文章编号:1007－6735(2024)04－0407－10　　DOI: 10.13255/j.cnki.jusst.20230308004

基于特征深层融合的吊装过程视频目标分割

周明君，　王朝立，　孙占全
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘要：吊装事故的频繁发生,对国家、社会、人民都造成了非常大的损害。根据吊装过程的视频信息，实现无人安全监控的关键是准确度和速度，提出了一种新的基于全局编码和非对称卷积的目标分割网络，研究视频图像的半监督目标分割问题。首先，将带有标签的视频图像输入网络，分别通过全局编码器与相似性编码器提取到互为补充的特征，从而获得对目标外观的有效表示；然后，通过非对称卷积将两个分支的特征进行深层融合；最后，采用残差上采样解码生成预测掩膜，实现对目标的分割。该方法在 DAVIS2017 数据集上的准确度为 0.675，综合指标为 0.708，帧率为 31 帧/s；在实验用吊装数据集上的准确度为 0.952，综合指标为 0.976，比基线方法高 5.1%，帧率为 26.16 帧/s。与其他网络方法进行了实验比较，验证了分割算法在准确度与速度方面的有效性。

关键词：深度学习；视频目标分割；特征深层融合；目标外观表示；吊装
中图分类号：TP 391　　**文献标志码：**A

Video object segmentation based on deep feature fusion for hoisting process

ZHOU Mingjun,　WANG Chaoli,　SUN Zhanquan
(*School of Optical−Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: The frequent occurrence of hoisting accidents has caused great damage to the country, society and people. According to the video information in hoisting process, the accuracy and speed are the key to realize unmanned safety monitoring system. A new object segmentation network based on global coding and asymmetric convolution was proposed to study semi-supervised video object segmentation task. Firstly, video frames with labels were input into the network, and complementary features were

收稿日期: 2023−03−08
基金项目: 国家自然科学基金资助项目(6217323)；国防科工局基础研究项目 (JCKY2019413D001)
第一作者: 周明君(1997−)，女，硕士研究生．研究方向：人工智能、视频目标分割与跟踪．E-mail：202440443@st.usst.edu.cn
通信作者: 王朝立(1965−)，男，教授．研究方向：非线性控制、机器人控制、人工智能．E-mail：clwang@usst.edu.cn

引文格式: 周明君，王朝立，孙占全．基于特征深层融合的吊装过程视频目标分割[J]．上海理工大学学报，2024，46(4)：407-416．DOI：10.13255/j.cnki.jusst.20230308004
Citation: ZHOU Mingjun, WANG Chaoli, SUN Zhanquan. Video object segmentation based on deep feature fusion for hoisting process[J]. Journal of University of Shanghai for Science and Technology, 2024, 46(4)：407-416.

extracted by global encoder and similarity encoder respectively, so that appearance of the target could be effectively represented. Then features of two branches were deeply fused through asymmetric convolution, and residual up-sampling decoding was used to generate prediction mask for target segmentation. The accuracy and overall indicator on DAVIS2017 dataset were 0.675 and 0.708 respectively, and frame rate was 31 frames per second. On hoisting dataset, the accuracy was 0.952, with the result of 0.976 overall indicator which was 5.1% higher than baseline, and frame rate was 26.16 frames per second. Compared to other methods on DAVIS2017 dataset and the hoisting dataset for experiment, the verification showed that the proposed method was effective in accuracy and speed.

Keywords: *deep learning*; *video object segmentation*; *feature deep fusion*; *object appearance representation*; *hoisting*

随着卷积网络的发展，计算机视觉领域的研究也越来越深入。视频目标分割 (video object segmentation, VOS) 是视频分析和编辑中的基本任务，也是计算机视觉中的一个基本问题和研究热点，在动作抓取[1]、自动驾驶、视频监控[2]、视频编辑等领域有着广泛的应用。因此，视频目标分割受到了极大的关注。在吊装作业过程中往往存在着很多安全隐患，由于人工目测监控视频效率低下，且检测的准确度无法得到保障，通过深度学习算法，实现对吊装作业安全监控的视频目标分割任务。

视频目标分割任务，是在给定视频序列的每一帧中，区分前景和背景像素，预测出目标区域像素的掩膜，对一个或多个目标对象进行跟踪和分割。相比视频目标跟踪[3-4]关注目标的位置尺寸关系，从而进行定位，视频目标分割更关注对目标的精确表示。现有的视频目标分割算法根据其学习方法，大致可以分为无监督学习和半监督学习两类。无监督视频对象分割，是在没有样本标注时，自动分割出每帧图像中最主要的对象。在没有任何先验的情况下，无监督方法很难从一个视频序列中识别出特定的感兴趣目标。相比之下，本文主要考虑的是半监督的视频目标分割，即需要给定第一帧或某几帧目标的真实分割掩膜，对后续帧中所有的目标进行分割。在整个视频序列中，目标随着时间会发生明显的外观变化，以及目标遮挡、快速运动等情况。此外，背景中可能还包含在视觉上或者语义上与目标相似的干扰物。因此，视频目标分割的一个重要问题就是如何对目标外观进行有效表示。

为了达到这个目的，之前大多数的方法通过在线微调独立处理每一帧，这种单帧模型能够得到较高的分割准确度。例如，OSVOS 方法[5]将图像分类上预训练的卷积网络用于视频目标分割任务，只使用第一帧作为参考，独立检测后续每一帧的目标。由于在线微调过程没有集成到网络的离线训练中，不能实现端到端训练。这种方法忽略了帧之间的信息，计算量大，测试时间长，在速度上无法达到实时。另外一类是基于掩膜传播的方法[6]，将前一帧预测的掩膜前馈传播，与当前帧的特征图进行级联，利用帧间信息，指导网络找出当前帧的目标。MSK 方法[7]提出将视频目标分割作为一个掩膜细化问题，通过卷积神经网络对前一帧预测的掩膜进行细化。AGAME 方法[8]引入外观模块，在前向传递中学习目标外观和背景的表示，外观模块的学习和预测阶段都是完全可微分的，使整个分割网络实现端到端训练。这类方法考虑了相邻帧中像素运动的时空联系，能够适应物体外观和位置变化相较平滑的运动，但是容易受到时间间断的影响。在目标遮挡、快速运动时，网络对前一帧的误差很敏感，传播变得不可靠，就会出现漂移问题。

最后一类方法是基于匹配的方法[9]，从给定的初始帧中学习目标的外观，提取初始帧和当前帧的特征，对每帧图像进行像素匹配的计算。这类算法对时间的依赖性不强，处理不匹配和漂移问题时具有较好的鲁棒性。然而，这种方法主要是基于初始帧的目标外观检测，常常不能适应外观的变化，当目标发生旋转等情况表现出不同的视觉表征，可能会导致无法匹配，并且难以区分具有相似外观的对象。RGMP 方法[10]将初始帧和前一帧的预测结果反馈到输入作为参考，使用前一帧预测的结果对网络微调。由于没有对目标外观训练或者外观模型过于简单，对目标的识别能力

不理想。

因此，目标外观变化出现新的视觉表征时，大多数模型对目标的识别能力下降。本文提出了新的全局编码器，利用图像中更丰富的语义信息指导提取目标外观特征，结合相似性编码器，建立有效的目标外观模型，增强分割网络的鲁棒性。随着对视频分割算法的深入研究，目前的方法虽然在准确度和运行速度上得到了很多改进和提升，但是大部分方法仅仅简单地将深层特征拼接，导致分割准确度不高。本文提出了一个深度特征融合模块，对深层特征进行融合，充分利用所获得的目标特征，增强网络对不同目标的判别力，从而提高视频目标分割的准确度。

本文的主要贡献如下：

a. 提出一种全局编码器模块，与相似性编码器互相补充，更好地对目标外观进行表示，提升网络对特定对象的识别能力，增强模型对于不同对象的判别能力；

b. 将非对称卷积引入模型，设计出一种深度特征融合的模块，生成准确的分割掩膜，提高视频目标分割网络的准确度；

c. 本文的算法与目前最先进的方法相比，在多个数据集上有较大的提高，充分证实本文算法具有能够有效区分前景和背景的优越性能。

1　基于特征深层融合的视频目标分割算法

1.1　概　述

本文提出一种基于目标特征提取与深度特征融合，用于吊装安全监控的快速视频目标分割。整体网络模型包含 4 个部分：相似性编码器 (target encoder) 和全局编码器 (global encoder)、深度特征融合模块 (feature fusion module, FFM)、基于残差上采样的解码器 (decoder) 和反馈回路，其主体如图 1 所示。其中，相似性编码器是基于孪生网络，通过当前帧与参考帧的相似性度量，得到相关性特征；全局编码器通过骨干网络提取当前帧的背景特征。两个编码器提取的特征互为补充，使网络更好地构建目标的表观模型。深度特征融合模块通过非对称卷积层，将相似性特征和全局特征进行融合，不是简单的相加或拼接，从而有效提高了准确度，且不会很大地增加网络的参数量。基于残差上采样的解码器，通过跳跃连接有效利用浅层特征，融合预测掩膜更新的特征，将目标特征还原，最终生成准确的分割掩膜。本文原始输入图像大小为 1 920×1 080，分别剪裁出包含目标区域大小为 127×127、303×303、257×257 的图像输入各支路，使网络更加关注目标，减少图像中的背景干扰。

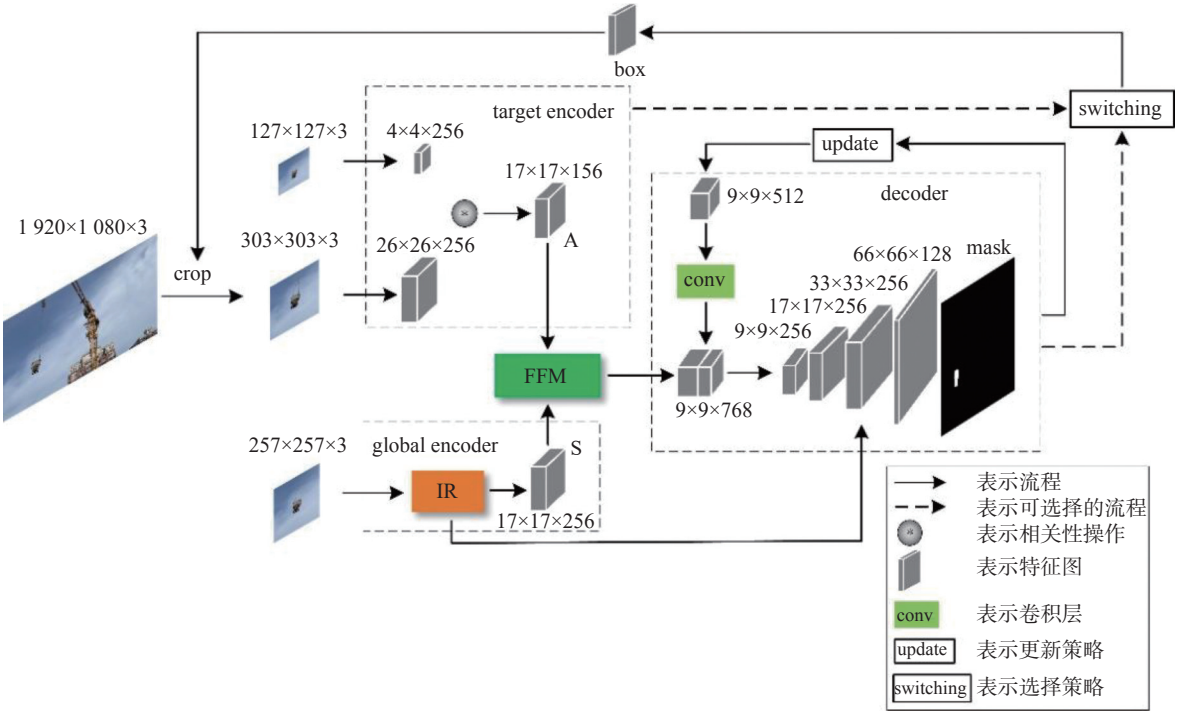

图 1　总体网络框架的示意图

Fig.1　Schematic diagram of the overall video object segmentation network

1.2 相似性编码器和全局编码器

相似性编码器和全局编码器分别是提取目标特征和全局特征的编码器。相似性编码器是基于AlexNet骨干网络的SiamFC++[11]跟踪网络架构，按照SAT[12]网络设置的剪裁策略，根据初始帧的目标区域和当前帧的目标区域，作为网络输入的模板图像和搜索图像。基于孪生网络，利用特征相关性对当前帧的物体与目标外观的相似性进行编码。如图1所示，模板图像大小为127×127，搜索图像大小为303×303，模板图像中给定的目标作为整个视频序列中感兴趣的目标，指导网络学习目标特征。

本文提出的全局编码器是基于特征提取(IR)模块捕获背景特征。如图2所示，模块中的骨干网络为ResNet50的变体，即ResNet50-M，每个卷积层stage1，stage2，stage3，stage4的具体结构如表1所示。本文受Inception-ResNet[13]的启发，根据卷积核大小改变视野范围。为了不丢失图像中所有的全局信息，获取更多背景与目标的特征，本文在特征提取骨干网络的stage1之后增加IR模块，在更浅层捕获更多的特征。IR模块使用了多个1×1、3×3的小卷积，对于同一张输入图像提取不同尺度的特征进行相加。通过1×1卷积，不仅是改变通道数来减少计算量，还是对每个像素点提取更多通道上的特征关系。在每个卷积层中加入ReLU激活函数，从特征图中捕获更多像素点之间复杂的非线性关系。IR模块中也利用了跳跃连接，进一步将浅层特征与深层特征进行相加，不丢失图像的全局特征。

类似地，首先，在目标周围剪裁一个相对较小的区域，减少背景的干扰，输入全局编码器中，通过stage1的卷积层与最大池化得到浅层特征。然后，在IR模块中的卷积层提取不同尺度的特征图相加，使用1×1的卷积将通道数改为256，从而与跳跃连接的浅层特征相加。在此模块中得到的特征后续再经过stage2，stage3，stage4捕获深层特征。全局编码器得到的全局特征图为相似性编码器的相关性特征图做了适当的补充，建立目标外观模型来帮助网络从背景干扰中识别目标，使分割网络更具有鲁棒性。

1.3 深度特征融合模块

大部分方法仅仅简单地将深层特征拼接，导致网络对不同目标的判别力低，分割准确度不高。在基线网络中，仅将深度网络提取的相似性

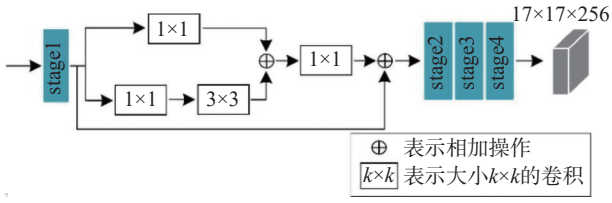

图2 特征提取模块

Fig.2 IR module

表1 ResNet50-M网络结构

Tab.1 Network structure of ResNet50-M

卷积层	输入大小	输出大小	卷积核大小
stage1	257×257×3	131×131×32	3×3, stride 2 {3×3, stride 1}×2 3×3 maxpooling
stage2	131×131×32	66×66×64	3×3, stride 1 {3×3, stride 1}×2 {3×3, stride 1}×2
stage3	66×66×64	33×33×128	3×3, stride 2 {3×3, stride 2}×2 {3×3, stride 1}×3
stage4	33×33×128	17×17×256	3×3, stride 2 {3×3, stride 2}×2 {3×3, stride 1}×5
stage5	17×17×256	9×9×512	3×3, stride 2 {3×3, stride 2}×2 {3×3, stride 1}×2

特征图和另一个分支的特征图，以简单的相加方式进行特征融合。因此，本文提出了深度特征融合模块，结构如图3所示。为了克服卷积运算的局部性，在编码阶段之后，本文引入了非对称卷积，将一个$k\times k$卷积核分解为两个非对称卷积$k\times 1$和$1\times k$的卷积核，利用一维卷积增加了网络深度。图像中的运动目标会发生旋转、形状改变、缩放等外观变化，通过$1\times k$、$k\times 1$非对称卷积的滑动窗口，以不同的方向在特征图上滑动进行卷积操作，捕获每个像素点在各个方向上的特征关系，使网络适应目标的外观变化。因此，将提出的深度特征融合模块，通过$1\times k+k\times 1$和$k\times 1+1\times k$(在本文中k=3)组成非对称卷积层，对深层特征图进行融合。模块中使用残差块(residual block)，如图4所示，是为了不丢失浅层的特征，而1×1的卷积层是为了进一步提取特征通道上的非线性关系。此外，使用非对称卷积也可以减少网络参数量的增加。

以两个编码器分支得到的全局特征图S和相似性特征图A作为输入。相似性特征图被送入卷积层进一步提取深层特征，之后与相同维度大小的全局特征图相加，再由多个3×3卷积块组成的

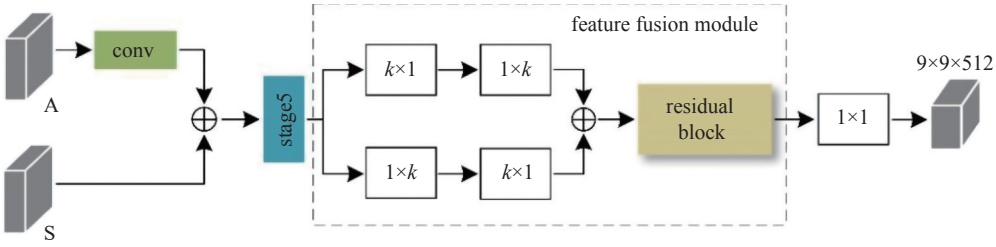

图 3　深度特征融合模块

Fig.3　Feature fusion module (FFM)

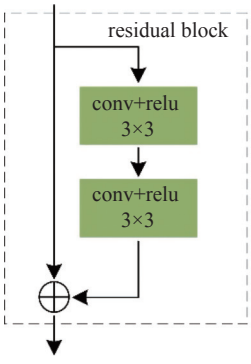

图 4　残差模块

Fig.4　Residual module

stage5 提取特征，经过 1×3、3×1 的非对称卷积层组成的深度特征融合模块，最终生成目标的特征图。

1.4　基于残差上采样的解码器

大部分现有的方法更关注特征提取编码的阶段，而忽略了解码阶段。为保证边缘细节等信息不丢失，在解码过程中，使网络充分利用浅层特征。首先将反馈回路更新的目标特征进行卷积操作，输出的通道数调整为 256，随后再与深度特征融合模块的特征级联。为了充分利用浅层网络的特征信息，根据残差学习的思想，为获得更多的空间语义信息，将全局编码器骨干网络提取的浅层特征以跳跃连接的方式，与深层特征进行双线性插值上采样，最终生成预测掩膜。

按照 SAT 方法的更新策略，反馈回路将预测的二值掩膜与输入图像相乘，并融合之前帧的预测掩膜，进行目标特征的更新 (update)。同时，根据该方法中剪裁策略 (crop) 和选择策略 (switching)，使用相似性编码器中添加的回归头或者分割网络的预测掩膜生成最小包围框 (box)，根据该包围框，确定待搜索目标的位置，裁剪出下一帧相对较小的目标区域的图像作为网络输入。

2　实　验

本文分别在吊装数据集和 DAVIS2017[14] 数据集上做了实验，并在消融实验中分析了本文提出的方法和各模块的作用。本文的方法在两个数据集上的结果都有所提高，在吊装数据集上综合指标为 0.976，帧率为 26.16 帧/s；在 DAVIS2017 数据集上综合指标为 0.708，帧率为 31 帧/s，充分证明了本文算法具有一定的竞争力。

2.1　实验数据集

本实验的吊装数据集是自己手工标注构建的，视频数据来源于手机相机，多个角度拍摄工地吊装的作业过程。从拍摄的所有视频中挑选出 6 个视频，包含 3 类不同的物体。按照 30 帧/s 从视频取出约 170～300 张的连续帧。利用基于 python 环境的图像标注工具 labelme，对每一个视频帧的吊装物体进行像素级的人工标注，生成.json 文件，得到目标的真实标注。自建的数据集包含 6 个视频序列，每段视频时长约为 10 s，总共 1 530 个视频帧，大小为 1 920×1 080。数据集分为训练集和验证集，各 3 段视频，分别有 889 张用于训练，641 张用于验证。每个视频序列包含一个目标，约 170～300 帧，用于半监督的视频目标分割任务。

DAVIS2017 是 DAVIS 在 2017 年发布的用于比赛的视频目标分割数据集，拍摄于现实场景，包括摄像机抖动、背景干扰和其他复杂环境的情况，是常用的数据集之一。DAVIS2017 数据集是像素级标注的多目标视频目标分割数据集，有 150 个视频序列，每个视频序列时长约 2～4 s，总共 10 459 个视频帧，其中包括 60 个训练集 (4 200 张) 和 30 个验证集 (2 023 张)。

2.2　网络训练细节

根据 SiamFC++训练策略，在目标跟踪数据集上训练相似性编码器。在 ImageNet[15] 上对全局编码器的骨干网络 ResNet50-M 进行训练。对于训练样本，采用 COCO[16] 训练集、DAVIS2017 训练集 (60 个视频) 和 Youtube-vos[17] 训练集 (3 471 个视

频)。整个训练过程有20个epoch，需要3 d。对于每个epoch，随机选择150 000张图像进行训练，从视频序列中随机选择一张模板图像和一张搜索图像。训练时batchsize大小设置为32，使用动量为0.9的SGD优化器，对预测掩膜使用交叉熵损失，训练和测试过程均在两块3080 GPU上进行。前两个epoch的学习率从10^{-5}线性增加到10^{-2}，后18个epoch使用余弦退火学习率。

2.3 消融实验

为了验证本文提出的特征提取(IR)模块和深度特征融合模块(FFM)的优化性能，在吊装数据集上做了消融实验，结果如表2所示。基线方法由相似性特征提取的孪生网络分支和显著性网络分支构成骨干网络，然后将得到的特征图直接相加，经过上采样解码，最终预测出目标的分割掩膜。消融实验在基线方法的基础上，分别增加了IR模块和FFM，*J&F*分别提高了0.7%和1.1%。结果表明，提出的两个模块对于网络目标分割的结果是有提升的。Pre表示编码器中提取特征的骨干网络在ImageNet大型数据集上进行预先训练，获得的权重用于训练网络。增加预训练、IR模块和FFM后得到的结果有明显的提升，比基线方法提高了5.1%。实验结果表明，本文提出的基于特征提取模块的全局编码器和深度特征融合模块的网络有利于视频目标分割任务。

表2 消融实验

Tab.2 Ablation experiment

方法	*J&F*	*J*	*F*
基线方法	0.925	0.900	0.949
基线方法+ IR	0.932	0.904	0.960
基线方法+ FFM	0.936	0.912	0.959
基线方法+ IR & FFM	0.943	0.918	0.967
基线方法+ IR & FFM & Pre	0.976	0.952	0.999

3 实验结果与分析

3.1 评价指标

本文采用基准数据集DAVIS2017的标准评价指标，即区域相似度Jaccard(*J*)和轮廓精度F-score(*F*)。区域相似度为预测的二值分割掩膜与标注真值之间的交并比，是比值的形式：分子是预测掩膜与标注真值的前景的交集；分母是两者的并集。区域相似度用于表示分割结果的准确度，其公式表示为

$$J=\frac{Y\cap T}{Y\cup T} \tag{1}$$

式中：Y表示预测值；T表示标注真值。

轮廓精度是描述预测分割结果的边界是否与标注真值的边界对应，其定义式为

$$F=\frac{2PR}{P+R} \tag{2}$$

式中：P表示precision，即精确率；R表示recall，即召回率。

本文采用区域相似度J和轮廓精度F的均值作为综合指标，记作*J&F*。

3.2 不同数据集上的结果

在吊装数据集上的实验结果如表3所示，对比了其他11种算法，包括2种在线学习(OL)的方法和9种离线学习的方法，其中，COSNet[18]是无监督学习的方法。本文算法在综合指标上结果为0.976，帧率为26.16帧/s，在所有对比方法中排名第二，相较领先于其他方法。排名第一的方法STM*[19]是在大型数据集上预训练网络权重，不仅有更多的训练样本，而且训练的时间更长。本文算法进行预训练后的综合指标与STM*方法的综合指标仅相差0.7%，但在速度上有很大的提升。而在没有预训练的情况下，本文算法的综合指标为0.943，STM网络得到的最好结果只有0.664。与SAT方法相比，本文增加了全局编码模块和深度特征融合模块，综合指标提升了5.1%。图5为不同方法在吊装数据集上的指标。可以直观看出，已有的一些方法虽然准确度提高，但是运行速度慢。本文方法实现了速度与准确度的平衡。

表3 不同方法在吊装数据集上的结果比较

Tab.3 Comparison of different methods on hoisting dataset

方法	OL	*J&F*	*J*	*F*	帧率/(帧·s^{-1})
OSVOS[5]	√	0.700	0.574	0.826	0.14
Metavos[20]	√	0.913	0.828	0.998	2.97
COSNet[18]	×	0.614	0.548	0.680	2.05
VM[21]	×	0.427	0.377	0.476	4.04
RVOS[22]	×	0.483	0.410	0.555	12.20
AGNN[23]	×	0.410	0.359	0.461	1.43
AGAME[8]	×	0.867	0.821	0.913	12.58
TVOS[24]	×	0.766	0.668	0.864	3.48
FRTM[25]	×	0.811	0.751	0.870	22.38
STM[19]	×	0.664	0.617	0.711	5.80
STM*[19]	×	0.983	0.967	0.999	5.82
SAT[12]	×	0.925	0.900	0.949	24.26
本文	×	**0.976**	**0.952**	**0.999**	**26.16**

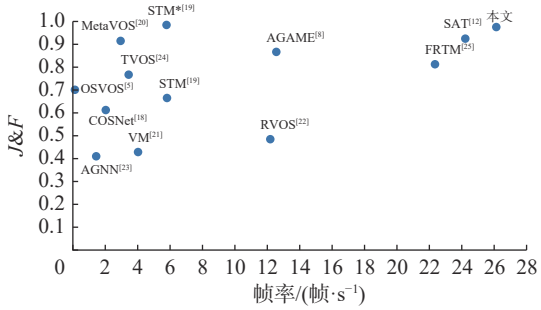

图 5　不同方法在吊装数据集上的综合指标与速度

Fig.5　Accuracy and speed of different methods on hoisting dataset

在 DAVIS2017 数据集上的实验结果如表 4 所示，与其他 21 种算法进行了对比，包括 6 种在线学习分割算法和 15 种离线学习分割算法。对于多目标视频分割任务，预测每个目标的概率图，并拼接在一起，通过 softmax 得到最终结果。相比单目标分割方法，多目标分割的实现更具有难度。本文算法的综合指标达到 0.708，帧率达到 31 帧/s，与其他视频分割算法相比，具有很强的竞争力。

3.3　可视化结果

在公开数据集 DAVIS2017 和实验用吊装数据集上，对实验结果进行可视化，直观地描述本文模型的分割效果。图 6 是在实验用吊装数据集上的分割结果可视化，第 1 列是输入网络的原始图像，第 2 列为人工制作的标签，第 3、4 列分别是通过基线方法得到的结果与本文模型的分割结果。进行比较发现，基线方法对目标分割的轮廓相对比较粗糙，当背景存在干扰物时，不能很好地区分物体与背景，会将一部分背景中的干扰物也分割出来，效果不好。而本文模型很好地将物体分割出来，结果更加精确。图 7 展示了 DAVIS2017 数据集的分割结果可视化。当目标的颜色特征与背景较为相似的时候，基线方法无法区分目标的轮廓，无法精确分割出物体轮廓；在目标快速运动以及背景中存在相似的物体时，如图 7 中跳街舞的人在脚部快速运动时出现的模糊以及背景中相似的行人，基线方法无法精确识别脚步动作与背景行人，而本文方法能够区分背景与目标，将轮廓分割得更明显；当物体出现被遮挡的情况下，本文方法更好地判断出属于前景的部分以及被遮挡的背景部分。

图 8 展示了在整个视频过程中，本文的模型

表 4　不同方法在 DAVIS2017 数据集上的结果比较

Tab.4　Comparison of different methods on DAVIS2017 dataset

方法	OL	*J*&*F*	*J*	*F*	帧率/(帧·s^{-1})
OSVOS[5]	√	0.488	0.462	0.514	—
OSVOS-S[26]	√	0.575	—	—	—
OnAVOS[27]	√	0.679	0.645	0.712	—
DyeNet[28]	√	0.682	0.658	0.705	—
OSMN[29]	√	0.548	0.525	0.571	—
CINM[30]	√	0.675	0.645	0.705	—
VM[21]	×	—	0.565	—	2.86
DyeNet[28]	×	0.625	0.602	0.648	—
RGMP[10]	×	0.667	0.648	0.686	—
FAVOS[31]	×	0.582	0.546	0.618	—
RVOS[22]	×	0.503	0.480	0.526	—
AGNN[23]	×	0.611	0.589	0.632	—
AGAME[8]	×	0.700	0.672	0.727	14.29
TVOS[24]	×	0.631	0.588	0.674	37
FRTM[25]	×	0.702	—	—	21.9
SiamMask[32]	×	0.550	0.543	0.585	35
SAT[12]	×	0.695	0.654	0.736	39
MAROI[33]	×	0.633	0.613	0.653	1.59
GCSeg[34]	×	0.670	0.655	0.684	1.09
MCGFAE[35]	×	0.687	0.665	0.718	—
TAB[36]	×	0.706	0.674	0.738	0.8
本文	×	**0.708**	**0.675**	**0.741**	**31**

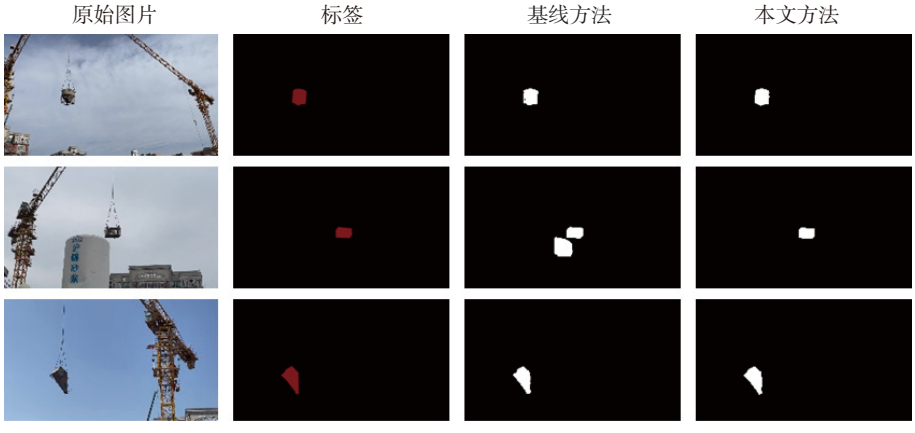

图 6　实验用吊装数据集上可视化结果的比较

Fig.6　Comparison of visual results on hoisting dataset

图 7 Davis2017 数据集上可视化结果的比较

Fig.7 Comparison of visual results on Davis2017 dataset

图 8 本文模型的分割效果可视化

Fig.8 Visualization of segmentation results of the proposed model

在每一帧上对目标的分割效果。前两行是在实验用吊装数据集上，物体发生旋转过程中，模型对目标的分割情况。后 4 行分别展示了在 DAVIS2017 数据集上，本文方法对不同目标的分割情况，以及当物体发生明显的外观变化、环境光线明暗的情况下，模型对目标的识别依然具有鲁棒性，以及很好的分割效果。

4　结　论

本文提出了一种基于目标特征提取与深度特征融合的视频目标分割算法，用于吊装安全监控。本文算法能有效地对目标外观进行表示，提升了网络对特定目标的识别能力，并取得较高的准确度。在分割任务中取得了良好的性能和实时的速度，实现端到端的半监督视频目标分割。

参考文献：

[1] ALLEN P K, TIMCENKO A, YOSHIMI B, et al. Automated tracking and grasping of a moving object with a robotic hand-eye system[J]. IEEE Transactions on Robotics and Automation, 1993, 9(2): 152−165.

[2] 聂志勇, 孙占冬. 基于深度学习的智能视频监控系统应用研究 [J]. 能源科技, 2023, 21(1): 9−13.

[3] ZHANG D W, FU Y W, ZHENG Z L. UAST: uncertainty-aware siamese tracking[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore: ICML, 2022: 26161–26175.

[4] PENG J L, WANG C G, WAN F B, et al. Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 145–161.

[5] CAELLES S, MANINIS K K, PONT-TUSET J, et al. One-shot video object segmentation[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 5320–5329.

[6] 章雪瑞, 孙凤铭, 袁夏. 视频目标分割中帧间相似性传播的研究 [J]. 计算机工程与应用, 2022, 58(6): 227−233.

[7] PERAZZI F, KHOREVA A, BENENSON R, et al. Learning video object segmentation from static images[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3491–3500.

[8] JOHNANDER J, DANELLJAN M, BRISSMAN E, et al. A generative appearance model for end-to-end video object segmentation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 8945–8954.

[9] 付利华, 赵宇, 姜涵煦, 等. 基于前景感知视觉注意的半监督视频目标分割 [J]. 电子学报, 2022, 50(1): 195−206.

[10] OH S W, LEE J Y, SUNKAVALLI K, et al. Fast video object segmentation by reference-guided mask propagation[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7376–7385.

[11] XU Y D, WANG Z Y, LI Z X, et al. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines[C]//Proceedings of the 2020 AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020: 12549–12556.

[12] CHEN X, LI Z X, YUAN Y, et al. State-aware tracker for real-time video object segmentation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9381–9390.

[13] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 4278–4284.

[14] PONT-TUSET J, PERAZZI F, CAELLES S, et al. The 2017 DAVIS challenge on video object segmentation[DB/OL]. [2017-04-03]. https://arxiv.org/abs/1704.00675v1.

[15] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248–255.

[16] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014: 740–755.

[17] XU N, YANG L J, FAN Y C, et al. YouTube-VOS: sequence-to-sequence video object segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 603–619.

[18] LU X K, WANG W G, MA C, et al. See more, know more: unsupervised video object segmentation with co-attention Siamese networks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3618–3627.

[19] OH S W, LEE J Y, XU N, et al. Space-time memory networks for video object segmentation with user guidance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 442−455.

[20] XU C Y, WEI L, CUI Z, et al. Meta-VOS: learning to adapt online target-specific segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 4760−4772.

[21] HU Y T, HUANG J B, SCHWING A G. VideoMatch: matching based video object segmentation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 56–73.

[22] VENTURA C, BELLVER M, GIRBAU A, et al. RVOS: end-to-end recurrent network for video object segmentation[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5272–5281.

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

[24] ZHANG Y Z, WU Z R, PENG H W, et al. A transductive approach for video object segmentation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 6947–6956.

[25] ROBINSON A, JÄREMO LAWIN F, DANELLJAN M, et al. Learning fast and robust target models for video object segmentation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 7404–7413.

[26] MANINIS K K, CAELLES S, CHEN Y, et al. Video object segmentation without temporal information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(6): 1515−1530.

[27] VOIGTLAENDER P, LEIBE B. Online adaptation of convolutional neural networks for video object segmentation[C]//Proceedings of the 2017 British Machine Vision Conference. London: BMVA Press, 2017.

[28] LI X X, CHANGE LOY C. Video object segmentation with joint re-identification and attention-aware mask propagation[C]//Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018: 93–110.

[29] YANG L J, WANG Y R, XIONG X H, et al. Efficient video object segmentation via network modulation[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 6499–6507.

[30] BAO L C, WU B Y, LIU W. CNN in MRF: video object segmentation via inference in a CNN-based higher-order Spatio-temporal MRF[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5977–5986.

[31] CHENG J C, TSAI Y H, HUNG W C, et al. Fast and accurate online video object segmentation via tracking parts[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7415–7424.

[32] WANG Q, ZHANG L, BERTINETTO L, et al. Fast online object tracking and segmentation: a unifying approach[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1328–1338.

[33] FU L H, ZHAO Y, SUN X W, et al. Video object segmentation based on motion-aware ROI prediction and adaptive reference updating[J]. Expert Systems with Applications, 2021, 167: 114153.

[34] LIU W D, LIN G S, ZHANG T Y, et al. Guided co-segmentation network for fast video object segmentation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(4): 1607−1617.

[35] LI X J, ZHENG W M, ZONG Y. Motion cues guided feature aggregation and enhancement for video object segmentation[J]. Neurocomputing, 2022, 493: 176−190

[36] WANG H, LIU W B, XING W W. A temporal attention based appearance model for video object segmentation[J]. Applied Intelligence, 2022, 52(2): 2290−2300.

（编辑：黄 娟）