定位与前述几种定位的主要区别是输入的模态不同，基于声音的视频定位首先需要理解语音中包含的信息，然后在视频中找到包含这些信息的对象。基于声音的视频定位的优势是声音往往跟视频同时出现，而不用人工标注它们之间的对应关系，因此能够轻易地获得大量的视频 – 语音对。

与前面基于文本和基于事件的视频定位不同，基于声音的视频定位着眼于空间定位的方法更多一些，还有一些方法关注基于声音的时序视频定位。首先，直接根据输入语音查询视频中发出该声音的时间，该问题通常被描述为跨模态定位的一个子集。该问题的代表方法有 AVDLN[94] 和 DCCA[95]。其次，另一个具体问题是视听事件定位问题，这个问题的目标是定位视频中可见和可听的事件。Ohishi 等人 [10] 提出了一种视频和语音的共分割方法，该方法采用了一个引导注意方案，利用信号中包含的相应音频和视频实体的时间接近性，有效地检测和利用音频和视频信息的时间共存。

总结

视频中的时序定位是一个基础性的、具有挑战性的多模态任务。在本文中，我们系统地总结了当前 TSGV 的研究进展，包括对现有方法进行分类，介绍当前基准数据集和评估指标，还提出了当前基准的局限性以及对未来研究方向的细致思考，希望能进一步促进 TSGV 的发展。对于未来工作，我们建议提出受数据偏差影响更小的数据集和评价指标，从而能够更可靠地评估模型，以及设计出在动态场景中具有鲁棒性和泛化能力的定位模型。 ■

王 鑫

CCF 专业会员。清华大学计算机系助理研究员。国家优秀青年科学基金获得者主要研究方向为媒体大数据，多媒体智能，机器学习。xin_wang@tsinghua.edu.cn

朱文武

CCF 会士。清华大学计算机系教授。AAAS/IEEE/ACM/SPIE Fellow，欧洲科学院外籍院士。主要研究方向为多媒体网络计算、大数据智能等。
wwzhu@tsinghua.edu.cn

参考文献

[1] Buch S, Escorcia V, Ghanem B, et al. 2017. End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos[C]// *British Machine Vision Conference 2017*.

[2] Cao D, Zeng Y, Liu M, et al.STRONG: Spatio-Temporal Reinforcement Learning for Cross-Modal Video Moment Localization[C]// *MM'20: The 28th ACM International Conference on Multimedia*. ACM, 2020.

[3] Cao D, Zeng Y, Wei X, et al.Adversarial Video Moment Retrieval by Jointly Modeling Ranking and Localization[C]// *MM'20: The 28th ACM International Conference on Multimedia*. ACM, 2020.

[4] Chen D, Fisch A, Weston J, et al. Reading Wikipedia to Answer Open-Domain Questions[C]// *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. ACL, 2017: 1870-1879.

[5] Chen J, Chen X, Ma L, et al. Temporally Grounding Natural Sentence in Video[C/OL]// *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL, 2018: 162-171. https://doi.org/10.18653/v1/D18-1015.

[6] Chen J, Ma L, Chen X, et al. 2019. Localizing

Natural Language in Videos[C]// *AAAI 2019*. AAAI Press, 2019: 8175-8182.

[7] Chen L, Lu C, Tang S, et al. Rethinking the Bottom-Up Framework for Query-Based Video Localization[C]// *AAAI 2020*. AAAI Press, 2020: 10551-10558.

[8] Chen S, Jiang W, Liu W, et al. Learning modality interaction for temporal sentence localization and event captioning in videos[C]// *European Conference on Computer Vision*. Springer, 2020: 333-351.

[9] Chen S, Jiang Y. Semantic Proposal for Activity Localization in Videos via Sentence Query[C/OL]// *AAAI 2019*. AAAI Press, 2019: 8199-8206. https://doi.org/10.1609/aaai.v33i01.33018199.

[10]Chen S, Jiang Y. Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language[C]// *European Conference on Computer Vision*. Springer, 2020: 601-618.

[11]Chen S, Jiang Y. Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning[C]// *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021: 8425-8435.

[12]Chen Z, Ma L, Luo W, et al. Look closer to ground better: Weakly-supervised temporal grounding of sentence in video[OL]. arXiv preprint arXiv:2001.09308 (2020). https://arxiv.org/abs/2001.09308

[13]Chen Z, Ma L, Luo W, et al. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video[C/OL]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, 2019: 1884-1894. https://doi.org/10.18653/v1/P19-1183

[14]Duan X, Huang W, Gan C, et al. Weakly Supervised Dense Event Captioning in Videos[C]// NeurIPS 2018. 2018: 3063-3073.

[15]Escorcia V, Soldan M, Sivic J, et al. Temporal localization of moments in video collections with natural language[OL]. arXiv preprint arXiv:1907.12763 (2019). https://arxiv.org/abs/1907.12763.

[16]Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 5277–5285. https://doi.org/10.1109/ICCV.2017.563

[17]Mingfei Gao, Larry Davis, Richard Socher, and Caiming Xiong. 2019. WSLLN:Weakly Supervised Natural Language Localization Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 1481–1487. https://doi.org/10.18653/v1/D19-1157

[18]Runzhou Ge, Jiyang Gao, Kan Chen, and Ram Nevatia. 2019. Mac: Mining activity concepts for language-based temporal localization. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 245–253.

[19]Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. 2019. ExCL: Extractive Clip Localization Using Natural Language Descriptions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 1984–1990. https://doi.org/10.18653/v1/N19-1198

[20]Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. IEEE Computer Society, 580–587. https://doi.org/10.1109/CVPR.2014.81

[21]Meera Hahn, Asim Kadav, James M. Rehg, and Hans Peter Graf. 2020. Tripping through time: Efficient Localization of Activities in Videos. In 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020. BMVA Press. https://www.bmvc2020-conference.com/assets/papers/0549.pdf

[22]Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. 2019. Read, Watch, and Move: Reinforcement Learning for Temporally Grounding Natural Language Descriptions in Videos. AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 8393–8400. https://doi.org/10.1609/aaai.v33i01.33018393

[23]Lisa Anne Hendricks, Oliver Wang, Eli Shechtman,

Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing Moments in Video with Natural Language. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 5804–5813. https://doi.org/10.1109/ICCV.2017.618

[24]Chiori Hori, Huda AlAmri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Irfan Essa, Dhruv Batra, and Devi Parikh. 2019. End-to-end Audio Visual Scene-aware Dialog Using Multimodal Attention-based Video Features. In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019. IEEE, 2352–2356. https://doi.org/10.1109/ICASSP.2019.8682583

[25]Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural Language Object Retrieval. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 4555–4564. https://doi.org/10.1109/CVPR.2016.493

[26]Jiabo Huang, Yang Liu, Shaogang Gong, and Hailin Jin. 2021. Cross-Sentence Temporal and Semantic Relations in Video Activity Localisation. arXiv preprint arXiv:2107.11443 (2021). https://arxiv.org/abs/2107.11443

[27]Bin Jiang, Xin Huang, Chao Yang, and Junsong Yuan. 2019. Cross-modal video moment retrieval with spatial and language-temporal attention. In Proceedings of the 2019 on international conference on multimedia retrieval. 217–225.

[28]Yifan Jiao, Zhetao Li, Shucheng Huang, Xiaoshan Yang, Bin Liu, and Tianzhu Zhang. 2018. Three-dimensional attention-based deep ranking model for video highlight detection. IEEE Transactions on Multimedia 20, 10 (2018), 2693–2705.

[29]Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. 2014. Fast saliency based pooling of fisher encoded dense trajectories. In ECCV THUMOS Workshop, Vol. 1. 5.

[30]Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl

[31]Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017. IEEE Computer Society, 706–715. https://doi.org/10.1109/ICCV.2017.83

[32]Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. Springer, 447–463.

[33]Tianwei Lin, Xu Zhao, and Zheng Shou. 2017. Single Shot Temporal Action Detection. In Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017. 988–996. https://doi.org/10.1145/3123266.3123343

[34]Zhijie Lin, Zhou Zhao, Zhu Zhang, Qi Wang, and Huasheng Liu. 2020. Weakly-Supervised Video Moment Retrieval via Semantic Completion Network. AAAI 2020, New York, NY, USA, February 7-12, 2020. AAAI Press, 11539–11546. https://aaai.org/ojs/index.php/AAAI/article/view/6820

[35]Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware Biaffine Localizing Network for Temporal Sentence Grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11235–11244.

[36]Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly Cross- and Self-Modal Graph Attention Network for Query-Based Moment Localization. In MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. 4070–4078. https://doi.org/10.1145/3394171.3414026

[37]Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, Kevyn

Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 15–24. https://doi.org/10.1145/3209978.3210003

[38]Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In 2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018. 843–851. https://doi.org/10.1145/3240508.3240549

[39]Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. DEBUG: A Dense Bottom-Up Grounding Approach for Natural Language Video Localization. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 5144–5153. https://doi.org/10.18653/v1/D19-1518

[40]Minuk Ma, Sunjae Yoon, Junyeong Kim, Youngjoon Lee, Sunghun Kang, and Chang D Yoo. 2020. Vlanet: Video-language alignment network for weakly-supervised video moment retrieval. In European Conference on Computer Vision. Springer, 156–171.

[41]Shugao Ma, Leonid Sigal, and Stan Sclaroff. 2016. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 1942–1950. https://doi.org/10.1109/CVPR.2016.214

[42]Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In Proceedings of the tenth ACM international conference on Multimedia. 533–542.

[43]Niluthpol Chowdhury Mithun, Sujoy Paul, and Amit K. Roy-Chowdhury. 2019. Weakly Supervised Video Moment Retrieval From Text Queries. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 11592–11601. https://doi.org/10.1109/CVPR.2019.01186

[44]Jonghwan Mun, Minsu Cho, and Bohyung Han. 2020. Local-Global Video-Text Interactions for Temporal Grounding. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. IEEE, 10807–10816. https://doi.org/10.1109/CVPR42600.2020.01082

[45]Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. 2020. Uncovering Hidden Challenges in Query-Based Video Moment Retrieval. In 31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020. BMVA Press. https://www.bmvc2020-conference.com/assets/papers/0306.pdf

[46]Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained Iterative Attention Network for Temporal Language Localization in Videos. In MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. 4280–4288. https://doi.org/10.1145/3394171.3414053

[47]Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding Action Descriptions in Videos. Transactions of the Association for Computational Linguistics 1 (2013), 25–36. https://doi.org/10.1162/tacl_a_00207

[48]Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. 2020. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2464–2473.

[49]Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In European conference on computer vision. Springer, 144–157.

[50]Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video Object Grounding Using Semantic Roles in Language Description. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. IEEE, 10414–10424. https://doi.org/10.1109/CVPR42600.2020.01043

[51]Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs. In 2016 IEEE

Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 1049–1058. https://doi.org/10.1109/CVPR.2016.119

[52]Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In European Conference on Computer Vision. Springer, 510–526.

[53]Bharat Singh, Tim K. Marks, Michael J. Jones, Oncel Tuzel, and Ming Shao. 2016. A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 1961–1970. https://doi.org/10.1109/CVPR.2016.216

[54]Yijun Song, Jingwen Wang, Lin Ma, Zhou Yu, and Jun Yu. 2020. Weakly-supervised multi-level attentional reconstruction network for grounding textual queries in videos. arXiv preprint arXiv:2003.07048 (2020). https://arxiv.org/abs/2003.07048

[55]Reuben Tan, Huijuan Xu, Kate Saenko, and Bryan A Plummer. 2021. Logan: Latent graph co-attention network for weakly-supervised video moment retrieval. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2083–2092.

[56]Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. 2021. Human-centric spatio-temporal video grounding with visual transformers. IEEE Transactions on Circuits and Systems for Video Technology (2021).

[57]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[58]Hao Wang, Zheng-Jun Zha, Xuejin Chen, Zhiwei Xiong, and Jiebo Luo. 2020. Dual Path Interaction Network for Video Moment Localization. In MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. 4116–4124. https://doi.org/10.1145/3394171.3413975

[59]Hao Wang, Zheng-Jun Zha, Liang Li, Dong Liu, and Jiebo Luo. 2021. Structured Multi-Level Interaction Network for Video Moment Localization via Language Query. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7026–7035.

[60]Jingwen Wang, Lin Ma, and Wenhao Jiang. 2020. Temporally grounding language queries in videos by contextual boundary-aware prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 12168–12175.

[61]Limin Wang, Yu Qiao, and Xiaoou Tang. 2014. Action recognition and detection by combining motion and appearance features. THUMOS14 Action Recognition Challenge 1, 2 (2014), 2.

[62]Weining Wang, Yan Huang, and Liang Wang. 2019. Language-Driven Temporal Activity Localization: A Semantic Matching Reinforcement Learning Model. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 334–343. https://doi.org/10.1109/CVPR.2019.00042

[63]Aming Wu and Yahong Han. 2018. Multi-modal Circulant Fusion for Video-to-Language and Backward. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, Jérôme Lang (Ed.). ijcai.org, 1029–1035. https://doi.org/10.24963/ijcai.2018/143

[64]Jie Wu, Guanbin Li, Xiaoguang Han, and Liang Lin. 2020. Reinforcement Learning for Weakly Supervised Temporal Grounding of Natural Language in Untrimmed Videos. In MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. 1283–1291. https://doi.org/10.1145/3394171.3413862

[65]Jie Wu, Guanbin Li, Si Liu, and Liang Lin. 2020. Tree-structured policy based progressive

reinforcement learning for temporally language grounding in video. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 12386–12393.

[66]Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. 2021. Boundary Proposal Network for Two-Stage Natural Language Video Localization. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 2986–2994.

[67]Huijuan Xu, Kun He, Bryan A Plummer, Leonid Sigal, Stan Sclaroff, and Kate Saenko. 2019. Multilevel language and vision integration for text-to-clip retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 9062–9069.

[68]Yuecong Xu, Jianfei Yang, and Kezhi Mao. 2019. Semantic-filtered Soft-Split-Aware video captioning with audio-augmented feature. Neurocomputing 357 (2019), 24–35.

[69]Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded Video Moment Retrieval with Causal Intervention. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery, 1–10. https://doi.org/10.1145/3404835.3462823

[70]Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 982–990. https://doi.org/10.1109/CVPR.2016.112

[71]Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-End Learning of Action Detection from Frame Glimpses in Videos. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2678–2687. https://doi.org/10.1109/CVPR.2016.293

[72]Yitian Yuan, Xiaohan Lan, Long Chen, Wei Liu, and Wenwu Zhu. 2021. A Closer Look at Temporal Sentence Grounding in Videos: Datasets and Metrics. arXiv preprint arXiv:2101.09028 (2021). https://arxiv.org/abs/2101.09028

[73]Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. 2019. Semantic Conditioned Dynamic Modulation for Temporal Sentence Grounding in Videos. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 534–544. https://proceedings.neurips.cc/paper/2019/hash/6883966fd8f918a4aa29be29d2c386fb-Abstract.html

[74]Yitian Yuan,Tao Mei, Peng Cui,and Wenwu Zhu.2017. Video summarization by learning deep side semantic embedding. IEEE Transactions on Circuits and Systems for Video Technology 29, 1 (2017), 226–237.

[75]Yitian Yuan, Tao Mei, and Wenwu Zhu. 2019. To find where you talk: Temporal sentence localization in video with attention based location regression. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. 9159–9166.

[76]Runhao Zeng, Haoming Xu, Wenbing Huang, Peihao Chen, Mingkui Tan, and Chuang Gan. 2020. Dense Regression Network for Video Grounding. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020. IEEE, 10284–10293. https://doi.org/10.1109/CVPR42600.2020.01030

[77]Bowen Zhang, Hexiang Hu, Joonseok Lee, Ming Zhao, Sheide Chammas, Vihan Jain, Eugene Ie, and Fei Sha. 2020. A Hierarchical Multi-Modal Encoder for Moment Localization in Video Corpus. arXiv preprint arXiv:2011.09046 (2020). https://arxiv.org/abs/2011.09046

[78]Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S. Davis. 2019. MAN: Moment Alignment Network for Natural Language Moment Retrieval via Iterative Graph Adjustment. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 1247–1257. https://doi.org/10.1109/CVPR.2019.00134

[79]Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. 2021. Video Corpus Moment Retrieval with Contrastive Learning. In Proceedings of the

44th International ACM SIGIR Conference on Research and Development in Information Retrieval (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery.

[80]Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. 2020. Span-based Localizing Network for Natural Language VideoLocalization.In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 6543–6554. https://doi.org/10.18653/v1/2020.acl-main.585

[81]Ke Zhang, Kristen Grauman, and Fei Sha. 2018. Retrospective encoders for video summarization. In Proceedings of the European Conference on Computer Vision (ECCV). 383–399.

[82]Mingxing Zhang, Yang Yang, Xinghan Chen, Yanli Ji, Xing Xu, Jingjing Li, and Heng Tao Shen. 2021. Multi-Stage Aggregated Transformer Network for Temporal Language Localization in Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12669–12678.

[83]Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 12870–12877.

[84]Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-Modal Interaction Networks for Query-Based MomentRetrievalinVideos.InProceedingsofthe42ndInternationalACMSIGIRConferenceonResearchandDevelopment in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019, Benjamin Piwowarski, Max Chevalier, Éric Gaussier, Yoelle Maarek, Jian-Yun Nie, and Falk Scholer (Eds.). ACM, 655–664. https://doi.org/10.1145/3331184.3331235

[85]Zhu Zhang, Zhijie Lin, Zhou Zhao, Jieming Zhu, and Xiuqiang He. 2020. Regularized Two-Branch Proposal Networks for Weakly-Supervised Moment Retrieval in Videos. In MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020. 4098–4106. https://doi.org/10.1145/3394171.3413967

[86]Zhu Zhang, Zhou Zhao, Zhijie Lin, Xiuqiang He,et al. 2020. Counterfactual Contrastive Learning for Weakly-Supervised Vision-Language Grounding. Advances in Neural Information Processing Systems 33 (2020), 18123–18134.

[87]Zhao Y, Xiong Y, Wang L, et al. Temporal action detection with structured segment networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2914-2923.

[88]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. 2017: 5783-5792.

[88]Chao Y W, Vijayanarasimhan S, Seybold B, et al. Rethinking the faster r-cnn architecture for temporal action localization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1130-1139.

[89]Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.

[90]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. 2017: 4325-4334.

[91]Zhai Y, Wang L, Tang W, et al. Two-stream consensus network for weakly-supervised temporal action localization[C]//European conference on computer vision. Springer, Cham, 2020: 37-54.

[92]Zhao H, Gan C, Rouditchenko A, et al. The sound of pixels[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 570-586.

[93]Tian Y, Shi J, Li B, et al. Audio-visual event localization in unconstrained videos[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 247-263.

[94]Andrew G, Arora R, Bilmes J, et al. Deep canonical correlation analysis[C]//International conference on machine learning. PMLR, 2013: 1247-1255.

[95]Ohishi Y, Tanaka Y, Kashino K. Unsupervised Co-Segmentation for Athlete Movements and Live Commentaries Using Crossmodal Temporal Proximity[C]//2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 9137-9142.