用于视频显着物体检测的流动引导递归神经编码器

李冠斌[1] 袁燮[1] 田浩伟[2] 王克泽[1] 梁琳[1,3*]

[1]中山大学 [2]浙江大学 [3]SenseTime Group Limited liguanbin@mail.sysu.edu.cn, xiey39@mail2.sysu.edu.cn, thwei@zju.edu.cn, wangkeze@mail2.sysu.edu.cn, linliang@ieee.org

摘要

最近，由于深度卷积神经网络，图像显着性检测取得了显着进展。 然而，从图像到视频扩展最先进的显着性检测器是具有挑战性的。 显着物体检测的性能受到物体或相机运动的影响以及视频中外观对比度的急剧变化。 在本文中，我们提出了流向引导回归神经编码器（FGRNE），这是一种用于视频显着物体检测的准确和端到端的学习框架。 它通过利用光流方面的运动信息和利用LSTM网络的顺序特征演进编码来增强每帧特征的时间相关性。 它可以被认为是一个通用框架，可以将任何基于FCN的静态显着性检测器扩展到视频显着物体检测。 强化实验结果验证了FGRNE各部分的有效性，并证实我们提出的方法明显优于DAVIS和FBMS公共基准的最新方法。

1. 介绍

突出物体检测旨在识别引起人们注意的图像或视频中视觉上最具特色的物体。 由于需要在许多计算机视觉应用中解决这个问题，例如图像和视频压缩，已经引起了很多关注[12]，对象分割[37]，视觉追踪[38]和人员重新识别[43]。 虽然基于图像的显着物体检测在过去的十年中得到了广泛的研究，但基于视频的显着物体检测的研究却少得多由于其高复杂度和缺乏大规模注释视频数据集。

图1. 静态图像显着性检测器的挑战以及基于视频的显着物体检测中时间相干性建模的有效性。

近年来，由于深度卷积神经网络（CNN）的成功部署，静态图像中显着物体检测的性能提高了很多[21, 10, 18, 20]。 尽管如此，直接将这些方法应用于视频显着物体检测并不容易，且具有挑战性。 显着物体检测的性能受到物体或相机运动的影响以及视频中外观对比度的急剧变化。 如图2的第二行所示。 1，现有技术的静止图像显着物体检测器（例如DSS [10]）由于无法保持连续帧之间的显着对象的视觉连续性和时间相关性而急剧恶化。

认知研究表明，视觉对比是导致特定区域在静态图像中显着突出的关键因素。 对于动态视频而言，由物体运动引起的连续帧之间的差异更能吸引人们的注意[13]。 这种时间信息已经在现有的视频显着对象检测方法中以图形模型的形式被利用[35, 3]或者简单地嵌入卷积神经网络[36]。 基于图形模型的方法通常使用生成框架，其首先根据帧内外观对比度推断初始显着性图

*前两位作者对本文作出同样的贡献。 通讯作者梁林。 这项工作得到国家重点发展计划资助项目2016YFB1001004，国家自然科学基金资助项目61702565，广东省自然科学基金项目资助项目2017A030312006的资助，并获得CCF-腾讯开放式研究基金资助。

信息[3]或帧间梯度流场[35]，并进一步将能量函数与一些启发式时空模型相结合，以鼓励输出显着图的跨帧一致性。 由于它们独立于训练数据和使用手工制作的低级特征，因此基于图形模型的方法应对具有复杂语义对比度和对象运动的视频是艰巨的。 尽管在这些方法中光流已经被利用，但它仅用于启发式后处理的离台模式。 最近，随着深度CNN在静态图像显着物体检测中的应用日益广泛，还有人试图将CNN扩展到视频显着物体检测[36, 16]。 他们只是将连续的帧图像连接起来，并将其反馈到卷积神经网络以进行时间相干性建模。 然而，由于卷积神经网络不具备记忆功能，因此这种原始帧图像的幼稚聚合以及之后的严重卷积操作不能很好地表征视频帧在时域中的连续动态演变。而且，这种简单的时空建模策略缺乏对物体运动的明确补偿，使得难以在保持时间相干性的同时（例如物体移动超出神经网络的接受区域）以剧烈运动来检测显着物体。

在这项工作中，我们提出了流向导回归神经编码器（FGRNE），这是一个端到端的学习框架，用于将任何基于FCN的静止图像显着性检测器扩展到视频显着物体检测。 它通过利用光流方面的运动信息和利用LSTM网络的顺序特征演进编码来增强每帧特征的时间相关性。具体而言，我们使用现成的基于FCN的图像显着性检测器（例如DSS [10]）作为我们的主机网络，用于特征提取和最终显着性推断，以及预先训练好的FlowNet [7]用于帧对之间的运动估计。 我们的FGRNE学习如何通过引入流向导特征变形和基于LSTM的时间相干特征编码来改善每帧特征。 最后一个时间步的输出特征映射被认为是我们的编码特征，并被馈送到主机网络的上部用于显着性推断。 此外，我们的FGRNE还涉及另一个LSTM模块，以改善大时间间隔帧对的估计光流。FGRNE的所有三个模块包括运动计算和更新，流向引导特征变形以及时间相关特征编码都是与主机网络端到端训练。

总之，本文有以下贡献：

- 我们引入流引导递归神经编码器框架来增强每帧特征表示的时间相干性建模，其可以被利用来扩展任何基于FCN的静止图像显着性检测器对视频显着物体检测。
- 我们建议在FGRNE框架中加入一个光流网络来估计每帧的运动，这在进一步用于特征变形中以明确补偿物体的运动。
- 我们建议利用我们的FGRNE中的ConvLSTM进行顺序特征编码，该特征编码可捕获时域中的外观对比度的演变，并且补充特征翘曲以改进视频突出对象检测的性能。

2. 相关工作

2.1. 静止图像显着对象检测

几十年来图像显着物体检测已经被广泛研究。 传统的方法可以分为基于低级特征的自下而上的方法[8, 15, 5]和高层次知识指导下的自上而下模型[14, 40, 22]。近年来，深刻的CNN将突出显示目标检测的研究推向了一个新的阶段，并成为该领域的主导研究方向。 基于深度CNN的方法可以进一步分为两类，包括基于区域的深度特征学习[19, 42, 32]和端到端完全卷积网络基于方法[20, 10, 18, 33, 17]。 第一类中的方法将图像分成区域，并将每个区域作为独立单元进行深度特征提取和显着性推断。 由于特征提取和存储中的重要冗余，它们通常是空间和时间消耗的。 为了克服这个缺陷，已经开发了基于深度FCN的模型，以端到端可训练的方式将原始输入图像直接映射到其相应的显着图。这些方法可以充分利用特征共享机制，在单个网络前向操作中产生每个区域的分层特征。 他们可以产生出众的显着图，并成为该领域最先进方法的基本组成部分。

与这些基于静止图像的显着物体检测方法相比，我们关注视频显着物体检测，其中包含时间和运动信息，以改进用于显着性推断的特征映射表示。 它可以被认为是将任何基于FCN的模型扩展到视频显着物体检测的通用框架，并且可以很容易地从静止图像显着物体检测器的改进中受益。

2.2. 视频突出对象检测

与静止图像中的显着性检测相比，由于有效时空模型的高度复杂性，检测视频显着对象更具挑战性

以及缺乏大规模注释的视频数据集。 研究界对此的研究远远不够。 对这个问题的早期方法可以看作是一些静态显着性模型的简单扩展，具有额外的时间特征[24, 9]。 更新的和值得关注的作品通常将视频显着性检测作为连续帧上的时空上下文建模问题，并将能量函数与手工规则结合起来，以鼓励输出显着性图的空间平滑度和时间一致性[3, 35, 6]。 然而，这些方法都属于无监督生成模型，并且依赖于手工制作的低级特征用于启发式显着性推断，因此无法处理需要知识和语义推理的复杂视频。 尽管最近Le等人未发表的作品 [16]提出将深度CNN特征纳入时空CRF框架中以提高时间一致性，但仍存在多级流水线的缺点及其高计算成本。 与我们最相关的工作是[33]，它利用第二个FCN来改善从初始静态FCN显着网络生成的显着图的时间相关性，方法是将连续帧对的连接以及初始显着性图和直接映射到精化显着性在转发网络操作中映射。 由于卷积神经网络没有记忆功能，因此无法很好地模拟时域中视频帧的连续演变。 而且，这种时空模型的粗略策略缺乏对物体运动的明确补偿，使得难以通过剧烈运动来检测显着物体。

相比之下，我们的方法考虑了特征级别中的时间信息而不是原始输入帧，并且合并了LSTM网络以自然地编码顺序特征演进。 整个框架都是端对端培训，推理过程非常高效。 此外，我们的方法可以进一步结合这种基于图形模型的后处理技术（例如CRF）来提高性能。

2.3. 基于光流的运动估计

光流估计两个连续帧之间的每像素运动，并广泛用于各种视频分析任务。 传统方法主要基于变分公式，主要处理小位移，并受高效视频应用的高计算成本的限制。 最近，基于深度学习的方法已被用于光流计算[7, 28, 11]。 最具代表性的工作是FlowNet [7]这表明CNN可以应用于高效的光流推断。 还有一些尝试将FlowNet融入当代深度学习框架中，以加强代表性流程的时间连续性，

视频功能，这带来了各种视频理解任务的性能改进，包括视频识别[45]，对象检测[44]和视频对象分割[29]。

光流在现有的视频显着物体检测模型中已经被利用，然而，它在后处理中被用作辅助运动特征或手工规则以改善时间相干性。 灵感来自[45, 44]，我们合并了光流以实现跨帧的特征翘曲并补偿由对象运动引起的变化。 然而，与这些努力不同，运动流在我们的框架中动态更新，并且特征变形的结果被用于时间特征编码而不是特征聚合。 此外，我们首先将光流融合到回归神经编码器中，以进行有效的时空特征学习，并且在视频显着物体检测任务中展现了其优越的性能。

3. 流引导复发神经编码器

给定视频帧序列I_i，i = 1, 2, ..., N，视频显着目标检测的目的是输出所有帧的显着图S_i，i = 1, 2, ..., N。 用于静态图像的最先进的显着物体检测器大多基于FCN结构[20, 23, 18, 10]。 给定一个预训练静态模型（例如DSS [10]模型），它可以被认为是一个特征提取模块N_{FEA}，后面跟着一个像素明显回归模块N_{REG}。 给定图像I的输出显着图S可以被计算为$S = N_{REG}(N_{FEA}(I))$。 由于在特征表示中缺乏时间相干性建模，将该模型直接应用于每个单独的帧通常会生成不稳定的且时间上不一致的显着性图。

我们提出的FGRNE旨在通过额外查看k个前帧的片段来增强特征表示的时间一致性。 鉴于参考 - $E_i = (N_{FEA}(I_i), N_{FEA}(I_{i-1}), ..., N_{FEA}(I_{i-k}))$。 由于物体运动及其外观对比的变化是两个核心内容，通过影响显着性的因素，建议的FGRNE采用了现成的FlowNet模型[7]和基于LSTM的特征编码器来分别处理这两个因素。

如图所示。 2，我们FGRNE的架构由三个模块组成，包括运动计算和更新，运动引导特征变形和时间相干特征编码。 具体而言，我们首先计算k个前帧相对于参考帧的每个帧的光流图。 每个流程图进一步以相反的顺序被馈送到LSTM用于运动细化。 其次，应用每个时间步骤更新的流程图来相应地扭曲特征映射。 最后，每个扭曲的特征被连续地馈送到另一个LSTM用于时间

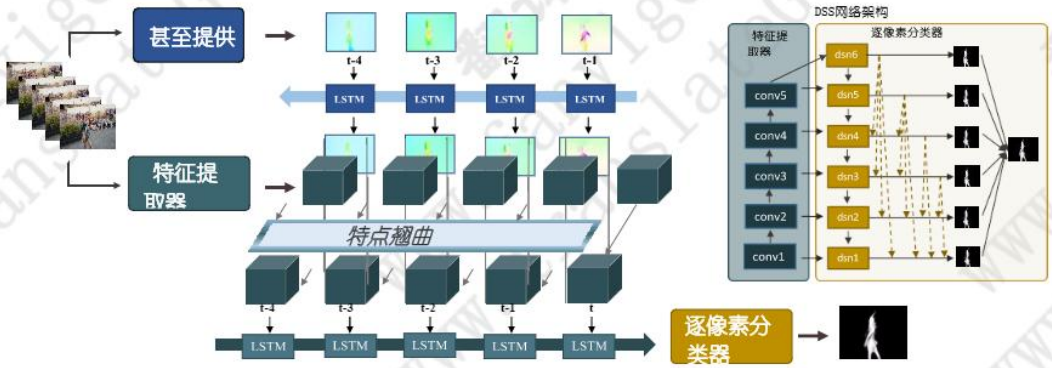

图2. 流向引导递归神经编码器的总体框架。 它包含一个LSTM，带有用于运动流更新的反向顺序输入，一个流向导特征变形模块和另一个用于时间相干特征编码的LSTM。

相干特征编码，产生结果特征F_i。 输出显着图因此被计算为$S_i = N_{REG}(F_i)$。

3.1. 运动计算和更新

给定参考I_i和k个前帧的窗口，我们首先应用嵌入式FlowNet F [7]给个人，同时估计k个初始流场$\{O_{i\to j} = F(I_i, I_j) \mid j = i-1, i-2, \ldots, i-k\}$相对于参考帧。 那里-

湍流场$O_{i\to j}$是两个通道的位置偏移图。 它计算每个像素的位移（u，v）

'中的空间位置（x′，y′），即（x′，y′）中的像素位置（x，y） u，y + v），其中u和v分别表示水平和垂直方向上的像素偏移量。

由于FlowNet最初是根据连续帧的配对数据进行训练的，因此可能不够准确，无法反映两个帧之间长时间间隔的运动关系。 直观地说，越接近参考帧，估计的运动流动越准确。 我们可以逐渐采用更接近帧的流程图来改进更大的时间间隔。 基于上述考虑，我们建议将ConvLSTM [39]与基于CNN的FlowNet共同学习流程图并按相反顺序进行细化。

ConvLSTM是传统完全连接的LSTM的延伸，它具有卷积结构

输入到状态和状态到状态的连接。 所有在ConvLSTM中传输的数据都可以看作3D张量，最后两个维度是空间维数 -
sions。 令$X_1, X_2, \ldots, X_t$表示输入到ConvLSTM

和 $H^1, H^2, \ldots, H^t$代表其隐藏状态。 在每个时间步，ConvLSTM的输出隐藏状态基于更新
根据其自己的输入以及来自之前输入的编码的过去状态，其被表述为

$$H_t = ConvLSTM(H_{t-1}, C_{t-1}, X_t), \quad (1)$$

C
ConvLSTM在前一时间步的记忆细胞状态在哪里。 关注[39]，ConvLSTM模块由输入门i_t，忘记门f_t和输出门o_t组成，总体更新方程可以列在2），其中''表示卷积运算符，''表示Hadamard乘积，$\sigma()$表示sigmoid函数：

$$i_t = \sigma(W_{xi} * X_t + W_{Hi} * H_{t-1} + W_{Ci} C_{t-1} + b_i)$$
$$f_t = \sigma(W_{xf} * X_t + W_{Hf} * H_{t-1} + W_{比照} C_{t-1} + b_f)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{Hc} * H_{t-1} + b_c)$$

$$O_t = \sigma(W_{xo} \times X_t + W_{洁} * H_{t-1} + W_{合作} O_{t-1} + b_o)$$
$$H_t = o_t \tanh(C_t) \quad (2)$$

为了用ConvLSTM更新光学流场，
LSTM层展开为k个流场的窗口，隐藏状态的大小设置为与内部相同，
放置流程图。 我们按照相反的顺序，即$X_{1:k} = O_{i,(i-1)}, O_{i,(i-2)}, \ldots, O_{i,(i-k)}$顺序地将k个初始运动流馈送到ConvLSTM单元。 隐藏的状态是
对更新的流场进行编码，该流场被进一步馈送到具有1×1的核心尺寸的卷积层以产生经细化的流程图$RO_{i\to j}$，表达为：

$$j = i - t$$
$$H_t = ConvLSTM(H_{t-1}, C_{t-1}, O_{i\to j}) \quad (3)$$
$$RO_{i\to j} = Conv_{1\times 1}(H_t)$$

3.2. 运动引导特征翘曲

由于[45]，给定精确的流程图$RO_{i\to j}$，通过应用以下翘曲函数将$j^{自}$帧上的特征映射$N_{FEA}(I_j)$翘曲到参考帧，

$$WarpF_{j\to i} = W(N_{FEA}(I_j), RO_{i\to j}) \quad (4)$$

$\mathrm{WarpF}_{i\ j}$ 引用从帧j到帧i扭曲的特征映射。 ( ) 是双线性翘曲函数 W ，应用于特征地图中每个通道的所有空间位置。 它被实现为双线性插值 - （I_j）在期望位置与光流$RO_{i\ j}$相关。

3.3. 时间相干特征编码

尽管特征变形操作可以补偿由物体或相机移动引起的特征未对齐。 表征视频帧的连续动态演化以及时域中的外观对比度的演变还不足以表征。 基于上述考虑，我们提出利用另一个ConvLSTM进行顺序特征编码。 具体来说，这个ConvLSTM需要一系列的变形特征（包括参考框架的特征）作为输入，即，$X_{1:k}$在等式（1） 被设定为 $X_{1:k}=\mathrm{WarpF}_{i\ k}(N_{i\ k})$，$\mathrm{WarpF}_{i\ (i\ k-1)}$，...，$\mathrm{WarpF}_{i\ (i\ 1)}$，$N_{FEA}$（$I_i$），并且通过计算前向隐藏序列时间特征编码从t = 1到t = k + 1，然后更新输出层。状态更新函数在（1）可以重写如下：

$$j = i - k + (t - 1) \qquad (\quad)$$

$$H_t = \mathrm{ConvLSTM}\ H_{t-1},\ C_{t-1},\ \mathrm{WarpF}_{i\to j},\ t \leqslant k \qquad (5)$$

$$H_{k+1} = \mathrm{ConvLSTM}\ (H_k,\ C_k,\ N_{FEA}\ (I_i))$$

隐藏的状态是迄今为止记住的未来的编码。 最后一步k + 1的隐藏状态是我们的最终特征编码。

4. 实验结果

4.1. 实验装置

4.1.1 数据集

我们评估我们的方法在两个公共数据集上的性能：弗莱堡 - 伯克利运动分割（FBMS）数据集[2, 25]和DAVIS [27]数据集。 FBMS数据集包含59个视频，其中包含720个注释稀疏注释帧。 DAVIS是一个新开发的视频对象分割数据集，它包含50个高质量和全高清视频序列，其中包含3,455个密集注释的像素级和每帧地面实况。 它是涵盖各种视频对象分割挑战（如遮挡，运动模糊和外观变化）的最具挑战性的基准之一。

存在另一个数据集SegTrack V2，它是来自原始SegTrack数据集的扩展数据集，30]，并包含有关鸟类，动物，汽车和人类的14个视频，其中包含1,066个密集注释的帧图像。 如提到的[36]，我们将整个SegTrackV2，FBMS和DAVIS的训练集合作为我们的训练集，以及在DAVIS和FBMS的测试集上评估我们的训练模型。

4.1.2 评价标准

与基于图像的显着物体检测类似，我们采用精确召回曲线（PR），最大F-measure和平均绝对误差（MAE）作为评估指标。 连续显着图将重新缩放到[0,255]，并使用区间内的所有整数阈值进行二值化。 在每个阈值处，通过比较二元显着性图和地面实际值可以获得一对精度和召回值。 PR曲线是根据数据集中所有图像的显着性图的平均精度和查全率获得的。 F-measure被定义为

$$F_\beta = \frac{(1+\beta^2)\cdot Precision\cdot 召回}{\beta^2 \bullet Precision + 召回}, \qquad (6)$$

其中β^2设定为0.3，如[1]。 我们报告从PR曲线计算出的最大F-measure（maxF）。 MAE被定义为平均像素绝对差异在二元地面真值G和显着图之间S [26]，

4.1.3 实施细节

我们提出的FRGNE已经在Mxnet上实施[4]，一个灵活的开源深度学习框架网络。 FGRNE兼容任何基于FCN的静止图像显着物体检测器。 在本文中，我们选择最先进的深度监督显着物体检测（DSS）[10]方法作为基线，并将嵌入了FGRNE的更新后的DSS作为视频显着物体检测的最终模型，并与其他基准进行比较。 在Section 4.3中，我们将在其他主机网络上列出我们提出的FGRNE的更多结果，以证明我们提出的算法的有效性。在训练期间，帧图像在馈送到网络之前被调整为256 * 512。 在推断时，我们将图像调整为256像素的较短边。我们使用SGD以0.9的动量对采用端到端模式的框架中的所有组件进行了培训。 学习率初始设置为2.5e-4，每8k轮训练时减少0.9。 丢失功能设置为与主机网络相同（例如，DSS [10]采用图像级别的平衡交叉熵损失）。 窗口大小k由内存限制，其默认值在我们的实验中设置为5。 我们还探讨了部分中不同设置的影响4.3。 实验在具有NVIDIA Titan X GPU和3.4GHz Intel处理器的工作站上执行。

DATASET	公	MST	MB+	RFCN	DHSNet	DCL	DSS	SAG	GF	DLVSD	FGRNE
	maxF	0.455	0.520	0.732	**0.778**	0.740	**0.775**	0.528	0.628	0.699	**0.798**
	MAE	0.165	0.183	**0.047**	**0.035**	0.061	**0.047**	0.080	0.067	0.064	**0.032**
	maxF	0.540	0.525	0.741	**0.744**	0.740	**0.760**	0.572	0.607	0.696	**0.783**
	MAE	0.179	0.204	0.089	**0.076**	0.133	**0.077**	0.145	0.101	**0.077**	**0.063**

表1. 定量结果的比较，包括最大F-measure（越大越好）和MAE（越小越好）。 最好的三个结果分别以红色，蓝色和绿色显示。

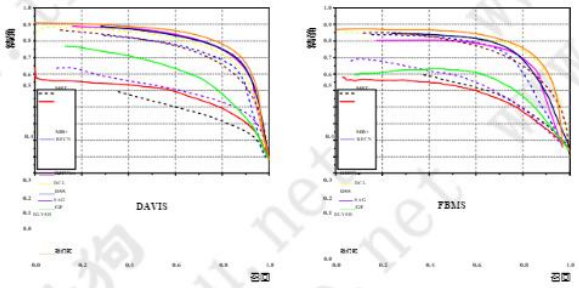

图3. DAVIS和FBMS上10种显着性检测方法的精度回忆曲线比较。 我们提出的FGRNE在两个测试数据集中始终优于其他方法。

4.2. 与艺术水平的比较

我们将我们的方法（FGRNE）与最近的9种最先进的方法（包括MST [31]，MB + [41]，RFCN [33]，DHSNet [23]，DCL [20]，DSS [10]，SAG [34]，GF [35]和DLVSD [36]。 前六个是用于静态图像的最新状态显着物体检测方法，而最后三个是基于视频的显着性模型。 为了公平比较，我们使用作者提供的实现或显着图。 我们还使用训练集对训练我们的FGRNE的所有公共静态显着性模型进行微调，并使用精炼模型进行比较。

视觉比较如图1所示。 4。 可以看出，基于深度学习的静态显着性模型在独立观看时可以生成看起来很有前途的显着性地图，当将整个序列放入时，它们并不令人惊讶地不一致。 虽然现有的基于视频的模型可以在对象运动相对较轻的视频上产生一致的结果，但它们仍然无法处理外观（对象或相机运动）发生显着变化的视频。 特别值得注意的是，我们提出的方法结合了现成的DSS [10]模型作为我们的基线，它可以学习如何改善具有时间相干性的原始特征，并最终产生比原始特征更好的优化结果。 一般来说，我们的方法在各种具有挑战性的情况下生成更加准确和一致的显着图。

作为定量评估的一部分，我们展示了图2中PR曲线的比较。 3。 如图所示，我们的方法（FGRNE）明显优于DAVIS和FBMS上的所有现有静态和动态显着物体检测算法。 此外，表中列出了最大F-measure和MAE的定量比较。 1.

我们提出的方法在FBMS和DAVIS上分别提高了最佳执行静态算法达到的最大F测量值5.24%和2.57%

MAE分别下降17.10%和8.57%。 什么时候

与表现最佳的基于视频的模型相比，我们的FGRNE将最大F-measure提高了12.50%，并且

FBMS和DAVIS数据集分别为14.16%，和

相应地，MAE降低18.18%和50%。 一个有趣的现象是当前最好的静态显着性模型实际上超越了基于状态视频的显着物体检测方法，因为它具有出色的完全卷积网络。

方法	S_p	S_s	S_c	S_t	循环	S_g
功能聚合？		✓		✓		
流动引导特征翘曲？				✓	✓	✓
用LSTM更新流程？						✓
使用LSTM进行功能编码？			✓		✓	✓
maxF	0.775	0.768	0.777	0.780	0.793	**0.798**
MAE	0.047	0.052	0.036	0.036	0.035	**0.032**
运行时间（毫秒）	97	112	137	162	184	191

表2. 流引导回归神经编码器的有效性。

4.3. 消融研究

4.3.1 流动引导递归神经编码器的有效性

正如章节中所讨论的 3我们提出的FGRNE涉及三个主要模块，包括运动流更新，运动引导特征翘曲和时间相干特征编码。 为了验证这三个模块各自的有效性和必要性，我们将FGRNE与表格中的五个变体进行比较。 2.

S_s是指从单帧基线模型生成的显着图。 为了便于比较，我们还使用我们使用的训练集的各个框架对模型进行微调。 它达到最大F_β= 0.775和MAE = 在DAVIS测试集中为0.047，这已经超越了大多数最先进的方法。 这表明微调的基线模型具有竞争力，并可作为评估的有效参考。 与我们的整个框架相比，结果显示，将FGRNE嵌入基线模型完全导致2.97％的F-measure增加，同时将MAE降低31.91％。

S_p是指基线模型上的朴素特征聚合算法。 参考帧的特征仅作为特征图的加权和更新

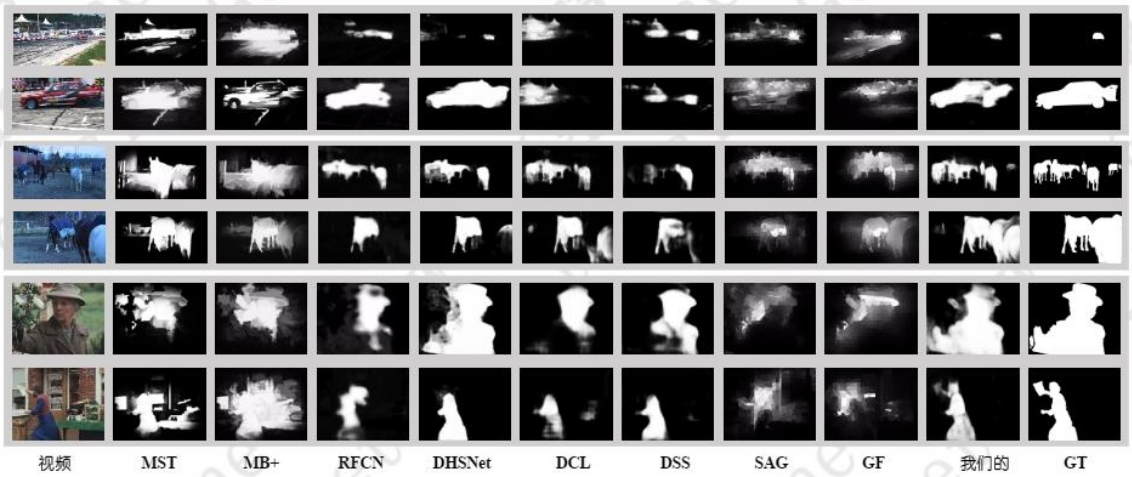

图4. 从最先进的方法（包括我们的FGRNE）生成的显着图的视觉比较。 地面实况（GT）显示在最后一列。 我们的模型始终生成最接近实际情况的显着图。

观看窗口，j帧w的权重设置为$\frac{1}{(F_i-F_j)^2+1}$。它表示为$F_i=\sum_{l=0}^{k}\frac{1}{l+1}N_{FEA}^{i\to j}$

与我们培训FGRNE一样，它也是端对端培训。 如表中所示，这种变体的F值下降到0.768，而MAE增加到0.052，这甚至对基线模型不利。 它表明这种天真的特征聚合不适合顺序特征建模。 我们推测其原因在于由于场景结构和外观的变化导致的特征不一致。

S_c是指在基线模型上的简单特征编码算法和FGRNE的简并变体。 运动更新模块关闭并且不使用流动运动，即运动流程$O_{i\to i}$在训练期间设置为全零。 该变体也以与FGRNE相同的方式进行端对端培训。 如表中所示，F-measure轻微上升至0.777，而MAE则大幅下跌23.40％至0.036。 但是，性能仍然比拟议的FGRNE差得多。这表明回归神经编码器可以学习利用先前帧的特征来改善参考帧的时间相干性。 但是，仅基于LSTM的特征编码是不够的。

S_a将运动引导特征变形添加到S_b的模型中，而不打开运动演变更新模块。 它实际上是一个流向导特征聚合程序。 它将F-measure增加1.56％至0.780，而MAE降低30.77％至S_b的表现0.036。 这意味着功能对齐是功能聚合之前的一项重要操作。 S_a的性能明显提高，也揭示了视频显着物体检测的运动建模的重要性。

S_r将运动引导特征变形添加到模型中S_c。 它是没有运动流更新的FGRNE的退化版本。 所有其他因素保持不变。 它在 -
使最大F值下降2.06％至0.793并下降

MAE减少2.78％至0.035与S_r的性能，这意味着运动引导特征变形的性能增益与基于LSTM的时间相干性建模是互补的。 事实上，物体运动和其外观对比度的变化是视频显着性的两个核心影响因素，这与我们提出的FGRNE中两个互补模块的设计完全一致。

S_f是指提出的FGRNE方法，它打开S_r中的运动流进化更新模块。 此外，F-指数上涨0.63％至0.798，而MAE下跌8.57％至0.032。 这证明了反向LSTM可以帮助改进运动流，这弥补了FlowNet在估计具有大时间间隔的帧对的光流时缺乏。

此外，我们还列出了我们提出的FGRNE的每个变体的运行时成本比较。 如图所示，将FGRNE合并到静态模型中每帧需要额外的94ms。 注意到特征提取是在给定窗口中所有帧的显着性推断期间共享的，并且我们的算法以滑动窗口模式运行。 因此，扩大窗口大小不会导致时间计算成本的严重增加。

4.3.2 灵敏度特征提取选择

正如部分所述 3，我们的FGRNE依赖于预先训练的静态显着性检测器作为我们的主机网络。 主机网络分为特征提取器和像素分类模块。 原则上，它可以在任何时候分裂

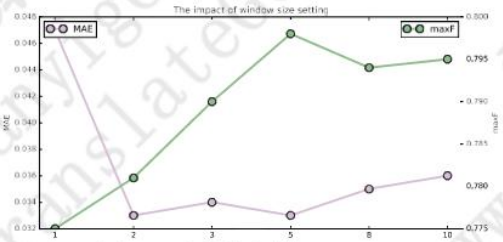

图5. 不同窗口大小设置的灵敏度分析

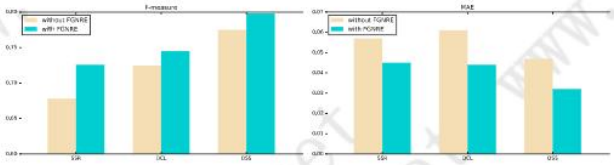

图6. 主机模型选择的灵敏度

因为主机网络是完全卷积的。 我们探索在最终结果的表现中添加FGRNE对不同级别特征提取的影响。 我们分别对主机DSS模型的Conv3 3，Conv4 3和Conv5 3的输出特征映射进行了特征编码实验。 实验结果表明，FGRNE能够提高特征地图各尺度上的时间相关性，选择Conv3 3，Conv4 3和Conv5 3的特征地图时，其最大值分别为0.777,0.79和0.798，其中包含FGRNE使用从Conv5 3中提取的特征可获得最大的性能增益，从而使F值增加2.97%，MAE减少8.57%

关于它的单帧静态版本。 -

4.3.3 对窗口大小设置的敏感度

我们提出的FGRNE学习通过利用窗口k前帧来促进编码特征的时间相关性。 受工作站内存限制，k可以设置为最大值10.我们已经探索了k = 1,2,3,5,8,10的不同设置对显着物体检测性能的影响。 结果在图。 5显示5和8个前帧的训练达到非常接近的精度，k = 5表现稍好。 默认情况下，我们在实验中训练和推理期间设置k = 5。

4.3.4 主持模型选择的敏感性

正如部分所述 3我们采用基于FCN的静态显着性检测器作为FGRNE的主机模型。 为了证明我们提出的方法可广泛应用于任何基于FCN的主机网络模型，我们申请将我们的FGRNE并入另外两个最近发布的基于FCN的显着物体检测方法，包括DCL [20]和MSRNet [18]。 对于后者，由于机器内存的限制，我们只对其单一版本即SSRNet进行实验。 如图所示。 6对F-measure和MAE的实验评估表明，我们的FGRNE可以通过训练有效地增强特征表示的时空一致性，大大提高了视频显着物体检测的性能。

DATASET	LVO	LVO+CRF	FSEG	LMP	SFL	OUS	OUS+CRF
DAVIS	70.9	75.9	70.7	70.0	67.4	73.0	77.1
FBMS	63.5	65.1	68.4	35.7	55.0	72.4	76.2

表3. 无监督视频对象分割在平均IoU方面的性能比较

5. 与无监督视频对象分割方法的比较

视频显着对象检测的问题设置与无监督视频对象分割的问题设置非常相似，除了其目标是计算每个像素的显着性概率值而不是二进制分类。 为了与最先进的无监督视频对象分割方法进行公平比较，我们将FGRNE与基于静态ResNet-101的基于像素的二进制分类模型相结合，该模型的特征从Conv5 x的最终输出特征图中提取。 我们根据平均IoU对DAVIS和FBMS数据集上的方法进行评估，并与一些最先进的方法进行比较。 如表所示 3，我们提出的方法优于LVO [29]，这是先前的技术水平，分别在DAVIS和FBMS的IoU测量上分别下降了2.96%和14.0%。 注意到如[29]，在DAVIS排行榜上报告的75.9%的mIoU值包括作为后处理的CRF，无CRF的LVO的结果是70.9，如其论文中所报道的。 为了公平比较，我们还在表格中报告了有和没有CRF的mIoU结果。 可以看出，我们用CRF提出的方法在DAVIS和FBMS上分别大大超过LVO 1.6%和16.90%。

6. 结论

在本文中，我们提出了一个视频显着物体检测的准确和端到端的框架。 我们提出的流动引导循环编码器旨在改善深度特征表示的时间相干性。 它可以被认为是一个通用的框架，将任何基于FCN的静态显着性检测器扩展到视频显着物体检测，并且可以很容易地从基于图像的显着物体检测方法的未来改进中受益。 此外，由于我们专注于学习增强功能编码，因此它可以轻松扩展到其他视频分析应用，值得在未来进行探索。

参考

[1] R. Achanta, S. Hemami, F. Estrada和S. Susstrunk。 频率调谐显着区域检测。 在CVPR，第1597-1604页。 IEEE，2009。 5

[2] T. Brox和J. Malik。 通过点轨迹的长期分析进行对象分割。 ECCV，第282-295页，2010。 5

[3] C. Chen, S. Li, Y. Wang, H. Qin, and A. Hao。 通过空间 - 时间融合和低秩相干扩散进行视频显着性检测。 TIP，26（7）：3156-3170,2017。 1,2,3

[4] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao,
B. Xu, C. Zhang和Z. Zhang。 Mxnet：适用于异构分布式系统的灵活且高效的机器学习库。 arXiv预印本arXiv：1512.01274,2015。 5

[5] M.-M. Cheng, NJ Mitra, X. Huang, PH Torr和S.-M. 胡。 基于全局对比度的显着区域检测。 TPAMI，37（3）：569-582,2015。 2

[6] Y. Fang, Z. Wang, W. Lin和Z. Fang。 视频显着性结合了时空线索和不确定性加权。
ITIP，23（9）：3910-3921，2014。 3

[7] P. Fischer, A. Dosovitskiy, E. Ilg, P.Haüsser, C.Hazırba, V. Golkov, P. van der Smagt, D. Cremers和T. Brox。 Flownet：利用卷积网络学习光流。 arXiv预印本arXiv：1504.06852,2015。 2,3,4

[8] D. Gao和N. Vasconcelos。 自下而上的显着性是一个判别过程。 在ICCV中，第1-6页。 IEEE，2007。 2

[9] C. Guo, Q. Ma和L. Zhang。 利用四元数傅里叶变换相位谱的时空显着性检测。 在CVPR中，1-8页。 IEEE，2008。 3

[10] Q. Hou, M.-M. Cheng, X.-W. 胡, A. Borji, Z. Tu和P. Torr。 深度监控短连接的显着物体检测。 arXiv预印本arXiv：1611.04849,2016。 1,2,3,5,6

[11] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and
T. Brox。 Flownet 2.0：使用深度网络进行光流估计的演变。 arXiv预印本arXiv：1612.01925,2016。 3

[12] L. Itti。 使用视觉注意力的神经生物学模型进行视频压缩自动化。 TIP，13（10）：1304-1318,2004。 1

[13] L. Itti, C. Koch和E. Niebur。 用于快速场景分析的基于显着性的视觉注意模型。 TPAMI，20（11）：1254-1259,1998。 1

[14] Y. Jia和M. Han。 与类别无关的对象级显着性检测。 在ICCV，2013年第1761-1768页。 2

[15] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li。 突出物体检测：区分性区域特征整合方法。 在CVPR，第2083-2090页，2013年。 2

[16] T.-N. Le和A. Sugimoto。 利用时空深度特征进行视频显着物体检测。 arXiv预印本arXiv：1708.01447,2017。 2,3

[17] G. Lee, Y.-W. Tai和J. Kim。 编码的低级别距离图和高级特征的深度显着性。 在CVPR，第660-668页，2016年。 2

[18] G. Li, Y. Xie, L. Lin, and Y. Yu。 实例级显着对象分割。 arXiv预印本arXiv：1704.03604,2017。 1,2,3,8

[19] G. Li和Y. Yu。 基于多尺度深度特征的视觉显着性。 在CVPR，第5455-5463页，2015年。 2

[20] G. Li和Y. Yu。 用于显着物体检测的深度对比学习。 在CVPR，第478-487页，2016年。 1,2,3,6,8

[21] G. Li和Y. Yu。 基于多尺度深cnn特征的视觉显着性检测。 TIP，25（11）：5012-5024,2016。 1

[22] Y.Li, X. Hou, C.Koch, JM Rehg和ALYuille。 显着物体分割的秘密。 在CVPR，第280-287页，2014年。 2

[23] N. Liu和J. Han。 Dhsnet：用于显着物体检测的深层次显着网络。 在CVPR，第678-686页，2016年。 3,6

[24] V. Mahadevan和N. Vasconcelos。 动态场景中的时空显着性。 TPAMI，32（1）：171-177，2010。 3

[25] P. Ochs, J. Malik和T. Brox。 通过长期视频分析对移动物体进行分割。 TPAMI，36（6）：1187-1200，2014。 5

[26] F. Perazzi, P. Krahenbuhl, Y. Pritch和A. Hornung。 显着性过滤器：用于显着区域检测的基于对比度的过滤。 在CVPR，2012年。 5

[27] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross和A. Sorkine-Hornung。 视频对象分割的基准数据集和评估方法。
在CVPR，第724-732页，2016年。 5

[28] A. Ranjan和MJ Black。 使用空间金字塔网络进行光流估计。 arXiv预印本arXiv：1611.00850,2016。 3

[29] P. Tokmakov, K. Alahari和C. Schmid。 使用视觉记忆学习视频对象分割。 arXiv预印本arXiv：1704.05737,2017。 3,8

[30] D. Tsai, M. Flagg, A. Nakazawa和JM Rehg。 使用多标签mrf优化的运动一致性跟踪。 IJCV，100（2）：190-202,2012。 5

[31] 厕所。 Tu, S.He, Q.Yang和S.-Y. 简。 使用最小生成树的实时显着对象检测。 在CVPR，第2334-2342页，2016年。 6

[32] L. Wang, H. Lu, X. Ruan和M.-H. 杨。 深度网络通过局部估计和全局搜索进行显着性检测。
在CVPR，第3183-3192页，2015年。 2

[33] L. Wang, L. Wang, H. Lu, P. Zhang和X. Ruan。 循环完全卷积网络的显着性检测。 在ECCV，第825-841页。 斯普林格，2016年。 2,3,6

[34] W. Wang, J. Shen和F. Porikli。 显着性感知测地视频对象分割。 在CVPR，第3395-3402页，2015。 6

[35] W. Wang, J. Shen和L. Shao。 使用局部梯度流优化和全局细化的一致视频显着性。
ITIP，24（11）：4185-4196,2015。 1,2,3,6

[36] W. Wang, J. Shen和L. Shao。 全卷积网络视频显着物体检测。 TIP，27（1）：38-49，2018。 1,2,5,6

[37] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao和S. Yan。 Stc：一种简单到复杂的弱监督语义分割框架。 TPAMI，39（11）：2314-2320，2017。 1

[38] H. Wu, G. Li和X. Luo。 用于概率性对象跟踪的加权注意块。 Visual Computer，30（2）：229-243,2014。 1

[39] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong和W.-c. 鸣。 卷积lstm网络：用于降水预报的机器学习方法。 在NIPS，第802-810页，2015年。 4

[40] J. Yang和M.-H. 杨。 通过联合crf和字典学习自上而下的视觉显着性。 在CVPR，第2296-2303页。 IEEE，2012。 2

[41] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price和R. Mech。 在80 fps下的最小障碍物显着物体检测。 在ICCV，第1404-1412页，2015年。 6

[42] R. Zhao, W. Ouyang, H. Li, and X. Wang。 多情境深度学习的显着性检测。 在CVPR，第1265-1274页，2015年。 2

[43] R. Zhao, W. Ouyang和X. Wang。 无人监督显着学习重新识别人。 在CVPR，第3586-3593,2013页。 1

[44] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei。 视频对象检测的流引导功能聚合。 arXiv预印本arXiv:1703.10025，2017。 3

[45] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei。 用于视频识别的深度特征流。 arXiv预印本arXiv:1611.07715,2016。 3,4