中国科学: 信息科学 2024 年 第 54 卷 第 3 期: 674–691
SCIENTIA SINICA Informationis
论文

基于时空层级查询的指代视频目标分割

兰猛[1], 张乐飞[1*], 杜博[1], 张良培[2]

1. 武汉大学计算机学院国家多媒体软件工程技术研究中心, 人工智能研究院, 多媒体网络通信工程湖北省重点实验室, 武汉 430072
2. 武汉大学测绘遥感信息工程国家重点实验室, 武汉 430072
* 通信作者. E-mail: zhanglefei@whu.edu.cn

收稿日期: 2023–02–08; 修回日期: 2023–04–24; 接受日期: 2023–08–02; 网络出版日期: 2024–03–14

国家自然科学基金优秀青年科学基金 (批准号: 62122060) 和国家杰出青年科学基金 (批准号: 62225113) 资助项目

摘要 本文针对当前指代视频目标分割方法缺乏目标时空一致性建模和目标时空表征学习不足等问题, 进行了深入的研究, 提出了基于时空层级查询的指代视频目标分割方法 (STHQ). 本文将指代视频目标分割看作基于查询的序列预测问题, 并提出两级查询机制进行目标的时空一致性建模和时空特征学习. 在第 1 阶段, 本文提出了帧级空间信息提取模块, 该模块使用语言特征作为查询独立地和视频序列中的每一帧在空间维度进行信息交互, 生成包含目标空间信息的实例嵌入; 在第 2 阶段, 本文提出时空信息聚合模块, 该模块使用视频级的可学习查询嵌入和第 1 阶段生成的实例嵌入在时空维度进行信息交互, 生成具有时空表征信息的视频级实例嵌入; 最后, 视频级实例嵌入线性变换为条件卷积参数, 并和视频序列中的每一帧进行卷积操作, 生成目标的掩码预测序列. 在该领域的 3 个基准数据集上的实验结果表明, 本文提出的 STHQ 方法超越了现有的方法, 实现了最佳的性能.

关键词 指代视频目标分割, 时空一致性建模, 时空特征学习, 跨模态特征交互, Transformer

1 引言

指代视频目标分割 (referring video object segmentation, RVOS) 旨在分割出一个视频序列中由一段文字语言所描述的目标对象 [1]. 这个新兴的多模态任务吸引了许多研究者的关注和探究, 因为它为人机交互提供了一个更加自然和友好的方式, 是研究跨模态视频分割大模型的基础性技术, 也是实现“通用人工智能” 的关键技术之一 [2]. 指代视频目标分割有着广泛的应用场景, 例如在文本驱动的视频编辑应用中, 指代视频目标分割可以根据文本提示将指代目标从背景中提取出来, 然后根据需求进行相应的编辑. 在智慧安防应用中, 可以根据文本描述的提示来迅速检索指定的目标, 实现目标快速的定位和跟踪. 而在机器人领域, 则可以输入文本指令指挥机器人完成对指定目标的定位和各种复杂任务. 相比于传统的依赖第一帧目标像素级标注来提供目标的先验位置信息的半监督视频目标分割任

引用格式: 兰猛, 张乐飞, 杜博, 等. 基于时空层级查询的指代视频目标分割. 中国科学: 信息科学, 2024, 54: 674–691, doi: 10.1360/SSI-2023-0030
Lan M, Zhang L F, Du B, et al. Spatio-temporal hierarchical query for referring video object segmentation (in Chinese). Sci Sin Inform, 2024, 54: 674–691, doi: 10.1360/SSI-2023-0030

© 2024《中国科学》杂志社

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

务 [3,4] 和仅依靠目标对象显著性特征的无监督视频目标分割任务 [5,6], RVOS 任务不仅要处理多样性的语言描述和视觉内容之间的跨模态理解, 还要考虑视频任务中目标在时间维度上的动态变化, 因此更加具有挑战性.

数据是基于深度学习的算法模型的发展基石 [7], 因此为了推动指代视频目标分割领域的发展, 研究者们提出了多个 RVOS 数据集和对应的基线模型. 例如, Khoreva 等 [8] 在常用的半监督视频目标分割数据集 DAVIS 2017 [9] 的基础上, 通过标注目标对象对应的语言描述, 构建了一个新的用于 RVOS 任务的数据集 Ref-DAVIS17, 并提出了一个联合了指代图像定位模型 [10] 和图像分割模型的 RVOS 基准模型. Seo 等 [1] 在半监督视频目标分割数据集 YouTube-VOS 2019 [11] 的基础上, 通过标注目标对象对应的语言描述, 构建了 RVOS 领域第一个大规模的基准数据集, 有效推动了该领域的发展. 在上述数据集的支持下, 国内外的研究者们提出了多种 RVOS 算法模型, 并不断地刷新着性能上界. Seo 等 [1] 提出的端到端的 URVOS 模型将指代图像分割和半监督视频目标分割两种策略组合在一起, 生成准确的目标掩码预测. 其中, 指代图像分割独立地预测每一帧的目标掩码, 而基于半监督视频目标分割的掩码传播机制则用于确保目标在时序空间上的一致性. 然而指代图像分割生成的掩码预测质量严重影响着 URVOS 的最终性能. Wu 等 [12] 提出在视频级特征、帧级特征和目标级特征上分别进行跨模态语义对齐, 然后将多级特征进行融合后用于生成最终的目标掩码预测序列. 最近, 基于多模态 Transformer 架构的 RVOS 方法 [13,14] 凭借其显著的性能优势逐渐引起了研究者的关注. 这类方法将 RVOS 任务看作是序列预测任务, 采用类似于 DETR [15] 的基于查询的 Transformer 架构 [16], 生成候选目标预测序列, 并利用二分匹配算法建立预测序列和真值之间的关系, 从而实现对模型进行监督训练. 其中, Botach 等 [13] 提出的 MTTR 架构将视觉特征和语言特征级联后输入 Transformer 的编码器中生成多模态特征, 然后将可学习的目标查询嵌入和多模态特征输入到 Transformer 的解码器中生成包含目标信息的目标实例嵌入, 这些目标实例嵌入经过掩码头部网络变换为动态卷积参数并和视频帧特征进行卷积运算, 生成最终的目标掩码预测序列. Wu 等 [14] 则提出将语言特征作为目标查询嵌入, 在 Deformable DETR [17] 框架基础上进入跨模态交互, 生成包含目标信息的实例嵌入表征, 并也采用条件卷积生成最终的目标掩码预测序列.

基于多模态 Transformer 的 RVOS 方法凭借其在多个数据集上优异的性能证明了将 RVOS 任务看作基于查询的序列预测问题的优越性. 然而当前的方法 [13,14] 为了减少模型的计算消耗均没有考虑对视频序列的帧间一致性进行建模, 如图 1(a) 所示, 当前方法在 Transformer 中采用帧间独立的跨模态交互, 而帧与帧之间缺乏信息的交互, 使得生成的实例嵌入缺乏目标的时空表征信息, 可能导致最终生成的目标掩码预测在时空一致性上存在缺陷. 在本文中, 时空一致性指的是在单帧中对目标对象的空间掩码预测在多帧的时序维度上的一致性. 因此, 如何高效地建模目标时空一致性和学习目标的时空表征, 从而更加全面地解译视频序列中的目标信息, 提升视频目标分割结果的时空一致性和准确性, 是当前基于 Transformer 的 RVOS 方法亟待解决的难点问题.

为了解决当前基于 Transformer 的 RVOS 方法在时空一致性建模和时空表征学习上存在的问题, 构建一个高效准确的指代视频目标分割算法框架, 本文基于多模态 Transformer 架构提出了时空层级查询策略. 如图 1(b) 所示, 在一个端到端的架构内, STHQ 分两级逐步进行目标的时空一致性建模和时空表征学习, 从而提升最终的视频目标分割性能. 具体来说, 在第 1 阶段, 我们提出帧级空间信息提取模块, 使用语言特征作为目标查询嵌入独立地和视频序列中的每一帧在空间维度进行信息交互, 使得生成的实例嵌入包含目标在空间维度上的表征信息. 在第 2 阶段, 本文提出时空信息聚合模块, 该模块学习一个视频级的查询嵌入, 并将第 1 阶段生成的只具有目标空间表征信息的实例嵌入在时空维度展开, 视频级的查询嵌入和展开后的实例嵌入在时空维度进行信息交互, 生成具有时空表征信息的

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

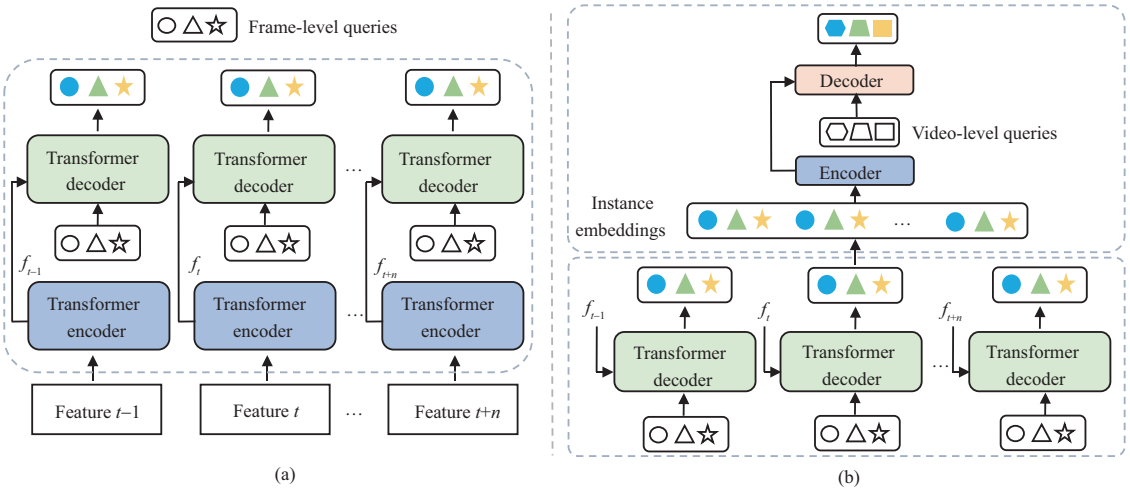


图 1 (网络版彩图) 基于查询的指代视频目标分割算法的对比. (a) 当前已有的 RVOS 方法; (b) 本文提出的基于时空层级查询的 RVOS 方法

Figure 1 (Color online) Comparison of query-based referring video object segmentation approaches. (a) Existing query-based methods; (b) our proposed spatio-temporal hierarchical query-based method

视频级实例嵌入, 从而实现对目标的时空一致性建模和时空表征学习. 同时, 由于避免了目标查询嵌入直接和图像帧在时空维度进行特征交互, 而是采用分级机制选择和特征维度更小的实例表征嵌入在时空维度进行信息交互, 从而显著减少了模型的计算负担, 提升了模型的运行效率. 最后, 视频序列中的每一帧共享由视频级实例嵌入线性变换生成的条件卷积核, 通过卷积运算生成视频序列的目标掩码预测序列. 代码: https://github.com/LANMNG/STHQ.

本文的主要贡献可总结如下:

• 提出了一种新颖的基于时空层级查询的指代视频目标分割算法, 解决了当前基于查询和多模态 Transformer 架构的 RVOS 方法缺乏目标时空一致性建模和时空表征学习不足等问题.

• 提出了两级查询机制, 以更加高效的形式构建了先空间后时空的目标时空一致性建模和时空表征学习体系.

• 在 3 个基准数据集上的实验结果表明本文提出的算法超越了现有方法实现了最佳性能, 证明了算法的有效性和先进性.

2 相关工作

2.1 指代视频目标分割

指代视频目标分割是一项非常具有挑战性的跨模态任务, 相比于以精细的目标掩码作为先验信息的半监督视频目标分割任务, RVOS 任务只能将形式灵活的语言描述作为目标的参考信息. 在指代视频目标分割的发展初期, 研究者们常常会从不同的角度来解决 RVOS 任务. 例如, Bellver 等 [18] 将 RVOS 任务看作是指代图像分割 [19] 在视频领域的扩展, 因此直接将指代图像分割算法应用在 RVOS 任务上, 但该算法忽略了目标在时序上的一致性, 导致算法的精度较低. Khoreva 等 [8] 提出先使用指代图像定位模型 [10] 定位出目标的边界框位置, 再在帧间增加目标边界框的时序一致性约束, 最后使用一个分割网络对目标进行分割. Seo 等 [1] 提出的 URVOS 模型将指代图像分割和半监督视频目标分

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

割两种方法联合在一起优化, 在指代图像分割的基础上引入了目标的时序关联性, 有助于生成准确的目标掩码预测. 然而上述两种 RVOS 算法过于依赖图像级目标处理算法的性能, 从而导致模型的性能依然受到限制. Wang 等 [20] 提出了一种非对称的交叉引导注意力网络来增强视觉和文本特征之间的交互, 提升最终的分割性能. Wu 等 [12] 提出多级表征学习机制, 在视频级、图像级和目标级 3 个层级分别进行视觉和文本的语义对齐, 最后将多级特征进行融合并生成目标预测结果. Ding 等 [21] 提出利用文本表征作为帧间目标信息传递的桥梁, 从而实现目标的时序一致性建模. 受视频实例分割 [22, 23] 和目标检测 [17, 24] 等领域基于查询的 Transformer 架构的启发, Botach 等 [13] 和 Wu 等 [14] 将 RVOS 任务看作序列预测问题, 并提出了相应的基于查询的解决方案, 显著地提高了 RVOS 的分割性能. 其中, Botach 等 [13] 提出的 MTTR 算法构建了一个多模态 Transformer 架构, 利用可学习的目标嵌入来检索并保存目标对象的表征信息, 然后将生成的实例嵌入转换为条件卷积核的参数并和图像特征进行卷积运算, 生成目标掩码预测序列. Wu 等 [14] 提出的 ReferFormer 算法以语言特征作为目标查询嵌入, 从而更加高效和准确地检索出图像帧中的目标信息. 然而上述基于多模态 Transformer 的方法在利用查询嵌入和图像特征进行交互时采取的都是帧间独立的方式, 忽视了对目标时序上的一致性建模, 虽然减少了计算消耗, 但缺乏帧间目标一致性建模导致的目标漂移和误判使得模型的性能下降. 因此, 本文针对当前基于查询的 RVOS 方法在目标时空一致性建模和时空表征上的不足等问题, 提出了基于时空层级查询的指代视频目标分割方法, 高效地建模目标的时空一致性, 充分地学习目标的时空表征, 使得模型实现了更好的视频目标分割性能.

2.2 基于查询的方法

基于查询的策略最早由目标检测算法 DETR [15] 提出, DETR 在 Transformer 的解码器中学习一个固定数量的目标查询集合, 用来建模图像帧中目标之间的位置关系以及图像的上下文特征, 实现了较好的目标检测性能, 并大力推动了目标检测领域的进一步发展 [17, 24]. 因为基于可学习的目标查询策略在目标特征建模上的优越性, 基于查询的策略被应用到计算机视觉的多个领域, 并取得了优异的性能. 例如, Cheng 等 [25] 在通用图像分割领域提出的 mask2former 框架在全景分割、实例分割和语义分割领域都刷新了性能的最佳值; Wang 等 [22] 提出的 VisTR 重构了基于查询的视频实例分割新范式, 性能大幅超越了传统方法. Sun 等 [26] 提出的 TransTrack 引入目标查询来检测新增目标, 实现了多目标视频跟踪的最佳性能. Yuan 等 [27] 提出的 PolyphonicFormer 利用目标查询建模深度和全景分割之间关系, 在深度感知视频全景分割领域取得了更加鲁棒和优良的性能. Misra 等 [28] 提出的 3DETR 模型则为基于点云的 3D 目标检测领域提供了新的角度和基线模型. Xu 等 [29] 提出的基于查询的可变形混合 Transformer 架构的 DeMT 算法则在多任务密集预测领域实现了较好的性能. 基于查询的策略在众多领域的广泛应用表明利用查询机制建模目标的表征信息和上下文关系的有效性, 因此本文延续了基于查询的策略, 构建了时空层级查询机制, 在指代视频目标分割领域实现了更好的性能.

3 本文算法

3.1 预备知识

多头注意力模块是由文献 [16] 提出的一种注意力计算模块, 可以有效建模特征中元素间的全局关系, 对于学习特征的全局表征具有重要作用, 被广泛地应用在自然语言处理和计算机视觉的各种任务

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

中. 多头注意力模块的计算过程为

$$\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}\left(\mathrm{head}_1,\ldots,\mathrm{head}_h\right)W^O, \tag{1}$$

head_i 的计算过程为

$$\mathrm{head}_i=\mathrm{Attention}(Q_i,K_i,V_i)=\mathrm{softmax}\left(\frac{Q_iK_i^{\mathrm{T}}}{\sqrt{d_k}}\right)V_i, \tag{2}$$

其中, h 为注意力头部的数量, $Q_i=QW_i^Q$, $K_i=KW_i^K$, $V_i=VW_i^V$. $W_i^Q\in\mathbb{R}^{d_{\mathrm{input}}\times d_k}$, $W_i^K\in\mathbb{R}^{d_{\mathrm{input}}\times d_k}$, $W_i^V\in\mathbb{R}^{d_{\mathrm{input}}\times d_v}$ 和 $W^O\in\mathbb{R}^{hd_v\times d_{\mathrm{input}}}$ 为可学习的参数.

多头注意力模块一般分为多头自注意力模块和多头互注意力模块. 多头自注意力模块中 Q,K,V 是经过同一个特征嵌入变换而来的, 可以相同, 而多头互注意力模块中 Q 和 K,V 可以不同.

多尺度可变形注意力模块为文献 [17] 中提出的计算效率更高收敛速度更快的多头注意力模块, 被广泛应用在计算机视觉领域的多种任务中. 多尺度可变形注意力模块的计算过程为

$$\mathrm{MSDeformAttn}\left(\boldsymbol{z}_q,\hat{\boldsymbol{p}}_q,\left\{\boldsymbol{x}^l\right\}_{l=1}^L\right)=\sum_{m=1}^M\boldsymbol{W}_m\left[\sum_{l=1}^L\sum_{k=1}^K A_{mlqk}\cdot\boldsymbol{W}'_m\boldsymbol{x}^l\left(\phi_l\left(\hat{\boldsymbol{p}}_q\right)+\Delta\boldsymbol{p}_{mlqk}\right)\right], \tag{3}$$

其中, $\boldsymbol{z}_q\in\mathbb{R}^C$ 为查询嵌入 (Q) 中索引值为 q 的特征嵌入, $\hat{\boldsymbol{p}}_q$ 为 $\boldsymbol{z}_q$ 的参考点坐标, $\left\{\boldsymbol{x}^l\right\}_{l=1}^L,\boldsymbol{x}^l\in\mathbb{R}^{C\times H_l\times W_l}$ 为输入的多尺度特征 (K). m 为注意力头部的总数 M 的索引, l 为输入特征总的层级数 L 的索引层级数, k 是采样点的索引. $\Delta\boldsymbol{p}_{mlqk}$ 和 A_{mlqk} 分别为第 m 个头部的第 l 层特征中的第 k 个采样点的位置偏移量和注意力参数, 均由 $\boldsymbol{z}_q$ 线性变换生成. ϕ_l 为尺度调整函数. $\boldsymbol{W}'_m\in\mathbb{R}^{C_h\times C}$ 和 $\boldsymbol{W}_m\in\mathbb{R}^{C\times C_h}$ 为可学习的参数, 且 $C_h=C/M$.

多尺度可变形注意力模块也可以分为多尺度可变形自注意力模块和多尺度可变形互注意力模块. 多尺度可变形自注意力模块中特征 Q 和 K 由相同的多尺度特征变换而来, 而多尺度可变形互注意力模块中的特征 Q 和 K 可以为不相同的特征, 但特征 K 必须为多尺度特征.

3.2 算法概览

给定一个包含 T 帧的视频序列 $\mathcal{V}=\{I_t\}_{t=1}^T$ 和一段包含 L 个单词的指代目标的语言描述 $\mathcal{E}=\{e_l\}_{l=1}^L$, 本文的模型旨在生成视频序列的每一帧中指代目标对象的掩码预测. 本文提出的 STHQ 算法的整体框架图如图 2 所示, STHQ 算法主要包含 4 个部分, 第 1 部分是特征提取模块, 图像编码器和文本编码器分别提取图像帧特征和语言文本特征, 并作为下一阶段的输入. 第 2 部分是 STHQ 方法中提出的层级查询机制的第 1 阶段, 即帧级空间信息提取模块, 在该模块中, 句子级文本特征作为语言查询嵌入和经过 Transformer 编码器编码后的图像帧特征在 Transformer 解码器中进行帧间独立的跨模态特征交互, 生成目标实例嵌入. 第 3 部分为层级查询机制的第 2 阶段, 即时空信息聚合模块, 在该模块中, 第 1 阶段生成的目标实例嵌入输入到实例编码器中并在时空维度上进行全局特征编码, 然后和可学习的视频级查询嵌入一起输入到实例解码器中, 生成视频级实例表征嵌入. 第 4 部分为实例序列分割模块, 视频级实例嵌入分别经过类别头部网络和掩码头部网络的线性变换后生成类别概率值和条件卷积核的参数. 条件卷积核和第 1 阶段 Transformer 编码器输出的经过特征金字塔网络的图像特征进行卷积操作, 生成目标掩码预测序列. 在训练阶段, 经过实例序列匹配机制实现预测序列和真值之间二分匹配, 并计算由分类损失函数和分割损失函数构成的模型总损失值, 通过梯度反向传播实现对模型的监督训练.

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

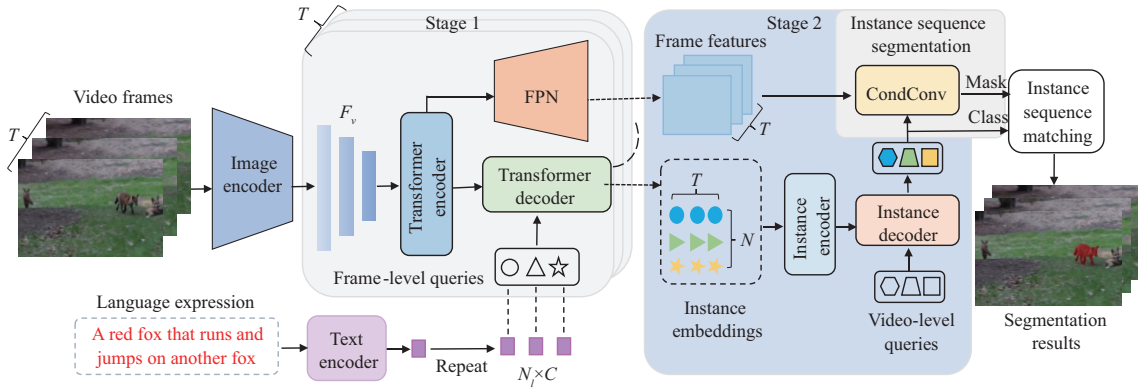


图 2 (网络版彩图) 本文提出的基于时空层级查询的指代视频目标分割算法整体框架. 在第 1 阶段, 句子级语言特征作为帧级查询嵌入和输入的视频序列中的每一帧在 Transformer 解码器中进行独立的特征交互, 生成目标实例嵌入. 在第 2 阶段, 实例编码器在时空维度上对实例嵌入进行编码, 输出的实例特征和可学习的视频级查询嵌入一起输入实例解码器中, 生成视频级实例嵌入, 视频级实例嵌入经过头部网络分别生成类别概率预测和条件卷积核. 最后视频序列中每一帧的特征和条件卷积核通过卷积运算生成目标掩码预测序列

Figure 2 (Color online) Overview of the proposed spatio-temporal hierarchical query-based referring video object segmentation method. In the first stage, the sentence-level language feature serves as the frame query embeddings and interacts with the frame features of the input video sequence independently, generating the instance embeddings. In the second stage, the instance encoder encodes the instance embeddings in the spatio-temporal dimension and the output instance feature along with the learnable video query embeddings is sent into the instance decoder to produce the video-level instance embeddings, which is converted to produce the class probability and the kernel weights of the conditional convolution via the head networks. Finally, the conditional convolutions are performed on the features of each frame in the video sequence and the video-level instance embeddings, which produce the mask prediction sequence of the target object

3.3 特征提取模块

特征提取模块包含两个针对不同模态输入的编码器. (1) 视觉编码器提取视频序列中每一帧的多尺度骨干特征 $\mathcal{F}_v = \{F_t\}_{t=1}^{T}$, F_t 包含了空间步长为 [8, 16, 32] 的多尺度骨干特征. 值得注意的是, 此处的视觉编码器可以是 2D 的空间编码器, 比如 ResNet [30], 也可以是 3D 的时空编码器, 例如 Video Swin Transformer [31]. 2D 的空间编码器独立地编码视频序列中的每一帧, 而 3D 编码器则在时空维度提取视频序列中每一帧的特征. (2) 文本编码器. 本文采用已有的文本编码模型 RoBERTa [32] 来编码输入的语言描述 $\mathcal{E} = \{e_l\}_{l=1}^{L}$, 分别生成句子级文本特征 $F_e^s \in \mathbb{R}^{1\times C}$ 和单词级文本特征 $F_e^w \in \mathbb{R}^{L\times C}$. 其中, 句子级文本特征 F_e^s 在时空层级查询中的第一阶段中作为语言查询嵌入输入到 Transformer 的解码器中, 而单词级文本特征 F_e^w 则在后续的特征金字塔网络中和视觉特征进行了跨模态交互.

3.4 帧级空间信息提取模块

在视频目标分割领域, 经典的 VisTR [22] 在 Transformer 的编码器和解码器中就在时空维度对图像特征进行编解码操作, 从而实现对目标的时空一致性建模和时空表征学习. 然而这种策略会造成巨大的资源消耗和计算负担, 导致模型的运行效率较低. 因此后续的相关工作一般会采用帧间独立编解码 [13, 14, 33], 并设计额外的模块用于增强帧间信息联系 [23, 25, 34]. 为了有效降低模型的计算消耗, 本文在时空层级查询机制的第 1 阶段设计了帧级空间信息提取模块对视频序列的图像帧进行独立的编码和基于语言查询的特征解码. 如图 2 中 Stage 1 所示, 帧级空间信息提取模块主要包含 3 个部分, Transformer 编码器、Transformer 解码器和一个特征金字塔网络 (feature pyramid network, FPN), 其中, Transformer 的编码器和解码器均采用了 Deformable DETR [17] 的编码器和解码器结构, 以实现高

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

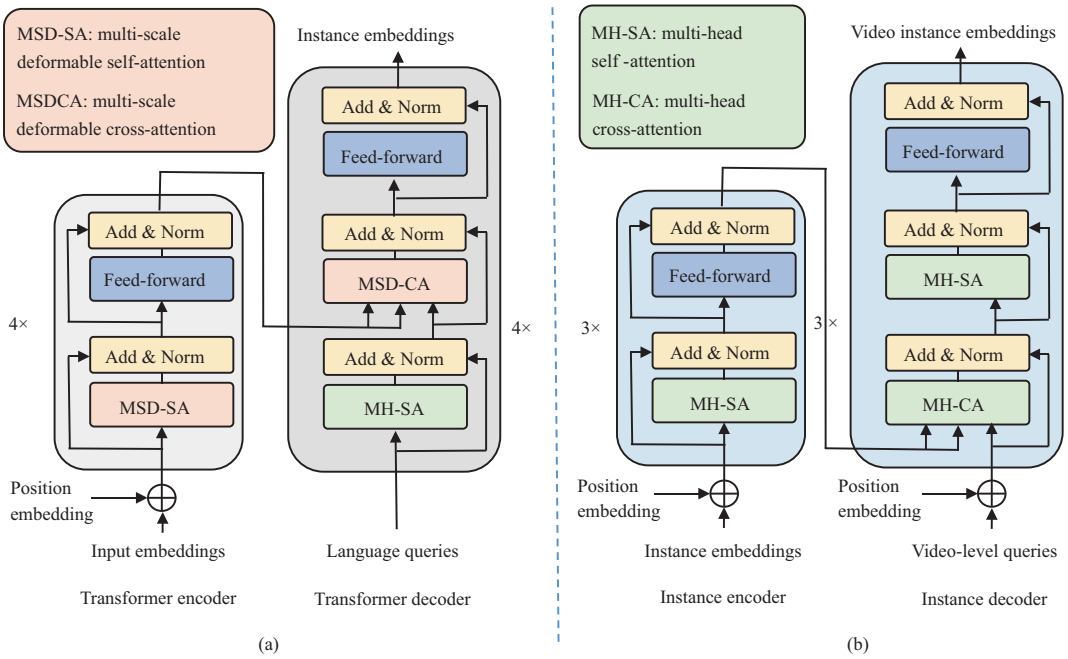


图 3 (网络版彩图) 时空层级查询机制中的编码器 – 解码器结构. (a) 帧级空间信息提取模块的编 – 解码器结构; (b) 时空信息聚合模块的编 – 解码器结构

Figure 3 (Color online) Encoder-decoder structures in the spatio-temporal hierarchical query strategy. (a) Encoder-decoder structure of frame-level spatial information extraction module; (b) encoder-decoder structure of spatio-temporal information aggregation module

效准确的全局像素级关系建模. 特征金字塔网络则采用了 ReferFormer [14] 中提出的跨模态特征金字塔网络用于生成跨模态图像帧特征 $\mathcal{F}_{\text{seg}} = \{f_{\text{seg}}^t\}_{t=1}^T$, 其中 $f_{\text{seg}}^t \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_d}$, H 和 W 为输入图像的高和宽.

Transformer 编码器由 4 个相同的编码层堆叠而成, 如图 3(a) 所示, 每一层包含一个多尺度可变形自注意力模块 [17] (multi-scale deformable self-attention module) 和一个全连接前馈网络 [16] (fully connected feed-forward network). 图像帧的多尺度特征 $\mathcal{F}_v = \{F_t\}_{t=1}^T$ 在输入到 Transformer 编码器前先由 1×1 的卷积层将通道数统一降到 $C = 256$, 并为每一帧的特征都加上固定的 2D 位置编码. 为了实现帧间独立并行运算, 视频图像特征将时序维度变换到了 batch 维度, 并将高度和宽度合并. 输入的多尺度特征经过 4 个编码层的特征学习后, 编码器的多尺度特征输出 $F_{\text{frame}}^{\text{en}}$ 将作为解码器的输入和语言查询嵌入进行解码操作.

Transformer 解码器由 4 个相同的解码层堆叠而成, 如图 3(a) 所示, 每一层包含一个多头自注意力模块 [16]、一个多尺度可变形互注意力模块 [17] 和一个全连接前馈网络, 如图 3(a) 所示. 在解码器部分, 引入 N_l 个目标查询嵌入来表征每一帧图像中可能的目标实例, 同时视频序列中的每一帧都有着相同的 N_l 个初始目标查询嵌入. 因为本模块以句子级文本特征作为目标查询嵌入, 因此将 $F_e^s \in \mathbb{R}^C$ 复制 $N_l \times T$ 份, 得到整个视频序列的目标查询嵌入 $Q_1^s \in \mathbb{R}^{T\times N_l\times C}$. Q_1^s 和编码器的输出 $F_{\text{frame}}^{\text{en}}$ 一起输入到 Transformer 解码器中, 在每一层中, Q_1^s 经过自注意力模块后和 $F_{\text{frame}}^{\text{en}}$ 在互注意力模块进行特征交互, 并经过前馈网络变换后得到每一层的输出. 在多尺度可变形互注意力模块中, 查询嵌入 Q_1^s 和多尺度特征 $F_{\text{frame}}^{\text{en}}$ 分别为式 (3) 中的查询嵌入 Q 和多尺度特征 K. 最后经过 4 层的解码运算得到 $N_q = T \times N_l$ 个实例嵌入, 即 $F_{\text{ins}} \in \mathbb{R}^{T\times N_l\times C}$.

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

3.5 时空信息聚合模块

由于帧级空间信息提取模块输出的目标实例嵌入 F_{ins} 的特征维度相较于图像特征已大幅降低, 同时目标实例嵌入中包含有语言描述所指代的目标的表征信息且没有密集图像帧特征的干扰, 因此, 在时空层级查询机制的第 2 阶段, 提出时空信息聚合模块, 在目标实例嵌入 $F_{\text{ins}} \in \mathbb{R}^{T\times N_l\times C}$ 上以较小的计算消耗建模目标的时空一致性和学习目标的时空表征, 并引入一个视频级目标查询嵌入和目标实例嵌入在时空维度进行特征交互, 生成视频级的目标实例嵌入, 用以预测最终的目标掩码序列. 如图 2 中 Stage 2 所示, 时空信息聚合模块主要包含 3 个部分, 实例编码器、实例解码器和实例序列分割模块.

实例编码器由 3 个相同的编码层堆叠而成, 如图 3(b) 所示, 每一层包含一个多头自注意力模块 [16] 和一个全连接前馈网络. 为了实现在目标实例嵌入的时空维度进行关系建模, 在输入实例编码器之前, 将实例嵌入的时空维度进行合并, 即 $F_{\text{ins}} \in \mathbb{R}^{T\times N_l\times C}$ 变换为 $F'_{\text{ins}} \in \mathbb{R}^{TN_l\times C}$. 实例嵌入 F'_{ins} 经过一个线性变换后输入到实例编码器中进行时空关系建模和目标的时空特征学习, 生成目标实例特征 $F^{\text{en}}_{\text{ins}} \in \mathbb{R}^{TN_l\times C}$. 值得注意的是, 由于 T 和 N_l 均较小, 实例编码器为了学习目标的全局时空信息, 并没有采用类似于 Swin Transformer [35] 的局部窗口注意力机制来降低计算量.

实例解码器由 3 个相同的解码层堆叠而成, 如图 3(b) 所示. 然而在每一层中, 实例解码器采用了不同于帧级空间信息提取模块的 Transformer 解码器所采用的自注意力模块、互注意力模块和前馈网络的排列顺序, 因为 Transformer 解码器的目标查询嵌入输入的是句子级文本特征, 而实例解码器中输入的可学习的视频级查询嵌入 $Q_2^v \in \mathbb{R}^{N_v\times C}$ 是随机初始化的 (N_v 为每一帧学习的目标嵌入数量), 不具有特定的特征表达能力, 先将 Q_2^v 输入到自注意力模块对解码过程意义不大. 因此本文调整了自注意力模块和互注意力模块的顺序, 让视频级查询嵌入 Q_2^v 先跟目标实例特征 $F^{\text{en}}_{\text{ins}}$ 在互注意力模块进行目标解码, 交互后的查询嵌入再经过自注意力模块和前馈网络的特征变换得到解码层的输出, 经过 3 个解码层后得到最后的视频级实例嵌入输出 $F^v_{\text{ins}} \in \mathbb{R}^{N_v\times C}$.

实例序列分割模块包含两个建立在实例解码器上的轻量级头部网络, 用来进一步转换解码器输出的视频级实例嵌入输出. 其中, 分类头部网络为一个线性变换层, 输出二元分类概率, 表明预测实例是否为文本语言所指代且在当前帧是可见的. 掩码头部网络为 3 个连续的线性层, 它的输出为 3 个 1×1 的条件卷积层 (conditional convolutions, CondConv) 的参数 [36], 其中前两个卷积层的输出通道为 8, 最后一个卷积层的输出通道为 1. 值得注意的是, 视频序列中的 T 帧共享视频级实例嵌入输出 F^v_{ins}, 即 F^v_{ins} 被复制 T 份得到 $N_q = N_v \times T$ 个实例嵌入序列后再输入到两个头部网络中, 从而为视频序列生成 N_q 个类别概率序列和 N_q 条件卷积层参数 $\Omega = \{\omega_i\}_{i=1}^{N_q}$. 由于条件卷积核是由视频级实例嵌入变换而来的, 包含有目标对象的信息, 因此本文将条件卷积核和在第 1 阶段生成的跨模态图像帧特征 $\mathcal{F}_{\text{seg}}$ 进行卷积操作, 并上采样生成和输入图像相同尺寸的二元分割掩码序列, 过程如式 (4) 所示:

$$\hat{s}_i = \text{Up}\left\{\hat{f}_i \circledast \omega_i\right\}_{i=1}^{N_q}, \tag{4}$$

其中, ω_i 为第 i 个条件卷积核, $\hat{f}_i$ 为与之相对应的图像帧特征, $\circledast$ 表示卷积操作. 将掩码预测序列按帧间顺序划分后形成了最终的视频序列的掩码预测结果 $\hat{S} \in \mathbb{R}^{T\times N_v\times H\times W}$.

3.6 实例序列匹配

在时空信息聚合模块中, 本文为视频序列中的 T 帧中的每一帧都生成了 N_v 个实例预测结果, 且保持着相同的顺序, 这使得可以将整个实例预测结果作为整体来使用实例匹配策略 [22] 对整个 RVOS

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

模型进行监督训练. 令 STHQ 的预测结果为 $\hat{y}=\{\hat{y}_i\}_{i=1}^{N_v}$, 第 i 个实例的预测结果可以表示为

$$\hat{y}_i=\left\{\hat{p}_i^t,\hat{s}_i^t\right\}_{t=1}^{T}, \tag{5}$$

其中, $\hat{p}_i^t$ 是一个概率值, 表明预测实例是否为文本语言所指代的目标且在第 t 帧是可见的. $\hat{s}_i^t\in\mathbb{R}^{H\times W}$ 为预测的二元分割掩码.

由于整个视频序列中只有一个文本指代的目标对象, 因此视频序列的真值实例序列可以表示为 $y=\{c^t,s^t\}_{t=1}^{T}$, 其中, c^t 为独热编码, 1 表示实例存在于第 t 帧, 否则为 0. 通过最小化式 (6) 的匹配损失, 本文可以找到最佳的实例预测结果作为正样本, 而将其他负样本作为空类, 通过最小化总的匹配损失实现模型的监督训练.

$$\hat{y}_{\text{pos}}=\underset{\hat{y}_i\in\hat{y}}{\arg\min}\mathcal{L}_{\text{match}}\left(y,\hat{y}_i\right), \tag{6}$$

其中,

$$\mathcal{L}_{\text{match}}\left(y,\hat{y}_i\right)=\lambda_{\text{cls}}\mathcal{L}_{\text{cls}}\left(y,\hat{y}_i\right)+\lambda_{\text{mask}}\mathcal{L}_{\text{mask}}\left(y,\hat{y}_i\right), \tag{7}$$

$\mathcal{L}_{\text{cls}}$ 是焦点损失函数 [37], $\mathcal{L}_{\text{mask}}$ 则是 DICE 损失函数 [38] 和二元掩码焦点损失函数按照 1:1 的权重组合而成的. 匹配损失在帧间独立计算, 并按照图像帧数量进行归一化处理.

3.7 推理过程

在推理时, 由于所有的图像帧都是共享相同的帧级语言查询嵌入和视频级目标查询嵌入, 所以本文提出的 STHQ 方法可以处理任意长度的视频序列. 对于较长的视频序列, 考虑到不同的运算平台的计算资源时, 可以将整个视频序列分割成多个视频片段, 逐步进行处理并最后整合在一起. 给定一个视频序列和对应的语言描述, STHQ 方法可以预测出 N_v 个实例序列, 对于每一个实例预测序列, STHQ 计算所有帧的类别概率值的平均值, 并得到一个均分集合 $\mathcal{P}=\{p_i\}_{i=1}^{N_v}$. STHQ 方法选择出类别概率均分最高的实例预测序列, 并选择该实例的分割掩码预测序列作为最终的分割结果.

4 实验

4.1 数据集和评价指标

数据集. STHQ 方法在 3 个基准数据集上实验, 分别是 Ref-YouTube-VOS 数据集 [1]、Ref-DAVIS17 数据集 [8] 和 A2D-Sentences 数据集 [39]. Ref-YouTube-VOS 数据集是一个大型的 RVOS 数据集, 其中训练集包含 3471 个视频和 12913 个语言描述, 验证集包括 202 个视频和 2096 个语言描述. Ref-DAVIS17 数据集是在 DAVIS17 数据集 [9] 的基础上通过给目标进行语言描述标注构建而成的. Ref-DAVIS17 的训练集和验证集分别包含 60 个和 30 个视频. 遵从文献 [13, 14] 的设定, STHQ 只采用验证集进行性能测试. 因为 Ref-DAVIS17 的验证集每个目标包含 4 个语言描述, 因此本文将 4 个语言描述对应性能评分的均值作为模型最后的性能评分. A2D-Sentences 数据集是通过为原始的 A2D 数据集 [40] 标注额外的语言描述构建而成的. A2D-Sentences 数据集是演员动作视频集, 包含 3754 个视频, 其中 3017 个训练集, 737 个测试集, 总共包含 6655 个关于视频中演员动作的语言描述, 每个视频中包含 3∼5 张带有像素级标注的图像帧.

评价标准. 按照标准的评价协议 [41], 本文采用区域相似性 $\mathcal{J}$, 轮廓精度 $\mathcal{F}$ 和它们的均值 $\mathcal{J}\&\mathcal{F}$ 作为 Ref-YouTube-VOS 和 Ref-DAVIS17 验证集的评价指标. 其中, 由于 Ref-YouTube-VOS 没有提供验

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

证集的真值标签, 因此需要将预测结果提交到官方服务器上进行性能验证. 对于 A2D-Sentences 数据集, 本文采用 precision@K, Ovrall IoU, Mean IoU 和 mAP 作为评价指标 [42].

4.2 实现细节

模型设计. 本文尝试了不同的视觉编码器, 包括 ResNet [30], Swin Transformer [35] 和 Video Swin Transformer [31], 而文本编码器使用的是 RoBERTa 模型 [32]. 帧间空间信息提取模块中引入的语言目标查询嵌入的数量 N_l 和时空信息聚合模块引入的视频级目标查询嵌入的数量 N_v 默认相同, 且都设置为 5.

训练细节. STHQ 模型的训练分为两个阶段. 在第 1 阶段, STHQ 模型在指代图像分割的数据集上进行预训练, 使用到的数据集包括 Ref-COCO [43], Ref-COCOg [43] 和 Ref-COCO+ [44]. 为了方便表述, 后文统一用 Ref-COCO 指代上述 3 个数据集. 在预训练阶段, 输入的图像帧数量 $T = 1$, 每张显卡的批处理大小为 4, STHQ 模型训练了 10 个周期, 其中在第 6 和 8 个周期学习率乘以系数 0.1. 在第 2 阶段, 本文针对不同的 RVOS 数据集采取了不同的微调策略. 对于 Ref-YouTube-VOS 数据集和 Ref-DAVIS17 数据集, 本文将预训练后的 STHQ 模型在 Ref-YouTube-VOS 训练集上微调 6 个周期, 其中在第 3 和 5 个周期将学习率乘以系数 0.1. 微调阶段, 每张显卡的批处理大小为 1, 每个输入的视频片段中包含 $T = 5$ 张从同一个视频序列中随机采样的图像帧, 同时包括随机水平翻转、随机裁剪和光度失真等数据增强策略会被应用在这些图像帧上. 所有的输入图像在被输入到模型前都会被调整尺寸大小, 尺寸的大小区间为 [360, 640]. STHQ 模型由 AdamW 优化器 [45] 进行优化, 视觉编码器的初始学习率设置为 5×10^{-5}, 文本编码器的初始学习率设置为 1×10^{-5}, 模型的其他部分的初始学习率设置为 1×10^{-4}. 对于 A2D-Sentences 数据集, 本文将预训练后的 STHQ 模型在 A2D-Sentences 训练集上进行微调, 微调策略和在 Ref-YouTube-VOS 训练集上微调的策略相同. 式 (7) 中不同损失函数的系数分别为 $\lambda_{\text{cls}} = 2$, $\lambda_{\text{mask}} = 5$. 本文采用了辅助损失函数用于监督时空层级查询的第一阶段生成的实例嵌入. 本文的所有实验都在同一台设备上进行, 模型的实现采用了 Pytorch 框架, 并在 4 块 A100 GPU 上进行分布式训练.

4.3 与现有方法的比较

本小节将介绍本文提出的 STHQ 方法和现有 RVOS 方法在 3 个基准数据集上的性能比较, 同时也会对本方法中的核心模块和超参数进行消融分析.

4.3.1 Ref-DAVIS17 验证集

本小节将本文提出的 STHQ 方法和最新的 RVOS 方法在 Ref-DAVIS17 验证集上的性能进行比较, 结果如表 1 [1,12,14,18,21,46] 所示. 根据使用的训练数据的不同, 性能的比较分为两个部分. 当仅使用 Ref-COCO 等指代图像分割数据集进行预训练时, 各方法在 Ref-DAVIS17 验证集上的性能比较如表 1 上半部分所示. 本文提出的 STHQ 方法在 $\mathcal{J}\&\mathcal{F}$ 上实现了 53.9% 的精度, 且分别比对比方法 ReferFormer [14] 和 MLSA [12] 高了 1.8% 和 5.7%. 由于只在指代图像分割数据集上训练而没有在 RVOS 数据集上进行微调, 因此该结果表明 STHQ 方法具有更好的泛化性. 当模型预训练并在 Ref-YouTube-VOS 训练集上进行微调时, 各方法在 Ref-DAVIS17 验证集上的性能比较如表 1 下半部分所示. 当骨干网络为 ResNet-50 时, STHQ 方法在 $\mathcal{J}\&\mathcal{F}$ 上达到了 59.6% 的精度, 比 ReferFormer 提高了 1.1%. 此外, 相比于在 Ref-DAVIS17 训练集上进行了额外微调的 MLSA [12], LBSTI [21] 和 URVOS [1] 等方法, 分别高了 1.7%, 5.3% 和 8%. 当骨干网络为更强的 Swin-Large Transformer (Swin-L) [35] 和时

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

表 1　本文方法和现有方法在 Ref-DAVIS17 验证集上的性能比较

Table 1　Comparison with state-of-the-art methods on Ref-DAVIS17 validation set[a)b)]

Method	Backbone	Training set	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS [1]	ResNet-50	Ref-COCO	44.1	41.2	47.0
RefVOS [18]	ResNet-50	Ref-COCO	45.1	–	–
MLSA [12]	ResNet-50	Ref-COCO	48.2	45.1	51.2
*ReferFormer [14]	ResNet-50	Ref-COCO	52.1	49.2	54.9
*STHQ	ResNet-50	Ref-COCO	**53.9**	**51.8**	**56.0**
CMSA + RNN [46]	ResNet-50	Ref-COCO+Ref-YTVOS	40.2	36.9	43.5
URVOS [1]	ResNet-50	Ref-COCO+Ref-YTVOS	51.6	47.3	56.0
*MLSA [12]	ResNet-50	Ref-COCO+Ref-YTVOS	52.7	50.0	55.4
LBSTI [21]	ResNet-50	Ref-COCO+Ref-YTVOS	54.3	–	–
MLSA [12]	ResNet-50	Ref-COCO+Ref-YTVOS	57.9	53.9	62.0
*ReferFormer [14]	ResNet-50	Ref-COCO+Ref-YTVOS	58.5	55.8	61.3
*STHQ	ResNet-50	Ref-COCO+Ref-YTVOS	**59.6**	**56.8**	**62.4**
*ReferFormer [14]	Swin-L	Ref-COCO+Ref-YTVOS	60.5	57.6	63.4
*STHQ	Swin-L	Ref-COCO+Ref-YTVOS	**61.3**	**58.4**	**64.2**
*ReferFormer [14]	Video-Swin-B	Ref-COCO+Ref-YTVOS	61.1	58.1	64.1
*STHQ	Video-Swin-B	Ref-COCO+Ref-YTVOS	**61.8**	**59.0**	**64.6**

a) * means no finetuning on Ref-DAVIS17 training set.

b) Bold indicates the best performance.

空编码器 Video-Swin-Base Transformer (Video-Swin-B) [31] 时, STHQ 方法在 $\mathcal{J}\&\mathcal{F}$ 上的精度分别为 61.3% 和最高的 61.8%, 并分别比相同设置下的 ReferFormer 高出了 0.8% 和 0.7%. 根据上述实验结果可知, 相比于基于传统编码器 – 解码器架构的 MLSA 和 LBSTI 等方法, STHQ 和 ReferFormer 等基于查询的方法在缺乏目标训练集微调的情况下, 依旧保持大幅的性能领先, 充分表明了基于查询的算法在网络架构上的先进性和算法优良的泛化性.

4.3.2　Ref-YouTube-VOS 验证集

本小节将介绍 STHQ 方法和最新的 RVOS 方法在更为复杂的 Ref-YouTube-VOS 验证集上的性能比较. 如表 2 [1, 12~14, 21] 所示, 当只使用 Ref-YouTube-VOS 训练集进行训练时, 以 Video-Swin-Tiny [31] 网络作为骨干网络的 STHQ 方法在 $\mathcal{J}\&\mathcal{F}$ 上实现了 57.0% 的精度, 比具有相同设置的 ReferFormer [14] 和 MTTR [13] 分别高出 1.2% 和 1.7%, 且在单项的 $\mathcal{J}$ 和 $\mathcal{F}$ 指标上均优于对比方法. 当使用 Ref-COCO 数据集预训练和 Ref-YouTube-VOS 训练集微调时, 本文测试了多种骨干网络下的 STHQ 方法的性能并和其他 RVOS 方法进行了比较. 如表 2 所示, 当骨干网络为 ResNet-50 时, STHQ 方法在 $\mathcal{J}\&\mathcal{F}$ 上的精度为 56.8%, 比相同设置下的 ReferFormer [14]、MLSA [12] 和 LBSTI [21] 等方法分别高出了 1.2%, 7.1% 和 7.4%. 当骨干网络分别为 Swin-Large, Video-Swin-Tiny 和 Video-Swin-Base 的 Transformer 网络时, STHQ 方法在 $\mathcal{J}\&\mathcal{F}$ 上分别实现了 63.2%, 61.3% 和最高的 63.5% 的精度, 比相同设置下的 ReferFormer 方法分别高出了 0.8%, 1.9% 和 0.6%. 上述结果表明本文提出的 STHQ 方法在不同的训练集和骨干网络设置下均超过了当前最新的 RVOS 对比方法, 达到了最佳的性能. 同时, 相比于缺乏目标时空关系建模和时空表征的 MTTR 和 ReferFormer 等基于查询的方法, 上述实验结果充分证明

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

表 2 本文方法和现有方法在 Ref-YouTube-VOS 验证集上的性能比较

Table 2 Comparison with state-of-the-art methods on Ref-YouTube-VOS validation set[a)]

Method	Backbone	Training set	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
MTTR [13]	Video-Swin-T	Ref-YTVOS	55.3	54.0	56.6
ReferFormer [14]	Video-Swin-T	Ref-YTVOS	55.8	54.8	56.9
STHQ	Video-Swin-T	Ref-YTVOS	**57.0**	**55.8**	**58.2**
URVOS [1]	ResNet-50	Ref-COCO+Ref-YTVOS	47.2	45.3	49.2
LBSTI [21]	ResNet-50	Ref-COCO+Ref-YTVOS	49.4	48.2	50.6
MLSA [12]	ResNet-50	Ref-COCO+Ref-YTVOS	49.7	48.4	51.0
ReferFormer [14]	ResNet-50	Ref-COCO+Ref-YTVOS	55.6	54.8	56.5
STHQ	ResNet-50	Ref-COCO+Ref-YTVOS	**56.8**	**55.8**	57.8
ReferFormer [14]	Swin-L	Ref-COCO+Ref-YTVOS	62.4	60.8	64.0
STHQ	Swin-L	Ref-COCO+Ref-YTVOS	**63.2**	**61.8**	**64.6**
ReferFormer [14]	Video-Swin-T	Ref-COCO+Ref-YTVOS	59.4	58.0	60.9
STHQ	Video-Swin-T	Ref-COCO+Ref-YTVOS	**61.3**	**59.8**	**62.8**
ReferFormer [14]	Video-Swin-B	Ref-COCO+Ref-YTVOS	62.9	61.3	64.6
STHQ	Video-Swin-B	Ref-COCO+Ref-YTVOS	**63.5**	**61.7**	**65.3**

a) Bold indicates the best performance.

表 3 本文方法和现有方法在 A2D-Sentences 测试集上的性能比较

Table 3 Comparison with state-of-the-art methods on A2D-Sentences test set[a)]

Method	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
Hu et al. [47]	VGG-16	34.8	23.6	13.3	3.3	0.1	47.4	35.0	13.2
ACAN [20]	I3D	55.7	45.9	31.9	16.0	2.0	60.1	49.0	27.4
CSTM [48]	I3D	65.4	58.9	49.7	33.3	9.1	66.2	56.1	39.9
CMPC-V [42]	I3D	65.5	59.2	50.6	34.2	9.8	65.3	57.3	40.4
MTTR [13]	Video-Swin-T	72.1	68.4	60.7	45.6	16.4	70.2	61.8	44.7
LBSTI [21]	ResNet-50	73.0	67.4	59.0	42.1	13.2	70.4	62.1	47.2
ReferFormer [14]	Video-Swin-T	82.8	79.2	72.3	55.3	19.3	77.6	69.6	52.8
STHQ	Video-Swin-T	82.4	**79.2**	**72.7**	**56.4**	**21.0**	**78.0**	**69.6**	**53.9**

a) Bold indicates the best performance.

了 STHQ 方法提出的时空层级查询机制在目标时空建模和表征上的有效性.

4.3.3 A2D-Sentences 测试集

本小节将 STHQ 方法和最新的 RVOS 方法在 A2D-Sentences 测试集上的测试结果进行对比, 结果如表 3 [13, 14, 20, 21, 42, 47, 48] 所示. 可以看到, 本文提出的 STHQ 方法在 P@0.6 ∼ P@0.9, Ovrall IoU, Mean IoU 和 mAP 等多项性能指标上均超越了所有的对比方法. 其中, STHQ 的 mAP 达到了 53.9%, 比采用相同骨干网络的基于查询的 ReferFormer 和 MTTR 方法分别高了 1.1% 和 9.2%, 同时性能也大幅领先于基于传统编码器 – 解码器架构的对比方法 LBSTI, CMPC-V [42] 和 CSTM [48] 等. 上述实验结果也证明了基于查询机制的 RVOS 算法的优越性以及 STHQ 方法提出的时空层级查询机制的有

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

表 4 不同查询策略对比实验

Table 4 Comparison of different query strategies[a)b)]

Strategy	FLSIEM	STIAM	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
ST-1	–	–	59.9	58.5	61.3	5.4
Frame-level	✓	–	59.0	57.5	60.4	10
STHQ	✓	✓	**61.3**	**59.8**	**62.8**	9.2

a) FLSIEM denotes frame-level spatial information extraction module, and STIAM means spatio-temporal information aggregation module.

b) Bold indicates the best performance.

效性.

4.3.4 模型分析

本小节介绍对 STHQ 方法的核心模块和重要超参数进行的消融分析, 实验结果均为以 Video-Swin-Tiny 为 backbone 的 STHQ 模型在 Ref-YouTube-VOS 验证集上的测试结果.

查询策略. 本文以 STHQ 为基础, 对比了 3 种不同的查询策略, 并对 STHQ 模型中的帧级空间信息提取模块和时空信息聚合模块进行了消融实验和分析, 实验结果如表 4 所示. 其中, ST-1 表示采用类似 VisTR [22] 的单阶段查询策略, 即将图像帧的时序和空间维度合并到一个维度, 使得在 Transformer 的编码器阶段模型可以直接学习目标的时空特征, 并在解码阶段让目标查询嵌入和目标的时空特征进行交互解码, 生成目标实例嵌入信息, 然后通过和 STHQ 相同的方式生成目标的掩码预测序列; Frame-level 表示 STHQ 模型只使用帧间空间信息提取模块提取帧级的目标实例嵌入信息, 然后利用和 STHQ 相同的条件卷积生成目标的掩码预测序列; STHQ 表示基于时空层级查询策略的完整 STHQ 模型. 由于时空信息聚合模块是基于帧级空间信息提取模块设计的, 无法进行单独验证, 因此要验证时空信息聚合模块对模型的性能, 可以通过对比完整的 STHQ 模型和 Frame-level 的模型的性能来体现时空信息聚合模块对 STHQ 模型性能的影响. 从表 4 可知, 本文提出的 STHQ 方法在 $\mathcal{J}\&\mathcal{F}$ 上实现了最佳的 61.3% 的精度, 且 FPS (frame per second) 达到了 9.2. ST-1 策略虽然在单阶段就实现了目标时空建模和特征学习, 但图像帧特征背景复杂, 特征学习和基于查询的解码过程难度相对较大, 同时对计算资源的消耗增多, 导致在相同的资源条件下, 模型在训练阶段使用的图像帧数 T 要比 STHQ 更小, 使得模型训练不够充分, 从而导致模型性能下降, $\mathcal{J}\&\mathcal{F}$ 精度比 STHQ 低 1.4%, 且模型的运行效率降低, FPS 只有 5.4. Frame-level 在解码生成实例嵌入信息的过程中缺乏对目标时空一致性的建模和学习, 导致 $\mathcal{J}\&\mathcal{F}$ 精度只有 59.0%, 比 STHQ 低 2.3%. 上述实验结果证明了本文提出的时空层级查询机制在目标时空关系建模和时空特征学习上的有效性和优越性.

特征金字塔. STHQ 中采用了跨模态特征金字塔网络生成本文增强的高分辨图像特征, 为了探究跨模态特征金字塔网络和标准的特征金字塔网络, 即移除掉跨模态特征金字塔网络中的跨模态融合模块对 STHQ 模型性能的影响, 本文测试了使用两种不同的特征金字塔网络时 STHQ 模型的性能, 结果如表 5 所示. 由表 5 可知, 在使用标准特征金字塔网络时, STHQ 模型的 $\mathcal{J}\&\mathcal{F}$ 精度为 59.8%, 比使用跨模态特征金字塔网络的 STHQ 模型的 $\mathcal{J}\&\mathcal{F}$ 精度降低了 1.5%. 这表明在特征金字塔网络中使用文本特征作为补充信息进行跨模态融合有利于精确的目标掩码预测.

训练帧数. 本文研究了在训练时不同的视频片段帧数 T 对 STHQ 模型的影响, 实验结果如表 6 所示. 当训练帧数为 1 时, 模型在训练阶段缺乏对视频时序关系的建模和表征学习, 因此 $\mathcal{J}\&\mathcal{F}$ 精度仅有 56.9%. 当训练的帧数逐渐增加时, 模型的性能也逐渐增加. 由于计算资源的限制, 实验中 STHQ

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

表 5 不同特征金字塔设置的对比实验

Table 5 Comparison of different FPN setting[a)]

Setting	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Standard FPN	59.8	58.2	61.4
Cross-modal FPN	**61.3**	**59.8**	**62.8**

a) Bold indicates the best performance.

表 6 不同训练帧数的对比实验

Table 6 Comparison of different training frames[a)]

Frame	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	56.9	55.4	58.4
3	58.4	57.0	59.8
5	**61.3**	**59.8**	**62.8**

a) Bold indicates the best performance.

表 7 不同目标查询数量的对比实验

Table 7 Comparison of different object query numbers[a)]

Query	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	60.4	59.0	61.9
3	60.9	59.4	62.4
5	**61.3**	**59.8**	**62.8**
7	61.0	59.5	62.4

a) Bold indicates the best performance.

表 8 不同文本编码器训练设置的对比实验

Table 8 Comparison of various text encoder training settings[a)]

Pretraining	Finetuning	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Frozen	Frozen	59.3	57.8	60.8
Frozen	Not Frozen	59.8	58.2	61.4
Not Frozen	Frozen	60.3	59.1	61.5
Not Frozen	Not Frozen	**61.3**	**59.8**	**62.8**

a) Bold indicates the best performance.

可以设置的最大训练帧数为 5, STHQ 的精度达到了最佳的 61.3%.

目标查询数量. 本文开展了多个实验来分析不同的目标查询数量对算法结果的影响, 测试了当目标查询数量分别为 1, 3, 5, 7 时 STHQ 的精度, 实验结果如表 7 所示. 由表 7 可知, 由于图像帧特征复杂的背景和目标不断变化的位置关系, 当目标查询嵌入的数量为 1 时, 对目标的表征和上下文信息学习不够充分, 使得 $\mathcal{J}\&\mathcal{F}$ 精度仅为 60.4%, 当目标嵌入数量逐渐增长到 5 时, STHQ 模型的 $\mathcal{J}\&\mathcal{F}$ 精度达到了最佳的 61.3%, 比目标查询嵌入的数量为 1 时 $\mathcal{J}\&\mathcal{F}$ 精度提升了 0.9%. 而当目标查询嵌入的数量增加到 7 时, 模型的精度开始下降, 说明目标嵌入数量为 5 时, STHQ 模型对于目标的表征和上下文学习达到了最佳的状态, 因此本文为了达到最佳的精度, 默认 STHQ 的目标查询数量为 5.

文本编码器训练设置. 当前基于查询的方法在模型的训练阶段对于文本编码器的训练设置各不相

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

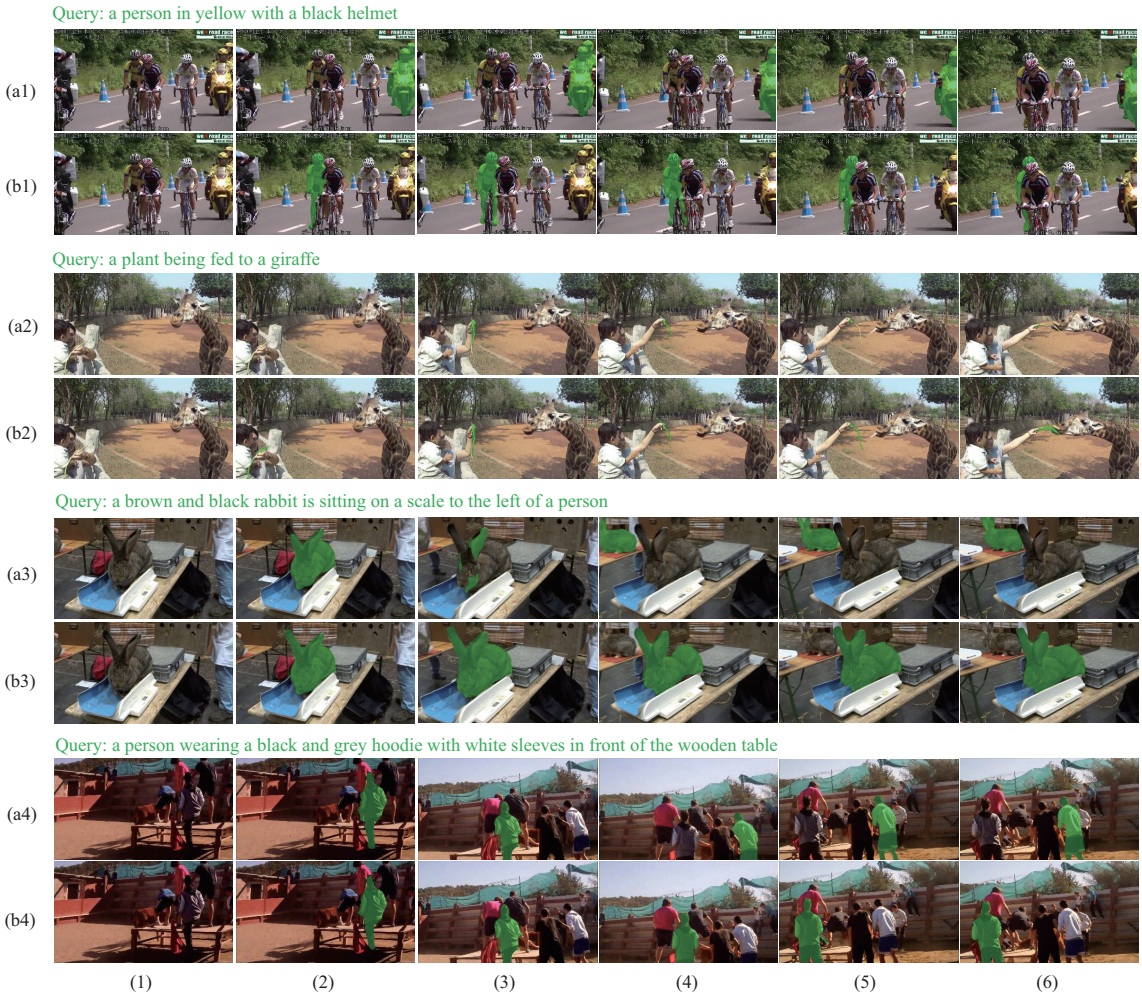


图 4　(网络版彩图) STHQ 和 ReferFormer 在 Ref-YouTube-VOS 验证集上的可视化结果对比图. (a1)∼(a4) ReferFormer 的可视化效果图, (b1)∼(b4) STHQ 的可视化效果图. Query 为目标描述语句. (1) 初始图像帧, (2)∼(6) 视频序列指代目标的分割结果

Figure 4　(Color online) Comparison of STHQ and ReferFormer visualization results on Ref-Yube-VOS validation set. (a1)∼(a4) Visualization results of ReferFormer, and (b1)∼(b4) visualization results of STHQ. Query is the description language of the target object. (1) Initial frame and (2)∼(6) segmentation results of the referring object in the video sequence

同, 如 MTTR 的文本编码器参数不参与模型的训练, ReferFormer 的文本编码器只参与预训练阶段的训练而不参与微调阶段的训练. 考虑到 Ref-COCO 数据集和 RVOS 数据集语言表述的特性以及和视频序列的相关性, 为了研究不同的文本编码器训练设置对 STHQ 模型的影响, 本文测试了多种训练设置组合的性能, 结果如表 8 所示. 由表 8 可知, 当文本编码器的参数在预训练和微调阶段均参与更新时, STHQ 的 $\mathcal{J}\&\mathcal{F}$ 精度达到了最佳的 61.3%, 而当文本编码器不参与模型训练而只使用预加载的参数进行特征提取时, STHQ 的 $\mathcal{J}\&\mathcal{F}$ 精度仅为 59.3%, 比两阶段参数均更新的 STHQ 模型精度降低了 2%. 只在预训练阶段和只在微调阶段进行文本编码器参数更新的 STHQ 模型的 $\mathcal{J}\&\mathcal{F}$ 精度分别为 60.3% 和 59.8%, 均比不更新文本编码器参数的模型精度高. 上述实验结果证明将文本编码器在新的语料库上进行参数更新是有效的, 且在预训练和微调阶段均进行参数更新的精度最佳.

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

4.3.5 可视化结果

图 4 展示了本文提出的具有目标时空表征能力的 STHQ 方法和缺乏目标时序关系建模和表征的 ReferFormer 方法的可视化效果对比图. 图 4(a1) 和 (b1) 的比较中, 图 4(a1) 中 ReferFormer 直接预测了错误的目标, 而图 4(b1) 中 STHQ 的预测结果不仅预测正确, 且在遮挡场景和细节等方面表现良好. 图 4(a2) 和 (b2) 的比较则显示 STHQ 方法对细小目标的预测效果要优于 ReferFormer. 图 4(a3) 和 (b3) 以及图 4(a4) 和 (b4) 的可视化结果比较则显示 ReferFormer 的目标分割在时序上缺乏一致性, 在视频序列后期会出现预测错误目标的情况, 而 STHQ 则能很好地在时序上分割同一个目标, 展现了 STHQ 在目标时空一致性和时空表征上的优异性能.

5 结论

本文针对当前指代视频目标分割方法在目标时空一致性建模和时空表征学习中存在的问题, 提出了一种基于时空层级查询的指代视频目标分割方法 STHQ. 通过帧级空间信息提取和时空信息聚合两个阶段的渐进式目标关系建模和表征学习, 生成了视频级的目标实例表征嵌入, 并通过和图像帧特征进行条件卷积生成最终的目标掩码预测序列. 提出的 STHQ 方法在 RVOS 领域的 3 个基准数据集上和现有的 RVOS 方法进行了性能比较, 精度超越了现有的方法, 证明了 STHQ 方法的优越性. 开展的消融实验也验证了所提出的时空层级查询机制的有效性. 未来, 我们将探究更为高效的目标时空建模和表征方式以及多模态预训练在 RVOS 任务上的有效性, 以进一步提升性能.

参考文献

1 Seo S, Lee J Y, Han B. URVOS: unified referring video object segmentation network with a large-scale benchmark. In: Proceedings of European Conference on Computer Vision, Glasgow, 2020. 208–223
2 Li X L. Multi-modal cognitive computing. Sci Sin Inform, 2023, 53: 1–32 [李学龙. 多模态认知计算. 中国科学: 信息科学, 2023, 53: 1–32]
3 Lan M, Zhang J, He F X, et al. Siamese network with interactive transformer for video object segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2022. 1228–1236
4 Lu X K, Wang W G, Danelljan M, et al. Video object segmentation with episodic graph memory networks. In: Proceedings of the European Conference on Computer Vision, Glasgow, 2020. 661–679
5 Lu X K, Wang W G, Shen J B, et al. Zero-shot video object segmentation with co-attention Siamese networks. IEEE Trans Pattern Anal Mach Intell, 2020, 44: 2228–2242
6 Zhao Z C, Zhang K H, Fan J Q, et al. Learning motion guidance for efficient unsupervised video object segmentation. Acta Autom Sin, 2023, 49: 1001–1009 [赵子成, 张开华, 樊佳庆, 等. 基于运动引导的高效无监督视频目标分割网络. 自动化学报, 2021, 49: 1001–1009]
7 Zhang B, Zhu J, Su H. Toward the third generation of artificial intelligence. Sci Sin Inform, 2020, 50: 1281–1302 [张钹, 朱军, 苏航. 迈向第三代人工智能. 中国科学: 信息科学, 2020, 50: 1281–1302]
8 Khoreva A, Rohrbach A, Schiele B. Video object segmentation with language referring expressions. In: Proceedings of Asian Conference on Computer Vision, Perth, 2019. 123–141
9 Pont-Tuset J, Perazzi F, Caelles S, et al. The 2017 DAVIS challenge on video object segmentation. 2017. ArXiv:1704.00675
10 Yu L C, Lin Z, Shen X H, et al. MAttNet: modular attention network for referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. 1307–1315
11 Xu N, Yang L J, Fan Y C, et al. YouTube-VOS: sequence-to-sequence video object segmentation. In: Proceedings of the European Conference on Computer Vision, Munich, 2018. 585–601
12 Wu D M, Dong X P, Shao L, et al. Multi-level representation learning with semantic alignment for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 4996–5005
13 Botach A, Zheltonozhskii E, Baskin C. End-to-end referring video object segmentation with multimodal transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 4985–4995

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

14 Wu J N, Jiang Y, Sun P Z, et al. Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 4974–4984
15 Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision, Glasgow, 2020. 213–229
16 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems, 2017
17 Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of International Conference on Learning Representation, 2021
18 Bellver M, Ventura C, Silberer C, et al. A closer look at referring expressions for video object segmentation. Multimed Tools Appl, 2023, 82: 4419–4438
19 Wang Z Q, Lu Y, Li Q, et al. CRIS: clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 11686–11695
20 Wang H, Deng C, Yan J C, et al. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019. 3939–3948
21 Ding Z H, Hui T R, Huang J S, et al. Language-bridged spatial-temporal interaction for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 4964–4973
22 Wang Y Q, Xu Z L, Wang X L, et al. End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 8741–8750
23 Hwang S J, Heo M R, Oh S W, et al. Video instance segmentation using inter-frame communication transformers. In: Proceedings of Advances in Neural Information Processing Systems, 2021. 13352–13363
24 Wang W, Zhang J, Cao Y, et al. Towards data-efficient detection transformers. In: Proceedings of the European Conference on Computer Vision, Tel Aviv, 2022. 88–105
25 Cheng B W, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 1290–1299
26 Sun P Z, Cao J K, Jiang Y, et al. TransTrack: multiple object tracking with transformer. 2020. ArXiv:2012.15460
27 Yuan H B, Li X T, Yang Y B, et al. PolyphonicFormer: unified query learning for depth-aware video panoptic segmentation. In: Proceedings of the European Conference on Computer Vision, Tel Aviv, 2022. 582–599
28 Misra I, Girdhar R, Joulin A. An end-to-end transformer model for 3D object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 2906–2917
29 Xu Y Y, Yang Y B, Zhang L F. DeMT: deformable mixer transformer for multi-task learning of dense prediction. 2023. ArXiv:2301.03461
30 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016. 770–778
31 Liu Z, Ning J, Cao Y, et al. Video Swin Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 3202–3211
32 Liu Y H, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. 2019. ArXiv:1907.11692
33 Wu J F, Jiang Y, Bai S, et al. SeqFormer: sequential transformer for video instance segmentation. In: Proceedings of the European Conference on Computer Vision, Tel Aviv, 2022. 553–569
34 Yang S S, Wang X G, Li Y, et al. Temporally efficient vision transformer for video instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 2885–2895
35 Liu Z, Lin Y T, Cao Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10012–10022
36 Tian Z, Shen C H, Chen H. Conditional convolutions for instance segmentation. In: Proceedings of European Conference on Computer Vision, Glasgow, 2020. 282–298
37 Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2980–2988
38 Milletari F, Navab N, Ahmadi S A. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the IEEE International Conference on 3D Vision, 2016. 565–571
39 Gavrilyuk K, Ghodrati A, Li Z, et al. Actor and action video segmentation from a sentence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 5958–5966
40 Xu C L, Xiong C M, Corso J J. Action understanding with multiple classes of actors. 2017. ArXiv:1704.08723
41 Perazzi F, Pont-Tuset J, McWilliams B, et al. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 724–732
42 Liu S, Hui T R, Huang S F, et al. Cross-modal progressive comprehension for referring segmentation. IEEE Trans

https://engine.scichina.com/doi/10.1360/SSI-2023-0030

Pattern Anal Mach Intell, 2021, 44: 4761–4775
43 Yu L C, Poirson P, Yang S, et al. Modeling context in referring expressions. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 69–85
44 Mao J H, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 11–20
45 Loshchilov I, Hutter F. Decoupled weight decay regularization. In: Proceedings of International Conference on Learning Representations, 2018
46 Ye L W, Rochan M, Liu Z, et al. Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 10502–10511
47 Hu R H, Rohrbach M, Darrell T. Segmentation from natural language expressions. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 108–124
48 Hui T R, Huang S F, Liu S, et al. Collaborative spatial-temporal modeling for language-queried video actor segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 4187–4196

Spatio-temporal hierarchical query for referring video object segmentation

Meng LAN[1], Lefei ZHANG[1*], Bo DU[1] & Liangpei ZHANG[2]

1. *National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, Hubei Key Laboratory of Multimedia and Network Communication Engineering, School of Computer Science, Wuhan University, Wuhan* 430072*, China*;
2. *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan* 430072*, China*
* Corresponding author. E-mail: zhanglefei@whu.edu.cn

Abstract In this paper, we propose a spatio-temporal hierarchical query-based referring video object segmentation (RVOS) method, called STHQ, to address the problems of lack of spatio-temporal consistency modeling and insufficient learning of spatio-temporal representation of the target in existing RVOS methods. We view the RVOS task as a query-based sequence prediction problem, and propose a two-level query mechanism for spatio-temporal consistency modeling and feature learning of the target. In the first stage, we devise the frame-level spatial information extraction module, which adopts language features as the query to interact independently with each frame of the video sequence in the spatial dimension, and generate instance embeddings containing spatial information about the target. In the second stage, we propose a spatio-temporal information aggregation module. The module uses the video-level learnable queries to interact with the instance embeddings generated in the first stage in the spatio-temporal dimension, and produces the video-level instance embeddings with spatio-temporal representation information. Finally, the video-level instance embeddings are linearly converted into the parameters of conditional convolution, which is used to perform convolution with the features of each frame in the video sequence and generate the mask prediction sequence of the target. The experimental results on three benchmarks show that our proposed STHQ outperforms the existing approaches and achieves state-of-the-art performance.

Keywords referring video object segmentation, spatio-temporal consistency modeling, spatio-temporal feature learning, cross-modal feature interaction, Transformer

https://engine.scichina.com/doi/10.1360/SSI-2023-0030