Modeling and Simulation 建模与仿真, 2025, 14(2), 236-244
Published Online February 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.142147

Hans汉斯

基于不确定性感知自适应伪标签的指代视频目标分割

张施明，陈智谦，米金鹏*

上海理工大学机器智能研究院，上海

收稿日期：2025年1月24日；录用日期：2025年2月17日；发布日期：2025年2月27日

摘　要

指代视频目标分割(Referring Video Object Segmentation, RVOS)是一项新兴的多模态任务，旨在通过理解给定指代表达的语义来分割视频片段中的目标区域。然而，基准数据集的标注是通过半监督方式收集的，仅提供了视频第一帧的真实目标掩码。为了在一个更综合的框架中探索未标记数据中的隐藏知识，本文引入了在线伪标签来解决RVOS问题。具体来说，使用之前训练阶段的即时学习检查点作为教师模型，在未标记的视频帧上生成伪标签，并将获得的伪标签用作训练数据的增强，以监督随后的训练阶段。为了避免伪标签带来的混淆，本文提出了一种不确定性感知的细化策略，根据模型预测的置信度自适应地修正生成的伪标签。本文在基准数据集Refer-YouTube-VOS和Refer-DAVIS$_{17}$上进行了广泛的实验来验证所提出的方法。实验结果表明，本文的模型与最先进的模型相比取得了具有竞争力的结果。

关键词

指代视频目标分割，伪标签，不确定性感知细化

Uncertainty-Aware Adaptive Pseudo-Labeling for Referring Video Object Segmentation

Shiming Zhang, Zhiqian Chen, Jinpeng Mi*

Institute of Machine Intelligence, University of Shanghai for Science and Technology, Shanghai

Received: Jan. 24th, 2025; accepted: Feb. 17th, 2025; published: Feb. 27th, 2025

Abstract

Referring video object segmentation (RVOS) is an emerging multimodal task aiming to segment

*通讯作者。

文章引用: 张施明, 陈智谦, 米金鹏. 基于不确定性感知自适应伪标签的指代视频目标分割[J]. 建模与仿真, 2025, 14(2): 236-244. DOI: 10.12677/mos.2025.142147

target regions in video clips by understanding the semantics of given referring expressions. While the annotations of the benchmark datasets are collected in a semi-supervised manner, which only provides the ground truth object masks on the first frame of videos. To explore the concealed knowledge in the unlabeled data in a more integrated framework, we introduce online pseudo-labeling to address RVOS. Specifically, we employ the on-the-fly learned checkpoints in the previous training epochs as the teacher model to produce the pseudo labels on the unlabeled video frames, and the obtained pseudo-labels are utilized as augmentation for the training data to supervise the subsequent training stage. To avert the confusion derived from pseudo-labels, we propose an uncertainty-aware refinement strategy to adaptively rectify the generated pseudo-labels based on the model prediction confidence. We conduct extensive experiments on the benchmark datasets Refer-YouTube-VOS and Refer-DAVIS$_{17}$ to validate the proposed approach. The experimental results demonstrate that our model achieves competitive results compared with state-of-the-art models.

Keywords

Referring Video Object Segmentation, Pseudo-Labeling, Uncertainty-Aware Refinement

1. 引言

指代视频目标分割(RVOS)建立在视频目标分割(Video Object Segmentation, VOS) [1] [2]的基础之上，它利用指代表达作为引导，通过理解给定指代表达的语义来对视频片段中的目标区域进行分割。作为理解视频中场景语义的一项重要任务，指代视频目标分割可广泛应用于视频编辑、视频检索、增强现实以及人机交互[3]。

与其他与指代表达式相关的多模态任务(例如指代表达定位[4] [5]和指代表达分割[6] [7])类似，在指代视频目标分割中，指代表达在分割视频序列所有帧中的目标区域时起着至关重要的作用。Khoreva 等人[8]探索了一种识别目标对象的替代方法，除了作为一种指出目标对象的更实用且更自然的方式之外，使用语言描述有助于避免偏差，还能使系统对复杂的动态变化和外观差异更具鲁棒性。Seo 等人[9]将视频和指称表达式作为输入，并估算出给定语言表达式在整个视频帧中所指代的目标掩码，恰当地结合两种注意力模型，联合执行基于语言的目标分割和掩码传播，以此来学习指代视频目标分割中的跨模态对应关系。虽然这些模型取得了不错的成果，但它们忽略了视频帧中区域候选的长时间关系。

与用于指代表达定位和指代表达分割的图像区域表达注释相比，在每个视频帧上收集目标掩码表达标签是一项耗时且昂贵的任务。为了减轻模型训练阶段繁琐的手动注释并减少工作量，方法[10]提出了一种两阶段、自上而下的指称视频目标分割解决方案，相关工作[11]采用一种新颖的多层次表示学习方法法探究视频内容的内在结构，以提供一组具有区分性的视觉嵌入，从而实现更有效的视觉－语言语义对齐，但都是以半监督的方式训练模型，该方式仅利用视频级多模态对应或第一帧注释上的真实值。半监督设置的核心挑战是如何有效地利用未标记样本中的有用信息。为了探索隐藏在未标记数据中的有用知识，本文借鉴了半监督学习方法[12]-[14]来解决 RVOS 问题。

伪标签[12]旨在借助预训练教师模型的引导，为未标注样本生成伪标签，生成的伪标签被用作训练数据的辅助部分，用于监督学生模型的训练。作为伪标签法的一种特定方案，自训练[15] [16]在训练阶段将自身作为教师来生成伪标签。受伪标签和自训练显著特性的启发，本文将在线伪标签法引入指代视频目

标分割中，以提升模型的分割性能。值得注意的是，这项工作是首次尝试通过伪标签来改进 RVOS。

本文为指代视频目标分割提出了一个基于在线伪标签法的框架。考虑到在视频帧上直接使用伪标签会加剧由类别间不平衡导致的模型偏差，本文提出了一种不确定性感知策略，以自适应地校正生成的伪标签。具体而言，首先将之前训练轮次中即时学习到的检查点用作教师，以便为未标注的视频帧生成伪标签。然后采用分割置信度作为指标，对预测置信度进行排序，并选择置信度较高的伪标签来扩充训练数据，用于监督后续的模型训练。通过自适应伪标签筛选，能够避免将不确定的标签引入模型训练中，并确保通过伪标签法提升模型性能。

这项工作的主要贡献总结如下：

(1) 本文通过在线伪标签提出了一个新颖的框架来解决视频目标分割(RVOS)问题，并且本项工作是首次尝试将伪标签引入视频目标分割领域。

(2) 本文提出了一种不确定性感知策略，该策略能够基于分割预测的不确定性自适应地校正生成的伪标签。

(3) 本文在广泛使用的基准数据集 Refer-YouTube-VOS [9]和 Refer-DAVIS$_{17}$ [8]上对所提出的方法进行了验证。实验结果证明了所提方法的有效性。

2. 实验方法

本文提出了一种具备不确定性感知的伪标签方法，以提升 RVOS 的效果。具体而言，将先前训练轮次所学到的检查点用作教师模型，为未标注的视频片段生成伪标签。为避免引入因预测确定性较低的伪标签所造成的噪声，本文提出了一种具备不确定性感知的伪标签优化策略，用以校正生成的伪标签。本文所提出的框架结构如图 1 所示。

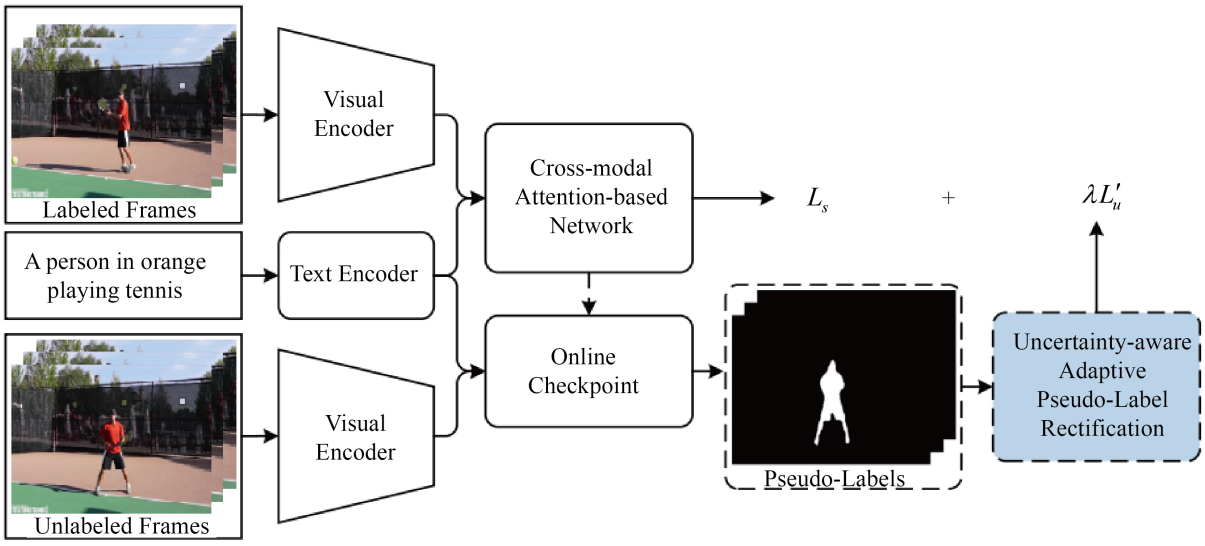

Figure 1. Overview of the adaptive pseudo-labeling
图 1. 自适应伪标签方法概述

2.1. 准备工作

RVOS 任务旨在通过理解给定的指代表达式 $\Lambda=\{e_i\}_{i=1}^{N}$ (包含 N 个单词)的语义，使用二值分割掩码 $O=\{o_i\}\in\mathbb{R}^{H\times W}$，在包含 M 帧的视频片段 $V=\{v_i\}_{i=1}^{M}$ 上对目标区域进行分割。

在实践中，对每个视频帧进行掩码到表达映射的注释并不容易，因此本文以半监督的方式分割目标区域，即利用每个视频的三个帧上的掩码表达注释进行模型训练。

最近发布的模型的主要策略是使用 Transformer 来提高模型性能。尽管这些模型取得了有希望的结果，但它们需要更强大的计算资源和更长的训练阶段。与基于 Transformer 的模型不同，本文采用联合指代视频目标分割(Unified referring video object segmentation, URVOS) [9]中引入的基线模型作为骨干，对视频帧中目标区域和指代表达之间的跨模态对应关系进行建模。URVOS 首先采用自注意层来获得视频帧和表达的联合特征表示，然后将联合特征表示馈送到跨模态注意模块中以获得跨模态注意特征 FC。为了捕获当前帧和前一帧的视觉表示之间的时间对齐信息，URVOS 开发了一个记忆注意力模块，以产生记忆注意力特征图 F_M。此外，F_C、F_M 和原始视频帧视觉特征图 F_V 是由特征金字塔解码器 D 处理以预测目标掩模。URVOS 通过 D 最小化定位真值掩码和解码 logit 之间的交叉熵(CrossEntropy)来训练模型。URVOS 的训练目标定义为：

$$\text{logit}_s = D\left(F_C, F_M, F_V\right) \tag{1}$$

$$L_s = \text{CrossEntropy}\left(\text{gt\_mask}, \text{logit}_s\right) \tag{2}$$

2.2. 在线伪标签

传统的伪标签利用静态教师模型在未标记的数据上生成伪标签，这需要一个预训练过程来获取教师模型。相比之下，本文的目标是在一个更完整的过程中，利用伪标签的未标记样本中的有用信息，在这个过程中，教师模型是在动态训练阶段学习的，而不是单独的训练程序。具体来说，本文使用在之前的训练周期中学习到的检查点作为教师来生成伪标签。随着模型训练的进行，可以基于更强大的教师模型逐步生成更好的伪标签。

给定一个包含 K 个标签对的训练集 $S_K = \left\{\left(v_i, o_i, e_i\right)\right\}_{i=1}^{K}$ 和 U 个未标签样本 $S_U = \left\{\left(v_i, e_i\right)\right\}_{i=K+1}^{K+U}$，使用在之前训练周期中学到的检查点 C 作为教师，为 S_U 生成伪掩码 $\tilde{o}_i$。计算无监督损失 L_u 如下：

$$\text{losit}_u = D\left(\tilde{F}_C, \tilde{F}_M, \tilde{F}_V\right) \tag{3}$$

$$L_u = \text{CrossEntropy}\left(\text{gt\_mask}, \text{logit}_u\right) \tag{4}$$

其中，$\tilde{F}_C$，$\tilde{F}_M$，$\tilde{F}_V$ 分别是获得的跨模态注意力特征、记忆注意力特征和选定的视频帧特征。使用生成的伪标签训练模型的最终损失定义为：

$$L = L_s + \lambda L_u \tag{5}$$

其中 λ 是一个超参数，用于平衡模型训练中 L_s 和 L_u 的贡献。

2.3. 不确定性感知伪标签细化

伪标签的原始假设是预训练的教师模型可以在未标记的数据上生成具有高预测置信度的伪标签，以增强原始训练数据。在此基础上，选择具有前 k 个最高预测样本的未标记数据作为后续训练阶段的辅助标记数据。另一方面，如果在所有未标签的帧上使用相同的置信度细化阈值过滤伪标签，则不可避免地会扩大模型对简单视频帧表达样本的偏差，或者带来新的噪声来降低模型性能。为了避免这个问题，本文提出了一种不确定性感知的伪标签细化策略，根据模型的预测不确定性自适应地校正生成的标签。

为了选择信息量最大的伪标签，本文建议根据掩模预测的不确定性自适应地校正生成的标签。受主动学习[17]的启发，主动学习采用模型预测的熵来衡量模型预测的不确定性。因此，可以通过以下方式计算模型预测不确定性来纠正标签：

$$\mu = \text{Entropy}\left(\varphi\left(\text{logit}_u\right)\right) \tag{6}$$

其中，φ 是 softmax 函数，用于将 logit 向量映射到概率分布。

基于获得的不确定性感知阈值，本文选择预测的 logit_s，logit_u，其熵值大于 μ 的作为精细化伪标签，以增强训练数据。因此，最终的训练损失由下式给出：

$$L'_u = \text{CrossEntropy}\left(\text{gt\_mask}, \text{logit}'_u\right) \tag{7}$$

$$L' = L_s + \mu L'_u \tag{8}$$

3. 实验

3.1. 数据集和评估指标

数据集：本文在两个基准数据集上进行了广泛的实验，即 Refer-YouTube-VOS 和 Refer-DAVIS$_{17}$，以验证本文提出的方法。Refer-YouTube-VOS 为从 YouTube-VOS [2]中选择的 3978 个视频收集了 15009 个相应的指代表达，并提供了两种类型的注释，即完整视频表达式和第一帧表达式。遵循其他 RVOS 模型，本文采用两种类型注释进行模型训练和验证。此外，Refer-YouTube-VOS 只发布训练集和验证集，因此也在验证分割上测试了本文的模型。Refer-DAVIS$_{17}$ 包括来自 DAVIS$_{17}$ [18]的 90 个视频，其中包含 1500 多个相应的指代表达。Refer-DAVIS$_{17}$ 包括 60 个视频的训练集，以及 30 个视频的验证集。

评估指标：与其它模型类似，遵循标准评估指标[19]来评估本文提出的方法。采用区域相似度(J) (%)来计算真实标签与预测分割之间的平均交并比(mean Intersection over Union，mIoU)，使用轮廓准确度(F) (%)来评估真实标签与预测之间的边界相似度，并使用它们的平均值($J\&F$) (%)同时评估区域相似度和轮廓准确度。

3.2. 实现细节

在视频帧中，通过随机采样选择五帧作为滑动窗口，以确保这五帧分布在视频帧的不同位置。在特定的训练周期后，只给出第一帧的真值，允许模型在完整的视频帧长度数据中传播和预测，并获得所有帧的伪标签。然后，选择前三个伪标签来替换原始随机采样的三帧作为训练数据。

本文采用 ResNet-50 [20]作为骨干网络来提取视频片段的深度特征表示。使用 Adam 优化器和初始学习率 1e-4 训练了总共 20 个 epoch 的模型，学习率在第十和第十五 epoch 衰减了 0.1。此外，采用 EMA [21]来提高模型的训练效率。

3.3. 与最先进的方法比较

为了评估所提出方法的性能，将本文的模型与采用 ResNet-50 作为主干网络并在 Refer-YouTube-VOS 和 Refer-DAVIS17 上训练模型的最优方法上进行了比较。

在 Refer-YouTube-VOS 上的比较：在 Refer-YouTube-VOS 数据集上，本文的模型与包括 RefVOS [22]、URVOS、CMPC-V [23]、YOFO [19]、LBDT [24]、VLIDE [25]、MLRLSA [11]和 Locator [26]在内的 SOTA 方法进行了比较。结果总结在表 1 中。

从表 1 中可以看出，本文的模型在三个指标上的性能优于其他模型。具体来说，本文的模型比基线模型 URVOS [19]分别提高了 8.55%、5.93%和 7.24%，并且超过了之前 SOTA 模型 Locator 分别提高了 5.02%、4.02%和 4.47%。根据比较结果，所提出的方法在 Refer-YouTube-VOS 数据集上比其他模型更有效。

在 Refer-DAVIS$_{17}$ 上的比较：在 Refer-DAVIS$_{17}$ 上，本文的方法与 Khoreava 等人[8]、URVOS、RefVOS、VLIDE、LBDT-4 和 MLRLSA (仅预训练)进行了比较。本文将结果列在表 2 中。

如表 2 所示，本文的模型在 Refer-DAVIS$_{17}$ 上实现了最佳的分割性能。所提出的方法在 J、F 和 J&F 指标上分别超越了基线模型 Khoreava 等人 13.55%、15.57%和 14.56%，并且在与 MLRLSA 相比时分别提高了 0.78%、1.48%和 1.13%。从表 1 和表 2 中列出的比较结果来看，很明显本文的模型提高了分割性能，并且在两个基准测试上表现更稳健。

Table 1. Performance (Acc%) on Refer-Youtube-VOS val set
表 1. 在 Refer-Youtube-VOS 验证集上的结果(准确率)

方法	来源	*J*	*F*	*J&F*
RefVOS	MTA 2023	39.50	-	-
URVOS	ECCV 2020	45.27	49.19	47.23
CMPC-V	TPAMI 2021	45.64	48.32	48.59
YOFO	AAAI 2022	47.50	49.68	48.59
LBDT	CVPR 2022	48.18	50.57	49.38
VLIDE	CVPR 2022	48.44	50.67	49.56
MLRLSA	CVPR 2022	48.43	50.96	49.70
Locator	TPAMI 2023	48.80	51.10	50.00
本文	-	**53.82**	**55.12**	**54.47**

Table 2. Performance (Acc%) on Refer-DAVIS17 val set
表 2. 在 Refer-DAVIS$_{17}$ 验证集上的结果(准确率)

方法	来源	*J*	*F*	*J&F*
Khoreava *et al.*	ACCV 2018	37.30	41.30	39.30
URVOS	ECCV 2020	41.23	47.01	44.12
RefVOS	MTA 2023	-	-	44.50
VLIDE	CVPR 2022	47.71	52.33	50.02
LBDT-4	CVPR 2022	-	-	**54.08**
MLRLSA	CVPR 2022	50.07	55.39	52.73
本文	-	**50.85**	**56.87**	53.86

本文在图 2 中列出了一些可视化结果，其中带有相关指称表达的正确分割样本位于虚线之上，而不正确的分割则列在虚线之下。

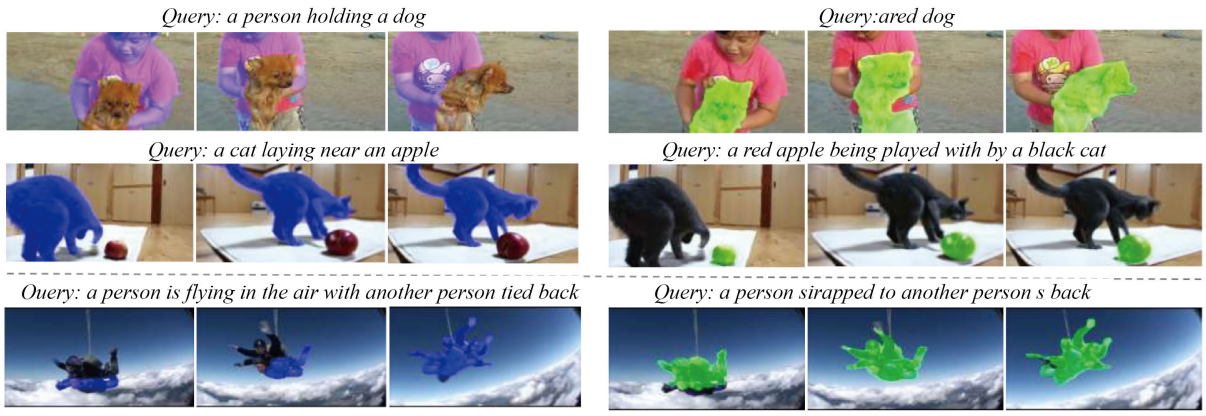

Figure 2. Qualitative visualization results acquire by the proposed approach on Refer-YouTube-VOS
图 2. 本文方法在 Refer-YouTube-VOS 上的定性可视化结果

3.4. 消融实验

为了验证所提出方法的好处，本文采用了三种不同的方式生成伪标签，即基于平均交并比(mIoU)、基于熵和不确定性感知的自适应伪标签。本文在两个基准数据集上进行了消融实验，并分别在表 3 和表 4 中总结了结果。

Table 3. Performance (Acc%) with different pseudo-label generation strategies on Refer-DAVIS17
表 3. 在 Refer-DAVIS17 上使用不同伪标签生成策略的性能(准确率)

模式	J	F	$J\&F$
mIoU	53.51	52.30	52.91
Entropy	50.16	48.39	49.28
Adaptive	53.45	52.95	53.20

基于 mIoU 的伪标签：本文首先使用 mIoU 作为生成伪标签的标准，并选择 mIoU 大于 0.5 的样本作为伪标签来扩充训练数据。

基于熵的伪标签：然后计算预测 logits 的熵，并选择熵值最小的三个样本作为伪标签。

不确定性感知的自适应伪标签：最后，本文学习一个自适应阈值来细化基于熵的伪标签，以提高模型的预测不确定性。本文采用 $\lambda = 1$ 时得到的结果与其他策略进行比较。

从表 3 可以看出，具有不确定性意识的自适应伪标签在 F 和 $J\&F$ 上分别超过了基于 mIoU 的策略 0.65%和 0.29%。从表 4 可以看出，具有不确定性意识的策略在三个指标上都优于其他策略。这些比较结果表明，所提出的具有不确定性意识的自适应伪标签在基准测试上更有效和稳健。

Table 4. Performance (Acc%) with different values of λ on Refer-YouTube-VOS
表 4. 在 Refer-YouTube-VOS 上使用不同 λ 值的性能(准确率)

模式	微调	J	F	$J\&F$
mIoU	×	46.69	50.70	48.69
	√	47.42	52.31	49.86
Entropy	×	41.69	46.03	43.86
	√	45.10	50.51	47.80
Adaptive	×	47.11	52.85	51.11
	√	50.85	56.87	53.86

3.5. 超参数设置

在本节中，本文通过设置方程(8)中权衡超参数 μ 的不同值来分析所提出的自适应伪标签方法的贡献。本文将 μ 设定为{0.05, 0.1, 0.5, 1.0, 5.0, 10.0}，并将获得的结果总结在表 5 中。

Table 5. Performance (Acc%) with different pseudo-label generation strategies on Refer-YouTuBe-VOS
表 5. 在 Refer-YouTuBe-VOS 上使用不同伪标签生成策略的性能(准确率)

μ	J	F	$J\&F$
0.05	32.97	23.67	28.32
0.1	36.82	26.69	31.75

续表

0.5	52.96	52.44	52.70
1.0	53.45	52.95	53.20
5.0	53.09	53.97	53.53
10.0	**53.82**	**55.12**	**54.47**

从表 5 可以看出，分割性能随着 μ 的变化而变化。当将 μ 设定为较小的值，例如，$\mu \in \{0.05, 0.1, 0.5\}$ 时，分割性能比 $\mu = 1$ 时更差。当使用 $\mu = 10$ 时，模型在三个指标上的分割性能优于 $\mu = 1$。因此，本文采用 $\mu = 10$ 获得的结果作为最佳结果，并在表 1 中与 SOTA 方法进行比较。

4. 结论

本文提出了一种基于不确定性感知的自适应伪标签方法来解决指代视频目标分割问题。所提出的架构利用在线伪标签来挖掘未标记样本中有用的信息，并使用模型预测置信度作为代理来改进生成的伪标签，以提高模型性能。具体来说，本文采用之前训练周期中学习到的检查点作为教师模型，在未标记样本上预测伪标签，然后使用生成的伪标签作为有标签训练数据的增强，以监督后续的训练过程。为了减轻伪标签引起的噪声影响，本文提出了一种模型预测不确定性感知策略来自适应地过滤生成的伪标签。此外，本文在基准数据集上进行了广泛的实验，实验结果证明了本文提出的方法对于指代视频目标分割的有效性。总而言之，基于模型不确定性的伪标签生成能促使模型开发更多有用的样本，而这些样本可以增强原始的数据，使得模型能学习到更多有益的知识，进而提升模型性能，为指代视频目标分割领域带来重要价值。

基金项目

国家自然科学基金(62106026, 62272170, 42130112)，上海市自然科学基金面上项目(23ZR1419300)。

参考文献

[1] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M. and Sorkine-Hornung, A. (2016) A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 724-732. https://doi.org/10.1109/cvpr.2016.85

[2] Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., *et al*. (2018) YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In: *Lecture Notes in Computer Science*, Springer, 603-619. https://doi.org/10.1007/978-3-030-01228-1_36

[3] Zhou, T., Porikli, F., Crandall, D.J., Van Gool, L. and Wang, W. (2023) A Survey on Deep Learning Technique for Video Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 7099-7122. https://doi.org/10.1109/tpami.2022.3225573

[4] Hu, R., Rohrbach, M., Andreas, J., Darrell, T. and Saenko, K. (2017) Modeling Relationships in Referential Expressions with Compositional Modular Networks. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 4418-4427. https://doi.org/10.1109/cvpr.2017.470

[5] Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., *et al*. (2018). MAttNet: Modular Attention Network for Referring Expression Comprehension. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1307-1315. https://doi.org/10.1109/cvpr.2018.00142

[6] Hu, R., Rohrbach, M. and Darrell, T. (2016) Segmentation from Natural Language Expressions. In: *Lecture Notes in Computer Science*, Springer, 108-124. https://doi.org/10.1007/978-3-319-46448-0_7

[7] Shi, H., Li, H., Meng, F. and Wu, Q. (2018) Key-Word-Aware Network for Referring Expression Image Segmentation. In: *Lecture Notes in Computer Science*, Springer, 38-54. https://doi.org/10.1007/978-3-030-01231-1_3

[8] Khoreva, A., Rohrbach, A. and Schiele, B. (2019) Video Object Segmentation with Language Referring Expressions. In: *Lecture Notes in Computer Science*, Springer, 123-141. https://doi.org/10.1007/978-3-030-20870-7_8

[9] Seo, S., Lee, J. and Han, B. (2020) URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In: *Lecture Notes in Computer Science*, Springer, 208-223. https://doi.org/10.1007/978-3-030-58555-6_13

[10] Liang, C., Wu, Y., Zhou, T., Wang, W., Yang, Z., Wei, Y. and Yang, Y. (2021) Rethinking Cross-Modal Interaction from a Top-Down Perspective for Referring Video Object Segmentation.

[11] Wu, D., Dong, X., Shao, L. and Shen, J. (2022) Multi-Level Representation Learning with Semantic Alignment for Referring Video Object Segmentation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orlean, 18-24 June 2022, 4986-4995. https://doi.org/10.1109/cvpr52688.2022.00494

[12] Li, H., Wu, Z., Shrivastava, A. and Davis, L.S. (2022) Rethinking Pseudo Labels for Semi-Supervised Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 1314-1322. https://doi.org/10.1609/aaai.v36i2.20019

[13] Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., *et al*. (2022) Semi-Supervised Semantic Segmentation Using Unreliable Pseudo-labels. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 4238-4247. https://doi.org/10.1109/cvpr52688.2022.00421

[14] Xu, Y., Shang, L., Ye, J., Qian, Q., *et al*. (2021) Dash: Semi-Supervised Learning with Dynamic Thresholding. *International Conference on Machine Learning*, Online, 18-24 July 2021, 11525-11536.

[15] Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., *et al*. (2019) Remixmatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring. 2019 *International Conference on Learning Representation*, New Orleans, 6-9 May 2019.

[16] Xie, Q., Luong, M., Hovy, E. and Le, Q.V. (2020) Self-Training with Noisy Student Improves ImageNet Classification. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 10684-10695. https://doi.org/10.1109/cvpr42600.2020.01070

[17] Settles, B. (2009) Active Learning Literature Survey. Computer Sciences Technical Report 1648.

[18] Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A. and Van Gool, L. (2017) The 2017 Davis Challenge on Video Object Segmentation.

[19] Li, D., Li, R., Wang, L., Wang, Y., Qi, J., Zhang, L., *et al*. (2022) You Only Infer Once: Cross-Modal Meta-Transfer for Referring Video Object Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **36**, 1297-1305. https://doi.org/10.1609/aaai.v36i2.20017

[20] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[21] Tarvainen, A. and Valpola, H. (2017) Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. 2017 *Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017.

[22] Bellver, M., Ventura, C., Silberer, C., Kazakos, I., Torres, J. and Giro-i-Nieto, X. (2022) A Closer Look at Referring Expressions for Video Object Segmentation. *Multimedia Tools and Applications*, **82**, 4419-4438. https://doi.org/10.1007/s11042-022-13413-x

[23] Liu, S., Hui, T., Huang, S., Wei, Y., Li, B. and Li, G. (2021) Cross-Modal Progressive Comprehension for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 4761-4775.

[24] Ding, Z., Hui, T., Huang, J., Wei, X., Han, J. and Liu, S. (2022) Language-Bridged Spatial-Temporal Interaction for Referring Video Object Segmentation. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 4954-4963. https://doi.org/10.1109/cvpr52688.2022.00491

[25] Feng, G., Zhang, L., Hu, Z. and Lu, H. (2022) Deeply Interleaved Two-Stream Encoder for Referring Video Segmentation.

[26] Liang, C., Wang, W., Zhou, T., Miao, J., Luo, Y. and Yang, Y. (2023) Local-Global Context Aware Transformer for Language-Guided Video Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 10055-10069. https://doi.org/10.1109/tpami.2023.3262578