arXiv:2303.07748v1 [cs.CV] 14 Mar 2023

生成引导的多层级视频定位统一网络

Xing Cheng MMU 快手 北京，中国
chengxing03@kuaishou.com

Xiangyu Wu MMU 快手 北京，中国
w
uxiangyu@kuaishou.com

Dong Shen MMU 快手 北京，中国
shendong
@kuaishou.com

Hezheng Lin MMU 快手 北京，中国
linhezheng@kuaishou.co
m

Fan Yang MMU 快手 北京，中国
yangfan
@kuaishou.com

摘要

视频定位旨在未剪辑视频中找到与查询描述最匹配的时间戳。主流方法可分为时刻级和片段级框架。时刻级方法从全局视角直接预测每个瞬时时刻作为边界的概率，通常在粗粒度定位中表现更佳。另一方面，片段级方法将不同时间窗口中的时刻片段聚合为候选片段，再推断出最相似的候选片段，因此在细粒度定位方面具有优势。

本文提出一种多层级统一框架，通过发挥时刻级和片段级方法各自的优势来提升性能。此外，我们在两个层级中均采用了一种新颖的生成引导范式。该范式引入多模态生成器，生成隐式边界特征和片段特征，随后将其视为查询，由判别器计算边界分数。生成引导的解决方案将视频定位从双模态匹配任务转变为跨模态注意力任务，突破了以往的框架，并取得了显著提升。所提出的生成引导的多层级统一网络（GMU）超越了以往方法，在多个采用不同特征的基准数据集上达到最先进水平，例如 Charades-STA、ActivityNet Captions。

CCS 概念

• **信息系统 → 信息检索中的新颖性；多媒体与多模态检索**。

关键词

视频定位；跨模态；生成；多层级；

允许免费制作本作品全部或部分内容的数字或纸质副本，用于个人或课堂使用，前提是这些副本不得用于或分发以获取利润或商业利益，且副本须附有本声明及首页的完整引文。非 ACM 所有的本作品组件，其版权必须得到尊重。允许在注明出处的情况下摘录。若要以其他方式复制、重新出版、发布至服务器或分发至列表，则须事先取得明确许可和/或获得

费用。请通过 permissions@acm.org 申请许可。会议简称 *"XX"*，*2018*年*6*月*3*日至*5*日，纽约州伍德斯托克©2022 计算机协会。ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00 https://doi.org/XXXXXXX.XXXXXXX

ACM 引用格式：

Xing Cheng、Xiangyu Wu、Dong Shen、Hezheng Lin 和 Fan Yang。2022。生成引导的多层级视频定位统一网络。载于《请根据版权确认邮件中的信息填写正确的会议标题（会议简称*"XX"*）。》ACM，美国纽约州纽约市，10 页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

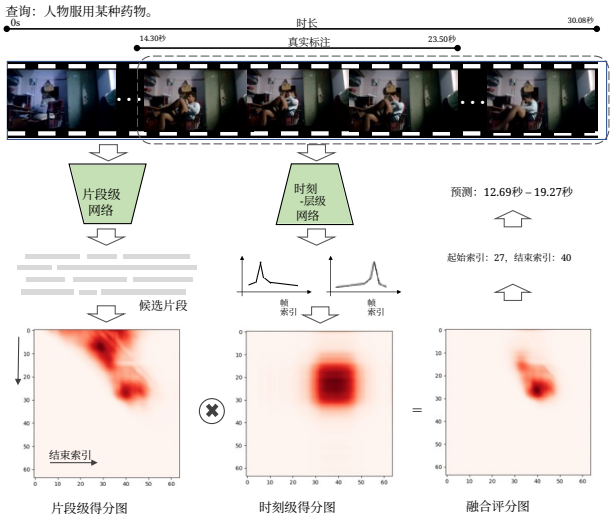


图1：视频定位任务简介及所提出 GMU 的整体框架。点的颜色越红，表示预测得分越高。片段级映射图中的 top-1 得分落入了错误的簇，而时刻级映射图在推导准确结果时较为模糊。将二者统一可以解决这两个缺陷。

视频定位 [6], 作为跨模态任务之一 [8, 10, 17,22, 39, 44],, 旨在精准定位与给定查询相对应的时间戳，例如图1所示的"人物正在服用某种药物"。我们需要推断出准确的时间戳 [14.3, 23.5]，其表示"服用某种药物"这一动作。以往的工作大致可分为时刻级方法和片段级方法。

时刻级方法 [19] 从全局视角处理整个视频，然后推断定位结果，或判断每个时刻成为边界的概率，如图2(a)所示。对起始概率和结束概率进行双仿射乘法 [4] 可得到得分图。这种方法在粗粒度定位方面表现良好，但由于采用了分开预测的策略，在细粒度定位方面表现不佳。片段级方法 [46] 将视频分割成不同的片段，然后检索最相似的候选片段，这称为片段级方法。它直接模拟真实标注（GT）的交并比（IoU）分布，并在对角线附近生成若干簇，因为整体真实标注都分布在对角线附近；正如 [36], 中所探讨的，错误的簇会导致较大偏差。参见图1，这种方法在粗粒度定位方面不如时刻级方法，而在细粒度定位方面则相反。我们在第4.3节进行了大量实验，并在第6节进行了可视化，以支持这一观点。

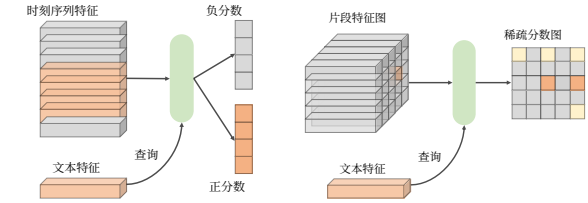
本文提出了一种多层级特征融合机制，以利用两种方法各自的优势。如图1所示，我们首先以一定的采样率对视频中的帧进行采样。我们采用预训练的 C3D [30], I3D [1], 或 VGG [29] 提取其特征。片段级组件枚举所有可能的候选片段并对其进行评分。时刻级组件则直接预测每个时刻位于边界处的概率。最后，我们通过哈达玛积融合这些得分图，从而得到更可靠的结果。进一步的实验表明，这种改进并非直接源于融合步骤本身，而是得益于纳入统一训练步骤后模型视觉和语义表征能力的增强。

此外，我们采用了生成引导架构，如图2(c)和图2(d)所示。在以往的片段级或时刻级方法中，视频定位任务在特征编码阶段分别独立处理视觉模态和文本模态，然后基于编码后的特征采用匹配策略 [19, 46]。由于不同模态之间存在异质性，学习一致的潜在空间可能颇具挑战。我们设计了生成器，在两个层级上生成目标特征；其中，时刻级的目标指边界，片段级的目标指目标片段。

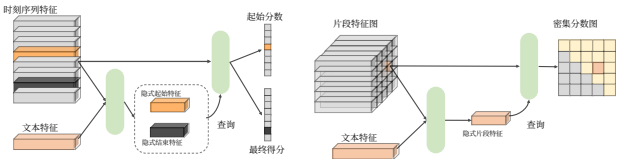
目标特征是跨模态表示，与视觉特征共享统一潜在空间。随后，它们作为查询输入判别器，而原始视觉特征则作为待匹配的值。这种架构显著缓解了跨模态异质性，消融实验表明其在视频定位任务中有效。

最后，以往的片段级方法通常采用稀疏预测策略。例如，如图 2(b) 所示，真实标注是片段特征图中坐标为 (2，3) 的橙色点。由于稀疏预测只考虑黄色点，因此预测结果会是 (2，2) 或 (2，4)，这意味着精度有所损失。为解决这一问题，我们在 BMN [13] 中使用采样掩码权重对帧进行稀疏采样，并采用图 2(d) 所示的密集预测策略，其中得分图中的所有点都会参与损失函数的计算。该策略显著提升了精度。我们的贡献总结如下：

- 我们提出了一种多层级统一框架，融合时刻级和片段级信息，并改进基础模块的表示能力。
- 我们摒弃常规做法，在两个层级上都采用生成引导架构，以增强跨模态特征的对齐，并引入注意力机制来优化视频定位任务。
- 通过多层级统一、生成引导，以及掩码语言模型（MLM）和密集预测等经验证有效的技巧，所提出的生成引导多层级统一框架（GMU）在时刻级和片段级组件上均优于基线方法，并在 Charades-STA 和 ActivityNet Caption 上达到最先进水平。



(a) 以往时刻级框架示意图。 (b) 以往片段级框架。



(c) 生成引导的时刻级框架示意图。(d) 生成引导的片段级框架示意图。

图 2：所提出的生成引导框架与以往直接方法的比较。在片段级结果图中，每个点表示一个候选片段；图的横轴表示起始索引，纵轴表示结束索引。黄色点表示纳入考虑的候选片段，橙色点表示模型从黄色点中选出的最终结果。

2 相关工作

本节首先详细介绍以往的时刻级 [9, 16, 19, 20, 32, 40, 42] 和片段级方法 [7, 37, 41, 45 47]。然后，我们还将介绍一些兼顾这两个层级的融合方法，以及其他一些跨模态领域中的生成方法。

时刻级。 </style}最初在 [6] 中提出，采用固定长度的滑动窗口将视频分割成短片段，并将其送入特征提取器，随后通过多模态交集模块和回归器进行定位推断。LGI [19]认为，上下文信息隐藏在中层语义阶段中，这些阶段由若干连续词语组成。它为多层级阶段设计了多种局部注意力机制，并通过全局上下文交集将它们统一起来。DRN [40]

将特征金字塔堆叠起来，以对视觉表示和语义表示进行相乘与交集运算。每一层都包含一个定位模块，其中有三个头，分别计算匹配分数、交并比得分和位置分数。

片段级。 2D-TAN [46] 枚举所有时间戳，并通过池化或卷积操作将其中的帧嵌入聚合为固定维度的特征。随后形成二维时间图，其中可以找到每个时间戳的表示。原始定位问题由此转化为检索任务。之后，研究者提出了基于 2D-TAN 的 MS-2D-TAN [45] 和 FMVC [7] 。MS-2D-TAN 将单尺度扩展为多尺度编码。FMVC 运用层次化语义引导注意力，挖掘语义角色树上的局部-全局依赖关系，从而更好地与候选片段匹配。

融合级。 DPIN [31] 首次提出了多层集成框架。BPNet [34] 采用相同的策略，并使用跨模态注意力。LPNet [33] 预先设计了一组人工设定的时刻候选项，以利用先验知识。它们都显著提升了效果。

生成方法。涉及视觉与自然语言的跨模态生成任务包括图像描述生成 [5, 35], 视频描述生成 [3, 8], 和文本生成图像 [18, 24]。这类任务通常采用编码器将输入转换为嵌入，再通过解码器显式地将其解码为另一种模态。在视频定位中，并不需要立即进行生成。然而，本文聚焦于生成框架：我们将原始解码器替换为判别器，并利用生成的嵌入隐式检索最优匹配。与以往架构相比，该框架显著提升了性能。

3 方法

GMU 的总体结构如图 3 所示。它包含四个组成部分：语义编码器、视觉基础模块、时刻级预测器和片段级预测器。它们分别用于编码文本、编码视觉帧、获取时刻级得分图和获取片段级得分图。本节将以 LGI 作为时刻级基线、以 2D-TAN 作为片段级基线，详细介绍各个模块。

3.1 语义编码器

为公平比较以往方法 [19], ，我们采用双向长短期记忆网络（BiLSTM）作为语义编码器。对于输入查询$Q = \{q_1, \ldots, q_L\}$（包含 L 个词），BiLSTM 将其编码为语义特征 s。为确保顶层网络保留语义信息，我们补充了掩码语言模型（MLM）任务。我们随机将一个词 q_i 替换为 $[mask]$，表示掩码词元，它是输出端需要预测的目标。BiLSTM 输出句子特征，并凭借自身优势自适应地利用长短期依赖关系： $s = BiLSTM(Q)$，其中 $s \in \mathbb{R}^d$。

3.2 视觉基础模块

该组件将视频编码为共享的视觉基础特征 V_f，供时刻级和片段级预测器使用。对于输入视频 V，首先按照固定间隔将其分割为时刻片段 $\{v_1, \ldots, v_T\}$，其中 T表示时刻片段的数量。它们由预训练的三维卷积神经网络（CNN）提取为 $V_o \in \mathbb{R}^{T \times d_i}$: $V_o = 3D_CNN(V)$，其中 d_i为初始维度。随后，通过设计的基础编码器将它们编码为 $V_f = \{f_1, \ldots, f_T\}$。不同于基础编码器仅由 FNN 构成的 2D-TAN [46], ，我们的方法还加入了可学习的位置嵌入。该过程可表示为:

$$V_f = ReLU(V_o W_{base} + W_{pos}) \tag{1}$$

其中， $W_{base} \in \mathbb{R}^{d_i \times d}$ 表示线性层中的参数，$W_{pos} \in \mathbb{R}^{T \times d}$ 表示位置嵌入。

3.3 时刻级预测器

在该组件中，输入 V_f 首先与语义特征 s 通过哈达玛积进行融合，以获得语义感知特征f_{sa}。随后，将其依次输入局部和全局注意力网络。因此，我们得到兼具全局视角和语义信息的整体特征 f_o。接着，我们设计了一个精细的时刻级生成器，将这些特征聚合为隐式起始特征 f_{start}、隐式结束特征 f_{end}、注意力分数 s_{att}和掩码词元特征 s_{mask}。生成的起始特征 f_{start}、结束特征 f_{end}与 V_f之间的相似度可直接产生起始分数 s_{start}和结束分数 s_{end}。整体过程可表示为:

$$f_{sa} = V_f \odot s \tag{2}$$

$$f_o = Attention(f_{sa}) \tag{3}$$

$$f_{start}, f_{end}, s_{att}, s_{mask} = Gen_{fl}(f_o) \tag{4}$$

$$s_{start} = Sim(V_f, f_{start}) \tag{5}$$

$$s_{end} = Sim(V_f, f_{end}) \tag{6}$$

$$S_m = BFM(s_{start}, s_{end}) \tag{7}$$

其中， $Attention$、Gen_{fl}表示局部-全局注意力和时刻层级生成器，具体介绍如下。 BFM 表示双仿射乘法，即 $S^i_m{}^j = \sqrt{s^i_{start}, s^j_{end}}$。 Sim表示用于衡量相似度的函数。其表达式为:

$$Sim(a, b) = W_a(||a|| \odot ||b||)W_b \tag{8}$$

其中， $a \in \mathbb{R}^d$， $b \in \mathbb{R}^{T \times d}$表示输入。 $W_a \in \mathbb{R}^{1 \times T}$和 $W_b \in \mathbb{R}^{d \times d}$ 表示全连接层的可学习参数。

3.3.1 **局部-全局注意力。**语义感知特征由残差块编码为局部注意力： $f_l = ResBlock(f_{sa})$， 其中 f_l表示局部注意力的输出。 $ResBlock$表示带有残差连接的 n_l 层基本块，其表达式为:

$$f_l^i = ReLU(f_l^{i-1} + BasicBlock(f_l^{i-1})) \tag{9}$$

其中， f_l^i表示残差块中第 i层的输出。 $BasicBlock$表示批归一化（BN）和修正线性单元之后的卷积网络。卷积网络的卷积核大小、步幅和填充分别设为 $\{1, 1, 0\}$。

我们采用n_g层非局部块（NLBlock），以进一步利用全局依赖关系:

$$\begin{aligned} f_o^i &= NLBlock(f_o^{i-1}) \\ &= f_o^{i-1} + (W_{vl} f_o^{i-1}) Softmax(\frac{(W_{qr} f_o^{i-1})^T (W_k f_o^{i-1})^T}{\sqrt{d}})^T \end{aligned} \tag{10}$$

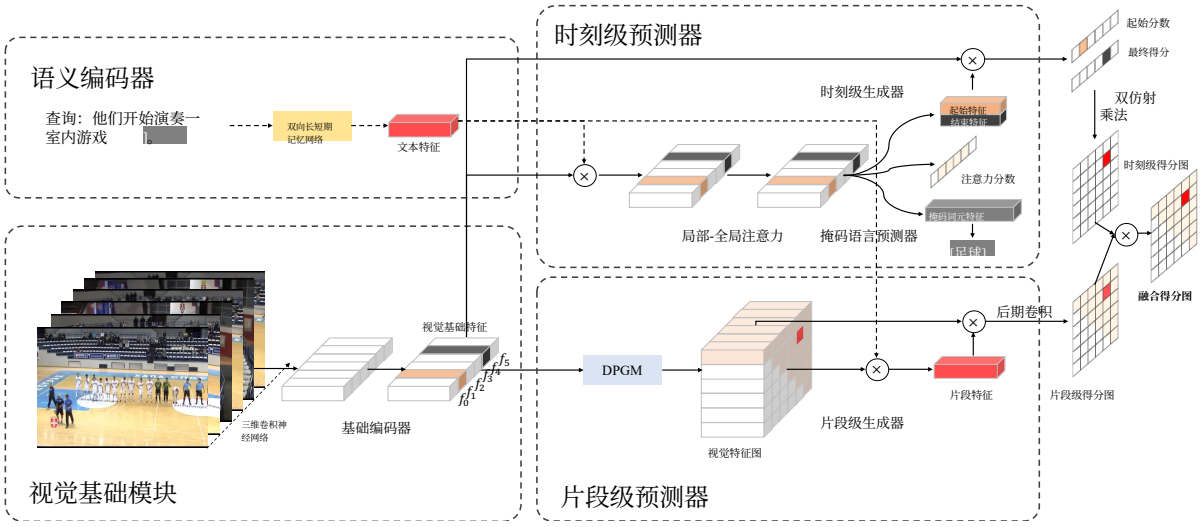


图 3：所提出 GMU 的架构。文本经过随机掩码后，由长短期记忆网络（LSTM）进行编码。视觉帧由预训练的三维卷积神经网络编码，随后输入所介绍的基于前馈神经网络（FNN）的模块。对于时刻级部分，设计了局部-全局注意力和一个生成器，用于生成隐式起始特征、结束特征、注意力分数和掩码词元特征。起始特征、结束特征与捷径视觉基础特征之间的相似度输入 FNN，以生成起始分数和结束分数；二者经双仿射乘法后生成时刻级得分图。对于 CLP，采用密集提议生成模块（DPGM）枚举由连续帧组成的所有可能片段。文本特征与视觉特征图融合后，融合特征进一步由片段级生成器编码，并生成隐式片段特征，该特征作为查询，用于在捷径视觉特征图中检索最相似的候选片段。时刻级得分图与片段级得分图的哈达玛积即为 GMU 的推理输出。

其中，W_{vl}、W_{qr}和 W_k是可学习矩阵，分别表示值、查询和键线性层。f_o^i表示第 i个非局部块的输出全局特征。在这种情况下，$f_o^0 = f_l^{n_l}$（全局注意力的输入）等于局部注意力的输出。

3.3.2 时刻级生成器。在该组件中，我们采用注意力池化策略 [19]，将多帧特征 $f_o \in \mathbb{R}^{T\times d}$聚合为一个特征 $f_a \in \mathbb{R}^d$，然后通过 FNN 将其投影为 f_{start}、f_{end}和 s_{mask}。

$$w_a = Softmax(FNN_{AP}(f_o)) \tag{11}$$

$$f_a = w_a f_o \tag{12}$$

$$f_{start}, f_{end} = FNN_{start}(f_a), FNN_{end}(f_a) \tag{13}$$

$$s_{mask} = FNN_{mask}(f_a) \tag{14}$$

其中，$w_a \in \mathbb{R}^T$是注意力权重，用于自适应地衡量每一帧的重要性分数。FNN_{AP}，FNN_{start}，FNN_{end}和 FNN_{mask}是彼此独立的前馈神经网络，可保留输入的原始维度。

3.4 片段级预测器（CLP）

所提出的 CLP 沿用了以往片段级方法的设计理念：根据坐标聚合特定帧，从而得到片段级特征图。我们并未通过堆叠卷积层来实现这一操作，而是引入了 [13] 中提出的采样掩码权重 W。此外，还采用了生成引导架构，以缩小视觉特征与语义特征之间的差距。

通常，首先使用密集提议生成模块（DPGM）和片段级生成器（CLG），分别生成视觉特征图 M_v 和隐式片段特征 f_c。随后，按照 $LateConv$，将 f_c 与捷径 M_v 进行哈达玛积运算，以得到片段级得分图（S_c）。$LateConv$ 由若干卷积层组成。此过程可大致表示为：

$$M_v = DPGM(V_f) \tag{15}$$

$$f_c = CLG(M_f) \tag{16}$$

$$S_c = LateConv(M_v \odot f_c) \tag{17}$$

3.4.1 密集提议生成模块（***DPGM***）。如图 4 所示，采样掩码权重 $W \in \mathbb{R}^{N\times T\times T\times T}$与视觉基础特征 $V_f \in \mathbb{R}^{d\times T}$进行矩阵乘法，生成采样后的候选片段特征图 $S_m \in \mathbb{R}^{N\times d\times T\times T}$，其中 N表示采样数量。值得注意的是，W中有三个维度等于 T，这是因为对于每个起始时间和结束时间，都有一个权重 $w \in \mathbb{R}^{T\times N}$。随后，我们对 N维度上的 S_m执行最大池化操作，得到视觉特征图 $M_v \in \mathbb{R}^{d\times T\times T}$。

$$M_v = Max(W \times V_f, dim = 0) \tag{18}$$

其中 $Max(\cdot\ dim = 0)$表示在 N维度上计算最大值。

3.4.2 片段级生成器（***CLG***）。 该模块旨在提取隐式片段特征 f_c，其与视觉信息之间的差距比语义特征更小。我们首先利用 $EarlyConv$ 扩大感受野，然后通过哈达玛积与语义特征 s 融合，以利用文本信息。

$$M_f = EarlyConv(M_v) \odot s \tag{19}$$

其中，$M_f \in \mathbb{R}^{T\times T\times d}$表示融合特征图。考虑到$M_f$中的真实标注在维度$d$的所有值都应达到最大值。我们采用 $MaxPool2d$ 来消除冗余特征。

$$f_c = MaxPool2d(M_f) \tag{20}$$

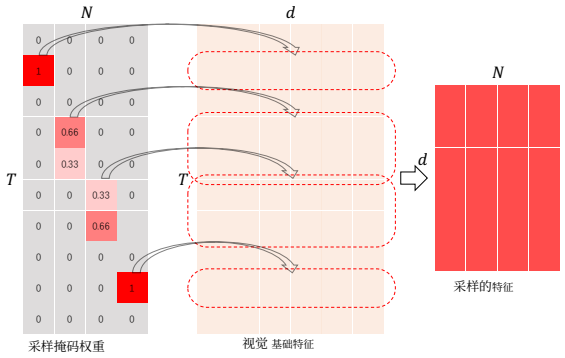


图 4：DPGM 中密集采样的示意图。我们采用 [13] 中提出的采样掩码权重 W，与视觉基础特征进行矩阵乘法。 W 中的非零值$w_{i,\ j}$ 以红色标注，表示以权重$w_{i,\ j}$对第 i个特征进行采样。

3.5 损失函数

对于每个视频及其查询，GMU 会输出起始分数 s_{start}、结束分数 s_{end}、掩码词元特征 s_{mask}、注意力权重 w_a 和片段级得分图 M_c。s_{start}和 s_{end}对应边界定位损失 L_{BL} [13]。其余分别对应交叉熵（CE）损失 L_{CE} [27]、时间注意力引导（TAG）损失 L_{TAG} [19] 和二元交叉熵（BCE）损失 L_{BCE} [46]。总损失由四部分组成。

$$\mathcal{L} = \mathcal{L}_{BL} + \mathcal{L}_{CE} + \mathcal{L}_{TAG} + \mathcal{L}_{BCE} \tag{21}$$

*3.5.1*边界定位损失。 s^i_{start}，s^i_{end}表示第 i 个时刻片段 m_i覆盖起始或结束时间戳的概率。请注意，m_i覆盖时段 $p_i = D_m[i-0.5, i+0.5]$，其中 $D_m = \frac{D}{T}$是每个时刻片段的时长。考虑到边界定义的模糊性，我们扩展起始和结束时刻片段的时长。我们使用 p_i与 $[t_s - 1.5D_m, t_s + 1.5D_m]$或 $[t_e - 1.5D_m, t_e + 1.5D_m]$的交并比作为标签，其中 t_s和 t_e表示起始和结束时间戳。

$$\mathcal{L}_{BL} = \mathcal{L}_{wbl}(s_{start}, l_{start}) + \mathcal{L}_{wbl}(s_{end}, l_{end}) \tag{22}$$

其中 l_{start}和 l_{end}表示起始和结束标签。L_{wbl}表示加权二元逻辑回归损失函数 [13]。

$$\mathcal{L}_{wbl}(s,l) = -\frac{1}{N_b}\sum_{i=1}^{N_b} \alpha^+ b_i \log s_i + \alpha^-(1-b_i)\log(1-s_i) \tag{23}$$

其中 $b_i = sign(l_i - \theta)$和 $N_b = B\cdot T$表示一个批次中的时刻数量 B。$\theta = 0.5$ 是一个阈值。 $\alpha^+ = \frac{N_b}{\sum b_i}$、 $\alpha^- = \frac{N_b}{N_b - \sum b_i}$ 分别~~表示正~~权重和负权重。

3.5.2 交叉熵损失。 我们将掩码语言建模（MLM）任务视为分类问题，并采用交叉熵损失来解决。其表达式为：$L_{CE} = \frac{1}{B}\sum_{i=1}^{B} \log s_{i,I_i}$，其中 I_i 表示掩码词元的索引。

3.5.3 时间注意力引导损失。 w_a 反映每一帧的重要性得分，真实标注范围内的时刻应具有更高的得分。我们采用 TAG 损失 [38]，引导注意力池化操作学习均衡的得分。其表达式为：$L_{TAG} = -\sum_{i=1}^{B} \frac{\sum_{j=1}^{T} w_a^{i,j} \log w_a^{i,j}}{\sum_{j=1}^{T} w_a^{i,j}}$，其中$\hat{w}^{i,j}_a$是一个二元得分，用于衡量第 j 个样本中的第 i 个时刻是否位于真实标注范围内。

3.5.4 二元交叉熵损失。 与之前基于片段的方法相同 [46]，我们采用缩放后的交并比得分 y 作为标签，通过二元交叉熵损失训练 CLP。

$$y_i = \begin{cases} 0 & o_i \le o_{min} \\ \dfrac{o_i - o_{min}}{o_{max} - o_{min}} & Otherwise \\ 1 & o_i \ge o_{max} \end{cases} \tag{24}$$

其中，o_i表示第 i 个片段候选项的原始交并比得分。o_{max}、o_{min}是两个阈值。因此，二元交叉熵损失可表示为：$L_{BCE} = \frac{1}{N_c}\sum_{i=1}^{N_c} y_i \log p_i + (1-y_i)\log(1-p_i)$，其中 $N_c = B\cdot T\cdot T$表示一个批次中的候选项总数。 p表示片段级得分图 M_c中的预测得分。

4 实验

4.1 数据集

Charades-STA [28]。 该数据集包含 6,672 段室内视频，时长介于 2.4 至 194.4 秒之间，共包含 16,128 个时刻片段，其时间戳均经过精确标注。未剪辑视频的平均时长不到 30 秒，而标注时刻片段的平均时长为 8.2 秒。因此，这是一个包含短视频的小型数据集。为公平比较，我们采用与 [19, 46] 相同的数据划分。训练集包含 12,408 个时刻片段，测试集包含 3,720 个时刻片段。

ActivityNet Caption [12]。 该数据集包含 2 万个未剪辑视频，时长从 1.1 秒到 975.0 秒不等，跨度很大。每个视频平均标注有 3.65 个查询，每个查询平均由 13.48 个词组成。官方划分按照 2:1:1 的比例进行划分。训练集、val_1 split 和 val_2 split 中分别有 37,421、17,505 和 15,031 个时刻片段。按照 [46]，我们将 val_2 split 视为测试集。总体而言，这是一个规模更大、视频更长的数据集。

4.2 实现细节

4.2.1 评估指标。 我们采用指标“R@*n*, IoU=*m*”和平均交并比来报告评估结果。我们设置 $n = 1$ 和 $m \in \{0。1，\ 0。3，\ 0。5，\ 0。7\}$。一些方法 [46] 还会报告 $n = 5$的结果，而该结果在推理过程中会显著受到非极大值抑制（NMS）阈值 [21] 的影响。较大的“R@5, IoU=0.5”可能意味着较小的“R@5, IoU=0.7”。在后处理的超参数设置不同的情况下进行比较并不公平。因此，我们重点关注 R@1 和平均交并比。

4.2.2 实现细节。 在 Charades-STA 上，我们使用 VGG、C3D 和 I3D 作为预训练三维卷积神经网络进行实验，以便进行公平比较。对于 AvtivityNet captions，我们采用 C3D 和 I3D 特征。Charades-STA 和 AvtivityNet captions 的片段数量 T和维度 d 均设为 64 和 512。所有实验均在搭载 Pytorch [23] 实现的 NVIDIA V100 工作站上进行。优化器为 Adam [11]，衰减因子为 0.5。学习率设为 1×10^{-4}，批量大小 B 固定为 32。Charades-STA 和 ActivityNet Captions 的阈值 o_{max} 和 o_{min} 分别设为 0.5 和 1.0。推理时，我们报告时刻级得分图 S_m、片段级得分图 S_c和融合得分图 $S_f = S_m \odot S_c$的结果。 n_l和 n_g设为 3 和 2，作为 [19]。第 3.4.2 节中的 $EarlyConv$由 8 层卷积组成，其输入通道、输出通道、卷积核大小、步幅、填充和膨胀分别为：第一层 {512, 512, 5, 1, 16, 1 512, 512, 5, 1, 0, 1 7 }，最后几层为 { }。至于由 9 层卷积构成的 $LateConv$，其前 8 层设置与 $EarlyConv$相同。最后一层设置为 {512, 1, 1, 1, 0, 0}。如图 5 所示，由于片段级收敛速度快于时刻级，因此在经过 $upsilon$ 个 epoch 后，我们将 L_{BCE}的权重设为 0.001。 $upsilon$在 Charades-STA 和 ActivityNet caption 上分别等于 9 和 7。

4.3 与以往最先进水平的比较

如表1和表2所示，我们将 GMU 各层级的性能与以往最先进水平进行比较。在两个基准数据集上，对于每种三维卷积神经网络特征，所提出的 GMU 均达到了最先进水平。

在 Charades-STA 上，采用 VGG特征的 GMU 融合图（GMU-F）以交并比=0.5 提升了 0.78%，以交并比=0.7 提升了 0.52%，超过了所有以往的最先进方法。此外，单独的时刻级和片段级结果也显著优于基线方法。与 LGI 相比，GMU-M 在交并比=0.5 上提升了 2.77%；与 2D-TAN 相比，GMU-C 的提升则达到 4.46%。采用 C3D特征时，交并比=0.7 和平均交并比上的提升分别扩大至 3.33% 和 2.52%。对于 I3D特征，GMU-F 在交并比=0.5、交并比=0.7 和平均交并比上均达到最先进水平。GMU-C 在交并比=0.5 上显著优于 2D-TAN，提升了 7.15%；在交并比=0.7 上则提升了 4.14%。在该基准数据集上，融合两个层级后，所有指标的性能均得到提升。

对于采用 C3D特征的 ActivityNetCaption，所提出的 GMU 在 IoU=0.5 和平均交并比方面达到了最先进水平。单独的片段级映射图表现出色，在三个指标上均取得第二高的成绩。它在 IoU=0.5 和平均交并比分别超过 2D-TAN 2.53% 和 2.44%。在 IoU=0.3 和 IoU=0.5 下，GMU-M 的表现也分别比 LGI 高出 4.48% 和 2.05%。采用 I3D特征时，GMU-F 在 IoU=0.5、0.7 和平均交并比方面达到最先进水平，而 GMU-C 位居第二。时刻级方法在 IoU=0.3 下仍表现最佳；与片段级得分图融合后，其表现降低了 0.11%。

表1：与以往方法在 Charades-STA 上的比较。请注意，LGI 和 2D-TAN 分别是 GMU 的时刻级和片段级基线。GMU-M、-C、-F 分别表示由时刻级、片段级和融合得分图得出的结果。“M”“C”和“F”分别代表时刻、片段和融合层级。

三维卷积神经网络层级		方法	R@1，交并比= 0.3	0.5	0.7	mIoU
VGG	M	LGI [19]	-	41.72	21.48	-
		CBLN [15]	-	43.67	24.44	-
		CPN [48]	**64.41**	46.08	25.06	43.90
		SSCS [2]	-	43.15	25.54	-
		GMU-M（我们的方法）	61.72	44.49	24.57	42.12
	C	2D-TAN [46]	-	39.70	23.31	-
		FVMR [7]	-	42.36	24.14	-
		MS-2D-TAN [45]	-	45.65	27.20	-
		GMU-C（本文方法）	60.75	46.18	26.16	41.83
	F	APGN [14]	-	44.23	25.64	-
		DPIN [31]	-	47.98	26.96	-
		GMU-F（我们的方法）	**64.41**	**48.76**	**27.72**	**44.22**
C3D	M	CTRL [6]	-	23.63	8.89	-
		CBP [32]	-	36.80	18.87	35.74
		GMU-M（我们的方法）	56.61	39.33	19.01	37.69
	C	FVMR [7]	-	38.16	18.22	-
		MS-2D-TAN [45]	-	41.10	**23.25**	-
		GMU-C（我们的方法）	55.67	41.85	22.63	38.16
	F	BPNet [34]	55.46	38.25	20.51	38.03
		LPNet [33]	**59.14**	40.94	21.13	39.67
		GMU-F（我们的方法）	58.28	**43.39**	**23.17**	**39.74**
I3D	M	LGI [19]	**72.96**	59.46	35.48	51.28
		SSCS [2]	-	60.75	36.19	-
		GMU-M（我们的方法）	72.02	56.51	32.85	50.19
	C	2D-TAN [46]	-	50.62	28.71	-
		MS-2D-TAN [45]	-	60.08	37.39	-
		GMU-C（我们的方法）	69.84	57.77	35.97	50.06
	F	BPNet [34]	65.48	50.75	31.64	46.34
		LPNet [33]	66.58	54.33	34.03	47.71
		GMU-F（我们的方法）	72.58	**61.18**	**37.45**	**52.18**

所提出的训练统一策略显著提升了所有层级的性能。纯时刻级和片段级方法均优于各自对应的基线方法。从结果的分布分析来看，在 IoU=0.3 时，时刻级方法始终优于片段级方法；而在 IoU=0.5 和 IoU=0.7 时则相反。时刻级方法的优势在于从上至下的视角进行整体判断，因此其在粗粒度定位上的召回率表现更好。片段级方法则从自底向上的视角进行预测，在细粒度定位上表现更优。将两者统一起来，能够兼具各自优势，并在各个精确度层级上取得平衡，从而持续提升平均交并比。

此外，我们还将 GMU 与其他基于多层级融合的方法进行了比较，结果依然具有竞争力。在所展示的

表2：与先前方法在ActivityNet描述上的比较。

三维卷积神经网络层级		方法	R@1，交并比= 0.3	0.5	0.7	mIoU
C3D	M	DRN [40]	-	45.45	24.39	-
		LGI [19]	58.52	41.51	23.07	41.13
		SSCS [2]	61.35	46.67	27.56	-
		IVG-DCL [20]	63.22	43.84	27.10	44.21
		GMU-M（我们的方法）	63.00	43.56	24.67	44.32
	C	2D-TAN [46]	59.45	44.51	26.54	-
		CMIN [47]	63.61	43.40	23.88	-
		FVMR [7]	60.63	45.00	26.85	-
		MS-2D-TAN [45]	61.04	46.16	**29.21**	-
		GMU-C（我们的）	61.58	47.04	28.98	45.11
	F	ReLoCLNet [42]	42.65	28.54	17.76	-
		BPNet [34]	58.98	42.07	24.69	42.11
		LPNet [33]	**64.29**	45.92	25.39	44.72
		GMU-F（我们的方法）	62.36	**47.08**	28.61	**45.48**
I3D	M	ExCL [9]	62.30	42.70	24.10	-
		VSLNet [43]	63.16	43.22	26.16	-
		GMU-M（我们的方法）	**63.17**	43.53	26.80	44.17
	C	TMLGA [26]	51.28	33.04	19.26	-
		MS-2D-TAN [45]	62.09	45.50	28.28	-
		GMU-C（本文方法）	62.20	46.11	28.37	45.01
	F	GMU-F（本文方法）	63.06	**46.48**	**28.70**	**45.39**

五项实验中，GMU 的优势主要体现在两方面：1. 它采用了更独立、有效且专业化的多层级分离预测器，并结合掩码语言模型任务和密集采样。2. 我们采用了“分别训练、联合预测”的策略。

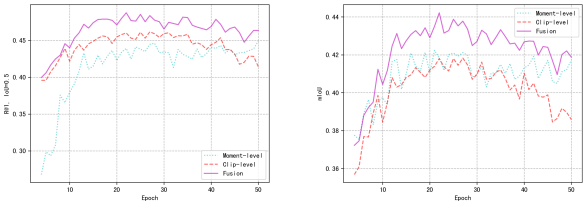
我们还在 TACOS [25] 上进行了实验，以验证 GMU 的泛化能力。请参见附录中的表1。

5 消融研究

消融研究的结果如图5、表3、表4和表5所示。

多层级统一策略。图5考察了各融合图的性能随训练轮数变化的情况。融合图方法始终优于单独的时刻级或片段级方法，这表明多层级统一策略是有效的。更重要的是，表3中第7行与第8行、第9行与第19行、第10行与第20行、第22行与第25行以及第23行与第26行的比较表明，共享编码器（SE）提升了各层级的性能。片段级性能大幅提升了约3%–5%，这说明统一两个层级的优势还体现在能够学习到更好的基础编码器，包括视觉基础模块和语义编码器。此外，图5(a)和图5(b)中的片段级性能在训练若干轮后有所下降，而时刻级性能则保持波动，并围绕稳定值变化。因此，有必要在经过 v 个训练轮数后降低 L_{BCE} 的权重。表3中的第27至30行探讨了其影响。

生成引导策略和超参数。如表3所示，对比第1–3行与第5–7行可以证明



(a) 各映射的平均交并比随训练轮次的变化。(b) 各映射的“R@1, IoU=0.5”随训练轮次的变化。

图5：在 Charades-STA 测试集上，生成的映射的性能随训练轮次的变化。

生成引导策略在不同配置下提升了时刻层级的性能。根据第23行、第24行、第26行和第27行，融合模型也很大程度上依赖该策略来进一步提升性能。该新颖框架的创新之处在于用生成特征替代原始语义特征，从而缩小查询与图库之间的差距。此外，来自原始视觉特征的捷径连接增强了对视觉信息的记忆。我们还测试了预分割时刻数量的影响 T；根据第1–7行和第11–14行，较大的 T 对两个层级都有显著益处。至于采样数量N，参照第14–18行，如果$N > 4$，较大的 N 带来的收益较小。为兼顾性能与时间成本，我们在训练期间将N 设为4。

损失函数。参见表4，各损失项都发挥着至关重要的作用。对于时刻级部分，仅使用边界定位损失 L_{BL} 的模型表现不佳。精心设计的辅助掩码语言模型任务 L_{BL} 使模型能够保留语义信息，并使“R@1, IoU=0.5”提高了3.11%。时间注意力引导损失 L_{TAG} 引导 *NLBlock* 在时刻级部分的注意力聚焦于真实标注中的时刻片段，使“R@1, IoU=0.5”和平均交并比分别提升了8.04%和3.64%。L_{BL}、L_{BL}和L_{TAG}的组合取得了最佳的时刻级性能。此外，包含四种损失项的融合模型优于其他所有组合。与两个独立层级相比，该模型表现出全面优势。

CLP 的组成部分。第4.3节中的结果表明，CLP 相较于片段级基线 2D-TAN 有显著提升。主要改动包括 DPGM、GG 和 SE，参见表5。2D-TAN 会聚合候选片段中的所有帧，并进行稀疏预测。我们对固定数量的帧进行采样，并进行密集预测。仅采用 DPGM 时，CLP 的表现就优于 2D-TAN，这与表3中第14–18行的分析一致。采样更多帧或全部帧并不会带来提升，而对所有候选片段进行密集预测则会带来提升。此外，若采用共享编码器（SE）和生成引导（GG）策略，单独的 2D-TAN 和所提出的 CLP 均能得到提升，这表明它们具有优越性。

6 进一步分析

为直观地反映多层级统一策略的影响，我们将生成的得分图和真实标注 IoU 图可视化，如

表3：相关策略和超参数的消融研究。SE 表示共享编码器的策略，包括视觉基础模块和语义编码器。GG 表示生成引导策略。 κ_e 和 κ_l 表示 *EarlyConv* 和 *LateConv* 的卷积层数量。 υ 表示过渡时期，在此之后我们将 CLP 部分的损失权重设为 0.001。 N 表示 W 中的采样数量。

层级	Row	策略 SE	策略 GG	超参数 κ_e	κ_l	υ	T	N	R@1 交并比=0.5	mIoU
时刻	1	✗	✗	-	-	-	16	-	38.01	36.76
	2	✗	✗	-	-	-	32	-	39.27	37.11
	3	✗	✗	-	-	-	64	-	39.81	37.56
	4	✗	✓	-	-	-	8	-	36.83	36.52
	5	✗	✓	-	-	-	16	-	38.92	37.02
	6	✗	✓	-	-	-	32	-	40.01	37.62
	7	✗	✓	-	-	-	64	-	40.23	37.87
	8	✓	✓	-	-	-	64	-	**44.49**	**42.12**
Clip	9	✗	✗	-	9	-	64	4	42.71	39.21
	10	✗	✗	-	17	-	64	4	42.60	38.98
	11	✗	✓	8	9	-	8	4	40.01	37.28
	12	✗	✓	8	9	-	16	4	41.32	38.40
	13	✗	✓	8	9	-	32	4	41.76	38.88
	14	✗	✓	8	9	-	64	4	42.53	39.07
	15	✗	✓	8	9	-	64	2	41.39	38.41
	16	✗	✓	8	9	-	64	8	42.49	39.12
	17	✗	✓	8	9	-	64	16	42.58	39.16
	18	✗	✓	8	9	-	64	32	42.74	39.47
	19	✓	✗	-	9	-	64	4	45.88	41.63
	20	✓	✗	-	17	-	64	4	45.52	41.59
	21	✓	✓	8	9	-	64	4	**46.18**	**41.83**
融合	22	✗	✗	-	9	10	64	4	43.82	40.79
	23	✗	✗	-	17	10	64	4	44.09	40.92
	24	✗	✓	8	9	10	64	4	44.81	41.58
	25	✓	✗	-	9	10	64	4	47.69	43.74
	26	✓	✗	-	17	10	64	4	47.54	43.26
	27	✓	✓	8	9	10	64	4	**48.76**	**44.22**
	28	✓	✓	8	9	9	64	4	48.23	43.92
	29	✓	✓	8	9	11	64	4	48.20	43.73
	30	✓	✓	8	9	✗	64	4	47.34	43.03

图6所示。前三列分别为时刻级、片段级和融合得分图，最后一列为真实标注IoU图。观察结果可得出以下结论：1）时刻级得分图中的高分点往往呈矩形分布，这是因为其由双仿射乘法计算得出，导致模型对唯一真实标注的判断模糊；2）片段级得分图中的高分分布倾向于聚集在对角线附近的几个簇中。如果排名第一的点落入错误的簇中，结果就会完全错误。 [36] 中探讨了这一相关现象。这是因为 CLP 直接从

表4：损失函数的消融研究。

方法	损失项 $\mathcal{L}_{BL}$	$\mathcal{L}_{CE}$	$\mathcal{L}_{TAG}$	$\mathcal{L}_{BCE}$	R@1 交并比=0.5	mIoU
时刻-层级	✓				31.80	34.06
	✓	✓			34.92	35.33
	✓		✓		39.84	37.70
	✓	✓	✓		40.23	37.87
片段级				✓	42.53	39.07
融合	✓			✓	44.67	41.78
	✓		✓	✓	47.01	42.89
	✓	✓	✓	✓	48.76	44.22

表5：片段级配置的消融研究。SC 和 MP 分别表示堆叠卷积和最大池化，它们被 2D-TAN 用作视觉聚合策略。GG 和 SE 的说明见表3。

方法	组件 SC	MP	DPGM	GG	SE	R@1，交并比= 0.5	0.7
2D-TAN（基线）	✓					39.70	23.31
		✓				39.81	23.25
	✓			✓	✓	44.21	23.44
		✓		✓	✓	44.73	23.69
片段级（我们的方法）			✓			42.71	22.36
			✓	✓		42.53	22.42
			✓		✓	45.52	25.77
			✓	✓	✓	46.18	26.16

基准数据集的整体真实标注分布，其中真实标注主要分布在对角线附近。3）真实标注 IoU 图总是向真实标注的右上方扩散。

单独来看，时刻级方法难以准确判断可能的边界；片段级方法则难以召回排名第一的点。因此，片段级方法的"R@1, IoU=0.3"低于时刻级方法。如第三列所示，融合得分图充分利用了两者的优点并避免了其缺点。时刻级方法中的冗余点和片段级方法中的错误簇都被显著消除，因此融合方法表现最佳，并在各项指标之间取得平衡。附录中的图1展示了更详细的结果。

7 结论

本文提出了一种名为 GMU 的新型网络，以有效解决视频定位任务。它统一了时刻级和片段级预测器，从而学习更稳健的视觉编码器和语义编码器。融合图结合了两者的优势，在低精确度和高精确度的视频定位要求下均达到最先进性能。此外，我们率先将生成引导范式引入视频定位。生成的隐式特征被用于替代原始语义特征，从而缩小不同模态之间的差距。捷径连接还保留了长期记忆，以便更好地保留

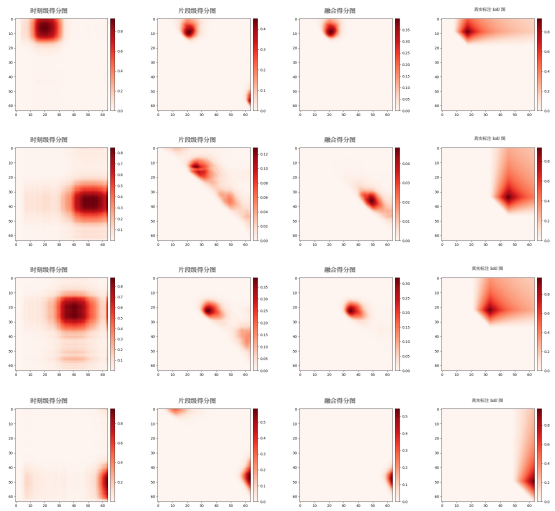


图 6：得分图和真实标注交并比图的可视化。它们取自 Charades-STA 测试集。对应的视频名称为：“YVKIV”、“ZHRPD”、“E6DLK”和“FPJ9D”。

视觉信息。未来，深入融合这两个层级并探索更灵活的生成结构，前景可期。

参考文献

[1] Joao Carreira 和 Andrew Zisserman。2017。动作识别何去何从？一种新模型与 Kinetics 数据集。载于 IEEE 计算机视觉与模式识别会议论文集。6299–6308。[2] Xinpeng Ding、Nannan Wang、Shiwei Zhang、De Cheng、Xiaomeng Li、Ziyuan Huang、Mingqian Tang 和 Xinbo Gao。2021。基于支持集的跨监督视频定位。载于 IEEE/CVF 国际计算机视觉大会论文集。11573–11582。[3] Jeffrey Donahue、Lisa Anne Hendricks、Sergio Guadarrama、Marcus Rohrbach、Subhashini Venugopalan、Kate Saenko 和 Trevor Darrell。2015。用于视觉识别与描述的长期循环卷积网络。载于 IEEE 计算机视觉与模式识别会议论文集。2625–2634。[4] T. Dozat 和 C. D. Manning。2016。用于神经依存句法分析的深度双仿射注意力。（2016）。[5] Ali Farhadi、Mohsen Hejrati、Mohammad Amin Sadeghi、Peter Young、Cyrus Rashtchian、Julia Hockenmaier 和 David Forsyth。2010。每张图片都在讲述一个故事：从图像生成句子。载于欧洲计算机视觉会议。Springer，15–29。[6] Jiyang Gao、Chen Sun、Zhenheng Yang 和 Ram Nevatia。2017。TALL：通过语言查询进行时间活动定位。载于 IEEE/CVF 国际计算机视觉大会论文集。5267–5275。[7] Junyu Gao 和 Changsheng Xu。2021。快速视频片段检索。载于 IEEE/CVF 国际计算机视觉大会论文集。1523–1532。[8] Lianli Gao、Zhao Guo、Hanwang Zhang、Xing Xu 和 Heng Tao Shen。2017。结合基于注意力的 LSTM 与语义一致性的视频描述生成。IEEE 多媒体汇刊 19，9（2017），2045–2055。[9] Soham Ghosh、Anuva Agarwal、Zarana Parekh 和 Alexander Hauptmann。2019。ExCL：使用自然语言描述进行抽取式片段定位。arXiv 预印本 arXiv:1904.02755（2019）。[10] Andrej Karpathy、George Toderici、Sanketh Shetty、Thomas Leung、Rahul Suk- thankar 和李飞飞。2014。基于卷积神经网络的大规模视频分类。载于 IEEE 计算机视觉与模式识别会议论文集。1725–1732。[11] Diederik P Kingma 和 Jimmy Ba。2014。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412.6980（2014）。[12] Ranjay Krishna、Kenji Hata、Frederic Ren、李飞飞和 Juan Carlos Niebles。2017。视频中的事件密集描述。载于 IEEE/CVF 国际计算机视觉大会论文集。706–715。

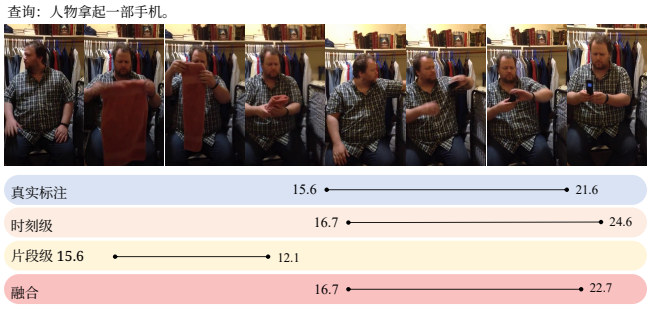
[13] Tianwei Lin、Xiao Liu、Xin Li、Errui Ding 和 Shilei Wen。2019。BMN：用于时序动作提议生成的边界匹配网络。载于 IEEE/CVF 国际计算机视觉大会论文集。3889–3898。[14] Daizong Liu、Xiaoye Qu、Jianfeng Dong 和 Pan Zhou。2021。用于视频时间句子定位的自适应提议生成网络。arXiv 预印本 arXiv:2109.06398（2021）。[15] Daizong Liu、Xiaoye Qu、Jianfeng Dong、Pan Zhou、Yu Cheng、Wei Wei、Zichuan Xu 和 Yulai Xie。2021。用于时序句子定位的上下文感知双仿射定位网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集。11235–11244。[16] Chujie Lu、Long Chen、Chilie Tan、Xiaolin Li 和 Jun Xiao。2019。DEBUG：一种用于自然语言视频定位的密集自底向上定位方法。载于《2019年自然语言处理实证方法会议与第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集》。5144–5153。[17] Huaishao Luo、Lei Ji、Ming Zhong、Yang Chen、Wen Lei、Nan Duan 和 Tianrui Li。2021。Clip4clip：关于 CLIP 端到端视频片段检索的实证研究。arXiv 预印本 arXiv:2104.08860（2021）。[18] Elman Mansimov、Emilio Parisotto、Jimmy Lei Ba 和 Ruslan Salakhutdinov。2015。利用注意力根据描述生成图像。arXiv 预印本 arXiv:1511.02793（2015）。[19] Jonghwan Mun、Minsu Cho 和 Bohyung Han。2020。用于时序定位的局部-全局视频文本交互。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10810–10819。[20] Guoshun Nan、Rui Qiao、Yao Xiao、Jun Liu、Sicong Leng、Hao Zhang 和 Wei Lu。2021。基于双重对比学习的干预式视频定位。载于 IEEE/CVF计算机视觉与模式识别会议论文集。2765–2775。[21] Alexander Neubeck 和 Luc Van Gool。2006。高效的非极大值抑制。载于第18届国际模式识别会议（ICPR’06），第3卷。IEEE，850–855。[22] Yingwei Pan、Ting Yao、Houqiang Li 和 Tao Mei。2017。利用迁移语义属性进行视频描述生成。载于 IEEE计算机视觉与模式识别会议论文集。6504–6512。[23] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。2019。PyTorch：一种命令式风格的高性能深度学习库。《神经信息处理系统进展》32（2019），8026–8037。[24] Scott Reed、Zeynep Akata、Xinchen Yan、Lajanugen Logeswaran、Bernt Schiele 和 Honglak Lee。2016。生成对抗式文本到图像合成。载于国际机器学习会议。PMLR，1060–1069。[25] Michaela Regneri、Marcus Rohrbach、Dominikus Wetzel、Stefan Thater、Bernt Schiele 和 Manfred Pinkal。2013。在视频中定位动作描述。《计算语言学协会会刊》1（2013），25–36。https://doi.org/10 1162 _ _00207 。/tacl a Cristian Rodriguez、Edison [26] Marrese-Taylor、Fatemeh Sadat Saleh、Hongdong Li 和 Stephen Gould。2020。利用引导注意力对视频中的自然语言查询进行无提议时序片段定位。载于 IEEE/CVF冬季计算机视觉应用会议论文集。2464–2473。[27] R. Rubinstein。1999。组合与连续优化的交叉熵方法。《应用概率方法论与计算》1，2（1999），127–190。[28] Gunnar A Sigurdsson、Gül Varol、Xiaolong Wang、Ali Farhadi、Ivan Laptev 和 Abhinav Gupta。2016。家中的好莱坞：通过众包收集活动理解数据。载于欧洲计算机视觉会议。Springer，510–526。[29] Karen Simonyan 和 Andrew Zisserman。2014。用于大规模图像识别的超深卷积网络。arXiv 预印本 arXiv:1409.1556（2014）。[30] Du Tran、Lubomir Bourdev、Rob Fergus、Lorenzo Torresani 和 Manohar Paluri。2015。利用三维卷积网络学习时空特征。载于 IEEE国际计算机视觉会议论文集。4489–4497。[31] Hao Wang、Zheng-Jun Zha、Xuejin Chen、Zhiwei Xiong 和 Jiebo Luo。2020。用于视频片段定位的双路径交互网络。计算机协会，美国纽约州纽约市，4116–4124。https://doi.org/10.1145/3394171.3413975[32] Jingwen Wang、Lin Ma 和 Wenhao Jiang。2020。通过上下文边界感知预测对视频中的语言查询进行时序定位。载于 AAAI人工智能会议论文集，第34卷。12168–12175。[33] Shaoning Xiao、Long Chen、Jian Shao、Yueting Zhuang 和 Jun Xiao。2021。利用可学习片段提议进行自然语言视频定位。arXiv 预印本 arXiv:2109.10678（2021）。[34] Shaoning Xiao、Long Chen、Songyang Zhang、Wei Ji、Jian Shao、Lu Ye 和 Jun Xiao。2021。用于两阶段自然语言视频定位的边界提议网络。载于 AAAI人工智能会议论文集，第35卷。2986–2994。[35] Quanzeng You、Hailin Jin、Zhaowen Wang、Chen Fang 和 Jiebo Luo。2016。利用语义注意力进行图像描述生成。载于 IEEE计算机视觉与模式识别会议论文集。

计算机视觉与模式识别。4651–4659。[36] Yitian Yuan、Xiaohan Lan、Xin Wang、Long Chen、Zhi Wang 和 Wenwu Zhu。2021。深入探究视频中的时序句子定位。第二届以人为中心的多媒体分析国际研讨会论文集（11月2021 10 1145 3475723 3484247）。https://doi.org/。/。[37] Yitian Yuan、Lin Ma、Jingwen Wang、Wei Liu 和 Wenwu Zhu。2019。用于视频时序句子定位的语义条件动态调制。arXiv 预印本 arXiv:1910.14303（2019）。[38] Yitian Yuan、Tao Mei 和 Wenwu Zhu。2019。寻找你说话的位置：利用基于注意力的位置回归进行视频时间句子定位。载于 AAAI人工智能会议论文集，第33卷。9159–9166。[39] Joe Yue-Hei Ng、Matthew Hausknecht、Sudheendra Vijayanarasimhan、Oriol Vinyals、Rajat Monga 和 George Toderici。2015。超越短片段：用于视频分类的深度网络。载于 IEEE计算机视觉与模式识别会议论文集。4694–4702。[40] Runhao Zeng、Haoming Xu、Wenbing Huang、Peihao Chen、Mingkui Tan 和 Chuang Gan。2020。用于视频定位的稠密回归网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集。10287–10296。[41] Da Zhang、Xiyang Dai、Xin Wang、Yuan-Fang Wang 和 Larry S Davis。2019。Man：通过迭代图调整实现自然语言片段检索的片段对齐网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集。1247–1257。

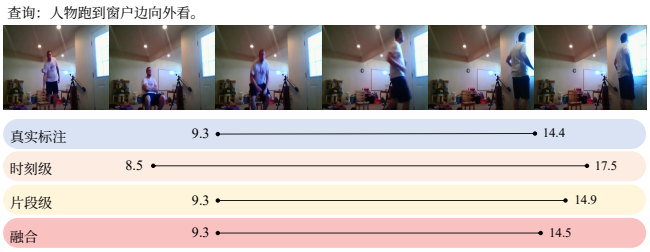
[42] Hao Zhang、Aixin Sun、Wei Jing、Guoshun Nan、Liangli Zhen、Joey Tianyi Zhou 和 Rick Siow Mong Goh。2021。《基于对比学习的视频语料库片段检索》。arXiv 预印本 arXiv:2105.06247（2021）。[43] Hao Zhang、Aixin Sun、Wei Jing 和 Joey Tianyi Zhou。2020。《用于自然语言视频定位的基于跨度的定位网络》。arXiv 预印本 arXiv:2004.13931（2020）。[44] Hong Jiang Zhang、Jianhua Wu、Di Zhong 和 Stephen W Smoliar。1997。《用于基于内容的视频检索与浏览的集成系统》。模式识别 30，4（1997），643–658。[45] Songyang Zhang、Houwen Peng、Jianlong Fu、Yijuan Lu 和 Jiebo Luo。2021。《用于自然语言片段定位的多尺度二维时间邻接网络》。IEEE模式分析与机器智能汇刊（2021）。[46] Songyang Zhang、Houwen Peng、Jianlong Fu 和 Jiebo Luo。2020。《学习用于自然语言片段定位的二维时间邻接网络》。载于《AAAI人工智能会议论文集》，第34卷。12870–12877。[47] Zhu Zhang、Zhijie Lin、Zhou Zhao 和 Zhenxin Xiao。2019。《用于视频中基于查询的片段检索的跨模态交互网络》。载于《第42届国际ACM SIGIR信息检索研究与发展会议论文集》。655–664。[48] Yang Zhao、Zhou Zhao、Zhu Zhang 和 Zhijie Lin。2021。《通过线段树实现时间视频定位的级联预测网络》。载于《IEEE/CVF计算机视觉与模式识别会议论文集》。4197–4206。

arXiv:2303.07748v1 [cs.CV] 14 Mar 2023

附录



（a）视频名称：ZHRPD。



(b) 视频名称：E6DLK

图 1：使用 I3D 特征在 Charades-STA 上的预测示例。

ACM 引用格式：
。2022。附录。载于论文集 请从版权确认邮件中填写正确的会议名称（会议简称“*XX*”）。ACM，美国纽约州纽约市，1页。
https://doi.org/XXXXXXX.XXXXXXX

1 TACOS 上的实验

1.1 数据集

TACoS。该数据集包含127个视频和17,344个在厨房中拍摄的标注片段。我们按照50%、25%和25%的比例划分训练集、验证集和测试集。

1.2 比较

如表1所示，GMU在TACoS的所有指标上均取得了优异的性能。在交并比=0.1和交并比=0.3指标上，GMU分别比MS-2D-TAN高出2.89%和1.08%，进一步证明了GMU采用多层级统一策略的稳健性。此外，与以往的多层级方法相比，GMU的性能也更为突出。相关说明见正文。

表1：与此前方法在 TACoS 上的比较

三维卷积神经网络	层级	方法	R@1，交并比= 0.1	0.3	0.5
C3D	M	CBP	-	27.31	24.79
		DRN	-	-	23.17
		CMIN	32.48	24.64	18.05
		IVG-DCL	49.36	38.84	29.07
		CBLN	49.16	38.98	27.65
		GMU-M（我们的方法）	53.13	41.27	28.75
	C	2D-TAN	47.59	37.29	25.32
		FVMR	53.12	41.48	29.12
		MS-2D-TAN	52.39	45.61	35.77
		GMU-C（我们的方法）	52.77	44.42	34.69
	F	BPNet	-	25.96	20.96
		APGN	-	40.47	27.86
		GMU-F（我们的方法）	**55.28**	**46.69**	**35.89**

2 预测示例

允许为个人或课堂用途免费制作本作品全部或部分内容的数字或纸质副本，前提是这些副本不得为营利或商业利益而制作或分发，并且副本须注明本声明及首页上的完整引文。对于本作品中不属于 ACM 所有的组件，必须尊重其版权。注明出处后，可以摘录。若要以其他方式复制、重新发布、上传至服务器或分发到列表，则须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。会议简称“*XX*”，*2018*年*6*月*3*日至*5*日，纽约州伍德斯托克© 2022 计算机协会。ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00 https://doi.org/XXXXXXX.XXXXXXX