计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):317−328 [doi: 10.15888/j.cnki.csa.009027]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

时空渐进式学习的视频显著性目标检测①

王星驰[1], 李军侠[2]

[1](南京信息工程大学 自动化学院, 南京 210044)
[2](南京信息工程大学 计算机与软件学院, 南京 210044)
通信作者: 李军侠, E-mail: jxli@nuist.edu.cn

摘　要: 视频显著性目标检测需要同时结合空间信息和时间信息, 连续地定位视频序列中与运动相关的显著性目标, 其核心问题在于如何高效地刻画运动目标的时空特征. 现有的视频显著性目标检测算法大多使用光流, ConvLSTM 以及 3D 卷积等提取时域特征, 缺乏对时间信息的连续学习能力. 为此, 设计了一种鲁棒的时空渐进式学习网络 (spatial-temporal progressive learning network, STPLNet), 以完成对视频序列中显著性目标的高效定位. 在空间域中使用一种 U 型结构对各视频帧进行编码解码, 在时间域中通过学习视频序列中帧间运动目标的主体部分和形变区域特征, 渐进地对运动目标特征进行编码, 能够捕捉到目标的时间相关性特征和运动趋向性. 在 4 个公开数据集上与 13 个主流的视频显著性目标检测算法进行一系列对比实验, 所提出的模型在多个指标 (max$F$, S-measure (S), MAE) 上达到了最优结果, 同时在运行速度上具有较好的实时性.

关键词: 视频显著性目标检测; 深度学习; 空间信息; 时间信息; 静态特征挖掘; 运动特征渐进学习

引用格式: 王星驰,李军侠.时空渐进式学习的视频显著性目标检测.计算机系统应用,2023,32(4):317–328. http://www.c-s-a.org.cn/1003-3254/9027.html

Spatial-temporal Progressive Learning for Video Salient Object Detection

WANG Xing-Chi[1], LI Jun-Xia[2]

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

Abstract: Video salient object detection (VSOD) can continuously locate motion-related salient objects in video sequences by combining spatial and temporal information. Its core lies in how to efficiently describe the spatial and temporal features of moving objects. Existing VSOD algorithms mainly use optical flow, ConvLSTM, and 3D convolution to extract time domain features, but their continuous learning ability of temporal information is insufficient. Therefore, a robust spatial-temporal progressive learning network (STPLNet) is proposed to realize the efficient localization of salient objects in the video sequences. In the spatial domain, the method uses a U-shaped structure to encode and decode each video frame. In the temporal domain, it progressively encodes the features of the moving objects by learning the features of subject parts and deformation regions about the moving objects between frames in the video sequences. In this way, the method can capture the time correlation features and motion tendency of the objects. A series of comparative experiments are carried out on four public datasets, with 13 mainstream VSOD algorithms involved. The proposed model achieves optimal results on multiple indicators including max$F$, S-measure (S), and MAE, and it has excellent real-time performance in running speed.

Key words: video salient object detection (VSOD); deep learning; spatial information; temporal information; static feature mining; motion feature progressive learning

① 基金项目: 科技创新 2030-“新一代人工智能”重大项目 (2018AAA0100400)
收稿时间: 2022-08-24; 修改时间: 2022-09-27; 采用时间: 2022-10-19; csa 在线出版时间: 2022-12-23
CNKI 网络首发时间: 2022-12-27

视频显著性目标检测旨在从动态视频序列中提取出最引人注意力的运动目标, 近年来引起了研究人员极大的兴趣. 作为智能视频处理中一项关键支撑技术, 为目标检测[1]、目标追踪[2]、场景理解[3] 等高层任务提供可靠数据, 视频显著性目标检测已被广泛应用于智能安防、航空航天、人机交互、自动驾驶、医疗诊断等各个领域. 视频显著性目标检测任务面临的主要挑战来自于背景干扰、光照变化、遮挡、尺度变化、运动模糊、目标变形等各种因素对运动目标的干扰, 以及人类在动态场景中视觉注意行为固有的复杂性.

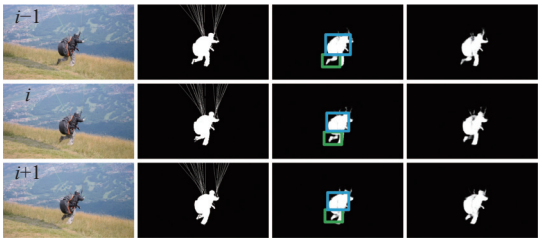
不同于仅考虑空间域的静态图像显著性目标检测, 在视频序列中目标的运动是吸引人类视觉注意力的关键要素, 因此, 视频显著性目标检测任务的核心问题在于如何高效地刻画运动目标的时空特征. 传统视频显著性目标检测算法主要依赖于手工设计的特征和先验知识 (颜色对比度、背景先验和形态学线索等), 以及来自光流估计[4] 的帧间运动信息. 受限于手工特征的表达能力, 传统算法的检测准确性较低, 同时光流估计的使用导致其检测速度较慢.

基于深度学习的方法是目前的主流算法, 它们通常使用光流估计, ConvLSTM[5] 以及 3D 卷积[6] 等逐步地对目标的运动信息进行编码, 进而提取时间特征. 例如, 文献 [7] 设计了一个双支路的结构以结合视频帧间信息和光流估计. 文献 [8] 对传统的 ConvLSTM 进行了改进, 在编码的过程中同时关注于选择性注意和注意力转移这两个时间动态特性. 文献 [9] 对多帧信息进行堆叠, 随后使用 3D 卷积获得时空信息. 由于卷积神经网络强大的特征学习及表达能力, 基于深度学习的视频显著性目标检测算法可以取得较好的检测结果, 但同时也存在一些不足之处. 基于光流的算法通过刻画运动目标在相邻帧之间的运动性获取时域信息, 对长时域信息的捕捉能力有限, 且计算复杂度较高难以获得实时速度. 3D 卷积在提取帧间时空信息时提取到的特征鲁棒性不强. 基于 ConvLSTM 的算法拥有较多的神经元, 在处理长序列时计算效率较低.

针对上述问题, 本文设计了一个鲁棒性较强的时空渐近式学习网络 (spatial-temporal progressive learning network, STPLNet) 来学习视频序列中的时空信息以高效且完整地检测出运动目标. 该网络由两个模块组成, 静态特征挖掘 (static feature mining, SFM) 和运动特征渐近学习 (motion feature progressive learning, MFPL). 其中 SFM 用于挖掘视频帧的静态空间特征, 而 MFPL 使用连续帧间指导的方式来渐进式地学习运动目标的时间特征. SFM 使用一种 U 型结构对输入序列进行下采样编码和上采样解码以挖掘其空间特征.在下采样编码时, 为了保证能够从输入序列中提取足够的特征而设置了多个残差块. 而在上采样解码时, 高层次特征充分融合低层次特征, 保证了信息的丰富度和完整度. MFPL 关注于运动目标时间特征的学习, 重点刻画输入序列中目标的运动特性并进一步预测其运动趋势. 通过设计运动主体特征提取 (motion subject feature extractor, MSFE) 和优化特征提取 (refinement feature extractor, RFE) 这两个模块对连续帧间目标的运动特征进行渐进式学习. MSFE 首先对视频序列中上一帧运动目标的主体部分进行特征提取, 将获得的特征融合到当前帧完成对当前时刻目标主体部分的加强. RFE 利用连续帧间上下文关系对目标的运动性进行学习, 使用上一帧的信息来引导当前帧目标进行形变特征的学习, 使其具有识别运动目标形变区域的能力且能够预测目标运动的趋向性. 随后进一步加强运动目标主体部分特征并对背景噪声进行抑制. 如图 1 中所示, 运动目标经过完整的 MSFE 和 RFE 学习后, 其主体部分 (图中较大框) 和形变区域 (图中较小框) 更好地被识别出且具有清晰的细节信息. 本文的贡献点总结如下.

(1) 提出了一种新颖的视频显著性目标检测算法, 对于静态空间特征和运动时间特征具有较强的特征表达能力, 实现了较好的显著性目标检测结果.

(2) 设计了运动特征渐近学习模块, 从运动目标的主体部分和形变区域入手, 以帧间指导的方式渐进式地学习时间特征, 其对目标的运动信息具有稳健的刻画能力.



(a) 序列 (b) 标签 (c) MSFE+RFE (d) MSFE

图 1 各模块有效性可视化对比

1 相关工作

本部分从图像显著性目标检测和视频显著性目标检测两方面介绍与本文相关的算法, 并进行相应分析.

1.1 图像显著性目标检测

传统的图像显著性目标检测算法大多基于手工设计的特征同时利用先验知识对显著目标进行检测. 文献 [10] 使用一种基于背景感知的显著性目标检测方法. 随着深度学习的发展, 基于卷积神经网络的图像显著性目标检测算法成为主流, 它们大多使用全卷积神经网络[11] 对输入进行像素级预测. 文献 [12] 对神经网络中的池化部分进行扩展, 基于 U 型结构识别出不同层潜在显著性目标的位置信息, 同时为具有高级语义信息的深层特征和含有丰富细节信息的浅层特征设计了一种更好的融合方式. 文献 [13] 针对输入显著性区域突出不均匀和边缘不清晰的问题, 使用了一种通道空间联合注意力机制来进行显著性目标检测. 文献 [14] 使用一种基于全局特征引导的显著性模型. 文献 [15] 提出一个新的单阶段模型, 从不同分辨率的图像中提取特征进行学习. 此外, 该算法还提出了一个高分辨率显著性目标检测数据集. 文献 [16] 提出了一个渐进式双注意力残差模型, 通过设计互补注意力特征图来指导模型进行残差学习, 对显著性预测图使用了一种渐进式地由粗到细的优化方式.

1.2 视频显著性目标检测

不同于图像显著性目标检测算法其仅需考虑静态空间特征, 视频显著性目标检测算法需要同时考虑输入序列的静态空间特征和帧间运动目标时间特征, 并对目标的运动信息进行重点刻画, 能够快速且准确地从输入序列中检测到显著性物体, 该任务更加具有挑战性. 近年来, 基于深度学习方法的算法有着较好的表现. 文献 [17] 设计了一种金字塔扩张双向 ConvLSTM 递归结构, 该模型处理特征时采用前向和后向 ConvLSTM 提取多尺度时空信息. 由于视频数据像素级标注获取困难, 文献 [18] 结合光流估计对数据集中部分未标记帧生成伪标签, 并使用 NonLocal[19] 和 ConvGRU[20] 进一步对视频帧间的时空相关性进行加强. 该学习策略仅需少量的被标记帧即可产生准确的显著性检测结果. 文献 [21] 将一个轻量的时间模型集成进空间支路, 粗略地定位与确信度高的显著运动相关的空间显著性区域, 同时空间支路本身能够以一种多尺度的方式反复优化时间模型. 文献 [22] 使用提取的边缘信息引导对时空特征的同步提取, 将深层纹理信息和浅层边缘信息进行结合. 文献 [23] 使用注意力模块在视频序列中学习对比特征, 无需高计算量的时间建模技术. 文献 [24] 基于视频显著性目标检测和图像显著性目标检测之间的相似性和差异性, 提出一个新颖的渐进式框架来定位和分割序列中的显著性目标区域而不使用任何视频标注信息.

本文对目标的运动特征进行渐进式学习, 可以较为准确地预测视频序列中的显著性目标.

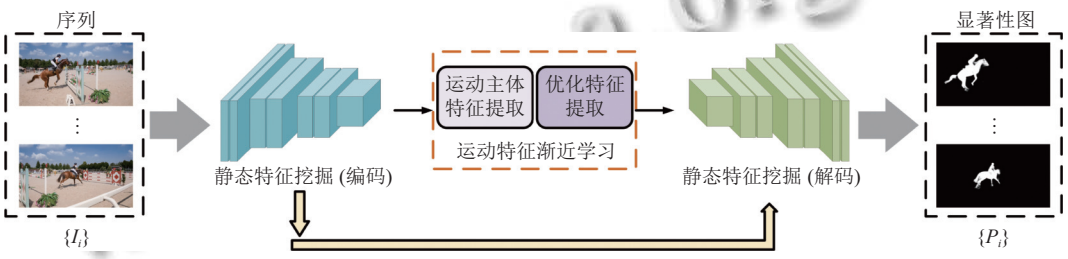


图 2 时空渐进式学习网络

2 本文方法

本文基于全卷积神经网络提出一种新颖的视频显著性目标检测算法—时空渐近式学习网络 (STPLNet). 如图 2 所示, 整个 STPLNet 主要由两个模块组成, 分别是静态特征挖掘 (SFM) 和运动特征渐近学习 (MFPL). SFM 使用一种 U 型结构对各帧进行编码和解码以挖掘静态空间特征. MFPL 在时间域上从运动目标主体部分和形变区域两方面入手, 达到对时间特征的渐进学习.

2.1 静态特征挖掘

如图 3 所示, 为了充分挖掘静态特征并综合考虑浅层细节特征与深层语义信息, 静态特征挖掘 (SFM) 模块分为编码器 (图 3 上半部分) 和解码器 (图 3 下半部分).

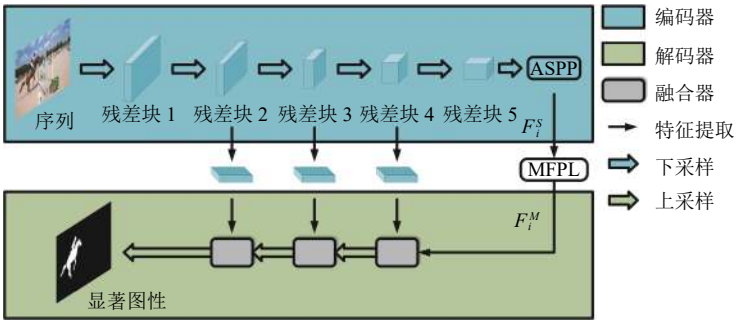


图 3 静态特征挖掘模块

编码器主要用于对输入序列中的各帧图片进行编码操作以挖掘其空间信息, 解码器则进行解码操作, 结合浅层信息将特征图恢复成原始尺寸并输出显著性预测图.

综合考虑模型的构建深度和卷积计算量, 基于目前较为成熟的 ResNet50[25] 结构对编码器进行搭建. 首先使用 ResNet50 中 5 个残差块对输入帧进行下采样, 在编码器的第 5 个残差块之后连接一个空洞空间金字塔池化 (atrous spatial pyramid pooling, ASPP)[26] 模块. ASPP 并行使用 3 种不同的空洞卷积对输入进行多尺度特征提取, 采样率分别是 6、12、18. 相比于普通的卷积, 空洞卷积由于存在采样率, 所以拥有更大的感受野, 能够对更大范围的信息进行卷积采样. 本文的 ASPP 模块对输入特征分别使用 3 种采样率的空洞卷积进行采样, 随后将相应结果进行拼接, 最后再经过卷积层将其尺寸恢复到输入大小. 通过这种方式可以获取输入的多尺度特征, 能够提取出更丰富的信息, 增强模型的决策能力和鲁棒性. 这里编码器的输入为连续视频帧 I_i, 其中i代表序列帧编号, 编码器处理过后的特征图表示为F_i^S, 随后将其输入到 MFPL 中完成时间域的特征学习. 整体流程如下:

$$F_i^S = \mathrm{SFM}_{\mathrm{encoder}}(I_i) \tag{1}$$

其中, $\mathrm{SFM}_{\mathrm{encoder}}$为静态特征挖掘模块中的编码器部分, I_i为输入图像, i为序列帧编号, F_i^S为经过$\mathrm{SFM}_{\mathrm{encoder}}$处理的输出特征图.

经过 MFPL 处理过后的特征图其尺寸为 256×28×28, 该特征图具有较好的感受野, 对原始输入图片的特征映射能力较好, 同时 256 维的深度也保证了信息的丰富度. 这些特征逐帧在解码器中与编码器提取的对应帧特征逐步融合并进行上采样, 使得深层信息和浅层信息能够充分融合, 从而加强特征的表达能力. 如图 3 所示, 对应编码器中的残差块 2、残差块 3、残差块 4, 本文设置了 3 个融合器. 融合器接收两种输入特征图, 使用逐像素相加方式对其进行融合, 随后对融合后的结果进行上采样. 将 MFPL 处理后的特征图与编码器残差块 4 提取的特征输入到融合器中, 首先融合器对这两种特征图进行融合, 随后对融合后的结果上采样使其尺寸匹配编码器残差块 3 提取的特征. 这样经过 3 次融合操作后, 特征图将会兼顾具有高级语义信息的深层特征和含有丰富细节信息的浅层特征. 最后将特征图尺寸恢复到输入大小, 完成解码操作. 整体流程如下:

$$P = \mathrm{SFM}_{\mathrm{decoder}}(F_i^M) \tag{2}$$

其中, $\mathrm{SFM}_{\mathrm{decoder}}$为静态特征挖掘模块的解码器部分, F_i^M为 MFPL 输出的特征图, P代表最终输出的显著性预测图, 其尺寸为 1×448×448 .

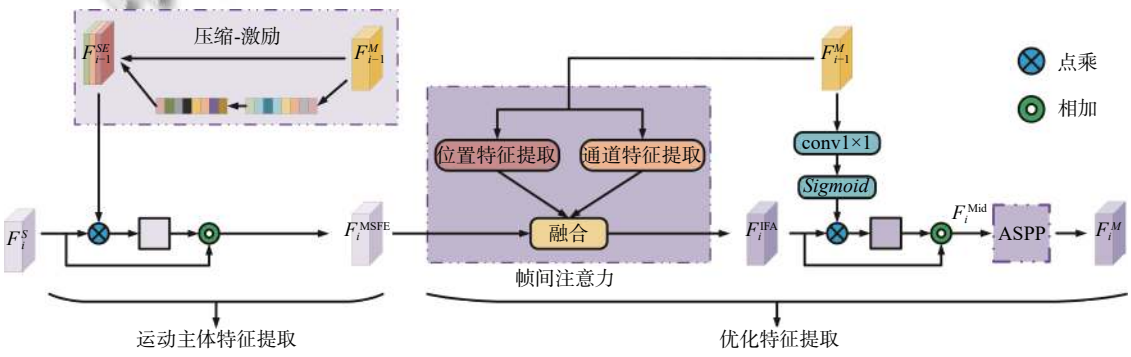


图 4 运动特征渐进学习模块

2.2 运动特征渐进学习

对于运动特征渐进学习 (MFPL) 模块来说, 其输入为连续视频序列帧F_i^S, 尺寸为 256×28×28, i为不同帧编号. 整体流程如下:

$$F_i^M = \mathrm{MFPL}(F_i^S, F_{i-1}^M) \tag{3}$$

其中, MFPL表示运动特征渐进学习模块, F_{i-1}^M为经过MFPL处理的上一帧特征图, F_i^M为当前帧输出.

相比于各视频帧目标的静态空间特征, MFPL 关注连续视频帧间运动目标的时间特征. 观察视频序列中目标的运动变化趋势, 本文将运动目标解析成主体部分和形变区域, 对其特征进行渐进学习. 基于视频序列的连续性以及目标的运动性, 使用上一帧特征来指导当前帧, 不断地对当前帧特征进行加强同时引导 MFPL 对形变特征进行学习, 这样各帧都受到了其过去帧的指导从而保证了对运动特征学习的连续性. 通过这种方式, MFPL 可以捕捉到视频序列中帧间运动目标特征的时间相关性, 从而对目标运动趋向性进行预测. MFPL 的结构如图 4 所示, 它由两个有效模块组成, 分别是运动主体特征提取 (MSFE) 和优化特征提取 (RFE). 其中, MFPL 的输入序列为F_i^S, 将经过完整 MSFE 和 RFE 指导学习后的最终输出表示为F_i^M.

2.2.1 运动主体特征提取

MSFE 旨在对连续帧中属于运动目标的主体部分特征进行指导加强. 首先将上一帧F_{i-1}^M输入到压缩激励网络 (SENet)[27] 对特征进行优化. SENet 对输入进行特征压缩, 将每个二维的通道变成一个实数用其映射全局信息, 随后基于通道之间的依赖关系对每一个输入通道进行筛选和权重评估. 这样 SENet 通过特征通道之间的相关性对F_{i-1}^M进行了校准, 从全局角度出发完善特征, 加强权重高的特征同时抑制权重较低的特征. 相应公式如下:

$$F_{i-1}^{SE} = \mathrm{SENet}(F_{i-1}^M) \tag{4}$$

其中, F_{i-1}^M为 MFPL 模块的输出特征图, 其尺寸为 256×28×28, 用F_{i-1}^{SE}表示经过 SENet 处理的F_{i-1}^M.

为了增强当前帧目标主体部分的特征, 对F_{i-1}^{SE}和F_i^S进行逐通道特征点乘且保持通道数不变. 将点乘后的特征和原始输入的F_i^S进行融合, 以保持当前帧中信息的完整度, 至此完成对视频序列中运动目标主体部分特征的提取和学习. 整个过程如下:

$$F_i^{\mathrm{MSFE}} = \begin{cases} F_i^S \times F_{i-1}^{SE} + F_i^S, i \neq 1 \\ F_i^S \times F_i^{SE} + F_i^S, i = 1 \end{cases} \tag{5}$$

其中, ×代表对应特征相乘, +表示对应特征相加, F_i^{MSFE}表示 MSFE 的输出结果, 对于第 1 帧来说使用其自身进行特征提取.

2.2.2 优化特征提取

由于视频序列中的目标在运动时部分区域会发生形变, 本文设计了 RFE 用来处理运动目标的形变区域同时进一步对主体部分进行特征增强. 首先设计一个帧间注意力 (inter-frame attention, IFA) 模块, 对序列中运动目标的形变区域进行学习. 如图 5 所示, IFA 由位置注意力和通道注意力两部分组成[28], 对输入序列的位置特征和通道特征进行学习. 位置注意力捕获一张特征图上任意位置之间的空间依赖性, 根据像素的特征相似性对其进行计算, 从而促进位置之间特征相互改进. IFA 对上一帧F_{i-1}^M进行位置特征计算, 随后将其融合到当前帧F_i^{MSFE}上, 使得当前帧捕获并学习到上一帧中目标像素位置之间的关系. 通道注意力利用任意通道之间的依赖性, 增强相互之间的特征映射关系以完成通道间的特征计算. 当上一帧F_{i-1}^{M}完成通道特征提取后, 再将其和当前帧F_i^{MSFE}融合以学习上一帧通道之间的特征表达, 使得当前帧通道之间拥有更好的依赖性.

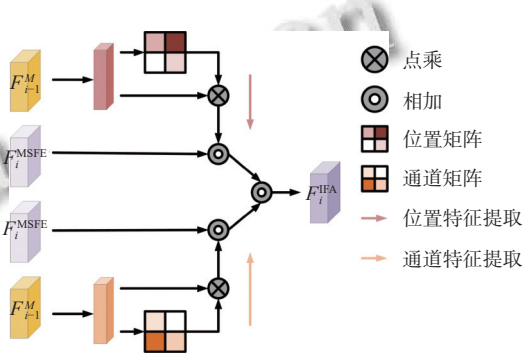


图 5 帧间注意力模块

IFA 通过对运动目标进行特征提取和权重学习, 能够捕捉到丰富的上下文关系. 这样使得当前帧能够关注到目标的运动趋向性, 根据连续帧间关系来识别出运动目标的形变区域并完成其在当前帧中的预测. 具体过程如下:

$$F_i^{\mathrm{IFA}}{}_{\mathrm{position}} = \mathrm{IFA}_{\mathrm{position}}\left(F_i^{\mathrm{MSFE}}, F_{i-1}^M\right) \tag{6}$$

$$F_i^{\mathrm{IFA}}{}_{\mathrm{channel}} = \mathrm{IFA}_{\mathrm{channel}}\left(F_i^{\mathrm{MSFE}}, F_{i-1}^M\right) \tag{7}$$

$$F_i^{\mathrm{IFA}} = F_i^{\mathrm{IFA}}{}_{\mathrm{position}} + F_i^{\mathrm{IFA}}{}_{\mathrm{channel}} \tag{8}$$

其中, $\mathrm{IFA}_{\mathrm{position}}$代表 IFA 对$F_i^{\mathrm{MSFE}}$中运动目标的位置特征进行学习, 其结果表示为$F_i^{\mathrm{IFA}}{}_{\mathrm{position}}$. $\mathrm{IFA}_{\mathrm{channel}}$代表 IFA 对$F_i^{\mathrm{MSFE}}$中运动目标的通道特征进行学习, 其结果表示为$F_i^{\mathrm{IFA}}{}_{\mathrm{channel}}$. +表示对应特征相加, F_i^{IFA}为帧间注意力模块的最终输出结果.

随后使用一个 1×1 卷积对F_{i-1}^{M}进行压缩, 以保持尺度不变并实现跨通道的信息共享和整合. 之后连接一个非线性 *Sigmoid* 激活函数, 增加网络的非线性拟合能力, 这样生成的单通道特征图拥有丰富的语义信息. 最后将它和F_i^{IFA}进行相乘, 并将结果与原始的F_i^{IFA}相结合. 该操作可以进一步加强当前帧运动目标的主体部分特征, 同时对无用的背景噪声进行抑制, 提高特征的表达能力. 相应过程如下所示:

$$F_i^{\mathrm{Mid}} = F_i^{\mathrm{IFA}} \times Sigmoid(F_{i-1}^{M}) + F_i^{\mathrm{IFA}} \tag{9}$$

其中, ×表示对应特征相乘, +代表对应特征相加, F_i^{Mid}为输出结果.

最后引入 ASPP 模块并将F_i^{Mid}输入其中, 提高模型在时间域特征学习的鲁棒性, 用F_i^{M}表示在 MFPL 中经过完整的时域特征学习后的输出结果.该特征图F_i^{M}的尺寸为 256×28×28, 将其输入到 SFM 模块中完成解码操作. 特征图首先和 SFM 编码器中残差块 4 输出的特征图 (尺寸为 256×28×28) 进行融合并上采样, 后续依次和残差块 3、残差块 2 的输出逐步进行相同操作. 最后将特征图尺寸恢复至 1×448×448, 以此作为显著性预测图进行输出, 完成解码操作.

本文提出的算法通过学习输入序列中运动目标的空间特征和时间特征, 以实现对显著性目标的准确预测. 模型首先在$\mathrm{SFM}_{\mathrm{encoder}}$中对帧内静态空间特征编码, 将结果$F_i^{S}$输入至MFPL中进行帧间运动目标时间特征的学习. 最后再使用$\mathrm{SFM}_{\mathrm{decoder}}$对$F_i^{M}$进行解码, 输出显著性预测图 P.

2.3 损失函数

本文算法所有的训练过程使用的损失函数为二值交叉熵损失函数[29], 其公式如下:

$$L_{\mathrm{BCE}} = -\sum_{\mathrm{i}} G_i \times \log P_i - \sum_{i} (1-G_i) \times \log(1-P_i) \tag{10}$$

其中, G为真值标签, P为预测图, i为像素点, ×为逐元乘法. 在训练过程中首先对静态特征挖掘模块进行预训练, 随后对整个模型 (静态特征挖掘模块和运动特征渐进学习模块) 同时进行微调训练.

3 实验与结果

本节将详细介绍具体的实验设置, 并对实验结果进行相应分析.

3.1 数据集和评价指标

算法实验环境为配置 NVIDIA 3080 GPU 的 Ubuntu 16 系统, 在 PyTorch 1.7 框架中完成代码编写. 首先使用 DUTS[30] 和 HKU-IS[31] 数据集对 SFM 进行预训练. DUTS 拥有 10 553 张训练图像, 其中包含重要的场景. HKU-IS 包含 4 447 张具有显著目标的图像. 预训练之后使用 DAVIS[32], FBMS[33], VOS[34] 和 DAVSOD[8] 数据集的训练集对整个模型进行微调训练. 最后在这 4 个数据集的测试集上进行测试. DAVIS 是一个经典视频数据集拥有 50 个不同序列共计 3 455 张像素级标注图片. FBMS 拥有 59 个视频序列, 共计 720 张图片且标注了像素级运动目标. VOS 拥有 200 个视频序列. DAVSOD 是一个具有挑战性的数据集, 其专门设计于视频任务.

本文使用 3 个通用指标对算法进行评价, 分别是 F-measure[35], S-measure[36] 和 MAE[37].

F-measure 是一个经典且有效的评价指标, 综合评价算法的准确率和召回率, 公式为:

$$F\text{-measure} = \frac{\left(1+\beta^2\right) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{11}$$

其中, $Precision$为精确率代表正确预测为正例占全部预测为正例的比率, $Recall$召回率表示正确预测正例占真值的比率, 设置β^2为 0.3, 以更加强调准确率. 在本文中报告最大的 F-measure 值, 记为 max$F$. 其值越接近 1, 表示算法结果越接近于标签.

S-measure 同时结合面向物体和面向区域的结构相似性度量对算法进行评价, S-measure 值同算法效果呈正相关, 最大不超过 1. 其值越高, 表示算法效果越好, 文中记为 S.

MAE 是平均绝对误差, 用来评价预测结果和标签的绝对误差平均值, 公式为:

$$MAE = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |P(x,y) - G(x,y)| \tag{12}$$

其中, x和y表示像素坐标位置, H和W代表预测图的高和宽. 其值越小, 表示预测结果P和标签G越相近, 而完全吻合时该值为 0, 文中记为 *MAE*.

3.2 模型分析

3.2.1 消融实验

本文算法由多个部分组成, 通过实施消融实验来验证各个模块的有效性. 如表 1 所示, 在 DAVIS, FBMS 和 VOS 数据集上使用全部 3 个指标进行评价, 保证了实验的充分性和准确性.

表 1 消融实验效果对比

模块			DAVIS		FBMS		VOS	
SFM	MSFE	RFE	max$F$	S	max$F$	S	max$F$	S
√			0.797	0.855	0.845	0.871	0.740	0.822
√	√		0.864	0.895	0.879	0.887	0.800	0.855
√		√	0.866	0.895	0.865	0.880	0.788	0.847
√	√	√	**0.875**	**0.900**	**0.881**	**0.890**	**0.804**	**0.856**

注: 加粗表示各列排名第一结果

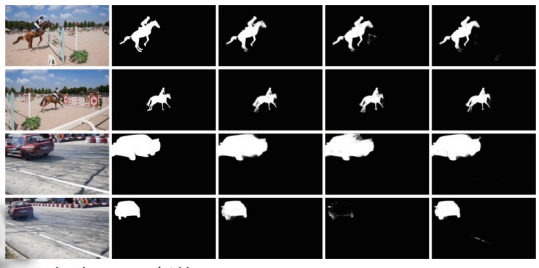
SFM: 首先移除 MFPL 同时保持 SFM 结构不变, 在此基础上测试算法对于静态空间特征的提取和学习能力. 如表 1 中第 1 行结果所示, 当仅有 SFM 起作用时算法也保持较好效果. 该结果验证了 SFM 能初步提取和学习输入序列的空间特征, 为算法进一步对时间域特征的学习打下了基础, 证明了所使用的预训练方案的有效性.

SFM+MSFE: MSFE 旨在针对运动目标主体部分进行时域特征学习. 从表 1 中第 2 行结果来看, 当 SFM 和 MSFE 共同作用时, 算法相比于仅使用 SFM 增加了对时间域特征的学习且结果有了初步改善. 该实验验证了所提出的 MSFE 能够有效地学习到帧间目标的时间特征, 对整体算法是至关重要的

SFM+RFE: RFE 对于序列中运动目标形变区域进行学习, 同时进一步优化主体部分特征. 如表 1 中第 3 行结果所示, 所提出的 RFE 也具有很好地时域特征学习能力, 是算法的重要组成部分.

SFM+MSFE+RFE: 当时域中的 MSFE 和 RFE 共同起作用时, 其同时对运动目标的主体部分和形变区域提取和学习特征, 使得算法能够完整地对时域中的运动特征进行处理. 如表 1 中第 4 行结果所示, 该实验充分证明所设计的 MSFE 和 RFE 共同作用时能够较好地对运动目标主体部分特征进行强化, 同时也较为准确地捕捉到目标的运动趋势并对其形变部分进行学习. 由此可见, MSFE 和 RFE 的共同作用对于整个算法是十分重要的且有利于精度提升, 进一步证明了所提模块的有效性.

本文也将相应的结果进行了可视化, 其也进一步证实了各模块的有效性. 如图 6 所示, 当所有模块都起效果时, 算法拥有最好的表现结果, 能够完整地定位到整个显著性目标区域.



(a) 序列 (b) 标签 (c) MSFE+RFE (d) MSFE (e) RFE

图 6 消融实验效果可视化对比

3.2.2 不同特征融合方式对比

为了进一步对处理时域特征的 MFPL 进行分析, 这里在 DAVIS, FBMS 和 VOS 数据集上验证不同特征融合方式对于算法最终性能的影响. 如图 4 所示, 对于 MFPL 中的两个圆环型结构 (代表特征图融合方式), 分别进行不同的融合操作实验: 拼接 (Concat) 和相加 (Add). 其中拼接融合方式要求参与的特征图拥有相同的通道数, 连接后通道数增加. 相加融合方式是具有相同通道数的特征图逐像素进行权重相加且不改变原通道数. 如表 2 中第 2 行结果可知, 在 DAVIS 数据集上相加融合方式的 max$F$ 值相比于拼接融合方式提高了 0.3%, 而 S 值也改善了 0.2%. 在 FBMS 和 VOS 数据集上, 相加融合方式的 max$F$ 和 S 值也均优于拼接融合方式. 从结果来看, 相比于拼接融合方式, 相加融合方式可以充分兼顾不同特征, 防止有效信息的丢失. 此外, 增加了对每个特征对应位置信息量的表述且保持特征图维度不变, 有益于后续训练. 该实验验证了在 MFPL 中, 使用相加融合方式更好地提升最终算法的性能.

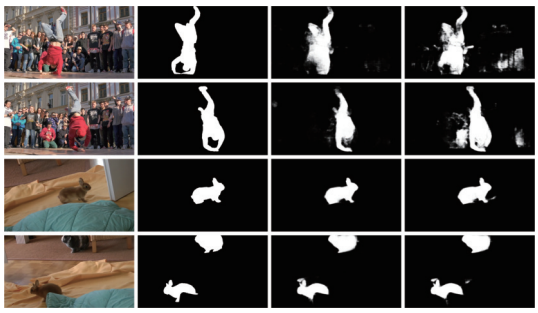
表 2 不同融合方式效果对比

融合方式	DAVIS		FBMS		VOS	
	max$F$	S	max$F$	S	max$F$	S
拼接	0.872	0.898	0.876	0.882	0.793	0.849
相加	**0.875**	**0.900**	**0.881**	**0.890**	**0.804**	**0.856**

注: 加粗表示各列排名第一结果

如图 7 可视, 对比二者生成的显著性图发现, 相加融合方式的预测结果更接近于标签同时对噪声产生了

较好的抑制效果. 综合考虑, 本文使用相加融合作为 MFPL 中的特征融合方式.



(a) 序列 (b) 标签 (c) 相加 (d) 拼接

图 7 不同融合方式效果可视化对比

3.3 与先进算法对比

3.3.1 定量分析

本文和近 5 年 13 个先进算法进行定量分析实验, 对比算法包括 SCOM[38], SCNN[39], DLVS[40], FGRN[41], MBNM[42], PDBM[17], SSAV[8], NHM[43], STFA[21], CAS[44], MSAA[45], STEG[22], PVSOD[24]. 为了实验的充分性和公平性, 所有用于对比的先进算法的评价指标结果都来自公开发表的论文, 保证相同的测试标准. 在 DAVIS, FBMS, VOS 和 DAVSOD 数据集上对所有算法进行统一评价. 如表 3 可视, 使用 max$F$, S, MAE 作为评价指标来验证不同算法的检测能力. 从实验结果来看, 本文算法在多个不同的数据集上都能够有较好的效果, 仅有少部分的指标稍显不足但与最好的指标也相差不远. 如表 3 所示, 在 DAVIS 数据集上拥有最好的 3 个评价指标结果. 其 max$F$ 值高出第 2 名算法 STFA 1%, S 值超出第 2 名算法 SSAV 0.7%. 在 FBMS 数据集中, 拥有最好的 max$F$ 值和 S 值. 具体来说, max$F$ 值高出第 2 名算法 SSAV 1.6%, S 值超出 1.1%. 在 VOS 和 DAVOSD 数据集上的结果依旧位居前列. 其中, VOS 上的 max$F$ 值和 S 值在所有算法中皆排名第一, 其 max$F$ 值和 S 值分别高出第 2 名算法 STFA 1.3% 和 0.6%. 在 DAVSOD 这个全新的具有挑战性的数据集上, 本文算法的 3 个评价指标均排在第 1 名, 优于其他算法, 其 max$F$ 值超出 2022 年最新的无监督视频显著性目标检测算法 PVSOD 0.5%, 同时 MAE 值降低了 0.6%.

该定量分析实验验证了所提出的 SFM 能够较好地从各帧中挖掘出静态空间特征, 并且 MFPL 能够针对运动目标主体部分和形变区域两方面提取和学习到足够的时间特征, 证明了本文提出算法的有效性.

表 3 不同算法效果对比

算法	DAVIS			FBMS			VOS			DAVSOD		
	max$F$	S	MAE	max$F$	S	MAE	max$F$	S	MAE	max$F$	S	MAE
$SCOM_{18}$	0.783	0.832	0.048	0.797	0.794	0.079	0.690	0.712	0.162	0.464	0.599	0.220
$SCNN_{18}$	0.714	0.783	0.064	0.762	0.794	0.095	0.609	0.704	0.109	0.532	0.674	0.128
$DLVS_{18}$	0.708	0.794	0.061	0.759	0.794	0.091	0.675	0.760	0.099	0.521	0.657	0.129
$FGRN_{18}$	0.783	0.838	0.043	0.767	0.809	0.088	0.669	0.715	0.097	0.573	0.693	0.098
$MBNM_{18}$	0.861	0.887	0.031	0.816	0.857	0.047	0.670	0.742	0.099	0.520	0.637	0.159
$PDBM_{18}$	0.855	0.882	0.028	0.821	0.851	0.064	0.742	0.818	0.078	0.572	0.698	0.116
$SSAV_{19}$	0.861	0.893	0.028	0.865	0.879	0.040	0.742	0.819	0.073	0.603	0.724	0.092
NHM_{20}	—	—	—	0.849	0.868	0.051	0.729	0.810	0.072	0.630	0.729	0.090
$STFA_{21}$	0.865	0.892	**0.023**	0.856	0.872	**0.038**	0.791	0.850	0.058	0.651	0.746	0.086
CAS_{21}	0.860	0.873	0.032	0.863	0.856	0.055	0.774	0.808	**0.051**	0.608	0.699	0.086
$MSAA_{21}$	0.844	0.880	0.031	—	—	—	—	—	—	—	—	—
$STEG_{21}$	—	—	—	0.813	0.837	0.066	—	—	—	—	—	—
$PVSOD_{22}$	0.844	0.869	0.041	0.862	0.873	0.042	0.729	0.811	0.074	0.659	0.744	0.085
本文	**0.875**	**0.900**	**0.023**	**0.881**	**0.890**	0.045	**0.804**	**0.856**	0.059	**0.664**	**0.747**	**0.079**

注: 加粗表示各列排名第一结果, 不同算法名称右下方数字为提出时间, —标注表示该算法未公开该指标

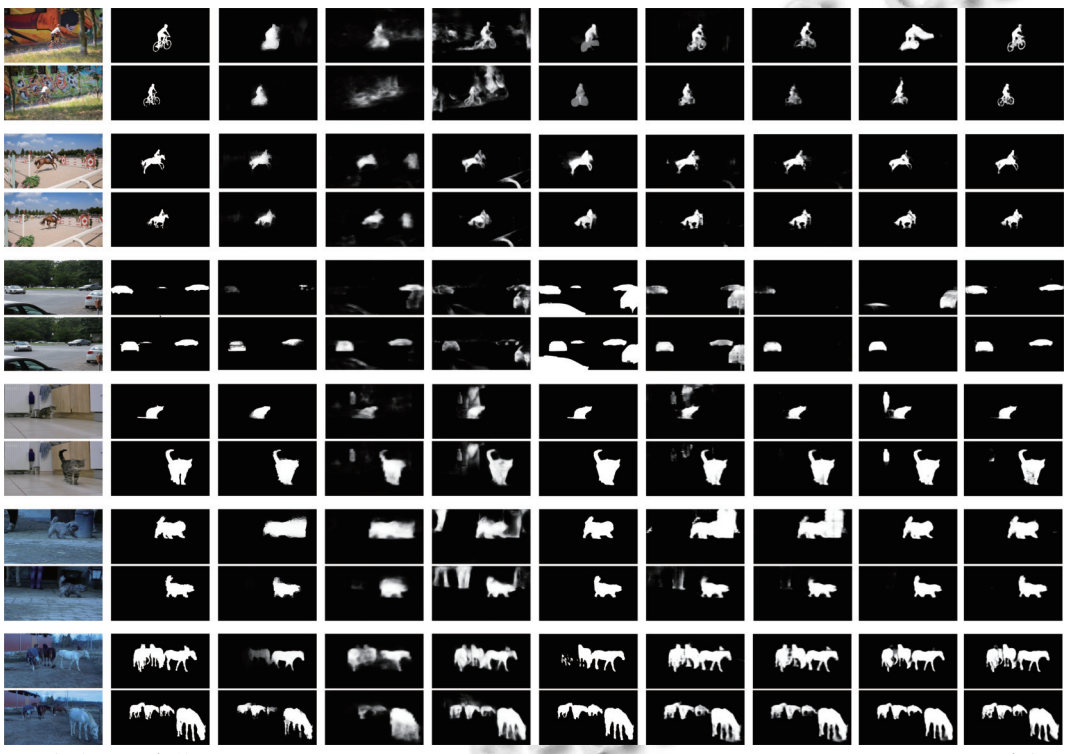
3.3.2 定性分析

鉴于篇幅有限, 这里使用具有较好检测效果的算法进行可视化定性比较, 其中使用算法 SCOM, DLVS, FGRN, MBNM, PDBM, SSAV, STFA 用作可视化比较. 如图 8 所示, 从可视化结果来看, 本文提出的算法具有较好的检测能力, 能够较为准确地从输入中识别出正确的显著性目标, 其输出的预测图和标签相似度较高. 从图中可以观察到, 对于 DAVIS 数据集中的 bmx-trees

序列和 horsejump-high 序列 (图 8 第 1–4 行), 显著目标较小且存在复杂背景, 识别难度较大. 其他算法难以准确地识别出该显著运动目标, 存在检测区域不全和将背景标记为前景的问题, 而本文算法可以准确地将目标识别出并具有较好的边缘. 而对低对比度的序列也能较好地进行识别, 例如在 FBMS 数据集的 dogs01 序列 (图 8 第 9、10 行) 中, 准确地检测出显著性目标, 并对背景噪声进行抑制. 在 horse05 序列 (图 8第 11、12 行) 中也找全了所有的目标, 相比于最新的 STFA 算法拥有更好的边缘. 这主要得益于所提出的算法能够对各帧的静态空间特征进行较为完善的挖掘, 同时也能够较好地捕捉到帧间上下文信息进而学习到运动目标的时间特征.

3.3.3 算法检测时间分析

为了综合评价算法的检测能力, 在多个数据集上测试其检测速度. 如表 4 所示, 在 DAVIS 数据集上的检测速度为 27 f/s, 在 DAVSOD 数据集上的检测速度最快, 达到了 32 f/s.



(a) 序列 (b) 标签 (c) SCOM (d) DLVS (e) FGRN (f) MBNM (g) PDBM (h) SSAV (i) STFA (j) 本文

图 8 不同算法效果可视化对比

表 4 算法在不同数据集上的检测速度 (f/s)

数据集	检测速度
DAVIS	27
FBMS	28
VOS	29
DAVSOD	**32**

注: 加粗表示最快检测速度

此外, 与 6 种不同的先进算法在 DAVIS 数据集上进行检测时间对比, 所有结果均来自公开发表的论文. 如表 5 中所示, 本文算法的检测速度为 27 f/s (单帧检测时间为 0.037 s), 满足实时性需求. 最新算法 MSAA 的单帧检测时间达到了 0.01 s, 但在多个数据集上的检测精度不如本文算法. 综合来看, 本文提出的算法在满足实时性的同时兼顾检测准确度, 在多个数据集上都有较好的表现.

表 5 不同算法的检测时间对比 (s)

算法	时间
SCOM	38.8
SCNN	38.5
DLVS	0.47
PDBM	0.05
SSAV	0.05
MSAA	**0.01**
本文	0.037

注: 加粗表示最短检测时间

4 结论

针对现有算法对视频序列中时间信息连续学习能力差的问题, 本文提出了一种时空渐进式学习网络(STPLNet)的视频显著性目标检测算法. 提出的静态特征挖掘模块主要用于提取和学习输入视频序列中各帧的空间特征, 运动特征渐进学习模块在于充分学习视频序列的时间特征. 在对时间特征学习时, 分析连续视频帧中运动目标的特性, 使用帧间指导学习的方式对其显著性主体部分进行特征增强, 对其形变区域进行运动性预测. 将 STPLNet 在 4 个数据集上与 13 个先进算法进行对比, 结果表明在 3 个评价指标 max$F$, S 和 MAE 下取得了综合最好的成绩, 同时检测时间满足实时性要求. 接下来将继续探索更有效的帧间信息指导方式, 提升时域特征的鲁棒学习能力, 以生成更准确的显著性预测结果.

参考文献

1 Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10778–10787. [doi: 10.1109/CVPR42600.2020.01079]

2 Wang Q, Zhang L, Bertinetto L, *et al*. Fast online object tracking and segmentation: A unifying approach. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 1328–1338. [doi: 10.1109/CVPR.2019.00142]

3 Anderson P, He XD, Buehler C, *et al*. Bottom-up and top-down attention for image captioning and visual question answering. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6077–6086. [doi: 10.1109/CVPR.2018.00636]

4 Wang WG, Shen JB, Shao L. Consistent video saliency using local gradient flow optimization and global refinement. IEEE Transactions on Image Processing, 2015, 24(11): 4185–4196. [doi: 10.1109/TIP.2015.2460013]

5 Shi XJ, Chen ZR, Wang H, *et al*. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 802–810.

6 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 4489–4497. [doi: 10.1109/ICCV.2015.510]

7 Li HF, Chen GQ, Li GB, *et al*. Motion guided attention for video salient object detection. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 7273–7282. [doi: 10.1109/ICCV.2019.00737]

8 Fan DP, Wang WG, Cheng MM, *et al*. Shifting more attention to video salient object detection. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 8546–8556. [doi: 10.1109/CVPR.2019.00875]

9 Min K, Corso J. TASED-Net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 2394–2403. [doi: 10.1109/ICCV.2019.00248]

10 包晓安, 朱晓芳, 张娜, 等. 基于背景感知的显著性目标检测算法. 计算机系统应用, 2018, 27(6): 103–110. [doi: 10.15888/j.cnki.csa.006428]

11 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440. [doi: 10.1109/CVPR.2015.7298965]

12 Liu JJ, Hou QB, Cheng MM, *et al*. A simple pooling-based design for real-time salient object detection. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3912–3921. [doi: 10.1109/CVPR.2019.00404]

13 陈维婧, 周萍, 杨海燕, 等. 通道-空间联合注意力机制的显著性检测模型. 计算机工程与应用, 2021, 57(19): 214–219. [doi: 10.3778/j.issn.1002-8331.2006-0238]

14 左保川, 张晴. 采用特征引导机制的显著性检测网络. 计算机工程与应用, 2021, 57(14): 201–208. [doi: 10.3778/j.issn.1002-8331.2011-0187]

15 Xie CX, Xia CQ, Ma MC, *et al*. Pyramid grafting network for one-stage high resolution saliency detection. Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11707–11716.

16 Zhang LQ, Zhang Q, Zhao R. Progressive dual-attention residual network for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 5902–5915. [doi: 10.1109/TCSVT.2022.3164093]

17 Song HM, Wang WG, Zhao SY, *et al*. Pyramid dilated

deeper convLSTM for video salient object detection. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 744–760. [doi: 10.1007/978-3-030-01252-6_44]

18 Yan PX, Li GB, Xie Y, *et al*. Semi-supervised video salient object detection using pseudo-labels. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 7283–7292. [doi: 10.1109/ICCV.2019.00738]

19 Wang XL, Girshick R, Gupta A, *et al*. Non-local neural networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7794–7803. [doi: 10.1109/CVPR.2018.00813]

20 Ballas N, Yao L, Pal C, *et al*. Delving deeper into convolutional networks for learning video representations. Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016. [doi: 10.48550/arXiv.1511.06432]

21 Chen CLZ, Wang GT, Peng C, *et al*. Exploring rich and efficient spatial temporal interactions for real-time video salient object detection. IEEE Transactions on Image Processing, 2021, 30: 3995–4007. [doi: 10.1109/TIP.2021.3068644]

22 Bi HB, Yang LN, Zhu HH, *et al*. STEG-Net: Spatiotemporal edge guidance network for video salient object detection. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(3): 902–915. [doi: 10.1109/TCDS.2021.3078824]

23 Chen YW, Jin XJ, Shen XH, *et al*. Video salient object detection via contrastive features and attention modules. Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2022. 536–545. [doi: 10.1109/WACV51458.2022.00061]

24 Xu BW, Liang HR, Ni WT, *et al*. Learning video salient object detection progressively from unlabeled videos. arXiv:2204.02008, 2022.

25 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

26 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587, 2017.

27 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141. [doi: 10.1109/CVPR.2018.00745]

28 Fu J, Liu J, Tian HJ, *et al*. Dual attention network for scene segmentation. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 3141–3149. [doi: 10.1109/CVPR.2019.00326]

29 De Boer PT, Kroese DP, Mannor S, *et al*. A tutorial on the cross-entropy method. Annals of Operations Research, 2005, 134(1): 19–67. [doi: 10.1007/S10479-005-5724-Z]

30 Wang LJ, Lu HC, Wang YF, *et al*. Learning to detect salient objects with image-level supervision. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3796–3805. [doi: 10.1109/CVPR.2017.404]

31 Li GB, Yu YZ. Visual saliency based on multiscale deep features. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 5455–5463. [doi: 10.1109/CVPR.2015.7299184]

32 Perazzi F, Pont-Tuset J, McWilliams B, *et al*. A benchmark dataset and evaluation methodology for video object segmentation. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 724–732. [doi: 10.1109/CVPR.2016.85]

33 Ochs P, Malik J, Brox T. Segmentation of moving objects by long term video analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(6): 1187–1200. [doi: 10.1109/TPAMI.2013.242]

34 Li J, Xia CQ, Chen XW. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. IEEE Transactions on Image Processing, 2018, 27(1): 349–364. [doi: 10.1109/TIP.2017.2762594]

35 Achanta R, Hemami S, Estrada F, *et al*. Frequency-tuned salient region detection. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 1597–1604. [doi: 10.1109/CVPR.2009.5206596]

36 Fan DP, Cheng MM, Liu Y, *et al*. Structure-measure: A new way to evaluate foreground maps. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 4558–4567. [doi: 10.1109/ICCV.2017.487]

37 Perazzi F, Krähenbühl P, Pritch Y, *et al*. Saliency filters: Contrast based filtering for salient region detection. Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 733–740. [doi: 10.1109/CVPR.2012.6247743]

38 Chen YH, Zou WB, Tang Y, *et al*. SCOM: Spatiotemporal constrained optimization for salient object detection. IEEE

Transactions on Image Processing, 2018, 27(7): 3345–3357. [doi: 10.1109/TIP.2018.2813165]

39 Tang Y, Zou WB, Jin Z, *et al*. Weakly supervised salient object detection with spatiotemporal cascade neural networks. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(7): 1973–1984. [doi: 10.1109/TCSVT.2018.2859773]

40 Wang WG, Shen JB, Shao L. Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing, 2018, 27(1): 38–49. [doi: 10.1109/TIP.2017.2754941]

41 Li GB, Xie Y, Wei TH, *et al*. Flow guided recurrent neural encoder for video salient object detection. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3243–3252. [doi: 10.1109/CVPR.2018.00342]

42 Li SY, Seybold B, Vorobyov A, *et al*. Unsupervised video object segmentation with motion-based bilateral networks. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 215–231. [doi: 10.1007/978-3-030-01219-9_13]

43 Cai JP, Lin S. A novel hybrid model for video salient object detection. Proceedings of 2020 International Conference on Computer Engineering and Intelligent Control. Chongqing: IEEE, 2020. 275–279. [doi: 10.1109/ICCEIC51584.2020.00059]

44 Ji YZ, Zhang HJ, Jie ZQ, *et al*. CASNet: A cross-attention siamese network for video salient object detection. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(6): 2676–2690. [doi: 10.1109/TNNLS.2020.3007534]

45 Xu MZ, Fu P, Liu B, *et al*. Multi-stream attention-aware graph convolution network for video salient object detection. IEEE Transactions on Image Processing, 2021, 30: 4183–4197. [doi: 10.1109/TIP.2021.3070200]

(校对责编: 孙君艳)