基于并行多方向注意力的无监督视频目标分割

樊佳庆[1)] 苏天康[2)] 张开华[3)] 刘青山[3)]

[1)](南京航空航天大学计算机科学与技术学院 南京 211106)

[2)](南京信息工程大学自动化学院 南京 210044)

[3)](南京信息工程大学计算机学院 南京 210044)

摘 要 时空特征传播对准确的无监督视频目标分割任务至关重要.但是,由于现实中视频的复杂性,导致时空特征学习与传播变得十分具有挑战性.在本文中,提出了两个新颖的模块分别用于增强视频中目标的空间和时间表示.具体来说,首先,针对当前帧,在空间上提出一个新颖的多方向注意力模块,旨在沿着水平、垂直与通道方向上分别提取注意力图.同时,设计了一个并行时序模块用于整合当前帧和之前帧的信息.该模块并行地计算出连续帧之间的二阶相似度,并且根据该相似度图重新对当前帧特征进行加权与增强.此外,该相似度图还直接生成一个有效的掩膜,用于进一步增广当前帧中目标的特征表示.接着,将上述空间和时间特征进行融合以获得最终增广的时空特征表示,并将其输入解码器来预测当前帧中待分割目标的掩膜.在三个主流无监督视频目标分割数据集上的大量实验结果表明,本文提出的方法与当前最新方法相比取得了领先的性能.相关代码将公布在 https://github.com/su1517007879/MP-VOS.

关键词 无监督视频目标分割;多方向注意力;时空调制;并行注意力

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02337

Unsupervised Video Object Segmentation via Parallel Multiple Direction Attention

FAN Jia-Qing[1)] SU Tian-Kang[2)] ZHANG Kai-Hua[3)] LIU Qing-Shan[3)]

[1)](*School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106)

[2)](*School of Automation, Nanjing University of Information Science and Technology, Nanjing* 210044)

[3)](*School of Computer Science, Nanjing University of Information Science and Technology, Nanjing* 210044)

Abstract The propagations of spatio-temporal representations are essential for accurate unsupervised video object segmentation. However, due to the complexity of realistic videos, the learning and propagation of spatio-temporal representations become very challenging. In this paper, both spatial and temporal representations of objects are enhanced by two modules, respectively. Specifically, in the current frame, this paper spatially proposes a novel multiple direction attention module, aiming to extract triple attention maps along the horizontal, vertical, and channel-level directions. Simultaneously, a parallel temporal module is designed to integrate the information from the previous video frame with current representations. It calculates the second-order similarity of the coherent frame pairs in a parallel way, and re-weights the current frame according to the similarity map. Also, the similarity map directly generates a valid mask to augment the representation of current frame. Furthermore, the spatial and temporal features are fused to achieve the augmented spatio-temporal representations and put into the decoder framework to predict the masks of current

收稿日期:2021-10-27;在线发布日期:2022-06-27.本课题得到科技创新 2030-"新一代人工智能"重大项目(2018AAA0100400)、国家自然科学基金项目(U21B2044,61825601,61876088)与江苏省 333 工程人才项目(BRA2020291)资助.**樊佳庆**,博士研究生,主要研究兴趣为视频目标分割.E-mail:jqfan@nuaa.edu.cn.**苏天康**,硕士研究生,主要研究兴趣为视频目标分割.**张开华**,博士,教授,中国计算机学会(CCF)会员,主要研究兴趣为计算机视觉与图像处理.**刘青山**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究兴趣为计算机视觉与模式识别.E-mail:qsliu@nuist.edu.cn.

frame. Extensive experimental results on three mainstream unsupervised VOS datasets show that the proposed model yields favorable performance. The source code is available at https://github.com/su1517007879/MP-VOS.

Keywords unsupervised video object segmentation; multiple direction attention; spatio-temporal modulation; parallel attention

1 引 言

无监督视频目标分割(Unsupervised Video Object Segmentation,UVOS)是指在无任何人机交互干预(例如:在第一帧视频中提供标注)的情况下,自动地在整个视频剪辑中分割出显著的目标区域[1-2]. UVOS在相关工作[3-4]中,有时也被称为零样本视频目标分割(Zero-shot Video Object Segmentation,ZVOS). 为了减少歧义,本文统一记作UVOS. UVOS是一项极具挑战性的任务,算法必须克服物体表观变化与遮挡等问题,并且区分随着时间变化的相似物体. 这种自动分割技术在许多领域产生了重要影响,包括运动分析、视频监控、机器视觉与车辆自动驾驶等.

传统的UVOS算法通常直接使用运动边界[5]、显著性[6]和点的轨迹[7]等来处理视频中的无监督目标分割问题. 虽然这类方法在一些特定视频帧上表现良好,但是在对整个视频序列的分割上面的表现不够稳定. 随着深度神经网络技术的快速发展,近年来的工作主要集中在利用先进的深度神经网络来学习视频中物体的时空特征,以此来提取出充足的目标表观信息[8-9]. 例如,Wang等人[10]证明了人类视觉注意力和视频中主要目标判定之间存在强有力的关联,并且为UVOS设计了动态变化的时空视觉注意力模型,在主流数据集上表现优异. 接着,Lu等人[11]借助时空协同注意力机制耦合全局的时空信息,也获得了优异的结果. Yang等人[12]提出建立当前帧和参考帧之间密集的关联,用于建模UVOS中的长时依赖关系. 该方法在相关标准数据集上表现出有竞争力的性能. 接着,Zhen等人[13]从连续的视频帧中提取出内在的关联,用于增强特征的判别能力,在UVOS任务上取得了领先的结果. 基于上述工作,Zhou等人[14]利用视频帧之间的光流信息,设计了一种运动注意力机制,来获得强大的目标时空特征表示,用于辅助视频中的无监督目标分割,在多个标准数据集上的实验验证了该方法的有效性. 尽管上述UVOS方法在性能方面取得了较大进展,但是仍然存在一些不足,有待进一步探索. 在时域建模方面,一些方法[15-16]使用的光流模型不仅需要借助大量外部数据集来训练,而且在线提取光流信息的计算开销较大. 进一步地,利用记忆单元提取多帧的时序信息[9,17]也会增大计算量并且占用大量显存. 此外,一些基于时空注意力机制的方法[4,11]是对时域和空域信息进行直接的耦合处理,在相邻帧表观有复杂变化时,直接耦合通常很难学习到鲁棒的时空特征. 而在空域建模方面,传统的$W\times H$型二维注意力机制[10]存在结构较复杂、实用性较差等问题.

针对上述问题,本文提出了一种基于并行多方向注意力的UVOS方法,旨在通过提升视频帧之间的特征传播能力,来获取更强大的目标时空特征表示. 具体来说,本文设计了两个新颖的模块来分别增强目标的空间和时间特征表示. 针对连续帧,提出了一个并行时序调制模块用于捕捉当前帧和之前帧的时域连续性,从而使得当前帧具有更稳定的时空表示. 针对当前帧,在空间上提出一个沿着水平、垂直与通道方向上的多方向注意力模块,分别提取注意力图. 本方法和相关方法[14,18]的不同之处在于,本方法不依赖外部数据训练的光流时空调制模型,而直接利用并行时序模型来调制当前帧的特征,此外,还提出了高效的多方向注意力机制来代替传统的二维注意力机制,极大降低了注意力的计算量. 在主流数据集上的实验结果验证了本文方法的有效性.

本文的主要贡献总结如下:

(1)提出了一种并行时序调制模型,来计算连续帧之间的二阶相似度,并且根据该相似度图重新对当前帧进行加权,最终整合当前帧和之前帧的信息,得到鲁棒的目标时空特征表示.

(2)设计了一种多方向注意力模块,用于捕获不同方向上的物体信息,增强当前帧特征的判别能力,最终提升像素级分割效果.

(3)提出的方法与当前热门的UVOS方法相比,在多个标准数据集上取得了较好的结果,验证了本文所提模块的有效性.

2　相关工作

2.1　无监督视频目标分割

传统的 UVOS 方法通常使用运动信息或者显著的线索来推断出前景的物体[3,18]. 最近的方法主要建立在全卷积神经网络上[19-21]，比如，文献[14,22]提出了两个分支的双流网络分别捕捉静态信息和动态信息，在多个数据集上面取得了良好的结果. 接着，Tokmakov 等人[23]利用了视觉记忆体来建模视频中的物体表观变化，获得了较鲁棒的时空特征，在数据集上的实验也验证了该方法的有效性. 进一步地，一些基于注意力机制的工作[10,20]使用了非局部(Non-local)结构进行全局的综合前景推理. Lu 等人[21]设计了一种情境图记忆网络来处理在线的物体表观变化问题，在多个数据集上取得了优异的表现.

2.2　基于时空特征融合的无监督视频目标分割

时空特征融合在视频目标分割任务中起着重要作用. 许多方法通过注意力机制建立情境信息来增强目标特征. 注意力机制在自然语言处理(NLP)[24-25]、语音辨识[26]和计算机视觉[27-28]任务中取得重大成功. 受此启发，Wang 等人[4]验证了视觉注意力行为的高度一致性并利用观察结果深入研究无监督视频目标分割背后的基本原理，在基准测试中取得了最高水平的表现. 接着，Wang 等人[10]提出一种新的排序注意力模块，以端到端的方式学习像素间的时空相似性来获得更好的视频目标分割性能. 在两个主流测试集上均实现了准确性和效率之间的良好平衡. 为了在全局上处理 UVOS 任务，Lu 等人[11]引入了全局协同注意力(Co-attention)机制，用于建模视频相邻帧之间内在的时空相关性. 该协同注意力机制使得分割模型专注于学习有区别的前景特征表示，在相关标准数据集上取得最好的结果. Zhou 等人[14]利用非对称注意力模块对运动信息进行建模，有效提升了对目标时空表观的感知能力，在四个具有挑战性的 UVOS 数据集上取得了当时最好的结果. 在小样本视频目标分割任务上，Chen 等人[29]提出了一种新颖的网络，将满秩注意力模块分解为两个小注意力模块，从而在图像和视频之间搭建了桥梁，有效提升了算法的性能. 与上述方法不同，本文避免了计算复杂的时空相关性，而直接利用简单的并行时序调制模块与多方向注意力模块来同时进行时空特征的增强与融合，在模型效率与分割精度方面都得到了极大提升.

3　本文方法

如图 1 所示，本文方法主要包含并行时序调制模块(分为并行二阶融合与时序表示调制)、多方向注意力模块及融合与预测模块. 经过这三个模块的共同协作，最终输出预测的分割目标掩膜序列. 详细的算法步骤如算法 1 所示.

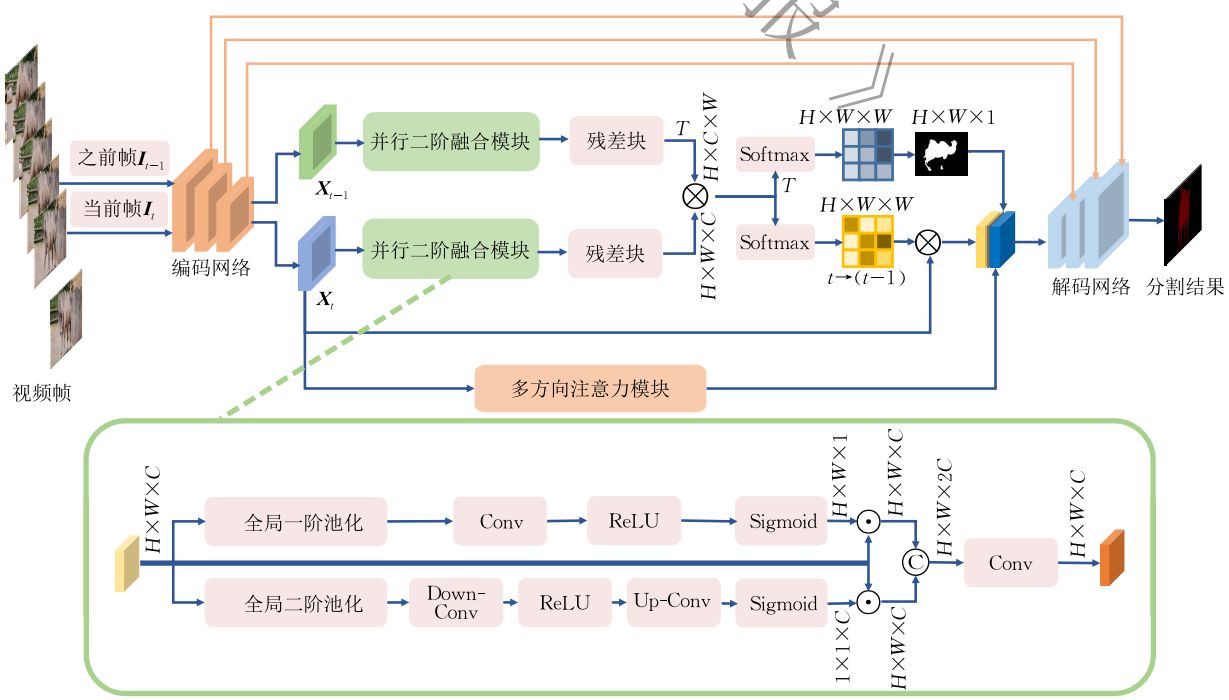


图 1　提出方法的流程图

算法 1. 基于并行多方向注意力的无监督视频目标分割.

输入:视频帧$\{\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_N\}$;标注好的掩膜标签$\{\boldsymbol{G}_1,\boldsymbol{G}_2,\cdots,\boldsymbol{G}_N\}$

输出:预测的掩膜$\{\boldsymbol{P}_1^{\boldsymbol{A}},\boldsymbol{P}_2^{\boldsymbol{A}},\cdots,\boldsymbol{P}_N^{\boldsymbol{A}}\}$

For $t=1$ to N Do

并行二阶融合模块:

1. 输入当前帧$\boldsymbol{I}_t$,提取当前帧特征$\boldsymbol{X}_t$;
2. 根据$\boldsymbol{X}_t$计算得到一阶表示$\boldsymbol{X}_t^1$,根据$\boldsymbol{X}_t$计算得到二阶表示$\boldsymbol{X}_t^2$;
3. 一阶二阶的融合表示$\bar{\boldsymbol{X}}_t\leftarrow\boldsymbol{X}_t^1\oplus\boldsymbol{X}_t^2$;

时序表示的调制:

4. 输入t帧与$t-1$帧的并行二阶融合特征$\bar{\boldsymbol{X}}_t$与$\bar{\boldsymbol{X}}_{t-1}$;
5. 对当前帧$\bar{\boldsymbol{X}}_t$和前一帧$\bar{\boldsymbol{X}}_{t-1}$按顺序用 Batch 级的矩阵乘法得到增强后表示$\boldsymbol{V}$,调换位置后再相乘得到掩膜$\boldsymbol{M}_{(t-1)\to t}$;

多方向注意力模块:

6. 根据当前帧特征$\boldsymbol{X}_t$,在水平、垂直与通道三个方向上分别计算注意力;
7. 把三个方向上的注意力加权到原特征上,得到方向感知特征$\boldsymbol{O}$;

融合与预测模块:

8. 对时序调制后的特征$\boldsymbol{V}$,掩膜$\boldsymbol{M}_{(t-1)\to t}$与方向感知特征$\boldsymbol{O}$进行融合,并预测得到当前帧的掩膜$\boldsymbol{P}_t^{\boldsymbol{A}}$;
9. 根据损失函数计算得到预测的掩膜$\boldsymbol{P}_t^{\boldsymbol{A}}$和真实值$\boldsymbol{G}_t$之间的差异程度,并在训练时最小化该值;
10. 更新$t-1$帧,$\bar{\boldsymbol{X}}_{t-1}\leftarrow\bar{\boldsymbol{X}}_t$;

End For

Return 预测的掩膜$\{\boldsymbol{P}_1^{\boldsymbol{A}},\boldsymbol{P}_2^{\boldsymbol{A}},\cdots,\boldsymbol{P}_N^{\boldsymbol{A}}\}$

3.1 并行时序调制模块

本调制模块是用来挖掘当前帧和之前帧的共有信息以提升当前帧所学特征的鲁棒性.如图 1 所示,输入当前帧(t 帧)和前一帧($t-1$ 帧)的特征图$\boldsymbol{X}_t$,$\boldsymbol{X}_{t-1}\in\mathbb{R}^{H\times W\times C}$,对它们分别进行并行二阶融合,以充分捕获时空上下文信息.

并行二阶融合模块. 如图 1 的下半部分所示,二阶融合模块分为全局一阶分支和全局二阶分支.全局一阶分支首先对当前帧特征$\boldsymbol{X}_t\in\mathbb{R}^{H\times W\times C}$沿通道$C$方向同时进行全局最大池化和全局平均池化,再沿通道方向串联得$\boldsymbol{W}_t\in\mathbb{R}^{H\times W\times 2}$,接着沿通道方向进行 1×1 卷积和 Sigmoid 得到空间权重图$\widetilde{W}_t\in\mathbb{R}^{H\times W\times 1}$,接着与原始特征$\boldsymbol{X}_t\in\mathbb{R}^{H\times W\times C}$进行相乘操作,其中各个空间位置对应相乘,得到空间一阶加权后的表示$\boldsymbol{X}_t^1\in\mathbb{R}^{H\times W\times C}$.类似地,在全局二阶支路方面,首先把输入的$\boldsymbol{X}_t\in\mathbb{R}^{H\times W\times C}$拆分成$C$个特征向量$\boldsymbol{X}_t=[\boldsymbol{x}_t^1,\boldsymbol{x}_t^2,\cdots,\boldsymbol{x}_t^C]$,单个列向量长度是$H\times W$.然后,重新排列成矩阵$\boldsymbol{F}\in\mathbb{R}^{C\times WH}$,得到样本的协方差矩阵:

$$\boldsymbol{\Sigma}=\boldsymbol{F}\bar{\boldsymbol{I}}\boldsymbol{F}^{\top}\tag{1}$$

这里的$\bar{\boldsymbol{I}}=\frac{1}{WH}\left(\boldsymbol{I}-\frac{1}{WH}\boldsymbol{1}\right)$,$\boldsymbol{I}$是$WH\times WH$的单位矩阵,$\boldsymbol{1}$是$WH\times WH$的全 1 矩阵.协方差归一化能显著增强特征表示的判别能力[30],所以本文先对式(1)中的协方差矩阵$\boldsymbol{\Sigma}$进行归一化,并且,它能够被特征值分解如下:

$$\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top}\tag{2}$$

这里的$\boldsymbol{U}$是一个正交矩阵,$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\cdots,\lambda_C)$是一个特征值对角矩阵.接着,协方差归一化特征可以通过如下公式获得:

$$\hat{\boldsymbol{F}}=\sqrt{\boldsymbol{\Sigma}}=\boldsymbol{U}\sqrt{\boldsymbol{\Lambda}}\boldsymbol{U}^{\top}\tag{3}$$

这里的$\sqrt{\boldsymbol{\Lambda}}=\mathrm{diag}(\sqrt{\lambda_1},\sqrt{\lambda_2},\cdots,\sqrt{\lambda_C})$.

得到的归一化后的协方差矩阵描述了通道级的特征关联.令$\hat{\boldsymbol{F}}=[\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_C]$,本文使用全局二阶池化操作得到通道级的描述子$\boldsymbol{z}\in\mathbb{R}^{C\times 1}$,其中第$c$个元素为

$$z_c=Gcp(\boldsymbol{f}_c)=\frac{1}{C}\sum_{i=1}^{C}\boldsymbol{f}_c(i)\tag{4}$$

这里的$Gcp(\cdot)$表示全局协方差池化操作.不同于一阶池化操作,二阶协方差池化利用了特征之间的依赖关系及其本身的数据特性,因而使得学习到的特征比一阶操作更具判别能力[19,30].与 SE-Net[28] 类似,本文在通道级描述子的基础上得到通道级的注意力向量$\boldsymbol{w}\in\mathbb{R}^{C\times 1}$,并且使之重新加权到原特征$\boldsymbol{X}_t\in\mathbb{R}^{H\times W\times C}$上,得到二阶池化后的特征$\boldsymbol{X}_t^2\in\mathbb{R}^{H\times W\times C}$,接着与一阶特征$\boldsymbol{X}_t^1\in\mathbb{R}^{H\times W\times C}$串联到一起,并通过 1×1 卷积得到融合后的表示$\bar{\boldsymbol{X}}_t\in\mathbb{R}^{H\times W\times C}$.

时序表示的调制. 同样地,$(t-1)$帧经过并行注意力模块也能得到融合的特征表示$\bar{\boldsymbol{X}}_{t-1}\in\mathbb{R}^{H\times W\times C}$,本文把$\bar{\boldsymbol{X}}_{t-1}$重新排列成$\bar{\boldsymbol{X}}_{t-1}^{*}\in\mathbb{R}^{H\times C\times W}$,接着对$\bar{\boldsymbol{X}}_t$和$\bar{\boldsymbol{X}}_{t-1}^{*}$进行 Batch 级的矩阵乘法[31]和 Softmax 操作就能得到$t\to(t-1)$的权重$\boldsymbol{W}_{t\to(t-1)}\in\mathbb{R}^{H\times W\times W}$.这个权重和当前帧表示$\boldsymbol{X}_t\in\mathbb{R}^{H\times W\times C}$相乘得到增强后的表示

$$\boldsymbol{V}=\boldsymbol{W}_{t\to(t-1)}\otimes\boldsymbol{X}_t\tag{5}$$

这里的$\otimes$表示 Batch 级的矩阵乘法.

同样地,把t和$(t-1)$帧调换位置,能得到类似的图$\boldsymbol{W}_{(t-1)\to t}\in\mathbb{R}^{H\times W\times W}$,本文直接生成一个掩膜$\boldsymbol{M}_{(t-1)\to t}\in\mathbb{R}^{H\times W\times 1}$,具体方法如下:

$$\boldsymbol{M}_{(t-1)\to t}(i,j)=\begin{cases}1, & \sum\limits_{k\in[1,W]}\boldsymbol{W}_{(t-1)\to t}(i,k,j)>\theta\\ 0, & \text{其他}\end{cases}\tag{6}$$

3.2　多方向注意力模块

多方向压缩操作. 为了保留位置信息，本文使用方向感知注意力机制沿着水平、垂直和通道三个维度编码位置关系，方向注意力感知模块网络结构如图 2 所示. 设当前帧输入为 $\boldsymbol{X}_t \in \mathbb{R}^{H\times W\times C}$，本文首先利用一维池化核$(H,1)$和$(1,W)$分别对水平和垂直方向进行编码，得到水平方向注意力 $a_h^c(h)$：

$$a_h^c(h)=\frac{1}{W}\sum_{k=1}^{W}x_t^c(h,k) \tag{7}$$

其中 $x_t^c(h,k)$指输入 $\boldsymbol{X}_t$的第 c 维通道的第(h,k)个位置的值. 类似地，垂直方向的注意力可以表示为 $a_v^c(v)$：

$$a_v^c(v)=\frac{1}{H}\sum_{l=1}^{H}x_t^c(l,v) \tag{8}$$

其中 v 指 $\boldsymbol{X}_t$的第 c 维通道的第 v 个宽度. 此外，通道方向注意力压缩表示为 a^c：

$$a^c=\frac{1}{H}\sum_{l=1}^{H}\sum_{k=1}^{W}x_t^c(l,k) \tag{9}$$

其中 a^c表示输入 $\boldsymbol{X}_t$的第 c 维通道的全局权重. 以上三个方向感知注意力能够捕获沿着所有方向的远程相关性并保留位置信息，有助于模型能够更准确地定位感兴趣的目标.

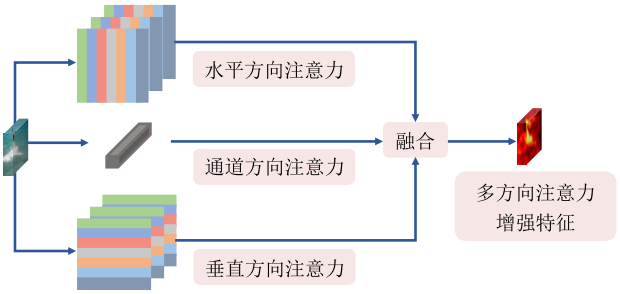


图 2　多方向注意力模块

方向感知激活. 在获得水平注意力图 $\boldsymbol{A}^h \in \mathbb{R}^{C\times H}$(见式(7))和垂直注意力 $\boldsymbol{A}^v \in \mathbb{R}^{C\times W}$(见式(8))后，本文拼接这两个方向得注意力图：

$$S=\rho(D(Cat[\boldsymbol{A}^h,\boldsymbol{A}^v])) \tag{10}$$

其中 $Cat[,]$表示沿着第二分量的拼接，$\rho(\cdot)$表示为非线性激活函数，$D(\cdot)$指沿着第一分量降维比为 r 的降维函数，输出 $\boldsymbol{S}\in\mathbb{R}^{\frac{C}{r}\times(H+W)}$同时在水平和垂直方向对空间信息进行编码. 然后本文将 $\boldsymbol{S}$ 分离成 $\boldsymbol{S}^h \in \mathbb{R}^{\frac{C}{r}\times H}$和 $\boldsymbol{S}^v\in\mathbb{R}^{\frac{C}{r}\times W}$，并且利用 1×1 卷积转化 $D^h(\cdot)$来获得水平注意力权重 $\boldsymbol{P}^h=\sigma(D^h(\boldsymbol{S}^h))$，其中输出为 $\boldsymbol{P}^h\in\mathbb{R}^{W\times H\times C}$，$\sigma(\cdot)$指 Sigmoid 激活函数. 接着将第 c 个通道的第(i,j)个输入特征值重新组合为

$$o_c(k,l)=x_c(k,l)\cdot p_c^h(k)\cdot p_c^v(l) \tag{11}$$

其中加强后的输出表示为 $\boldsymbol{O}_{h,v}\in\mathbb{R}^{W\times H\times C}$，$(\cdot)$表示标量乘法.

为了进一步编码通道方向注意力，本文获得强化特征表示为 $\boldsymbol{O}_c\in\mathbb{R}^{W\times H\times C}$：

$$\boldsymbol{O}_c=\boldsymbol{X}_t\cdot\sigma(\varphi_1(\mathrm{ReLU}(\varphi_2(a)))) \tag{12}$$

其中 $\varphi_1(\cdot)$和 $\varphi_2(\cdot)$为线性变换，$\sigma(\cdot)$指 Sigmoid 激活函数，$(\cdot)$表示通道元素相乘. 最后，本文对 $\boldsymbol{O}_{h,v}$和 $\boldsymbol{O}_c$ 进行元素相加，获得方向感知注意力特征 $\boldsymbol{O}=\boldsymbol{O}_{h,v}\oplus\boldsymbol{O}_c$，其中$\oplus$指元素级相加.

3.3　融合与预测模块

为了融合上述特征表示，本文使用串联 Concatenate 和 1×1 卷积操作，

$$\boldsymbol{A}=Conv_{1\times1}(Cat(\boldsymbol{V},\boldsymbol{M}_{(t-1)\to t},\boldsymbol{O})) \tag{13}$$

其中的 $\boldsymbol{V}$ 与 $\boldsymbol{M}_{(t-1)\to t}$分别由式(5)、(6)计算得到. 接着，把融合后的目标时空表示输入解码网络，预测得到最后的掩膜. 本任务是有监督任务，训练集中每一帧都有对应的真实值标签，在训练时采用特定的损失函数用于监督整个训练. {本文采用交叉熵损失函数，其定义如下：

$$\mathcal{L}_{CE}(\boldsymbol{P}_t^{\boldsymbol{A}},\boldsymbol{G}_t)=-\sum_{i,j}[\log(\boldsymbol{P}_t^{\boldsymbol{A}}(i,j))\boldsymbol{G}_t(i,j)+\log(1-\boldsymbol{P}_t^{\boldsymbol{A}}(i,j))(1-\boldsymbol{G}_t(i,j))] \tag{14}$$

这里的 $\boldsymbol{P}_t^{\boldsymbol{A}}$表示根据式(13)增广后的特征 $\boldsymbol{A}$ 预测得到的掩膜，$\boldsymbol{G}_t$指代标注好的掩膜真实值，(i,j)代表像素点位置，下标 t 表示当前帧. 该损失主要度量预测掩膜和真实值之间的相似度，使得预测值更逼近于真实值.

4　实验评估与分析

4.1　实验设置

本文使用的骨干网络是 ResNet-101，在 ImageNet 数据集上预训练后的模型作为特征提取器. 本文的训练数据由两个部分组成：(1) DAVIS-2016[3] 包含 30 个视频序列，大约有 2000 帧训练数据；(2) Youtube-VOS 数据集[32] 拥有 10 000 多帧的训练数据. 由于 Youtube-VOS 数据集量比较大，标注比较粗糙，而 DAVIS-2016 数据集的量比较小，但是标注精确度非常高. 所以，本文首先在 Youtube-VOS 数据集上进行训练，然后在 DAVIS-16 数据集上进行微调. 为了和其它算法的训练集保持相同规模，本文在 Youtube-VOS 与 DAVIS-2016 上分别选取了 9000 张和 2000 张训练视频帧. 此外，对于原始的训练帧，本文进行了随机翻转(−10°到 10°)和缩放裁剪等数

据增广方法，获取一定量的训练数据. 整个网络采用SGD优化方法来优化，特征提取模块学习率采用1e-4，在融合时采用1e-3的学习率，权重衰减率取1e-5，批次大小设置为5，训练迭代30次，微调迭代20次. 实验在PyTorch1.6.0框架下完成，计算平台配置有一块AMD-Ryzen 3950X的CPU、32 GB内存和单块RTX 2080Ti的GPU.

4.2 测试数据集及评估指标

本文分别在三个无监督视频目标分割测试集上做了评估实验，包括FBMS-59数据集[7]、DAVIS-2016数据集[3]与ViSal测试集[37]. DAVIS-2016数据集包含50个序列，其中的30个序列有完整的标注作为训练集，另外20个序列则作为测试集. 在评估指标方面，根据DAVIS-2016数据集工具箱中的建议，本文采用区域相似度 J，边缘准确率 F 与二者平均值 $J\&F$ 来评价方法的性能. 区域相似度 J 表示模型预测的掩膜和真实掩模之间的交并比 $J=\frac{A\cap B}{A\cup B}$，$A$ 与 B 分别表示预测区域和真实区域. 边缘准确率 F 由边缘精度 P_c 和边缘召回率 R_c 计算得到 $F=\frac{2P_c\times R_c}{P_c+R_c}$.

FBMS-59数据集使用的标注策略是间隔20帧标注一帧的稀疏标注，有59个视频组成，30个测试视频和29个训练视频. 区域相似度用于评估此数据集上的算法. ViSal数据集是有17个测试视频，共193精确标注的视频帧，无训练集. 在ViSal数据集上，本文报告平均绝对误差（MAE）和 F-measure值，具体可参考文献[38].

4.3 定量实验分析

DAVIS-2016数据集上的结果：表1左侧数据展示了提出的方法和最近的UVOS方法的对比结果. 本文提出的MP-VOS在平均 $J\&F$、区域相似度 J 和边缘精度 F 三个指标上的得分分别是83.6%、83.7%和83.4%，均超过了表1中的其它对比方法. 其中，得分第三高的DFNet[12]取得了82.6%、83.4%和81.8%的分数，本文的MP-VOS分别取得了1%、0.3%和1.6%的优势. 而且对比最近的方法FSNet也有0.4%的综合优势. 这些提升是由于本文提出的MP-VOS既考虑了当前帧的多方向的注意力信息，又通过并行时序调制模块，实现了时序线索的增强. 对比只使用光流作为时序信息的MATNet方法，本方法显著地提升了无监督视频目标分割的精度. 此外，在 MAE 和 F_m 指标上，如表2所示，本文提出的MP-VOS也显著高于其它对比方法，该结果也从另外的角度验证了并行时序调制和多方向注意力机制的有效性.

表1 在DAVIS-2016和FBMS数据集上的区域相似度和边缘精度评估结果（下划单线、双线、曲线的加粗字体表示前三名的得分）

Method	DAVIS-2016			FBMS
Metrics	$J\&F$	J	F	J
LMP[17]	68.0	70.0	65.9	—
LVO[23]	74.0	75.9	72.1	—
PDB[33]	75.9	77.0	74.5	74.0
MBNM[34]	79.5	80.4	78.5	73.9
AGS[4]	78.6	79.7	77.4	—
COSNet[11]	80.0	80.5	79.4	75.6
AGNN[20]	79.9	80.7	79.1	—
AnDiff[12]	81.1	81.7	80.5	—
WCS-Net[35]	81.5	82.2	80.7	—
MATNet[14]	81.6	82.4	80.7	**76.1**
DFNet[13]	**82.6**	**83.4**	**81.8**	**82.3**
FSNet[36]	**83.2**	**83.4**	**83.1**	—
MP-VOS(Ours)	**83.6**	**83.7**	**83.4**	**75.9**

FBMS-59数据集上的结果：表1最右侧一列和表2的中间列分别展示了对比方法的区域相似度 J、MAE 和 F_m 这三个结果. 本文方法在区域相似度 J 上取得了75.9%的分数，仅次于DFNet和MATNet. 这是因为DFNet使用了当前序列的所有视频帧作为当前帧目标特征的补充，在视频序列较长的情况下，计算量会大大增加，因此该方法的计算量会变得极大，实用性也会大大降低. 并且，该算法还使用了CRF（条件随机场）策略对分割结果进行后处理，同样削弱了实用性. 类似地，MATNet则应用了光流网络提取帧与帧之间的光流信息，使得该方法的精度严重地依赖于光流模型的准确程度. 另外，该方法无法处理在线的UVOS任务，因为它计算光流时计算的是整个视频的所有光流. 但是，本文的方法不依赖CRF后处理和光流的辅助，在简单有效的框架下就能获得较好的精度.

如表2所示，在 MAE 指标的对比中，MP-VOS以0.059取得了第4，仅次于TENET、MBNM和DFNet. TENET方法在DAVIS-2016数据集上的 MAE 值和 F_m 值都落后于本文的MP-VOS方法，而在FBMS和ViSal数据集上超过了本文的方法，这主要是因为TENET采用了额外的视频显著性检测领域里的训练数据，能够得到比本文方法更充足的图片训练量. 而本文的方法只使用了UVOS的训练集，所以在部分指标的 MAE 上面稍低于TENET方法. 由于MBNM和DFNet有光流提取和CRF后

处理等的辅助，故 MBNM 的 *MAE* 得分稍高于本文的方法. 但是，本文提出的 MP-VOS 依然获得了 84.2%的 F_m 得分，显著高于 MBNM、DFNet 的 81.6% 和 83.3%，这充分验证了本文所提方法 MP-VOS 是简单有效的.

表 2　在 DAVIS-2016、FBMS 与 ViSal 数据集上的平均绝对误差(*MAE*)和 *F*-measure 值的对比结果(下划单线、双线、曲线的加粗字体表示前三名的得分)

Method	DAVIS-2016		FBMS		ViSal	
Metrics	*MAE*	F_m	*MAE*	F_m	*MAE*	F_m
FCNS[8]	0.053	72.9	0.100	73.5	0.041	87.7
FGRNE[9]	0.043	78.6	0.083	77.9	0.040	85.0
TENET[39]	**0.019**	**90.4**	**0.026**	**89.7**	**0.014**	**94.9**
MBNM[34]	0.031	86.2	**0.047**	81.6	0.047	—
PDB[33]	0.030	84.9	0.069	81.5	0.022	91.7
AnDiff[12]	0.044	80.8	0.064	81.2	0.030	90.4
DFNet[13]	**0.018**	**89.9**	**0.054**	**83.3**	**0.017**	**92.7**
MP-VOS (Ours)	**0.014**	**92.4**	0.059	**84.2**	**0.019**	**92.1**

ViSal 数据集上的结果：与 FBMS 的结果类似，本文提出的 MP-VOS 以 0.019(*MAE*)与 92.1%(*F*-measure)的得分稍落后于 TENET 与 DFNet. 和 FBMS 数据集中的情况类似，在光流估计和 CRF 后处理的辅助下，TENET 与 DFNet 以 0.005 和 0.002 的微弱优势领先于本文提出的算法. 但是，本文方法 MP-VOS 在不借助复杂的前后处理的情况下仍然取得了 0.019 和 92.1 的 *MAE* 和 F_m 值，并且，在更主流的数据集 DAVIS-2016 上的 *MAE* 和 F_m 也值都显著高于 TENET 与 DFNet，这些都充分说明了本文所提方法具有较强的性能. 综上所述，本文提出的基于并行多方向注意力的无监督视频目标分割方法(MP-VOS)在 DAVIS-2016、FBMS-59 与 ViSal 数据集上都取得了较好的实验结果. 在舍弃复杂后处理的情况下，本文提出的方法在对比主流 UVOS 方法时，依然拥有良好的竞争力.

4.4　消融实验分析

表 3 列出了本文在 DAVIS-2016 数据集上的消融实验结果. 在去除并行时序调制策略之后，本文的 J 和 F 得分为 80.6%和 78.4%，J 和 F 分别下降了 3.1%和 5.0%，该结果显示出并行时序调制策略在时空特征传播中的积极作用. 为了验证并行时序调制模块中的一阶支路的重要性，本文进一步去除了一阶分支，可以从表 3 中看到结果从 83.6%下降到了 83.4%，说明了一阶分支对整个并行分支起到积极的作用. 此外，本文在去除了多方向注意力机制后，得到 80.0%，比原始版本下降了 3.6%，说明了多方向注意力机制对于 UVOS 也至关重要. 为了进一步分析多方向注意力机制，本消融实验分别单独去除了三个方向上的注意力. 得到的实验结果显示，当去除横向注意力时，性能下降最多，说明在这个三个方向的注意力中，横向注意力的作用最明显. 接着，本文把骨干网络从 ResNet101 替换成 ResNet50，速度从 36.2 fps 提升到了 56.3 fps，但是性能却大大降低了，变为 78.0%，这从侧面说明了骨干网络对 UVOS 性能的重要影响，骨干网络越深效果越好. 此外，相比于基准方法[14]，本文方法在 J 和 F 上分别有 1.3%与 2.7%的较大提升，说明了所提算法的整体性能优势. 综上所述，本文提出的并行时序调制. 和多方向注意力模块对整体性能都有积极的作用，而且，并行时序调制策略比多方向注意力模块对整个 UVOS 系统的影响更大.

表 3　在 DAVIS-2016 数据集上的消融实验

	平均 $J\&F$	J	F
去除并行时序调制	79.5	80.6	78.4
去除并行一阶分支	83.4	83.5	83.2
去除多方向注意力	80.0	81.0	79.0
去除横向注意力	78.9	80.1	77.8
去除纵向注意力	79.1	80.3	77.9
去除通道方向注意力	79.5	80.7	78.2
骨架网络替换成 ResNet50(56.3 fps)	78.0	79.0	77.0
基准方法(Baseline)	81.6	82.4	80.7
原始版本(36.2 fps)	83.6	83.7	83.4

4.5　定性实验分析

图 3 展示了本文算法的定性实验的结果. 从上到下的视频序列依次是 DAVIS2016 数据集(Blacksman、Dog、Drift-straight 和 Parkour)、Youtube-objects 测试集(Cat_0003、Dog_0005 和 Motorbike_0009)与 Visal 数据集(Car、Panda 和 Snow_leopards). 如图 3 所示，本文提出的方法可以较好应对多种实际场景下面临的问题，包括快速运动场景(如 Blacksman、Parkour 与 Motorbike_0009 序列)、背景杂乱场景(Dog 与 Panda 序列)、剧烈形变场景(Drift-straight、Car 与 Dog_0005 序列)、部分遮挡场景(Snow_leopards 与 Cat_0003 序列). 例如，在第四行的 Parkour 序列中，即使特技运动员飞快奔跑，本文提出的并行时序调制依然能有效捕捉鲁棒的物体特征表示，从而获得准确的视频目标分割结果. 在第六行的 Dog_0005 序列中，小狗在镜头中不断移动，产生了各种形变，甚至身体大部分已经离开镜头的视野，但本文提出的 MP-VOS 方法通过横向和纵向的注意力机制显著地增强了运动物体的表示，使得模型应对表观变

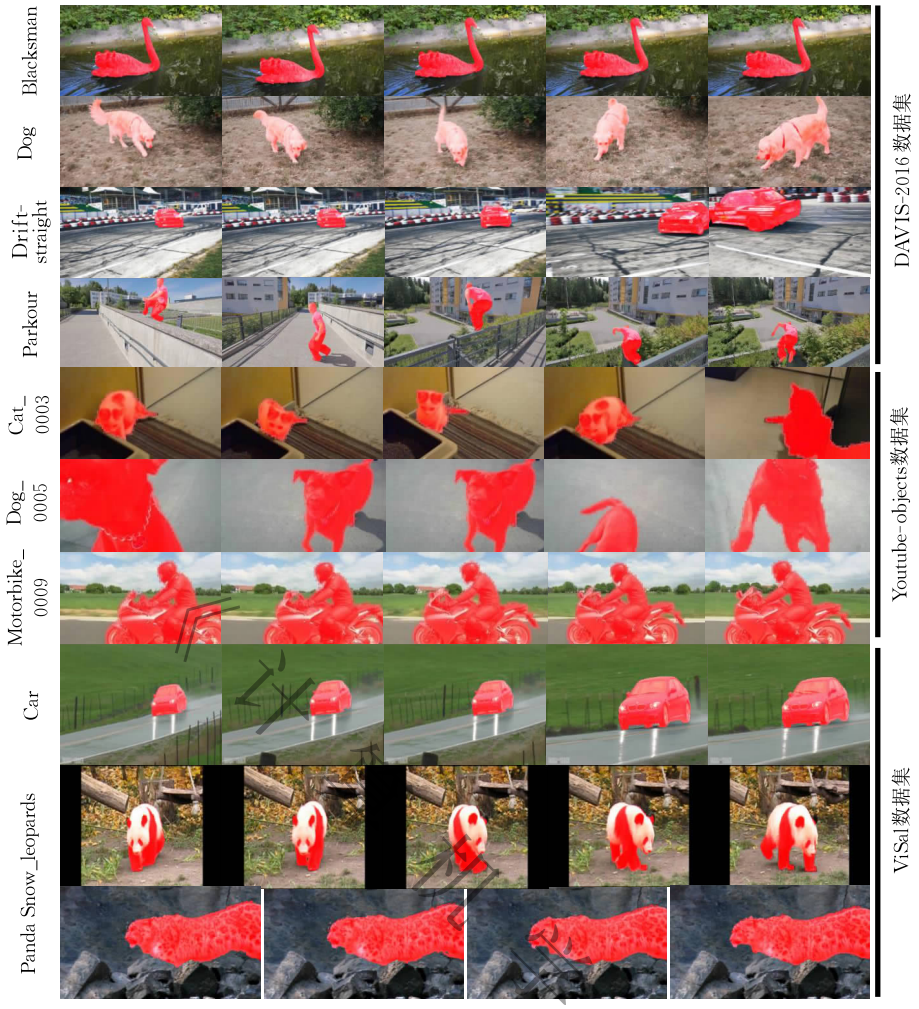


图 3 定性实验结果

化的能力增强，最终所提的算法依然可以完整地分割出理想的小狗轮廓. 以上的定性实验充分验证了本文方法在特定场景中的有效性.

4.6 损失曲线分析

图 4 展示了训练过程中损失函数的演化图，从图中可以看出总体训练过程是较平稳的. 具体地，在训练初始阶段损失值下降幅度较大，说明选取的学习率是合适的. 在训练到一定阶段之后，损失值的下降趋向于平稳，但仍在持续下降，说明模型选取的损失函数能正常收敛. 此外，图中的曲线波动与 Batch size 的设置相关，本文选取的 Batch size 比较合适，所以训练过程中的曲线波动较小.

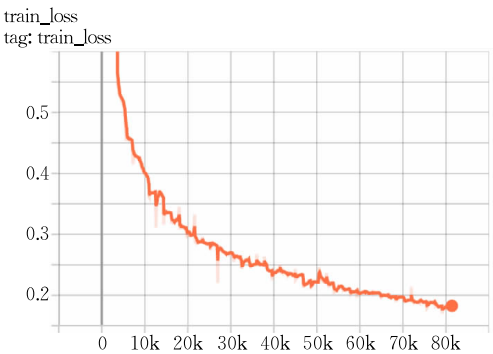


图 4 训练过程中的损失变化图

5 结 论

本文提出了一种基于多方向注意力和并行集成表示的高效无监督视频目标分割方法. 首先，本文设计了一种并行时序调制机制，来计算连续帧之间的相似度，并根据该相似度重新精炼当前帧的特征表示. 接着，将精炼后的当前帧表示与之前帧的信息整合到一起，得到鲁棒的目标时空表示. 此外，本文又进一步提出了一种多方向注意力策略，用于捕捉不同方向上的物体显著信息，增强当前帧目标表示的判别能力. 最后，本文在三个无监督视频目标分割数据集上进行了实验评估，得到的实验结果充分验证了本文提出方法的有效性.

致　谢　感谢指导老师的细心指导，感谢评审专家的审稿意见和提出的专业性建议！

参　考　文　献

[1] 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)

[2] Huang Shu-Ying, Hu Wei, Yang Yong, et al. A low-exposure image enhancement based on progressive dual network model. Chinese Journal of Computers, 2021, 44(2): 384-394(in Chinese)
(黄淑英，胡威，杨勇等. 基于渐进式双网络模型的低曝光图像增强方法. 计算机学报，2021，44(2)：384-394)

[3] 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

[4] Wang W, Song H, Zhao S, et al. Learning unsupervised video object segmentation through visual attention//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3064-3074

[5] Papazoglou A, Ferrari V. Fast object segmentation in unconstrained video//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia, 2013: 1777-1784

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

[7] Ochs P, Malik J, Brox T. Segmentation of moving objects by long term video analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(6): 1187-1200

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

[9] Li G, Xie Y, Wei T, et al. Flow guided recurrent neural encoder for video salient object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 3243-3252

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

[11] Lu X, Wang W, Ma C, et al. See more, know more: Unsupervised video object segmentation with co-attention Siamese networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3623-3632

[12] Yang Z, Wang Q, Bertinetto L, et al. Anchor diffusion for unsupervised video object segmentation//Proceedings of the 2019 IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 931-940

[13] Zhen M, Li S, Zhou L, et al. Learning discriminative feature with CRF for unsupervised video object segmentation//Proceedings of the European Conference on Computer Vision. Cham, Swiss: Springer, 2020: 445-462

[14] Zhou T, Li J, Wang S, et al. MATNet: Motion-attentive transition network for zero-shot video object segmentation. IEEE Transactions on Image Processing, 2020, 29: 8326-8338

[15] Yang S, Zhang L, Qi J, et al. Learning motion-appearance co-attention for zero-shot video object segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 1564-1573

[16] Yang C, Lamdouar H, Lu E, et al. Self-supervised video object segmentation by motion grouping//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 7177-7188

[17] Tokmakov P, Alahari K, Schmid C. Learning motion patterns in videos//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 3386-3394

[18] Keuper M, Andres B, Brox T. Motion trajectory segmentation via minimum cost multicuts//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3271-3279

[19] Dai T, Cai J, Zhang Y, et al. Second-order attention network for single image super-resolution//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 11065-11074

[20] Wang W, Lu X, Shen J, et al. Zero-shot video object segmentation via attentive graph neural networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 9236-9245

[21] Lu X, Wang W, Danelljan M, et al. Video object segmentation with episodic graph memory networks//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 661-679

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

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in Neural Information Processing Systems. Long beach, USA, 2017: 5998-6008

[25] Ramachandran P, Parmar N, Vaswani A, et al. Stand-alone self-attention in vision models. arXiv preprint arXiv: 1906.05909, 2019

[26] Chan W, Jaitly N, Le Q, Vinyals O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China, 2016: 4960-4964

[27] Wang X, Girshick R, Gupta A, He K. Non-local neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7794-7803

[28] Hu J, Shen L, Sun G. Squeeze-and-excitation networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132-7141

[29] Chen H, Wu H, Zhao N, et al. Delving deep into many-to-many attention for few-shot video object segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14040-14049

[30] Li P, Xie J, Wang Q, et al. Is second-order information helpful for large-scale visual recognition//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2070-2078

[31] Wang L, Wang Y, Liang Z, et al. Learning parallax attention for stereo image super-resolution//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 12250-12259

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

[33] Song H, Wang W, Zhao S, et al. Pyramid dilated deeper ConvLSTM for video salient object detection//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 715-731

[34] Li S, Seybold B, Vorobyov A, et al. Unsupervised video object segmentation with motion-based bilateral networks//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 207-223

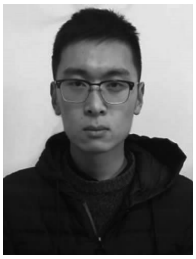
[35] Zhang L, Zhang J, Lin Z, et al. Unsupervised video object segmentation with joint hotspot tracking//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 490-506

[36] Ji G, Fu K, Wu Z, et al. Full-duplex strategy for video object segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 4922-4933

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

[38] Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 3203-3212

[39] Ren S, Han C, Yang X, et al. TENet: Triple excitation network for video salient object detection//Proceedings of the European Conference on Computer Vision. Edinburgh, Scotland, 2020: 212-228



FAN Jia-Qing, Ph. D. candidate. His research interests focus on video object segmentation.

SU Tian-Kang, M. S. candidate. His research interests focus on video object segmentation.

ZHANG Kai-Hua, Ph. D., professor. His research interests include computer vision and image processing.

LIU Qing-Shan, Ph. D., professor. His research interests include computer vision and pattern recognition.

Background

Video object segmentation refers to highlighting specific target objects in a given video. This paper focuses on unsupervised video object segmentation (UVOS), also known as zero-shot video object segmentation (ZVOS). The UVOS model can automatically segment the target as accurately as possible in all other frames without the manual annotation in the first frame, and at the same time, it is best to achieve real-time processing and less memory usage.

Since temporal information must be transmitted from front to back among video frames, how to effectively mine and transmit the spatio-temporal features of video frames has become a research hotspot in UVOS. Current mainstream UVOS methods use dual-stream network, memory bank, spatio-temporal attention mechanism or 3D convolution to depict spatio-temporal correlations among video sequences. Although previous UVOS methods have made great progress in performance, there are still some shortcomings, such as: (1) some optical flow models used by UVOS methods need to

be trained with a large number of external data sets, and this borrowed optical flow model has a great impact on the accuracy of the final video segmentation; (2) The computation cost of online optical flow model is high, which affects the real-time performance of online UVOS; (3) Extracting temporal information of multiple frames by memory unit will occupy a lot of computing resources and memory usage; (4) Some methods based on spatial and temporal attention mechanism can not fully capture temporal information due to the interference between spatial and temporal cues; (5) The traditional W∗H two-dimensional attention mechanism has some problems such as complicated structure and large computation.

To relieve the above problems, this paper proposes a UVOS method based on parallel multi-directional attention, which aims to obtain a more powerful spatio-temporal representation of target by improving the feature propagation capability between video frames. Specifically, two novel modules are designed to enhance the spatial and temporal representation of objects respectively. For coherent frames, a parallel temporal modulation module is proposed to capture the temporal continuity of the current frame and the previous frame, so that the current frame has a more stable spatio-temporal representation. For the current frame, a multi-direction attention module along the horizontal, vertical and channel directions is designed to extract the attention diagram respectively. The difference between the proposed method and other methods is that the training of our method is not dependent on external optical flow model, and the direct use of parallel time-series model to complement the current frame. In addition, it also puts forward more efficient multi-directional attention mechanism instead of the traditional two-dimensional attention mechanism, greatly reducing the computation of attention. Experimental results on mainstream data sets verify the effectiveness of the proposed method.

This work is supported by the National Key Research and Development Program of China under Grant No. 2018AAA-0100400, in part by the National Natural Science Foundation of China under Grant Nos. U21B2044, 61825601, 61876088, in part by the 333 High-level Talents Cultivation Project of Jiangsu Province (BRA2020291).