一种改进的视频分割网络及其全局信息优化方法

张 琳 [1, 2, 3] 陆 耀 [1, 2] 卢丽华 [1, 2] 周天飞 [1, 2] 史青宣 [4]

摘 要 提出了一种基于注意力机制的视频分割网络及其全局信息优化训练方法. 该方法包含一个改进的视频分割网络, 在对视频中的物体进行分割后, 利用初步分割的结果作为先验信息对网络优化, 再次分割得到最终结果. 该分割网络是一种双流卷积网络, 以视频图像和光流图像作为输入, 分别提取图像的表观信息和运动信息, 最终融合得到分割掩膜 (Segmentation mask). 网络中嵌入了一个新的卷积注意力模块, 应用于卷积网络的高层次特征与相邻低层次特征之间, 使得高层语义特征可以定位低层特征中的重要区域, 提高网络的收敛速度和分割准确度. 在初步分割之后, 本方法提出利用初步结果作为监督信息对表观网络的权值进行微调, 使其辨识前景物体的特征, 进一步提高双流网络的分割效果. 在公开数据集 DAVIS 上的实验结果表明, 该方法可准确地分割出视频中时空显著的物体, 效果优于同类双流分割方法. 对注意力模块的对比分析实验表明, 该注意力模块可以极大地提高分割网络的效果, 较本方法的基准方法 (Baseline) 有很大的提高.

关键词 视频物体分割, 卷积神经网络, 注意力机制, 全局信息优化

引用格式 张琳, 陆耀, 卢丽华, 周天飞, 史青宣. 一种改进的视频分割网络及其全局信息优化方法. 自动化学报, 2022, **48**(3): 787−796

DOI 10.16383/j.aas.c190292

An Improved Video Segmentation Network and Its Global Information Optimization Method

ZHANG Lin[1, 2, 3] LU Yao[1, 2] LU Li-Hua [1, 2] ZHOU Tian-Fei[1, 2] SHI Qing-Xuan[4]

Abstract This paper presents an attention-based video segmentation network and its global information optimization training method. We propose an improved segmentation network, and use it to compute initial segmentation masks. Then the initial masks are considered as priors to finetune the network. Finally, the network with the learnt weight generates fine masks. Our two-stream segmentation network includes appearance branch and motion branch. Fed with image and optical flow image separately, the network extracts appearance features and motion features to generate segmentation mask. An attention module is embedded in the network, between the adjacent high level feature and low level feature. Thus the high level features locate the semantic region for the low level feature, speeding up the network convergence and improving segmentation quality. We propose to optimize the initial masks to finetune the original appearance network weights, making the network recognize the object and improving the network performance. Experiments on DAVIS show the effectiveness of the segmentation framework. Our method outperforms the traditional two-stream segmentation algorithms, and achieves comparable results with algorithms on the dataset′s leaderboard. Validation experiment illustrates our attention module greatly improves the network performance than the baseline.

Key words Video object segmentation, convolutional neural network (CNN), attention mechanism, global information optimization

Citation Zhang Lin, Lu Yao, Lu Li-Hua, Zhou Tian-Fei, Shi Qing-Xuan. An improved video segmentation network and its global information optimization method. *Acta Automatica Sinica*, 2022, **48**(3): 787−796

收稿日期 2019-04-10 录用日期 2019-07-30
Manuscript received April 10, 2019; accepted July 30, 2019
国家自然科学基金 (61273273), 国家重点研发计划 (2017YFC0112001) 资助
Supported by National Natural Science Foundation of China (61273273) and National Key Research and Development Program of China (2017YFC0112001)
本文责任编委 桑农
Recommended by Associate Editor SANG Nong

1. 北京理工大学计算机学院 北京 100081 2. 智能信息技术北京市重点实验室 北京 100081 3. 北方电子设备研究所 北京 100083 4. 河北大学网络空间安全与计算机学院 保定 071000

1. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081 2. Beijing Laboratory of Intelligent Information Technology, Beijing 100081 3. The Institute of North Electronic Equipment, Beijing 100083 4. School of Cyber Security and Computer, Hebei University, Baoding 071000

视频物体分割[1−3] 是计算机视觉领域中的重要研究方向, 与其他任务诸如行为分析[4]、视频内插[5] 等有紧密联系. 当输入一个视频时, 视频分割算法针对视频中的每一帧图像计算出一幅分割掩膜, 该掩膜可提取图像中具有显著运动特征的前景. 由于前景物体的外表变形、遮挡和背景杂乱等困难, 视频物体分割是一个具有挑战性的问题. 而分割过程中无需先验信息及人工干预的无监督视频物体分割更为困难.

为解决无监督视频分割问题, 本文提出了一种改进的双流视频物体分割网络, 并利用该网络产生初步的分割结果. 作为视频分析中常用的网络结

构[6−8], 双流网络可以并行分析视频中的时域−空域信息. 本文提出具有相同分支结构的双流分割网络, 同时对表观及运动做出分割, 并通过融合得到分割结果.

为使高层的特征指导低层特征提取更具判别力的特征, 本方法提出在网络中加入注意力模块[9−12]. 该注意力模块用于主干 (Backbone) 网络的相邻特征层之间, 可将高层特征转化为与低层特征具有相同维度的注意力张量 (Tensor), 强化高层特征所指定的更具语义信息的特征维度, 同时弱化与目标不相关的特征, 使低层特征具有更强的判别力, 实现高层语义特征对低层特征的监督. 实验表明, 加入注意力模块后网络的收敛速度更快, 且网络的分割效果得到提高.

经过分割后, 视频中有些图像的分割结果较准确, 有些图像则较差. 为了对初始分割做优化, 文献 [13] 将交互图像分割中的优化方法[3] 用于视频物体分割任务中. 基于初始分割结果, 此类方法针对每个视频前景物体的表观特征建立基于图模型 (Graph) 的能量函数并优化. 然而图模型方法无法准确建模表观变化大的运动物体.

本文提出利用初始结果作为先验对表观分支网络进行权值微调的方法. 利用阈值对初始结果进行挑选, 选择其中的可靠像素作为信息监督网络训练过程, 可以使得表观网络识得视频前景物体, 同时避免被不可靠像素所误导.

本文的主要贡献可归纳如下: 1) 提出了一种视频物体分割方法, 首先利用双流卷积网络对视频分割, 得到初步的分割结果; 进而利用初步结果对分割网络的表观分支做权值微调, 使其适应该视频中的前景物体; 再次使用新权值下的分割网络对视频做分割, 得到最终结果. 通用数据集 DAVIS 上的实验显示该方法具有很好的分割能力, 能够准确的对视频中的运动物体进行分割. 2) 提出了一个简单却有效的卷积注意力模块, 该模块可以用于分割网络中并提高卷积神经网络 (Convolutional neural network, CNN) 的表现能力. 3) 提出了利用初步分割结果作为先验信息对网络进行微调的方法, 该方法可以使分割网络学习到视频中前景物体的表观特征, 提高分割效果.

1 相关工作

1.1 视频分割

为解决视频物体分割问题, 很多有效算法被提出. 根据分割算法中提取特征所使用的方法, 可将其分类为: 1) 基于非深度特征 (Non-deep learning) 的分割算法[14−22]; 2) 基于深度 (Deep learning) 特征的分割算法[23−32]. 此外, 根据人与算法的交互程度, 两类算法均具有 3 种子类别: 1) 无监督 (Unsupervised) 分割算法: 没有任何先验信息, 全自动的分割算法. 2) 半监督 (Semi-supervised) 分割算法: 需要由人指定分割区域, 通常以视频中第 1 帧图像的真值给出. 3) 全监督 (Supervised) 分割算法: 需要人与算法的多次交互, 以修正长时间分割中的误差.

非深度学习分割方法使用人工定义的描述子, 通过对整个视频上的运动特征[19−20]、表观特征[14, 18] 或二者的结合[15, 17] 综合分析产生分割预测. 作为无监督分割方法, FST (Fast object segmentation in unconstrained video)[21] 通过分析运动特征得到具有显著相对运动的前景区域. 更进一步, 文献 [20] 提出综合分析运动边缘、表观边缘与超像素, 得到时空边缘概率图像, 利用测地距离对其优化得到更好的分割预测. 文献 [17] 是基于提议 (Proposal) 的分割方法. 该方法首先调用文献 [22] 的方法产生许多粗糙候选物体提议, 并使用支持向量机 (Support vector machine, SVM) 筛选出更为可靠的提议集合, 且进一步使用条件随机场 (Conditional random field, CRF) 进行了优化. VOSA (Video object segmentation aggregation)[18] 是一种集成方法, 该方法首先利用已有的方法对每一帧图像得到一组分割结果, 由于不同的方法具有不同的优势和劣势, 每一组分割结果中都包含较好的和较差的结果. 然后利用所定义的能量函数来优化不同分割结果在最终结果中的权重, 最终得到最优结果. 半监督分割方法利用跟踪[15] 或传播[14] 方式将已知的真值传递到整个视频中. 如文献 [15] 将跟踪与分割置于同一框架下, 将分割任务定义为对于物体部件的跟踪. OFL (Video segmentations via object flow)[14] 则是基于图的分割方法, 在每幅图像内建立图,同时在图像间建立更高层次的图, 并在图上建立能量函数, 通过优化能量函数得到视频分割结果.

得益于近年发展快速的卷积神经网络技术, 很多基于卷积神经网络的分割方法[23−32] 相继提出, 并且超越了大部分传统方法的效果. 无监督分割网络需考虑物体的表观特征和运动特征, 因此文献 [24−25] 提出利用双流网络来进行视频分割. 两支网络的输入分别为视频图像和由光流编码出的 RGB 图像, 以此来进行表观特征的提取和运动特征的提取. 光流分支的加入可以对运动进行分析, 优化最终的结果. 文献 [26] 则是在相邻图像的表观网络顶部加入卷积长短期记忆 (Long short-term menory, LSTM) 模块, 以此编码时域信息, 从提取到的表观特征中寻求运动显著的区域选择为前景. 半监督视频分割

方法[23, 27–28] 则是利用先验信息 (通常是第 1 帧图像的真值) 使其在整列视频上扩展, 得到所有图像的分割预测. 算法 MSK (MaskTrack: Learning video object segmentation from static images)[23] 是在输入光流图像和表观分支之外, 额外输入了当前图像的前一帧 ($T-1$ 帧)的分割结果. 利用上一帧的结果对下一帧进行约束, 并提高下一帧的分割准确度. 算法 OSVOS (One-shot video object segmentation and optical flow)[27] 在测试集中利用第 1 帧的真值微调 (Finetune) 母网络的权重, 使得网络对于该视频中的运动物体敏感, 从而得到准确的视频分割结果. 此外, 文献 [28] 提出基于孪生网络 (Siamese network) 的快速分割方法, 该网络将视频的第 1 帧图像与其真值一起作为参考图像成为孪生网络中一支的输入, 同时将当前图像与前一帧图像的掩膜作为另一支的输入, 实现参考图像对目标的分割引导.

1.2　注意力机制

注意力在人类的感知系统中具有很重要的地位[33–35]. 人类使用视觉感知外界时不会将所有的注意力同时平均分配在视野中的所有位置, 而是将注意力集中于显著的区域, 同时弱化非显著区域的细节, 以更好地构建图像来理解图像的含义.

近年来有很多与注意力相关的研究, 试图将注意力过程应用于卷积神经网络中来提高网络表现力. 文献 [36] 提出了采用残差连接的注意力模块, 同时提出增加更多的注意力模块可以显著提升网络的性能的同时降低计算量. 文献 [36] 同时探讨了空间注意力 (Spatial attention) 和通道注意力 (Channel attention) 及其联合方式对于分类效果的影响, 并用实验证明混合联合方式效果最好. 文献 [37–38] 利用通道注意力模块来选择更具有分辨能力的特征, 使得网络中更有判别力的特征得到加强, 并提高图像分割效果.

2　双流卷积分割网络

本方法实现视频物体分割需要 3 个阶段, 如图 1 所示. 在图 1 中, 左侧为双流分割网络示意图, 右侧为全局信息优化策略, 伪训练集指利用初次分割结果所构建的集合.首先将图像和对应的光流图像输入双流卷积分割网络得到初步分割结果; 进而利用本文所提出的全局信息优化方法, 将上一步得到的分割结果作为先验信息, 监督双流网络中表观分支的微调, 经过训练后, 该网络可学习到视频中前景物体的表观特征; 最后使用新权值下的双流网络处理输入视频, 可得到最终的分割结果.

本节详细解释双流卷积分割网络的结构及其中的卷积注意力模块, 下一节将介绍利用先验信息实现全局信息优化的训练策略.

2.1　卷积注意力模块

给定一个高层特征 $F^h \in \mathbf{R}^{C^h \times H \times W}$ 作为输入, 其相邻的低层特征可表示为 $F^l \in \mathbf{R}^{C^l \times 2H \times 2W}$, 本文的卷积注意力模块可利用高层特征 F^h 推理得到通道注意力张量 $M \in \mathbf{R}^{C^l \times 1 \times 1}$, 其网络结构如图 2 所展示. 完整的注意力推理过程可表示为

$$F' = M(F^h) \otimes F^l \tag{1}$$

每个维度上的注意力值沿着空间的维度扩展, 使注意力强度可与低层特征相乘.

图 2 中 Feature 代表低层特征, 而 Side-prep 层代表由高层特征通过卷积后得到的特征层. 来自最高层的特征直接通过一次核为 3×3 的卷积操作得到特征维度为 16 的 Side-prep 层. 而中间层特征, 首先接受来自上层的注意力加强之后再卷积得到 Side-prep 层 (结构如图 3 所示). Side-prep 层可降低高层特征的维度, 且使其得到更强的深度语义信息提炼.

该 Side-prep 特征首先经过双线性插值操作将其尺寸放大到与低层特征具有相同大小, 使其可与

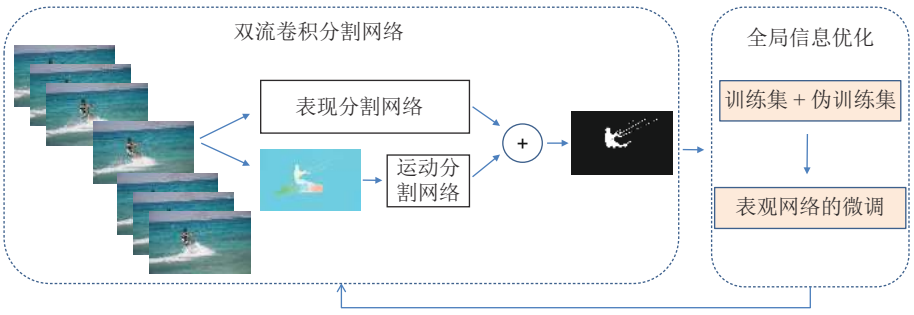


图 1　基于注意力的视频物体分割方法框架图

Fig. 1　The framework of proposed video object segmentation method with attention mechanism

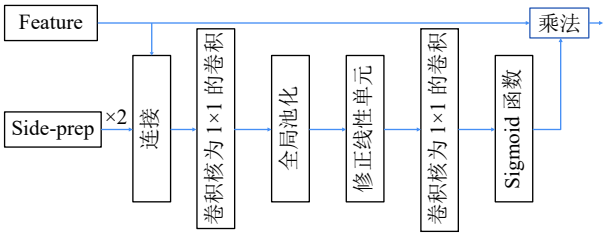


图 2 卷积注意力模块的结构

Fig. 2 The architecture of the convolutional attention module

低层特征融合. 二者以连接方式融合后的特征经由一次核大小为 1×1 卷积操作之后, 进行全局池化. 后续通过激活函数层, 与再一次核大小为 1×1 卷积操作, 最终通过一个 Sigmoid 层得到值为 0 到 1 之间的注意力张量.

在图像分割网络中, 经过多次卷积操作最终将输出一个概率图像, 该图定义了图像中的每个位置的像素成为每个类别的概率[37]. 如式 (2) 所示, 位于最终的概率图中的分数是所有特征图中所有通道特征的和.

$$y_k = F(x; w) = \sum_{i=1}^{D_1} \sum_{j=1}^{D_2} w_{i,j} x_{i,j} \tag{2}$$

在式 (2) 中, x 代表网络中的特征, w 代表卷积核, $k \in 1, 2, \cdots, K$ 代表特征通道的个数, D_1, D_2 分别代表像素在两个维度的位置.

$$\theta_i(y_k) = \frac{\exp(y_k)}{\sum\limits_{j=1}^{K} \exp(y_j)} \tag{3}$$

如式 (3) 所示, θ 代表预测概率, y 表示网络的输出. 最终预测到的像素标签是具有最大概率的类别标签.

如式 (1) 所示, 原来的低层特征 F^l 在注意力张量的影响下改变了它原有的特征通道的权重, 由高层特征所指定的有意义的区域通道被加强, 同时其余通道的权重则被削弱, 实现了高层特征对低层特征的监督.

2.2 视频分割网络

本文的特征提取网络采用了 OSVOS 网络结构. OSVOS 是半监督视频分割方法, 将其特征网络在数据集 DAVIS 的训练集中进行母网络训练. 在测试中, 首先利用测试视频第 1 帧图像的真值对网络参数进行微调, 使网络对该视频中的前景物体敏感, 进而用于视频中其他图像的分割.

本文采用其母网络结构为基础, 且在测试集中并不对网络参数进行微调. 本文所使用特征提取网络如图 3 所示. 可以看到本文的主干网络具有 5 个阶段, 从低层到高层分别具有{2, 2, 3, 3, 3}个卷积层, 在本网络中不特别指出的卷积层所使用的均是大小为 3×3 的卷积核. 在阶段内相邻卷积层间皆具有激活函数层, 每个阶段之间具有池化层. 随着卷积层的层数变多, 网络可以提取到更高层的语义信息, 而低层的特征具有丰富的细节信息.

本网络同时提取第 2 ~ 5 阶段的特征并用于分割, 如图 3 所示. 高层特征通过注意力层后转化为注意力张量, 与相邻低层特征相乘后通过 3×3 卷积成为 Side-prep 层. 来自高层的 Side-prep 特征通过注意力模块产生了新的注意力图, 该注意力张量作用于相邻的低层特征, 实现了高层特征对低层特征的监督, 其结构如图 2.

针对视频中的运动分割问题, 只有在视频图像序列中连续出现并且产生运动的物体才是前景物体. 以城市街景为例, 停在路边的车从表观分析属

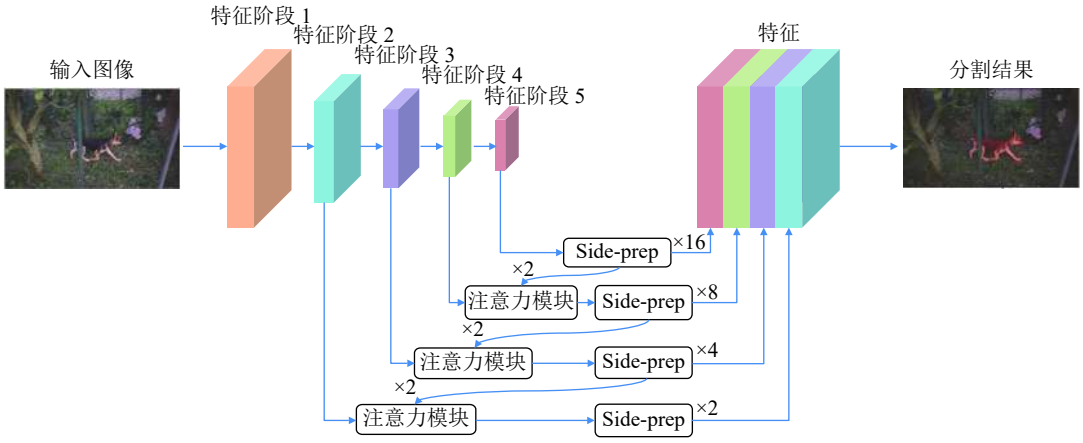


图 3 表观的特征提取网络

Fig. 3 The framework of appearance feature extractor network

于有意义的物体, 然而由于它在视频中没有发生运动, 则运动物体分割任务把该物体识别为背景. 因此只具有表观意义特征不能筛选其中的运动物体, 此时本文加入了运动特征分支网络对视频序列中的运动进行分析.

运动特征网络的输入为利用相邻图像计算得到的光流图像, 且与上段所述的表观网络具有相同的网络结构. 运动分割网络可以大致提取光流图像中产生运动突出的部分, 并将运动显著部分以前景表示, 而运动不显著的部分作为背景. 经过表观和运动两支网络的独立分割将得到两幅分割图像, 最终本文通过将两幅分割图像相加的方式将其融合并得到最后的分割结果.

3　全局信息优化训练策略

在视频序列中, 前景物体的姿态变化、视角变化、遮挡及出现等均会引起前景物体外表形态的变化. 因此在同一视频中某些图像的分割效果较好, 某些图像的分割效果较差. 众所周知, 尽管视频中的物体的表观产生了一定的变化, 而其 RGB 的特征是具有一定规律的, 因此本节提议利用前述步骤所计算得到的前景概率图 (分割结果) 作为先验信息, 对分割结果进行优化.

对分割结果的全局优化需要利用初步分割结果作为先验, 其目的是综合视频时域中的全局信息, 调整表观网络使得前景物体更显著. 经过双流分割网络的处理, 视频中的每幅图像都得到一幅前景概率图. 定义 $X=(x_j, j=1, \cdots, |X|)$ 作为图像中的所有像素, 定义 $Y=(y_j, j=1, \cdots, |X|)$, $y_j \in \{0, 1\}$ 代表所有像素的真值标签, $\Pr(y_j=1|X; W, w^{(m)})$ 表示经过初次分割后每个像素作为前景像素的概率.

定义阈值 α 及 β, $\Pr(y_j=1|X; W, w^{(m)})>\alpha$ 的像素定义为可相信的前景像素 ($y_j=1$), $\Pr(y_j=1|X; W, w^{(m)})<\beta$ 的像素定义为可相信的背景像素 ($y_j=0$. 而 $\beta \le \Pr(y_j=1|X; W, w^{(m)}) \le \alpha$ 被定义为需要忽视的像素 ($y_j=ignored$). 此时测试集图像及其处理后的分割结果构成伪训练集, 可以用于分割网络的微调.

采用比较极端的阈值是由于初步的分割结果中有很多误分割像素, 例如真值为前景的像素被分割为背景像素或反之. 使用错误的标签对网络进行微调会使得网络产生混乱, 因此 “不可靠像素” (前景概率 $\beta \le \Pr(y_j=1|X; W, w^{(m)}) \le \alpha$ 的像素) 需要定义为需要忽略的像素 (Ignored label). 这些像素在训练过程中将不对误差的梯度回传产生任何影响. 图 4 显示了在初步前景概率图中前景像素、背景像素和忽略像素的定义. 图 4 中被浅灰色掩膜所确定的区域内像素为正样本, 被深灰色区域确定的区域内像素为被忽略的样本, 其余未被标注的像素为负样本.

图 4　先验图像中的样本选择

Fig. 4　Our training examples selection

由于初步分割结果并不一定可靠, 因此不能使用该结果单独对表观网络进行训练. 此外, 由于本文在选取图像中正样本时仅考虑其前景概率, 并不考虑空间连续性, 因此会引起分割结果中产生空洞, 前景掩膜难以产生流畅的边缘. 本文选择将该先验作为训练集的补充数据, 混合后进行训练. 这种方法保证了该先验数据可以被网络习得, 同时训练集中的大量可靠数据可以降低不可靠数据对网络权值的影响.

利用初次分割结果对表观网络进行微调之后, 使用本文所提出的双流分割网络对视频进行再次分割, 得到最终的分割结果.

4　训练细节

本文利用像素级别的二分类交叉熵 (cross-entroy-loss) 为损失函数对网络进行训练. 该损失函数定义为

$$L(W) = -\beta\left(\sum_{j\in Y_+}\lg\Pr\left(y_j=1|X; W, w^{(m)}\right)\right) - (1-\beta)\left(\sum_{j\in Y_-}\lg\Pr\left(y_j=0|X; W, w^{(m)}\right)\right) \tag{4}$$

其中, L 表示预测与真值间的损失, W 表示网络中所包含的所有需要学习的参数, $X=(x_j, j=1, \cdots, |X|)$ 表示输入图像中的所有像素, $Y=(y_j, j=1, \cdots, |X|)$, $y_j \in \{0, 1\}$ 表示某次训练中的真值, $\beta=|Y_-|/(|Y_+|+|Y_-|)$ 是用于权衡正样本像素和负样本像素的参数.

计算 Sigmoid 函数 $\theta(\cdot)$ 在图像每个像素 j 上的激励值即为 $\Pr(y_j=1|X; W, w^{(m)})=\theta(a_j^{(m)} \in [0, 1])$.

本文网络的基础 CNN 网络是预先在 ImageNet 上针对图像分类进行训练得到的权值, 此网络是不能直接用于分割的. 因此首先将此网络在 DAVIS

数据集的训练集中进行二分类分割训练. 本文的表观网络以训练集中的图像作为输入, 并配合数据集中的标注做为真值. 当训练运动网络时, 首先利用已有的算法 (FlowNet[39]) 计算视频光流, 之后将光流图像作为运动网络的输入.

使用随机梯度下降 (Stochastic gradient descent, SGD) 方法, 配合动量 (Momentum) 为 0.9, 我们将网络进行了 160 次迭代训练 (Epoch). 其中所使用的数据通过翻转和缩放进行数据扩充. 网络中的学习率为 10^{-8}, 并且随着训练过程逐步减小. 两个分支网络具有相同的训练过程, 结果由分支网络的分割结果通过融合得到.

初次分割结果之后, 需要表观网络进行微调. 在本文中阈值常数定义为 $\alpha = 0.95$, $\beta = 10^{-8}$, 利用此两个阈值为初次得到的概率图中的每个像素给定标签, 使其成为正样本或负样本或被忽略像素. 调整数据后加入原始训练数据中, 在与前一阶段中使用相同超参数的情况下, 我们将网络进行了 5 次迭代训练.

5 实验结果及分析

本算法在 DAVIS[40] 数据集上进行实验来验证方法性能. DAVIS 是 2016 年提出的视频分割数据集, 每个视频中的图像序列均为 480 p 高清分辨率, 并具有精确的像素级别标签. 该数据集共包含 50 个视频, 其中 30 个视频构成训练集, 20 个视频构成测试集. 涵盖遮挡、背景混乱、高速运动等造成的运动物体难以分割的视频.

本文采用 DAVIS 数据集定义的 3 种方式作为定量评价标准: 区域相似度 $\mathcal{J}$、轮廓准确度 $\mathcal{F}$ 和时域稳定性 $\mathcal{T}$. 与交并比 (Intersection over union, IoU) 相似, $\mathcal{J}$ 度量算法得到的分割结果与真值在区域上的匹配度. 定义 M 为算法分割结果, G 为对应的真值, 则 $\mathcal{J}$ 可以表示为 $\mathcal{J} = \dfrac{M \cap G}{M \cup G}$. $\mathcal{F}$ 定义为 $\mathcal{F} = \dfrac{\in \mathcal{P}_\rfloor \mathcal{R}_\rfloor}{\mathcal{P}_\rfloor + \mathcal{R}_\rfloor}$, 其中 P_c 和 R_c 分别代表利用 M 和 G 的轮廓点计算出的精确度 (Precision) 和召回率 (Recall). 时域稳定性 $\mathcal{T}$ 用于评价算法是否会在视频中不同帧产生不稳定的分割结果, 该评价标准是在相邻视频帧之间采用动态时间弯曲 (Dynamic time warping) 计算得到的.

5.1 注意力模块的有效性实验

本文提出了一种新的用于引导低层特征训练的注意力模块, 为了验证该模块的有效性, 本文设置了对比实验, 将本文中的表观网络、运动分割网络和不加入注意力模块的 OSVOS 母网络以及普通的全卷积神经网络 (Fully convolutional network, FCN) 网络分别对 DAVIS 数据集进行分割实验, 得到表 1 的实验结果.

表 1 有效性对比实验

Table 1 Ablation experiments results

	方法	ours_m	ours_a	Baseline	FCN
$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	**0.595**	0.552	0.501	0.519
	Recall $\mathcal{O}\uparrow$	**0.647**	0.645	0.558	0.528
	Decay $\mathcal{D}\downarrow$	0.010	−0.029	**−0.046**	0.059
$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	**0.568**	0.493	0.458	0.482
	Recall $\mathcal{O}\uparrow$	**0.648**	0.487	0.426	0.448
	Decay $\mathcal{D}\downarrow$	0.063	**−0.035**	−0.025	0.054
$\mathcal{T}$	Mean $\mathcal{M}\downarrow$	0.689	0.721	**0.679**	0.829

表 1 中, ours_m 表示本文中所提出的运动分割网络, ours_a 表示文中所使用的表观分割网络. 该网络是在 OSVOS 母网络的基础上, 在相邻阶段的特征层间加入注意力模块形成. Baseline 即为本文所使用的 OSVOS 母网络. FCN 与二者具有相似的特征结构不同的跳跃连接方式.

对比实验表明, 以光流图像为输入的运动分割网络具有最好的分割效果 (0.595), 比另外三种基于图像的分割效果都要好, 这是由于视频图像序列内容复杂, 通常图像中不仅包含运动前景物体, 还包含其他非前景物体的显著物体, 因此仅仅根据表观特征难以确定运动显著的前景物体. 而运动分割网络以光流图像为输入, 可以提取出运动显著的区域, 且光流图像的质量极大影响运动分割的效果.

ours_a 和 OSVOS 母网络的分割结果显示, 加入注意力模块后表观特征网络的分割能力明显提高 (+10.2%). 本网络 (0.552) 的平均区域相似度 $\mathcal{J}$ 比经典的 FCN (0.519) 提高了 (+6.4%). 该实验显示, 注意力模块对于提高分割的准确率非常有效.

5.2 与其他算法的定量对比实验

本节展示了本算法与现有算法在数据集 DAVIS 上的定量比较结果. 参与比较的是本算法 (ours) 与去除全局优化后的本文网络 ours_n, fseg[8], fst[21], msg[41], lmp[42], tis[43], nlc[44], 和 cvos[45]. 表 2 中的数据, 大部分由数据集 DAVIS 的公开网站中所提供. 对于没有提供数值结果的方法, 本文使用原作者所提供的分割结果计算得到.

表 2 显示了本算法与其他算法结果的比较. 其中粗体字数值代表在该评价标准下该算法在所有算法中是效果最好的算法.

表 2　定量实验结果

Table 2　Quantitative experiments results

	方法	ours	ours_n	lmp	msg	fseg	fst	tis	nlc	cvos
$\mathcal{J}$	Mean $\mathcal{M}\uparrow$	**0.713**	0.710	0.700	0.533	0.707	0.558	0.626	0.551	0.482
	Recall $\mathcal{O}\uparrow$	0.798	0.791	**0.850**	0.616	0.835	0.649	0.803	0.558	0.540
	Decay $\mathcal{D}\downarrow$	**−0.036**	−0.007	0.013	0.024	0.015	−0.000	0.071	0.126	0.105
$\mathcal{F}$	Mean $\mathcal{M}\uparrow$	0.684	**0.695**	0.659	0.508	0.653	0.511	0.596	0.523	0.447
	Recall $\mathcal{O}\uparrow$	0.772	**0.809**	0.792	0.600	0.738	0.516	0.745	0.519	0.526
	Decay $\mathcal{D}\downarrow$	**−0.009**	0.004	0.025	0.051	0.018	0.029	0.064	0.114	0.117
$\mathcal{T}$	Mean $\mathcal{M}\downarrow$	0.534	0.589	0.572	0.301	0.328	0.366	0.336	0.425	**0.250**

在最重要的平均区域相似度 $\mathcal{J}$ 上, 本算法得到的分数为 0.713, 比同为双流视频分割的 fseg (0.707) 的评分提高了 0.85%; 比另一种深度学习方法 lmp (0.700)提高了 1.86%.

在轮廓准确度 $\mathcal{F}$ 中, 本算法的评分 (0.684) 也比较好. 比第 2 名的 lmp (0.659) 高 0.025 分, 比第 3 名的 fseg (0.653) 高 0.031 分. 表 2 的 $\mathcal{F}$ 区域显示, 本算法在轮廓处的准确度很高, 同时召回率也比较高. 说明本算法得到的预测分割可以比较准确地寻得视频中运动物体的轮廓.

全局优化后的结果 ours 与未经优化的网络分割结果 ous_n 相比, 平均区域相似度得到了一定程度的提高 (0.003), 然而召回率略有下降 (−0.011). 这是由于使用初次分割的结果作为先验知识网络进行微调时, 仅考虑了像素的前景概率, 没有考虑空间上相邻像素之间的关系, 因此会造成所学习到的网络对前景区域内像素相邻关系的判断减弱.

表 2 显示本网络的时域稳定性较差, 这是由于本算法利用光流信息分析视频间的运动信息, 而光流具有不稳定性, 快速运动、慢速运动、背景和相机的相对运动等都会导致光流图像的误差增大, 最终引起运动分割的效果变差. 由于光流具有短时特征, 其影响只会在视频中某些图像中出现, 造成时域不稳定.

5.3　与其他算法的结果图像对比

图 5 展示了本文算法与几种算法的定性结果对

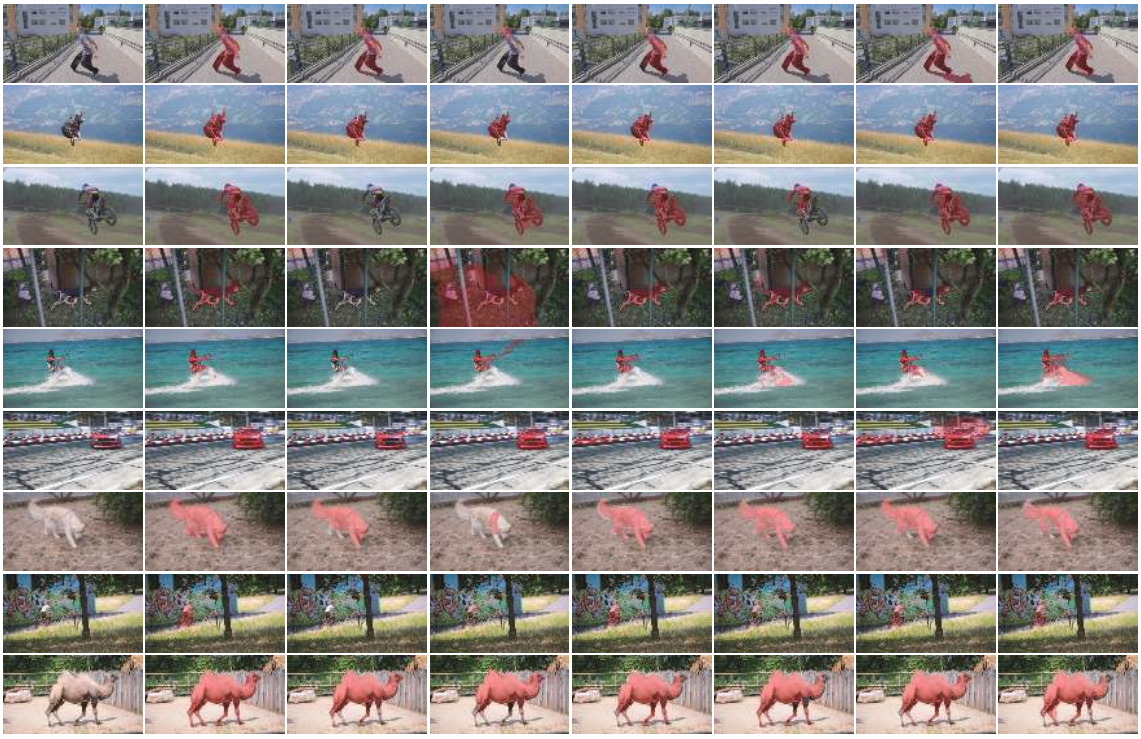


图 5　定性比较结果

Fig. 5　Qualitative results comparison

比. 所有图像均来自 DAVIS 数据集. 图 5 中第 1 列表示输入图像, 第 2 列是本算法的分割结果, 第 3 ~ 8 列是对比方法在输入图像上的分割结果. 第 1 行和第 2 行的分割对象比较清晰, 本算法可以很好地分割得到前景物体, 且不含有影子等噪声. 第 3 ~ 6 行的前景皆具有丰富的细节, 而本文算法都很好地分割到了完整的前景物体, 且具有比较完整的轮廓. 第 7 行和第 9 行是由前景背景高度相似造成的分割困难, 本算法正确提取了物体区域并完整将其分割. 第 8 行中骑自行车的人被复杂的树枝遮挡, 本文算法可以找到其中运动的人, 并且较为细致地在人和树枝密集交错的部分分割出人的区域.

6 结束语

本文提出了一种新的视频物体分割方法. 该方法包含双流视频分割网络和一种全局信息优化方法. 首先利用双流分割网络处理输入视频, 可得到初步的分割结果, 进而利用这些分割掩膜作为监督信息, 对分割网络中的表观分支进行微调, 继而利用新的权值对视频进行分割, 得到最终结果. 本文提出了一种新的注意力模块, 该模块可以利用多层神经网络中的高层特征对低层特征实现注意力区域引导, 提高图像分割的准确度. 同时提出了利用全局信息对原始网络进行微调的方法, 该方法可以比较好地综合视频中所有图像的表观特征, 针对运动前景物体调整表观分割网络, 并提高分割准确度. 在未来的研究工作中, 我们将在全局优化过程中考虑样本空间中像素之间的位置关系, 减轻全局优化中召回率下降的问题.

References

1 Chu Yi-Ping, Zhang Yin, Ye Xiu-Zi, Zhang San-Yuan. Adaptive video segmentation algorithm using hidden conditional random fields. *Acta Automatica Sinica*, 2007, **33**(12): 1252−1258
(褚一平, 张引, 叶修梓, 张三元. 基于隐条件随机场的自适应视频分割算法. 自动化学报, 2007, **33**(12): 1252−1258)

2 Liu Long, Han Chong-Zhao, Liu Ding, Liang Ying-Fu. A new video moving object segmentation algorithm based on Gibbs random field. *Acta Automatica Sinica*, 2007, **33**(6): 608−614
(刘龙, 韩崇昭, 刘丁, 梁盈富. 一种新的基于吉布斯随机场的视频运动对象分割算法. 自动化学报, 2007, **33**(6): 608−614)

3 Rother C, Kolmogorov V, Blake A. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics*, 2004, **23**(3): 309−314

4 Hu Zhi-Lan, Jiang Fan, Wang Gui-Jin, Lin Xing-Gang, Yan Hong. Anomaly detection based on motion direction. *Acta Automatica Sinica*, 2008, **34**(11): 1348−1357
(胡芝兰, 江帆, 王贵锦, 林行刚, 严洪. 基于运动方向的异常行为检测. 自动化学报, 2008, **34**(11): 1348−1357)

5 Lu Zhi-Hong, Guo Dan, Wang Meng. Motion-compensated frame interpolation based on weighted motion estimation and vector segmentation. *Acta Automatica Sinica*, 2015, **41**(5): 1034−1041
(鲁志红, 郭丹, 汪萌. 基于加权运动估计和矢量分割的运动补偿内插算法. 自动化学报, 2015, **41**(5): 1034−1041)

6 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014. 568−576

7 Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 1933−1941

8 Jain S D, Xiong B, Grauman K. FusionSeg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 2117−2126

9 Li X X, Loy C C. Video object segmentation with joint re-identification and attention-aware mask propagation. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 93−110

10 Zhang P P, Liu W, Wang H Y, Lei Y J, Lu H C. Deep gated attention networks for large-scale street-level scene segmentation. *Pattern Recognition*, 2019, **88**: 702−714

11 Zhao H S, Zhang Y, Liu S, Shi J P, Loy C C, Lin D H, et al. PSANet: Point-wise spatial attention network for scene parsing. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 270−286

12 Song C F, Huang Y, Ouyang W L, Wang L. Mask-guided contrastive attention model for person re-identification. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1179−1188

13 Jang W D, Lee C, Kim C S. Primary object segmentation in videos via alternate convex optimization of foreground and background distributions. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 696−704

14 Tsai Y H, Yang M H, Black M J. Video segmentation via object flow. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 3899−3908

15 Wen L Y, Du D W, Lei Z, Li S Z, Yang M H. JOTS: Joint online tracking and segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 2226−2234

16 Xiao F Y, Lee Y J. Track and segment: An iterative unsupervised approach for video object proposals. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 933−942

17 Perazzi F, Wang O, Gross M, Sorkine-Hornung A. Fully connected object proposals for video segmentation. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 3227−3234

18 Zhou T F, Lu Y, Di H J, Zhang J. Video object segmentation aggregation. In: Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME). Seattle, USA: IEEE, 2016. 1−6

19 Fragkiadaki K, Zhang G, Shi J B. Video segmentation by tracing discontinuities in a trajectory embedding. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 1846−1853

20 Wang W G, Shen J B, Yang R G, Porikli F. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(1): 20−33

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

22 Krahenbuhl P, Koltun V. Geodesic object proposals. In: Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. 725−739

23 Perazzi F, Khoreva A, Benenson R, Schiele B, Sorkine-Hornung A. Learning video object segmentation from static images. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 3491−3500

24 Tokmakov P, Alahari K, Schmid C. Learning video object segmentation with visual memory. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 4491−4500

25 Cheng J C, Tsai Y H, Wang S J, Yang M H. SegFlow: Joint learning for video object segmentation and optical flow. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017. 686−695

26 Song H M, Wang W G, Zhao S Y, Shen J B, Lam K M. Pyramid dilated deeper ConvLSTM for video salient object detection. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 744−760

27 Caelles S, Maninis K K, Pont-Tuset J, Leal-Taixe L, Cremers D, Van Gool L. One-shot video object segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 5320−5329

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

29 Cheng J C, Tsai Y H, Hung W C, Wang S J, Yang M H. Fast and accurate online video object segmentation via tracking parts. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7415−7424

30 Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W, Lu H Q. Dual attention network for scene segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 3146−3154

31 Sun T Z, Zhang W, Wang Z J, Ma L, Jie Z Q. Image-level to pixel-wise labeling: From theory to practice. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press, 2018. 928−934

32 Chen L C, Papandreou G, Kokkinos I, Murphy K, Yuille A L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, **40**(4): 834−848

33 Li K P, Wu Z Y, Peng K C, Ernst J, Fu Y. Tell me where to look: Guided attention inference network. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 9215−9223

34 Woo S, Park J, Lee J Y, Kweon I S. CBAM: Convolutional block attention module. In: Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018. 3−19

35 Corbetta M, Shulman G L. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 2002, **3**(3): 201−215

36 Wang F, Jiang M Q, Qian C, Yang S, Li C, Zhang H G, et al. Residual attention network for image classification. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 6450−6458

37 Yu C Q, Wang J B, Peng C, Gao C X, Yu G, Sang N. Learning a discriminative feature network for semantic segmentation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1857−1866

38 Li H C, Xiong P F, An J, Wang L X. Pyramid attention network for semantic segmentation. In: Proceedings of the 2018 British Machine Vision Conference. Newcastle, UK: BMVA Press, 2018. Article No. 285

39 Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, et al. FlowNet: Learning optical flow with convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 2758−2766

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

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

42 Tokmakov P, Alahari K, Schmid C. Learning motion patterns in videos. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE, 2017. 531−539

43 Griffin B, Corso, J. Tukey-inspired video object segmentation. In: Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE, 2019. 1723−1733

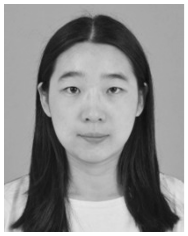
44 Faktor A, Irani M. Video segmentation by non-local consensus voting. In: Proceedings of the 2014 British Machine Vision Conference. Nottingham, UK: BMVA Press, 2014.

45 Taylor B, Karasev V, Soattoc S. Causal video object segmentation from persistence of occlusions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE, 2015. 4268−4276

张 琳 北京理工大学计算机学院博士研究生. 北方电子设备研究所助理研究员. 主要研究方向为视频物体显著性分析与视频分割.
E-mail: zhanglin@bit.edu.cn

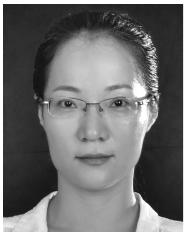
(**ZHANG Lin** Ph.D. candidate at the School of Computer Science and Technology, Beijing Institute of Technology, and assistant research fellow at the Institute of North Electronic Equipment. Her research interest covers video saliency and video segmentation.)

陆 耀 北京理工大学计算机学院教授. 主要研究方向为视觉神经计算, 图像图形处理与视频分析, 模式识别和机器学习. 本文通信作者.
E-mail: vis_yl@bit.edu.cn
(**LU Yao** Professor at the School of Computer Science and Technology, Beijing Institute of Technology. His research interest covers neural network, image processing and video analysis, pattern recognition, and machine learning. Corresponding author of this paper.)



卢丽华 北京理工大学计算机学院博士研究生. 主要研究方向为单人及群体行为识别和视频分割.
E-mail: lulihua@bit.edu.cn
(**LU Li-Hua** Ph.D. candidate at the School of Computer Science and Technology, Beijing Institute of Technology. Her research interest covers collective activity recognition, action recognition, and video segmentation.)

周天飞 北京理工大学计算机学院博士. 主要研究方向为运动物体跟踪, 视频分割及行为识别.
E-mail: ztfei.debug@gmail.com
(**ZHOU Tian-Fei** Ph.D. at the School of Computer Science and Technology, Beijing Institute of Technology. His research interest covers visual tracking, video segmentation, and action recognition.)



史青宣 河北大学网络空间安全与计算机学院副教授. 主要研究方向为计算机视觉, 模式识别, 机器学习.
E-mail: shiqingxuan@bit.edu.cn
(**SHI Qing-Xuan** Associate professor at the School of Cyber Security and Computer, Hebei University. Her research interest covers computer vision, pattern recognition, and machine learning.)