结合在线归纳和直推推理的快速视频目标分割方法

徐 凯[1)] 李国荣[1)] 洪德祥[1)] 张维刚[2)] 齐元凯[2)] 黄庆明[1)]

[1)](中国科学院大学计算机科学与技术学院 北京 100190)

[2)](哈尔滨工业大学(威海)计算机科学与技术学院 山东 威海 264209)

摘 要 视频目标分割任务是通过算法自动获得视频序列中感兴趣目标对应的像素级区域.因为存在目标表观变化、尺度变化、相似目标干扰、遮挡等困难,所以视频目标分割是一个非常有挑战的任务.现有的方法按照对给定的视频第一帧真实标签的利用方式不同可以分为两类:一类是基于在线归纳学习的方法;另一类是基于直推推理的方法.基于在线归纳学习的方法为了获得准确的结果,在测试阶段,利用给定的初始帧分割图来在线地微调整个网络,导致时间消耗较大,很难满足实时需求.此外,基于直推推理的方法在建模时序推理规则时需要使用大量的合成数据或者标注数据,增加了算法训练的成本.为了充分利用基于在线归纳学习和基于直推推理的两类算法的优点,同时避免两种方法的缺点,本文提出了一个新的结合在线归纳学习和直推推理的快速视频目标分割算法,该网络由直推推理分支和在线归纳分支组成.具体的,直推推理分支可以通过视频前若干帧图像和对应的分割图建模视频短期内的时序变换和运动信息,从而推理出当前帧的分割结果,其学到的时序特征可以指导网络提高视频分割的稳定性.直推推理分支的预训练过程中只需要使用无标注的原始视频数据,不需要使用任何的合成或标注信息.在线归纳分支根据参考帧在线训练,学到目标表观的判别性特征,提供长期的表观判别力.为了提高测试速度,不同于以往的方法,本文没有利用第一帧在线微调整个网络,而是通过在线更新一个非常轻量的模板网络.轻量模板网络提供粗略的分割结果作为注意力图作用到时序特征和当前帧的图像特征,然后经过解码网络生成最终更加精细的分割结果.通过大量的实验,表明本文的方法取得了当前较优的效果,在 DAVIS-2017 和 YouTube-VOS 数据集上分别达到了$\mathcal{J}\&\mathcal{F}$指标的 72.9%和 73.8%,在 DAVIS-2016 数据集上速度达到 18 帧每秒.

关键词 视频目标分割;自监督学习;注意力机制;视频预测;在线学习

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02117

A Fast Video Object Segmentation Method Based on Inductive Learning and Transductive Reasoning

XU Kai[1)] LI Guo-Rong[1)] HONG De-Xiang[1)] ZHANG Wei-Gang[2)] QI Yuan-Kai[2)] HUANG Qing-Ming[1)]

[1)](*School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing* 100190)

[2)](*School of Computer Science and Technology, Harbin Institute of Technology, Weihai, Shandong* 264209)

Abstract With the development of the multimedia and Internet technology, people can easily shoot a large number of video and photos through camera equipment and upload to the Internet now. The computer vision is a very important research field, whose goal is to make the computers to understand the content of video like humans. Video Object Segmentation (VOS) is one of the basic tasks in computer vision. The video object segmentation task aims to automatically obtain the pixel-level mask corresponding to the object of interest in the video sequence by the algorithm.

收稿日期:2021-07-26;在线发布日期:2022-04-13.本课题得到国家重点研发计划(2018YFE0118400)、国家自然科学基金(61772494,61976069,61902092)、中科院青促会项目资助.**徐 凯**,博士研究生,中国计算机学会(CCF)会员,主要研究方向为视频目标分割、视频自监督学习.E-mail:xukai16@mails.ucas.edu.cn.**李国荣**(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究方向为视觉目标跟踪与分割、群体计数、事件分析.E-mail:liguorong@ucas.ac.cn.**洪德祥**,硕士研究生,主要研究方向为视频目标分割、视频自监督学习.**张维刚**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为图像处理、视频分析、模式识别.**齐元凯**,博士,中国计算机学会(CCF)会员,主要研究方向为视觉目标跟踪、视频分割、机器学习.**黄庆明**,博士,教授,中国计算机学会(CCF)会士,国家杰出青年科学基金入选者,IEEE Fellow,主要研究领域为多媒体计算、图像处理、模式识别、机器学习、计算机视觉.

Video object segmentation has various important applications, including automatic driving, video surveillance, video editing, video understanding and so on. Video object segmentation is a highly challenging problem since appearance change, scale change, target occlusion, non-rigid shape change, background interference etc. The existing methods can be divided into two categories according to the different utilization of the real label of the first frame of a given video. One is based on online inductive learning, the other is based on transductive reasoning learning. Previous methods based on online inductive learning finetune the whole network on the given initial frame segmentation mask in the inference stage in order to obtain accurate results, resulting in large time consumption and difficult to meet the real-time requirements. In addition, previous methods based on transductive reasoning need to use a large amount of synthetic data or annotation data when modeling video temporal reasoning rules, which increases the cost of algorithm training. In order to make full use of the advantages of two kinds of algorithms based on online inductive learning and transductive reasoning, and avoid the disadvantages of the two methods, a fast video object segmentation method combining online inductive learning and transductive reasoning is proposed in this paper. Specifically, the transductive reasoning branch is pre-trained by the video prediction self-supervised learning method. It can model the short-term temporal transformation and motion information of the video through the images and segmentation results of the few previous frames, so as to infer the segmentation results of the current frame. The learned temporal features can guide the network to improve the stability of video segmentation. In the pre-training process of the transductive reasoning branch, only the original video data need to be used without any additional synthetic data or manual annotation. The online inductive learning branch is trained online according to the reference frame to learn the appearance discriminant feature of the target and provide long-term appearance discriminant power. In order to improve the test speed, different from the previous methods, the inductive learning branch proposed in this paper does not use the first frame to fine tune the whole network online, but updates a very lightweight template network online to make the template network produce a rough object segmentation mask. This rough segmentation mask is used as the attention map of attention mechanism, and the final segmentation result is generated through the decoding module together with the temporal features and the image features of the current frame. Several experiments are carried out on three challenging datasets (i. e., DAVIS-2016, DAVIS-2017 and YouTube-VOS) to show that our method achieves competitive performance against the state-of-the-arts.

Keywords video object segmentation; self-supervised learning; attention mechanism; video prediction; online learning

1 引 言

视频分割任务的目标是通过算法自动获得视频序列中目标的像素级区域. 视频目标分割有着广泛的应用,包括自动驾驶[1-2]、视频监控[3]、视频编辑、视频压缩[4]等. 视频目标分割任务根据测试阶段指定目标的方式不同可以分为两类,即单样本视频目标分割任务和零样本视频目标分割任务. 单样本视频目标分割是指在测试阶段给定第一帧中感兴趣目标的真实分割图,算法在后续视频帧中分割该目标对应的区域. 而零样本视频目标分割任务在测试时不给定任何信息,算法需要自动识别视频中的前景目标进行分割. 本文中,我们主要关注单样本视频目标分割任务. 由于存在目标表观动态变化、尺度变化、相似目标干扰、遮挡等情况,单样本视频目标分割任务存在很大的困难.

随着近些年深度学习技术的发展,视频目标分割任务有了很大的提升[5]. 但是由于单样本视频目标分割任务是目标类别开放的,即测试时需要分割

的目标不限于训练集中出现过的类别，可以是任意的类别. 如何利用第一帧中给定的感兴趣目标的分割图，得到后续视频中该物体准确的分割结果，是基于深度学习的视频目标分割任务的难点之一. 针对这个问题，现有的方法按照对给定的视频第一帧真实标签的利用方式不同可以分为两类. 一类是基于在线归纳学习的方法[6-10]，另一类是基于直推推理的方法[11-19]. 基于在线归纳学习的方法是指先离线训练一个通用的前景分类器，测试时根据给定的第一帧中的目标标签来在线训练网络，使得网络具有对指定目标的判别能力. 直推推理的方法是通过离线训练过程，学习由参考帧及其目标分割图推理当前帧中相同目标的分割结果的推理规则. 测试时不需要在线训练网络，只需要利用学到的推理规则，根据参考帧和对应的分割结果直接推理当前帧的分割结果.

基于在线归纳学习和基于直推推理的两类算法各有优缺点. 基于在线归纳学习方法的优点是，由于利用参考帧的样本直接微调网络，能够学到对于特定目标比较强的判别特征，对于后续目标变化不大的情况能够获得更加准确的分割结果，对于目标遮挡等情况能够恢复对目标的分割. 缺点是对于时序变化较大的目标，从参考帧中学到的判别特征会失去判别能力. 此外，在测试阶段，利用给定的初始帧分割图在线微调整个网络，导致时间消耗较大，很难满足实时的需求.

基于直推推理方法的优点是，由于使用之前帧的分割结果作为指导信息推理当前帧中的分割结果，所以对于视频中连续变化的目标有很好的分割能力. 缺点是对于目标遮挡等突然变化的目标会失去分割能力并且在后续视频中无法恢复对目标的分割. 此外，在建模视频时序推理规则时需要使用大量的合成数据或者标注数据进行训练，增加了算法训练的成本.

为了充分利用基于在线归纳学习和基于直推推理的两类算法的优点，同时避免两种方法的缺点，本文提出了一个新的结合在线归纳学习和直推推理的快速视频目标分割算法. 该网络由直推推理分支和在线归纳分支组成. 其中，直推推理分支可以通过视频前若干帧图像和对应的分割图建模视频短期内的时序变换和运动信息，从而推理出当前帧的分割结果，其学到的时序特征可以指导网络提高视频分割的稳定性. 为了解决以往基于直推推理的方法需要使用大量的合成数据或者标注数据进行训练的问题，本文采用自监督学习的方法对直推推理分支进行预训练. 直推推理分支只需要使用很容易获取的大量原始视频数据预训练，然后利用少量标注的视频分割数据集对该分支网络进行微调.

在线归纳分支是一个全卷积网络，根据参考帧在线训练，学到目标表观的判别性特征，提供长期的表观判别力. 在线归纳分支由三个子模块组成，即特征提取网络、轻量化模板网络和解码网络. 为了解决以往基于在线归纳学习方法需要在线地训练整个网络，导致时间消耗较大，难以满足实时需求的问题，受 Robinson 等人所提方法[20]的启发，本算法在测试阶段不再在线训练整个在线归纳分支，而是在线更新一个轻量模板网络. 轻量模板网络提供粗略的分割结果作为注意力图作用到时序特征和当前帧的图像特征，然后经过解码网络生成最终更加精细的分割结果. 由于轻量模板网络只由两层卷积组成，更新轻量模板网络的时间消耗较小，所以本方法的运行速度每秒达 18 帧，接近实时的运行速度. 在 DAVIS-2017 和 YouTube-VOS 数据集上的实验结果证明本文的方法取得了领先的效果.

本文的贡献可以总结为以下四点：

(1) 提出了一个新的结合直推推理和在线归纳的视频目标分割算法，这个方法结合了直推推理和在线归纳两类方法的优点而避免了两类算法的缺点.

(2) 直推推理分支由自监督的方式预训练，预训练过程只需使用原始视频数据而不需要人工标注. 直推推理分支可以获得视频的表观变化和运动信息，提高分割结果在视频中的稳定性.

(3) 针对在线归纳分支，设计了新的解码网络能够融合直推推理分支提供的短期时序信息. 引入只更新轻量模板网络而不用微调整个网络的策略，提高了视频目标分割速度.

(4) 本方法在 DAVIS-2017 和 YouTube-VOS 数据集上取得了领先的效果，并且算法的运行速度达到 18 帧每秒. 通过广泛的实验分析验证了本文提出的各个模块的有效性.

2　相关工作

由于单样本视频目标分割任务是类别开放的，所以在训练数据集上离线训练后的网络，在测试时，需要利用视频中给定的第一帧的目标信息才能具有对该特定目标的分割能力. 根据如何利用给定的唯一标注样本使网络具有分割指定目标的能力，基于

深度卷积网络的方法主要分为两类，一类是基于在线归纳学习的方法，另一类是基于直推推理的方法.基于直推推理的方法又可以分为基于分割图传播的方法和基于特征匹配的方法.下面将分别介绍这几类方法的相关工作.

2.1 基于在线归纳学习的方法

在使用深度卷积网络技术解决视频目标分割任务的初期，主要是采用基于在线归纳学习的方法.这类方法首先使用现有的视频目标分割任务数据集的训练集离线训练一个通用的前景分割网络.由于通用分割网络并不具有分割特定目标的能力，所以在测试时，需要使用给定的第一帧样本在线训练这个网络，使得网络具有分割指定目标的能力，然后在后续帧中分割该物体.

OSVOS[6]是这个方向的开拓性工作，它先在图像分割数据集上离线训练一个分割网络，然后在测试时，使用第一帧的图像微调整个网络.OSVOS-S[7]引入实例分割网络的提供的语义信息进一步提高性能.OnAVOS[8]通过对随后的视频帧进行额外的微调使得这一理念得到扩展.Lucid[21]研究在线微调的第一帧图像的数据增广方法.这些方法在对后续视频帧的分割当中只利用单帧信息而忽略了时序信息.为了利用时序信息，提高视频结果的稳定性，一些方法[22-24]进一步整合光流作为一个额外的时序线索.基于在线归纳学习的方法取得了令人满意的分割性能，但是由于需要在线训练网络，耗时较多，无法满足实际的应用需求.FRTM[20]提出只在线更新一个轻量模版网络，从而可以提高运算速度.

2.2 基于直推推理的方法

基于直推式推理的方法通过离线训练，隐式地学习由参考帧和对应真实的标注推理当前帧的分割结果的推理规则，测试时不需要在线训练网络，只需要根据参考帧和对应的标注信息直接推理当前帧分割结果.由于基于直推式推理的方法不需要在线训练网络，所以该类算法的运行速度一般比基于在线归纳学习的方法快.直推式推理的方法又可以分为两类，基于分割图传播的方法和基于特征匹配的方法.

2.2.1 基于分割图传播的方法

基于分割图传播方法[9,11-14]的思想是将前一帧输出的分割图作为分割当前帧图像的指导信息.由于视频的平滑连续性，相邻两帧之间变化不大，所以前一帧中目标的位置和表观与当前帧中的相同目标相近.前一帧中目标的分割图与当前帧要分割的目标有较大的重叠.因此网络经过离线训练可以学到相应的推理规则，其本质是根据前一帧分割图所指示目标的位置和表观，分割当前帧中与之最为接近的目标.

Perazzi 等人[9]将前一帧输出的分割图和当前帧的图像串联在一起作为输入，离线学习的模型通过对前一帧分割输出的细化来预测目标分割图，但是该方法为了学习特定目标的特征仍然需要在线微调网络.一些方法进一步取消了第一帧微调，只利用离线训练.RGMP[12]为了得到特定目标的特征，采用双流网络结构，其中一个分支和方法[9]一样输入将当前帧图像与前一帧的分割图，另一个分支输入第一帧的图像和分割图.第一帧的图像和分割图作为特定目标的信息，使得网络可以不通过在线微调就能感知需要分割目标特定信息.Yang 等人[14]将第一帧的目标的表观信息作为视觉信息以及把前一帧目标位置作为空间信息作为输入，得到网络调制参数，从而显式地调制分割网络，调制后的网络具有对当前帧图像更好的分割能力.A-GAME[11]为了提高网络对相似干扰目标的鲁棒性，引入了表观生成模块提供更有判别性的特征.RANet[13]结合分割图传播和方法特征匹配方法，进一步提高分割准确性.SAT[25]方法在之前帧分割结果作为指导的基础上，结合目标跟踪算法输出的目标框缩小分割的范围，同时根据当前帧分割状态调整目标框的生成策略.

2.2.2 基于特征匹配的方法

另一类基于直推式推理的方法是基于特征匹配的方法，这类方法同样不需要在第一帧图像上进行网络微调.这些方法[13,15-19]的思想是通过离线训练特征提取网络，把原始视频帧映射到特征空间.测试时第一帧目标的特征作为模版，通过特征检索匹配的方法对当前帧的特征或者像素进行分类，从而实现目标分割.Chen 等人[15]对于编码后的特征采用最近邻分类的方法对当前帧的像素进行前背景分类，即当前帧的像素点对应的特征与第一帧的所有特征点求距离，与之最近的点为前景点则分类为前景，反之分类为背景.Hu 等人[16]则把最近邻分类方法改进为当前帧的特征点与第一帧的特征点求相似度矩阵，然后对相似度矩阵经过 Softmax 函数得到分割图.FEELVOS[18]中当前帧的特征不只与第一帧的特征进行全局匹配，而且与前一帧的特征进行局部匹配，匹配图与特征串联经过卷积网络输

出最终的分割结果. TVOS[26] 进一步扩展匹配的帧数，不只在第一帧与前一帧匹配，和多个先前帧的特征进行匹配. STM[17] 同样与多个先前帧的特征进行匹配，但是匹配的结果不直接用来传播分割结果，而是将匹配后的特征经过解码网络输出分割结果. STM[17] 取得了当前领先的算法性能，但是它的训练需要额外的合成数据并且由于特征匹配的算法复杂度较高，即使没有在线微调的过程，也难以达到实时的运行速度. GC[27] 方法对 STM[17] 进行加速，不再将多个历史帧特征进行保存和分别匹配，而是将多帧的特征整合为一个全局特征，测试时只需与全局特征进行匹配.

这些基于直推式推理的方法只利用离线训练，不进行在线更新，往往需要使用大量合成数据集进行离线训练. 与这些方法不同的是，本文的直推推理分支只通过原始视频数据进行自监督预训练，然后利用少量标注的视频分割数据集对该分支网络进行微调.

3　本文方法

在本节中，首先介绍一下视频目标分割任务的符号化定义和整体网络结构，然后分别详细阐述所设计的两个网络分支，即直推推理分支和在线归纳分支，最后介绍网络的训练细节.

3.1　整体网络结构

给定一个视频序列 $\mathcal{V}=\{I_0, I_1, \cdots, I_t, \cdots\}$ 和第一帧目标的真实分割图 S_0，视频目标分割任务是需要得到后续视频帧的分割结果，即 $\mathcal{S}=\{\hat{S}_1, \cdots, \hat{S}_t, \cdots\}$，其中 $\hat{S}_t$ 是对应第 t 帧图像 I_t 的分割结果，$\hat{S}_t$ 和 S_t 分别表示算法输出的分割图和真实的分割图. 如图 1 所示，本文的分割网络由两个分支组成，即直推推理分支和在线归纳分支. 直推推理分支输入视频的前 δ 帧视频图像和对应的分割图 $\{[I_{t-\delta}, \hat{S}_{t-\delta}], \cdots, [I_{t-1}, \hat{S}_{t-1}]\}$，其中 $[,]$ 表示矩阵串联操作. 直推推理分支从前 δ 帧视频图像和对应的分割图中学习具有空间时序信息的判别特征，从而能够建模目标的表观的动态变化和运动信息，指导网络提高视频分割的稳定性. 在线归纳分支是一个全卷积网络，输入当前帧的图像 I_t 以及直推推理分支得到的判别特征，得到最终的分割结果. 在线归纳分支由三个子模块组成，即特征提取网络、轻量化模板网络和解码网络. 接下来，将详细地介绍这两个分支.

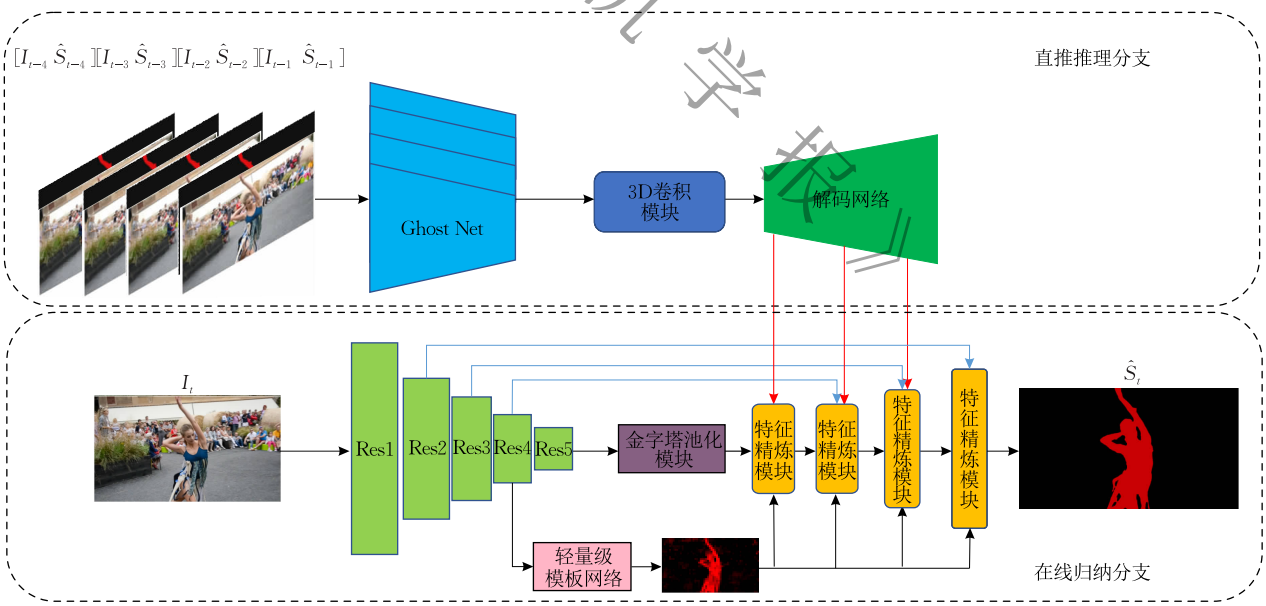


图 1　测试阶段整体网络结构示意图

3.2　直推推理分支

3.2.1　直推推理分支的网络结构

如图 2 所示，本文的直推推理分支由五个部分组成，即特征提取网络、3D 时序特征融合网络、解码网络、分割头网络和 RGB 头网络组成. 其中，分割头网络和 RGB 头网络只在训练过程使用，测试时不使用.

为了提高本文方法的运行速度，特征提取网络采用轻量级的 GhostNet[28] 网络. 直推推理分支的输入是视 RGB 图像和分割图串联组成的矩阵，RGB 图像有 3 个通道而分割图为 1 个通道，所以输入矩阵总共 4 个通道. 所以需要把 GhostNet[28] 网络第一个卷积层的输入通道数由 3 改为 4. 为了提高输出特征

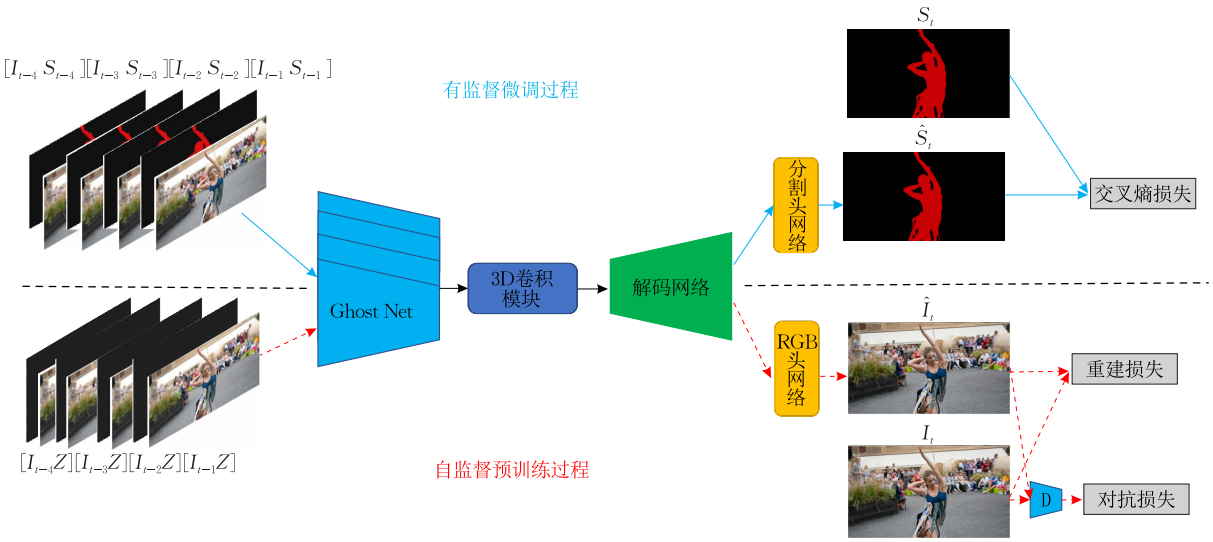


图 2 直推推理分支训练过程示意图

图的分辨率，把 GhostNet[28] 网络第二个模块的下采样步长由 2 改为 1，所以特征提取网络的总步长由 32 变为 16. 以测试过程为例，前 δ 帧视频图像和对应的分割图串联$\{[I_{t-\delta},\hat{S}_{t-\delta}],\cdots,[I_{t-1},\hat{S}_{t-1}]\}\in R^{4\times H\times W}$，分别输入到 δ 个共享参数的特征提取网络，分别得到对应的特征$\{z_{t-\delta},\cdots,z_{t-1}\}\in R^{c\times h\times w}$. 其中，$H$、$W$ 表示图像的高和宽，c、h、w 表示特征图的通道数、特征图的高和特征图的宽，其中 $h=H/16$，$w=W/16$.

3D 时序特征融合网络由两层 3D 卷积网络组成，能够融合多帧的时序特征. 将 δ 帧特征串联成一个时序特征 $z_f\in R^{c\times\delta\times h\times w}$ 输入时序特征融合网络，得到融合多帧特征后的时序特征 $f\in R^{c\times h\times w}$，然后输入到解码网络.

解码网络由 3 个上采样模块组成，每个上采样模块由 3×3 卷积层、批归一化层、ReLU 激活层和 2 倍上采样层组成. 每个上采样模块将特征图的分辨率提高两倍，通过解码网络得到解码后的特征 $f^d\in R^{c\times H/2\times W/2}$.

在训练阶段，解码后的特征分别通过分割头网络得到预测的当前帧的分割图 $\hat{S}_t\in R^{1\times H\times W}$，或者通过 RGB 头网络得到预测的当前帧的 RGB 图像 $\hat{I}_t\in R^{3\times H\times W}$. 分割头网络和 RGB 头网络由两个 3×3 卷积层和一个 2 倍上采样层组成，两个卷积层之间包含批归一化层和 ReLU 激活层. 在测试阶段，解码网络得到的多层特征送入到在线归纳分支，指导网络提高视频分割的准确性和稳定性.

3.2.2 直推推理分支的自监督预训练

视频目标分割任务的标签是像素级别，所以标注非常昂贵. 现在视频目标分割任务中存在少量标注好的训练数据，比如 DAVIS-2017，训练集有 60 段视频标注数据. 直接使用少量的标注数据训练网络会造成网络过拟合，对于未知类别的泛化性较差. 本文提出在视频预测任务预训练的网络参数基础上，在少量的视频目标分割数据上对网络进行微调. 通过在大规模数据上的自监督训练，可以学到有效的、通用的特征表示，减少对于标注数据的依赖. 在少量的视频目标分割数据上对网络进行微调，可以充分利用这些标注数据，将预训练好的特征表示迁移到视频目标分割任务. 直推推理分支的训练分为自监督预训练过程和有监督微调过程. 图 2 中下半部分红色(虚线)表示直推推理分支的自监督预训练过程，上半部分蓝色(实线)表示直推推理分支的有监督微调过程.

由于现有的基于直推推理的方法需要使用大量的合成数据或者标注数据进行训练，增加了训练成本. 本文算法为了减少对合成数据或者标注数据的依赖，采用自监督学习的方法对直推推理分支进行预训练. 视频预测任务是一种视频自监督学习方法，通过输入视频前若干帧，预测视频后一帧的方式，使得网络可以有效建模视频的动态信息. 同时训练过程中只需要原始视频数据，不需要任何的人工标注信息. 由于视频数据在网络上大量存在且容易获取，可以使用大规模视频数据通过自监督方式训练有效的模型. 所以本文中直推推理分支采用视频预测自监督方式预训练，可以在不需要任何人工标注信息的情况下，有效建模视频中的动态变化信息，有助于提高视频分割结果的稳定性和准确性.

如图 2 下半部分所示，为了提高视频预测质量，本文采用生成对抗方式训练. 把直推推理分支作为生成器 G，另外构建一个判别器 D. 判别器 D 用来判断图像是真实的视频帧还是生成器 G 生成的视频帧. 而生成器 G 的目标是生成尽量真实的当前帧图像，使得判别器 D 无法区分生成的图像. 本方法采用 Inception-V3[29] 网络作为判别器，并把最后的全连接层的输出维度修改为 2 维. 预测 t 时刻的视频帧时，需要输入生成器前 δ 帧图像 $\{I_{t-\delta},\cdots,I_{t-1}\}\in R^{3\times H\times W}$. 但是由于有监督微调训练过程以及测试过程需要输入前 δ 帧视频图像和对应的分割图串联 $\{[I_{t-\delta},\hat{S}_{t-\delta}],\cdots,[I_{t-1},\hat{S}_{t-1}]\}\in R^{4\times H\times W}$. 为了和有监督微调训练以及测试过程的输入维度保持一致，所以在自监督预训练过程使用 1 通道的零矩阵填补. 如图 2 所示，当时间 t 时，生成器输入前 δ 帧图像 $\{[I_{t-\delta},Z],\cdots,[I_{t-1},Z]\}\in R^{4\times H\times W}$，其中 Z 表示 1 通道的零矩阵. 最后通过 RGB 头网络生成当前帧的图像，即 $\hat{I}_t=G(\{[I_{t-i},Z]\}_{i=1}^{\delta})$. 然后判别器 D 判别生成的图像是否是真实图像. 生成器 G 和判别器 D 交替训练. 具体的，固定生成器 G 的参数，优化判别器 D，即最小化判别器的错误概率，损失函数的具体形式如下：

$$\min_{W_D}-\log(1-D(\hat{I}_t))-\log D(I_t) \tag{1}$$

其中 $\hat{I}_t$ 是生成器 G 根据前 δ 帧预测生成的当前帧图像，I_t 是真实的当前帧图像. 然后固定判别器 D 的参数，优化生成器 G. 生成器 G 的损失函数由两部分组成：重建损失和判别损失. 损失函数的形式如下：

$$\min_{W_G}\|I_t-\hat{I}_t\|_2-\lambda_{\text{adv}}\cdot\log D(\hat{I}_t) \tag{2}$$

其中，第一项是重建损失，惩罚生成帧和真实帧的差异. 第二项是对抗损失，最大化判别器 D 把生成的图像判别为真实图像的概率. λ_{adv} 是预定义的参数，用来平衡两种损失. 如上所述，判别器 D 和生成器 G 交替优化，使得生成器 G 即直推推理分支可以建模视频序列中的动态变化信息.

3.2.3　直推推理分支的有监督微调训练

为了充分利用现有的少量标注数据，本节将在少量的视频目标分割数据上对直推推理分支进行微调，将通过视频预测自监督学习方式预训练好的特征提取迁移到视频目标分割任务.

直推推理分支的有监督微调训练过程如图 2 上半部分所示，输入前 δ 帧图像和对应的分割图，即把自监督预训练过程中的零矩阵换成对应的分割图 $\{[I_{t-\delta},S_{t-\delta}],\cdots,[I_{t-1},S_{t-1}]\}\in R^{4\times H\times W}$. 然后通过分割头网络输出当前帧的分割图 $\hat{S}_t$. 将二值交叉熵损失作为损失函数：

$$\mathcal{L}(S_t,\hat{S}_t)=-\sum_{s_{i,j}=1}\log P(\hat{s}_{i,j}=1)-\sum_{s_{i,j}=0}\log P(\hat{s}_{i,j}=0) \tag{3}$$

其中 $s_{i,j}$ 和 $\hat{s}_{i,j}$ 为真实分割图 S_t 和预测分割图 $\hat{S}_t$ 对应坐标 (i,j) 的元素值，$\hat{s}_{i,j}=1$ 表示坐标 (i,j) 位置像素预测为前景，而 $\hat{s}_{i,j}=0$ 表示坐标 (i,j) 位置像素预测为背景.

3.3　在线归纳分支

3.3.1　在线归纳分支的网络结构

直推推理分支可以通过视频前若干帧图像和对应的分割图建模视频短期内的时序信息，对于连续变化的目标有较好的分割能力. 然而对于目标遮挡等突然变化的目标无法很好地处理，所以本算法提出一个在线归纳分支. 在线归纳分支利用参考帧的样本直接微调网络，能够学到对于特定目标表观的判别特征，建模长期的表观信息，对于目标遮挡等情况能够恢复正确的分割.

现有的一些基于在线归纳的方法[6-8]，为了准确分割目标，会在第一帧给定的真实分割图上在线训练整个网络，这导致算法耗时较大，无法满足实时要求. 受 Robinson 等人[20] 启发，本文的在线归纳分支引入了轻量级目标网络，在测试时只在线训练这个轻量级的目标网络，其他部分网络只需要离线训练，并且测试时保持固定. 经过在线训练的轻量级目标网络可以学到对于特定目标表观的判别能力，估计出一个粗略的目标分割图，这个粗略的结果作为注意力图指导分割网络得到更精确的结果. 如图 1 下半部分所示，整个在线归纳分支由三个部分组成：特征提取网络、轻量级目标网络、分割网络组成.

特征提取网络使用 ResNet101[30] 网络，由 5 个残差连接模块组成. 输入当前帧的图像 I_t，得到对应的特征，即 $x_t=\mathcal{F}_n(I_t)$，其中 x_t 表示特征，$\mathcal{F}_n$ 表示特征提取网络. x^d 表示特征提取网络不同深度的层提取的特征，d 表示网络深度，每个深度的特征分辨率下降 1/2.

轻量级目标网络由两层卷积网络组成，第一层为 1×1 卷积，减少特征通道数，第二层为 3×3 卷积输出粗略的分割结果. $\mathcal{T}_n$ 表示轻量级的目标网络，则 $a=\mathcal{T}_n(x^4)$，其中 $a\in R^{1\times H/16\times W/16}$ 表示轻量级目标网络输出的粗略的分割结果，输入为特征提取网络第 4 个残差模块对应的特征.

分割网络由一个金字塔池化模块[31]和四个注意力特征精炼模块组成.分割网络输入特征提取网络的特征 x、直推推理分支的时序特征 f 和轻量级目标网络输出的粗略的分割结果 a,输出最终的分割结果,即 $\hat{S}=\mathcal{S}_n(x,f,a)$,其中 $\mathcal{S}_n$ 表示分割网络.

注意力特征精炼模块网络结构如图 3 上半部分所示,每个注意力特征精炼模块输入上层注意力特征精炼模块输出的特征 r^{d-1}、通过跳跃连接来自特征提取网络的特征 x^d、来自直推推理分支的时序特征 f^d 和轻量级目标网络输出的粗略分割图 a,输出融合精炼后的特征 r^d.具体的,来自特征提取网络的特征 x^d 经过一个 1×1 的卷积减少特征通道数为 q,与来自上一个注意力特征精炼模块输出的特征 r^{d-1} 经过上采样 2 倍后逐点相加.来自直推推理分支的时序特征 f^d 经过双线性插值放缩成与 x^d 相同的尺寸,并与前一步相加得到的特征在通道维度串联.串联后得到的特征通过一个 3×3 的卷积层融合多个来源的特征,并输入到注意力模块.注意力模块的网络结构如图 3 下半部分所示,采用残差连接网络结构.输入特征经过两层 3×3 卷积,并与注意力图逐点相乘,得到的特征与原始输入特征相加后作为当前注意力特征精炼模块的输出特征 r^d.

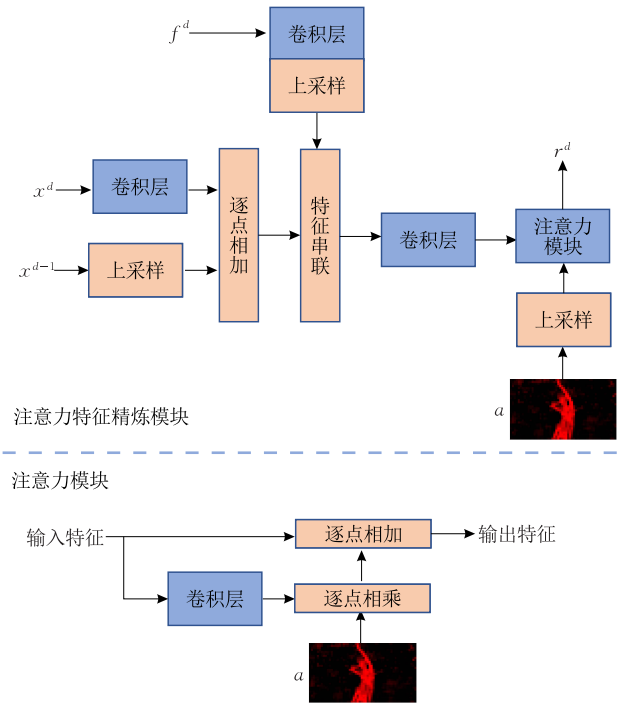


图 3 注意力特征精炼模块的网络结构图

3.3.2 在线归纳分支的离线循环训练

通过之前的介绍,我们已经了解了本文方法的网络结构,其中直推推理分支 G,经过自监督的预训练和微调后,接入在线归纳分支,在离线训练阶段固定参数.在线归纳分支中的特征提取网络 $\mathcal{F}_n$ 是在 ImageNet[32]图像分类数据集上预训练,同样在离线训练阶段固定参数.本节介绍如何离线训练分割网络 $\mathcal{S}_n$.由于轻量级目标网络是在测试时对每个测试视频都要单独训练,所以离线训练过程要模仿测试过程.一个训练样本由以下内容构成,从数据集中随机选择一个视频,然后随机选取一帧图像 I_s 和对应的分割图 S_s,用来训练轻量级目标网络 $\mathcal{T}_n$.在本文的实验中,输入直推推理分支的视频帧数 δ 设为 4,所以采样等间隔的 4 帧图像和对应的分割图 $\{[I_{t-4\eta},S_{t-4\eta}],\cdots,[I_{t-\eta},S_{t-\eta}]\}$ 作为直推推理分支的输入,其中 η 是间隔长度.当前帧 I_t 输入在线归纳分支,对应的真实分割图 S_t 用来求损失.训练轻量级目标网络的损失函数为

$$L_T=v_s\|a_s-S_s\|^2+\lambda_g\cdot\|w\|^2 \tag{4}$$

其中,$a_s=\mathcal{T}_n(\mathcal{F}_n(I_s))$ 是轻量级网络输出的粗略分割图,w 是轻量级目标网络的参数,λ_g 用来控制正则项的强弱.v_s 是用来平衡前景和背景的权重,v_s 的计算公式如下:

$$v_s=\begin{cases}\dfrac{p}{\hat{p}}, & S(n)=1\\ \dfrac{1-p}{1-\hat{p}}, & S(n)=0\end{cases} \tag{5}$$

其中,n 是像素位置索引,$\hat{p}$ 是前景像素数目占总像素数目的比例,$p=\max(p_{\min},\hat{p})$,本文实验设置 $p_{\min}=0.1$.为了提高优化速度,采用高斯牛顿法[33]来优化轻量级目标网络参数 w.

得到训练好的轻量级目标网络后,再训练分割网络 $\mathcal{S}_n$.在测试时,当前帧得到的分割图,在分割下一帧时,要作为前一帧的分割图输入到直推推理分支中.由于网络输出的分割图可能和真实的分割图不完全一致,所以如果使用真实的分割图作为直推推理分割的输入训练分割网络,会导致网络对于真实的分割图过拟合.如果网络输出的分割图出现误差,分割网络没有容错能力会导致误差累积问题.为了提高网络的鲁棒性,本文采取了循环训练的方式,即在一个视频中当前帧预测的分割结果,要作为直推推理分支的下一步的输入.离线循环训练的过程如图 4 所示.训练样本组成在前述的基础上还要增加 $I_{t+\eta}$,$I_{t+2\eta}$,$I_{t+3\eta}$,$I_{t+4\eta}$ 以及对应的分割图 $S_{t+\eta}$,$S_{t+2\eta}$,$S_{t+3\eta}$,$S_{t+4\eta}$.最终训练分割网络的损失函数采用二值交叉熵损失,具体为

$$L_S=\sum_{i=0}^{4}BCE(\hat{S}_{t+i\eta},S_{t+i\eta}) \tag{6}$$

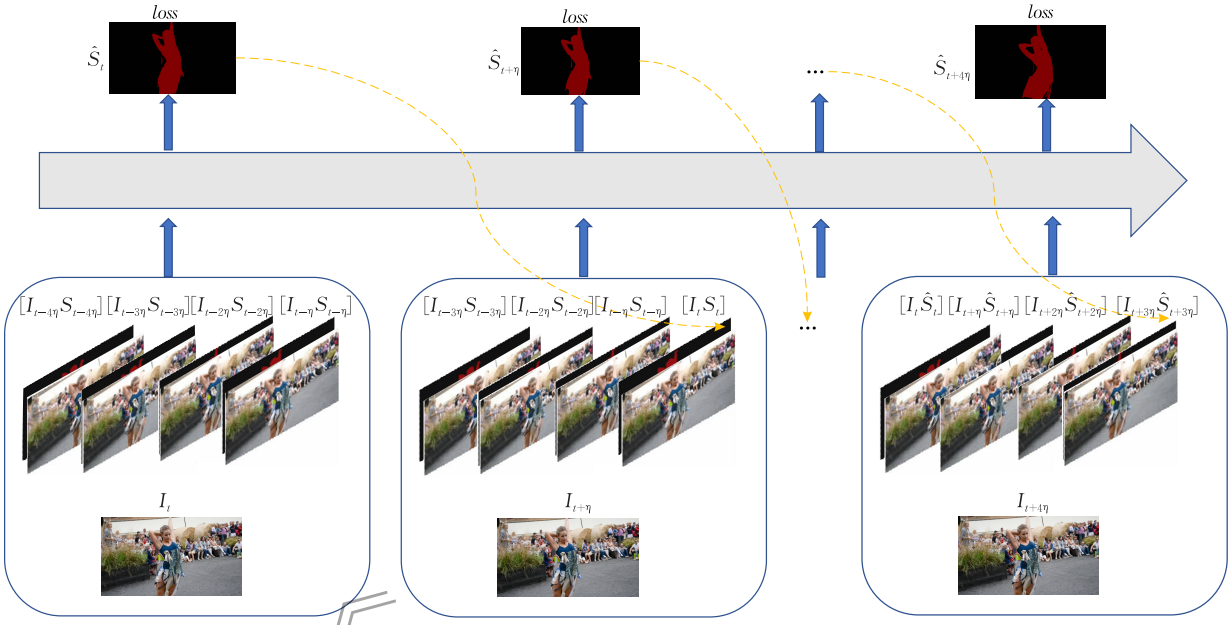


图 4　离线循环训练过程示意图

3.4　测试过程

为了直观理解，我们把测试过程总结为算法 1. 在测试时，当 $t=0$，首先根据给定的第 0 帧图像 I_0 和对应的分割图 S_0，通过数据增广的方式构造数据集 $M_0=\{(x_k,S_k)\}_{k=1}^{K}$，其中 $x_k=\mathcal{F}_n(I_k)$. 增广方式为先把图像的目标区域移除，然后使用快速图像修复方法[34]修复图像背景，然后对目标区域的图像做随机仿射变换生成新的图像和对应的分割图. 首先在数据集 M_0 上优化轻量级目标网络. 然后第 0 帧图像 I_0 和对应的分割图 S_0 串联并且复制 4 份 $\{[I_0,S_0],[I_0,S_0],[I_0,S_0],[I_0,S_0]\}$ 作为直推推理分支的初始输入. 在后续第 t 帧时，算法生成的 $t-1$ 帧分割结果 $\hat{S}_{t-1}$ 作为直推推理分支的输入，即 $\{[I_{t-4},\hat{S}_{t-4}],[I_{t-3},\hat{S}_{t-3}],[I_{t-2},\hat{S}_{t-2}],[I_{t-1},\hat{S}_{t-1}]\}$. 为了使轻量级目标网络对目标的变化更加鲁棒，引入轻量级目标网络的在线更新. 当前帧的特征 x_t 和分割结果 $\hat{S}_t$ 加入到更新轻量级目标网络的数据集中形成 M_t，同时每隔 t_s 帧对轻量级目标网络优化一次. M_t 的最大容量设为 $K_{\max}$，当 M_t 达到最大值时，则删除最早的样本.

算法 1.　测试过程.

输入：视频 $\mathcal{V}=\{I_0,\cdots,I_t,\cdots\}$，第一帧真实分割图 S_0

输出：后续视频帧的分割结果 $\mathcal{S}=\{\hat{S}_1,\cdots,\hat{S}_t,\cdots\}$

1. 构造初始增广数据集 $M_0=\{(x_k,S_k)\}_{k=1}^{K}$
2. 根据式(4)在数据集 M_0 上训练轻量级模板网络 $\mathcal{T}_n$
3. FOR $t=1,2,3\cdots$
4. 　$x_t=\mathcal{F}_n(I_t)$
5. 　$a_t=\mathcal{T}_n(x_t)$
6. 　IF $t<4$ THEN
7. 　　$f_t=G([I_0,S_0]_{\times(4-t)},\cdots,[I_{t-1},\hat{S}_{t-1}])$
8. 　ELSE
9. 　　$f_t=G([I_{t-4},\hat{S}_{t-4}][I_{t-3},\hat{S}_{t-3}],$
　　$[I_{t-2},\hat{S}_{t-2}][I_{t-1},\hat{S}_{t-1}]$
10. 　END IF
11. 　$\hat{S}_t=\mathcal{S}_n(x_t,f_t,a_t)$
12. 　$M_t=M_{t-1}+\{\hat{S}_t,x_t\}$ 将最新的 $\hat{S}_t$ 和 x_t 加入到 M_t
13. 　IF $t \bmod t_s=0$ THEN
14. 　　根据式(4)在数据集 M_t 上优化轻量级模板网络 $\mathcal{T}_n$
15. 　END IF
16. END FOR

4　实验结果与分析

本节首先介绍实验使用的参数设置、数据集，然后与现有的最好的视频目标算法进行对比分析，最后对本文方法的各个模块的作用进行消融实验分析.

4.1　实验细节设置

输入直推推理分支的视频帧数 δ 设为 4；离线训练分割网络时视频帧间隔 η 从 1，2，3 中随机选择；判别损失的权重 λ_{adv} 设为 0.01；分割网络的中间特征通道数 q 设为 64；在线更新轻量级模板网络的时间间隔 t_s 设为 8；测试时 M_0 的样本容量 K 设为 5；M_t 的最大容量 $K_{\max}$ 设为 80. 直推推理分支的自监督

预训练使用 YouTube-VOS[10] 数据集的训练集的原始视频数据，总共 3471 个视频，不使用数据集中的标注信息. 直推推理分支的有监督微调训练使用 DAVIS-2017[35] 数据集的训练集进行训练，共 60 个视频序列. 在线归纳分支的离线循环训练使用和测试数据集对应的训练集，比如在 DAVIS-2017 的测试集测试，就在 DAVIS-2017 数据集的训练集训练.

4.2 评价指标

为了综合评价视频目标分割方法，本文使用三个常用的评价指标，即区域相似度$\mathcal{J}$、轮廓准确性$\mathcal{F}$和总体得分$\mathcal{J}\&\mathcal{F}$.

(1) 区域相似度$\mathcal{J}$. 衡量算法分割区域的准确性，是算法输出的分割图和真实分割图相交区域的面积与相并面积的比值. 给定算法估计的分割图 $\hat{S}$ 和真实的分割图 S，区域相似度$\mathcal{J}$的计算公式为

$$\mathcal{J}=\frac{S\cap\hat{S}}{S\cup\hat{S}}.$$

(2) 轮廓准确性$\mathcal{F}$. 衡量算法分割目标轮廓的准确性，是算法输出的分割图轮廓上的像素点和真实分割图轮廓上像素点之间准确率 P_c 和召回率 R_c 的 $F1$ 度量，计算公式为

$$\mathcal{F}=\frac{2P_cR_c}{P_c+R_c}.$$

(3) 总体得分$\mathcal{J}\&\mathcal{F}$. 该指标是区域相似度$\mathcal{J}$和轮廓准确性$\mathcal{F}$的平均得分，计算公式为

$$\mathcal{J}\&\mathcal{F}=\frac{\mathcal{J}+\mathcal{F}}{2}.$$

4.3 数据集

为了验证本文方法的有效性，我们在 YouTube-VOS[10]、DAVIS-2017[35]、DAVIS-2016[36] 三个有挑战性的数据集上进行了广泛的实验. YouTube-VOS[10] 是大规模的多目标视频目标分割数据集，它包括 3471 个训练视频和 474 个验证视频. 验证视频中包含 91 类目标，其中 26 类目标不在训练集中出现. 出现在训练集中的目标类别被称为已见(seen)目标. 若目标类别没有出现在训练集中，只出现在验证集，则该类别被称为未见(unseen)类别. DAVIS-2017[35] 数据集是一个有挑战的多目标视频目标分割数据集，训练集有 60 个视频序列，验证集有 30 个视频序列. DAVIS-2016[36] 数据集是单目标视频目标分割数据集，总共有 50 个视频序列组成，其中 30 个用来训练，20 个用来测试. 三个数据集都包含尺度变化、严重遮挡、表观变化、背景杂乱等困难样本.

4.4 与现有其他算法的对比结果

4.4.1 DAVIS-2017 数据集上的对比结果

本文首先将所提算法与现有算法在 DAVIS-2017 数据集上进行对比实验. 如表 1 所示，对比方法分为只是用 DAVIS-2017 数据集进行训练和使用额外数据集(例如 YouTube-VOS 数据集和合成数据集等)训练两类. 本文的方法在平均得分$\mathcal{J}\&\mathcal{F}$、区域相似度$\mathcal{J}$、轮廓准确性$\mathcal{F}$三个指标上分别取得了 72.9%、70.0%、75.8%，均为只使用 DAVIS-2017[35] 数据集训练的方法中的最佳效果. 同时也超过了大多数使用额外训练数据的方法例如 AGSS-VOS[37]、A-GAME[11]、FEELVOS[18]. 本文的方法指标仅次于 STM[17] 方法，但是 STM 方法额外的使用了 YouTube-VOS[10] 数据集和合成数据集进行训练. 如果同样只使用 DAVIS-2017[35] 数据集进行训练，本文的方法则比 STM(DV17)[17] 平均指标高 1.3%. 通过实验结果可以表明，现有的基于直推推理的方法依赖于额外的合成数据或者标注数据进行训练. 而本文方法的直推推理分支使用了自督学习预训练，可以减少对于合成数据或者标注数据的依赖.

表 1 本文算法及其对比方法在 DAVIS-2017 数据集上的结果

方法	额外训练数据		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
	+YouTube-VOS	+合成数据			
OSNM[14]		✓	54.8	52.5	57.1
FAVOS[38]			58.2	54.6	61.8
OSVOS[6]			60.3	56.6	63.9
STCNN[39]			61.7	58.7	64.6
RANet[13]		✓	65.7	63.2	68.2
RGMP[12]		✓	66.7	64.8	68.6
OnAVOS[8]			67.9	64.5	71.2
OSVOS-S[7]			68.0	64.7	71.3
FRTM[20]			68.8	—	—
GC[27]			71.4	69.3	73.5
STM(DV17)[17]		✓	71.6	69.2	74.0
TVOS[26]			72.3	69.9	74.7
本文算法			**72.9**	**70.0**	**75.8**
AGSS-VOS[37]	✓		67.4	64.9	69.9
A-GAME[11]	✓	✓	70.0	67.2	72.7
FEELVOS[18]	✓		71.5	69.1	74.0
SAT[25]	✓		72.3	68.6	76.0
STM[17]	✓	✓	81.7	79.2	84.3

4.4.2 DAVIS-2016 数据集上的对比结果

本文算法及其对比方法在 DAVIS-2016 数据集上的结果如表 2 所示. 在表 2 中最后一列展示了本文方法和其他对比方法在 DAVIS-2016 数据集上的平均运行速度. 在本文的速度计算中，包含了算法 1 测试过程中所有步骤的时间消耗，其中步骤 1 构造增广数据集 M_0 耗时 0.08 s，步骤 2 训练轻量

级模板网络$\mathcal{T}_n$耗时 0.39 s. 本文算法的速度是在 Nvidia V100 GPU 下测得，对比方法的速度从原始论文中获得，其中 STM 算法速度中小括号内的数值是使用与本文实验相同硬件的实测值. 本文在 DAVIS-2016[36] 数据集上，本文的方法在平均得分$\mathcal{J}\&\mathcal{F}$、区域相似度$\mathcal{J}$、轮廓准确性$\mathcal{F}$三个指标上分别取得了 84.6%、84.9%、84.3%. 超过了大多数不使用额外训练数据集的方法 OSNM[14]、OSVOS[6]、FAVOS[38]、FRTM[20]、RGMP[12]、STCNN[39]，甚至超过了额外使用 YouTube-VOS 数据集训练的方法 A-GAME[11] 和 FEELVOS[18]. 本文的方法仅次于 OnAVOS[8]、OSVOS-S[7]、STM(DV17)[17]，其中 OnAVOS[8]、OSVOS-S[7] 需要在线微调整个网络，所以导致算法运行速度慢，无法满足实时需要.

表 2　本文算法及其对比方法在 DAVIS-2016 数据集上的结果

方法	额外训练数据		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	速度/(帧/秒)
	+YouTube-VOS	+合成数据				
OSNM[14]		✓	73.5	74.0	72.9	7.14
OSVOS[6]			80.2	79.8	80.6	0.11
FAVOS[38]			81.0	82.4	79.5	0.56
FRTM[20]			81.7	—	—	21.90
RGMP[12]		✓	81.8	81.5	82.0	7.70
STCNN[39]			83.8	83.8	83.8	0.25
OnAVOS[8]			85.5	**86.1**	84.9	0.08
OSVOS-S[7]			86.0	85.6	86.4	0.22
STM(DV17)[17]		✓	**86.5**	84.8	**88.1**	6.25(8.7)
本文算法			84.6	84.9	84.3	18.00
FEELVOS[18]	✓		81.7	81.1	82.2	2.22
A-GAME[11]	✓	✓	—	82.0	—	14.30
SAT[25]	✓		83.1	82.6	83.6	**39.00**

本文方法的运行速度达到 OnAVOS[8]、OSVOS-S[7] 的 225 倍和 81 倍. STM(DV17)[17] 方法需要使用额外的合成数据集训练，并且算法运行速度小于本文方法的 1/2.

4.4.3　YouTube-VOS 数据集上的对比结果

表 3 展示了本文方法和现有算法在 YouTube-VOS[10] 数据集上的对比结果. 本文的方法性能超过了所有不使用额外训练数据的方法 OnAVOS[8]、RVOS[40]、OSVOS[6]、S2S[10]、A-GAME[11]、FRTM[20]. 本文的方法在总体得分$\mathcal{G}$上甚至超过了使用额外合成数据集的训练的 PReMVOS[24] 方法 6.9%. 本文的方法仅次于 STM[17] 方法，但是 STM[17] 算法需要用额外的合成数据集训练，且算法运行速度远远低于本文的方法.

表 3　本文算法及其对比方法在 YouTube-VOS 数据集上的结果

方法	合成数据	$\mathcal{G}$	Seen		Unseen	
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
OnAVOS[8]		55.2	60.1	62.7	46.1	51.4
RVOS[40]		56.8	63.6	67.2	45.5	51.0
OSVOS[6]		58.8	59.8	60.5	54.2	60.7
SAT[25]		63.6	67.1	70.2	55.3	61.7
S2S[10]		64.4	71.0	70.0	55.5	61.2
A-GAME[11]		66.1	67.8	69.5	60.8	66.2
TVOS[26]		67.8	67.1	69.4	63.0	71.6
FRTM[20]		72.1	72.3	76.2	65.9	74.1
GC[27]		73.2	72.6	75.6	68.9	75.7
本文算法		**73.8**	**73.9**	**77.9**	**67.5**	**75.7**
PReMVOS[24]	✓	66.9	71.4	—	56.5	—
STM[17]	✓	79.4	79.7	84.2	72.8	80.9

图 5 是本文方法与现有方法的定性效果，从图 5 中可以看出本文的方法在目标遮挡(第一、二行)，尺度变换(第三、四行)，等困难场景均能取得较好的效果. 与现有算法 RANet[13] 和 TVOS[26] 相比本文算法更加鲁棒.

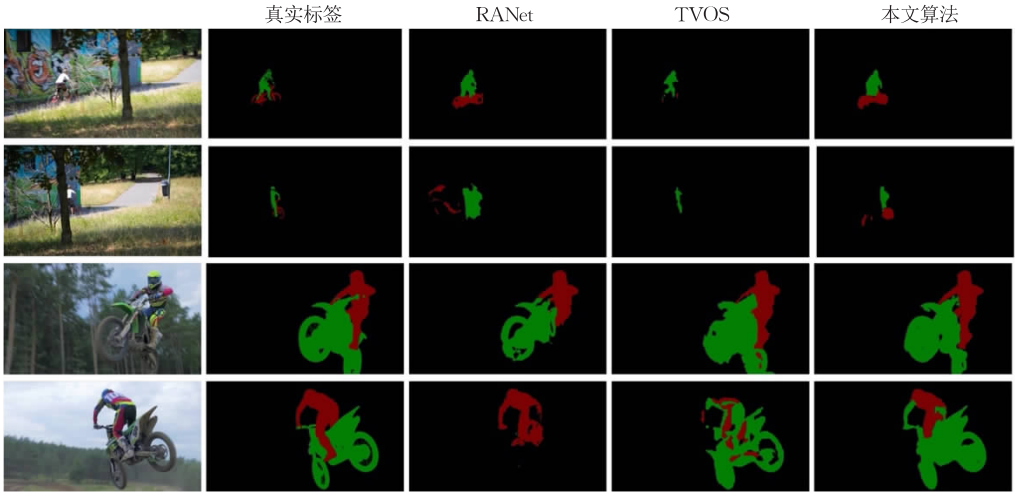


图 5　本文方法与现有方法的定性效果图比较

4.5 消融实验与分析

为了证明本文提出方法不同部件的作用，我们在 DAVIS-2017 数据集上进行了一系列的消融对比实验. 我们设计了四个变体方法，分别是：

(1) 取消直推推理分支. 不使用直推推理分支，只使用在线归纳分支.

(2) 取消自监督预训练. 训练直推推理分支时，不使用自监督预训练，只在 DAVIS-2017 视频分割数据集上训练.

(3) 取消有监督微调. 训练直推推理分支时，只使用自监督预训练，不在 DAVIS-2017 视频分割数据集上微调网络.

(4) 取消循环训练. 在离线训练分割网络 $\mathcal{S}_n$ 时，不使用循环训练方式.

表 4 展示了不同变体方法在 DAVIS-2017 上的定量指标. 图 7 展示了消融实验中不同变体方法的定性效果对比图. 通过对比表 4 第一行和第五行，我们发现使用直推推理分支会使平均得分 $\mathcal{J}\&\mathcal{F}$ 提高 2.7%. 说明直推推理分支可以有效地建模视频中的表观动态和运动信息，帮助产生更加准确和稳定的分割结果. 从图 7 中第一行的结果同样可以看出，不使用直推推理分支的变体方法会导致不同目标之间相互分割错误. 通过对比表 4 第二行和第五行，发现训练直推推理分支时使用自监督预训练是非常必要的，因为 DAVIS-2017 训练集只有 60 段视频，所以只使用 DAVIS-2017 从头开始训练直推推理分支会导致过拟合. 从图 7 中第二行的结果可以直观看到，不使用自监督预训练的方法会导致目标分割错误和分割不完全. 通过对比表 4 第三行和第五行，发现训练直推推理分支时，在 DAVIS-2017 数据集上微调网络也非常重要，因为可以缩小视频预测任务和视频分割任务之间的差异. 从图 7 中第三行的结果可以看出，不在 DAVIS-2017 数据集上微调网络直推推理分割会导致目标分割不完整. 通过对比表 4 第四行和第五行，发现训练分割网络 $\mathcal{S}_n$ 时，采用循环训练策略非常有必要，因为循环训练可以使网络对分割结果的偏差有更强的包容性和纠错性. 从图 7 中第四行的结果可以看出不使用循环训练会导致分割不完全，并且随着误差累积，分割误差会越来越大. 图 6 是 DAVIS-2017 数据集中视频序列每帧 $\mathcal{J}$ 和 $\mathcal{F}$ 指标的平均变换情况折线图，从中我们可以进一步得到相同的结论. 本文提出的完整方法对比其他变体方法在后半部分视频帧中取得了大幅度的领先，说明本文完整的方法具有更好的鲁棒性.

表 4 DAVIS-2017 数据集上的消融实验结果

变体方法	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
取消直推推理分支	70.2	67.4	73.0
取消自监督预训练	65.9	64.1	67.8
取消有监督微调	68.7	65.8	71.5
取消循环训练	65.3	64.1	67.4
本文完整算法	**72.9**	**70.0**	**75.8**

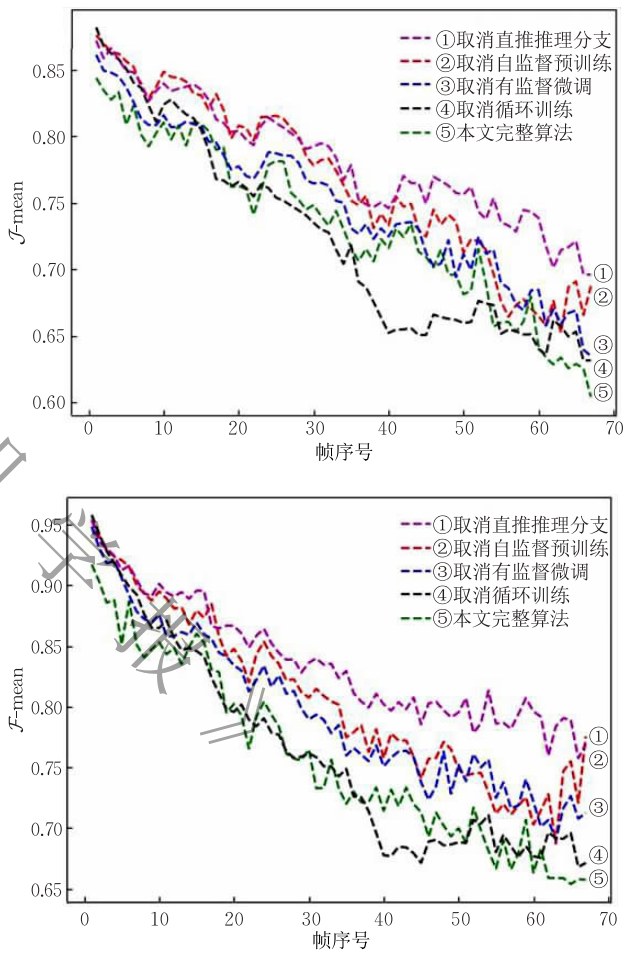


图 6 DAVIS-2017 数据集中视频序列每帧 $\mathcal{J}$ 和 $\mathcal{F}$ 指标的平均变换情况折线图

表 5 展示了分割网络中中间特征的通道数最佳取值的对比实验. 从表 5 可以看出，通道数 q 取 64 效果最好. 取值 32 时，由于网络表达能力不足会导致分割准确度降低. 取值超过 64 时，导致网络过拟合，同样造成测试准确度的下降. 图 8 展示了本文方法的两个失败样本，由此可见本文方法对于绳子等细小的物品分割还不够细致.

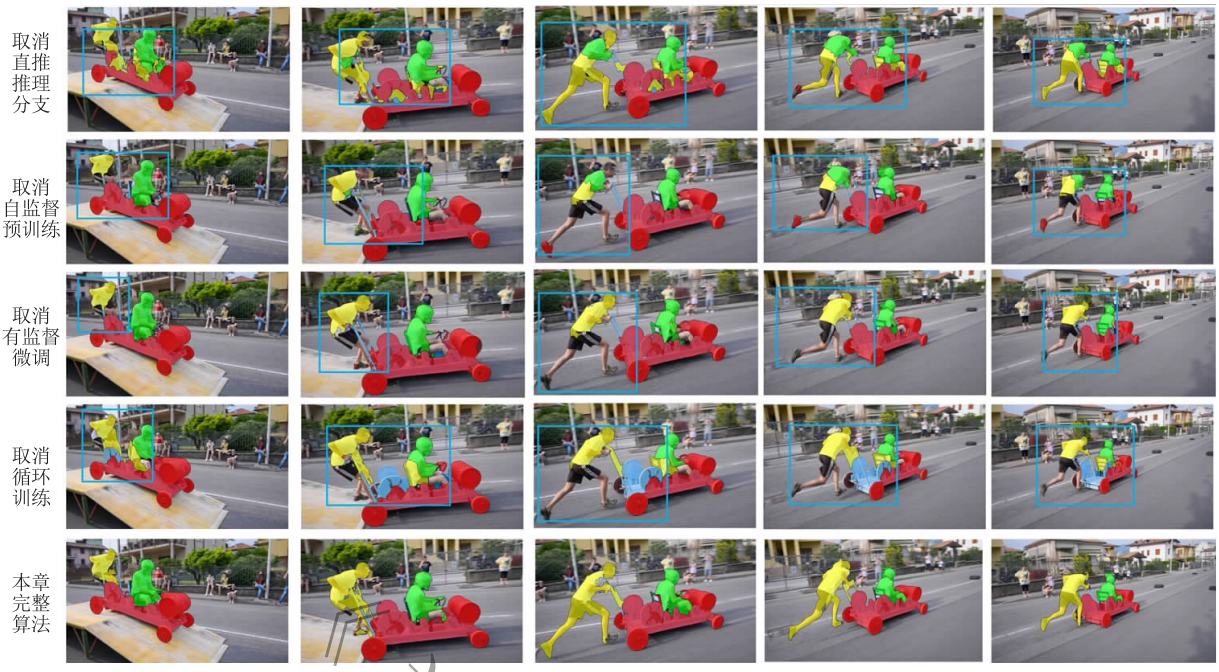


图 7　消融实验定性对比结果图

表 5　分割网络中间特征通道数 q 的取值分析

q 取值	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
32	71.9	69.3	74.4
64	**72.9**	**70.0**	**75.8**
128	72.2	69.7	74.6
256	70.7	68.3	73.0

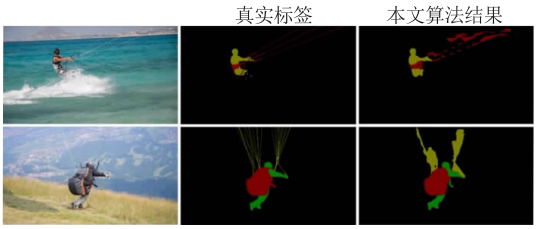


图 8　失败样本展示

5　总结与展望

本文提出了一种新的结合直推推理和在线归纳的快速视频目标分割方法，该方法由直推推理分支和在线归纳分支组成. 本文方法能够结合两类方法的优点并且避免了各自的缺点. 其中，直推推理分支可以建模视频的短期时序信息，对于目标的连续变化有较好的分割能力，并且提高视频分割的稳定性. 直推推理分支通过自监督学习的方式预训练，减少了对于标注数据和合成数据的依赖. 在线归纳分支通过利用参考帧优化网络，使得网络对目标表观具有较强的判别能力. 但是现有的基于在线归纳的方法，需要在给定的第一帧分割图上在线训练整个网络，导致时间消耗较大，很难满足算法实时需求. 为了提高测试速度，本文通过在线更新一个非常轻量的模板网络，使得轻量模板网络产生一个粗略的目标分割图. 这个粗略的分割图作为注意力图与时序特征和当前帧的图像特征一起经过解码模块产生最终的分割结果. 通过大量的实验证明，本文分割网络的测试速度能接近实时，并在多个数据集上取得了优于现有方法的性能. 进一步研究自监督学习，并通过自监督学习提高视频分割任务的性能是我们未来的研究方向.

参　考　文　献

[1] Ros G, Ramos S, Granados M, et al. Vision-based offline-online perception paradigm for autonomous driving//Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA, 2015: 231-238

[2] Saleh K, Hossny M, Nahavandi S. Kangaroo vehicle collision detection using deep semantic segmentation convolutional neural network//Proceedings of the 2016 International Conference on Digital Image Computing: Techniques and Applications. Gold Coast, Australia, 2016: 1-7

[3] Erdélyi A, Barát T, Valet P, et al. Adaptive cartooning for privacy protection in camera networks//Proceedings of the 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance. Seoul, Korea, 2014: 44-49

[4] Liu Li-Jie, Cai De-Jun, Weng Nan-Shan. A motion-oriented

video object segmentation algorithm. Chinese Journal of Computers, 2000, 23(12): 1326-1231(in Chinese)
(刘李杰, 蔡德钧, 翁南钐. 一种面向运动的视频对象分割算法. 计算机学报, 2000, 23(12): 1326-1231)

[5] Chen Jia, Chen Ya-Song, Li Wei-Hao, et al. Application and prospect of deep learning in video object segmentation. Chinese Journal of Computers, 2021, 44(3): 609-631(in Chinese)
(陈加, 陈亚松, 李伟浩等. 深度学习在视频对象分割中的应用与展望. 计算机学报, 2021, 44(3): 609-631)

[6] Caelles S, Maninis K K, Pont-Tuset J, et al. One-shot video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 221-230

[7] Maninis K K, Caelles S, Chen Y, et al. Video object segmentation without temporal information. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(6): 1515-1530

[8] Voigtlaender P, Leibe B. Online adaptation of convolutional neural networks for video object segmentation//Proceedings of the British Machine Vision Conference. London, Britain, 2017: 1-13

[9] Perazzi F, Khoreva A, Benenson R, et al. Learning video object segmentation from static images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2663-2672

[10] Xu N, Yang L, Fan Y, et al. YouTube-VOS: Sequence-to-sequence video object segmentation//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 585-601

[11] Johnander J, Danelljan M, Brissman E, et al. A generative appearance model for end-to-end video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8953-8962

[12] Oh S W, Lee J Y, Sunkavalli K, et al. Fast video object segmentation by reference-guided mask propagation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7376-7385

[13] Wang Z, Xu J, Liu L, et al. RANet: Ranking attention network for fast video object segmentation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 3978-3987

[14] Yang L, Wang Y, Xiong X, et al. Efficient video object segmentation via network modulation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 6499-6507

[15] Chen Y, Pont-Tuset J, Montes A, et al. Blazingly fast video object segmentation with pixel-wise metric learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1189-1198

[16] Hu Y T, Huang J B, Schwing A G. VideoMatch: Matching based video object segmentation//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 54-70

[17] Oh S W, Lee J Y, Xu N, et al. Video object segmentation using space-time memory networks//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 9226-9235

[18] Voigtlaender P, Chai Y, Schroff F, et al. FEELVOS: Fast end-to-end embedding learning for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 9481-9490

[19] Vondrick C, Shrivastava A, Fathi A, et al. Tracking emerges by colorizing videos//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 391-408

[20] Robinson A, Lawin F J, Danelljan M, et al. Learning fast and robust target models for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 7406-7415

[21] Khoreva A, Benenson R, Ilg E, et al. Lucid data dreaming for video object segmentation. International Journal of Computer Vision, 2019, 127(9): 1175-1197

[22] Cheng J, Tsai Y H, Wang S, et al. SegFlow: Joint learning for video object segmentation and optical flow//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 686-695

[23] Hu P, Wang G, Kong X, et al. Motion-guided cascaded refinement network for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1400-1409

[24] Luiten J, Voigtlaender P, Leibe B. PReMVOS: Proposal-generation, refinement and merging for video object segmentation//Proceedings of the Asian Conference on Computer Vision. Perth, Australia, 2018: 565-580

[25] Chen X, Li Z, Yuan Y, et al. State-aware tracker for real-time video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9384-9393

[26] Zhang Y, Wu Z, Peng H, et al. A transductive approach for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6949-6958

[27] Li Y, Shen Z, Shan Y. Fast video object segmentation using the global context module//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 735-750

[28] Han K, Wang Y, Tian Q, et al. GhostNet: More features from cheap operations//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1580-1589

[29] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2818-2826

[30] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[31] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2881-2890

[32] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252

[33] Danelljan M, Bhat G, Khan F S, et al. ATOM: Accurate tracking by overlap maximization//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4660-4669

[34] Telea A. An image inpainting technique based on the fast marching method. Journal of Graphics Tools, 2004, 9(1): 23-34

[35] Pont-Tuset J, Perazzi F, Caelles S, et al. The 2017 DAVIS challenge on video object segmentation. arXiv preprint arXiv: 1704.00675, 2017

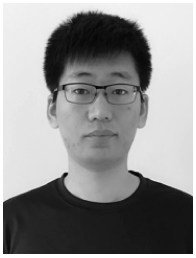
[36] Perazzi F, Pont-Tuset J, McWilliams B, et al. A benchmark dataset and evaluation methodology for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 724-732

[37] Lin H, Qi X, Jia J. AGSS-VOS: Attention guided single-shot video object segmentation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 3949-3957

[38] Cheng J, Tsai Y H, Hung W C, et al. Fast and accurate online video object segmentation via tracking parts//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2018: 7415-7424

[39] Xu K, Wen L, Li G, et al. Spatiotemporal cnn for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1379-1388

[40] Ventura C, Bellver M, Girbau A, et al. RVOS: End-to-end recurrent network for video object segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5277-5286



XU Kai, Ph. D. candidate. His research interests include video object segmentation, video self-supervised learning.

LI Guo-Rong, Ph. D., associate professor. Her research interests include video object tracking and segmentation, group counting and event analysis.

HONG De-Xiang, M. S. candidate. His research interests include video object segmentation, video self-supervised learning.

ZHANG Wei-Gang, Ph. D., professor. His research interests include image processing, video analysis, pattern recognition.

QI Yuan-Kai, Ph. D. His research interests include object tracking, video segmentation, machine learning.

HUANG Qing-Ming, Ph. D., professor. His research interests include multimedia computing, image processing, pattern recognition, machine learning, computer vision.

Background

The video object segmentation task aims to automatically obtain the pixel level mask corresponding to the object of interest in the video sequence through the algorithm. Video object segmentation has a variety of important applications, including automatic driving, video surveillance, video editing and so on. Video object segmentation (VOS) is a highly challenging problem since appearance changes, scale change, occlusion etc.

The existing methods can be divided into two categories according to the different utilization of the real label of the first frame of a given video. One is based on online inductive learning, the other is based on transductive reasoning learning. Previous methods based on online inductive learning finetune the whole network on the given initial frame segmentation mask in the inference stage in order to obtain accurate results, resulting in large time consumption and difficult to meet the real-time requirements. In addition, previous methods based on transductive reasoning need to use a large amount of synthetic data or annotation data when modeling video temporal reasoning rules, which increases the cost of algorithm training. In order to make full use of the advantages of two kinds of algorithms based on online inductive learning and transductive reasoning, and avoid the disadvantages of the two methods, a fast video object segmentation method combining online inductive learning and transductive reasoning is proposed in this paper. Specifically,

the transductive reasoning branch is pre-trained by the video prediction self-supervised learning method. It can model the short-term temporal transformation and motion information of the video through the images and segmentation results of the few previous frames, so as to infer the segmentation results of the current frame. The learned temporal features can guide the network to improve the stability of video segmentation. In the pre-training process of the transductive reasoning branch, only the original video data need to be used without any additional synthetic data or manual annotation. The online inductive learning branch is trained online according to the reference frame to learn the appearance discriminant feature of the target and provide long-term appearance discriminant power. In order to improve the test speed, different from the previous methods, the inductive learning branch proposed in this paper does not use the first frame to fine tune the whole network online, but updates a very light-weight template network online to make the template network produce a rough object segmentation mask. This rough segmentation mask is used as the attention map of attention mechanism, and the final segmentation result is generated through the decoding module together with the temporal features and the image features of the current frame.

This work is supported by the National Key Research and Development Program (No. 2018YFE0118400), the National Natural Science Foundation of China (Nos. 61772494, 61976069, 61902092) and the Youth Promotion Association of Chinese Academy of Sciences.