中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)05-1233-19

论文引用格式:Luo S H, Yuan X and Liang Y S. 2024. Multiframe spatiotemporal attention-guided semisupervised video segmentation. Journal of Image and Graphics, 29(05):1233-1251(罗思涵, 袁夏, 梁永顺. 2024. 多帧时空注意力引导的半监督视频分割. 中国图象图形学报, 29(05):1233-1251)[DOI:10.11834/jig.230606]

多帧时空注意力引导的半监督视频分割

罗思涵[1],袁夏[1*],梁永顺[2]

1. 南京理工大学计算机科学与工程学院,南京　210094; 2. 南京理工大学数学与统计学院,南京　210094

摘　要:**目的**　传统的半监督视频分割多是基于光流的方法建模关键帧与当前帧之间的特征关联。而光流法在使用过程中容易因遮挡、特殊纹理等情况产生错误,从而导致多帧融合存在问题。为了更好地融合多帧特征,本文提取第1帧的外观特征信息与邻近关键帧的位置信息,通过Transformer和改进的PAN(path aggregation network)模块进行特征融合,从而基于多帧时空注意力学习并融合多帧的特征。**方法**　多帧时空注意力引导的半监督视频分割方法由视频预处理(即外观特征提取网络和当前帧特征提取网络)以及基于Transformer和改进的PAN模块的特征融合两部分构成。具体包括以下步骤:构建一个外观信息特征提取网络,用于提取第1帧图像的外观信息;构建一个当前帧特征提取网络,通过Transformer模块对当前帧与第1帧的特征进行融合,使用第1帧的外观信息指导当前帧特征信息的提取;借助邻近数帧掩码图与当前帧特征图进行局部特征匹配,决策出与当前帧位置信息相关性较大的数帧作为邻近关键帧,用来指导当前帧位置信息的提取;借助改进的PAN特征聚合模块,将深层语义信息与浅层语义信息进行融合。**结果**　本文算法在DAVIS(densely annotated video segmentation)-2016数据集上的J和F得分为81.5%和80.9%,在DAVIS-2017数据集上为78.4%和77.9%,均优于对比方法。本文算法的运行速度为22帧/s,对比实验中排名第2,比PLM(pixel-level matching)算法低1.6%。在YouTube-VOS(video object segmentation)数据集上也取得了有竞争力的结果,J和F的平均值达到了71.2%,领先于对比方法。**结论**　多帧时空注意力引导的半监督视频分割算法在对目标物体进行分割的同时,能有效融合全局与局部信息,减少细节信息丢失,在保持较高效率的同时能有效提高半监督视频分割的准确率。

关键词:视频目标分割(VOS);特征提取网络;外观特征信息;时空注意力;特征聚合

Multiframe spatiotemporal attention-guided semisupervised video segmentation

Luo Sihan[1], Yuan Xia[1*], Liang Yongshun[2]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing　210094, China;

2. School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing　210094, China

Abstract: **Objective**　Video object segmentation (VOS) aims to provide high-quality segmentation of target object instances throughout an input video sequence, obtaining pixel-level masks of the target objects, thereby finely segmenting the target from the background images. Compared with tasks such as object tracking and detection, which involve bounding-box level tasks (using rectangular frames to select targets), VOS has pixel-level accuracy, which is more condu-

收稿日期:2023-09-11;**修回日期**:2024-01-03;**预印本日期**:2024-01-10

*通信作者:袁夏　yuanxia@njust.edu.cn

基金项目:国家自然科学基金项目(12071218)

Supported by:National Natural Science Foundation of China (12071218)

cive to locating the target accurately and outlining the details of the target's edge. Depending on the supervision information provided, VOS can be divided into three scenarios: semisupervised VOS, interactive VOS, and unsupervised VOS. In this study, we focus on the semisupervised task. In the scenario of semisupervised VOS, pixel-level annotated masks of the first frame of the video are provided, and subsequent prediction frames can fully utilize the annotated mask of the first frame to assist in computing the segmentation results of each prediction frame. With the development of deep neural network technology, current semisupervised VOS methods are mostly based on deep learning. These methods can be divided into the following three categories: detection-, matching-, and propagation-based methods. Detection-based object segmentation algorithms treat VOS tasks as image object segmentation tasks without considering the temporal association of videos, believing that only a strong frame-level object detector and segmenter are needed to perform target segmentation frame by frame. Matching-based works typically segment video objects by calculating pixel-level matching scores or semantic feature matching scores between the template frame and the current prediction frame. Propagation-based methods propagate the multiframe feature information before the prediction frame to the prediction frame and calculate the correlation between the prediction frame feature and the previous frame feature to represent video context information. This context information locates the key areas of the entire video and can guide single-frame image segmentation. Motion-based propagation methods have two types: one introduces optical flow to train the VOS model, and the other learns deep target features from the previous frame's target mask and refines the target mask in the current frame. Existing semisupervised video segmentation is mostly based on optical flow methods to model the feature association between key frames and the current frame. However, the optical flow method is prone to errors due to occlusions, special textures and other situations, leading to issues in multiframe fusion. Aiming to integrate multiframe features, this study extracts the appearance feature information of the first frame and the positional information of the adjacent key frames and fuses the features through the Transformer and the improved path aggregation network (PAN) module, thereby learning and integrating features based on multiframe spatiotemporal attention. **Method** In this study, we propose a semisupervised VOS method based on the fusion of features using the Transformer mechanism. This method integrates multiframe appearance feature information and positional feature information. Specifically, the algorithm is divided into the following steps: 1) appearance information feature extraction network: first, we construct an appearance information feature extraction network. This module, based on CSPDarknet53, is modified and consists of CBS (convolution, batch normalization, and Silu) modules, cross stage partial residual network (CSPRes) modules, residual spatial pyramid pooling (ResSPP) modules, and receptive field enhancement and pyramid pooling (REP) modules. The first frame of the video serves as the input, which is passed through three CBS modules to obtain the shallow features $\boldsymbol{F}_s$. These features are then processed through six CSPRes modules, followed by a ResSPP module, and finally another CBS module to produce the output $\boldsymbol{F}_d$, representing the appearance information extracted from the first frame of the video. 2) Current frame feature extraction network: we then build a network to extract features from the current frame. This network comprises three cascaded CBS modules, which are used to extract the current frame's feature information. Simultaneously, the Transformer feature fusion module merges the features of the current frame with those of the first frame. The appearance information from the first frame guides the extraction of feature information from the current frame. Within this, the Transformer module consists of an encoder and a decoder. 3) Local feature matching: with the aid of the mask maps from several adjacent frames and the feature map of the current frame, local feature matching is performed. This process determines the frames with positional information that has a strong correlation with the current frame and treats them as nearby keyframes. These keyframes are then used to guide the extraction of positional information from the current frame. 4) Enhanced PAN feature aggregation module: finally, the input feature maps are passed through a spatial pyramid pooling (SPP) module that contains max-pooling layers of sizes 3×3, 5×5, and 9×9. The improved PAN structure powerfully fuses the features across different layers. The feature maps undergo a concatenation operation, which integrates deep semantic information with shallow semantic information. By integrating these steps, the proposed method aims to improve the accuracy and robustness of VOS tasks. **Result** In the experimental section, the proposed method did not require online fine tuning and postprocessing. Our algorithm was compared with the current 10 mainstream methods on the DAVIS-2016 and DAVIS-2017 datasets and with five methods on the YouTube-VOS dataset. On the DAVIS-2016 dataset, the algorithm achieved commendable performance, with a region similarity score J and contour accuracy score F of

81.5% and 80.9%, respectively, which is an improvement of 1.2% over the highest-performing comparison method. On the DAVIS-2017 dataset, J and F scores reached 78.4% and 77.9%, respectively, an improvement of 1.3% over the highest-performing comparison method. The running speed of our algorithm is 22 frame/s, ranking it second, slightly lower than the pixel-level matching (PLM) algorithm by 1.6%. On the YouTube-VOS dataset, competitive results were also achieved, with average J and F scores reaching 71.2%, surpassing all comparison methods. **Conclusion** The semisupervised video segmentation algorithm based on multiframe spatiotemporal attention can effectively integrate global and local information while segmenting target objects. Thus, the loss of detailed information is minimized; while maintaining high efficiency, it can also effectively improve the accuracy of semisupervised video segmentation.

Key words: video object segmentation (VOS); feature extraction network; appearance feature information; spatiotemporal attention; feature aggregation

0 引　言

视频目标分割(video object segmentation, VOS)的目的是从整个视频背景中精确找到并分割感兴趣的物体。根据其监督类型,视频目标分割可以粗略地分为两类:无监督视频目标分割和半监督视频目标分割。本文专注于半监督任务,半监督视频目标分割旨在通过在第1帧中给定的像素级分割掩码,陆续分割后续帧中的目标(李瀚 等,2021)。半监督VOS因其广泛应用而受到了极大关注。半监督VOS可以应用于各种具有挑战性的领域,包括视频理解(Lin等,2019;Wu等,2019;Zolfaghari等,2018)、物体跟踪(刘雨亭 等,2022;Bolme等,2010;Fan等,2019)和视频编辑(汪水源 等,2021;Girgensohn等,2000)。由于目标特定的信息仅在第1帧中给出,并且目标可能经历快速移动和剧烈变形,如何充分利用有限的信息进行准确分割变得极具挑战性。

早期的半监督视频物体分割方法(Caelles等,2017;Voigtlaender和Lebibe,2017;Bao等,2018)通过调整预训练的分割网络来学习评估短语中的目标特定外观。这些在线微调方法可有效提升分割性能,但存在以下两个缺陷:推理实时性较差以及在目标物体形变较大的场景下性能有所下降。以OSVOS(one-shot VOS)(Caelles等,2017)为例,该方法仅根据视频序列第1帧对网络进行微调,难以有效分割形变较大的目标物。许多旨在避免在线微调的工作(Voigtlaender等,2019;Hu等,2018;Yang等,2020;Oh等,2018,2019;Wang等,2019)已经实现了性能和运行时间的平衡。这些工作主要可分为基于像素匹配的工作和基于特征传播的工作。基于像素匹配的方法(Voigtlaender等,2019;Hu等,2018;Yang等,2020)利用标注掩码帧和后续帧之间的像素级匹配分类对像素进行细分。基于像素匹配的方法可有效提取标注帧信息的全局特征,从而提升算法定位和分割目标的性能。然而,该方法仅能提取关键帧之间像素级别的相似性特征,无法提取完整视频的场景信息、物体变化等全局特征。在背景复杂或物体运动外观形状变化较快的场景下,该方法的视频分割性能显著下降。基于特征传播的方法(Oh等,2018,2019;Wang等,2019)是将标注掩码在时序上传播到后续帧中。该方法取得了良好的分割效果,但存在以下两个缺陷:一是模型参数较多,导致占用内存较大,难以部署在内存紧张的小型移动设备上;二是当前帧的分割结果图会依赖前序帧的分割结果,导致分割的误差会累积传播,降低算法的分割性能。另外,融合的方法有很多,但目前大多数方法都是基于复杂的卷积网络,需要大量参数。为了解决上述问题,一些方法与上述两种方法相结合,同时应用第1帧和前一帧的信息。虽然使用第1帧和前一帧可以捕捉目标的帧内特征,但简单的组合处理策略可能无法探索不同帧之间的复杂关系。因此,要获得目标物体的全局相关信息是非常困难的。大多数现有方法都无法以令人满意的精度和速度来解决VOS任务,因此仍然需要更有效的方法来达到VOS任务更好的速度与精度权衡。

实际上,在一段视频序列中,由于目标物体是实际存在的实体,尽管其因为运动、形变和遮挡等因素,外观会与第1帧发生部分变化,但是在颜色、纹理、大小等形态上,还会存在某些共性。因此,可以借助第1帧提取出的关于颜色、纹理等外观特征信息,用于修正后续帧在外观特征提取方面的细节。由于目标物体的运动往往是连续且循序渐进的,尽管目标物体因为连续运动会导致视角发生变化,但

是相邻帧的目标物体的位置特征往往具有相关性，因此可以用邻近帧的位置信息进一步指导当前帧的目标分割。针对上述方法存在的推理速度较低、缺乏全局语义信息指导和误差累积传播等问题，本文提出一种多帧时空注意力引导的半监督视频分割方法，有效提取并融合局部信息与视频全局信息，平衡运行速度与分割效果。该方法由视频预处理（即外观特征提取网络和当前帧特征提取网络）以及基于Transformer和改进的路径聚合网络（path aggregation network，PAN）模块的特征融合两部分构成。

本文主要贡献如下：1）提出一种基于Transformer的特征融合的帧间特征聚合方法，该方法利用邻近帧数帧计算的掩码图，与当前分析帧的特征图进行局部特征匹配，可以对不同帧不同位置的像素进行自适应学习与特征聚合；2）使用局部特征匹配机制以及改进的PAN特征聚合模块对深层语义特征和浅层语义特征进行较好的特征融合，实现对目标物体的精确分割；3）在YouTube-VOS（video object segmentation）数据集和DAVIS（densely annotated video segmentation）数据集进行了充分的实验，验证了该方法的有效性和优越性。

1 相关工作

本文将分别从基于在线微调、基于匹配、基于传播等角度对已有研究工作进行总结。

1.1 基于在线微调的工作

基于在线微调的方法（Caelles等，2017；Khoreva等，2017；Li和Loy，2018；Luiten等，2019；Maninis等，2019；Perazzi等，2017）主要在视频的第1帧上进行微调，以提取原始目标，然后逐帧分割视频。OSVOS（Caelles等，2017）使用预训练的网络，并使用测试视频第1帧对其进行微调。OnAVOS（online adaptation VOS）（Voigtlaender和Leibe，2017）利用在线自适应策略改进了OSVOS。与OSVOS不同，OnAVOS通过预测帧的高置信度结果和明确的背景对网络进一步微调。OSVOS-S（semantic one-shot video object segmentation）（Maninis等，2019）用语义信息改进OSVOS。LucidTracker（lucid data dreaming for object tracking）（Khoreva等，2017）引入了一种用于在线微调的数据增强机制。DyeNet（video object segmentation with joint re-identification and attention-aware mask propagation）（Li和Loy，2018）集成了实例重新标识和时序传播方法，并使用在线微调来提高性能。PReMVOS（proposal-generation，refinement and merging for video object segmentation）（Luiten等，2019）结合了实例分割（He等，2017）、光流（Dosovitskiy等，2015；Ilg等，2017）、细化和reID（re-identification）技术（Xiao等，2017）以及广泛的微调技术，获得了令人满意的性能。总之，在线微调对于VOS任务非常有效。因此，随后的方法（Bao等，2018；Li和Loy，2018；Perazzi等，2017）将在线微调视为提高VOS性能的常规技术。但是，在线微调的方法在实际应用中计算代价非常大。本文设计了一个高效率的网络结构针对VOS问题，该网络在DAVIS-2016、DAVIS-2017和YouTube-VOS数据集上获得了具有竞争力的精度（J score分别为81.5%、78.4%与69.3%），且比在线微调的方法快很多（是OSVOS速度的4.4倍，是LWL（learning what to learn）的7.3倍）。

1.2 基于匹配的工作

基于匹配的工作（Chen等，2018；Hu等，2018；Voigtlaender等，2019；Wang等，2019；Yang等，2020）通过计算模板帧和当前预测帧之间的像素级或语义特征匹配分数来进行视频物体分割。具体地，匹配分数是通过计算当前预测帧中每个像素的像素空间和模板帧中最近的邻居像素得到的，如Chen等人（2018）采用的像素级别度量学习计算匹配分数。这些基于匹配的方法首先通过骨干网络获取参考帧和当前帧的嵌入，然后进行像素级比较以获取当前帧的特定目标信息，最后输入到分割解码器中获得结果。一些研究（Wang等，2019；Girgensohn等，2000）选择第1帧作为模板帧，因为其掩码真值已知且目标物体在后续帧中可见。Hu等人（2018）提出了一个软匹配层来计算当前帧和第1帧的前景和背景相似性。然而，由于缺乏时间信息，当视频物体外观发生重大变化时，性能迅速下降。为了解决这个问题，一些研究（Voigtlaender等，2019；Yang等，2020）考虑视频的连续性，选择当前帧的前一帧作为模板帧。Voigtlaender等人（2019）采用语义像素嵌入的方式进行全局匹配（第1帧和当前帧之间的匹配）和局部匹配（前一帧和当前帧之间的匹配）以获得稳定的预测。Yang等人（2020）提出了协作式前景—背景整合网络，平等处理第1帧和前一帧中的前景和背景信息，以有效处理不匹配和漂移的问题。上述研究

表明，利用匹配进行VOS的思想是有效的，但基于匹配的方法只考虑了两帧之间的相关性，而没有充分利用完整视频序列中的全局背景。在本文工作中，首先从第1帧的全局背景中提取出目标物体的外观信息特征，并以此作为指导，为后面的视频分割提供信息。

1.3　基于传播的工作

基于传播的工作（Oh等，2018，2019；Wang等，2019）利用之前得到的掩码或之前学习的特征嵌入来提升VOS性能。有两种基于运动传播的方法：一种是引入光流来训练VOS模型，另一种是从对象掩码中学习深层对象特征的前一帧并细化当前帧的对象掩码。主要思想都是利用目标物体的时空一致性，使用以往的信息来辅助当前帧的掩码预测。Oh等人（2018）设计了一个连体结构网络，在传播过程中堆叠第1帧、前一帧和当前帧的特征。而Wang等人（2019）使用像素级的匹配技术，将前一帧的掩码作为解码器的输入，以接收更多信息重构预测掩码。同时一些研究通过捕捉相邻帧间的运动信息来表达帧间相关性，并结合单帧图像物体的外观显著性信息来分割目标物体，运动信息通过计算相邻帧间的光流向量来表示。RTNet（reciprocal transformations network）（Ren等，2021）模型计算预测帧和相邻帧从前往后的运动光流和从后往前的运动光流信息，然后将两幅RGB（red，green，blue）图像和两幅光流图像分别提取的特征进行两两融合。MATNet（motion-attentive transition network）（Zhou等，2020）模型使用互注意力机制融合RGB图像提供的外观特征和光流图像提供的运动特征获取显著性特征，取得了较好的结果。这些方法取得了一些进展，但由于视频中的快速运动或目标遮挡，仍然存在着漂移问题。与之前工作通过掩码传播显式地利用空间—时间一致性不同，STMVOS（space-time memory VOS）（Oh等，2019）引入时空记忆网络来保存和读取以前的特征嵌入，以指导当前的掩码预测。实验证明，STMVOS相比之前的方法具有更好的性能。但由于存储的特征和掩码，STMVOS会占用大量的内存资源。另一方面，STMVOS依靠大量模拟数据对模型进行预训练以提高性能，如果没有大量模拟数据的支持，其性能会明显下降。与STMVOS相比，本文引入了基于Transformer的特征融合模块，融合第1帧、前一帧与当前帧的信息特征以指导掩码预测，且无需预先进行模拟收集数据，避免了额外的计算开销。

2　本文方法

2.1　基于Transformer特征融合的半监督视频分割

为了解决VOS问题，本文提出一种基于Transformer的特征融合的半监督视频目标分割方法，相比于传统的循环神经网络（recurrent neural network，RNN）和卷积神经网络（convolutional neural network，CNN），Transformer模块使用自注意力机制来捕捉第1帧图像与当前帧图像之间的依赖关系，有效提取与当前帧有关联的在第1帧上的特征信息，从而提高本文算法的视频分割性能。通过使用Transformer机制融合多帧外观特征信息和位置特征信息，增强模型的准确分割能力。

本文方法的结构如图1所示，主要包括以下4个模块：外观信息特征提取网络模块、当前帧特征提取网络模块、Transformer特征融合模块和特征聚合网络模块。其中，外观信息特征提取网络模块根据第1帧中分割出的物体信息，提取第1帧中物体表面的特征，作为外观特征的指导信息，CBS（convolution、batch normalization，Silu）模块是由卷积、特征层归一化和激活函数组成。CSPRes（cross stage partial residual network）是交叉阶段部分连接残差网络模块，它结合了交叉阶段部分连接网络（cross stage partial，CSP）和残差网络（residual network，ResNet）。ResSPP（residual spatial pyramid pooling）是残差空间金字塔池化模块，它结合了残差连接（residual connection）和空间金字塔池化（spatial pyramid pooling）（2.2节）。当前帧特征提取网络模块中，会根据当前帧的信息，修正当前帧目标物体的特征细节。Transformer特征融合模块通过联合邻近帧数帧的掩码图，与当前帧的特征图进行局部特征匹配，选取对当前帧位置参考较大的帧的计算结果，参与后续特征提取的过程（2.3节）。特征聚合网络模块对上述网络提取的全局与局部信息进行融合，输出当前帧目标掩码图，进一步提高分割准确率（2.4节）。最后，本文描述当前帧是如何被输出的（2.5节）。

2.2　外观信息特征提取网络模块

一般来说，视频中目标物体的外观信息（包括物体的颜色、大小、纹理等）在目标物体运动的过程中，

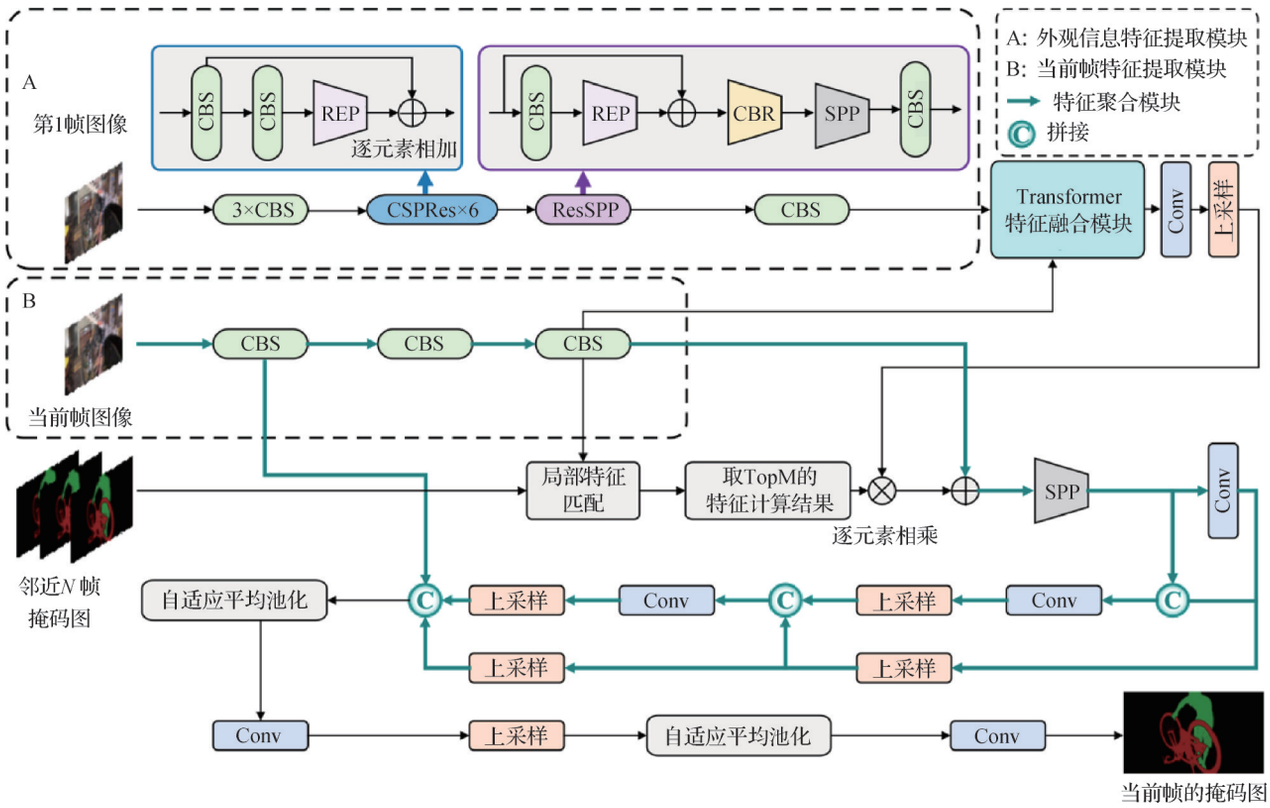


图1　本文方法整体结构图

Fig. 1　The overall structure of the proposed method

不会发生大范围的变化，因此其可帮助后续特征的提取和聚合。本文在半监督视频目标分割的过程中，设计了外观信息特征提取网络模块。此模块选用视频的第1帧进行特征提取，并将提取到的特征信息通过Transformer特征融合模块与当前帧在视频分割过程中提取到的特征进行结合，以更好地获取目标物体的外观信息等特征。

具体来说，为了获得具有更大分辨率带有语义信息的特征图，外观信息特征提取网络模块基于CSPDarknet53（Bochkovskiy 等，2020）进行改进，由CBS模块、CSPRes模块、ResSPP模块和金字塔池化模块（receptive field enhancement and pyramid pooling，REP）组成。本模块以视频第1帧$\boldsymbol{I}_0$作为输入，经过3个CBS（3 × CBS）模块，得到浅层特征$\boldsymbol{F}_s$，具体为

$$\boldsymbol{F}_s = 3 \times f_\theta(\boldsymbol{I}_0) \tag{1}$$

式中，f_θ表示CBS模块。

CBS模块的结构在图2中有所展示，由卷积层、特征图归一化层和Silu（sigmoid linear unit）激活函数层组成。

浅层特征$\boldsymbol{F}_s$经过n个CSPRes（n × CSPRes）模块和1个ResSPP模块后获得更深层的特征$\boldsymbol{F}_d$，具体为

$$\boldsymbol{F}_d = f_R(n \times f_C(\boldsymbol{F}_s)) \tag{2}$$

式中，f_R表示ResSPP模块，f_C表示CSPRes模块。

CSPRes整体模块结构图如图2所示，其中，CSPRes模块经过一个CBS模块，然后将输出的结果分成两个分支，一个分支经过1个CBS和1个REP模块之后与另外一个分支（第1个CBS模块的输出）的特征图对应位置相加；接着经过1个感受野增强和REP模块，该模块的设计目标是提高感受野（receptive field）和金字塔池化（pyramid pooling）的性能，从而改善目标分割的准确性和鲁棒性。REP模块将特征图分别送入一个1 × 1的卷积层和一个3 × 3的卷积层，并将两个特征图分别归一化，再将对应位置的特征图的值相加，经过一个Silu激活函数层输出。REP模块的结构在图2中有所展示。

ResSPP整体模块结构图如图3所示，ResSPP模块包括3部分：1）首先将输入分成两条路径，其中一条路径经过CBS模块和REP模块，与另一条路径上

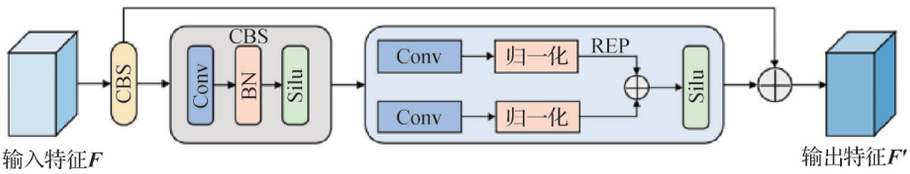


图 2　CSPRes 模块结构图

Fig. 2　CSPRes module structure diagram

的特征图对应位置的值相加；2）经过卷积、归一化、ReLU（rectified linear unit）激活函数，即 CBR（convolution，batch normalization，ReLU）模块；3）将输入分别经过 3 个卷积核（convolutional kernel）大小为 5 × 5、9 × 9、13 × 13 的最大池化层的输出结果进行拼接（concatenation，concat）操作，最终再经过一个 CBS 模块输出 $\boldsymbol{F}_d$，作为整个外观信息特征提取网络模块的输出。

经过精心设计的外观信息特征提取网络模块，本文设计的网络对视频中物体的外观信息完成基本的特征和知识提取，为后面视频帧的分割提供了丰富的提示信息。

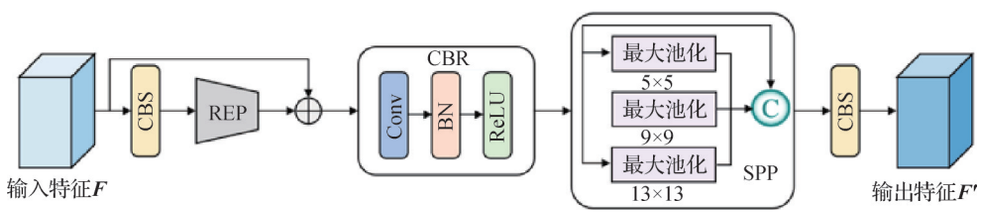


图 3　ResSPP 模块结构图

Fig. 3　ResSPP module structure diagram

2.3　当前帧特征提取网络模块与 Transformer 特征融合模块

当前帧的图像经过 3 个 CBS 模块进行下采样操作，得到当前帧的特征提取结果。因为当前帧与上一帧不会有太大变化，为了提高在当前帧的特征提取效率并且不流失太多细节，本文利用相对简单的 3 个网络模块进行提取。具体来说，当前帧特征提取网络由 3 个级联的 CBS 模块组成。由于充分利用了已有的第 1 帧和前一帧信息，当前帧特征提取网络可以参考其他帧的时序信息，从而得到表征能力更强的语义特征。

为了充分利用外观信息特征提取网络和当前帧特征提取网络的丰富特征，提出了基于 Transformer 的特征融合模块，通过对不同帧不同位置的像素进行特征学习，获取特征之间的关联，从而完成特征融合。同时为了使目标分割的效果更好，本文算法借助外观信息特征提取网络提取的外观信息，使用 Transformer 特征融合模块，对当前帧的外观信息进行微调，同时通过局部特征匹配网络联合邻近帧数帧的掩码图，与当前帧的特征图进行局部特征匹配，选取对当前帧位置参考较大的帧的计算结果，最后将综合得到的信息经过特征聚合网络进行特征融合，使得通过当前帧的全方位的信息生成的掩码图的细节更细化、准确率更高。该 Transformer 特征融合模块由一个编码器和一个解码器组成，如图 4 所示。

2.3.1　编码器

编码器主要用于对邻近帧中有关联的像素信息进行建模，在本模块的工作过程中，编码器将多帧特征进行拼接以获取包含多帧信息的特征，将其输入到注意力模块，使得多帧特征之间进行交错时空融合，特征通过残差的方式与注意力模块输出结果进行相加（addition，Add）操作并进行归一化，然后输入前馈网络，再次通过残差的方式与输出结果进行 Add 操作并进行归一化，然后得到编码之后的特征。本模块不但对前后帧中同一位置的特征进行融合，而且对不同深层语义与浅层语义的特征进行特征融合，可以获得更丰富的视频特征信息，与光流法相比，本文方法提取的特征更丰富、表征能力更强。

其中，交错时空注意力机制接收包含多帧信息的特征 $\boldsymbol{f}_E$，$\boldsymbol{f}_E$ 复制成 3 份相同的特征作为注意力机制的输入，分别为 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$。该注意力机制的具体实现过程如图 5 所示。

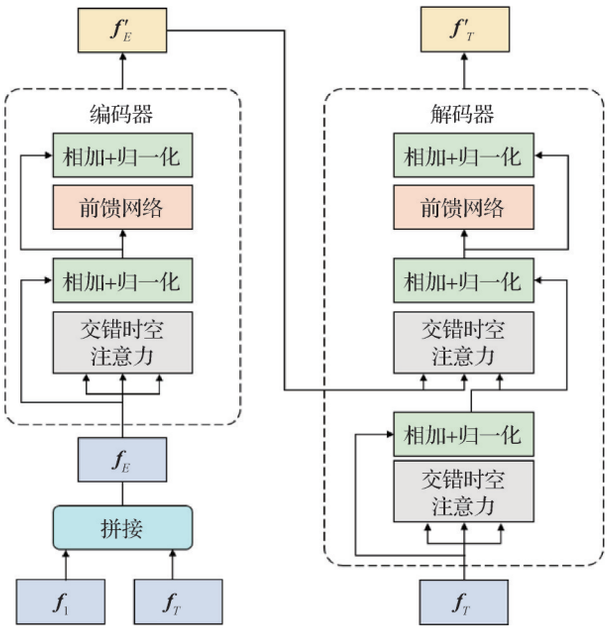


图4　Transformer特征融合模块网络结构

Fig. 4　Transformer feature fusion module network structure

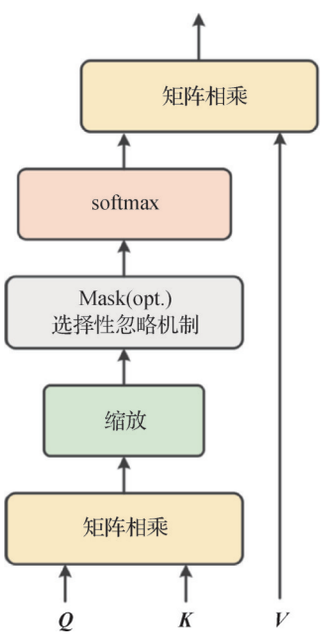


图5　交错时空注意力

Fig. 5　Cross-temporal attention

输入由维度为d_k的查询($\boldsymbol{Q}$)和键($\boldsymbol{K}$)以及维度为d_v的值($\boldsymbol{V}$)组成。计算查询与所有键的点积，每个点积除以$\sqrt{d_k}$，并应用softmax函数来获得值的权重。在实践中，同时计算一组查询的注意力函数，这些查询被打包成矩阵($\boldsymbol{Q}$)。键($\boldsymbol{K}$)和值($\boldsymbol{V}$)也被打包成矩阵$\boldsymbol{K}$和$\boldsymbol{V}$。输出矩阵计算为

$$Attention\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

2.3.2　解码器

解码器以当前帧的原始特征作为输入，输入到注意力模块，将输出结果与当前帧特征进行Add操作并归一化。其输出结果与编码器输出结果同时输入到注意力模块，并将结果同样进行Add操作并归一化，将输出结果输入到前馈网络并进行残差操作。经过解码器操作，本文将当前帧的原始特征与编码器提取的特征进行融合，最终得到的时空特征，包含邻近帧特征丰富的时间与空间信息。

2.4　特征聚合网络模块

随着模型的深度逐渐增加，在获取更多高级语义特征信息的同时，也会逐渐丢失低级语义信息中的空间信息。为了更好地综合高级语义与低级语义信息，本文算法设计了特征聚合网络模块，用于将高级语义信息与低级语义信息进行结合，从而更好地实现目标物体分割的任务。特征聚合网络总体模块如图6所示，其中四边形表示每一步得到的特征图，本文在特征聚合模块引入了空间金字塔池化模块(spatial pyramid pooling，SPP)(He等，2015)和经过本文改进的路径聚合网络(PAN)(Liu等，2018)。

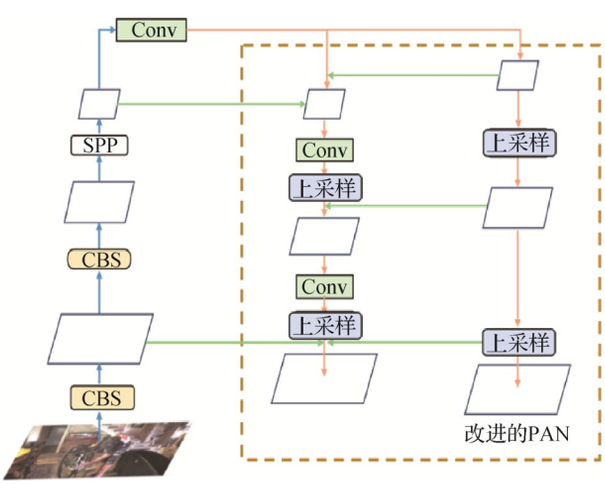


图6　特征聚合网络模块

Fig. 6　Feature aggregation network module

SPP模块包含3个最大池化层，将输入的特征图经过滑窗大小分别为3×3、5×5、9×9的最大池化层，再将不同尺度的特征图进行Concat操作。SPP模块代替了卷积层后的常规池化层，可以增加感受野，更能获取多尺度特征的融合，训练速度也较快。

改进的PAN结构对不同层次的特征进行强有力的融合，其在特征金字塔网络（feature pyramid network，FPN）（Lin等，2017）的基础上增加了自顶向下的特征金字塔结构，保留了更多的浅层位置特征，将整体特征提取能力进一步提升。原始的PAN中相同尺寸的特征图融合是通过Add操作，此处改进是：将Add操作改为Concat操作。二者的区别是Add操作是对相同位置的特征图上的值进行相加，得到的新特征图尺寸不变、数值变化；Concat操作是对具有相同尺寸的特征图进行通道拼接。通过该操作，得到的特征图长宽不变，而通道数则会变为参与拼接的各特征图通道数之和。Concat操作的优点是可以将多个特征图连接成一个更高维度的特征图，保留拼接的特征图的原始值，可以在更高维度上处理数据，得到的信息更全面。最后，将特征图进行Concat操作，对特征图基于通道数拼接，以更充分全面地利用不同层次的特征。

2.5　当前帧掩码图输出

经过上面4个模块处理后，本文最后输出当前帧掩码图。在输入特征聚合模块之前，本文将邻近N帧的掩码图和CBS的输出进行局部特征匹配，并取TopM（Top-M accuracy）的特征计算结果，并进行相乘。通过局部特征匹配，选取有价值的邻近帧的物体位置信息，让当前帧更多关注这些位置上目标的分割，减少了全画面搜索的时间，提高分割效率。局部特征匹配过程如图7所示。

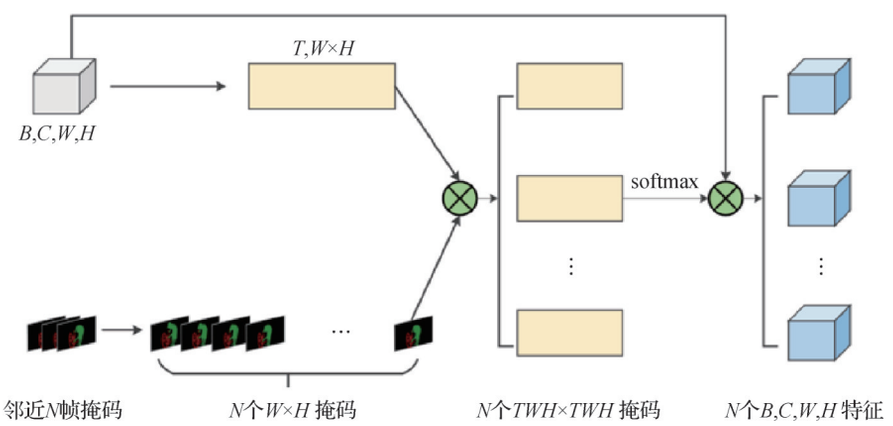


图7　局部特征匹配

Fig. 7　Local feature matching

局部匹配特征具体步骤如下：

1）将当前帧特征提取模块输出的影像的特征维度从（B，C，W，H）转换成（T，WH），此处B为batch-size，C，W，H分别为特征的通道数、宽度、高度，T为B和C的乘积；将n个最近邻的掩码展开，每个掩码维度转换成（T，WH）；

2）将每个变化后的掩码的转置与变化后的影像特征做乘法（WH，T）×（T，WH），最终生成N个匹配后的权重矩阵，维度为（WH，WH），然后将权重矩阵做softmax，进行归一化；

3）将变化后的影像特征与匹配后的权重矩阵做乘法，获得匹配后的特征为N个（T，WH），然后将每个特征维度还原成（B，C，W，H），最终获得特征为N个（B，C，W，H）维度的匹配特征；

4）对每个（B，C，W，H）维度的匹配特征做sum（summation）运算，获取值最大的前M个匹配特征。

此处，N与训练视频的邻近帧变化幅度密切相关：若视频的帧之间变化较快（例如风景转瞬即逝），则N的值取较小；若动作变化较缓，N的值取较大。总之，N基于所有训练视频综合取一个合适的值，具体大约是邻近3 s左右的帧数，以25帧/s为例，N的值大约取75。然后，本文利用特征聚合模块进行特征聚合。具体来说，上一步相乘的结果和CBS的输出相加，并传给SPP模块。SPP模块的输出分两条路径进行。第1条路径先通过一个卷积模块，并进行两个上采样模块；而第2条路径将SPP的输出与卷积模块进行Concat，然后其输出通过一个卷积模块和上采样模块，与第1条路径经过第1个上采样模块的输出进行Concat，再通过一个卷积模块和上采样模块，得到的输出与CBS的输出和第1条路径的输出进行Concat。最后，本文将输出经过5个模块，分别为自适应平均池化、卷积模块、上采样、自适应

平均池化和卷积模块，最终得到输出为当前帧的掩码图。

3 实验结果与分析

3.1 数据集

本文选用DAVIS数据集和YouTube-VOS数据集测试模型的性能。其中，训练集由DAVIS数据集和YouTube-VOS数据集中的训练集构成，测试集由DAVIS测试集和YouTube-VOS测试集分别测试，分别衡量不同模型在两个不同的数据集上的性能表现。

DAVIS数据集有两个版本，DAVIS-2016（Perazzi等，2016）由50个高质量、全高清的视频序列组成，总共3 455标注帧，视频帧率为24帧/s，1 080 p分辨率，含有多个视频目标分割挑战，如摄像机抖动、背景混杂、遮挡以及其他复杂状况。DAVIS数据集密集标注视频分割数据集，区分前后景的二元标注，是面向目标级分割的VOS数据集。DAVIS-2017（Pont-Tuset等，2017）是DAVIS-2016的多目标扩展，共有150个序列，训练集为其中60个视频序列，测试集为剩下的90个视频序列。本文第1个测试结果和第2个测试结果是分别基于DAVIS-2016、DAVIS-2017进行模型的测试。

YouTube-VOS（Xu等，2018）数据集是ECCV（European Conference on Computer Vision）2018视频目标分割比赛所用的数据集，与流行的由120个视频组成的DAVIS基准相比，YouTube-VOS的数据量是其30倍。具体而言，数据集包含训练集中的3 471个视频（65个类别）、验证集中的507个视频（额外的26个不可见类别）和测试集中的541个视频（额外的29个不可见类别）。由于不可见对象类别的存在，YouTube-VOS测试集非常适合于度量不同方法的泛化能力。本文第3个测试结果是基于YouTube-VOS的测试集视频序列进行模型的测试。

3.2 训练细节

本文方法基于Pytorch开源框架实现，损失函数由加权二分类交叉熵损失函数和softmax损失函数组成，具体为

$$J=-\left[y\log\hat{y}+(1-y)\log(1-\hat{y})\right] \tag{4}$$

$$f_{\text{softmax}}(g_i)=\frac{\mathrm{e}^{g_i}}{\sum_{c=1}^{C}\mathrm{e}^{g_c}} \tag{5}$$

式中，y表示真实标签（0或1），$\hat{y}$表示模型预测为正类的概率，g_i为第i个节点的输出值，C为输出节点的个数，即分类的类别个数。

优化器采用Adam优化器，在模型训练的初始阶段，batchsize大小设置为8，学习率大小设置为1×10^{-4}；当训练到15万次初步收敛时，batchsize大小不变，将学习率降低到5×10^{-5}；当模型训练到20万次初步稳定之后，将学习率降低到1×10^{-5}，并最终将模型训练至完全收敛，模型训练阶段结束。

本文在Linux（Ubuntu 18.04）系统上进行实验，机器的CPU为Intel E5-2680，内存为12 GB，GPU为4块Nvidia TITAN Xp。

3.3 评价指标

本文采用Perazzi等人（2016）所建议的3个评价指标：J score（J）、F score（F）和它们的平均值Avg（$J\&F$）。J score表示预测的掩码$\boldsymbol{M}$和真实标注$\boldsymbol{G}$之间相交与联合区域之比，衡量了区域相似度。F score表示真值掩码和预测掩码之间的轮廓精度，是综合准确率和召回率的评价指标，衡量了分割边界的准确程度。

此外，本文采用算法每秒处理的视频帧数（frames per second，FPS）作为额外的评价指标，用来衡量算法的运行速度。

3.4 消融实验

为了获得具有更大分辨率带有语义信息的特征图，本文选用较轻量级的特征提取网络用于外观特征提取。同时，为了挑选更适合本方案的算法网络，对当前主流的特征提取网络在DAVIS-2017数据集上进行对比实验，主要选取了ResNet50（residual network）（He等，2016），VGG16（Visual Geometry Group）（Simonyan和Zisserman，2015），GoogLeNet（LeCun等，1998），Darknet19（Redmon和Farhadi，2017），Efficient-B2（rethinking model scaling for convolutional neural networks）（Tan和Le，2019），Darknet53（Redmon和Farhadi，2018），CSPDarknet53（Bochkovskiy等，2020）和本方案中改进的CSPDarknet53网络，记为CSPDarknet53*。

本文对第1帧的特征提取网络使用上面的网络

结构进行替换做消融实验，表 1 显示了不同特征提取网络的实验结果。本文提出的特征提取网络在现有常用的特征提取网络中表现最好，具有最高的区域相似度和轮廓精度（J = 83.4%，F = 82.2%）。因此，在后面的实验中选用本文的 CSPDarknet53*作为外观特征提取网络。

表 1　不同的特征提取网络的实验结果对比

Table 1　Comparison of experimental results of different feature extraction networks

/%

网络	*Avg*(*J*&*F*)	∇*Avg*(*J*&*F*)	*J*	*F*
ResNet50（He 等，2016）	68.6	−14.2	68.4	68.7
VGG16（Simonyan 和 Zisserman，2015）	67.5	−15.3	67.0	67.9
GoogLeNet（LeCun 等，1998）	75.7	−7.1	75.2	76.1
DarkNet19（Redmon 和 Farhadi，2017）	81.4	−1.4	81.0	81.8
Efficient-B2（Tan 和 Le，2019）	78.9	−3.9	78.8	79.0
Darknet53（Redmon 和 Farhadi，2018）	82.4	−0.4	82.6	82.1
CSPDarknet53（Bochkovskiy 等，2020）	82.1	−0.7	82.3	81.9
CSPDarknet53*（本文）	**82.8**	**0.0**	**83.4**	**82.2**

注：加粗字体表示各列最优结果，∇表示与平均结果的差值。

本文方法在对当前帧进行特征提取时通过 Transformer 特征融合模块，学习第 1 帧与邻近帧的特征，为了验证本文中 Transformer 模块对分割效果的影响，本文接着设计了两个消融实验来验证 Transformer 模块的有效性。分别是：1）去掉 Transformer 特征融合模块；2）将 Transformer 特征融合模块替换为注意力机制模块 CBAM（convolutional block attention module）（Woo 等，2018）。与当前方法进行实验对比结果如表 2 所示。可以发现，Transformer 比空间注意力模块和去掉模块的效果都要好（$Avg(J\&F)$ = 82.8%）。

图 8 展示了本文标准模型与消融实验模型变体的分割结果定性对比。可以看出，将 Transformer 模块去掉之后，分割结果变得尤为粗糙，出现目标物体

表 2　Transformer 融合模块消融实验结果对比

Table 2　Comparison of ablation experiment results of Transformer fusion module

/%

策略	*Avg*(*J*&*F*)	∇*Avg*(*J*&*F*)	*J*	*F*
去掉 Transformer 模块	78.9	−3.9	79.2	78.6
用 CBAM 注意力模块替换 Transformer 模块	80.1	−2.7	80.3	79.9
本文	**82.8**	**0.0**	**83.4**	**82.2**

注：加粗字体表示各列最优结果，∇表示与平均结果的差值。

(a) 无Transformer　(b) 空间注意力　(c) 本文　(d) 真值

图 8　本文模型消融实验分割结果定性对比

Fig. 8　Qualitative comparison of segmentation results of model ablation experiments in this paper ((a) no Transformer; (b) spatial attention; (c) ours; (d) ground truth)

复杂区域漏检的情况。如果将Transformer特征融合模块替换为空间注意力模块，则会出现过分割的情况，使模型的分割结果变差，这充分说明了Transformer模块对视频帧预测有较大影响。

本文方法在特征聚合网络模块中引入了PAN结构和SPP模块，并对PAN结构进行了改进，为了分别验证本文改进的PAN结构和SPP模块对分割效果的影响，再次设计了两个消融实验来验证改进的PAN结构和SPP模块的有效性。针对PAN结构，本文的消融实验是使用改进前原始的PAN结构和本文改进后的PAN结构进行对比，该实验对比结果如表3所示。可以观察到，相比原始的PAN结构，改进的PAN结构可有效提升分割性能（$Avg(J\&F)$ = 82.8%）。图9的两组分割结果定性对比图展示了使用原始的PAN结构存在较多漏分割区域，而改进的PAN结构可有效减少漏检区域，显著提升分割效果，这充分说明改进的PAN结构相对原始PAN结构对视频帧预测效果更好。

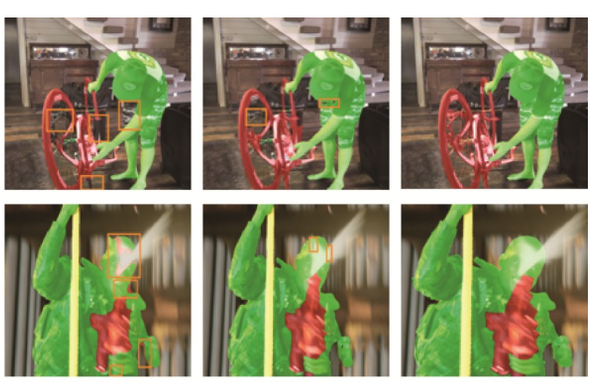
表3　PAN结构消融实验结果对比

Table 3　Comparison of ablation experiment results of PAN structure

/%

策略	$Avg(J\&F)$	$\nabla Avg(J\&F)$	J	F
原始PAN	79.9	−2.9	79.6	80.2
改进PAN	**82.8**	**0.0**	**83.4**	**82.2**

注：加粗字体表示各列最优结果，∇表示与平均结果的差值。



(a) 原始PAN　(b) 改进的PAN(本文)　(c) 真值

图9　PAN结构消融实验分割结果定性对比

Fig. 9　Qualitative comparison of segmentation results of PAN structure ablation experiments ((a) original PAN; (b) improved PAN (ours); (c) ground truth)

针对SPP模块，本文的消融实验是去掉SPP模块和本文使用SPP模块进行对比，该实验对比结果如表4所示。可以观察到，去掉SPP模块后分割指标低于使用SPP模块的结果指标。图10的两组分割结果定性对比图展示了相比于无SPP模块的算法，采用SPP模块算法的分割性能有所提升，有效减少了漏检和混检区域，这说明SPP模块对视频分割效果的提升起到了一定帮助。

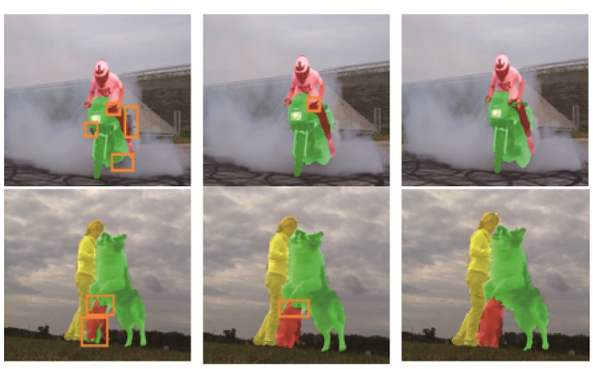
表4　SPP模块消融实验结果对比

Table 4　Comparison of ablation experiment results of SPP module

/%

策略	$Avg(J\&F)$	$\nabla Avg(J\&F)$	J	F
无SPP	80.5	−2.3	80.6	80.3
有SPP(本文)	**82.8**	**0.0**	**83.4**	**82.2**

注：加粗字体表示各列最优结果，∇表示与平均结果的差值。



(a) 无SPP　(b) 有SPP(本文)　(c) 真值

图10　SPP模块消融实验分割结果定性对比

Fig. 10　Qualitative comparison of segmentation results of SPP model ablation experiments ((a) no SPP; (b) with SPP (ours); (c) ground truth)

3.5　主流算法对比

3.5.1　实验结果定量分析

表5—表7分别展示了本文模型与当前主流的视频目标分割方法在DAVIS-2016、DAVIS-2017和YouTube-VOS数据集的对比结果。

表5展示了10种模型使用DAVIS-2016训练集训练后的模型在DAVIS-2016测试集上的各项指标结果与本文模型的对比。其中第2列表示是否需要在线微调。有些方法将在训练集训练得到的模型权重直接进行测试，不对测试集作权重调整，这种方法就是属于不需要在线微调的方法，如RANet(ranking attention network)(Wang等，2019)、SAT(state-aware tracker)(Chen等，2020)、CFBI(collaborative VOS by

表 5　DAVIS-2016 测试集上的定量结果

Table 5　The quantitative results on DAVIS-2016 test set

网络	微调	*J*/%	*F*/%	*Avg*(*J*&*F*)/%	帧速率/(帧/s)
OSVOS(Caelles 等,2017)	√	70.3	72.7	71.5	5
PLM(Shin Yoon 等,2017)	√	74.0	73.2	73.6	**30**
STMVOS(Oh 等,2019)	×	80.2	79.7	80.0	15
RANet(Wang 等,2019)	×	78.1	77.5	77.8	5
SAT(Chen 等,2020)	×	78.9	78.4	78.7	10
LWL(Bhat 等,2020)	√	79.7	79.0	79.4	3
CFBI(Yang 等,2020)	×	71.6	70.4	71.0	10
MATNet(Zhou 等,2020)	×	80.2	79.4	79.8	7
BATMAN(Yu 等,2022)	×	80.7	79.5	80.1	20
RHMNet(Zhou 等,2023)	×	81.2	79.4	80.3	18
本文	×	**81.5**	**80.9**	**81.2**	22

注:加粗字体表示各列最优结果。“√”表示使用在线微调,“×”表示不使用在线微调。

表 6　DAVIS-2017 测试集上的定量结果

Table 6　The quantitative results on DAVIS-2017 test set

网络	微调	*J*/%	*F*/%	*Avg*(*J*&*F*)/%	帧速率/(帧/s)
OSVOS(Caelles 等,2017)	√	72.2	76.4	74.3	5
PLM(Shin Yoon 等,2017)	√	73.4	74.3	73.9	**30**
STMVOS(Oh 等,2019)	×	76.7	76.4	76.6	15
RANet(Wang 等,2019)	×	74.6	73.9	74.3	5
SAT(Chen 等,2020)	×	75.2	74.6	74.9	10
LWL(Bhat 等,2020)	√	76.3	75.7	76.0	3
CFBI(Yang 等,2020)	×	68.0	66.3	67.2	10
MATNet(Zhou 等,2020)	×	77.2	76.6	76.9	7
BATMAN(Yu 等,2022)	×	75.8	77.4	76.6	20
RHMNet(Zhou 等,2023)	×	77.6	77.2	77.4	18
本文	×	**78.4**	**77.9**	**78.2**	22

注:加粗字体表示各列最优结果。“√”表示使用在线微调,“×”表示不使用在线微调。

foreground-background integration)(Yang 等,2020)等。另一些方法(如OSVOS、PLM、LWL)基于训练得到的参数,对测试集上的数据进行参数微调整。一般来说,在线微调能够使模型更好地提取测试集视频特有的特征,但会增加时间成本。本文模型属于不需要微调的方法,在保证分割速度的基础上还能取得优于在线微调类方法的分割效果,因为本文模型在训练阶段提取视频中目标物体抽象特征(如外观信息特征),因此能够适应测试集中物体的各种变化。

从表 5 可以看出,在线微调的方法,如 OSVOS、PLM(pixel-level matching)(Shin Yoon 等,2017)和 LWL(Bhat 等,2020)在 1 s 内快的可以分割 30 幅图像(PLM),慢的可以分割 3～5 幅图像(OSVOS 和 LWL),耗时较长,而本文模型直接加载离线权重来预测分割,平均每秒可以分割 22 幅图像,相比于 LWL 平均每秒分割 3 幅图像,本文模型的速度是 LWL 的 7.3 倍。且在速度提升的同时,区域相似度 *J*

表7 YouTube-VOS测试集上的定量结果

Table 7 The quantitative results on YouTube-VOS test set

网络	$Avg(J\&F)$/%	J_{seen}/%	F_{seen}/%	J_{unseen}/%	F_{unseen}/%	帧速率/(帧/s)
OSVOS(Caelles等,2017)	69.1	66.2	76.2	61.8	72.1	2
PLM(Shin Yoon等,2017)	62.3	67.4	60.1	63.4	58.3	**13**
STMVOS(Oh等,2019)	58.7	61.9	57.4	59.5	55.9	7
CFBI(Yang等,2020)	56.9	57.2	54.7	59.1	56.6	4
RMNet(Xie等,2021)	59.3	62.5	59.3	58.3	57.2	6
RHMNet(Zhou等,2023)	70.5	68.1	**78.3**	66.5	69.1	8
本文	**71.2**	**69.3**	75.9	**66.7**	**72.7**	10

注:加粗字体表示各列最优结果。

和轮廓精度F仍然领先于这种在线微调方法,所以本文模型的分割质量上也是远远优于这些在线微调方法。而STM模型需要大量存储前面帧的特征并向后传播来辅助当前帧掩码预测的方法,而且STM模型取得区域和边缘平均准确性为80.0%的分割效果是采用了大量由图像模拟生成视频进行预训练,然后再使用DAVIS-2017训练集进行训练调优得到的。而本文模型仅使用DAVIS-2017训练集数据进行训练(无预训练)就得到81.2%的分割效果,区域相似度J更是从79.7%提升到80.9%,提升了1.5%。RHMNet(reliability-hierarchical memory network)(Zhou等,2023)模型与STM类似,采用了两阶段训练方法,首先在静态图像上训练,然后再在DAVIS-2017训练集上训练。相对于RANet和CFBI这种基于像素匹配的方法,本文模型的分割效果也有一定提升,这得益于本文模型中除了局部匹配机制还增加了特征聚合模块,能够融合低级语义信息和高级语义信息,更好地提取视频的整体信息来指导分割。

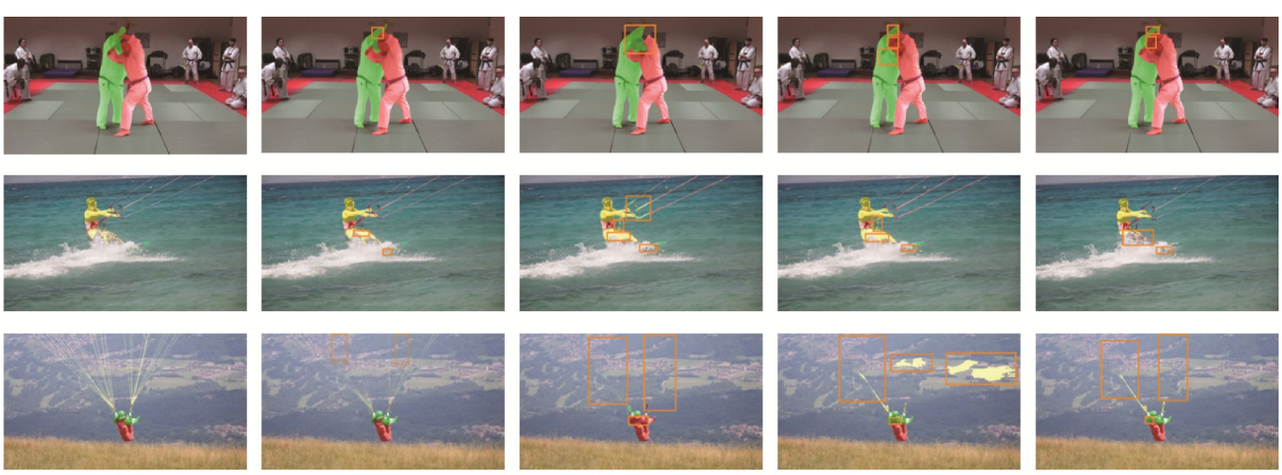
表6展示了上述10种模型在DAVIS-2017测试集上的指标结果,由于DAVIS-2017测试集中包含的是多目标物体的视频,所以模型在DAVIS-2017上的分割效果相对于DAVIS-2016测试集有所下降,但是本文模型仍优于PLM模型5.8%。对于添加了额外数据集进行预训练的STM和RHMNet模型,本文模型仅使用DAVIS-2017进行训练,但是分割效果仍然分别提升了2%和1%。对于借助从运动到外观的信息流的MATNet模型,本文模型以1.69%的略微差距领先,MATNet使用的双流编码器中,使用输入图像和光流场特征作为输入,但是光流法在使用过程中容易因遮挡、特殊纹理等情况产生错误,从而导致多帧融合存在问题。而本文提出的基于Transformer的特征融合模块,用来提取第1帧、邻近帧的特征,可以学习并融合不同帧的特征,避免了光流法容易出现的错误问题。

表7展示了OSVOS、PLM、STMVOS、CFBI,RMNet(regional memory network)(Xie等,2021)、RHMNet这6种模型与本文模型在YouTube-VOS测试集上的结果对比。由于YouTube-VOS测试集中会出现训练集中没有出现过的目标物体,所以区域相似度(J)和轮廓精度(F)都分为目标物体出现过(J_{seen}和F_{seen})和未出现过(J_{unseen}和F_{unseen})两种情况。可以看出,本文模型在YouTube-VOS测试集上的分割效果在各项指标中仍然优于STMVOS模型。对于在训练集中出现过的物体,本文模型在边缘准确率略低于RHMNet模型,其他评价指标均高于RHMNet模型。总的来说,本文方法与近几年较新的方法CFBI、MATNet和RHMNet方法相比,在DAVIS-2016、DAVIS-2017测试集和YouTube-VOS测试集上$Avg(J\&F)$均有提高;同时在推理速度方面,本文在单帧的推理速度也都高于CFBI、MATNet和RHMNet。CFBI使用的是基于像素级匹配的方法,MATNet利用运动信息来增强时空对象表示,RHMNet是以涂鸦监督的形式对目标进行分割,本文方法借助更精确的邻近帧的分割物体的位置信息,并通过局部特征匹配来修正基于轮廓的位置信息,因此轮廓精度的衡量指标F值更高。

3.5.2　实验结果定性分析

图 11 展示了 4 种模型（本文、OSVOS、STMVOS、CFBI）的分割视频单帧图像对比。第 1 行在搏斗视频中，本文、OSVOS、STMVOS 以及 CFBI 方法均受到影响，但是本文方法分割结果误差最小，其他方法在两个目标分割之间耦合区域都出现大范围判断错误情况，而本文模型则能够减少这种情况，两个目标物体之间的分割也更加清晰顺滑；图 11 第 2 行代表目标物体和背景物体同时运动时的复杂情况，在水浪和运动员都是运动状态的情况下，OSVOS、STMVOS 方法均将水浪判断为运动员的一部分，CFBI 方法丢失了完整目标物体，模型分割的结果较为粗糙，只有本文方法较为成功地将水浪和运动员区分开，仅存在一些细微的误差，这是由于本文方法充分利用提取的第 1 帧的外观信息特征，成功区分目标物体与背景物体，并且通过局部特征匹配，邻近帧的物体位置信息被提取，分割算法关注当前帧相应位置的分割，减少了全画面搜索的时间，提高了分割效率；图 11 第 3 行代表目标物体之间运动并且互相耦合的复杂情况，可以看出相对于 OSVOS 模型，本文模型正确预测了人肩部属于跳伞的人，而 OSVOS 模型将其判断为背包的一部分，STMVOS 模型大幅度扩大了目标物体的边缘，存在背景像素误判情况，将背景上的建筑判断为绳索部分，而 CFBI 模型未将背包的肩带部分完整地分割出来，其他 3 个方法均为将绳索部分较为完整地分割出来，大幅度缺失绳索的分割，只有本文模型比较好地将绳索区分出来，这是由于本文模型中使用了特征聚合模块，将低级语义信息与高级语义信息相互融合，加强了全局感知能力，能够更好地理解视频场景的意义。

图 12 展示了本文、SAT 模型和 MATNet 模型对分割整段视频的效果对比，显示了视频每隔 10 帧的分割结果。图 12（a）是汽车移动转弯视频的分割结果，可以看出本文算法对于形变和快速运动的处理能力；图 12（b）是游动的 5 条金鱼的分割结果，可以看出，在相似目标互相遮挡和背景繁杂的情况下，本文依然能取得良好的结果；图 12（d）（e）是对于图 12（c）的同一段视频 SAT 模型和 MATNet 模型的分割结果，可以看出，相对于 SAT 模型和 MATNet 模型，本文模型对于目标物体复杂区域漏检误检情况很少或几乎没有，而 SAT 模型和 MATNet 模型对于物体耦合情况的分割结果变差，尤其到视频后半段，3 个物体重叠部分增多时，SAT 模型和 MATNet 模型都将腿部部位混淆。从定性的实验结果可以看出，本文方法基于 Transformer 的指导学习算法在整体和一些细节处分割的效果明显好于对比评测的其他算法。



(a) 真值　(b) 本文　(c) OSVOS　(d) STMVOS　(e) CFBI

图 11　多种视频目标分割模型分割效果定性对比

Fig. 11　Qualitative comparison of segmentation effects of various video object segmentation models ((a) ground truth; (b) ours; (c) OSVOS; (d) STMVOS; (e) CFBI)

4　结　论

为了充分利用多帧视频数据间的信息相关性，本文提出了一种多帧时空注意力引导的半监督视频分割方法，该方法由视频预处理（即外观特征提取网络和当前帧特征提取网络）以及基于 Transformer 和改进的 PAN 模块的特征融合两部分构成。本文一

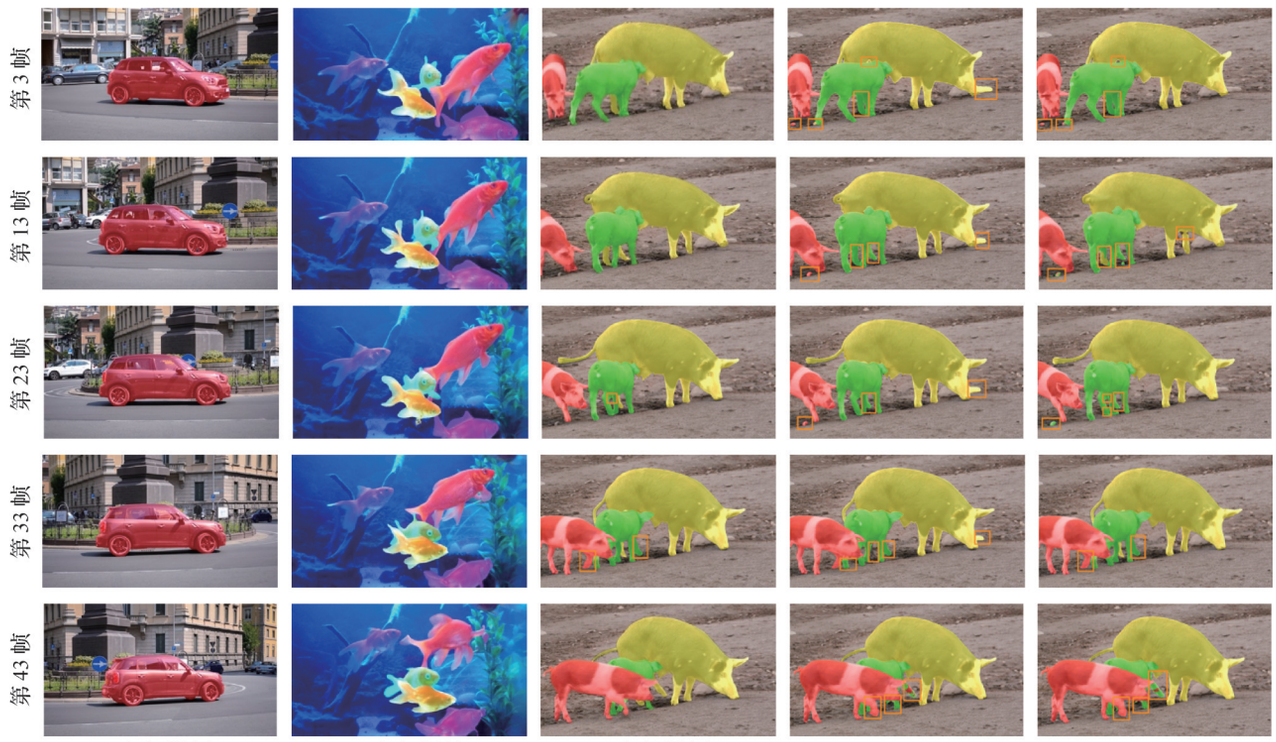


图 12　多种模型在多个视频目标分割数据集上分割完整视频结果对比

Fig. 12　Comparison of the results of multiple models segmenting complete videos on multiple video target segmentation datasets ((a)ours(vedio1);(b)ours(vedio2);(c)ours(vedio3);(d)SAT;(e)MATNet)

方面参考第 1 帧的外观特征信息，使用 Transformer 模块对当前帧的外观信息进行一定程度的修正；另一方面参考邻近数帧中的位置信息，用来指导当前帧的位置信息提取，既参考了视频帧之间的位置关联，又不过多依赖过往帧，从而造成训练资源的浪费。在 DAVIS-2016、DAVIS-2017 和 YouTube-VOS 数据集上的实验证明了所提方法的有效性和优越性。本文方法参考第 1 帧与邻近数帧的信息，指导当前帧更好地提取外观与位置信息，然后使用所设计的特征聚合模块，将深层语义信息与浅层语义信息进行很好地结合，进一步融合目标物体特征，从而实现更快、分割准确率更好的视频分割算法。

本文分割算法在对当前帧进行目标分割时参考了邻近帧的位置特征，所以当分割目标出现被其他物体大面积遮挡且持续一段时间时，效果可能不好。因此，未来工作包括目标物体在遇到大面积遮挡的问题时和目标物体的位置、纹理特征在短时间内剧烈变化的问题时，算法依然能保持较高的精度和效率。此外，针对当前长时间遮挡分割效果欠佳的问题，融合邻近几帧的特征，生成更具参考价值的邻近帧目标位置信息，用来指导当前帧位置信息的提取。

参考文献(References)

Bao L C, Wu B Y and Liu W. 2018. CNN in MRF: video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 5977-5986 [DOI: 10.1109/CVPR.2018.00626]

Bhat G, Lawin F J, Danelljan M, Robinson A, Felsberg M, Van Gool L and Timofte R. 2020. Learning what to learn for video object segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 777-794 [DOI: 10.1007/978-3-030-58536-5_46]

Bochkovskiy A, Wang C Y and Liao H Y M. 2020. Yolov4: optimal speed and accuracy of object detection [EB/OL]. [2023-07-21]. https://arxiv.org/pdf/2004.10934.pdf

Bolme D S, Beveridge J R, Draper B A and Lui Y M. 2010. Visual object tracking using adaptive correlation filters//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE: 2544-2550 [DOI: 10.1109/CVPR.2010.5539960]

Caelles S, Maninis K K, Pont-Tuset J, Leal-Taixé L, Cremers D and Van Gool L. 2017. One-shot video object segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Rec-

ognition (CVPR). Honolulu, USA: IEEE: 5320-5329 [DOI: 10.1109/CVPR.2017.565]

Chen X, Li Z X, Yuan Y, Yu G, Shen J X and Qi D L. 2020. State-aware tracker for real-time video object segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9381-9390 [DOI: 10.1109/CVPR42600.2020.00940]

Chen Y H, Pont-Tuset J, Montes A and Van Gool L. 2018. Blazingly fast video object segmentation with pixel-wise metric learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1189-1198 [DOI: 10.1109/CVPR.2018.00130]

Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V, Smagt P V D, Cremers D and Brox T. 2015. Flownet: learning optical flow with convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 2758-2766 [DOI: 10.1109/ICCV.2015.316]

Fan H, Lin L T, Yang F, Chu P, Deng G, Yu S J, Bai H X, Xu Y, Liao C Y and Ling H B. 2019. LaSOT: a high-quality benchmark for large-scale single object tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5369-5378 [DOI: 10.1109/CVPR.2019.00552]

Girgensohn A, Boreczky J, Chiu P, Doherty J, Foote J, Golovchinsky G, Uchihashi S and Wilcox L. 2000. A semi-automatic approach to home video editing//Proceedings of the 13th Annual ACM Symposium on User Interface software and Technology. San Diego, USA: ACM: 81-89 [DOI: 10.1145/354401.354415]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

He K M, Zhang X Y, Ren S Q and Sun J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1904-1916 [DOI: 10.1109/TPAMI.2015.2389824]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

Hu Y T, Huang J B and Schwing A G. 2018. VideoMatch: matching based video object segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 56-73 [DOI: 10.1007/978-3-030-01237-3_4]

Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A and Brox T. 2017. FlowNet 2.0: Evolution of optical flow estimation with deep networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1647-1655 [DOI: 10.1109/CVPR.2017.179]

Khoreva A, Benenson R, Ilg E, Brox T and Schiele B. 2017. Lucid data dreaming for object tracking. International journal of computer vision, 127(9): 1175-1197 [DOI: 10.1007/s11263-019-01164-6]

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

Li H, Liu K H, Liu J J and Zhang X Y. 2021. Multitask framework for video object tracking and segmentation combined with multi-scale interframe information. Journal of Image and Graphics, 26(1): 101-112 (李瀚, 刘坤华, 刘嘉杰, 张晓晔. 2021. 实时视觉目标跟踪与视频对象分割多任务框架. 中国图象图形学报, 26(1): 101-112) [DOI: 10.11834/jig.200519]

Li X X and Loy C C. 2018. Video object segmentation with joint re-identification and attention-aware mask propagation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 93-110 [DOI: 10.1007/978-3-030-01219-9_6]

Lin J, Gan C and Han S. 2019. TSM: temporal shift module for efficient video understanding//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7082-7092 [DOI: 10.1109/ICCV.2019.00718]

Lin T Y, Dollár P, Girshick R, He K, Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2117-2125 [DOI: 10.1109/CVPR.2017.106]

Liu S, Qi L, Qin H F, Shi J P and Jia J Y. 2018. Path aggregation network for instance segmentation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8759-8768 [DOI: 10.1109/CVPR.2018.00913]

Liu Y T, Zhang K H, Fan J Q and Liu Q S. 2022. Spatiotemporal feature fusion network based multi-objects tracking and segmentation. Journal of Image and Graphics, 27(11): 3257-3266 (刘雨亭, 张开华, 樊佳庆, 刘青山. 2022. 时空特征融合网络的多目标跟踪与分割. 中国图象图形学报, 27(11): 3257-3266) [DOI: 10.11834/jig.210417]

Luiten J, Voigtlaender P and Leibe B. 2019. PReMVOS: proposal-generation, refinement and merging for video object segmentation//Proceedings of the 14th Asian Conference on Computer Vision (ACCV). Perth, Australia: Springer: 565-580 [DOI: 10.1007/978-3-030-20870-7_35]

Maninis K K, Caelles S, Chen Y, Pont-Tuset J, Leal-Taixé L, Cremers D and Van Gool L. 2019. Video object segmentation without temporal information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(6): 1515-1530 [DOI: 10.1109/TPAMI.2018.2838670]

Oh S W, Lee J Y, Sunkavalli K and Kim S J. 2018. Fast video object segmentation by reference-guided mask propagation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Rec-

ognition. Salt Lake City, USA: IEEE: 7376-7385 [DOI: 10.1109/CVPR.2018.00770]

Oh S W, Lee J Y, Xu N and Kim S J. 2019. Video object segmentation using space-time memory networks//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9225-9234 [DOI: 10.1109/ICCV.2019.00932]

Perazzi F, Khoreva A, Benenson R, Schiele B and Sorkine-Hornung A. 2017. Learning video object segmentation from static images//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 3491-3500 [DOI: 10.1109/CVPR.2017.372]

Perazzi F, Pont-Tuset J, McWilliams B, Van Gool L, Gross M and Sorkine-Hornung A. 2016. A benchmark dataset and evaluation. methodology for video object segmentation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 724-732 [DOI: 10.1109/CVPR.2016.85]

Pont-Tuset J, Perazzi F, Caelles S, Arbeláez P, Sorkine-Hornung A and Van Gool L. 2017. The 2017 DAVIS challenge on video object segmentation [EB/OL]. [2023-07-21]. https://arxiv.org/pdf/1704.00675.pdf

Redmon J and Farhadi A. 2017. YOLO9000: better, faster, stronger//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6517-6525 [DOI: 10.1109/CVPR.2017.690]

Redmon J and Farhadi A. 2018. Yolov3: an incremental improvement [EB/OL]. [2023-07-21]. https://arxiv.org/pdf/1804.02767.pdf

Ren S C, Liu W X, Liu Y T, Chen H X, Han G Q and He S F. 2021. Reciprocal transformations for unsupervised video object segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15430-15439 [DOI: 10.1109/CVPR46437.2021.01520]

Shin Yoon J, Rameau F, Kim J, Lee S, Shin S and So Kweon I. 2017. Pixel-level matching for video object segmentation using convolutional neural networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2186-2195 [DOI: 10.1109/ICCV.2017.238]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA: ICLR: 1-14[DOI: 10.48550/arXiv.1409.1556]

Tan M X and Le Q V. 2019. EfficientNet: rethinking model scaling for convolutional neural networks//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 6105-6114

Voigtlaender P, Chai Y N, Schroff F, Adam H, Leibe B and Chen L C. 2019. FEELVOS: fast end-to-end embedding learning for video object segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9473-9482 [DOI: 10.1109/CVPR.2019.00971]

Voigtlaender P and Leibe B. 2017. Online adaptation of convolutional neural networks for video object segmentation//Proceedings of the British Machine Vision Conference (BMVC). London, UK: BMVA Press: 1-16 [DOI: 10.5244/C.31.116]

Wang S Y, Hou Z Q, Wang N, Li F C, Pu L and Ma S G. 2021. Video object segmentation algorithm based on adaptive template updating and multi-feature fusion. Opto-Electronic Engineering, 48(10): #210193 (汪水源, 侯志强, 王囡, 李富成, 蒲磊, 马素刚. 2021. 基于自适应模板更新与多特征融合的视频目标分割算法. 光电工程, 48(10): #210193 [DOI: 10.12086/oee.2021.210193]

Wang Z Q, Xu J, Liu L, Zhu F and Shao L. 2019. RANet: ranking attention network for fast video object segmentation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 3977-3986 [DOI: 10.1109/ICCV.2019.00408]

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: Convolutional block attention module//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 3-19 [DOI: 10.1007/978-3-030-01234-2_1]

Wu C Y, Feichtenhofer C, Fan H Q, He K M, Krahenbuhl P and Girshick R. 2019. Long-term feature banks for detailed video understanding//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 284-293 [DOI: 10.1109/CVPR.2019.00037]

Xiao T, Li S, Wang B C, Lin L and Wang X G. 2017. Joint detection and identification feature learning for person search//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 3376-3385 [DOI: 10.1109/CVPR.2017.360]

Xie H Z, Yao H X, Zhou S C, Zhang S P and Sun W X. 2021. Efficient regional memory network for video object segmentation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1286-1295 [DOI: 10.1109/CVPR46437.2021.00134]

Xu N, Yang L J, Fan Y C, Yue D C, Liang Y C, Yang J C and Huang T. 2018. YouTube-VOS: a large-scale video object segmentation benchmark [EB/OL]. [2023-07-21]. https://arxiv.org/pdf/1809.03327.pdf

Yang Z X, Wei Y C and Yang Y. 2020. Collaborative video object segmentation by foreground-background integration//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 332-348 [DOI: 10.1007/978-3-030-58558-7_20]

Yu Y, Yuan J L, Mittal G, Li F X and Chen M. 2022. BATMAN: bilateral attention Transformer in motion-appearance neighboring space for video object segmentation//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 612-629 [DOI: 10.1007/978-3-031-19818-2_35]

Zhou T F, Wang S Z, Zhou Y, Yao Y Z, Li J W and Shao L. 2020.

Motion-attentive transition for zero-shot video object segmentation// Proceedings of the AAAI Conference on Artificial Intelligence. New York，USA：AAAI：13066-13073［DOI：10.1609/aaai. v34i07. 7008］

Zhou Z K，Mao K G，Pei W J，Wang H P，Wang Y W and He Z Y. 2023. Reliability-hierarchical memory network for scribble-supervised video object segmentation［EB/OL］.［2023-09-07］. https://arxiv.org/pdf/2303.14384.pdf

Zolfaghari M，Singh K and Brox T. 2018. ECO：Efficient convolutional network for online video understanding//Proceedings of the 15th European Conference on Computer Vision. Munich，Germany：Springer：713-730［DOI：10.1007/978-3-030-01216-8_43］

作者简介

罗思涵，女，硕士研究生，主要研究方向为视频目标分割。
E-mail：luosihan@njust.edu.cn

袁夏，通信作者，男，副教授，主要研究方向为智能机器人环境理解与自主导航和视觉显著性分析。
E-mail：yuanxia@njust.edu.cn

梁永顺，男，研究员，主要研究方向为分数阶微积分和图像处理。E-mail：808884903@qq.com