E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)03-0842-13

论文引用格式:Ren X L, Zhang F F, Zhou W T and Zhou L. 2025. Complementary multi-type prompts for weakly-supervised temporal action location. Journal of Image and Graphics, 30(3):0842-0854(任小龙, 张飞飞, 周琬婷, 周玲. 2025. 多类型提示互补的弱监督时序动作定位. 中国图象图形学报, 30(3):0842-0854)[DOI:10.11834/jig.240354]

多类型提示互补的弱监督时序动作定位

任小龙[1],张飞飞[1*],周琬婷[2],周玲[3]

1. 天津理工大学计算机科学与工程学院,天津　300380; 2. 北京邮电大学人工智能学院,北京　100876;
3. 澳门科技大学计算机科学与工程学院,澳门　999078

摘　要:**目的**　弱监督时序动作定位仅利用视频级标注来定位动作实例的起止时间并识别其类别。目前基于视觉语言的方法利用文本提示信息来提升时序动作定位模型的性能。在视觉语言模型中,动作标签文本通常被封装为文本提示信息,按类型可分为手工类型提示(handcrafted prompts)和可学习类型提示(learnable prompts),而现有方法忽略了二者间的互补性,使得引入的文本提示信息无法充分发挥其引导作用。为此,提出一种多类型提示互补的弱监督时序动作定位模型(multi-type prompts complementary model for weakly-supervised temporal action location)。**方法**　首先,设计提示交互模块,针对不同类型的文本提示信息分别与视频进行交互,并通过注意力加权,从而获得不同尺度的特征信息;其次,为了实现文本与视频对应关系的建模,本文利用一种片段级对比损失来约束文本提示信息与动作片段之间的匹配;最后,设计阈值筛选模块,将多个分类激活序列(class activation sequence,CAS)中的得分进行筛选比较,以增强动作类别的区分性。**结果**　在3个具有代表性的数据集THUMOS14、ActivityNet1.2和ActivityNet1.3上与同类方法进行比较。本文方法在THUMOS14数据集中的平均精度均值(mean average precision,mAP)(0.1:0.7)取得39.1%,在ActivityNet1.2中mAP(0.5:0.95)取得27.3%,相比于P-MIL(proposal-based multiple instance learning)方法分别提升1.1%和1%。而在ActivityNet1.3数据集中mAP(0.5:0.95)取得了与对比工作相当的性能,平均mAP达到26.7%。**结论**　本文提出的时序动作定位模型,利用两种类型文本提示信息的互补性来引导模型定位,提出的阈值筛选模块可以最大化利用两种类型文本提示信息的优势,最大化其辅助作用,使定位的结果更加准确。

关键词:弱监督时序动作定位(WTAL);视觉语言模型;手工类型提示;可学习类型提示;分类激活序列(CAS)

Complementary multi-type prompts for weakly-supervised temporal action location

Ren Xiaolong[1], Zhang Feifei[1*], Zhou Wanting[2], Zhou Ling[3]

1. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin　300380, China;
2. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing　100876, China;
3. School of Computer Science and Engineering, Macao University of Science and Technology, Macao　999078, China

Abstract: **Objective** Weakly supervised temporal action localization uses only video-level annotations to locate the start

收稿日期:2024-06-26;**修回日期**:2024-09-04;**预印本日期**:2024-09-11

*通信作者:张飞飞　feifeizhang@email.tjut.edu.cn

基金项目:国家自然科学基金项目(62376196,62036012,U23A20387,62106262,62202331,62206200,62276118,62376037);天津市自然科学基金项目(24JCJQJC00190,22JCYBJC00030)

Supported by: National Natural Science Foundation of China(62376196, 62036012, U23A20387, 62106262, 62202331, 62206200, 62276118, 62376037); Natural Science Foundation of Tianjin, China(24JCJQJC00190, 22JCYBJC00030)

and end times of action instances and identify their categories. Only video-level annotations are available in weakly supervised environments; thus, directly designing a loss function for the task is impossible. Therefore, the existing work generally adopts the strategy of "localization by classification" and utilizes multi-example learning for training. However, this process has some limitations: 1) Localization and classification are two different tasks, revealing a notable gap between them; therefore, localization based on classification results may affect the final performance. 2) In weakly-supervised environments, fine-grained supervisory information to effectively distinguish between actions and backgrounds in videos is lacking, thereby posing a remarkable challenge for localization. Visual language models have recently received extensive attention. These models aim to model the correspondence between images and texts for more comprehensive visual perception. Specific textual prompts can improve the performance and robustness of the models to effectively apply large models to downstream tasks. Visual language-based approaches currently utilize auxiliary textual prompt information to compensate for supervisory information and improve the performance and robustness of temporal action localization models. In visual language models, action label text is regularly encapsulated as textual prompts, which can be categorized into Handcrafted Prompts and Learnable Prompts. Handcrafted Prompts comprise fixed templates and action labels (e. g., "a video of {class}"), which can learn a more generalized knowledge of the action class but lacks the specific knowledge of the relevant action. In contrast, Learnable Prompts comprise a set of learnable vectors, which can be adjusted and optimized during the training process. Therefore, the learnable type cues can learn more specific knowledge. The two types of text cues complement each other, improving the capability to discriminate different categories. However, the existing methods ignore the complementarity between the two, hindering the introduced text cues to maximize its guiding role and bringing certain noise information, which leads to inaccurate localization of action boundaries. Therefore, this paper proposes a multitype prompt complementary model for weakly-supervised temporal action location, which maximizes the guiding effect of textual cues to improve the accuracy of action localization. The methodology of this paper focuses on improving the accuracy of action location by maximizing the guidance of textual cues. **Method** First, this paper designs a prompt interaction module from the complementarity of textual prompts, which matches manual and learnable type prompts with action segments to obtain different similarity matrices. Then, the intrinsic connection between textual and segment features is mined through the attention layer and fused to realize the interaction between different types of textual prompts. Additionally, text cues must be effectively matched with action segments to play their guiding role. Therefore, this paper introduces the segment-level contrast loss, which is used to constrain the matching between cue messages and action segments. Finally, this paper designs a threshold filtering module to filter the class activation sequence (CAS) obtained from the guidance of different types of textual cue messages according to the threshold value. The final CAS is obtained by summing the CAS obtained after multilayer filtering according to a specific scale parameter, with the CAS obtained using only video-level features, covering the parts of each sequence with higher confidence scores, thus realizing the complementary advantages between different types of text cue information. **Result** Extensive experiments on three representative datasets, THUMOS14, ActivityNet1. 2, and ActivityNet1. 3, validate the effectiveness of the method proposed in this paper. Among them, different mAP(0. 1∶0. 5), mAP(0. 1∶0. 7), and mAP(0. 3∶0. 7) on THUMOS14 datasets achieved 58. 2%, 39. 1%, and 47. 9% average performance, respectively, which is comparable to P-MIL (proposal-based multiple instance learning) average performance by up to 1. 1%. In the ActivityNet1. 2 datasets, the method proposed in this paper achieves a performance of 27. 3% at mAP (0. 5∶0. 95), which is an average improvement of nearly 1% compared to the state-of-the-art method. The mAP (0. 5∶0. 95) of ActivityNet1. 3 datasets achieves comparable performance to the same work, with an average mAP of 26. 7%. In addition, numerous ablation experiments were conducted on the THUMOS14 datasets, and the experimental results proved the effectiveness of the modules. **Conclusion** This paper proposes a new weakly supervised temporal action localization model based on the complementarity of multiple types of cues, which alleviates the problem of inaccurate localization boundaries by leveraging the complementarity between different types of textual cues. The cue interaction module is designed in a targeted way to realize the interaction of different types of textual cue information. In addition, this paper introduces clip-level contrast loss to model the correspondence between text cue information and video, effectively constraining the matching between the introduced text cue information and action clips. Finally, this paper designs a multilayer nested threshold screening module, which can maximize the advantages of two different types of

textual cue information, avoid the interference of noisy information on the model, and maximize the auxiliary role of textual information. Extensive experiments on two challenging datasets demonstrate the effectiveness of the method proposed in this paper.

Key words: weakly supervised temporal action location (WTAL); visual language model; handcrafted prompts; learnable prompts; class activation sequence (CAS)

0 引　言

时序动作定位(temporal action location, TAL)是视频理解中的重要任务之一(Gaidon等, 2013),广泛用于视频监控、交通监管和动作检测等领域。目前大多数研究集中于全监督的环境设置,尽管全监督方法取得了显著效果,但是大规模的帧级注释限制了其在现实场景的应用。为了克服这一限制,弱监督时序动作定位(weakly-supervised temporal action location, WTAL)任务受到了研究者越来越多的关注。与全监督方法相比,弱监督时序动作定位任务仅需要视频级的数据标注,这很大程度上减少了标注成本,同时更贴合实际(Liu等, 2019)。尽管弱监督时序动作定位任务的研究已经取得一定进展,但这一任务仍面临重要的挑战:缺乏细粒度的监督信息。具体而言,弱监督环境设置下,缺少帧级注释,无法针对任务直接设计损失函数,因此现有工作普遍采用"以分类求定位"的策略(Ma等, 2021),利用多示例学习(multiple instance learning, MIL)(Liu等, 2021)进行训练。针对上述挑战,目前基于视觉语言的方法(Li等, 2023; Ju等, 2022),如图1(a)所示,直接引入文本提示信息作为辅助,利用文本与动作片段的交互得到分类激活序列(class activation sequence, CAS)用于后续定位。其中, Li等人(2023)引入手工类型提示和可学习类型提示,并利用掩码的思想将提示信息中的标签掩盖以迫使模型学习更全面的表征信息。Ju等人(2022)在手工类型提示的引导下利用知识蒸馏的方式将CLIP(contrastive language-image pre-training)(Radford等, 2021)中更具泛化的文本信息传递给模型以补充监督信息。但此类方法没有考虑不同类型文本提示信息间的互补性,并会引入一定的噪声信息,从而造成定位边界不准确的问题。如图1(b)所示,其中手工类型提示由固定模板和动作标签组成,如"*a video of* {*class*}",其能学习动作类别更为泛化的知识,但缺乏相关动

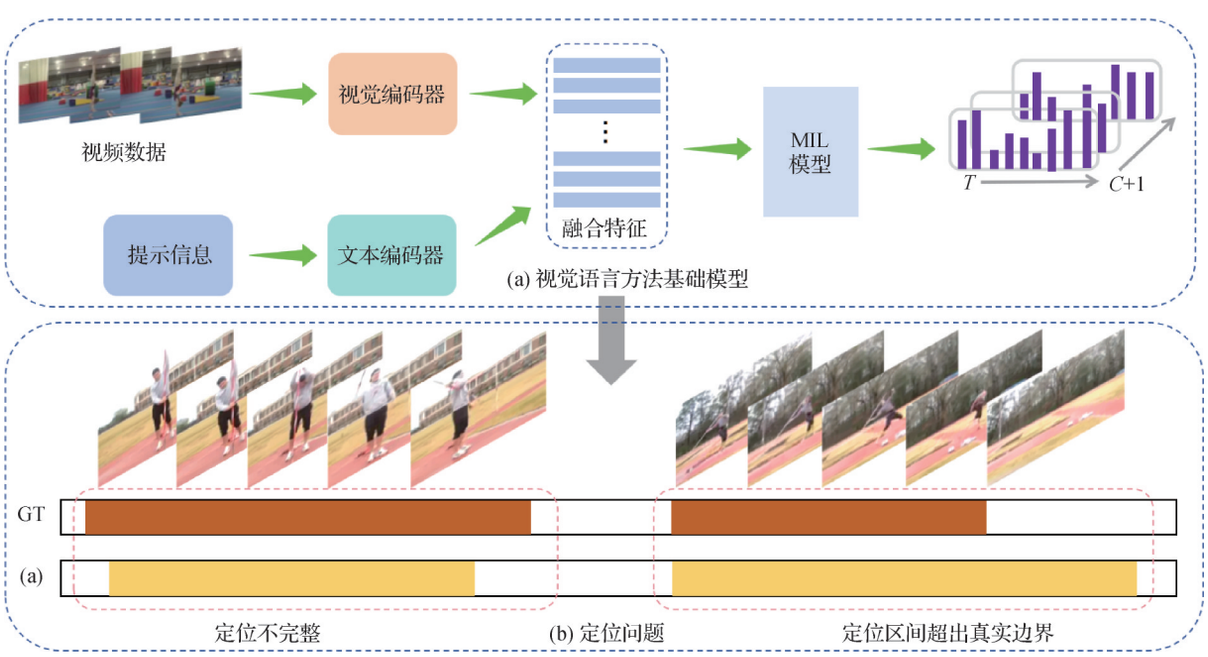

图1　视觉语言方法基础模型以及定位问题

Fig. 1　The base model of visual language approach and the location problem

((a) the base model of visual language approach; (b) the location problem)

作的具体知识，而可学习类型提示是由一组可学习的向量组成，可以在训练的过程中调整和优化。因此，可学习类型提示能够学到更为具体的知识。所以两种不同类型的文本提示信息可以互为补充，以提高对不同动作类别的判别能力，充分发挥其引导作用。

针对弱监督时序动作定位任务中的难点以及现有方法的不足，本文提出一种基于多类型提示互补的弱监督时序动作定位模型（multi-type prompts complementary model for weakly-supervised temporal action location）。首先，从文本提示信息间的互补性出发，设计了提示交互模块，分别将手工类型提示和可学习类型提示与动作片段进行匹配，以获取不同的相似矩阵，然后经过注意力层挖掘文本特征和片段特征的内在联系并进行融合。与此同时，文本提示信息需要更好地与动作片段匹配才能发挥其引导作用。为此，引入了片段级的对比损失，用于约束提示信息与动作片段的匹配。最后，设计阈值筛选模块，根据阈值筛选不同类型文本提示信息引导所得到的分类激活序列（CAS）。然后，将经过多层筛选后得到的CAS按照特定的比例参数与仅用视频级特征得到的CAS以加和的方式得到最终的CAS。其涵盖了各序列中具有较高置信分数的部分，从而实现不同类型文本提示信息之间的优势互补。

综上所述，本文工作总结如下：1）提出了一种新的基于多类型提示互补的弱监督时序动作定位模型，该模型借助不同类型文本提示信息间的互补性来缓解定位边界不准确的问题。通过针对性设计的提示交互模块，两种不同类型的文本提示信息可以有效地引导模型定位。2）引入片段级的对比损失，实现了文本提示信息和视频对应关系的建模。3）设计了多层嵌套的阈值筛选模块，该模块可以充分利用两种不同类型文本提示信息的优势，避免噪声信息对模型的干扰，最大化文本信息的辅助作用。4）在THUMOS14、ActivityNet1.2与ActivityNet1.3数据集上的大量实验结果表明，本文提出的框架具有优越的性能。

1 相关工作

1.1 全监督时序动作定位

现有的全监督方法可以大致分为两阶段和一阶段两类方法。两阶段方法（Xu 等，2020；Shou 等，2017）首先需要生成与动作类别无关的候选动作区间，然后对每个动作区间的动作进行分类并细化其边界。因此两阶段的方法往往具有较好的性能。一阶段的方法（Cheng 和 Bertasius，2022；Gao 等，2022）是直接对视频中的动作片段进行预测并生成与动作相关的动作区间。尽管全监督下的时序动作定位任务取得了显著效果，但是这些方法都需要实例级的时序标注，这项工作十分耗时耗力，因此在实际应用的过程中受限。

1.2 弱监督时序动作定位

为了缓解标注成本过高的问题，弱监督时序动作定位任务受到越来越多的关注。在弱监督环境下仅有视频级的标注，Wang等人（2017）首次将多实例学习的框架引入到任务中。按此框架大致可以分为基于擦除的、基于伪标签的、基于多分支注意力的、基于不确定性的方法和基于视觉语言的方法。基于擦除的方法（Singh 和 Lee，2017）将最具判别性的部分区域擦除，迫使模型关注不具区分性的区域，而获取更全面的特征信息；基于伪标签的方法（Ren 等，2023；Li 等，2022b）利用视频片段的分类结果，生成相应的伪标签作为监督信息来提高定位性能；基于多分支注意力的方法（Zhai 等，2023；Liu 等，2019）设置多分支和多注意力，学习不同的分类信息和注意力信息，形成知识补充，让网络不仅关注最显著的部分；基于不确定性的方法（Lee 等，2020；Islam 等，2021；Chen 等，2022）利用不确定性估计来提高视频中前景和背景的区分能力；最近出现的基于视觉语言的方法（Li 等，2023；Ju 等，2022）借助了视频信息以外的文本提示信息来辅助模型进行定位。然而，这类方法并没有考虑不同类型提示信息之间的互补性。相比之下，本文设计的提示交互模块，分别将手工类型提示和可学习类型提示与动作片段匹配，然后通过注意力机制加权，挖掘不同尺度的特征信息，以获取不同类型文本提示信息引导下的分类激活序列（CAS）。

1.3 视觉语言模型

视觉语言模型受到广泛关注，这些模型旨在对图像和文本之间的对应关系进行建模，从而实现更全面的视觉感知。为了更好地将大模型应用在下游任务，如视频问答（Tapaswi 等，2016）、视频接地（He 等，2019）、视频字幕（Wu 等，2023）等，特定的文本提示信息可以提高模型的性能和鲁棒性。例如，在

CLIP(Radford 等,2021)中利用的手工类型提示模板“*a video of*{*class*}”。然而这类文本提示信息仅仅是对动作标签的简单扩充,无法学习到有关动作类别中更为泛化的知识。Zhou 等人(2022)提出的可学习类型提示,可以在训练过程中调整和优化,因此可以弥补手工类型提示的不足。为了最大化两种文本提示信息的作用,本文借助多层筛选的思想,设计相应的阈值筛选模块,将不同类型提示信息经过处理得到的分类激活序列通过设置阈值的方式进行筛选,并将筛选后得到的分类激活序列用于定位,实现了不同类型文本提示信息的充分利用。

2 模型方法

本文方法框架如图 2 所示,框架由 3 个部分组成:特征提取部分(2.2)、提示交互模块(2.3)和阈值筛选模块(2.4)。

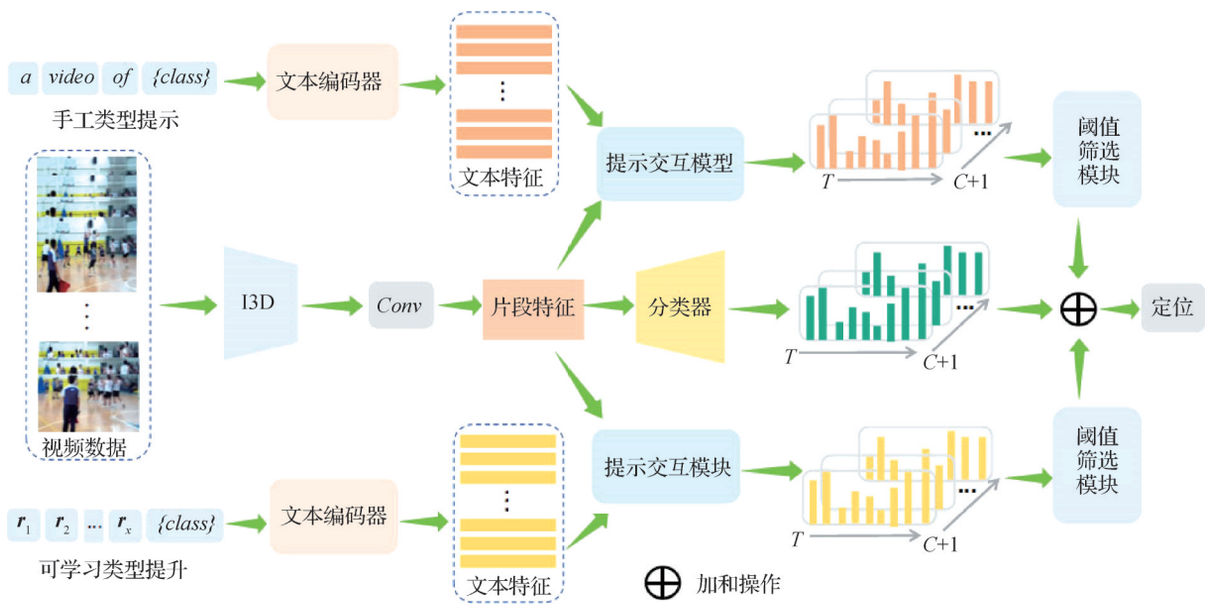

图 2 基于多类型提示互补的定位模型框架图

Fig. 2 Framework map of a localization model based on complementary multi-type prompts

2.1 整体结构

2.1.1 弱监督时序动作定位任务的问题表述

弱监督时序动作定位指在未修剪过的视频中定位动作实例的起止时间。动作实例可以表示为:$\{(c_i, s_i, e_i, q_i)\}$,s_i和e_i分别表示第i个动作的起始时间和结束时间,c_i和q_i表示当前动作的类别和置信度得分,且$c_i \in \{1, \cdots, C+1\}$,$C$表示已知动作类别的个数,$C+1$表示未知类。在训练过程中,给定$N$个没有裁剪过的视频$\boldsymbol{V}$,将其记为$\{\boldsymbol{V}_j\}_{j=1}^{N}$,以及对应的视频级标签:$\{y_j \in \mathbf{R}^C\}_{j=1}^{N}$,$\mathbf{R}^C$为包含所有已知动作类别的标签集合。如果第$j$个动作出现在视频中则$y_i = 1$,否则$y_i = 0$。

2.2 特征提取模块

2.2.1 视频特征提取

依照之前的工作(Ren 等,2023;Li 等,2022a),先将未经裁剪的视频$\boldsymbol{V}$划分为T个片段,然后将T个片段送入预训练的 I3D 网络(Carreira 和 Zisserman,2017)提取图像特征和光流特征,分别记为$\boldsymbol{X}_r \in \mathbf{R}^{T\times D}$和$\boldsymbol{X}_f \in \mathbf{R}^{T\times D}$,$D$为特征的维度。然后将 RGB 特征和光流特征融合得到融合后的特征:$\boldsymbol{X}_{rf} \in \mathbf{R}^{T\times 2D}$,然后经过卷积层、ReLU(rectified linear unit)和 dropout 层处理后得到最终的视频级特征$\boldsymbol{X} \in \mathbf{R}^{T\times 2D}$,即

$$\boldsymbol{X} = Conv(ReLU(\boldsymbol{X}_{rf})) \tag{1}$$

2.2.2 文本特征提取

由于弱监督环境下仅有视频级的标签,因此本文将对应的动作标签与提示模板*a video of*{*class*}相结合得到手工类型提示信息$\boldsymbol{P}_1$,并且为了弥补手工类型提示信息无法获得有关动作具体知识的不足,根据 Zhou 等人(2022)提出的 CoOp(context optimization),本文又引入了可学习类型提示。具体而言,将

x 个上下文向量 $\boldsymbol{R}=\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_x\}$ 作为可学习提示符，其中 $\boldsymbol{R}$ 从标准正态分布中随机采样生成得到，并将第 i 个类别的词向量添加到 $\boldsymbol{R}$ 中得到 $\boldsymbol{P}_2=\{\boldsymbol{r}_1,\boldsymbol{r}_2,\cdots,\boldsymbol{r}_x,\boldsymbol{c}_i\}$，而可学习类型提示的数量取决于动作类别的数量。最后分别将两种类型提示信息 $\boldsymbol{P}_1$ 和 $\boldsymbol{P}_2$ 送入到文本编码器 θ 中，得到对应的文本嵌入 $\boldsymbol{E}_{\mathrm{h}}\in\mathbf{R}^{(C+1)\times 2D}$ 和 $\boldsymbol{E}_{\mathrm{l}}\in\mathbf{R}^{(C+1)\times 2D}$，其中，$C+1$ 表示背景类。

$$\boldsymbol{E}_{\mathrm{h}}=\theta(\boldsymbol{P}_1),\ \boldsymbol{E}_{\mathrm{l}}=\theta(\boldsymbol{P}_2) \tag{2}$$

2.3 提示交互模块

为了对视频片段和文本的对应关系进行建模，本文设计了提示交互模块，框架如图 3 所示。

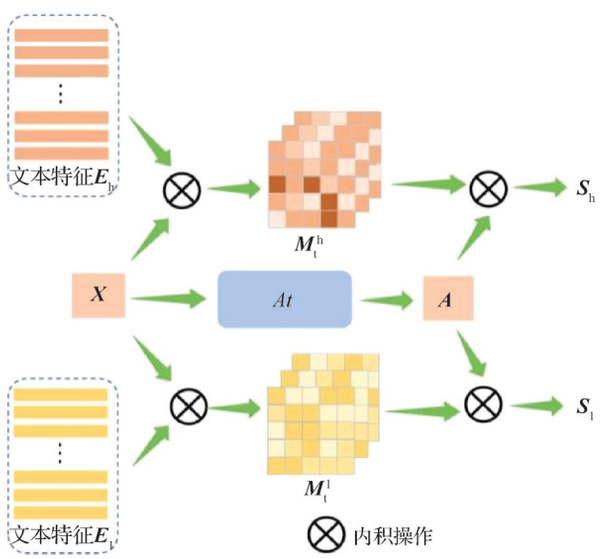

图 3 提示交互模块图

Fig. 3 Prompts interaction module map

现有视觉语言方法（Li 等，2023；Ju 等，2022）采用直接将不同类型文本提示信息与动作片段匹配的方式。而本文考虑到不同类型提示信息间的互补性，在将不同类型文本提示信息与视频特征进行匹配之后，分别利用生成的注意力权重 $\boldsymbol{A}$ 加权，以得到不同类型提示信息引导下的 CAS。并且基于 Lee 等人（2020）背景抑制的方法表明，使用注意力权重可以有效抑制背景对动作分类的干扰。其次，本文通过相似矩阵来衡量文本信息与动作片段的对应关系，过程具体为

$$\boldsymbol{A}=At(\boldsymbol{X}) \tag{3}$$

$$\boldsymbol{M}_{\mathrm{t}}^{\mathrm{h}}=\boldsymbol{X}\times\boldsymbol{E}_{\mathrm{h}},\ \boldsymbol{M}_{\mathrm{t}}^{\mathrm{l}}=\boldsymbol{X}\times\boldsymbol{E}_{\mathrm{l}} \tag{4}$$

$$\boldsymbol{S}_{\mathrm{h}}=\boldsymbol{A}\times\boldsymbol{M}_{\mathrm{t}}^{\mathrm{h}},\ \boldsymbol{S}_{\mathrm{l}}=\boldsymbol{A}\times\boldsymbol{M}_{\mathrm{t}}^{\mathrm{l}} \tag{5}$$

式中，$\boldsymbol{S}_{\mathrm{h}}$ 和 $\boldsymbol{S}_{\mathrm{l}}$ 分别表示手工类型提示和可学习类型提示引导下生成的 CAS。$At(\cdot)$ 表示由 3 个 1D 卷积层、ReLU、dropout 层和池化层组成的注意力机制。$\boldsymbol{M}_{\mathrm{t}}^{\mathrm{h}}$ 和 $\boldsymbol{M}_{\mathrm{t}}^{\mathrm{l}}$ 表示手工类型提示和可学习类型提示与片段交互后的相似矩阵，×代表内积运算。

此外，为了约束视频片段与文本提示信息的匹配，本文设计了片段级对比损失。具体而言，本文将得到的视频级特征 $\boldsymbol{X}$ 和两种不同类型的提示信息表示 $\boldsymbol{E}_{\mathrm{h}}\in\mathbf{R}^{1\times 2D}$ 与 $\boldsymbol{E}_{\mathrm{l}}\in\mathbf{R}^{1\times 2D}$，使用同一批数据 $\boldsymbol{E}_{\mathrm{h}}\in\mathbf{R}^{(C+1)\times 2D}$，$\boldsymbol{E}_{\mathrm{l}}\in\mathbf{R}^{(C+1)\times 2D}$ 来计算损失。之后本文基于 InfoNCE（information noise contrastive estimation）（van den Oord 等，2019）损失得到了片段级对比损失，具体为

$$L(\boldsymbol{X},\boldsymbol{E})=-\frac{1}{T}\sum_{i=1}^{N}\log\frac{\exp\dfrac{\partial(\boldsymbol{x}_i,\boldsymbol{e}_i)}{\tau}}{\sum\limits_{j}^{N}\exp\dfrac{\partial(\boldsymbol{x}_j,\boldsymbol{e}_j)}{\tau}} \tag{6}$$

$$\partial(\boldsymbol{x}_i,\boldsymbol{e}_i)=\frac{\boldsymbol{x}_i\times\boldsymbol{e}_i^{\mathrm{T}}}{\|\boldsymbol{x}_i\|\times\|\boldsymbol{e}_i^{\mathrm{T}}\|} \tag{7}$$

式中，$\boldsymbol{E}$ 包含 $\boldsymbol{E}_{\mathrm{h}}$ 和 $\boldsymbol{E}_{\mathrm{l}}$，$\tau$ 为训练过程中优化的温度参数。$\partial(\cdot,\cdot)$ 为余弦相似度公式。最后，由于有两种不同类型的提示信息，所以分别利用 InfoNCE 损失计算损失后得到 L_{ct}，定义为

$$L_{\mathrm{ct}}=L(\boldsymbol{X},\boldsymbol{E}_{\mathrm{h}})+L(\boldsymbol{X},\boldsymbol{E}_{\mathrm{l}}) \tag{8}$$

2.4 阈值筛选模块

与传统的多实例学习框架（Wang 等，2017）类似，本文也利用经分类得到的 CAS 计算置信分数。但考虑到不同类型文本提示信息的优势不同，因此，经提示交互模块得到的两类 CAS：$\boldsymbol{S}_{\mathrm{h}}$ 和 $\boldsymbol{S}_{\mathrm{l}}$，需要经过合理的筛选才能充分发挥不同类型提示信息的引导作用。为此，本文通过筛选各序列中置信度高的概率得分而设计了阈值筛选模块。首先，将视频级特征 $\boldsymbol{X}$ 经过分类器 f 处理，得到基准 CAS，将其记为 $\boldsymbol{S}\in\mathbf{R}^{T\times(C+1)}$。然后分别将 3 种不同的 CAS：$\boldsymbol{S}$，$\boldsymbol{S}_{\mathrm{h}}$，$\boldsymbol{S}_{\mathrm{l}}$ 经过归一化处理，得到 $\boldsymbol{S}'$，$\boldsymbol{S}'_{\mathrm{h}}$，$\boldsymbol{S}'_{\mathrm{l}}$，可表示为

$$\boldsymbol{S}=f(\boldsymbol{X}) \tag{9}$$

$$\boldsymbol{S}'=\frac{\exp(\boldsymbol{S})}{\sum\limits_{j}\exp(\boldsymbol{S})} \tag{10}$$

为了筛选具有较高置信度的概率分数，本文设置了相应的阈值矩阵 $\boldsymbol{m}_{\mathrm{h}}$ 和 $\boldsymbol{m}_{\mathrm{l}}$，用 ε 和 δ 分别表示 CAS 的上界与下界，这里取上界 0.4，下界 0.1，矩阵 $\boldsymbol{m}_{\mathrm{h}}$，$\boldsymbol{m}_{\mathrm{l}}$ 分别用于筛选序列 $\boldsymbol{S}'_{\mathrm{h}}$ 和 $\boldsymbol{S}'_{\mathrm{l}}$ 中高于上界的部分。在阈值矩阵中，将高于上界的部分取值为 1，其余部

分则为0。借助阈值矩阵筛选各序列中的最大值，以提高对应动作类别的得分，具体过程为

$$S_{\mathrm{h}}'' = S_{\mathrm{h}} \otimes \boldsymbol{m}_{\mathrm{h}}, S_{\mathrm{l}}'' = S_{\mathrm{l}} \otimes \boldsymbol{m}_{\mathrm{l}} \tag{11}$$

$$S_{\mathrm{up}} = \max(S_{\mathrm{h}}'' \cap S_{\mathrm{l}}'') \tag{12}$$

式中，S_{h}''和S_{l}''表示符合大于阈值条件的序列值，⊗表示逐元素乘法运算，S_{up}表示经过上界筛选后得到的CAS。

接下来对两序列的下界进行筛选，对应的阈值矩阵为$\boldsymbol{m}_{\mathrm{h}}'$，$\boldsymbol{m}_{\mathrm{l}}'$。与上界筛选同理，利用对应阈值矩阵筛选得到两序列中小于阈值的部份，这部分代表了对某一动作类别低置信度的得分，具体过程为

$$S_{\mathrm{h}}''' = S_{\mathrm{h}} \otimes \boldsymbol{m}_{\mathrm{h}}', S_{\mathrm{l}}''' = S_{\mathrm{l}} \otimes \boldsymbol{m}_{\mathrm{l}}' \tag{13}$$

$$S_{\mathrm{down}} = \min(S_{\mathrm{h}}''' \cap S_{\mathrm{l}}''') \tag{14}$$

式中，S_{h}'''和S_{l}'''表示符合小于阈值条件的序列值，S_{down}表示经过下界筛选后得到的分类激活序列。然后，将各序列中处于上、下界之间的部分进行比较，将其中较大的值取出作为新的分类激活序列S_{mid}，可表示为

$$S_{\mathrm{mid}} = \max((S_{\mathrm{h}} - S_{\mathrm{h}}'' - S_{\mathrm{h}}'''), (S_{\mathrm{l}} - S_{\mathrm{l}}'' - S_{\mathrm{l}}''')) \tag{15}$$

最终，本文将经过筛选的3类序列进行加和得到新的分类激活序列$\tilde{S}$，定义为

$$\tilde{S} = S_{\mathrm{up}} + S_{\mathrm{down}} + S_{\mathrm{mid}} \tag{16}$$

由于文本提示信息仅仅起到引导的作用，因此新得到的CAS：$\tilde{S}$并不能够作为定位的主要依据而直接使用。为此，本文在S的基础上与$\tilde{S}$按照一定的比例进行加和，得到最终用于定位的激活序列$\hat{S} \in \mathbf{R}^{T \times (C+1)}$，定义为

$$\hat{S} = S + \alpha \tilde{S} \tag{17}$$

式中，α为比例参数。

2.5 模型的训练和推理

2.5.1 训练阶段

与现有方法类似（Zhai等，2023；Liu等，2019），本文也利用多实例学习来训练模型，具体而言，首先将最终得到的分类激活序列$\hat{S} \in \mathbf{R}^{T \times (C+1)}$在时序上应用Top-$K$聚合策略（Paul等，2018），并进行softmax操作，得到预测的视频级分类得分$\hat{\boldsymbol{y}}_c$，然后通过交叉熵损失对其进行监督，因此，最终的损失为

$$\hat{S} = f_{\mathrm{softmax}}\left(\frac{1}{k}(\hat{S})\right) \tag{18}$$

$$L_{\mathrm{cls}} = \sum_{c=1}^{C+1} - \boldsymbol{y}_c \log \hat{\boldsymbol{y}}_c \tag{19}$$

$$L = L_{\mathrm{cls}} + \lambda L_{\mathrm{ct}} \tag{20}$$

3 实验结果与分析

3.1 数据集

THUMOS14（Idrees等，2017）包含20个动作类别的413个未裁剪视频，其中有200个验证视频和213个测试视频，每个视频中平均含有15个实例。本文根据之前的工作，用验证集进行训练、测试集进行评估。ActivityNet1.2是一个大型数据集，包含来自100个动作类别的4 819个训练视频、2 383个验证视频和2 480个测试视频，其中参考现有工作（Ren等，2023）在验证集上评估所提方法的效果。而ActivityNet1.3（Heilbron等，2015）包含大量人类在真实环境中未剪辑的动作视频（熊成鑫等，2020），来自200个动作类别的10 024个训练视频，4 962个验证视频和5 044个测试视频，每个视频平均含有1.6个动作实例。本节用所有的训练视频来训练模型，并在所有的测试视频评估本文方法。

3.2 评估标准

为了评估定位的性能，本文遵循标准的方法，在不同的交并比（intersection over union，IoU）上使用平均精度均值（mean average precision，mAP）。当类别预测正确且IoU值超过设置的阈值时，生成的动作区间才可以认为是正确的。

3.3 实施细节

本文使用在Kinetics-400上预训练的I3D（Carreira和Zisserman，2017）网络进行特征提取，用TVL1（Duval等，2009）算法从RGB帧中提取光流。并用Adam（Kingma和Ba，2017）优化器进行训练，在THUMOS14数据集中，学习率设置为0.000 3，优化权重为0.001，批次大小为10，对每个视频的随机采样片段数为750，$\boldsymbol{P}_2$数量为20，x = 300表示序列长度，对于ActivityNet1.2和ActivityNet1.3数据集，学习率设置为0.000 02，采样片段分别为150和400，$\boldsymbol{P}_2$分别为100和200。对于超参数的设置，τ为0.05，α为0.2，λ为0.05。

3.4 与最先进方法的比较

本文方法与最先进的WTAL方法在不同IoU阈值上进行比较。在THUMOS14数据集上的结果如

表 1 所示，在多阈值设置条件下，效果均有明显提升，分别在 mAP(0.1∶0.5)、mAP(0.1∶0.7)、mAP(0.3∶0.7)上实现了 58.2%、39.1%、47.9%的平均性能。特别是在高阈值 mAP(0.3～0.7)下，超过了 P-MIL 方法 0.9%。对于更复杂的数据集 ActivityNet1.2 与 ActivityNet1.3，实验结果分别如表 2 和表 3 所示。ActivityNet1.2 中本文方法在 mAP(0.5∶0.95)上实现了 27.3%的性能，与 P-MIL 方法相比平均提升近 0.8%，ActivityNet1.3 的实验结果也显示有一定的提升。但当 IoU 取值为 0.95 时，此时的性能要低于对比方法。本文认为性能下降的原因是随着 IoU 取值的增大，模型的定位难度也随之增大，同时 ActivityNet1.3 视频数量约为 ActivityNet1.2 的两倍，且训练集、验证集和测试集中动作类别差异较大，此时仅靠文本提示信息中所包含的语义信息，并不能够真正弥补监督信息的缺失，因此高阈值条件下文本提示信息的作用会受到限制。

表 1　在 THUMOS14 数据集上的实验结果

Table 1　Experimental results on the THUMOS14 dataset

/%

方法	mAP@IoU							平均		
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1～0.5	0.1～0.7	0.3～0.7
ASL(Ma 等,2021)	67.0	0.0	51.8	0.0	31.1	0.0	11.4	0.0	0.0	40.3
CoLA(Zhang 等,2021)	66.2	59.5	51.5	41.9	32.2	22.0	13.1	50.3	32.1	40.9
ACSNet(Liu 等,2021)	0.0	0.0	51.4	42.7	32.4	22.0	11.7	0.0	32.0	0.0
ACGNet(Yang 等,2022)	68.1	62.6	53.1	44.6	34.7	22.6	12.0	52.6	33.4	42.5
FTCL(Gao 等,2022)	69.6	63.4	55.2	45.2	35.6	23.7	12.2	53.8	34.4	43.6
ASM-Loc(He 等,2022)	71.2	65.5	57.1	46.8	36.6	25.2	13.4	55.4	35.8	45.1
T-SN(Wang 等,2023)	73.0	68.2	60.0	47.9	37.1	24.4	12.7	57.2	36.4	46.2
Boosting-WTAL(Li 等,2023)	0.0	0.0	56.2	47.8	39.3	27.5	15.2	0.0	37.2	0.0
TFE-DCN(Ju 等,2022)	72.3	66.5	58.6	49.5	40.7	27.1	13.7	57.5	37.9	46.9
P-MIL(Ren 等,2023)	71.8	67.5	58.9	49.0	40.0	27.1	15.1	57.4	38.0	47.0
本文	73.1	67.3	59.4	49.8	41.3	28.9	15.9	**58.2**	**39.1**	**47.9**

注：加粗字体表示各列最优结果。

表 2　在 ActivityNet1.2 数据集上的实验结果

Table 2　Experimental results on the ActivityNet1.2 dataset

/%

方法	mAP@IoU			平均
	0.5	0.75	0.95	
TEN(Li 等,2022b)	41.6	24.8	5.4	25.2
AUMN(Zhai 等,2022)	42.0	25.0	5.6	25.5
UGCT(Yang 等,2021)	41.8	25.3	5.9	25.8
D2-Net(Narayan 等,2021)	42.3	25.5	5.8	26.0
ACGNet(Yang 等,2022)	41.8	26.0	5.9	26.1
DGCNN(Shi 等,2022)	42.0	25.8	6.0	26.2
CO_2-Net(Hong 等,2021)	43.3	26.3	5.2	26.4
P-MIL(Ren 等,2023)	44.2	26.1	5.3	26.5
本文	**44.9**	**27.2**	**6.1**	**27.3**

注：加粗字体表示各列最优结果。

表 3　在 ActivityNet1.3 数据集上的实验结果

Table 3　Experimental results on the ActivityNet1.3 dataset

/%

方法	mAP@IoU			平均
	0.5	0.75	0.95	
DCC(Li 等,2022a)	38.8	24.2	5.7	24.3
FTCL(Gao 等,2022)	40.0	24.3	6.4	24.8
RSKP(Huang 等,2022)	40.6	24.6	5.9	25.0
ASM-Loc(He 等,2022)	41.0	24.9	6.2	25.1
TFE-DCN(Ju 等,2022)	41.4	24.8	6.4	25.3
P-MIL(Ren 等,2023)	41.8	25.4	5.2	25.5
Boosting-WTAL(Li 等,2023)	41.8	**26.0**	6.0	26.0
T-SN(Wang 等,2023)	41.8	25.7	**6.5**	26.3
本文	**42.1**	**26.0**	6.2	**26.7**

注：加粗字体表示各列最优结果。

3.5 消融实验

3.5.1 各部分的有效性

本文框架主要包含3个部分：1)提示交互模块，利用手工类型提示和可学习类型提示之间的互补性，引导模型定位；2)阈值筛选模块用于最大化两种提示的作用；3)视频片段与文本匹配的对比损失，记为L_{ct}。

为了验证框架中各部分的有效性，本文进行了全面的消融实验，如表4所示。具体而言，基准模型是仅含有视频信息的弱监督时序动作定位任务的基础模型；手工类型提示表示由手工模板与标签组成的提示信息；可学习类型提示表示由一组可学习的向量组成的提示信息；L_{ct}表示用于约束视频片段和提示信息匹配的损失。当手工类型提示与可学习类型提示同时引用时，表示本文方法中提出的提示交互模块。通过比较基准模型与(e)两种设定下的性能，可以得出以下结论：引入一定的文本辅助信息比仅用视频信息的弱监督时序动作定位模型效果要好。对比(a)和(b)两种设定下的方法，可以看出引入不同类型的提示信息对模型有不同的影响，虽然在不同范围内的mAP值上均有一定程度的提升，但是引入手工类型提示信息后，高阈值条件下的效果要低于“基准模型”。因此，手工类型提示引导下所学到的知识对于困难动作的识别并没有起到帮助作用。高阈值条件下的效果要低于“基准模型”，因此，手工类型提示引导下所学到的知识对于困难动作的识别并没有起到帮助作用。将(a)、(b)、(c)三者相比，发现在(c)的设定下性能有了明显提升，由此验证了提示交互模块的有效性。此外，在逐步添加阈值筛选模块和L_{ct}之后，本文方法的性能平均提高了近1%。通过比较(c)与(e)可以验证对比损失L_{ct}的必要性。另外，在(d)的基础上添加阈值筛选模块，得到了本文方法的最优性能，并且(d)与前面的几种设定相比，性能出现了下降，本文认为是在L_{ct}的约束下，当出现不同的提示信息，而没有设计合理的筛选方案，可能会干扰模型的定位，由此验证了阈值筛选模块的有效性。

表4 在THUMOS14数据集上各部分的消融实验

Table 4 Ablation experiments on different components on the THUMOS14 dataset

/%

方法	手工类型提示	可学习类型提示	阈值筛选模块	L_{ct}	mAP@IoU							平均
					0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.1～0.7
基准模型	–	–	–	–	71.2	66.1	57.8	48.3	39.9	27.6	14.8	46.6
(a)	√	–	–	–	71.8	66.5	58.2	48.3	39.5	27.9	14.3	46.7
(b)	–	√	–	–	72.5	66.9	58.4	48.7	40.2	28.3	15.1	47.2
(c)	√	√	√	–	72.8	67.1	58.8	49.4	40.6	28.5	15.4	47.6
(d)	√	√	–	√	72.4	66.9	58.5	48.6	39.8	28.0	14.7	47.0
(e)	√	√	√	√	73.1	67.3	59.4	49.8	41.3	28.9	15.9	**47.9**

注：加粗字体表示最优结果，“√”表示框架中具有的部分，“–”表示不具有。

3.5.2 不同手工类型提示模板的消融实验

表5为不同手工类型提示模板的实验结果。其中模板“*a video of*{*class*}”实现了最优的性能。通过比较不同手工模板，如：“*a*{*class*}”与其他模板相比性能要低，而其他模板之间的性能差距不大。本文认为手工类型提示信息的引导作用与其模板所包含的语义信息有关，语义信息越丰富对模型定位的帮助越大，对于模板“*a*{*class*}”，其包含的语义信息相比其他模板要少。并且，有些模板并不一定适合当前模型，因此不同模板对性能的影响是比较明显的。

3.5.3 关于不同阈值设置的消融实验

本文设计的阈值筛选模块中，主要选取了高于上界和低于下界的部分，对应的实验结果如图5所示，在图中，*Max*表示上界值，*Min*表示下界值，这里对*Max*的取值设置范围为0.4～0.9，用于筛选高置信度的分类激活序列；*Min*的范围为0.05～0.3，则代表了低置信度的CAS。实验结果表明，在上界取值0.4、下界取值0.1时，实现了最优性能。另外通

表 5　在 THUMOS14 数据集上使用不同手工类型提示模板的性能比较

Table 5　Performance comparison of different handcrafted prompt templates on the THUMOS14 dataset

/%

手工类型提示模板	mAP@IoU					平均
	0.3	0.4	0.5	0.6	0.7	
a video of	58.6	49.1	40.3	28.3	15.1	38.3
a	58.6	48.9	40.2	28.1	15.1	38.2
a video of action	58.8	49.2	40.6	28.4	15.4	38.5
a video about	59.1	49.5	41.0	28.4	15.4	38.7
a video of the	59.4	49.8	41.3	28.9	15.9	**39.1**

注：加粗字体为最优值。

过图 4，可以更直观地看出不同阈值条件下性能的变化。随着上界 *Max* 的取值越来越大或下界的取值 *Min* 越来越小，性能都会出现明显的下降。通过观察不同激活序列的概率取值情况得出结论：上界取值增大，其符合大于上界取值的数量也会随之减少，下界取值减少也是同样的道理。因此，这样筛选出来的激活序列所包含的取值情况较少，所以并不能起到很好的筛选作用。为此，本文在一定程度上降低上界值、提高下界值，避免对其强干预，以充分发挥两种文本提示信息的互补作用。

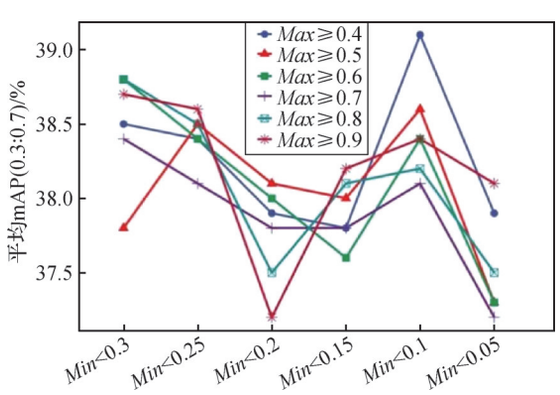

图 4　在 THUMOS14 数据集上不同阈值条件的可视化结果

Fig. 4　Visualization results under different threshold conditions on the THUMOS14 dataset

3.5.4　关于比例参数 α 的消融实验

为了测试不同比例参数设置下得到的分类激活序列对性能的影响，本文对 α 选取了 10 种不同取值，对应的实验结果如图 5 所示。α 的取值越高表明文本提示信息引导下的分类激活序列占比就越高，由实验结果可以看出，占比越高性能反而降低，这验证了文本提示信息并不能够占据主导地位，而是起到辅助引导的作用。当 $\alpha = 0.2$ 时，本方法实现了最好的性能。

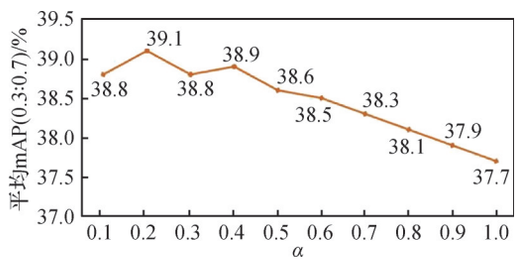

图 5　在 THUMOS14 数据集上不同 α 取值的实验结果

Fig. 5　Results for different values of α on the THUMOS14 dataset

3.5.5　关于计算效率的实验

本文通过计算模型的训练时间与测试时间衡量所提方法的计算效率，实验结果如表 6 所示。在训练阶段可以看出本文方法增加训练时间 1.1 s，而测试阶段用时均为 24 s 左右，因此所提方法与原模型在 mAP(0.1:0.7) 提升 1.2% 的基础上，训练和测试效率与原模型相当。

表 6　在 THUMOS14 数据集上的计算效率

Table 6　Computational efficiency on the THUMOS14 dataset

方法	mAP(0.1:0.7)/%	训练时间/s	测试时间/s
基准模型	37.9	**33.6**	**24.1**
本文	**39.1**	34.7	24.7

注：加粗字体表示各列最优结果。

3.5.6　可视化实验

为了更直观地展示所提方法的优势，图 6 将不同类型视频片段的结果可视化。其中包含了 3 个动作实例的检测结果，第 1 行代表视频动作片段，下面两行依次是视频中动作的真实定位区间，基准模型生成的定位结果以及本文方法最终生成的定位结果。根据定位的结果可以看出，本文方法能够提高定位边界的准确性，同时可以纠正错误的预测。因此，从图中可以看出，将不同类型的文本提示信息引入到基准模型中，并充分结合二者的优势，有助于生成更准确的定位结果，并在一定程度上抑制背景片段的影响。

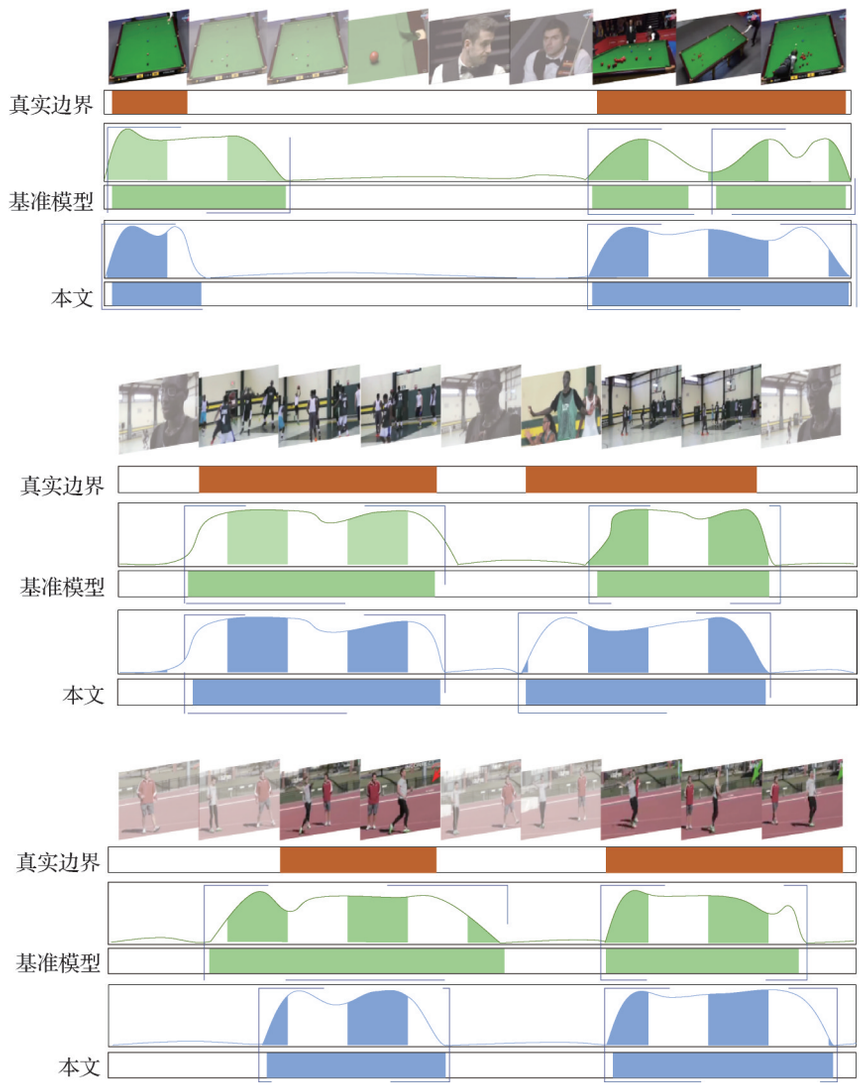

图6　不同类型视频片段的可视化结果

Fig. 6　Visualization results for different types of video clips

4　结　论

本文提出一种基于多类型提示互补的时序动作定位模型。其中,设计提示交互模块以实现不同类型文本提示信息的交互。并且,借助片段级对比损失来约束提示信息与动作片段的匹配。最后的阈值筛选模块,最大限度地利用两种文本提示信息之间的互补性。该方法在3个具有挑战性数据集上的大量实验证明了方法的有效性。

相比现有方法,本文方法虽然实现了显著的性能提升,但仍存在一些不足的地方亟待改进。在具体实验过程中,可以看出仅靠文本提示信息对模型定位能力的提升是有限的,特别是在数据集更复杂、定位难度更大的情况下。考虑到文本提示信息的局限性,仅依靠提示包含的语义信息可能无法全面理解和处理复杂的实际情境。为了更好地解决这个问题,需要提供更丰富的语义信息引导模型定位,并增强视频中动作和背景的区分能力。在之后的工作中,将进行以下尝试:1)除了引入一定外部辅助信息的手段之外,尝试采用不同方法挖掘视频本身的动作信息、文本信息等,以丰富语义信息解决复杂情况下模型鲁棒性较差的问题;2)探究如何更好地区分视频中的动作与背景,以抑制背景对动作定位的干扰。

参考文献(References)

Carreira J and Zisserman A. 2017. Quo Vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Con-

ference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chen M Y, Gao J Y, Yang S C and Xu C S. 2022. Dual-evidential learning for weakly-supervised temporal action localization//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 192-208 [DOI: 10.1007/978-3-031-19772-7_12]

Cheng F and Bertasius G. 2022. TALLFormer: temporal action localization with a long-memory transformer [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2204.01680.pdf

Duval V, Aujol J F and Gousseau Y. 2009. The TVL1 model: a geometric point of view. Multiscale Modeling and Simulation, 8(1): 154-189 [DOI: 10.1137/090757083]

Gaidon A, Harchaoui Z and Schmid C. 2013. Temporal localization of actions with actoms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11): 2782-2795 [DOI: 10.1109/TPAMI.2013.65]

Gao J Y, Chen M Y and Xu C S. 2022. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2203.16800.pdf

He B, Yang X T, Kang L, Cheng Z Y, Zhou X and Shrivastava A. 2022. ASM-Loc: action-aware segment modeling for weakly-supervised temporal action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2203.15187.pdf

He D L, Zhao X, Huang J Z, Li F, Liu X and Wen S L. 2019. Read, watch, and move: reinforcement learning for temporally grounding natural language descriptions in videos//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 8393-8400 [DOI: 10.1609/aaai.v33i01.33018393]

Heilbron F C, Escorcia V, Ghanem B and Niebles J C. 2015. ActivityNet: a large-scale video benchmark for human activity understanding//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 961-970 [DOI: 10.1109/CVPR.2015.7298698]

Hong F T, Feng J C, Xu D, Shan Y and Zheng W S. 2021. Cross-modal consensus network for weakly supervised temporal action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2107.12589.pdf

Huang L J, Wang L and Li H S. 2022. Weakly supervised temporal action localization via representative snippet knowledge propagation [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2203.02925.pdf

Idrees H, Zamir A R, Jiang Y G, Gorban A, Laptev I, Sukthankar R and Shah M. 2017. The THUMOS challenge on action recognition for videos "in the wild". Computer Vision and Image Understanding, 155: 1-23 [DOI: 10.1016/j.cviu.2016.10.018]

Islam A, Long C J and Radke R. 2021. A hybrid attention mechanism for weakly-supervised temporal action localization//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 1637-1645 [DOI: 10.1609/aaai.v35i2.16256]

Ju C, Zheng K H, Liu J X, Zhao P S, Zhang Y, Chang J L, Wang Y F and Tian Q. 2022. Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2212.09335.pdf

Kingma D P and Ba J L. 2017. Adam: a method for stochastic optimization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/1412.6980.pdf

Lee P, Uh Y and Byun H. 2020. Background suppression network for weakly-supervised temporal action localization//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: IEEE: 11320-11327 [DOI: 10.1609/aaai.v34i07.6793]

Li G Z, Cheng D, Ding X P, Wang N N, Wang X Y and Gao X B. 2023. Boosting weakly-supervised temporal action localization with text information [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2305.00607.pdf

Li J J, Yang T Y, Ji W, Wang J and Cheng L. 2022a. Exploring denoised cross-video contrast for weakly-supervised temporal action localization//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 19882-19892 [DOI: 10.1109/CVPR52688.2022.01929]

Li Z Q, Ge Y X, Yu J R and Chen Z M. 2022b. Forcing the whole video as background: an adversarial learning strategy for weakly temporal action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2207.06659.pdf

Liu D C, Jiang T T and Wang Y Z. 2019. Completeness modeling and context separation for weakly supervised temporal action localization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1298-1307 [DOI: 10.1109/CVPR.2019.00139]

Liu Z Y, Wang L, Zhang Q L, Tang W, Yuan J S, Zheng N N and Hua G. 2021. ACSNet: action-context separation network for weakly supervised temporal action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2103.15088.pdf

Ma J W, Gorti S K, Volkovs M and Yu G W. 2021. Weakly supervised action selection learning in video [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2105.02439.pdf

Narayan S, Cholakkal H, Hayat M, Khan F S, Yang M H and Shao L. 2021. D2-net: weakly-supervised action localization via discriminative embeddings and denoised activations [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2012.06440.pdf

Paul S, Roy S and Roy-Chowdhury A K. 2018. W-TALC: weakly-supervised temporal activity localization and classification [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/1807.10418.pdf

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2103.00020.pdf

Ren H, Yang W F, Zhang T Z and Zhang Y D. 2023. Proposal-based multiple instance learning for weakly-supervised temporal action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2305.17861.pdf

Shi H C, Zhang X Y, Li C S, Gong L X, Li Y and Bao Y J. 2022. Dynamic graph modeling for weakly-supervised temporal action localization//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 3820-3828 [DOI: 10.1145/3503161.3548077]

Shou Z, Chan J, Zareian A, Miyazawa K and Chang S F. 2017. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1417-1426 [DOI: 10.1109/CVPR.2017.155]

Singh K K and Lee Y J. 2017. Hide-and-seek: forcing a network to be meticulous for weakly-supervised object and action localization [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/1704.04232.pdf

Tapaswi M, Zhu Y K, Stiefelhagen R, Torralba A, Urtasun R and Fidler S. 2016. MovieQA: understanding stories in movies through question-answering//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4631-4640 [DOI: 10.1109/CVPR.2016.501]

van den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/1807.03748.pdf

Wang L M, Xiong Y J, Lin D H and Van Gool L. 2017. UntrimmedNets for weakly supervised action recognition and detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6402-6411 [DOI: 10.1109/CVPR.2017.678]

Wang Y, Li Y D and Wang H B. 2023. Two-stream networks for weakly-supervised temporal action localization with semantic-aware mechanisms//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18878-18887 [DOI: 10.1109/CVPR52729.2023.01810]

Wu W H, Luo H P, Fang B, Wang J D and Ouyang W L. 2023. Cap4Video: what can auxiliary captions do for text-video retrieval? [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2301.00184.pdf

Xiong C X, Guo D and Liu X L. 2020. Temporal proposal optimization for temporal action detection. Journal of Image and Graphics, 25(7): 1447-1458 (熊成鑫, 郭丹, 刘学亮. 2020. 时域候选优化的时序动作检测. 中国图象图形学报, 25(7): 1447-1458) [DOI: 10.11834/jig.190440]

Xu M M, Zhao C, Rojas D S, Thabet A and Ghanem B. 2020. G-TAD: sub-graph localization for temporal action detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10153-10162 [DOI: 10.1109/CVPR42600.2020.01017]

Yang W F, Zhang T Z, Yu X Y, Qi T, Zhang Y D and Wu F. 2021. Uncertainty guided collaborative training for weakly supervised temporal action detection//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 53-63 [DOI: 10.1109/CVPR46437.2021.00012]

Yang Z C, Qin J and Huang D. 2022. ACGNet: action complement graph network for weakly-supervised temporal action localization//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 3090-3098 [DOI: 10.1609/aaai.v36i3.20216]

Zhai Y H, Wang L, Tang W, Zhang Q L, Zheng N N, Doermann D, Yuan J S and Hua G. 2023. Adaptive two-stream consensus network for weakly-supervised temporal action localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4136-4151 [DOI: 10.1109/TPAMI.2022.3189662]

Zhai Y H, Wang L, Tang W, Zhang Q L, Zheng N N and Hua G. 2022. Action coherence network for weakly-supervised temporal action localization. IEEE Transactions on Multimedia, 24: 1857-1870 [DOI: 10.1109/TMM.2021.3073235]

Zhang C, Cao M, Yang D M, Chen J and Zou Y X. 2021. CoLA: weakly-supervised temporal action localization with snippet contrastive learning [EB/OL]. [2024-06-26]. https://arxiv.org/pdf/2103.16392.pdf

Zhou K Y, Yang J K, Loy C C and Liu Z W. 2022. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9): 2337-2348 [DOI: 10.1007/s11263-022-01653-1]

作者简介

任小龙,男,硕士研究生,主要研究方向为时序动作定位。
E-mail:a15563259217@stud.tjut.edu.cn
张飞飞,通信作者,女,教授,主要研究方向为计算机视觉。
E-mail:feifeizhang@email.tjut.edu.cn
周琬婷,女,副研究员,主要研究方向为医学图像分析。
E-mail:wanting.zhou@bupt.edu.cn
周玲,女,助理教授,主要研究方向为计算机视觉。
E-mail:lzhou@must.edu.mo