自动化学报
ACTA AUTOMATICA SINICA

基于运动引导的高效无监督视频目标分割网络

赵子成 张开华 樊佳庆 刘青山

Learning Motion Guidance for Efficient Unsupervised Video Object Segmentation

ZHAO Zi-Cheng, ZHANG Kai-Hua, FAN Jia-Qing, LIU Qing-Shan

在线阅读 View online: https://doi.org/10.16383/j.aas.c210626

基于运动引导的高效无监督视频目标分割网络

赵子成 [1] 张开华 [1] 樊佳庆 [1] 刘青山 [1]

摘 要 大量基于深度学习的无监督视频目标分割 (Unsupervised video object segmentation, UVOS) 算法存在模型参数量与计算量较大的问题, 这显著限制了算法在实际中的应用. 提出了基于运动引导的视频目标分割网络, 在大幅降低模型参数量与计算量的同时, 提升视频目标分割性能. 整个模型由双流网络、运动引导模块、多尺度渐进融合模块三部分组成. 具体地, 首先, RGB 图像与光流估计输入双流网络提取物体外观特征与运动特征; 然后, 运动引导模块通过局部注意力提取运动特征中的语义信息, 用于引导外观特征学习丰富的语义信息; 最后, 多尺度渐进融合模块获取双流网络的各个阶段输出的特征, 将深层特征渐进地融入浅层特征, 最终提升边缘分割效果. 在 3 个标准数据集上进行了大量评测, 实验结果表明了该方法的优越性能.

关键词 无监督视频目标分割, 运动引导, 局部注意力, 互注意力

引用格式 赵子成, 张开华, 樊佳庆, 刘青山. 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, 2023, **49**(4): 872−880

DOI 10.16383/j.aas.c210626

Learning Motion Guidance for Efficient Unsupervised Video Object Segmentation

ZHAO Zi-Cheng[1] ZHANG Kai-Hua[1] FAN Jia-Qing[1] LIU Qing-Shan[1]

Abstract Numerous unsupervised video object segmentation (UVOS) algorithms based on deep learning have superfluous model parameters and expensive computational overhead, which limits the applications of the algorithms in practice. To relieve the issues, this paper proposes an unsupervised video object segmentation network based on motion guidance, which can significantly reduce the number of model parameters and calculations, and improve the performance of segmentation. The multi-scale progressive fusion module consists of three parts. Specifically, RGB image and optical flow estimation are fed into the dual flow network to extract object appearance features and motion features. Then, the motion guidance module extracts semantic information from motion features through local attention to guide semantical appearance features learning. Finally, the multi-scale progressive fusion module obtains output features of each stage of dual flow network, and gradually integrates deep features with shallow features. Extensive evaluations are conducted on three mainstream datasets, and the results show the superior performance of the proposed method.

Key words Unsupervised video object segmentation (UVOS), motion guidance, local attention, co-attention

Citation Zhao Zi-Cheng, Zhang Kai-Hua, Fan Jia-Qing, Liu Qing-Shan. Learning motion guidance for efficient unsupervised video object segmentation. *Acta Automatica Sinica*, 2023, **49**(4): 872−880

无监督视频目标分割 (Unsupervised video object segmentation, UVOS) 目的是在没有任何人为干预的情况下从视频中自动分割出显著的对象. 这种自动分割主要目标的任务近年来受到了广泛的关注, 并在计算机视觉的许多领域产生了巨大的影响, 包括监控、机器人和自动驾驶等.

传统方法通常使用手工特征来解决这一问题, 例如运动边界[1]、稀疏表示[2]、显著性[3−4] 和点轨迹[2, 5−6]. 尽管上述算法取得了一定的成功, 但在准确发现整个视频序列中最显著的对象方面还不够理想. 随着深度学习的兴起, 最近的几项研究试着将这一问题建模为零目标帧问题[7−8]. 这些方法通常从大规模的训练数据中学习一个强大的对象表示, 然后调整模型来测试视频, 而不需要任何注释.

尽管上述方法取得了突破性的进展, 但是仍然存在问题. 上述方法使用重量级网络提取更好的特征表示, 例如基于 ResNet101 网络的 DeepLab v3 网络[9] 同时使用复杂的机制, 捕获显著物体. 这些导致了较大的模型参数量, 较高的模型计算量, 较

收稿日期 2021-07-06 录用日期 2021-10-18
Manuscript received July 6, 2021; accepted October 18, 2021
科技创新 2030 —— "新一代人工智能"重大项目 (2018AAA0100400), 国家自然科学基金 (61876088, U20B2065, 61532009), 江苏省 333 工程人才项目 (BRA2020291) 资助
Supported by National Key Research and Development Program of China (2018AAA0100400), National Natural Science Foundation of China (61876088, U20B2065, 61532009), and 333 High-level Talents Cultivation of Jiangsu Province (BRA2020291)
本文责任编委 黄华
Recommended by Associate Editor HUANG Hua
1. 南京信息工程大学自动化学院 南京 210044
1. School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044

慢的模型训练与推理速度, 限制了算法在实际场合中的应用.

如何高效捕获显著物体是网络轻量化的关键. 在最近的研究中, 使用互注意力机制在不同视频帧之间捕获相似物体[10], 取得了较好效果, 但不能区分背景中与显著目标相似的物体, 且计算量较大. 基于人对运动物体的敏感性, 可以利用运动信息捕获显著物体. 同时由于视频中物体缓慢移动的先验信息, 基于局部匹配的运动信息提取方法较为高效, 因此本文使用光流估计网络提取运动信息.

同时, 提取 RGB 图像中物体的外观特征来补充运动信息缺少的具体细节, 提升最终分割效果. 因为 RGB 图像与光流估计存在像素点对应的关系, 光流估计中的运动信息又包含了显著物体的大致位置与轮廓信息, 所以可以在运动信息中使用局部注意力机制得到卷积权重, 引导外观特征学习语义, 减低 RGB 图像支路的特征提取难度. 这种运动信息引导外观信息学习的方法, 使得本文算法在使用轻量级特征提取器的同时, 可以获得良好的特征提取质量, 降低了模型参数量与模型计算量. 最后, 将提取的多个阶段特征送入多尺度渐进融合模块, 经过卷积与上采样的组合, 不断增强高分辨率特征的语义信息, 得到更加准确的分割结果.

本文主要贡献如下:

1) 提出一种轻量级无监督视频目标分割算法, 大幅缩小模型参数量与模型计算量, 显著提升了无监督视频目标分割算法的速度.

2) 基于运动先验信息, 设计出一种基于局部注意力的运动引导模块, 通过局部注意力提取运动信息中的语义信息, 并以卷积权重的形式引导外观特征学习语义, 最终提升分割性能.

3) 与当前最先进的方法相比, 本文方法在多个标准数据集上取得了具有竞争力的实验结果, 表明了本文算法的有效性, 取得速度与精度的平衡.

1 相关工作

1.1 无监督视频目标分割

早期的无监督视频目标分割模型通常分析点轨迹[2, 5−6]、物体建议[11]、运动边界[1] 或显著性信息[3−4] 来推断目标, 但是受制于数据集、算力等多方面的限制, 效果不理想. 近年来, 得益于大型数据集的建立[12−13] 与全卷积分割网络发展, 多种方法提出用零目标帧解决方案来解决这一问题.

一种分割显著物体的方法是通过视频显著物体检测[14]. 该方法对预先训练好的语义分割网络进行微调, 提取空间显著性特征, 然后训练卷积长短期记忆 (Convolution long short-term memory, ConvLSTM) 捕捉时序信息. 随着注意力机制的出现, 新研究使用带有互注意力机制的孪生网络[10], 在视频不同帧之间获取空间与时序信息进行推理, 但是不能很好区分背景中与显著物体相似的物体. 双流网络也是一种流行的选择[15−16], 融合运动与外观信息一起进行对象推理. 例如, 运动注意转换网络 (Motion-attentive transition network, MATNet)[17] 中, 使用互注意力机制在双流网络各个阶段之间融合运动与外观特征获取显著性特征, 取得了较好结果. 然而, 这些研究使用的互注意力机制带来巨大的计算量问题, 这限制了实际场合中的应用.

1.2 互注意力机制

神经网络中的注意机制受到人类感知的启发, 在深层神经网络中得到了广泛研究. 通过端到端的训练, 注意机制允许网络有选择地注意输入的子集. 例如, 利用多上下文注意进行人体姿势估计[18], 利用空间注意和通道注意两种方法, 来动态选择一个图像部分作为图像描述[19]. 最近, 视觉和语言任务中的共同注意机制得到了研究, 例如视觉问答[20] 和视觉对话[21]. 在这些工作中, 共同注意机制被用来挖掘不同模式之间的潜在相关性. 例如, 在之前的视觉问答研究[20] 中创建了一个模型, 该模型联合执行问题引导的视觉注意和图像引导的问题注意. 这样, 学习的模型可以选择性地聚焦于图像区域和文档片段. 本文的注意力模型是受这些文献启发, 它被用来在具有先验信息的特征之间挖掘信息, 以一个更优雅的网络架构来捕捉运动信息, 引导外观信息学习显著性特征.

2 本文方法

如图 1 所示, 本文提出一种端到端网络, 主要由双流网络、运动引导模块、多尺度渐进融合模块三个部分组成.

2.1 双流网络

本文构建一个双流网络来提取运动特征与外观特征, 这在许多相关的视频任务中被证明是有效的. 不同于以往研究, 本文使用轻量网络替代基于 ResNet101 的 DeepLab v3 网络[9], 并在轻量网络的不同阶段, 插入运动引导模块来增强外观特征的语义. 考虑到对推理速度与分割效果的平衡, 本文使用 MobileNet v2 网络[22] 作为双流网络的每条支路的特征提取器. 对于双流网络, 给定一张图片 $\boldsymbol{I}_a \in \mathbf{R}^{3\times H\times W}$ 与对应的光流估计 $\boldsymbol{I}_m \in \mathbf{R}^{3\times H\times W}$, 双流网络在第 i 阶段提取外观特征 $\boldsymbol{V}_{a,i} \in \mathbf{R}^{C\times H\times W}$ 与运动特

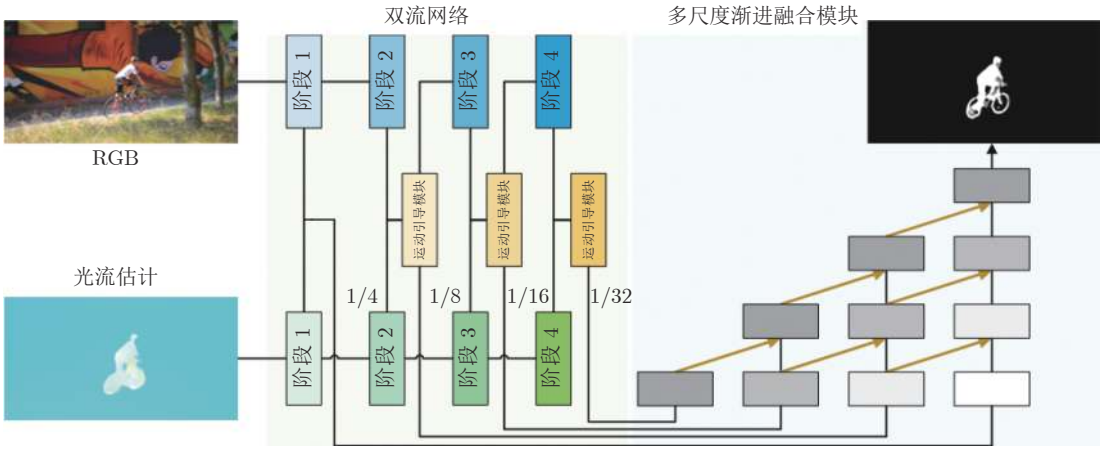


图 1 网络框架图

Fig. 1 Figure of network structure

征 $\boldsymbol{V}_{m,i} \in \mathbf{R}^{C\times H\times W}$, 送入第 i 个阶段的运动引导模块增强外观特征:

$$\boldsymbol{U}_{a,i} = F_{MG}(\boldsymbol{V}_{a,i}, \boldsymbol{V}_{m,i}) \tag{1}$$

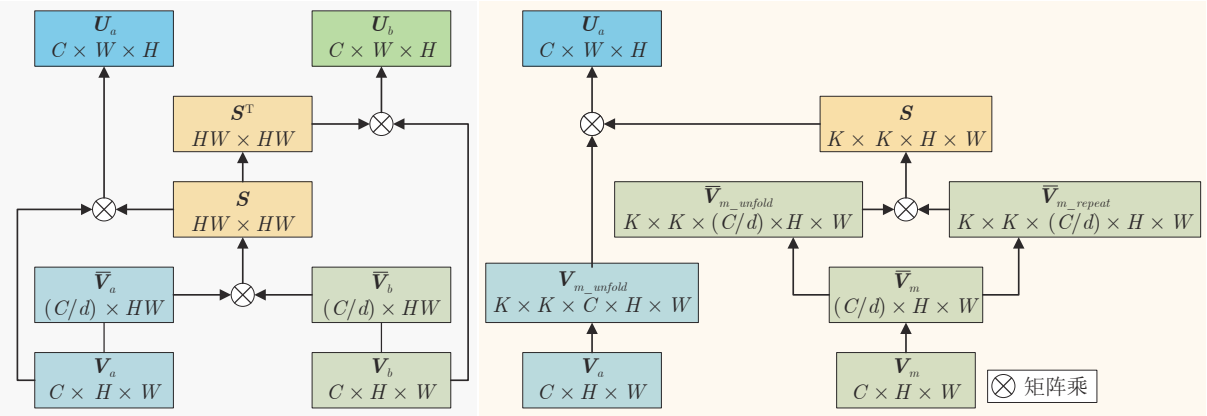
式中, $F_{MG}(\cdot)$ 表示运动引导模块, $\boldsymbol{U}_{a,i} \in \mathbf{R}^{C\times H\times W}$ 表示增强后的第 i 个阶段外观特征. 特别地, 在双流网络的第 1 阶段不设置运动引导模块, 以保留浅层特征的细节信息. 对于网络第 i 阶段的增强外观特征 $\boldsymbol{U}_{a,i}$ 与运动特征 $\boldsymbol{V}_{m,i}$, 在通道维度拼接得到 $\boldsymbol{U}_i = Concat(\boldsymbol{U}_{a,i}, \boldsymbol{U}_{m,i}) \in \mathbf{R}^{2C\times H\times W}$ 后, 送入多尺度渐进融合模块, 得到最终分割图.

2.2 运动引导模块

互注意力机制被广泛应用于提取不同模态特征中的关联信息, 在协同注意力孪生网络 (Co-attention siamese networks, COSNet)[10] 中使用互注意力机制, 提取同一视频中多帧的特征之间的关联信息, 在 MATNet[17] 中使用互注意力机制, 将外观特征转化为运动注意力表示. 互注意力机制的大量运用在取得良好结果的同时, 有着计算量巨大的问题, 因此改进互注意机制可以带来可观的效率提升.

朴素的互注意力机制如图 2(a) 所示. 特征 $\boldsymbol{V}_a \in \mathbf{R}^{C\times H\times W}$ 与特征 $\boldsymbol{V}_b \in \mathbf{R}^{C\times H\times W}$ 送入互注意力模块, 通过 1×1 卷积压缩通道到 C/d, 后调整维度得到 $\overline{\boldsymbol{V}}_a \in \mathbf{R}^{C\times HW}$ 与 $\overline{\boldsymbol{V}}_b \in \mathbf{R}^{C\times HW}$, 计算 $\overline{\boldsymbol{V}}_a$ 与 $\overline{\boldsymbol{V}}_b$ 中特征点相似度, 得到相似度矩阵 $\boldsymbol{S} \in \mathbf{R}^{HW\times HW}$, 矩阵 $\boldsymbol{S}$ 与其转置矩阵 $\boldsymbol{S}^{\mathrm{T}}$ 分别与 $\boldsymbol{V}_a$ 与 $\boldsymbol{V}_b$ 做矩阵乘法, 恢复空间维度后得到增强后的特征 $\boldsymbol{U}_a \in \mathbf{R}^{C\times H\times W}$ 与特征 $\boldsymbol{U}_b \in \mathbf{R}^{C\times H\times W}$.

本文从加权求和的角度, 分析互注意力机制的优势. 在互注意力中, 特征 $\boldsymbol{U}_b$ 中的每个特征点 $\boldsymbol{I}_b \in$



(a) 互注意力模块
(a) Co-attention module

(b) 运动引导模块
(b) Motion guidance module

图 2 注意力结构

Fig. 2 Attention structure

$\mathbf{R}^{C\times1\times1}$, 由一组权重 $\boldsymbol{W}\in\mathbf{R}^{1\times HW}$ 对特征 $\overline{\boldsymbol{V}}_a$ 中的每一个特征点加权求和得到, 这组权重 $\boldsymbol{W}$ 由 $\overline{\boldsymbol{V}}_b$ 中对应位置的特征点 $\boldsymbol{I}_b$ 与 $\overline{\boldsymbol{V}}_a$ 中所有特征点的相似度矩阵归一化得到. 这种方式类似多层感知机 (Multi-layer perceptron, MLP), 全局的计算获得全局的感受野. 不同的是, 在 MLP 中的权重是可学习参数, 互注意力中以相似度方式定义权重, 不需要进行学习, 降低了过拟合风险.

互注意力机制获得了全局的感受野, 同时避免了像 MLP 一样增加可学习参数, 但是存在计算量较大的问题. 使用局部替代全局将大幅减少计算量, 在合理利用特征先验信息的情况下, 不会导致模型性能下降. 类似卷积对 MLP 的改进, 本文使用滑窗的方式得到局部注意力.

具体地, 计算运动特征 $\boldsymbol{V}_m$ 中的每一个特征点 $\boldsymbol{I}_m$ 与其周围 K 窗口内的特征点之间的相似度, 归一化得到相似度矩阵 $\boldsymbol{W}\in\mathbf{R}^{1\times K\times K}$. 外观特征 $\boldsymbol{V}_a$ 中对应位置的特征点 $\boldsymbol{I}_a$, 使用相似度矩阵 $\boldsymbol{W}$ 对其周围 K 窗口内的特征点做加权求和, 得到 $\boldsymbol{U}_a$ 的中特征点 $\dot{\boldsymbol{I}}_a$. 通过这种方式, 运动特征通过局部注意力提取语义信息获得加权权重, 并通过传递权重给外观特征引导加权求和的方式引导学习高级语义.

通过现有框架实现的运动引导模块并行计算如图 2(b) 所示. 外观特征 $\boldsymbol{V}_a\in\mathbf{R}^{C\times H\times W}$ 按 im2col 方式展开并调整维度得到 $\overline{\boldsymbol{V}}_{a\_unfold}\in\mathbf{R}^{K\times K\times C\times H\times W}$. 运动特征 $\boldsymbol{V}_m\in\mathbf{R}^{C\times H\times W}$ 经过一层 1×1 卷积压缩通道得到 $\overline{\boldsymbol{V}}_m\in\mathbf{R}^{(C/d)\times H\times W}$, 按 im2col 方式展开 $\overline{\boldsymbol{V}}_m$ 并重新排列维度得到 $\overline{\boldsymbol{V}}_{m\_unfold}\in\mathbf{R}^{K\times K\times(C/d)\times H\times W}$, 复制 $\overline{\boldsymbol{V}}_m$ 特征点 $K\times K$ 次并重新排列维度得到 $\overline{\boldsymbol{V}}_{m\_repeat}\in\mathbf{R}^{K\times K\times(C/d)\times H\times W}$. 特征 $\overline{\boldsymbol{V}}_{m\_unfold}$ 与特征 $\overline{\boldsymbol{V}}_{m\_repeat}$ 在通道维度上做相似度, 得到相似度矩阵 $\boldsymbol{S}\in\mathbf{R}^{(K\times K)\times H\times W}$. 特征 $\overline{\boldsymbol{V}}_{a\_unfold}$ 与相似度矩阵 $\boldsymbol{S}$ 做矩阵乘, 得到特征 $\boldsymbol{U}_a\in\mathbf{R}^{C\times H\times W}$.

本文的运动引导模块类似卷积, 不同点在于, 通过相似度方式定义的滑窗卷积权重对于特征图中的每一个特征点是动态的, 且不需要进行学习.

对比互注意力模块, 运动引导模块大幅降低了计算量, 不同输入尺寸下计算量对比如表 1 所示. 同时, 运动引导模块可以通过限制最大关联距离 (滑窗的大小 K) 来平衡模型对运动信息的提取能力与对背景噪声的抑制能力. 具体地, 过小的 K 无法获得足够的运动信息; 过大的 K 增加计算量, 并可能提取到与前景物体相似的背景物体的运动信息. 特别地, 当 K 为 1 时, 运动引导模块退化为运动特征 $\boldsymbol{V}_m$ 与外观特征.

表 1　不同模块每秒浮点运算数对比

Table 1　Comparison of floating-point operations per second of different modules

输入尺寸 (像素)	互注意模块 (MB)	运动引导模块 (MB)
$64\times64\times16$	10.0	2.3
$64\times32\times32$	153.1	9.0

$\boldsymbol{V}_a$ 进行逐元素点乘, 不具备在运动特征 $\boldsymbol{V}_m$ 中获得局部注意力的能力. 因此, 选取合适的 K 对于运动引导模块十分重要. 除了直接调整 K 的取值, 还可以调整模块的堆叠层数来模拟较大 K 值模块的效果, 这进一步降低了计算量, 提升了最终效果. K 取值和模块堆叠次数对模型性能的影响, 将在第 3.6 节进行实验分析.

2.3　多尺度渐进融合模块

双流网络不同阶段提取的特征拥有不同的分辨率, 包含不同层次的语义信息, 合理使用这些特征显得尤为重要. 之前的研究采取 UNet[23] 方式的上采样融合策略, 同时使用空洞空间卷积池化金字塔增大各个阶段的感受野, 但是忽略了不同阶段特征在语义层面的融合差别.

如图 3 所示, 分割结果图包含的语义信息可以看作高级语义的子集, 低分辨率的深层语义特征融合高分辨率的浅层语义特征, 得到高分辨率的高级语义特征. 但是, 随着融合的不断进行, 待融合的低分辨率特征与高分辨率特征之间的语义鸿沟将会加大, 这不利于融合权重的学习, 降低了分割性能.

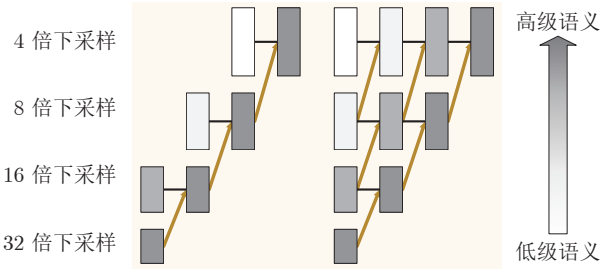


图 3　UNet 方式的上采样与多尺度渐进融合模块

Fig. 3　Upsampling module and multi-scale progressive fusion module in UNet mode

因此, 本文提出多尺度渐进融合模块, 采取不断将高级语义融合进高分辨特征的策略. 双流网络提取的多阶段特征分别送入处理不同分辨率特征的多阶段支路中, 并在每个阶段之间会由低分辨率特征向高分辨率特征进行融合.

具体地, 对于第 $j-1$ 阶段第 i 条支路特征 $\boldsymbol{U}_{j-1,i}\in\mathbf{R}^{2C\times(H/2)\times(W/2)}$, 先进行 2 倍上采样, 后与

第 $j-1$ 阶段第 $i-1$ 条支路特征 $\boldsymbol{U}_{j-1,i-1}\in\mathbf{R}^{C\times H\times W}$ 在通道维度进行拼接, 再送入两层残差结构调整通道数量, 并进行融合语义信息, 最终得到第 j 阶段第 $i-1$ 个特征 $\boldsymbol{U}_{j,i-1}\in\mathbf{R}^{C\times H\times W}$.

$$\boldsymbol{U}_{j,i-1}=F_{conv}(Concat(\boldsymbol{U}_{j-1,i-1},Up(\boldsymbol{U}_{j-1,i}))) \quad (2)$$

$$F_{conv}(*)=F_{res}(F_{res}(*)) \quad (3)$$

式中, $*$ 代表输入特征. 通过这种方式, 降低融合特征语义之间差距, 提升了最终分割结果.

3 实验设置与结果分析

3.1 训练细节

本文采用在 ImageNet 数据集上预训练的 MobileNet v2 网络[22] 作为双流网络特征提取器, 使用二值交叉熵损失函数作为训练的损失函数. 训练数据分为 Youtube-VOS 数据集[13] 和 DAVIS-16 数据集[12] 两部分. 因为 Youtube-VOS 数据集实际存在类别标签, 不利于类别无关的视频分割任务的训练, 且分割标注的准确度低于 DAVIS-16 数据集, 所以本文选择在 Youtube-VOS 数据集上预训练模型, 在 DAVIS-16 数据集上微调模型进行测试. 同时为了保证公平, 在 Youtube-VOS 数据集上采用间隔抽帧的方式, 得到 9 000 张训练图像, 加上 DAVIS-16 数据集的 2 000 张训练图像, 共计 11 000 张训练图像, 与其他算法训练集规模持平. 本文使用在研究中常用的 PWCNet 网络预先处理数据集, 得到光流估计图像.

本文使用常用的数据增广策略, 对于每一张训练图片, 随机翻转后采取 $-10^{\circ}\sim 10^{\circ}$ 的随机角度旋转图片, 后裁剪并缩放到 384×672 像素尺寸. 网络预训练阶段微调阶段均使用随机梯度下降优化器, 特征提取器与运动引导模块使用 1×10^{-4} 的学习率, 多尺度渐进融合模块使用 1×10^{-3} 的学习率, 学习率衰减率和权重衰减率分别为 0.9 和 5×10^{-4}, 批量大小均为 10 (张/批). 预训练迭代 25 轮, 微调迭代 10 轮. 使用 PyTorch 1.6.0 框架搭建网络, 并在 1 张 GeForce GTX 2080 Ti GPU 上训练并测试模型.

3.2 数据集

本文在 DAVIS-16 数据集[12]、FBMS 数据集[2] 和 ViSal 数据集[24] 上测试模型性能.

1) DAVIS-16 数据集由 50 个视频组成, 30 个视频用于训练, 20 个视频用于测试.

2) FBMS 数据集由 59 个视频组成, 29 个用于训练, 30 个用于测试. 采用每 20 帧标注一帧的稀疏标注策略.

3) ViSal 数据集由 17 个测试视频组成, 共 193 帧标注图片.

3.3 评价指标

对于无监督视频分割任务, 本文采用 DAVIS-16 的标准评价指标, 区域相似度 J 和轮廓精度 F. 其中, J 为分割结果和标注真值掩模的交并比:

$$J=\frac{|\boldsymbol{M}\cap\boldsymbol{GT}|}{|\boldsymbol{M}\cup\boldsymbol{GT}|} \quad (4)$$

式中, $\boldsymbol{M}$ 表示预测的分割结果, $\boldsymbol{GT}$ 表示分割真值掩模.

F 将掩模视为系列闭合轮廓的集合, 计算基于轮廓的 F 度量:

$$F=\frac{2P\times R}{P+R} \quad (5)$$

式中, P 为准确率, R 为召回率.

另外, 本文采用综合指标 $J\&F$, 表示两者的均值:

$$J\&F=\frac{J+F}{2} \quad (6)$$

本文使用平均绝对误差 (Mean absolute error, MAE) 和 F_β 评价模型, 对视频显著性进行检测.

MAE 描述了二值显著性图与真图的像素级的直接比较:

$$\text{MAE}=\frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\|\boldsymbol{S}_{i,j}-\boldsymbol{G}_{i,j}\| \quad (7)$$

式中, $\boldsymbol{S}$ 表示二值显著性图, $\boldsymbol{G}$ 表示真图, W 和 H 对应图像的宽和高.

F_β 是综合准确率和召回率的评价指标, 可以较为全面地反映算法的性能:

$$F_\beta=\frac{(1+\beta^2)P\times R}{\beta^2P+R} \quad (8)$$

式中, 加权调和参数 β^2 常被设置为 0.3.

3.4 结果对比

表 2 对比了本文算法与其他几种先进算法在 DAVIS-16 数据集[12] 和 FBMS 数据集[2] 上的表现. 在 DAVIS-16 数据集中, 本文采用 J、F 和 $J\&F$ 评价指标作为参考. 在 FBMS 数据集上, 本文采用 J 指标作为参考. 本文算法没有采用除去翻转之外的任何后处理方法, 例如 COSNet[10]、MATNet[17]中, 使用的条件随机场后处理方法. 本文算法在 DAVIS-16 数据集上, 以 $J\&F=83.6\%$ 位于第 1. 在 FBMS 数据集上, 以 $J=75.9\%$ 位于第 2, 仅次于

表 2　不同方法在 DAVIS-16 和 FBMS 数据集的评估结果 (%)

Table 2　Evaluation results of different methods on DAVIS-16 and FBMS datasets (%)

方法	DAVIS-16			FBMS
	$J\&F$	J	F	J
LMP[25]	68.0	70.0	65.9	—
LVO[16]	74.0	75.9	72.1	—
PDB[14]	75.9	77.0	74.5	74.0
MBNM[26]	79.5	80.4	78.5	73.9
AGS[27]	78.6	79.7	77.4	—
COSNet[10]	80.0	80.5	79.4	75.6
AGNN[7]	79.9	80.7	79.1	—
AnDiff[28]	81.1	81.7	80.5	—
MATNet[17]	81.6	82.4	80.7	**76.1**
本文算法	**83.6**	**83.7**	**83.4**	75.9

MATNet, 相差 0.2%. 本文算法在 DAVIS-16 数据集上取得较好结果, 主要归功于两个方面: 1) 运动引导模块的局部注意力抑制了大量背景噪声; 2) 多尺度渐进融合模块配合相对较大的输入分辨率, 提升分割结果. 值得注意的是, 本文算法在 FBMS 数据集的指标明显低于在 DAVIS-16 数据集上的指标, 是由于光流估计网络 PWCNet 在 FBMS 数据集上效果不佳, 无法获得较好的运动信息引导外观特征.

视频显著性检测任务的目的, 在于通过联合空间和时间信息实现视频序列中与运动相关的显著性目标的连续提取. 由于无监督视频目标分割与视频显著性检测的任务相似性, 本文同样测试模型 DAVIS-16[12]、FBMS[2]、ViSal[24] 三个数据集上的视频显著性检测指标, 使用 MAE 和 F_β 指标作为依据, 结果如表 3 所示. 本文算法在 DAVIS-16数据集上获得了最好的指标, 同时在 FBMS 数据和 ViSal 数据集获得具有竞争力的指标, 表明了本文方法的有效性.

由于本文算法选择了轻量级网络与局部注意力模块, 除去在标准数据集上的良好表现外, 同样在模型参数量与模型推理速度上具有优势. 表 4 对比了本文算法与两种最先进方法的模型参数量、模型计算量与推理时延. 算法测试不使用后处理, 同时为了排除不同数据加载方式对模型推理速度的干扰, 本文仅测试输入对应分辨率且批量为 1 的随机矩阵时模型的推理速度. 首先模型推理 10 轮预热, 然后推理 60 轮统计用时, 分别去掉用时最高与最低的 20 轮, 统计剩余 20 轮的平均时间得到推理时延. 通过表 3 的对比实验可以看出, 本文算法有效降低

表 3　不同方法在 DAVIS-16、FBMS 和 ViSal 数据集的评估结果 (%)

Table 3　Evaluation results of different methods on DAVIS-16、FBMS and ViSal datasets (%)

方法	DAVIS-16		FBMS		ViSal	
	MAE	F_β	MAE	F_β	MAE	F_β
FCNS[29]	5.3	72.9	10.0	73.5	4.1	87.7
FGRNE[30]	4.3	78.6	8.3	77.9	4.0	85.0
TENet[31]	1.9	90.4	**2.6**	**89.7**	**1.4**	**94.9**
MBNM[26]	3.1	86.2	4.7	81.6	4.7	—
PDB[14]	3.0	84.9	6.9	81.5	2.2	91.7
AnDiff[28]	4.4	80.8	6.4	81.2	3.0	90.4
本文算法	**1.4**	**92.4**	5.9	84.2	1.9	92.1

表 4　不同方法的模型参数量、计算量与推理时延

Table 4　Model parameters, computation and infer latency of different methods

算法	COSNet[8]	MATNet[17]	本文算法
输入尺寸 (像素)	473 × 473	473 × 473	384 × 672
参数量 (MB)	81.2	142.7	6.4
计算量 (GB)	585.5	193.7	5.4
时延 (ms)	65	78	15

了模型参数与模型计算量, 这在实际应用中具有更多的优势. 同时, 本文算法在更高分辨率输入图像的情况下, 推理时延只有 15 ms, 对比同样使用运动特征的 MATNet[17] 方法, 推理速度提升 5.2 倍. 考虑到本文算法内存消耗较少, 因此在相同设备上具有更大的并发量.

为了验证本文算法的高效性, 本文测试对比算法在 GeForce GTX2080 Ti 上的运行性能, 结果如表 5 所示.

表 5　不同方法在 GTX2080 Ti 上的性能表现

Table 5　Performance of different methods on GTX2080 Ti

方法	并发量	每秒帧数	时延 (ms)
MATNet[17]	18	16	62.40
本文算法	**130**	**161**	**6.21**

得益于较低的参数量和计算量, 在充分利用 11 GB 显示器存储情况下, 本文算法具有更高的并发能力, 可以同时处理 130 帧图片, 对比 MATNet 提升 7.2 倍. 同时, 本文算法每秒帧数达到 161 帧/秒, 平均推理时延只有 6.21 ms. 表明了本文算法的高效性.

3.5　分割结果对比

图 4 对比展示了本文算法与其他方法的分割结

图 4 分割结果对比展示

Fig. 4 Comparative display of segmentation results

果. 由图 4 可以看出, 本文算法可以较好抑制背景噪声.

3.6 消融实验

表 6 展示了本文算法在 DAVIS-16 数据集[12]上的消融实验结果, FG 代表运动引导模块, U 代表多尺度渐进融合模块. 使用 J 指标与 F 指标作为主要依据. 基线模型基于 MobileNet v2 网络[22] 的双流网络提取运动与外观特征, 并在网络的每个阶段通过运动与外观特征矩阵点乘融合语义, 最终通过 UNet[23] 方式上采样得到分割结果. 基线模型在 DAVIS-16 数据集上, 仅得到了 $J=$ 75.8% 和 $F=73.5\%$ 的结果. 通过加入多尺度渐进融合模块改善了边缘 F 指标, 由 73.5% 上升至 75.6%. 通过加入运动引导模块大幅改善了分割性能. 同时, 本文通过在双流网络中插入不同参数的运动引导模块, 探索运动引导模块的最佳效果. 如表 7 所示, 通过加入 K 为 3 的运动引导模块, 模型取得了大幅度的性能提升, 对比加入多尺度渐进融合模块的基线模型, J 指标提升了 6.7%, F 指标提升了 6.8%.

表 6 运动引导模块与多尺度渐进融合模块的消融实验 (%)

Table 6 Ablation experiment on motion guidance module and multi-scale progressive fusion module (%)

指标	本文算法	无 FG	FG
J	83.7	75.8	76.1
F	83.4	73.5	75.6

通过实验可以看出, 随着 K 值的扩大, 模型性能出现先升后降的现象. 这主要是随着 K 的增加, 局部注意力获得的运动信息变多, 受到背景噪声的影响也在变大. 最终在 K 为 7 时, 取得较好平衡.

类似卷积中使用多层 3×3 卷积模拟更大卷积的方式, 本文也探索了堆叠运动引导模块带来的影响. 通过堆叠两层 K 为 3 的模块, 模拟了 K 为 5 的模块效果, 降低计算量的同时, 获得了更好的结果

表 7 不同核 K 大小与堆叠次数对比

Table 7 Comparison of different Kernel sizes and cascading times

K	堆叠层数	J (%)	F (%)
3	1	82.8	82.4
3	2	83.4	82.7
3	3	**83.7**	**83.4**
3	4	83.5	83.2
5	1	83.2	82.6
7	1	83.4	82.7
9	1	83.1	82.4

表现. 本文将此归结于, K 为 5 的模块实际只进行了 1 次语义提取, 替换为相似的 K 为 3 的模块可以进行 2 次提取语义信息, 最终性能超过 K 较大时的模型. 同时, 堆叠运动引导模块同样出现了随着 K 的增大, 性能先升后降的现象. 由表 7 可以看出, 本文算法选择堆叠 3 层 K 为 3 的运动引导模块的模型, 作为本文的最终模型.

3.7 分割结果展示

图 5 展示了本文算法的分割结果. 可见本文算法在多种挑战场景下性能出色. 在第 1 行中, 本文算法可以较好区分显著前景与背景中相似物体; 在第 2 行和第 3 行中, 本文算法可以从嘈杂背景中准确分割显著物体; 在第 4 行和第 5 行中, 本文算法可以较好处理物体遮挡情况; 在第 6 行中, 本文可以较好处理多个显著前景目标. 可视化结果表明了本文算法的有效性.

4 结束语

本文提出了一种基于运动引导的无监督视频目标分割算法. 首先, 通过双流网络提取运动与外观特征; 然后, 经过运动引导模块引导外观特征学习显著的特征, 从而避免重量级特征提取器与互注意力机制带来的巨大计算量; 最后, 多尺度渐进融合模块不断将高级语义融入到浅层特征中, 得到最终

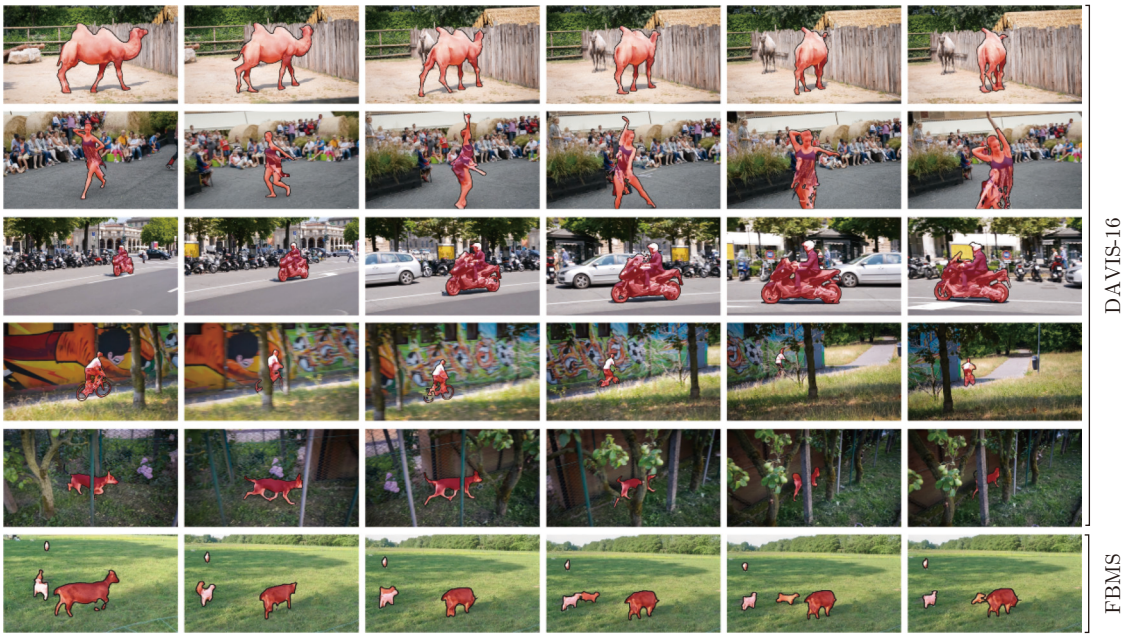


图 5 分割结果展示

Fig. 5 Display of segmentation results

预测的分割结果. 在多个标准评测数据集上的实验结果, 都充分验证了本文算法的优越性.

References

1 Papazoglou A, Ferrari V. Fast object segmentation in unconstrained video. In: Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia: IEEE, 2013. 1777−1784

2 Huang Hong-Tu, Bi Du-Yan, Hou Zhi-Qiang, Hu Chang-Cheng, Gao Shan, Zha Yu-Fei, et al. Research of sparse representation-based visual object tracking: A survey. *Acta Automatica Sinica*, 2018, **44**(10): 1747−1763
(黄宏图, 毕笃彦, 侯志强, 胡长城, 高山, 查宇飞, 等. 基于稀疏表示的视频目标跟踪研究综述. 自动化学报, 2018, **44**(10): 1747−1763)

3 Wang W, Shen J, Porikli F. Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE, 2015. 3395−3402

4 Qian Sheng, Chen Zong-Hai, Lin Ming-Qiang, Zhang Chen-Bin. Saliency detection based on conditional random field and image segmentation. *Acta Automatica Sinica*, 2015, **41**(4): 711−724
(钱生, 陈宗海, 林名强, 张陈斌. 基于条件随机场和图像分割的显著性检测. 自动化学报, 2015, **41**(4): 711−724)

5 Ochs P, Brox T. Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In: Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 1583−1590

6 Su Liang-Liang, Tang Jun, Liang Dong, Wang Nian. A video co-segmentation algorithm by means of maximizing submodular function and RRWM. *Acta Automatica Sinica*, 2016, **42**(10): 1532−1541
(苏亮亮, 唐俊, 梁栋, 王年. 基于最大化子模和 RRWM 的视频协同分割. 自动化学报, 2016, **42**(10): 1532−1541)

7 Ventura C, Bellver M, Girbau A, Salvador A, Marques F, Giroinieto X. RVOS: End-to-end recurrent network for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 5277−5286

8 Wang W, Lu X, Shen J, Crandall D J, Shao L. Zero-shot video object segmentation via attentive graph neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 9236−9245

9 Chen L C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv: 1706.05587, 2017.

10 Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 3623−3632

11 Faktor A, Irani M. Video segmentation by non-local consensus voting. In: Proceedings of the British Machine Vision Conference. Nottingham, UK: 2014.

12 Perazzi F, Pont-Tuset J, McWilliams B, Van-Gool L, Gross M, Sorkine-Hornung A. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 724−732

13 Xu N, Yang L J, Fan Y C, Yang J C, Yue D C, Liang Y C, et al. Youtube-VOS: Sequence-to-sequence video object segmentation. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: 2018. 585−601

14 Song H, Wang W, Zhao S, Shen J, Lam K M. Pyramid dilated deeper ConvLSTM for video salient object detection. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: 2018. 715−731

15 Jampani V, Gadde R, Gehler P V. Video propagation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 451−461

16 Tokmakov P, Alahari K, Schmid C. Learning video object segmentation with visual memory. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 4481−4490

17 Zhou T, Li J, Wang S, Tao R, Shen J. Matnet: Motion-attentive transition network for zero-shot video object segmentation.

IEEE Transactions on Image Processing, 2020, **29**: 8326−8338

18 Chu X, Yang W, Ouyang W, Ma C, Yuille A L, Wang X. Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 1831−1840

19 Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 5659−5667

20 Lu J, Yang J, Batra D, Parikh D. Hierarchical question-image co-attention for visual question answering. arXiv preprint arXiv: 1606.00061, 2016.

21 Wu Q, Wang P, Shen C, Reid I, Van-Den-Hengel A. Are you talking to me? Reasoned visual dialog generation through adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6106−6115

22 Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L C. MobileNet v2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 4510−4520

23 Ronneberger O, Fischer P, Brox T. UNet: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention. Munich, Germany: 2015. 234−241

24 Wang W, Shen J, Shao L. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 2015, **24**(11): 4185−4196

25 Tokmakov P, Alahari K, Schmid C. Learning motion patterns in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017. 3386−3394

26 Li S, Seybold B, Vorobyov A, Lei X, Kuo C C J. Unsupervised video object segmentation with motion-based bilateral networks. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: 2018. 207−223

27 Wang W, Song H, Zhao S, Shen J, Zhao S, Hoi S C, et al. Learning unsupervised video object segmentation through visual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 3064−3074

28 Yang Z, Wang Q, Bertinetto L, Hu W, Bai S, Torr P H. Anchor diffusion for unsupervised video object segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 931−940

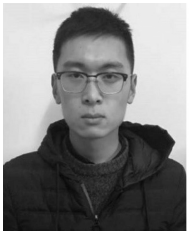
29 Wang W, Shen J, Shao L. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 2017, **27**(1): 38−49

30 Li G, Xie Y, Wei T, Wang K, Lin L. Flow guided recurrent neural encoder for video salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 3243−3252

31 Ren S, Han C, Yang X, Han G, He S. TENet: Triple excitation network for video salient object detection. In: Proceedings of the European Conference on Computer Vision. Edinburgh, Scotland: 2020. 212−228

赵子成 南京信息工程大学自动化学院硕士研究生. 主要研究方向为视频目标分割, 深度学习.
E-mail: 20191222013@nuist.edu.cn
(**ZHAO Zi-Cheng** Master student at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers video object segmentation and deep learning.)

张开华 南京信息工程大学自动化学院教授. 主要研究方向为视频目标分割, 视觉追踪. 本文通信作者.
E-mail: zhkhua@gmail.com
(**ZHANG Kai-Hua** Professor at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers video object segmentation and visual tracking. Corresponding author of this paper.)



樊佳庆 南京信息工程大学自动化学院硕士研究生. 主要研究方向为视频目标分割. E-mail: jqfan@nuaa.edu.cn
(**FAN Jia-Qing** Master student at the School of Automation, Nanjing University of Information Science and Technology. His main research interest is video object segmentation.)

刘青山 南京信息工程大学自动化学院教授. 主要研究方向为视频内容分析与理解. E-mail: qsliu@nuist.edu.cn
(**LIU Qing-Shan** Professor at the School of Automation, Nanjing University of Information Science and Technology. His research interest covers video content analysis and understanding.)