运动提示引导自适应学习无监督视频目标分割

韩志冬,胡升龙,宋慧慧*,张开华
(南京信息工程大学自动化学院,江苏南京 210044)

摘 要: 现有无监督视频目标分割(Unsupervised Video Object Segmentation,UVOS)方法多采用像素级密集匹配策略,通过对齐融合多帧之间或单帧与光流之间的信息来提升模型性能. 然而,在遮挡、相机抖动、运动模糊等挑战性场景中,光流估计误差易产生大量错误匹配,导致融合后的时空表征易过拟合运动噪声. 为此,本文提出一种运动提示引导的自适应学习 UVOS 框架. 通过设计一种无监督光流提示生成算法,将光流编码的密集运动信息转换为稀疏点和框提示,借助提示学习引导分割一切模型(Segment Anything Model,SAM)通过本文设计的两个轻量级适配器来自适应学习,从而获得更为鲁棒的时空表征,增强模型的抗噪能力. 为获得有效的提示,设计了一种无监督运动提示生成算法. 该算法基于光流特征计算一系列统计量,筛选出显著区域,再利用运动边缘信息去除伪显著区域的干扰,并设定自适应阈值进行过滤,生成提示显著运动目标所在区域的点和框坐标. 为提升SAM在下游UVOS任务中的泛化性,提出一种自适应表征学习SAM模型. 通过设计两个轻量级特征适配器,从SAM的通用知识库中自适应学习与下游UVOS任务相关的知识,以准确地粗定位目标. 针对SAM基于纯Transformer架构在细节处理上的不足,基于卷积神经网络(Convolutional Neural Networks,CNN)架构设计了表观聚焦细化模块. 由SAM得到的定位注意力图渐进式地引导细化过程,使模型的注意力从全局粗定位聚焦到局部细化,最终得到更加精确的分割掩码. 本文方法在DAVIS16(DAVIS 2016)、FBMS(Financial and Business Management System)和YTOBJ(YouTube-OBJects)三个主流数据集上进行了充分验证. 结果表明:本文方法在区域相似度指标上较当前先进方法分别提升了1.8%、1.6%和2.6%,充分表明了本文方法的有效性.

关键词: 无监督视频目标分割(UVOS);光流噪声;分割一切模型(SAM);提示学习;自适应表征学习;运动外观解耦;多模态

基金项目: 国家自然科学基金(No.62276141)

中图分类号: TP391.41 **文献标识码:** A **文章编号:** 0372-2112(2025)07-2305-19

电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250138

Motion-Prompts Guided Adaptive Learning for Unsupervised Video Object Segmentatio

HAN Zhi-dong, HU Sheng-long, SONG Hui-hui*, ZHANG Kai-hua
(*School of Automation*, *Nanjing University of Information Science and Technology*, *Nanjing*, *Jiangsu* 210044, *China*)

Abstract: Existing unsupervised video object segmentation (UVOS) methods often employ pixel-level dense matching strategies to enchance model performance by aligning and fusing features among multiple frames or between a single frame and its corresponding optical flow. However, in challenging scenarios such as occlusion, camera shak, and motion blur, optical flow estimation errors can easily generate numerous erroneous matches, leading to overfitting of the fused spatio-temporal representations to motion noise. To address this issue, we propose a motion-prompts guided adaptive learning UVOS framework. By designing an unsupervised motion-prompts generation algorithm, the dense motion information encoded by optical flow is transformed into sparse point and box prompts. With the help of prompt learning, the segment anything model (SAM) is guided to adaptively learn through two lightweight adapters designed in this paper, thereby obtaining more robust spatio-temporal representations and enhancing the model's noise resistance capability. To obtain effective prompts, we design an unsupervised motion-prompt generation algorithm. This algorithm calculates a series of statistical measures from the optical flow features to identify salient regions, then utilizes motion edge information and an adaptive

收稿日期:2025-02-26;录用日期:2025-06-30;责任编辑:覃怀银
*通讯作者:宋慧慧

threshold to eliminate pseudo-salient regions, ultimately generating the point and box coordinates that highlight the locations of motion-salient objects. To enhance the generalization ability of SAM in downstream UVOS tasks, an adaptive representations learning SAM model is proposed. By incorporating two light-weight feature adapters, the model adaptively extracts knowledge relevant to the downstream UVOS task from SAM's general knowledge base, enabling accurate coarse localization of objects. To overcome the lack of attention to details in pure Transformer-based SAM, a convolutional neural networks (CNN)-based feature focusing refinement module guided by the location map is designed. The localization attention map generated by SAM progressively guides the refinement process, shifting the model's focus from global coarse localization to local refinement, and ultimately producing more accurate segmentation masks. Our method has been thoroughly validated on three mainstream datasets: DAVIS 2016 (DAVIS16), financial and business management system (FBMS), and YouTube-Objects (YTOBJ). Compared with current state-of-the-art methods, our approach achieves improvements of 1.8%, 1.6%, and 2.6% in the region similarity metric, respectively, thereby fully demonstrating the effectiveness of our proposed method.

Key words: unsupervised video object segmentation (UVOS); optical flow noise; segment anything model (SAM); prompt learning; adaptive representation learning; decouple appearance-motion learning; multi-modality

Foundation Item(s): National Natural Science Foundation of China (No.62276141)

1 引言

视频目标分割(Video Object Segmentation,VOS)是计算机视觉领域中一项极为重要且极具挑战的任务,已广泛应用于自动驾驶、运动分析、目标检测和视频编辑等多个场景[1~4]. 根据定义方式的不同,VOS可分为半监督VOS(第一帧掩码引导)、无监督VOS(无引导,Unsupervised VOS,UVOS)、弱监督VOS(初始目标框引导)、交互式VOS(人机交互引导)和参考VOS(文本引导)等[5~9]. 其中,UVOS需在无任何先验提示的前提下,完全依靠算法自身从视频序列中分割出显著运动的前景目标,同时需克服运动不显著、遮挡、目标外观变化大等挑战,并保证分割目标在时间上的连贯性,因此更具挑战性.

传统UVOS方法[10~13]主要借助运动信息、边缘信息或显著信息来分割目标. 文献[11]在每一帧中生成候选区域,并利用运动信息增强这些候选区域,随后通过相邻帧的候选区域扩展当前帧的候选区域,以此来表示主要对象最可能出现的区域. 文献[13]利用光流信息,通过计算最小栅栏距离,获取目标的显著性先验信息和边界信息,并以此为线索扩散运动信息. 然而,由于传统算法过度依赖人工设计的特征,只能提取到一般性的先验知识,因此在复杂场景中的分割结果较粗糙.

近年来,深度学习在UVOS领域取得了突破性进展,大量相关工作涌现. 根据运动信息利用方式的不同,这些方法可分为两类设计范式:第一类是隐式利用运动信息范式[14~18],旨在通过挖掘帧间运动目标的相关性信息来隐式地利用运动信息;第二类是显式利用运动信息范式[6,19~22],通过光流估计得到光流图,其表示连续视频帧之间每个像素点的运动矢量,能够有效捕捉目标的时序动态特征,编码密集运动信息,从而直接显式利用运动信息. 第一类范式主要采用的模型包括图卷积神经网络(Convolutional Neural Networks,CNN)[23]、孪生网络[24]和3D神经网络[25]等. 文献[15,16]将UVOS定义为在视频图上进行信息迭代融合的过程,通过图结构传递任意帧之间的关系,从而能够挖掘视频帧之间更加丰富的关系,甚至可以跨帧获取相关性,进而更准确地捕获目标. 文献[14]强调视频帧的内在相关性,借助孪生网络架构,设计协同注意力模块以捕捉两帧之间目标的协同显著性信息. 文献[17,18]将3D卷积应用于UVOS,通过学习强判别力的时空表征来生成高质量分割掩码.

上述工作通过学习密集匹配的方式对齐并融合连续或任意顺序视频帧中显著运动目标的表征. 如图1(a)所示,在遮挡挑战场景下,这种密集匹配策略容易产生大量错误匹配,将遮挡物(如树干)误认为是目标的一部分,从而导致遮挡部分被错误分割. 此外,由于这些方法没有显式利用运动信息(如光流图),在解码过程中,如果目标背景较为嘈杂,会分散针对显著运动目标的注意力,难以充分挖掘显著运动信息.

为此,如图1(b)所示,第二类范式通过设计双流网络[26],分别处理视频帧和光流信息,从而显式利用运动信息. 两分支之间的特征通过密集匹配的方式融合外观与运动信息,生成鲁棒的时空表征. 然而,尽管显式利用运动信息在一定程度上提升了性能,但目标静止、遮挡等场景下的错误光流估计会引入运动噪声. 当这类噪声与外观特征进行密集匹配时,生成的时空表征会显著降质. 例如,如图1(b)所示,光流图中前车轮的估计缺失导致在融合匹配过程中前车轮的分割失败.

为了解决密集匹配在有限训练数据下抗噪能力弱和易受噪声干扰,进而对模型学习到的现有偏见过拟合的问题,本文提出了运动提示引导分割一切模型

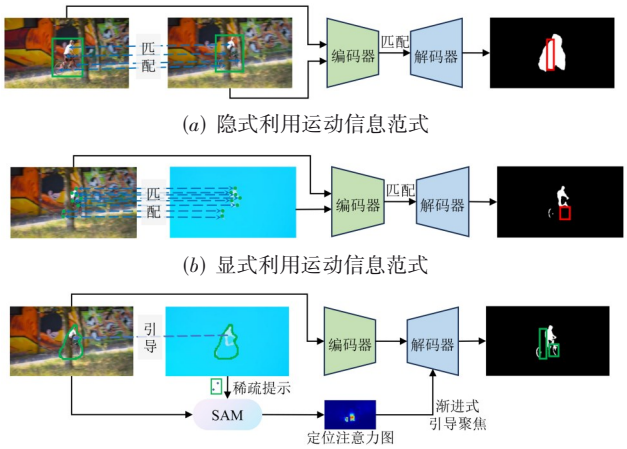
(Segment Anything Model,SAM[27])进行自适应表征学习UVOS框架. 如图1(*c*)所示,通过将密集运动信息转换为稀疏点和框提示的方式,避免引入过多噪声,从而降低对光流估计质量的要求. 具体而言,本文的主要贡献包括:首先,设计了一种无监督光流提示生成算法,通过对光流特征计算一系列统计量来对显著性连通区域进行打分;随后,利用运动边缘信息进一步去除伪显著区域的干扰,并通过设定自适应阈值对得分区域进行过滤,生成提示显著运动目标所在区域的点和框坐标;其次,提出了一种自适应表征学习的SAM模型,用于粗定位待分割目标. 通过设计两个轻量级特征适配器,从SAM编码的知识库中自适应学习与下游UVOS任务相关的知识,以提升SAM在UVOS任务中的泛化性,而无需付出高成本获取大量高质量训练数据;最后,设计了一个由SAM定位结果引导的表观聚焦细化模块,将SAM输出的定位注意力图用于逐级渐进式引导局部细化,使模型注意力从全局粗定位聚焦到局部细化,即模型先根据提示信息在全局对待分割前景目标的位置和范围进行初步定位,而后模型的注意力仅需聚焦于目标而无需分散在全局区域,从而精确地对目标的完整结构和边界进行精细调整,最终实现了更加精确的目标分割结果. 整个网络模型结构简单,无需精心、复杂设计,在包含DAVIS16(DAVIS 2016)[28]、FBMS(Financial and Business Management System)[2]和YTOBJ(YouTube-Objects)[29]在内的三个主流测试数据集上都达到了目前最先进的性能. 本文的主要贡献总结如下:

(1)提出了一种实现UVOS任务的新范式,旨在解耦外观和运动信息的学习,充分学习到不同模态的有效表征,有效避免了运动噪声对外观的干扰. 通过大量实验对比发现,本文方法在运动不显著场景下,效果提升最为明显,从而为解决UVOS相关问题提供一个新的解决视角.

(2)设计了一种无监督光流提示生成算法,将光流图这种密集运动信息转化为稀疏的点和框坐标提示信息,用于提示基础大模型SAM进行目标定位. SAM会根据提示优先定位出提示范围内最为显著的目标. 不同于以往的匹配模式,这类稀疏提示能够避免大部分运动噪声,从而减少对运动噪声的敏感度,提升模型抗噪能力.

(3)设计了两个轻量级自适应表征学习适配器,旨在将SAM中经过海量数据训练获得的大量通用知识,经过适配器微调,转换为更加适合本文特定下游任务的知识,极大地提升了SAM在UVOS任务上的泛化性.

(4)设计了一个由SAM定位图引导的聚焦模块和轻量化卷积注意力模块组成的表观聚焦细化模块. SAM粗定位到目标之后,仅需经过一个编码器和本文设计的表观聚焦细化模块,即可让模型的注意力从全局粗定位聚焦到局部细化,从而获得更加精确的分割结果.



(*a*) 隐式利用运动信息范式

(*b*) 显式利用运动信息范式

(*c*) 稀疏点和框提示范式(本文方法)

图1　本文方法与主流范式对比图

2　相关工作

2.1　UVOS

传统UVOS工作[10~13]较为依赖人工设计的低级特征,将运动、边缘、显著性、候选区域、测地线技术、点轨迹等作为线索. 然而,这些特征的表征能力有限,难以自适应复杂场景.

随着深度学习在UVOS领域的成功应用,其应对复杂场景的能力得到了极大提升. 文献[15,16]设计了图神经网络来记忆视频序列,通过设计节点之间的信息传递和聚合模块或频繁读写内存来存储更多帧的信息,从而挖掘视频帧之间丰富的相关性,取得了良好的分割结果. 文献[14]强调视频帧之间存在强相关性,借助孪生网络分别提取不同视频帧表征,并通过设计协同注意力模块来计算帧间的相关性,以分割出具有连贯区域的结果. 然而,上述工作未显式利用运动信息,在面对复杂场景下嘈杂背景的干扰时,难以充分挖掘显著运动信息. 为此,文献[6,19~22]显式使用光流图编码运动信息,通过设计双流网络分别提取并融合外观和运动信息,生成鲁棒的时空表征,实现了领先的分割效果. 其中,文献[19]设计了运动衰减模块来抑制运动信息的表达;文献[6]设计了特征对齐模块和特征自适应模块来对齐两种模态的特征,并有选择地表达运动信息;文献[21]则通过设计数据读取策略来减少模型对运动信息的过度依赖. 这些工作关注到了运动噪声对外观信息的干扰,但抑制运动信息的同时,也限制了模型性能的进一步提升.

2.2 SAM与提示学习

SAM[27]凭借其海量的高质量训练数据，以及点、框、掩码、文本等多种形式的语义提示设计，具备了强大的零样本泛化性能. 在无需微调的前提下，SAM能够在多种图像分割任务上取得先进的性能. 因此，SAM被众多研究者视为一个强大的外部知识库，并被广泛应用于图像分割、图像编辑、目标检测、图像标注、视频跟踪等[30~33]诸多领域，且均取得了良好的效果.

生成有效的提示对SAM在新任务上的高泛化性至关重要，因此，提示学习吸引了大量研究者的关注. 文献[30]首次将SAM成功扩展到医学图像分割领域，并通过标记放射学中的最长直径来获取框提示. 文献[31]为克服像素级数据标注成本高的难题，利用现有遥感检测数据集中的位置和类别提示，生成了一个基于SAM的大规模遥感图像分割数据集. 文献[32]受提示学习启发，提出一种学习类别提示嵌入的SAM提示新方法，用于生成遥感图像的语义分割结果. 文献[33]也提出了一种新的提示类型，通过引入一个视频显著目标检测网络，自动为前景对象生成轨迹提示，引导SAM逐帧生成视频掩码.

2.3 基础模型微调自适应学习

大模型自适应学习主要包括模型部分参数微调[34,35]和设计适配器微调[36~42]两种范式，旨在让大模型更好地适应下游任务.

前者在训练过程中解冻原模型部分参数，并随训练过程更新该部分参数，使模型更好地适应下游任务需求. 例如，文献[34]冻结住SAM图像编码器中部分参数，通过更新提示编码器和掩码解码器部分参数进行模型微调，让SAM更好地适应下游任务需求. 文献[35]则是通过有选择地更新图片编码器中的部分参数和掩码解码器的参数来实现类似目的.

后者在不更新原模型参数的基础上，通过设计轻量级适配器，将大模型编码的通用知识编码为下游任务所需特定知识，从而更好适应下游任务. 文献[36]在文本编码器后加入跨模态多层线性感知机(MultiLayer Perceptron，MLP[43])适配器模块，并在图像编码器的后端加入密集注意力适配器模块，通过微调两个适配器，从而学习到更加高效的下游知识. 文献[37]在原模型的MLP分支旁并行加入一个适配器分支，通过冻住原MLP分支参数，只更新并行适配器分支，从而学习有效的下游任务知识.

3 本文方法

如图2所示，本文方法主要包括运动提示定位模块(3.1节)和表观聚焦细化模块(3.2节)两部分. 其中，运动提示定位模块包含运动提示生成模块和自适应表征学习SAM两部分.

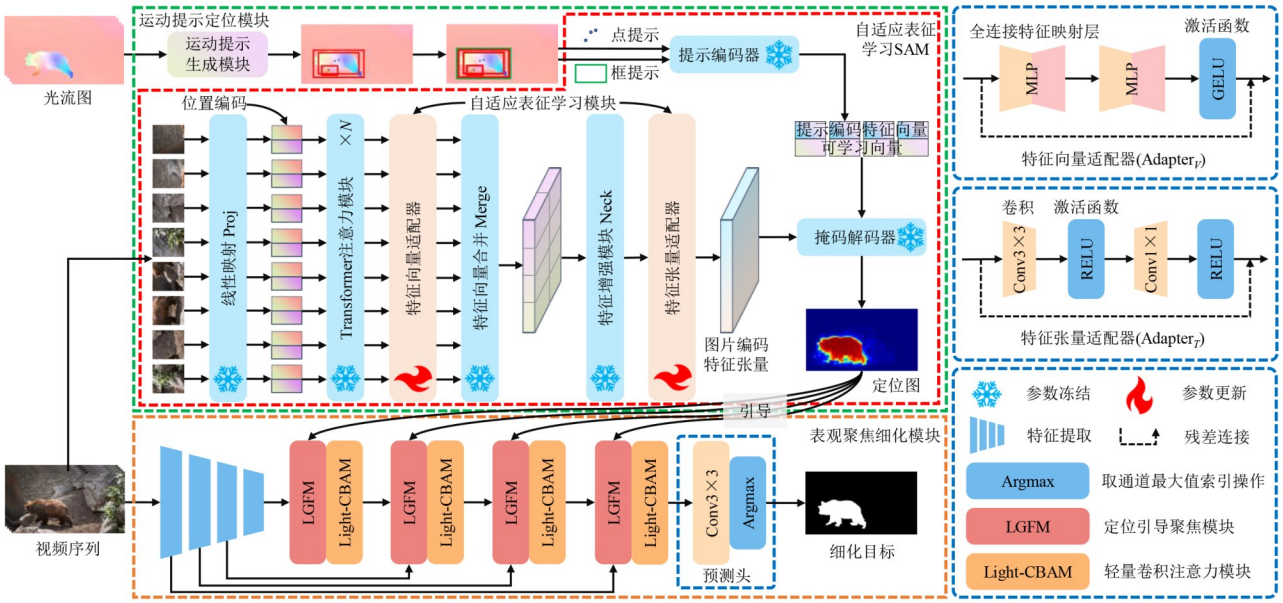


图2 本文方法网络结构图

首先，给定时刻t的视频帧$\boldsymbol{I}_t \in \mathbb{R}^{H\times W\times 3}$和其对应的光流图$\boldsymbol{F}_t \in \mathbb{R}^{H\times W\times 3}$，其中，$H$、$W$、3分别代表图片高度、宽度与通道数. 首先，对$\boldsymbol{F}_t$计算一系列均值、最大值等统计量，得到运动显著性图，并获取显著性得分最高的前k个连通区域的框坐标$R_t \in \mathbb{R}^{k\times 4}$以及点坐标$P_t \in \mathbb{R}^{k\times 2}$. 其中，框坐标为左上角及右下角横纵坐标值，点坐标采用框内有效像素点的重心横纵坐标值. 其次，将R_t和P_t作为提示输入SAM，得到粗定位结果L_t. 最

后,用L_t逐级渐进式引导表观聚焦细化模块聚焦目标,关注细节并解码预测,得到最终输出掩码M_t.

3.1　运动提示定位模块

3.1.1　运动提示生成模块

相关研究[30,44]表明,SAM在下游任务上的性能受制于提示信息的准确性. 为了使SAM充分适应下游UVOS任务,本文设计了一种运动提示生成算法,在单目标和多目标场景中均能准确计算待分割目标的提示信息,详见算法1和图3. 首先,通过对光流图计算均值、最大值等一系列统计量,对显著性连通区域打分,得到显著运动特征图;其次,通过运动边界约束,过滤掉部分伪显著区域;最后,将显著性得分最高的前k个运动显著区域的外接矩形框和重心坐标作为输出,用于提示SAM分割提示范围内的目标.

SAM具备强大的零样本(Zero-Shot)泛化性能,只要提示覆盖目标的部分区域,即可获得准确的定位结果. 算法1中步骤6得到的矩形框也可用于提示,然而,该矩形框易受噪声干扰,从而影响提示的准确性. 具体原因在于,步骤9利用运动边缘信息过滤掉部分伪显著区域,如果直接使用步骤6得到的矩形框,在部分场景中会引入错误或冗余的提示信息. 其中一个代表性场景如图3所示. 在该场景下,由于拍摄时相机的运动,通过步骤6得到的矩形框虽然能够很好地过滤掉非显著目标(树)区域,但由于左侧部分树叶边缘信息也被估计出来,步骤6得到的矩形框虽然包含了需要分割的整体目标,但相对于主体目标偏大,容易引入噪声. 在实验过程中,本文也尝试直接使用步骤6得到的矩形框作为提示,但其性能相较于使用步骤10~16得到的所有显著区域外接矩形框有所下降,因此最终未采用该策略.

算法1　运动提示生成算法

输入:光流图

输出:前k个高分框和点坐标

1. 分别计算$\boldsymbol{F}_t$三个通道所有像素均值$\{m_i\}_{i=1}^3=\{\boldsymbol{F}_t[:,:,i]\}_{i=1}^3$
2. $\boldsymbol{F}_t$减去每个像素所在通道的均值,再求平方,最终得到距离图$\mathcal{D}_t[:,:,i]=(\boldsymbol{F}_t[:,:,i]-m_i)^2$
3. $\mathcal{D}_t$沿通道维度求和后再开算数平方根,得到显著运动响应图$D=\sqrt{\mathrm{sum}(\mathcal{D}_t)}$
4. 计算光流图的灰度图$G_t=\mathrm{Gray}(\boldsymbol{F}_t)$
5. 利用Canny算子提取运动边缘信息$B_t=\mathrm{Canny}(G_t)$
6. 根据边缘图B_t计算得到一个外接矩形框R_o,包含所有边缘像素点
7. 计算D中的最大值和平均值$d_{\max}=\max(D), d_{\mathrm{mean}}=\mathrm{mean}(D)$
8. 计算阈值$\tau=\theta*d_{\max}+(1-\theta)*d_{\mathrm{mean}}$,并保留所有符合$D\geqslant\tau$的像素点组成的连通区域,记作Areas　　// 实验中θ取值为0.15
9. 利用步骤6得到的矩形框R_o约束Areas,只保留R_o范围内的连通区域,记作areas
10. 初始化三个空数组scores, boxes, points = [], [], [],分别存放显著目标的显著性得分、框坐标和点坐标
11. FOR area in areas
12. 　计算area的得分,并加入到scores中
13. 　计算area的外接矩形框坐标,并加入到boxes中
14. 　计算area的重心点坐标,并加入到points中
15. END FOR
16. 根据scores对points和boxes进行降序排序,得分最高的前k个框坐标R_t和点坐标P_t为最终返回结果.
17. return R_t,P_t

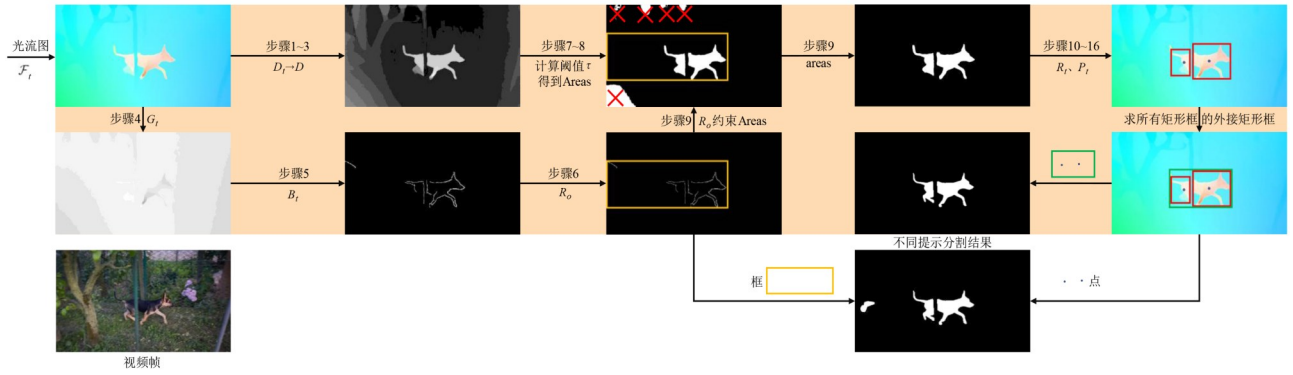


图3　运动提示生成算法流程示意图

首先,本方法基于光流图对运动显著性进行建模,通过阈值过滤,将运动强度高于阈值τ的区域视为显著性区域,其余部分归类为非显著性区域. 这种策略的核心动机在于区分静止背景与潜在运动目标,为后续生成提示提供候选区域. 其次,本文选择将提示点设置为矩形框内有效像素点的重心,主要考虑以下三点:(1)该策略能够获得一个更具代表性的中心位置,减少极端偏移带来的干扰;(2)重心点计算简单且稳定,具有较好的鲁棒性;(3)相比于使用框中心点或随机采样点,重心点更贴近真实目标的密集分布区域. 如图4所示,其中绿色为重心点,蓝色为中心点,不难发现,重心点能更准确地落在目标上,而中心点则有一定概率落在目标之外,造成提示偏差,给模型学习带来负面影响.

图4 重心点与中心点提示对比示意图

3.1.2 自适应表征学习SAM

以往工作[6,19~22]大都通过密集匹配光流与外观特征，融合得到时空表征，这难免会受运动噪声影响. 为此，本模块通过为SAM引入稀疏运动提示，并通过设计自适应表征学习适配器，引导其自适应学习有效的外观信息，从而解耦运动与外观学习，有效避免了运动噪声对模型性能的影响.

本文提出的自适应表征学习适配器设计如图2右侧所示. 在原SAM中，输入视频帧经过多个Transformer块后，生成一系列特征向量. 本文针对这些特征向量设计了一个适配器Adapter_V，其主要由两个MLP层[43]和高斯误差线性单元(Gaussian Error Linear Units，GELU)激活函数[45]组成，通过对特征向量进行两次非线性变换映射，提升其表征力. 接着，对输出的特征向量进行特征合并，得到一个特征张量. 该张量输入SAM的Neck模块，再输入本文设计的特征张量适配器Adapter_T进行特征变换和自适应学习. 该适配器通过一个3×3和一个1×1卷积进行特征维度变换和特征提取，并加入了RELU激活函数[46]，以增加特征非线性表达能力. 经过以上两个适配器，最终输出的特征$\mathcal{F}_i$更加符合本文下游任务需要.

上述通过构建轻量级适配器微调SAM，在减少可学习参数量的同时，提升SAM对下游UVOS任务的自适应性，从而带来更加精准的定位结果. SAM主要由一个基于ViT[47](Vision Transformer)的图像编码器Enc_i、一个提示编码器Enc_p和一个轻量掩码解码器Dec_m组成. 首先，Enc_i对视频帧$\boldsymbol{I}_t$编码，得到图像编码张量$\boldsymbol{E}_i$；其次，Enc_p对提示R_t和P_t进行编码，得到提示编码向量$\boldsymbol{E}_p$；最后，将图像编码张量和提示编码向量一起输入解码器Dec_m，得到最终定位结果L_t. 其过程表示如下：

$$\boldsymbol{E}_i=\text{Enc}_i(\boldsymbol{I}_t)\in\mathbb{R}^{h\times w\times c} \tag{1}$$

$$\boldsymbol{E}_p=\text{Enc}_p(R_t,P_t)\in\mathbb{R}^{n\times c} \tag{2}$$

$$L_t=\text{Dec}_m\Big(\text{CrossAttn}\Big(\boldsymbol{E}_i,\text{Concat}(\boldsymbol{E}_p,\boldsymbol{E}_t)\Big)\Big) \tag{3}$$

其中，h、w和c为特征张量的高度、宽度和通道数；n为提示编码向量的特征维度；P_t和R_t为算法1得到的点和框提示. 接着，提示编码向量$\boldsymbol{E}_p$和可学习向量$\boldsymbol{E}_t$通过Concat拼接后，与$\boldsymbol{E}_i$通过交叉注意力[48]CrossAttn对图像特征和提示特征进行充分交互，建立上下文相关性. 可学习的$\boldsymbol{E}_t$使得提示信息在训练过程中更加灵活，自适应性更佳.

由于本文方法在原SAM模型中加入了两个自适应表征学习适配器，图像编码器Enc_i编码的具体流程如下：

$$\boldsymbol{E}_i'=\text{Adapter}_V\Big(\text{Transformers}\Big(\text{Proj}(\boldsymbol{I}_t)\Big)\Big) \tag{4}$$

$$\boldsymbol{E}_i=\text{Adapter}_T\Big(\text{Neck}\Big(\text{Merge}(\boldsymbol{E}'_i)\Big)\Big) \tag{5}$$

其中，Adapter_V为本文设计的特征向量适配器；Adapter_T为本文设计的特征张量适配器；Proj为线性映射过程；Merge为特征向量合并过程；Neck为特征增强过程，具体可见图2. 具体地，Merge操作将h个形如$(b\times c\times 1\times w)$的特征向量在第三个维度进行拼接，得到形状为$(b\times c\times h\times w)$的特征张量. 其中$b$为批大小，$c$为通道数，$h$为向量个数，$w$为单个向量的维度大小. h和w亦可理解为拼接完成后的特征张量的高度和宽度.

3.2 表观聚焦细化模块

SAM采用了巨量训练数据，微调使其在下游UVOS任务上的泛化性极大提升. 通过以上自适应表征学习

SAM生成的定位结果L_t可生成较准确的定位结果.但由于SAM基于纯Transformer架构设计,采用了自注意力与交叉注意力机制,主要关注全局上下文依赖关系建模,难以充分捕获到局部信息,而局部信息对提升分割精度至关重要.因此,本文提出基于CNN设计的表观聚焦细化模块,先对视频帧提取多级特征,充分捕获不同尺度下的信息,再通过L_t逐级引导特征聚焦在目标区域,进一步细化局部分割细节.

如图2和图5所示,该表观聚焦细化模块包含一个多级特征编码器[49]和一个轻量化表观聚焦细化模块.首先,对于输入视频帧使用骨干网络Enc_m进行多尺度特征提取.其次,对多尺度特征采用定位引导聚焦模块(Location-Guided Focus Module,LGFM)进行引导聚焦,由运动提示定位模块输出的定位结果L_t渐进式引导特征聚焦于目标区域.再次,通过设计轻量化的卷积块注意力模块(Light Convolutional Block Attention Module,LightCBAM),让网络同时兼顾全局和通道的学习,专注于学习目标细节信息,从而减少周围背景的影响,提升分割精度.最后,预测头CNN_D对逐层聚焦后的特征进行预测.其中,表观聚焦细化模块的输入是运动提示定位模块输出的定位图、当前层高分辨率特征、由定位图引导前层特征聚焦后的低分辨率特征.详细过程表示如下:

$$\{\boldsymbol{X}_t^1,\boldsymbol{X}_t^2,\boldsymbol{X}_t^3\boldsymbol{X}_t^4\}=\mathrm{Enc}_m(\boldsymbol{I}_t) \tag{6}$$

$$\boldsymbol{X}_t^i=\begin{cases}\mathrm{LGFM}(\boldsymbol{X}_t^i,\boldsymbol{X}_t^i\odot L_t), & \text{if } i=4\\ \mathrm{LGFM}(\boldsymbol{X}_t^i,\boldsymbol{X}_t^i\odot L_t,\boldsymbol{X}_t^{i+1}), & \text{otherwise}\end{cases} \tag{7}$$

$$\boldsymbol{X}_t^i=\mathrm{LightCBAM}(\boldsymbol{X}_t^i) \tag{8}$$

$$M_t=\mathrm{CNN}_D(\boldsymbol{X}_t^1) \tag{9}$$

其中,$\odot$为逐元素相乘.

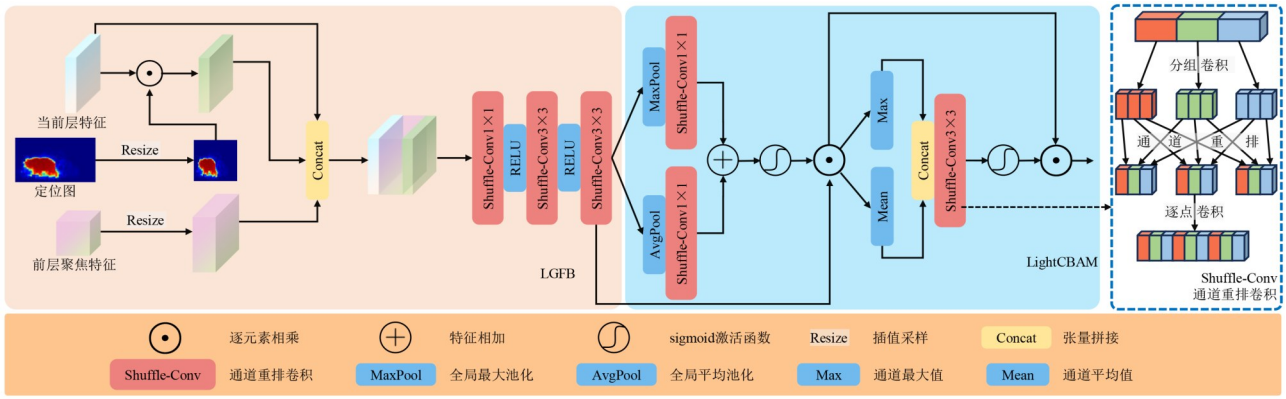


图5　表观聚焦细化模块结构示意图

3.3　损失函数

本模型的损失函数由二值交叉熵(Binary Cross-Entropy,BCE)损失函数$\mathcal{L}_{\mathrm{BCE}}$和交叉熵(Cross-Entropy,CE)损失函数$\mathcal{L}_{\mathrm{CE}}$[43]组成:

$$\mathcal{L}=\mathcal{L}_{\mathrm{BCE}}+\mathcal{L}_{\mathrm{CE}} \tag{10}$$

$$\mathcal{L}_{\mathrm{BCE}}=-\sum^{i,j}\Big[M_{\mathrm{gt}}(i,j)*\ln\sigma\big(L_t(i,j)\big)+\big(1-M_{\mathrm{gt}}(i,j)\big)*\ln\big(1-\sigma\big(L_t(i,j)\big)\big)\Big] \tag{11}$$

$$\mathcal{L}_{\mathrm{CE}}=-\sum^{i,j}\Big[M_{\mathrm{gt}}(i,j)*\ln M_t(i,j)+\big(1-M_{\mathrm{gt}}(i,j)\big)*\ln\big(1-M_t(i,j)\big)\Big] \tag{12}$$

其中,$\mathcal{L}_{\mathrm{BCE}}$和$\mathcal{L}_{\mathrm{CE}}$分别用来监督学习运动提示定位模块和表观聚焦细化模块中的参数;M_{gt}为真实掩码;i和j为像素点的横坐标和纵坐标;σ为sigmoid激活函数[43].

在UVOS中,标签通常含有不确定性,$\mathcal{L}_{\mathrm{BCE}}$与$\mathcal{L}_{\mathrm{CE}}$的结合可在鲁棒性与准确性之间取得平衡.本文选择$\mathcal{L}_{\mathrm{BCE}}$与$\mathcal{L}_{\mathrm{CE}}$的组合损失,主要考虑两点:(1)$\mathcal{L}_{\mathrm{BCE}}$可强化模型对前景区域的响应,尤其在前背景比例不平衡时具有更好的优化效果,这一损失函数主要用于监督运动提示定位模块,并且与SAM源码中保持一致;(2)$\mathcal{L}_{\mathrm{CE}}$则适合处理多类监督信号,能够提升整体的分类稳定性,这一损失函数主要用于监督表观聚焦细化模块.由于无监督目标分割任务最终只需要得到二值分割结果,分类较为简单,常用$\mathcal{L}_{\mathrm{BCE}}$和$\mathcal{L}_{\mathrm{CE}}$损失函数,使用更加复杂的损失函数或许会对模型的优化引入更多的不确定因素.

4　实验

4.1节介绍实验相关设置,包括数据集使用方式、模型评估指标和本文方法实现细节.4.2节从定量和定性的角度,在性能指标和可视化结果上与其他主流工作进行对比,以验证本文所提方法的有效性.4.3节介绍本文所提方法的相关消融实验,对比模型中的一些重要参数和模块变化对指标的影响.

4.1　实验设置

4.1.1　数据集

为了更加充分地验证本文所提方法的有效性,本

文在三个广泛用于UVOS任务的数据集对本文方法进行评估，这三个数据集分别是DAVIS16、FBMS和YTOBJ. 在训练阶段，本文使用了DAVIS16训练集和DUTS(DUT Salient object detection)[50]数据集. 在测试阶段，本文使用了DAVIS16、FBMS和YTOBJ三个数据集的测试集.

DUTS包含超过10 000张训练数据，且每张图像真实掩码的边界标注都比较精确，是最大的显著性检测数据集之一.

DAVIS16是VOS任务最常用的数据集之一，其包含30个训练视频序列和20个验证视频序列，每一帧都包含对目标真实掩码的标注.

FBMS包含59个视频序列，共720帧. 在部分序列中，有多个对象被标注为显著对象.

YTOBJ数据集由从YouTube视频网站收集到的大量视频序列组成，由于无训练集和测试集的区分，它仅用于测试. 其包含10个对象类超过20 000帧图像，每类包含9~24个视频序列.

4.1.2 评估指标

UVOS任务常用的评测指标[28]为$\mathcal{J}$和$\mathcal{F}$. 其中，$\mathcal{J}$用于评估区域相似度，是衡量真实掩码和预测掩码之间准确性的指标，其表示为

$$\mathcal{J}=\left|\frac{M_{\mathrm{gt}}\bigcap M_{\mathrm{pred}}}{M_{\mathrm{gt}}\bigcup M_{\mathrm{pred}}}\right| \tag{13}$$

其中，M_{gt}为真实掩码；M_{pred}为模型最终预测的掩码. $\mathcal{F}$是一个衡量真实掩码和预测掩码之间边界轮廓之间吻合度的指标，其表示为

$$P=\left|\frac{M_{\mathrm{gt}}\bigcap M_{\mathrm{pred}}}{M_{\mathrm{pred}}}\right|,R=\left|\frac{M_{\mathrm{gt}}\bigcap M_{\mathrm{pred}}}{M_{\mathrm{gt}}}\right| \tag{14}$$

$$\mathcal{F}=\left|\frac{2\times P\times R}{P+R}\right| \tag{15}$$

$\mathcal{G}$代表$\mathcal{J}$和$\mathcal{F}$的平均值，其表示为

$$\mathcal{G}=\left|\frac{\mathcal{J}+\mathcal{F}}{2}\right| \tag{16}$$

4.1.3 实现细节

训练数据由两部分组成：(1)DAVIS16包含30个视频序列，2 079张图像及其对应的光流图和真实掩码标注；(2)DUTS包含15 572张图像和真实掩码标注. 为防止模型对DAVIS16视频序列过拟合，每个训练周期对30个视频序列分别随机读取一次. DUTS与DAVIS16按1∶1的数量读取，即每个训练周期对DUTS随机读取30次.

DUTS由非连续视频序列组成，无法生成光流，因此，本文将这类情况归为光流估计失效的情况. 在光流估计失效的情况下，本文的光流提示生成算法无法得到点和框提示. 因此，在这种情况下，SAM采用无提示模式，以防错误提示带来的误差累积.

为了保持本文方法与主流方法在对比上的公平性，本文遵循文献[6,21]等主流工作[14~16,18,19,51~60]中的相关实验设置，使用循环成对场变换(Recurrent All-pairs Field Transforms，RAFT)[61]离线生成的光流图作为本文运动信息的来源.

在训练阶段，当读取数据来自DAVIS16时，通过算法1得到点和框提示；当读取数据来自DUTS时，无需光流，SAM采用无提示模式，通过自适应学习定位到理想的显著目标. 得益于DUTS数据集中大量人工标注的高质量真实掩码，这种训练策略极大地提升了模型在无提示模式下的分割精度和性能.

考虑到目标在时间维度上通常保持结构连续性，过强的数据增强会引入不符合视频时序真实分布的变形，反而干扰模型对运动和形状的建模. 因此，本文的数据增强策略仅采用简单的随机水平翻转和垂直翻转. 不仅有助于提升模型的泛化能力，避免过拟合，同时还能保持帧间一致性. 优化器使用自适应矩估计(Adaptive moment estimation，Adam)[62]，初始学习率为1×10^{-4}. 1×10^{-4}作为许多任务常用的初始学习率，经实验验证，在本文网络结构下，可以在保持训练稳定性的同时实现较快的收敛. 批量大小batch-size为4，即每一批从训练数据中随机读取4张图像. 实验硬件条件为Linux操作系统、32 GB内存、8核16线程的i7-10700K处理器、单张RTX3090显卡，计算统一设备架构(Compute Unified Device Architecture，CUDA)版本为10.2，Pytorch版本为1.12.1.

4.2 实验对比

4.2.1 定量实验

表1和表2列举了本文方法与目前主流工作在DAVIS16、FBMS和YTOBJ上的定量对比结果，可见，本文方法在三个数据集上都表现最优. 较当前先进的引导槽注意力网络(Guided Slot Attention Network，GSANet)方法，在DAVIS16数据集的$\mathcal{J}$指标上，本文方法提升了1.8%. 同时，本文方法的泛化性也较为出色，在FBMS和YTOBJ数据集中，尽管存在较多光流估计较差的场景，本文方法仍取得了先进的结果. 具体而言，在FBMS数据集的$\mathcal{J}$指标上，本文方法较时序运动选项(Temporal Motion Option，TMO)提升了3.3%；在YTOBJ数据集的$\mathcal{J}$指标上较层次特征对齐网络(Hierarchical Feature Alignment Network，HFAN)提升了2.6%，并在10类场景中的6类上，性能均达到最优或次优. 其中，红色、绿色、蓝色加粗字体依次代表在相应指标上性能排名前三的工作.

在表1和表2提及的工作中，协同注意力孪生网络(CO-attention Siamese Network，COSNet)、注意力图神经

表1　本文方法与主流工作在DAVIS16和FBMS数据集上的定量实验对比结果

方法	发表	分辨率	光流图	后处理	FPS	DAVIS16/%			FBMS/%		
						$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
PDB[51]	ECCV'2018	473×473		√	20.0	77.2	74.5	75.9	74.0	—	—
AGS[52]	CVPR'2019	473×473		√	10.0	79.2	77.4	78.6	—	—	—
AGNN[15]	ICCV'2019	473×473		√	3.6	80.7	79.1	79.9	—	—	—
COSNet[14]	CVPR'2019	473×473		√	—	80.5	79.4	80.0	75.6	—	—
AnDiff[53]	CVPR'2019	—			—	81.7	80.5	81.1	—	—	—
MATNet[19]	AAAI'2020	473×473	√	√	20.0	82.4	80.7	81.6	76.1	—	—
GraphMem[16]	ECCV'2020	384×640	√		5.0	82.5	81.2	81.9	—	—	—
FSNet[54]	ICCV'2021	352×352	√	√	12.5	83.4	83.1	83.3	—	—	—
F2Net[55]	AAAI'2021	473×473			10.0	83.1	84.4	83.7	77.5	—	—
AMCNet[20]	ICCV'2021	384×384	√	√	17.5	84.5	84.6	84.6	76.5	—	—
TransportNet[56]	ICCV'2021	512×512	√		12.5	84.5	85.0	84.8	78.7	—	—
RTNet[57]	CVPR'2021	384×672	√	√	—	85.6	84.7	85.2	—	—	—
D2Conv3D[18]	WACV'2022	480×854			4.5	85.5	86.5	86.0	—	—	—
HFAN[6]	ECCV'2022	—	√		11.0	86.0	87.3	86.7	76.1	75.5	75.8
PMN[58]	WACV'2023	—	√		—	85.4	86.4	85.9	77.7	77.4	77.6
TMO[21]	WACV'2023	384×384	√		—	85.6	86.6	86.1	**79.9**	**82.7**	**81.3**
GSANet[59]	CVPR'2024	512×512	√		—	**87.0**	**88.4**	**87.7**	79.2	79.4	79.3
DPA[60]	CVPR'2024	352×352	√		—	**86.3**	**87.4**	**86.9**	**81.2**	**82.1**	**81.6**
本文方法	—	384×384	√		6.8	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**

表2　本文方法与主流工作在YTOBJ数据集上的定量实验对比结果　　单位:%

方法	飞机	鸟	船	车	猫	牛	狗	马	摩托	火车	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
PDB[51]	78.0	80.0	58.9	76.5	63.0	64.1	70.1	**67.6**	58.4	35.3	65.5	—	—
AGS[52]	**87.7**	76.7	**72.2**	78.6	69.2	64.6	73.3	64.4	62.1	48.2	69.7	—	—
AGNN[15]	81.1	75.9	70.7	78.1	67.9	69.7	**77.4**	67.3	**68.3**	47.8	70.8	—	—
COSNet[14]	81.1	75.7	71.3	77.6	66.5	69.8	76.8	67.4	**67.7**	46.8	70.5	—	—
MATNet[19]	72.9	77.5	66.9	79.0	73.7	67.4	75.9	63.2	62.6	51.0	69.0	—	—
GraphMem[16]	86.1	75.7	68.6	**82.4**	65.9	**70.5**	77.1	**72.2**	63.8	47.8	71.4	—	—
AMCNet[20]	78.9	80.9	67.4	**82.0**	69.0	69.6	75.8	63.0	63.4	**57.8**	71.1	—	—
RTNet[57]	84.1	80.2	70.1	79.5	71.8	70.1	71.3	65.1	**64.6**	53.3	71.0	—	—
HFAN[6]	84.7	80.0	72.0	76.1	**76.0**	**71.2**	76.9	**71.0**	64.3	**61.4**	**73.4**	**72.8**	**73.1**
PMN[58]	86.6	**85.8**	66.7	77.3	75.4	69.1	75.1	66.5	58.4	56.9	71.8	70.8	71.3
TMO[21]	85.7	80.0	70.1	78.0	73.6	70.3	76.8	66.2	58.6	47.0	71.5	71.6	71.6
GSANet[59]	86.0	80.9	**73.2**	77.9	73.1	**70.8**	75.6	63.0	63.0	57.1	72.1	70.6	71.3
DPA[60]	**87.5**	**85.6**	70.1	77.7	**81.2**	69.0	**77.8**	61.9	62.1	55.3	**73.7**	**74.0**	**73.9**
本文方法	**88.2**	**81.7**	**73.9**	**83.2**	**77.3**	70.1	**78.4**	66.8	63.1	**70.3**	**75.3**	**74.7**	**75.0**

网络(Attentive Graph Neural Network,AGNN)、情景图记忆(episodic Graph Memory,GraphMem)网络、动态扩张三维卷积(Dynamic Dilated Conv3D,D2Conv3D)等未显式利用运动信息,主要依靠模型捕捉帧间协同显著信息来确定待分割目标.而当遇到背景嘈杂和遮挡时,模型对显著目标的注意力就会被分散,导致难以充分挖掘到显著运动信息.运动注意力转换网络(Motion-Attentive Transition Network,MATNet)、HFAN、TMO、GSANet、双重原型注意力(Dual Prototype Attention,DPA)等显式利用运动信息,但同时也意识到运动噪声对分割结果的影响,为此设计模块来抑制运动信息的表达,但同时也抑制了模型的性能.以上方法均采用像素级密集匹配策略,然而,在遮挡、相机抖动、运动模糊等挑战性场景中,易产生大量错误匹配.为此,本文方法将光流编码的密集运动信息转换为稀疏点和框提示,提示SAM通过自适应学习,获得更为鲁棒的时空表征,增强模型抗噪能力.

表3列举了基于SAM改进的相关工作与本文方法

在DAVIS17(DAVIS 2017)[63]上的定量对比结果.DAVIS17相较于DAVIS16,增加了多目标视频序列,更具挑战性.相比于无监督方式的基于分割任意模型的无监督视频目标分割(Unsupervised Video Object segmentation via Segment Anything Model,UVOSAM)[33]和解耦视频分析(DEcoupled Video Analysis,DEVA)[64],本文方法的性能大幅领先.例如,相比于性能最优的UVOSAM,本文方法在$\mathcal{G}$指标上高出7.7个百分点.此外,即使相比于半监督工作SAM-Track(Segment Anything Model for Tracking)[65]、SAM-PT(Segment Anything Model with Point Tracking)[44]和TAM(Tracking Anything Module)[66],本文方法的性能仍大幅领先.例如,相比于性能最优的SAM-PT,本文方法在$\mathcal{G}$指标上高出7.2个百分点,这充分证明了本文方法在遇到多目标场景时,表现依然出色.其中,红色、绿色、蓝色加粗字体依次代表在相应指标上性能排名前三的工作.

表3 基于SAM改进的相关工作在DAVIS17上的定量实验对比结果 单位:%

方法	SAM-Track[64]	SAM-PT[44]	TAM[65]	UVOSAM[33]	DEVA[66]	本文方法
类型	半监督	半监督	半监督	无监督	无监督	无监督
$\mathcal{J}_\mathcal{M}\uparrow$	75.3	**76.5**	69.8	**75.5**	—	**85.9**
$\mathcal{F}_\mathcal{M}\uparrow$	**83.1**	**82.3**	76.4	82.0	—	**87.4**
$\mathcal{G}_\mathcal{M}\uparrow$	**79.2**	**79.4**	73.1	78.9	73.4	**86.6**

本文方法充分考虑了目标的运动和显著性信息,设计了提示生成算法以准确定位目标,并且基于设计的适配器对SAM进行了针对UVOS任务的微调.而表3中的三个半监督方法SAM-Track、SAM-PT和TAM均是针对视频跟踪任务提出的,由于SAM本身具备分割能力,论文中的实验部分在视频分割相关数据集上验证了其分割性能,但这些方法并非专门针对UVOS任务提出的解决方案.表3中提及的两个无监督方法UVOSAM和DEVA,均借助了一个显著目标检测网络来确定待分割目标,以此作为线索提示SAM进行分割,因此最终的分割性能依赖于显著目标检测网络的准确性,而显著目标检测网络在检测时,并没有考虑目标的运动信息,在遮挡、相机抖动、运动模糊场景下,极易分割错误目标.因此,本文针对SAM进行了一系列创新,使其在基于SAM改进的相关工作中性能达到领先.

由图6和表4可知,通过算法1中步骤6直接得到的矩形框可能包含大量背景,导致模型对无关内容进行定位和分割.而通过步骤10~16优化后的矩形框能够更准确地提示目标(如表4中加粗字体所示),并尽可能包含更少的背景,从而为模型带来更好的提示效果,进而获得更佳的定位和分割效果,在各个性能指标上也表现更佳.表5则为本文选择重心点而不是中心点作为提示点提供了更直观的定量对比.其中,加粗字体代表在相应指标上性能最优的值.

表4 步骤6与步骤10~16矩形框优化前后对指标影响的定量实验 单位:%

提示类型	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
步骤6	87.9	89.6	88.8	81.8	82.2	82.0	74.6	73.9	74.3
步骤10~16	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**	**75.3**	**74.7**	**75.0**

表5 中心点与重心点提示对指标影响的定量实验 单位:%

提示类型	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
中心点	86.1	88.3	87.2	80.6	80.8	80.7	73.3	72.5	72.9
重心点	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**	**75.3**	**74.7**	**75.0**

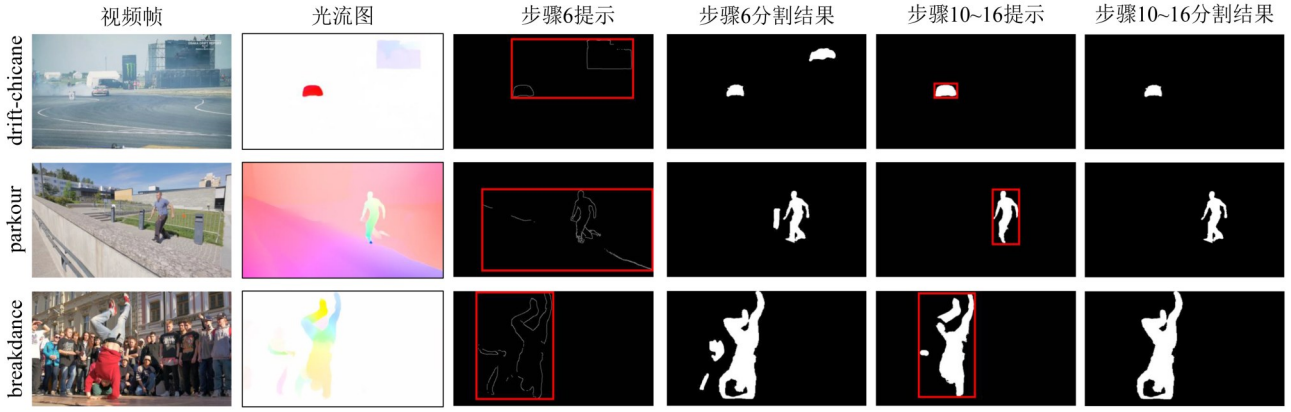


图6 步骤6与步骤10~16矩形框优化前后对分割结果影响对比示意图

最后,为了验证算法1中参数θ设置的合理性,并分析其对最终分割性能的影响,本文进行了敏感性分析实验,具体如表6所示. 在保持其他设置不变的前提下,分别选取$\theta\in\{0.05,0.1,0.12,0.14,0.15,0.16,0.18,0.2,0.25\}$进行实验. 实验结果表明:在$\theta$取值在0.12~0.18时,模型表现较为稳定,上下浮动较不大;当θ设置过低(如0.05)或过高(如0.25)时,分割精度明显下降. 综合考虑精度和鲁棒性,本文选择θ=0.15作为默认设置,在多个数据集上均获得较好的性能,如表6中加粗字体所示.

表6　算法1中参数θ取值对指标影响的定量实验　单位:%

θ取值	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}$ ↑	$\mathcal{F}_\mathcal{M}$ ↑	$\mathcal{G}_\mathcal{M}$ ↑	$\mathcal{J}_\mathcal{M}$ ↑	$\mathcal{F}_\mathcal{M}$ ↑	$\mathcal{G}_\mathcal{M}$ ↑	$\mathcal{J}_\mathcal{M}$ ↑	$\mathcal{F}_\mathcal{M}$ ↑	$\mathcal{G}_\mathcal{M}$ ↑
0.25	85.2	87.3	86.3	79.6	79.8	79.7	72.8	72.2	72.5
0.20	86.8	88.5	87.7	80.1	80.5	80.3	73.1	72.9	73.0
0.18	88.1	89.8	89.0	81.7	81.9	81.8	74.1	73.7	73.9
0.16	88.2	90.0	89.1	82.0	82.3	82.2	74.5	74.1	74.3
0.15	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**	**75.3**	**74.7**	**75.0**
0.14	88.3	90.2	89.3	82.2	82.3	82.3	74.8	74.2	74.5
0.12	88.0	90.1	89.1	81.8	81.8	81.8	74.5	73.8	74.2
0.10	87.8	89.7	88.8	81.1	81.3	81.2	73.9	73.1	73.5
0.05	82.3	85.0	83.7	78.5	78.8	78.7	70.2	69.9	70.1

4.2.2　定性实验

图7为本文方法在DAVIS数据集上的libby、kite-surf、dance-twirl、dogs-jump等8个场景连续视频帧上的可视化定性结果. 其中包含目标遮挡(libby、kite-surf)、快速移动(libby、kite-surf、drift-chicane、horsejump-high)、背景复杂(libby、dance-twirl)等多种挑战性场景. 本文方法的分割性能表现出色,边缘处分割精度较高. 在dance-twirl场景下,背景含有大量观众,本文方法仍能分割出前景中的舞者,而不受背景干扰影响;在kite-surf场景下,浪花之后露出的冲浪板部分占比较小,本文方法仍能有效识别并准确分割出冲浪板部分;在libby场景下的小狗被铁杆挡住的部分身体为非显著运动部分;在goat场景下的山羊和周围环境的色彩相似;在drift-chicane场景下的跑车身后的白烟会对目标造成干扰,针对这些极具挑战性的场景,本文方法也能有效克服挑战,实现精确分割. 以上定性实验结果充分证明了本文方法在各种复杂场景中的有效性.

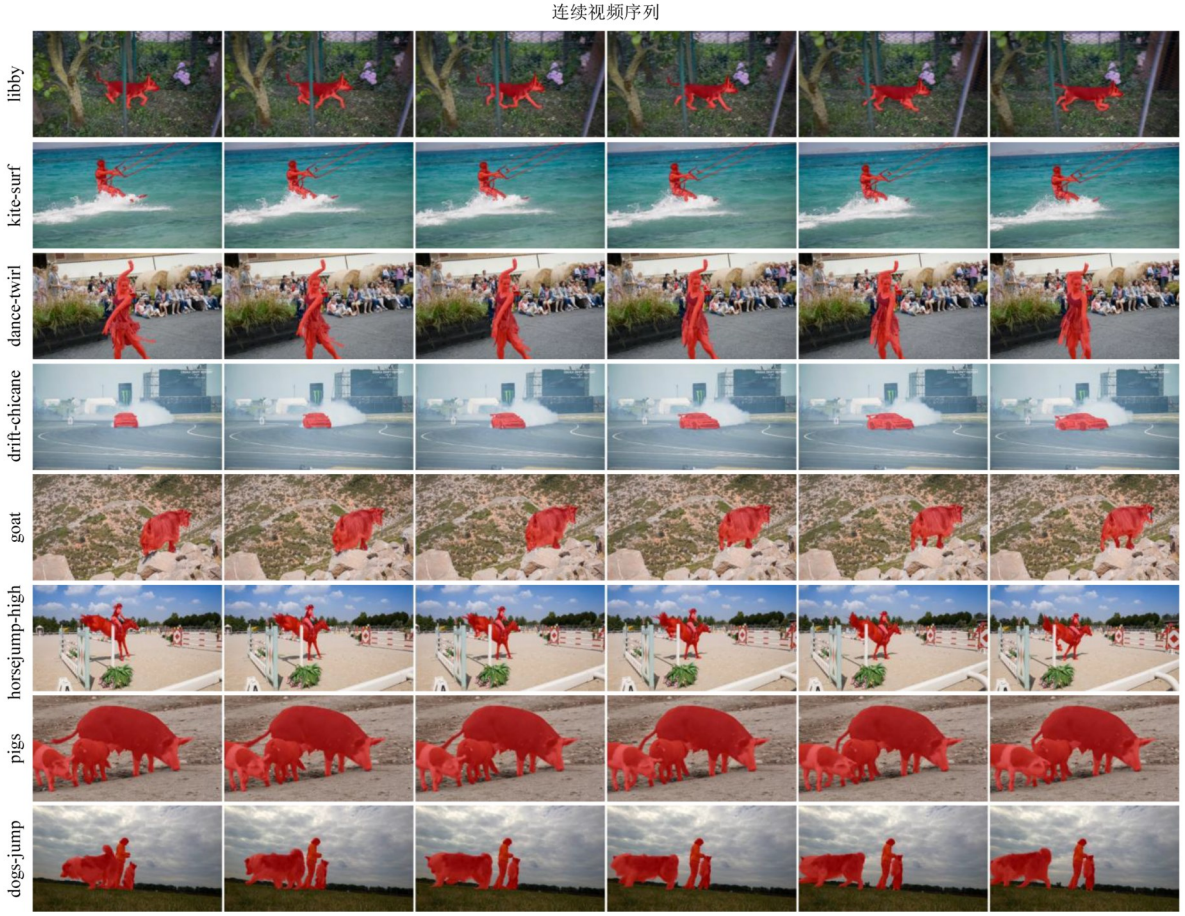


图7　本文方法在DAVIS数据集上部分场景下的连续分割效果

图8展示了本文方法与其他先进工作在DAVIS16数据集上的定性实验对比结果. 可见,所有对比工作均能分割出主体目标,但在定位准确度和分割精度上,本文方法表现最优. 在breakdance和parkour场景下,光流估计存在噪声干扰,例如舞者背后的观众、跑酷者在太阳底下的阴影,这些干扰导致其他对比工作分割准确度较低. 在bmx-trees、libby和kite-surf场景中,存在遮挡问题,例如树挡住了骑车的人、铁杆挡住了小狗、浪花挡住了冲浪板. 部分工作将树和铁杆错误地分割为目标的一部分,从而影响了最终分割精度,而其他工作未能分割出浪花后的冲浪板. 在blackswan场景下,天鹅尾部有一撮白色羽毛与整体黑色有明显区别,但从语义上二者属于同一目标. 针对这类挑战,除本文工方法之外,其他工作均未成功分割.

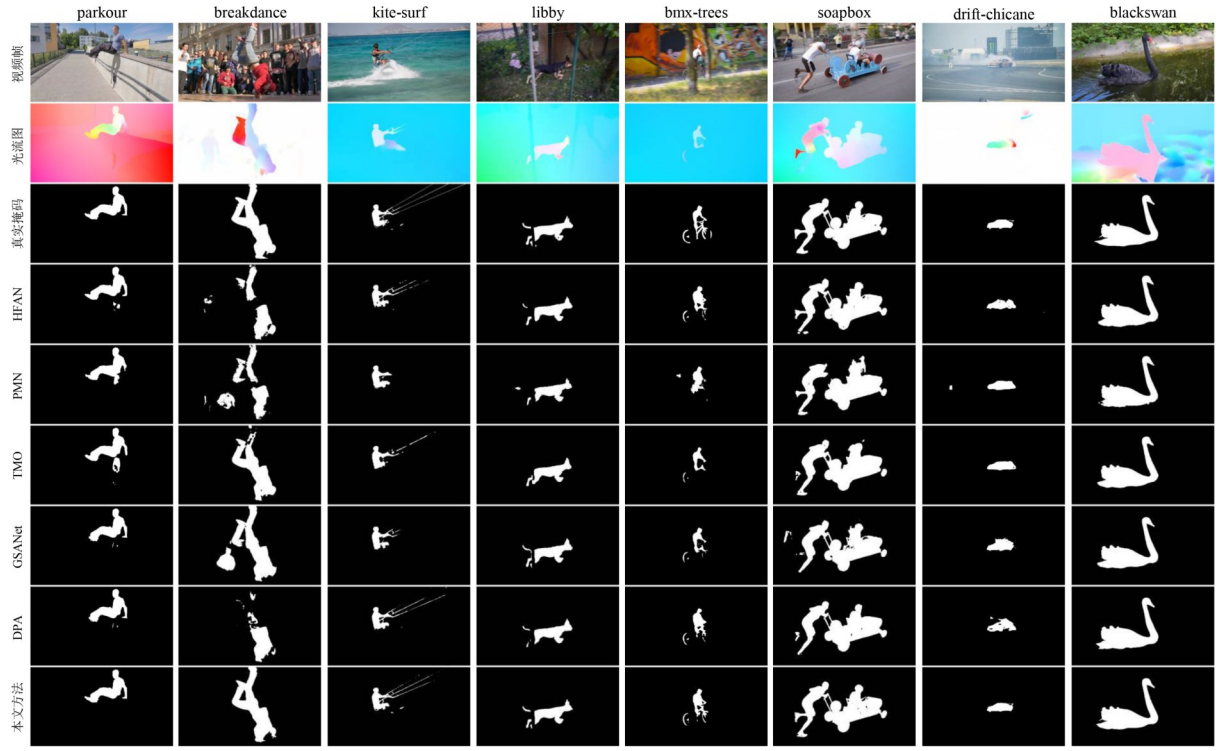


图8 本文方法与其他先进工作在DAVIS16数据集上的定性分割结果对比

图9为本文方法与其他先进工作在FBMS和YTOBJ上的定性实验对比结果. FBMS和YTOBJ存在大量光流估计较差或不准确的场景,从图9中的8个场景对比可见,本文方法在光流估计较差或不准确的情况下,仍取得了较为精准的分割.

从图8和图9的对比中可知,其他工作的分割结果中存在散点现象,而本文方法较为良好,原因在于,其他工作的密集匹配策略易受到光流噪声干扰,当光流估计出背景噪声时,这种密集匹配策略也将部分背景区域分割,造成散点现象. 本文方法得益于SAM对提示范围内的目标在整体语义理解上的精确性. 图10生动展示了在提示范围内,SAM能够进行精确的语义理解,准确定位到目标的整体,并将目标与背景区分开来,不受背景噪声干扰. 因此,SAM生成的定位注意力图能够有效地将目标和背景区分开. 在聚焦细化阶段,该注意力图引导多级特征进行渐进式细化分割,从而使模型专注于目标本身,忽略背景噪声的干扰,避免了散点的出现.

另外,由SAM输出的定位注意力图,在聚焦细化阶段引导多级特征渐进式细化分割. 边缘分割的准确性和区域分割一致的联系性,如不出现空洞,主要得益于:(1)定位注意力图本身所关注的区域就是连续的,其能够引导多级特征聚焦于目标本身,从而大大减少了模型对背景的关注;(2)多级特征中的低层特征感受野较小,保留了目标的边缘细节(如纹理、轮廓),而高层特征感受野较大,蕴含目标本身的语义信息(如目标形状、大小);(3)聚焦细化模块通过编码器-解码器结构设计和残差跳跃连接,将浅层的边缘细节与深层语义融合,避免单一尺度特征导致的边缘模糊或空洞;(4)损失函数将引导模型内部区域分割的一致,加强边缘轮廓的优化.

对于刚性目标,当提示信息仅覆盖其运动部分时,SAM也能得到出色的定位效果. 因此,对于刚性目标,尽管提示框不包含静止部分,但只要借助运动信息得

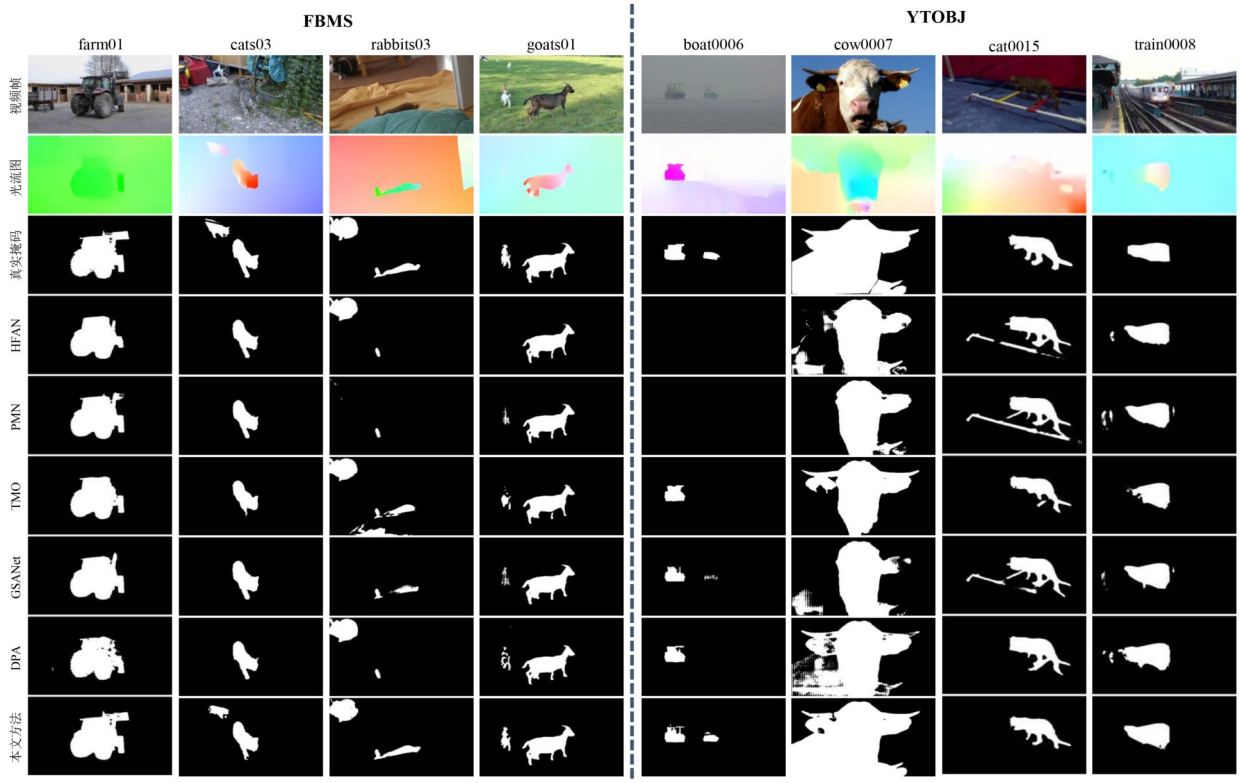


图9　本文方法与其他先进工作在FBMS和YTOBJ数据集上的定性分割结果对比

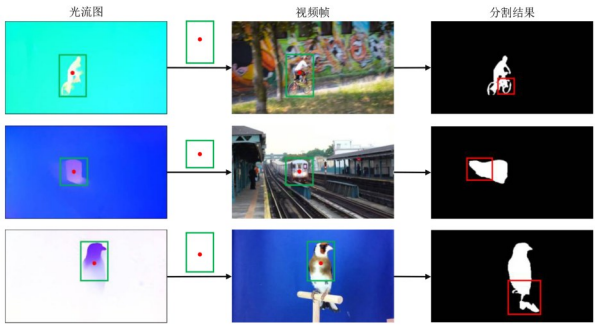


图10　刚性目标局部运动不显著场景分割结果

到包含部分目标的提示框,SAM就能凭借其强大的零样本泛化性能,通过部分提示信息引导模型准确定位到目标整体. 图10选取了一组刚性目标局部运动不显著的场景进行验证. 当光流估计对于刚性目标的静态(运动不显著)部分估计失败时,算法1仅计算出目标运动部分的提示信息. 然而,SAM仍可进行精确的语义理解,定位到刚性目标的静止部分.

4.3　消融实验

一些工作表明,SAM微调前后对下游任务性能的提升有显著影响. 表7通过对比SAM微调前后对下游UVOS任务性能的影响,进行了消融实验. 如表7中加粗字体所示,实验结果表明:微调后,SAM在UVOS任务上的表现更加出色,性能提升较为明显. 这同时证明了本文设计的两个轻量级特征适配器能够很好地将SAM中的有用知识自适应地转化为更适合下游UVOS任务的知识,极大地提升了SAM在UVOS任务上的泛化性. 这两个适配器的参数量和计算量分别为0.3 M和1.35 Gflops.

表7　SAM微调前后消融对比　　单位:%

SAM	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
微调前	82.2	84.0	83.1	75.7	75.5	75.6	67.5	65.2	66.4
微调后	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**	**75.3**	**74.7**	**75.0**

表8对算法1中点提示和框提示的组合方式进行了消融实验. 当点提示和框提示单独作用时,模型性能有所下降;而当点提示和框提示一起作用时,二者优势互补,达到了最优的实验结果,如加粗字体所示.

表9对模型中的两个创新设计模块进行了消融实验. 当仅使用运动提示定位模块时,在DAVIS16数据集

表8　不同提示组合消融对比　　单位:%

提示	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
点	87.2	89.2	88.2	80.9	81.6	81.3	73.7	73.4	73.6
框	86.5	88.6	87.6	80.2	81.1	80.7	73.0	73.0	73.0
点+框	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**	**75.3**	**74.7**	**75.0**

上的效果仍较为突出,但在FBMS和YTOBJ数据集上的泛化性能稍弱. 加入表观聚焦细化模块后,如加粗字体所示,整个模型的性能得到了进一步提升. 该消融实验充分验证了本文设计模型的定位模块能够准确定位目标,而细化模块轻量化的轻量化设计在降低参数量的同时,提升了模型的分割精度.

图11和图12针对表8和表9进行了更加直观的可视化示意. 由图11可知,当点提示和框提示单独作用时,分割结果要么不完整,要么会包含部分背景区域;只有当二者共同作用并相互约束时,分割效果才完整且准确. 图12展示了不同模块组合的消融可视化结果,虽然肉眼较难直观看出细微差别,但从图片下方的$\mathcal{J}$和$\mathcal{F}$指标上可以更加直观看出分割效果的区别. 加入表观聚焦细化模块后,性能有了进一步提升.

表9 不同模块组合消融对比 单位:%

运动提示定位模块	表观聚焦细化模块	DAVIS16			FBMS			YTOBJ		
		$\mathcal{J}_M\uparrow$	$\mathcal{F}_M\uparrow$	$\mathcal{G}_M\uparrow$	$\mathcal{J}_M\uparrow$	$\mathcal{F}_M\uparrow$	$\mathcal{G}_M\uparrow$	$\mathcal{J}_M\uparrow$	$\mathcal{F}_M\uparrow$	$\mathcal{G}_M\uparrow$
√		87.8	89.7	88.8	81.8	82.1	82.0	74.5	74.0	74.3
√	√	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**	**75.3**	**74.7**	**75.0**

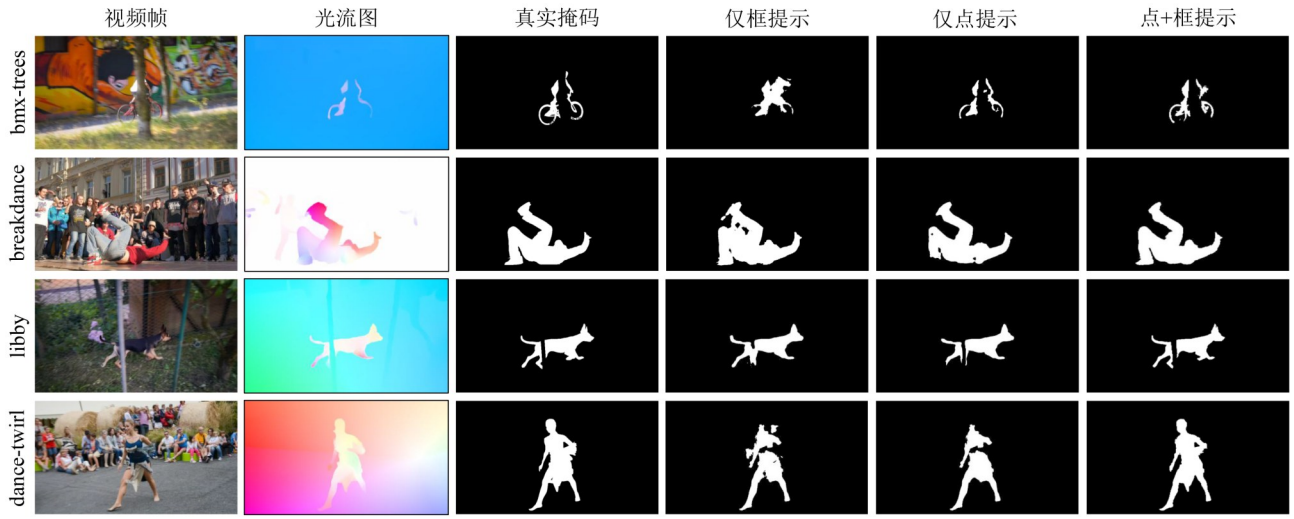


图11 不同提示组合消融可视化示意图

图12 不同模块组合消融可视化示意图

表10和表11对表观聚焦细化模块进行了相关消融实验. 表10对不同类型卷积的使用情况进行了消融对比,展示了各个模块的计算量和参数量的变化. 由表10可知,当采用通道重排卷积(Shuffle-Conv)代替传统卷积(Conv)来设计表观聚焦细化模块时,其计算量和参数量分别降低了88.0%和50.9%. 而由表11可知,在计算量和参数量大幅下降、节省计算和开销的同时,本文方法在三个数据集上的各项性能指标均变化不大.

表12针对轻量化卷积注意力模块中采用最大池化和平均池化并行设计进行了消融实验. 当仅使用其中一种池化时,性能均有所下降;只有采用并行设计,共同作用时,性能才达到最佳,如表12中加粗字体所示. 此外,并行结构的设计主要出于以下考虑:(1)互补性,平均池化能够捕捉图像整体的背景特征分布,有助于捕获全局语义信息;而最大池化更关注局部响应强烈的区域,尤其对前景目标的边缘、纹理等显著细节具有更强的敏感性. 二者结合可以互补全局与局部特征,有助于细化前景目标的结构轮廓与边界信息.(2)效率与效果平衡,该设计不引入额外的可学习参数,仅通过结构并行的方式增强通道注意力表示,符合本文模块轻量化的设计目标.

表10　表观聚焦细化模块轻量化计算量和参数量消融对比

卷积类型	Conv		Shuffle-Conv	
	计算量/GFlops	参数量/M	计算量/GFlops	参数量/M
表观聚焦细化模块	24.19	13.80	2.90	6.78
LGFB	21.68	7.15	0.39	0.13
CBAM	7.24	2.89	0.08	0.07

表11　表观聚焦细化模块轻量化性能指标消融对比 单位:%

卷积类型	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
Conv	88.8	90.8	89.8	82.2	82.9	82.6	75.2	74.9	75.1
Shuffle-Conv	88.6	90.5	89.6	82.5	82.9	82.7	75.3	74.7	75.0

表12　不同池化组合消融对比 单位:%

池化类型	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
最大池化	87.8	89.5	88.7	81.8	82.5	82.2	74.9	74.2	74.6
平均池化	87.2	89.1	88.2	81.5	82.0	81.8	74.1	73.5	73.8
最大池化+平均池化	**88.6**	**90.5**	**89.6**	**82.5**	**82.9**	**82.7**	**75.3**	**74.7**	**75.0**

表13和表14针对较为先进的轻量卷积技术和特征融合技术进行了相关消融实验. 从表13和表14中的各项性能指标可知,新的卷积技术和新的特征融合技术能够助力本文方法在性能上更进一步.

表13　不同轻量卷积技术消融对比 单位:%

卷积类型	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
ShuffleConv[67]	88.6	90.5	89.6	82.5	82.9	82.7	75.3	74.7	75.0
DYConv[68]	88.8	90.7	89.8	82.6	83.1	82.9	75.5	74.9	75.2
GhostConv[69]	89.0	91.0	90.0	82.8	83.4	83.1	75.5	75.0	75.3
FADC[70]	88.9	90.9	89.9	82.8	83.5	83.2	75.7	75.3	75.5
FDConv[71]	89.2	91.0	90.1	83.0	83.5	83.3	75.8	75.3	75.6

表14　不同特征融合技术消融对比 单位:%

特征融合类型	DAVIS16			FBMS			YTOBJ		
	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$	$\mathcal{J}_\mathcal{M}\uparrow$	$\mathcal{F}_\mathcal{M}\uparrow$	$\mathcal{G}_\mathcal{M}\uparrow$
FPN[72]	88.6	90.5	89.6	82.5	82.9	82.7	75.3	74.7	75.0
PAN[73]	88.8	90.6	89.7	82.5	83.1	82.8	75.4	75.1	75.3
BiFPN[74]	89.1	91.0	90.1	82.6	83.2	82.9	75.7	75.2	75.5
FreqFusion[75]	89.2	91.2	90.2	82.9	83.5	83.2	75.8	75.4	75.6

图13以注意力热力图的方式展示了本文方法的有效性. 与真实掩码相比,定位模块的输出确实能够较好地定位目标,使得注意力都集中在目标区域;而细化模块的输出进一步聚焦细化定位模块的注意力到目标区域,达到进一步细化分割的目的. 两个模块相互协作,极大地提升了本文所提模型的分割性能.

5　结论

本文提出了一种运动提示引导自适应学习的UVOS框架,旨在通过设计光流提示生成算法,将运动信息转换为稀疏点和框的坐标信息,用于提示SAM,通过轻量级适配器进行自适应学习,该框架能够获得更为鲁棒的时空表征. 实验结果表明:该方法在DAVIS16、FBMS和YTOBJ三个数据集上取得了领先性能,尤其是在较为复杂的场景(如遮挡、运动不显著场景)中表现出显著优势.

本文通过引入SAM,推动了大模型和提示学习在UVOS任务中的应用,为相关研究方向所面临的挑战提供了解决的新视角. 未来研究方向包括进一步设计更加精准的提示生成算法,以提升在复杂场景下的定位精度,特别是在光流估计失效或目标遮挡等挑战性场景下;此外,当前UVOS任务相关的数据集规模仍有限,

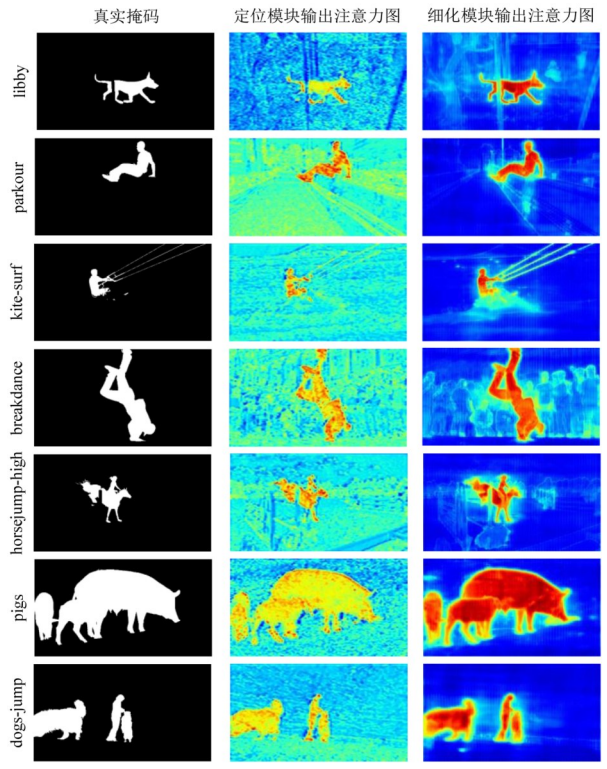


图 13 不同模块输出注意力可视化热力效果图

这对模型性能的提升构成了一定瓶颈. 如何充分利用SAM经过海量数据训练所获得的强大零样本泛化性能和通用分割能力,在少量训练数据下实现更优的分割效果,也是未来需要重点深入研究的问题. 以上展望构成了本文后续深入研究的重要内容.

参考文献

[1] MOHAMED E, EWAISHA M, SIAM M, et al. Instance-motseg: Real-time instance motion segmentation for autonomous driving[EB/OL]. (2021-05-26)[2025-02-26]. https://arxiv.org/abs/2008.07008.

[2] OCHS P, MALIK J, BROX T. Segmentation of moving objects by long term video analysis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(6): 1187-1200.

[3] DRAYER B, BROX T. Object detection, tracking, and motion segmentation for object-level video segmentation[EB/OL]. (2016-08-10)[2025-02-26]. https://arxiv.org/abs/1608.03066.

[4] SIDDIQUE A, LEE S. Object-wise video editing[J]. Applied Sciences, 2021, 11(2): 671-692.

[5] WANG W G, SHEN J B, PORIKLI F, et al. Semi-supervised video object segmentation with super-trajectories[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(4): 985-998.

[6] PEI G S, SHEN F M, YAO Y Z, et al. Hierarchical feature alignment network forUnsupervised video object segmentation[C]//Computer Vision - ECCV 2022. Cham: Springer, 2022: 596-613.

[7] LIN F C, XIE H T, LI Y, et al. Query-memory re-aggregation for weakly-supervised video object segmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2038-2046.

[8] MIAO J X, WEI Y C, YANG Y. Memory aggregation networks for efficient interactive video object segmentation[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10363-10372.

[9] WU J N, JIANG Y, SUN P Z, et al. Language as queries for referring video object segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 4964-4974.

[10] BANICA D, AGAPE A, ION A, et al. Video object segmentation by salient segment chain composition[C]//2013 IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE, 2013: 283-290.

[11] ZHANG D, JAVED O, SHAH M. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2013: 628-635.

[12] WANG W G, SHEN J B, YANG R G, et al. Saliency-aware video object segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(1): 20-33.

[13] HU Y T, HUANG J B, SCHWING A G. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation[C]//Computer Vision - ECCV 2018. Cham: Springer, 2018: 813-830.

[14] LU X K, WANG W G, MA C, et al. See more, know more: Unsupervised video object segmentation with co-attention Siamese networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3618-3627.

[15] WANG W G, LU X K, SHEN J B, et al. Zero-shot video object segmentation via attentive graph neural networks[C]// 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 9235-9244.

[16] LU X K, WANG W G, DANELLJAN M, et al. Video object segmentation with episodic graph memory networks[C]// Computer Vision - ECCV 2020. Cham: Springer, 2020: 661-679.

[17] MAHADEVAN S, ATHAR A, OŠEP A, et al. Making a case for 3d convolutions for object segmentation in videos[EB/OL]. (2023-09-01)[2025-02-26]. https://arxiv.org/abs/2008.11516.

[18] SCHMIDT C, ATHAR A, MAHADEVAN S, et al. D2Conv3D: Dynamic dilated convolutions for object segmentation in videos[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2022: 1929-1938.

[19] ZHOU T F, LI J W, WANG S Z, et al. MATNet: Motion-attentive transition network for zero-shot video object segmentation[J]. IEEE Transactions on Image Processing, 2020, 29: 8326-8338.

[20] YANG S, ZHANG L, QI J Q, et al. Learning motion-appearance co-attention for zero-shot video object segmentation[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 1544-1553.

[21] CHO S, LEE M, LEE S, et al. Treating motion as option to reduce motion dependency in unsupervised video object segmentation[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2023: 5129-5138.

[22] 苏天康, 宋慧慧, 樊佳庆, 等. 深度信号引导学习混合变换器的高性能无监督视频目标分割[J]. 电子学报, 2023, 51(5): 1388-1395.
SU T K, SONG H H, FAN J Q, et al. Learning depth signal guided mixed transformer for high-performance unsupervised video object segmentation[J]. Acta Electronica Sinica, 2023, 51(5): 1388-1395. (in Chinese)

[23] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2017-02-22)[2025-02-26]. https://arxiv.org/abs/1609.02907.

[24] BROMLEY J, BENTZ J W, BOTTOU L, et al. Signature verification using a "Siamese" time delay neural network[M]//Advances in Pattern Recognition Systems Using Neural Network Technologies. Singapore: World Scientific, 1994: 25-44.

[25] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 4489-4497.

[26] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//The 28th International Conference on Neural Information Processing Systems - Volume 1. New York: ACM, 2014: 568-576.

[27] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 3992-4003.

[28] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 724-732.

[29] PREST A, LEISTNER C, CIVERA J, et al. Learning object class detectors from weakly annotated video[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 3282-3289.

[30] MA J, HE Y T, LI F F, et al. Segment anything in medical images[J]. Nature Communications, 2024, 15: 654.

[31] WANG D, ZHANG J, DU B, et al. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model[J]. Advances in Neural Information Processing Systems, 2023, 36: 8815-8827.

[32] CHEN K Y, LIU C Y, CHEN H, et al. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 4701117.

[33] ZHANG Z H, WEI Z C, ZHANG S F, et al. Uvosam: A mask-free paradigm for unsupervised video object segmentation via segment anything model[EB/OL]. (2024-06-06)[2025-02-26]. https://arxiv.org/abs/2305.12659.

[34] CUI C, DENG R N, LIU Q, et al. All-in-SAM: From weak annotation to pixel-wise nuclei segmentation with prompt-based finetuning[J]. Journal of Physics: Conference Series, 2024, 2722(1): 012012.

[35] PENG Z L, XU Z Q, ZENG Z L, et al. SAM-PARSER: Fine-tuning SAM efficiently by parameter space reconstruction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(5): 4515-4523.

[36] LI Y L, ZHANG J, TENG X, et al. Refsam: Efficiently adapting segmenting anything model for referring video object segmentation[EB/OL]. (2024-09-03)[2025-02-26]. https://arxiv.org/abs/2307.00997.

[37] CHEN S, GE C, TONG Z, et al. Adaptformer: Adapting vision transformers for scalable visual recognition[J]. Advances in Neural Information Processing Systems, 2022, 35: 16664-16678.

[38] SUNG Y L, CHO J, BANSAL M. VL-ADAPTER: Parameter-efficient transfer learning for vision-and-language tasks[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5217-5227.

[39] PAN J, LIN Z, ZHU X, et al. St-adapter: Parameter-efficient image-to-video transfer learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 26462-26477.

[40] WU J D, FU R, FANG H H, et al. Medical sam adapter: Adapting segment anything model for medical image segmentation[EB/OL]. (2023-12-29) [2025-02-26]. https://arxiv.org/abs/2304.12620.

[41] CHEN T, ZHU L, DING C, et al. Sam fails to segment anything - sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more[EB/OL]. (2023-05-02)[2025-02-26]. https://arxiv.org/abs/2304.09148.

[42] GONG S Z, ZHONG Y, MA W A, et al. 3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation[J]. Medical Image Analysis, 2024, 98: 103324.

[43] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[44] RAJIČ F, KE L, TAI Y W, et al. Segment anything meets point tracking[EB/OL]. (2023-12-03)[2025-02-26]. https://arxiv.org/abs/2307.01197.

[45] HENDRYCKS D, GIMPEL K. Gaussian error linear units (gelus)[EB/OL]. (2023-06-06)[2025-02-26]. https://arxiv.org/abs/1606.08415.

[46] NAIR V, HINTON G E. Rectified linear units improve restricted Boltzmann machines[C]//The 27th International Conference on International Conference on Machine Learning. New York: ACM, 2010: 807-814.

[47] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[EB/OL]. (2021-06-03) [2025-02-26]. https://arxiv.org/abs/2010.11929.

[48] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//The 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000 - 6010.

[49] XIE E Z, WANG W H, YU Z D, et al. Segformer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.

[50] WANG L J, LU H C, WANG Y F, et al. Learning to detect salient objects with image-level supervision[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 3796-3805.

[51] SONG H M, WANG W G, ZHAO S Y, et al. Pyramid dilated deeper ConvLSTM for video salient object detection[C]// Computer Vision - ECCV 2018. Cham: Springer, 2018: 744-760.

[52] WANG W G, SONG H M, ZHAO S Y, et al. Learning unsupervised video object segmentation through visual attention[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 3059-3069.

[53] YANG Z, WANG Q, BERTINETTO L, et al. Anchor diffusion for unsupervised video object segmentation[C]// 2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 931-940.

[54] JI G P, FAN D P, FU K R, et al. Full-duplex strategy for video object segmentation[J]. Computational Visual Media, 2023, 9(1): 155-175.

[55] LIU D Z, YU D D, WANG C H, et al. F_2Net: Learning to focus on the foreground for unsupervised video object segmentation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2109-2117.

[56] ZHANG K H, ZHAO Z C, LIU D, et al. Deep transport network for unsupervised video object segmentation[C]// 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 8761-8770.

[57] REN S C, LIU W X, LIU Y T, et al. Reciprocal transformations for unsupervised video object segmentation[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 15430-15439.

[58] LEE M, CHO S, LEE S, et al. Unsupervised video object segmentation *via* prototype memory network[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2023: 5913-5923.

[59] LEE M, CHO S, LEE D, et al. Guided slot attention for unsupervised video object segmentation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 3807-3816.

[60] CHO S, LEE M, LEE S, et al. Dual prototype attention for unsupervised video object segmentation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 19238-19247.

[61] TEED Z, DENG J. RAFT: Recurrent all-pairs field transforms for optical flow[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 402-419.

[62] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2017-01-30) [2025-02-26]. https://arxiv.org/abs/1412.6980.

[63] PONT-TUSET J, PERAZZI F, CAELLES S, et al.

The 2017 davis challenge on video object segmentation[EB/OL]. (2018-03-01)[2025-02-26]. https://arxiv.org/abs/1704.00675.

[64] CHENG H K, OH S W, PRICE B, et al. Tracking anything with decoupled video segmentation[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 1316-1326.

[65] CHENG Y, LI L, XU Y, et al. Segment and track anything[EB/OL]. (2023-05-11)[2025-02-26]. https://arxiv.org/abs/2305.06558.

[66] YANG J, GAO M, LI Z, et al. Track anything: Segment anything meets videos[EB/OL]. (2023-04-28) [2025-02-26]. https://arxiv.org/abs/2304.11968.

[67] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6848-6856.

[68] CHEN Y P, DAI X Y, LIU M C, et al. Dynamic convolution: Attention over convolution kernels[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11027-11036.

[69] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More features from cheap operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1577-1586.

[70] CHEN L W, GU L, ZHENG D Z, et al. Frequency-adaptive dilated convolution for semantic segmentation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 3414-3425.

[71] CHEN L W, GU L, LI L, et al. Frequency dynamic convolution for dense image prediction[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Conference. Piscataway: IEEE, 2025: 30178-30188.

[72] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 936-944.

[73] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8759-8768.

[74] TAN M X, PANG R M, LE Q V. EfficientDet: Scalable and efficient object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10778-10787.

[75] CHEN L W, FU Y, GU L, et al. Frequency-aware feature fusion for dense image prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10763-10780.

作者简介

韩志冬　男,1999年12月出生于江苏省扬州市. 主要研究方向为视频目标分割.
E-mail: hanzd128@qq.com

宋慧慧　女,1986年6月出生于山东省聊城市. 现为南京信息工程大学自动化学院教授. 主要研究方向为遥感图像处理.
E-mail: songhuihui@nuist.edu.cn

胡升龙　男,2000年5月出生于江苏省宿迁市. 主要研究方向为协同显著性目标检测.
E-mail: hslnuist@163.com

张开华　男,1983年3月出生于山东省日照市. 现为南京信息工程大学自动化学院教授. 主要研究方向为视觉目标跟踪与分割.
E-mail: zhkhua@gmail.com