一种基于全双工策略的视频目标分割方法

季葛鹏 [1,2] 傅可人 [3] 吴哲 [4] 范登平 [1,*] 沈建冰 [5] 邵岭 [1]

[1] 起源人工智能研究院（IIAI） [2] 武汉大学计算机学院 [3] 四川大学计算机学院 [4] 鹏城实验室 [5] 澳门大学计算机与信息学院

http://dpfan.net/FSNet/

Abstract

表观信息与运动信息是视频目标分割 *(VOS)* 中相当重要的两个信息源. 先前的方法着重于使用基于单工策略的解决方案，这类方法降低了特征线索内和特征线索间交互能力的最大上限. 本文提出了一种名为全双工策略网络（***FSNet***）的新颖架构，它设计了一个关系交叉注意力模块（*RCAM*），用来实现在嵌入子空间中的双向信息传递. 引入的双向提纯模块（*BPM*），进一步用来更新时空嵌入特征的不一致性，有效地提升了模型的鲁棒性. 通过在全双工策略中考虑**相互约束**，本文的 *FSNet* 在特征融合和特征解码阶段前，可以同时执行跨膜态的特征传递（即：传输和接收），使其对视频目标分割中各种高难度场景（如：动态模糊、遮挡）具有更高的鲁棒性. 在五个通用评价基准（即：$DAVIS_{16}$、*FBMS*、*MCL*、*SegTrack-V2* 以及 $DAVSOD_{19}$）上的实验结果表明，*FSNet* 在视频目标分割和视频显著目标检测任务中胜过当前最前沿的方法.

1. 介绍

视频目标分割任务 [12,31,101,104] 是计算机视觉中智能分析视频的一个基础研究课题，目标是在像素级别上描述每一帧中移动的目标[1]. 该任务已经在机器人操纵 [1]、自动驾驶 [58]、视频剪辑 [33]、医疗 [35]、光流估计 [17]、交互式分割 [9,28,60]、无监督多目标视频分割 [75] 和视频描述 [65] 等领域中被广泛应用. 根据是否在输入时的第一帧人为提供对应的目标信息，视频目标分割任务分为两类设定即：半监督 [95] 视频目标分割和无监督 [59] 视频目标分割. 本文针对无监督的环境设定（即：零次学习的视频目标分割 [123, 124]）. 对于半监督视频目标分割任务，建议读者参考阅读以下相关文献 [5,8,43,53,73,76,112,114,118,119].

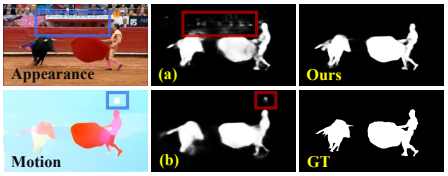

图 1: 单工策略（即：(a) 表观优化运动或者 (b) 运动优化表观）和本文**双工策略**的视觉对比图。相比之下，本文的 *FSNet* 提供一种交互的方式来利用在全双工策略**相互约束**下的表观和运动信息，从而提供了更准确的结构细节并缓解了短期的特征漂移 [115].

近年来，在视频内容理解上涌现了诸多使用视频帧间关联的表观信息（如：颜色帧 [117]）和运动信息（如：光流 [32,83] 和像素轨迹 [78]）的方法. 然而，短期的依赖估计（如：单步运动信息 [32,83]）产生了不可靠的预测结果，招致了一些通病 [29]（如：扩散，噪音和形变），这使得基于表观信息的模型（如：循环神经网络 [59,85]）的表现结果受到了模糊前景或杂乱背景的严重影响. 随着时空嵌入的传播，这些冲突容易使误差积累，从而形成短期内的特征漂移问题 [115].

*通讯作者：范登平 *(dengpfan@gmail.com)*. 本工作是季葛鹏在 IIAI 实习期间，由范登平研究员指导下完成的.

[1]后续章节中将等价地使用 '前景物体' 和 '目标物体'.

本文为 ICCV2021 论文 [36] 的中文翻译版.

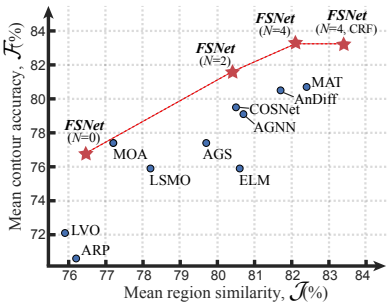

图 2: 平均边缘准确率分数 ($\mathcal{F}$) 与平均区域相似度分数 ($\mathcal{J}$) 在 DAVIS$_{16}$ [71] 数据集上的对比. 圆圈代表无监督 VOS 方法. 本文 *FSNet* 的四种变体模型使用**加粗斜体**表示，其中 'N' 表示双向提纯模块的数量. 与最佳的无监督 VOS 模型 (即：经 CRF 处理后的 MAT [124]) 相比，本文的 *FSNet* (N=4, CRF) 模型以很大的优势取得了新的最优性能.

早期的解决方案使用了*无方向约束的策略* [15, 34, 38, 85, 108]，其对表观及运动特征进行独立编码后直接执行融合. 然而，由于运动和表观特征是从分离分支中获取的两种独立模态，这种隐式的建模策略会导致特征上的冲突. 一个合理的想法是使用引导的方式去融合两者. 因此，近来的诸多方法采用了*单工策略* [29, 50, 54, 62, 68, 88, 124]，这类模型是基于表观信息或者基于运动信息引导的. 根据先前在认知心理学 [40, 87, 105] 和计算机视觉 [34, 93] 的相关研究可知，尽管这两个策略有着不错的进展，但它们都未能推断出在动态观察中引导人类视觉注意分配的外观线索和动作线索之间的相互约束.

对于同一个目标，本文认为其表观特性与运动特性应具有一定程度上的同质性. 直观来看，在图. 1 中，表观图（见左上）和运动图（见左下）的前景区域从本质上来看共享着与感知相关的模式，它包含了语义结构和运动姿态. 然而，在独立模态中的误导知识，如：斗牛场中的静态观众和电视中的动态水印（见蓝框）会在特征传递时产生不准确性，因而污染了最终的预测结果（见红框）.

为了缓解上述矛盾，有必要引入一种新的模态传输方案，而不是对不同模态进行独立地嵌入。受上述启发，本文借鉴了无线通讯领域的思路，提出了*全双工*[2]的方案. 如 图. 4 (c) & 图. 5 (c) 展示了跨运动信息和表观信息的双向注意力的框架，它在统一的架构中显示地融合了表观模式和运动模式. 如 图. 1 的第一行，本文的全双工策略网络 (***FSNet***) 相较于单工策略方法在视觉效果上表现更佳. 为了理解该学习策略的优异性原因，本文全面探究了单工和全双工策略的框架，并做出了以下贡献:

1. 本文强调了时空表征中全双工策略的重要性. 具体而言，使用一个名为关系交叉注意力 (RCAM) 的双向交互模块，从表观和运动分支提取具鉴别性的特征，以确保两者之间的相互约束.
2. 为了进一步提升模型的鲁棒性，本文提出了一个双向提纯模块 (BPM)，它配备了交错递减连接 (IDC) 以自动更新时空嵌入中的不一致性.
3. 本文在五个主流数据集上展现了模型的优越性能，特别是在 DAVIS$_{16}$ [71] 榜单中，本文的 *FSNet* (N=4, CRF) 在 $\mathcal{F}$ 评价指标上超越了前沿模型 MAT [124] 2.4%，而使用的训练数据更少 (本文-13K *vs.* MAT-16K).

2. 相关工作

2.1. 无监督视频目标分割

尽管出现了许多关于半监督视频目标分割的工作 [7, 14, 37, 69, 89, 107] (即假设目标物体的标注已经在第一帧给出)，其他研究人员仍在尝试更具挑战性的无监督视频目标分割任务. 早期的无监督视频目标分割模型采用具有低语义性的手工特征来进行启发式的图像分割推理，例如：长程稀疏轨迹点 [6, 22, 63, 79, 97]、候选目标 [47, 48, 57, 72]、显著性先验 [19, 92, 94]、光流 [88] 或者超像素点 [23, 24, 109]. 同样地，由于缺乏语义信息和上层内容的理解，这类传统模型的泛化性较差，导致在高度动态和复杂场景下具有较低的精确度. 近期，由于循环神经网络模型 [3, 81, 85, 99, 112, 122] 能更好的获取长程依赖并且采用了深度学习的技术，从而渐渐成为主流. 在该情况下，无监督视频目标分割任务被定义为时间维度上的循环建模问题，即在长程上下文找中协同地利用空间特征.

[2]在同个通道上，信息可以同时被接受和传递 [4].

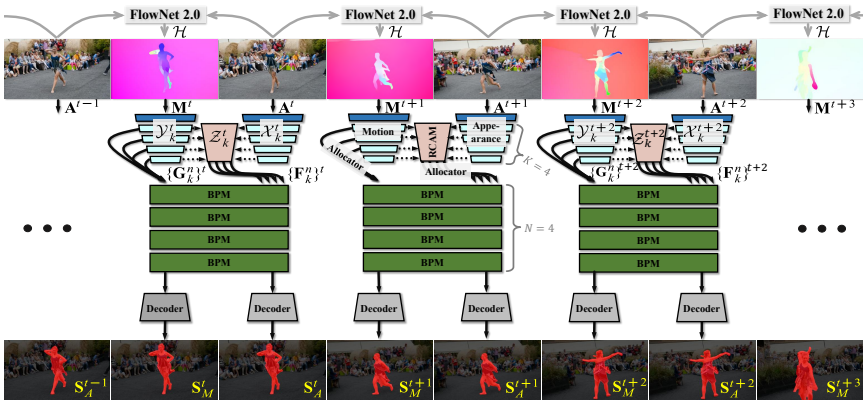

图 3: 本文的 *FSNet* 模型流程图. 关系交叉注意力模型（RCAM）使用了全双工策略在运动信息和表观信息之间提取更具鉴别性的表征. 然后堆叠了四个双向提纯模块（BMP）来进一步重新校准运动特征和表观特征之间的不一致性. 最后使用解码器来生成最后的预测图.

如何结合运动信息和表观特征是该领域长期以来尚未解决的难题。为此，Tokmakov 等人 [84] 提出了简单使用从视频获取的运动模式. 然而，由于其模型严重依赖于光流信息的引导，无法在两个相似的相邻帧中准确地分割出目标. 为了解决这个问题，许多工作 [15, 80, 85] 融合了来自平行网络中的时间特征和空间特征，也就是利用隐式建模的策略融合了来自时间分支与空间分支中的独立特征. Li 等人 [51] 提出了一个多阶段处理的方法来解决无监督视频目标分割问题，它使用了一个固定的表观信息网络来生成目标，并传入基于运动的双边预测器中来分割目标物体.

2.2. 基于注意力机制的视频目标分割

基于注意力的视频目标分割任务与无监督视频目标分割任务十分接近，因其旨在从视频片段中获取与注意力相关联的目标物体. 传统方法 [30, 98, 111, 125] 首先基于不同手工特征来计算单帧上的静态特征以及动态特征的显著度，然后进行时空优化来保持连续帧之间的一致性. 近期的工作 [45, 61, 96] 通常会以一种端到端的方式学习一个高级语义表征并执行时空检测. 许多方案都采用对时间信息考量的深度网络，如 ConvLSTM [21, 49, 81]、以光流或相邻帧作为输入 [50, 96]、3D 卷积 [45, 61] 或直接对特征进行时间维度上的拼接 [46]. 此外，长程的影响常常与深度学习结合并作为其考量因素. Li 等人 [52] 本文的关键帧策略用来定位具代表性高质量帧中的显著目标，并向较难检测的非关键帧传播其显著性. Chen 等人 [10] 利用了长程时空信息来改善显著性检测，其中 '超出当前帧范围' 的高质量帧与当前帧匹配，并且将两种信息均传入神经网络中来进行后续分类. 除了考虑如何更好地利用时间信息外，其他研究人员也尝试去处理视频显著性目标检测中的各种问题，例如减少标注数据量的需求 [113]、开发半监督的方法 [82] 或者研究相对显著性 [102]. Fan 等人 [21] 提出了一个基于显著性转移的 ConvLSTM 并用于视频显著性目标检测，以及一个带有高质量数据标注的注意力一致数据集.

3. 方法

3.1. 概要

假设一段视频包含 T 个连续帧 $\{\mathbf{A}^t\}_{t=1}^{T}$. 首先使用基于光流的生成器 $\mathcal{H}$（即：FlowNet 2.0 [32]）生成 $T-1$ 个光流图 $\{\mathbf{M}^t\}_{t=1}^{T-1}$，其通过两个相邻帧计算所得（即：$\mathbf{M}^t = \mathcal{H}[\mathbf{A}^t, \mathbf{A}^{t+1}]$）. 本方案在整体流程中舍弃了最后一帧以确保输入数量匹配. 因此，本文的流程以表观图片 $\{\mathbf{A}^t\}_{t=1}^{T-1}$ 和与其匹配的光流图片 $\{\mathbf{M}^t\}_{t=1}^{T-1}$ 作为输入. 首先，$\mathbf{M}^t$ 和 $\mathbf{A}^t$ 成对输入两个独立的 ResNet-50 [27] 分支中（即：图. 3 中的运动块和表观块）. 从 K 层中获得的表观特征 $\{\mathcal{X}_k\}_{k=1}^{K}$ 与运动特征 $\{\mathcal{Y}_k\}_{k=1}^{K}$ 被传入关系交叉注意力模块（RCAM），从而让网络能够嵌入基于时空的

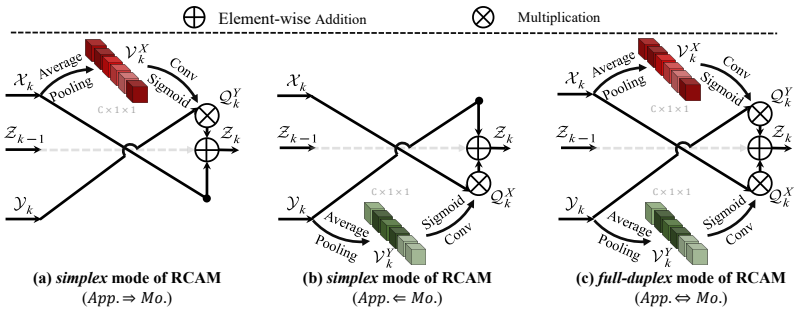

图 4: 关系交叉注意力模块（RCAM）和 单工（a & b）及全双工（c）策略的图示

交叉模态特征. 接着使用 N 个叠加的双向提纯模块（BPMs）对混合特征 $\{\mathbf{F}_k^n\}_{n=1}^N$ 和运动特征 $\{\mathbf{G}_k^n\}_{n=1}^N$ 进行提纯，以获取具有鉴别性的信息. 最后，采用两个解码器 $\mathbf{S}_M^t$ 和 $\mathbf{S}_A^t$ 生成最终的预测图.

3.2. 关系交叉注意力模块

如 § 1 中所述，单一模态（即：运动或表观）所引导的刺激信号可能会导致模型做出错误的决策. 为了减缓这个情况，本方案基于通道维度的注意力机制设计了关系交叉注意力模块（RCAM），对从两个模态获得有效的压缩信息进行提纯并互相调节彼此. 如图. 4 (c) 所示，关系交叉注意力模块的两个输入 $\{\mathcal{X}_k\}_{k=1}^K$ 和 $\{\mathcal{Y}_k\}_{k=1}^K$ 是从标准 ResNet-50 [27] 网络中获得的. 具体而言，针对第 k 层，本方案对 $\mathcal{X}_k$ 和 $\mathcal{Y}_k$ 进行全局平均池化（GAP）操作来获得基于通道维度的向量 $\mathcal{V}_k^X$ 和 $\mathcal{V}_k^Y$. 接下来使用两个具有可学习参数 $\mathbf{W}_\phi$ 和 $\mathbf{W}_\theta$ 的 1×1 卷积层（即：$\phi(x;\mathbf{W}_\phi)$ 和 $\theta(x;\mathbf{W}_\theta)$）生成两个具有鉴别性的全局描述子. Sigmoid 激活函数为 $\sigma[x] = e^x/(e^x+1)$，$x \in \mathbb{R}$ 随后用于转换描述子至 $[0,1]$ 区间，即生成有效的注意力向量用于通道维度的加权. 然后对 $\mathcal{X}_k$ 与 $\sigma\left[\theta(\mathcal{V}_k^Y;\mathbf{W}_\theta)\right]$ 之间进行外积相乘 $\otimes$，用来生成候选特征 $\mathcal{Q}_k^X$，反之亦然，如下所述：

$$\mathcal{Q}_k^X = \mathcal{X}_k \otimes \sigma\left[\theta(\mathcal{V}_k^Y;\mathbf{W}_\theta)\right], \tag{1}$$

$$\mathcal{Q}_k^Y = \mathcal{Y}_k \otimes \sigma\left[\phi(\mathcal{V}_k^X;\mathbf{W}_\phi)\right]. \tag{2}$$

接着结合 $\mathcal{Q}_k^X$、$\mathcal{Q}_k^Y$ 和融合后的低层语义特征 $\mathcal{Z}_{k-1}$ 来进行深层的特征提取. 在 ResNet-50 关联的第 k 层块中 $\mathcal{B}_k[x]$ 应用了逐元素相加 $\oplus$，最终获得了包含时空关联的融合特征 $\mathcal{Z}_k$：

$$\mathcal{Z}_k = \mathcal{B}_k\left[\mathcal{Q}_k^X \oplus \mathcal{Q}_k^Y \oplus \mathcal{Z}_{k-1}\right], \tag{3}$$

其 $k\in\{1:K\}$ 表示在主干中不同层次的特征. 注意，$\mathcal{Z}_0$ 表示全零张量. 在模型实现中，文献 [103,120] 建议使用高四层的金字塔特征，即 $K=4$.

3.3. 双向提纯模块

上述关系交叉注意力模块除了用于融合跨模态特征外，本文进一步引入了双向提纯模块（BPM）来提升模型的鲁棒性. 采用动作识别 [77] 和显著性检测 [106] 中的标准做法，本文的双向提纯阶段是由 N 个双向提纯模块叠加而成. 如图. 3 所示，首先使用特征分配器 $\psi_{\{F,G\}}(x;\mathbf{W}_\psi^{\{F,G\}})$ 对来自先前阶段的特征进行统一化：

$$\mathbf{F}_k^n = \psi_F(\mathcal{Z}_k;\mathbf{W}_\psi^F),\ \mathbf{G}_k^n = \psi_G(\mathcal{Y}_k;\mathbf{W}_\psi^G), \tag{4}$$

其中，$k\in\{1:K\}$ 和 $n\in\{1:N\}$ 分别表示特征的不同层级和双向提纯模块的数量. 具体而言，$\psi_{\{F,G\}}(x;\mathbf{W}_\psi^{\{F,G\}})$ 是由两个带有 32 个 3×3 卷积核大小的卷积层所构成. 值得注意的是，特征分配器的引入有助于减少计算负担，同时有利于不同特征之间的逐像素操作.

本文考虑在双向提纯模块中使用一种双向注意力的方案（详见 图. 5 (c)），它由两个单工策略（详见 图. 5 (a & b)）所构成. 一方面，运动特征 $\mathbf{G}_k^n$ 中包含时间线索并且可以通过拼接操作来丰富融合特征 $\mathbf{F}_k^n$. 另一方面，可以通过与融合特征 $\mathbf{F}_k^n$ 相乘来抑制运动特征 $\mathbf{G}_k^n$ 中的干扰信息. 此外，为获取鲁棒的表征，本文提出一种有效的跨模态融合策略，

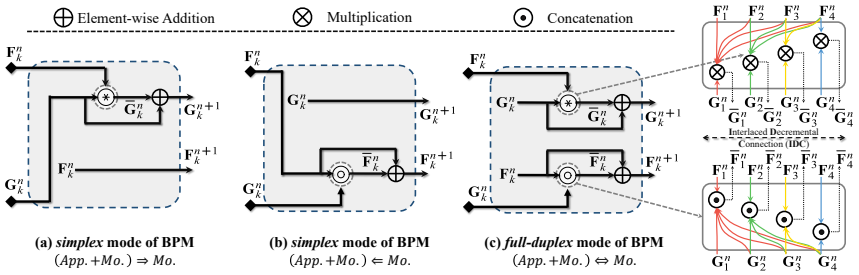

图 5: 双向提纯网络（BPM）中单工策略及全双工策略的图示.

其通过自顶向下 [55] 的交错递减连接 (IDC)，将低层次弱语义的特征传播到高层次富语义的特征之中. 具体而言，时空特征融合分支（详见 图. 5 (b)）作为其中的第一部分，可被定义为：

$$\mathbf{F}_k^{n+1} = \mathbf{F}_k^n \oplus \bigcup_{i=k}^{K} [\mathbf{F}_k^n, \mathcal{P}(\mathbf{G}_i^n)], \tag{5}$$

其中 $\mathcal{P}$ 代表接有一个 1×1 卷积层的上采样操作，用来重新缩放候选引导特征图，使其保持与 $\mathbf{F}_k^n$ 一致的分辨率. 符号 $\oplus$ 和 $\bigcup$ 分别表示在 IDC 策略中[3]逐元素相加和特征拼接操作，随后接有一个具有 32 个 1×1 卷积核大小的卷积层. 另一方面，本文将时间维度的特征重校准分支（详见 图. 5 (a)）定义为：

$$\mathbf{G}_k^{n+1} = \mathbf{G}_k^n \oplus \bigcap_{j=k}^{K} [\mathbf{G}_k^n, \mathcal{P}(\mathbf{F}_j^n)], \tag{6}$$

其中，$\bigcap$ 表示 IDC 策略中元素级别的特征相乘操作，紧随一个带有 32 个 1×1 卷积核大小的卷积层.

3.4. 解码器

在特征融合和以多层金字塔交互重新校准后，最后一个双向提纯模块产生了两组具有 32 个固定通道数的鉴别性特征（即：$\mathbf{F}_k^N$ 和 $\mathbf{G}_k^N$）. 本文在 U-Net [74] 的各个跳层连接中集成了金字塔池化模块 (PPMs) [121]，作为网络中的解码器，具体实现时，本文只采用了高四层（$K = 4$）. 由于特征从高层向低层进行融合，全局信息同时在解码器的不同尺度中得以保留：

$$\hat{\mathbf{F}}_k^N = \mathcal{C}[\mathbf{F}_k^N \odot \mathcal{UP}(\hat{\mathbf{F}}_{k+1}^N)], \tag{7}$$

$$\hat{\mathbf{G}}_k^N = \mathcal{C}[\mathbf{G}_k^N \odot \mathcal{UP}(\hat{\mathbf{G}}_{k+1}^N)]. \tag{8}$$

$\mathcal{UP}$ 在这里表示金字塔池化层和上采样操作，$\odot$ 表示特征间的拼接操作. 然后通道数经过一个卷积层 $\mathcal{C}$ 从 64 降低至 32. 最后，在上游输出（即：$\hat{\mathbf{F}}_1^N$ & $\hat{\mathbf{G}}_1^N$）后使用一个带有 32 个 1×1 卷积核大小的卷积层和一个 sigmoid 激活函数来生成视频帧在 t 时刻的预测图（即：$\mathbf{S}_A^t$ & $\mathbf{S}_M^t$）.

3.5. 训练

对于 t 时刻的视频帧给定一组预测 $\mathbf{S}^t \in \{\mathbf{S}_A^t, \mathbf{S}_M^t\}$ 和其对应的标注 $\mathbf{G}^t$，本文使用标准的二值交叉熵损失函数 $\mathcal{L}_{bce}$ 来衡量输出与目标的差异，其计算过程如下：

$$\begin{aligned}\mathcal{L}_{bce}(\mathbf{S}^t, \mathbf{G}^t) = -\sum_{(x,y)} [&\mathbf{G}^t(x,y)\log(\mathbf{S}^t(x,y)) \\ &+ (1-\mathbf{G}^t(x,y))\log(1-\mathbf{S}^t(x,y))],\end{aligned} \tag{9}$$

其中，(x, y) 表示视频帧中的位置坐标. 整体的损失函数被定义为：

$$\mathcal{L}_{total} = \mathcal{L}_{bce}(\mathbf{S}_A^t, \mathbf{G}_t) + \mathcal{L}_{bce}(\mathbf{S}_M^t, \mathbf{G}_t). \tag{10}$$

由于实验结果显示融合表观和运动线索会带来更好的表现，所以采用 $\mathbf{S}_A^t$ 作为最终的预测结果.

3.6. 实现细节

训练设定： 本方法使用 PyTorch [67] 框架实现并用 NVIDIA RTX TITAN 显卡进行加速. 所有的输入尺寸统一调整为 352×352 大小. 为了提升模型的稳定性和泛化性，在训练阶段本文采用多尺度（$\{0.75, 1, 1.25\}$）的训练策略 [26]. 根据 表. 4的实验

[3]例如在 $k = 2$, $K = 4$ 的设定下，$\mathbf{G}_2^n = \bigcup_{i=2}^{K=4}[\mathbf{F}_2^n, \mathcal{P}(\mathbf{G}_i^n)] = \mathbf{F}_2^n \odot \mathcal{P}(\mathbf{G}_2^n) \odot \mathcal{P}(\mathbf{G}_3^n) \odot \mathcal{P}(\mathbf{G}_4^n)$.

表 1: 视频目标分割任务中 DAVIS_{16} [71] 验证集的评测结果对比，包括本文的 *FSNet* 与 14 个前沿的无监督模型和 7 个半监督模型. '*w*/ Flow': 是否使用了光流算法. '*w*/ CRF': 是否使用了条件随机场算法 [42] 做后处理. 最好的分数使用**粗体**标记.

	Unsupervised																Semi-supervised						
Metric	***FSNet* (Ours)**		MAT [124]	AGNN [91]	AnDiff [115]	COSNet [56]	AGS [99]	EpO+ [18]	MOA [80]	LSMO [86]	ARP [41]	LVO [85]	LMP [84]	SFL [15]	ELM [44]	FST [66]	CFBI [116]	AGA [37]	RGM [107]	FEEL [89]	FA [14]	OS [7]	MSK [69]
w/ Flow	✔	✔	✔					✔	✔	✔	✔	✔	✔	✔	✔	✔							✔
w/ CRF	✔		✔	✔		✔	✔	✔	✔	✔		✔	✔						✔		✔		✔
Mean-$\mathcal{J}\uparrow$	**83.4**	82.1	82.4	80.7	81.7	80.5	79.7	80.6	77.2	78.2	76.2	75.9	70.0	67.4	61.8	55.8	**85.3**	81.5	81.5	81.1	82.4	79.8	79.7
Mean-$\mathcal{F}\uparrow$	**83.1**	83.0	80.7	79.1	80.5	79.5	77.4	75.5	77.4	75.9	70.6	72.1	65.9	66.7	61.2	51.1	**86.9**	82.2	82.0	82.2	79.5	80.6	75.4

表 2: 视频显著目标检测任务中多个数据集评测结果对比，包括本文的 *FSNet* 与 13 个前沿模型. '†' 代表生成的显著图不使用 CRF 后处理技术（为了评测对比公平）. 'N/A' 代表结果图无法获取.

		DAVIS_{16} [71]				MCL [39]				FBMS [64]				DAVSOD_{19}-Easy35 [21]			
	Model	$S_\alpha\uparrow$	$E_\xi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{max}\uparrow$	$F_\beta^{max}\uparrow$	$\mathcal{M}\downarrow$
2018	MBN [51]	0.887	0.966	0.862	0.031	0.755	0.858	0.698	0.119	0.857	0.892	0.816	0.047	0.646	0.694	0.506	0.109
	FGRN [49]	0.838	0.917	0.783	0.043	0.709	0.817	0.625	0.044	0.809	0.863	0.767	0.088	0.701	0.765	0.589	0.095
	SCNN [82]	0.761	0.843	0.679	0.077	0.730	0.828	0.628	0.054	0.794	0.865	0.762	0.095	0.680	0.745	0.541	0.127
	DLVS [96]	0.802	0.895	0.721	0.055	0.682	0.810	0.551	0.060	0.794	0.861	0.759	0.091	0.664	0.737	0.541	0.129
	SCOM [13]	0.814	0.874	0.746	0.055	0.569	0.704	0.422	0.204	0.794	0.873	0.797	0.079	0.603	0.669	0.473	0.219
2019∼2020	RSE [111]	0.748	0.878	0.698	0.063	0.682	0.657	0.576	0.073	0.670	0.790	0.652	0.128	0.577	0.663	0.417	0.146
	SRP [16]	0.662	0.843	0.660	0.070	0.689	0.812	0.646	0.058	0.648	0.773	0.671	0.134	0.575	0.655	0.453	0.146
	MESO [110]	0.718	0.853	0.660	0.070	0.477	0.730	0.144	0.102	0.635	0.767	0.618	0.134	0.549	0.673	0.360	0.159
	LTSI [10]	0.876	0.957	0.850	0.034	0.768	0.872	0.667	0.044	0.805	0.871	0.799	0.087	0.695	0.769	0.585	0.106
	SPD [52]	0.783	0.892	0.763	0.061	0.685	0.794	0.601	0.069	0.691	0.804	0.686	0.125	0.626	0.685	0.500	0.138
	SSAV [21]	0.893	0.948	0.861	0.028	0.819	0.889	0.773	0.026	0.879	0.926	0.865	0.040	0.755	0.806	0.659	0.084
	RCR [113]	0.886	0.947	0.848	0.027	0.820	0.895	0.742	0.028	0.872	0.905	0.859	0.053	0.741	0.803	0.653	0.087
	PCSA [25]	0.902	0.961	0.880	0.022	N/A	N/A	N/A	N/A	0.868	0.920	0.837	**0.040**	0.741	0.793	0.656	0.086
	***FSNet*†(Ours)**	**0.920**	**0.970**	**0.907**	**0.020**	**0.864**	**0.924**	**0.821**	**0.023**	**0.890**	**0.935**	**0.888**	0.041	**0.773**	**0.825**	**0.685**	**0.072**

可知 $N=4$（双向提纯模块的数量）达到了最好的效果. 本文使用随机梯度下降法（SGD）来优化整个网络，其动量设定为 0.9、学习率设定为 $2e^{-3}$、权重衰减设定为 $5e^{-4}$.

测试设定和运行时间： 对于给定帧及其对应的光流图，图像缩放为 352×352 并输入对应的分支. 与文献 [56,99,124] 相似，本方法使用随机条件场（CRF）[42] 做后处理. 不包括光流图生成和 CRF 后处理的时间，本模型的测试速度可达平均 0.08 秒/帧.

4. 实验

4.1. 无监督视频目标分割以及视频显著目标检测

数据集： 本文采用四个被广泛使用的视频目标分割数据集对本文的模型进行评估. DAVIS_{16} [71] 是其中最受欢迎的数据集，包含了 50 个高质量且密集标注的视频片段（其中 30 段用来训练、20 段用来验证）. MCL [39] 数据集包含了 9 个视频片段，主要用作测试. FBMS [64] 包含了 59 段自然场景下的视频，其中 29 段用来训练，30 段用来测试. SegTrack-V2 [48] 是最早的视频目标分割数据集之一，含有 13 个视频. 此之，DAVSOD_{19} [21] 是针对视频显著性目标检测任务而设计的. 这是最具挑战性的视觉注意力一致数据集，且包含了高质量的标注和多元化的属性标签.

评价指标： 本文采用了六个标准的评价指标：区域相似度 ($\mathcal{J}$) [71] 的平均值、边缘准确率 ($\mathcal{F}$) [71] 的平均值、结构相似度指标 (S_α, α=0.5) [11]、增强匹配指标的最大值 (E_ξ^{max}) [20]、F 指标的最大值 (F_β^{max}, β^2=0.3) [2] 和平均绝对误差 (MAE, $\mathcal{M}$) [70].

训练细节： 参照 [50] 中类似的多任务训练设定，本文的训练过程分为三个阶段：(*i*) 首先使用著名的静态显著性数据集 DUTS [90] 来训练表观分支以避免网络过拟合，该过程类似于文献 [21,81,96]，(*ii*) 接着使用生成的光流图训练运动分支，(*iii*) 最后，加载由上述两个子任务所获得的空间分支与时间分支的预训练权重，在 DAVIS_{16}（30 段视频）和 FBMS（29 段视频）上端到端地训练整个网络. 最后一个步骤大约花费 4 个小时并且在批大小为 8 的情况下 20 个周期后达到收敛状态.

测试细节： 本文在 DAVIS_{16} 的验证集（20 段视频）、FBMS 的测试集（30 段视频）、DAVSOD_{19}（35 段视频），全部的 MCL（9 段视频）和全部的 SegTrack-V2

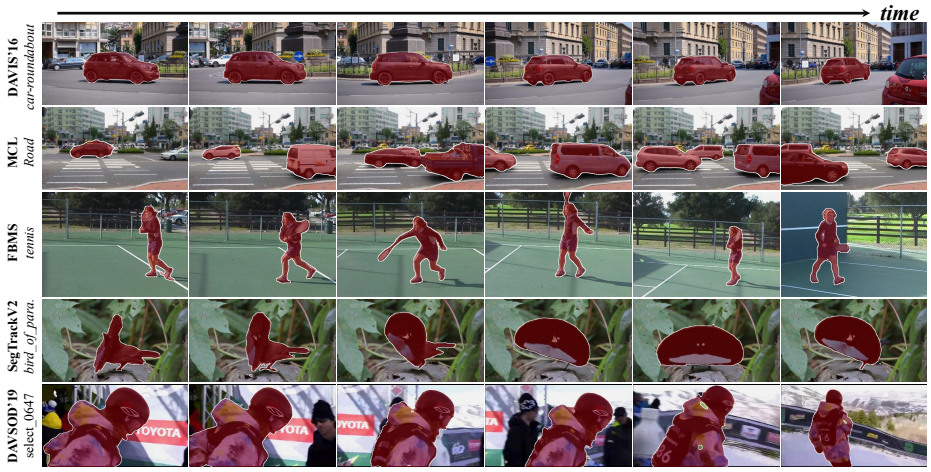

图 6: 在五个数据集上的定性对比结果，包括 DAVIS_{16} [71]、MCL [39]、FBMS [64]、SegTrack-V2 [48] 和 DAVSOD_{19} [21].

(13 段视频) 上参照标准的评价基准 [21, 71] 来测试本文模型.

DAVIS_{16} 上的评估： 如 表. 1 所示，将本文的 *FSNet* 与最前沿的 14 个无监督视频目标分割模型在 DAVIS_{16} 的排行榜上进行对比. 本文也对比了近期 7 个半监督方法. 为了对比的公平性，根据 [115] 的建议，本文也采用 0.5 作为阈值获得最终二值分割图. 本文的 *FSNet* 以 2.4% $\mathcal{F}$ 评价指标和 1.0% $\mathcal{J}$ 评价指标的差距超越了当前最佳的模型（AAAI'20-MAT [124]）. 值得注意的是，即便是在半监督模型使用了第一帧的标注信息后，本文的无监督视频目标分割模型也超越了它（CVPR'19-AGA [37]）.

本文也与当前 13 个最佳的视频显著目标检测模型经行比较. 所有的显著图[4]是从标准的评价基准 [21] 中获取的. 由 表. 2可见，本文的方法在所有的评价指标上一致地超越了 2018 年以后的所有模型. 针对 S_α 和 F_β^{max} 两个评价指标，本文的方法比最佳的 AAAI'20-PCAS [25] 模型提升了约 2.0%.

MCL 上的评估： 这个数据集的低清图像中有很多模糊边缘，这是由于目标快速移动所导致的. 因此，整体的性能表现低于 DAVIS_{16} 的结果. 如 表. 2 所示，本文的方法在如此困难的环境中仍然表现突出，与 ICCV'19-RCR [113] 和 CVPR'19-SSAV [21] 模型相比，仍有 3.0~8.0% 的提升.

[4]注意：所有比较的显著图包括本文模型，均不是二值的.

FBMS 上的评估： 这是最受欢迎的视频目标分割数据集之一，它含有多元属性。如，交互的目标和动态的背景以及非逐帧标注. 如表. 2 中所示，本文模型在 $\mathcal{M}$ 评价指标上取得了具有竞争力的表现. 此之，与先前表现最好的 SSAV [21] 模型相比，本文的模型在 $\mathcal{S}_\alpha$ (0.890 *vs.* SSAV=0.879) 和 E_ξ^{max} (0.935 *vs.* SSAV=0.926) 指标上都表现更好，使其更贴近于 [11, 20] 中所描述的人类视觉系统（HVS）模型.

SegTrack-V2 上的评估. 此为传统方法中最早的视频目标分割数据集. 因此，在此数据集上评测的深度无监督视频目标分割模型相当少. 本文仅比较最顶尖的三个模型：AAAI'20-PCAS [25] (S_α=0.866)、ICCV'19-RCR [113] (S_α=0.842) 和 CVPR'19-SSAV [21] (S_α=0.850). 本方法到达了最优异的表现 (S_α=0.870).

DAVSOD_{19} 上的评估. 大多数 DAVSOD_{19} 的视频与 DAVIS_{16} 的类似都包含了大量的单一（显著）目标. 我们发现 *FSNet* 超越了已知的所有算法. 本文的模型在 $\mathcal{S}_\alpha$ 以 3.2% 的优势大幅超越了当前最优的方法（即：AAAI'20-PCAS）.

定性结果. 在以上五个数据集的定性结果可于 图. 6 所见，验证了本文方法达到了高质量的无监督视频目标分割和视频显著性目标检测结果. 如第一行所示，因为右下方的红色车辆在边缘缓慢的移动，所以它并没有被标识出来. 然而，在全双工的策

表 3: 针对四个子模块在 DAVIS_{16} 和 MCL 数据集上的消融实验 (§ 4.2.1, § 4.2.2, & § 4.2.3). 此处设置双向提纯模块的数量为 $N = 4$.

	Component Settings				DAVIS_{16}		MCL	
	Appearance	Motion	RCAM	BPM	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
App.	✔				0.834	0.047	0.754	0.038
Mo.		✔			0.858	0.039	0.763	0.053
Vanilla	✔	✔			0.894	0.027	0.808	0.041
Rel.	✔	✔	✔		0.900	0.025	0.833	0.031
Bi-Purf.	✔	✔		✔	0.904	0.026	0.855	0.023
FSNet	✔	✔	✔	✔	**0.920**	**0.020**	**0.864**	**0.023**

表 4: 在 DAVIS_{16} [71] 和 MCL [39] 数据集上针对双向提纯模块的个数 N 所进行的消融实验，并给出了其参数量和 FLOPs 的对比.

	Param. (M)	FLOPs (G)	Runtime (s/frame)	DAVIS_{16}		MCL	
				$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
$N = 0$	0.000	0.000	0.03	0.900	0.027	0.833	0.031
$N = 2$	0.507	1.582	0.05	0.911	0.026	0.843	0.028
$N = 4$	1.015	3.163	0.08	**0.920**	**0.020**	**0.864**	**0.023**
$N = 6$	1.522	4.745	0.10	0.918	0.023	0.863	0.023
$N = 8$	2.030	6.327	0.13	0.920	0.023	0.864	0.023

略模型中同时双向地考虑了表观和运动信息，他可以自动地预测位于中间较小的车辆. 总的来说，对于这些挑战性的情况，如：动态背景（1^{st} 和 5^{th} 行），遮挡（2^{nd} 行），快速运动（3^{rd} 行），和形变（4^{th} 行），本文的模型都能够预测出正确目标中精细的细节. 从这个角度来看，展现了 *FSNet* 是个通用于无监督视频目标分割和视频显著性目标检测的模型.

4.2. 消融实验

4.2.1 模态选择

在本文的框架中探索不同刺激下（即：只使用表观 *vs.* 只使用运动）所产生的影响. 这里尝试只采用视频帧或只采用光流图（用 [32] 生成）来训练 ResNet-50 [27] 主干网络和本文的解码器（详见 § 3.4）. 如 表. 3所示，在 DAVIS_{16} 数据集中的 S_α 评价指标上，*Mo.* 结果比 *App.* 稍好，说明相比于“视频帧”，模型可以从“光流”种获取更多视觉信息. 尽管如此，在 MCL 数据集的 $\mathcal{M}$ 评价指标上 *App.* 超越了 *Mo.*. 这激发着我们去探索如何同时有效的使用表观信息和运动信息.

4.2.2 关系交叉注意力模块的有效性

为了验证关系交叉注意力模块（Rel.）的有效性，以原始的融合策略替换本文的融合策略，它由一个通道拼接操作和一个卷积操作来直接融合两个模态信息. 不出所料（表. 3），本文的 Rel. 变体在 DAVIS_{16} 和 MCL 数据集上均比原始融合策略表现的好. 在这里指出关系交叉注意力模块的两个重要特性：(i) 实现了双向修正和注意力 (ii) 在前向传播中缓解了当前帧的错误累积与放大.

表 5: 针对单工和双工策略在 DAVIS_{16} [71] 和 MCL [39] 数据集上的消融实验. 设置双向提纯模块的数量为 $N = 4$.

	Direction Setting		DAVIS_{16}		MCL	
	RCAM	BPM	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$	$\mathcal{S}_\alpha \uparrow$	$\mathcal{M} \downarrow$
simplex	$App. \Rightarrow Mo.$	$(App. + Mo.) \Rightarrow Mo.$	0.896	0.026	0.816	0.038
	$App. \Rightarrow Mo.$	$(App. + Mo.) \Leftarrow Mo.$	0.902	0.025	0.832	0.031
	$App. \Leftarrow Mo.$	$(App. + Mo.) \Rightarrow Mo.$	0.891	0.029	0.806	0.039
	$App. \Leftarrow Mo.$	$(App. + Mo.) \Leftarrow Mo.$	0.897	0.028	0.840	0.028
full-dup.	$App. \Leftrightarrow Mo.$	$(App. + Mo.) \Leftrightarrow Mo.$	**0.920**	**0.020**	**0.864**	**0.023**

4.2.3 双向提纯模块的有效性

为了阐明双向提纯模块（当 $N = 4$ 时）的有效性，本文衍生出两个不同的变体模型：Rel. 和 *FSNet*，表示框架中是/否包含双向提纯模块. 根据 表. 3数据，观察到有双向提纯模块的模型以 2.0∼3.0% 超越了没有双向提纯模块的模型. 这归因于双向提纯模块中交互递减连接的使用，有效地促进了不同信息的融合. 同样的，本文移除了关系交叉注意力模块来得到另一组设定（Vanilla 和 Bi-Purf.）来测试双向提纯模块模块的鲁棒性. 结果表明，即便使用了双向的原始融合策略（Bi-Purf.），仍能提升模型的稳定性和泛化能力. 这个特点是来自整体网络中的前向提纯操作和反向重新校准操作.

4.2.4 级联的双向提纯模块数量

直观来看，使用越多层的双向提纯模块相级联应该有更好的表现. 研究以及评测结果在表. 4展示，其 $N = \{0, 2, 4, 6, 8\}$. 注意到 $N = 0$ 表示**没有**使用双向提纯模块. 从表. 4 很明显可以观察到，越多的双向提纯模块对应着更好的预测结果，但是分数在 $N = 4$ 后达到饱和. 而且过多的双向提纯模块（即：$N > 4$）会导致更高的模型复杂度，并增加了过拟合的风险. 权衡后，实验中默认设定 $N = 4$.

4.2.5 全双工策略的有效性

为探究在全双工策略下关系交叉注意力模块和双向提纯模块的有效性，本文研究了两种无方向性（即：单工，见 图. 4 & 图. 5）的变体网络. 在 表. 5 中，

符号 $\Rightarrow$，$\Leftarrow$，和 $\Leftrightarrow$ 代表在关系交叉注意力模块和双向提纯模块中特征传递的方向. 具体地，$App. \Leftarrow Mo.$ 代表光流分支的注意力向量对表观分支产生权重上的影响，反之亦然. $(App. + Mo.) \Leftarrow Mo.$ 则表示运动信息对来自表观和运动分支的融合特征的引导. 对比结果表明所设计的模块（关系交叉注意力模块和双向提纯模块）在全双工策略中协同合作，并超越所有单工（单向）的设定.

5. 结论

本文提出了一个简单却有效的全双工策略网络（*FSNet*），它充分利用了表观信息和运动信息的优点来解决视频目标分割问题. 该架构包含了一个关系交叉注意力模块（RCAM）和一个高效的双向提纯模块. 前者用来提取来自双模态的特征，而后者用来逐步重新校准不准确的特征. 在双向提纯模块中，交错递减连接在向低语义的精细特征传播高语义的粗略特征中，担任了相当重要的角色. 本文全面验证了 *FSNet* 中的每个子模块并提供了一些有趣的发现. 最后，*FSNet* 作为统一的解决方案，显著地推动视频目标分割和视频显著性检测的进展. 如何在复杂场景中使用一个高效的类 Transformer [100, 126] 架构中学习长/短期信息似乎是一项有趣的工作.

参考文献

[1] Alexey Abramov, Karl Pauwels, Jeremie Papon, Florentin Wörgötter, and Babette Dellen. Depth-supported real-time video segmentation with the kinect. In *IEEE WACV*, pages 457–464, 2012. 1

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *IEEE CVPR*, pages 1597–1604, 2009. 6

[3] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. In *ICLR*, 2016. 2

[4] Dinesh Bharadia, Emily McMilin, and Sachin Katti. Full duplex radios. In *ACM SIGCOMM*, pages 375–386, 2013. 2

[5] Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *ECCV*, 2020. 1

[6] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *ECCV*, pages 282–295, 2010. 2

[7] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE CVPR*, pages 221–230, 2017. 2, 6

[8] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019. 1

[9] Bowen Chen, Huan Ling, Xiaohui Zeng, Gao Jun, Ziyue Xu, and Sanja Fidler. Scribblebox: Interactive annotation framework for video object segmentation. In *ECCV*, 2020. 1

[10] Chenglizhao Chen, Guotao Wang, Chong Peng, Xiaowei Zhang, and Hong Qin. Improved robust video saliency detection based on long-term spatial-temporal information. *IEEE TIP*, 29:1090–1100, 2019. 3, 6

[11] Ming-Ming Chen and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 2021. 6, 7

[12] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *IEEE CVPR*, pages 9384–9393, 2020. 1

[13] Yuhuan Chen, Wenbin Zou, Yi Tang, Xia Li, Chen Xu, and Nikos Komodakis. Scom: Spatiotemporal constrained optimization for salient object detection. *IEEE TIP*, 27(7):3345–3357, 2018. 6

[14] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *IEEE CVPR*, pages 7415–7424, 2018. 2, 6

[15] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *IEEE ICCV*, pages 686–695, 2017. 2, 3, 6

[16] Runmin Cong, Jianjun Lei, Huazhu Fu, Fatih Porikli, Qingming Huang, and Chunping Hou. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE TIP*, 28(10):4819–4831, 2019. 6

[17] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. Every Frame Counts: Joint Learning of Video Segmentation and Optical Flow. In *AAAI*, pages 10713–10720, 2020. 1

[18] Muhammad Faisal, Ijaz Akhter, Mohsen Ali, and Richard Hartley. Exploiting geometric constraints on dense trajectories for motion saliency. In *IEEE WACV*, 2020. 6

[19] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014. 2

[20] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 6, 2021. 6, 7

[21] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *IEEE CVPR*, pages 8554–8564, 2019. 3, 6, 7

[22] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi. Video segmentation by tracing discontinuities in a trajectory embedding. In *IEEE CVPR*, pages 1846–1853, 2012. 2

[23] Fabio Galasso, Roberto Cipolla, and Bernt Schiele. Video segmentation with superpixels. In *ACCV*, pages 760–774, 2012. 2

[24] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *IEEE CVPR*, pages 2141–2148, 2010. 2

[25] Yuchao Gu, Lijuan Wang, Ziqin Wang, Yun Liu, Mingming Cheng, and Shaoping Lu. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, 2020. 6, 7

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI*, 37(9):1904–1916, 2015. 5

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, pages 770–778, 2016. 3, 4, 8

[28] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *ECCV*, 2020. 1

[29] Ping Hu, Gang Wang, Xiangfei Kong, Jason Kuen, and Yap-Peng Tan. Motion-guided cascaded refinement network for video object segmentation. *IEEE TPAMI*, pages 1400–1409, 2020. 1, 2

[30] Yuan-Ting Hu, Jia-Bin Huang, and Alexander G Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*, pages 786–802, 2018. 3

[31] Xuhua Huang, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Fast video object segmentation with temporal aggregation network and dynamic template matching. In *IEEE CVPR*, pages 8879–8889, 2020. 1

[32] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE CVPR*, pages 2462–2470, 2017. 1, 3, 8

[33] Suyog Dutt Jain and Kristen Grauman. Click carving: Segmenting objects in video with point clicks. In *IJCV*, 2016. 1

[34] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *IEEE CVPR*, pages 2117–2126, 2017. 2

[35] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Debesh Jha, Huazhu Fu, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, 2021. 1

[36] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *IEEE ICCV*, 2021. 1

[37] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *IEEE CVPR*, pages 8953–8962, 2019. 2, 6, 7

[38] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for object tracking. In *IEEE CVPRW*, 2017. 2

[39] Hansang Kim, Youngbae Kim, Jae-Young Sim, and Chang-Su Kim. Spatiotemporal saliency detection for

video sequences based on random walk with restart. *IEEE TIP*, 24(8):2552–2564, 2015. 6, 7, 8

[40] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. 1987. 2

[41] Yeong Jun Koh and Chang-Su Kim. Primary object segmentation in videos based on region augmentation and reduction. In *IEEE CVPR*, pages 7417–7425, 2017. 6

[42] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011. 6

[43] Meng Lan, Yipeng Zhang, Qinning Xu, and Lefei Zhang. E3SN: Efficient End-to-End Siamese Network for Video Object Segmentation. In *IJCAI*, pages 701–707, 2020. 1

[44] Dong Lao and Ganesh Sundaramoorthi. Extending layered models to 3d motion. In *ECCV*, pages 435–451, 2018. 6

[45] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, volume 1, page 3, 2017. 3

[46] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE TIP*, 27(10):5002–5015, 2018. 3

[47] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *IEEE ICCV*, pages 1995–2002, 2011. 2

[48] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE ICCV*, pages 2192–2199, 2013. 2, 6, 7

[49] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *IEEE CVPR*, pages 3243–3252, 2018. 3, 6

[50] Haofeng Li, Guanqi Chen, Guanbin Li, and Yizhou Yu. Motion guided attention for video salient object detection. In *IEEE ICCV*, pages 7274–7283, 2019. 2, 3, 6

[51] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*, pages 207–223, 2018. 3, 6

[52] Y. Li, S. Li, C. Chen, A. Hao, and H. Qin. Accurate and robust video saliency detection via self-paced diffusion. *IEEE TMM*, 22(5):1153–1167, 2020. 3, 6

[53] Yu Li, Zhuoran Shen, and Ying Shan. Fast video object segmentation using the global context module. In *ECCV*, 2020. 1

[54] Fanqing Lin, Yao Chou, and Tony Martinez. Flow adaptive video object segmentation. *Image and Vision Computing*, 94:103864, 2020. 2

[55] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE CVPR*, pages 2117–2125, 2017. 5

[56] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In *IEEE CVPR*, pages 3623–3632, 2019. 6

[57] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *IEEE CVPR*, pages 670–677, 2012. 2

[58] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *IJRR*, 36(1):3–15, 2017. 1

[59] Sabarinath Mahadevan, Ali Athar, Aljosa Osep, Sebastian Hennen, Laura Leal-Taixé, and Bastian Leibe. Making a case for 3d convolutions for object segmentation in videos. In *BMVC*, 2020. 1

[60] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *IEEE CVPR*, pages 10366–10375, 2020. 1

[61] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *IEEE ICCV*, pages 2394–2403, 2019. 3

[62] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *IEEE CVPR*, pages 6819–6828, 2018. 2

[63] Peter Ochs and Thomas Brox. Higher order motion models and spectral clustering. In *IEEE CVPR*, pages 614–621, 2012. 2

[64] Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE TPAMI*, 36(6):1187–1200, 2013. 6, 7

[65] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *IEEE CVPR*, pages 6504–6512, 2017. 1

[66] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *IEEE ICCV*, pages 1777–1784, 2013. 6

[67] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 5

[68] Qinmu Peng and Yiu-Ming Cheung. Automatic video object segmentation based on visual and motion saliency. *IEEE TMM*, 21(12):3083–3094, 2019. 2

[69] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *IEEE CVPR*, pages 2663–2672, 2017. 2, 6

[70] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE CVPR*, pages 733–740, 2012. 6

[71] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE CVPR*, pages 724–732, 2016. 2, 6, 7, 8

[72] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *IEEE ICCV*, pages 3227–3234, 2015. 2

[73] Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. Learning fast and robust target models for video object segmentation. In *IEEE CVPR*, 2020. 1

[74] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 5

[75] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *ECCV*, 2020. 1

[76] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, 2020. 1

[77] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J Black. On the integration of optical flow and action recognition. In *GCPR*, pages 281–297, 2018. 4

[78] Jianbo Shi et al. Good features to track. In *IEEE CVPR*, pages 593–600, 1994. 1

[79] Jianbo Shi and Jitendra Malik. Motion segmentation and tracking using normalized cuts. In *IEEE ICCV*, pages 1154–1160, 1998. 2

[80] Mennatullah Siam, Chen Jiang, Steven Lu, Laura Petrich, Mahmoud Gamal, Mohamed Elhoseiny, and Martin Jagersand. Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In *IEEE ICRA*, pages 50–56, 2019. 3, 6

[81] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018. 2, 3, 6

[82] Yi Tang, Wenbin Zou, Zhi Jin, Yuhuan Chen, Yang Hua, and Xia Li. Weakly supervised salient object detection with spatiotemporal cascade neural networks. *IEEE TCSVT*, 29(7):1973–1984, 2018. 3, 6

[83] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 1

[84] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *IEEE CVPR*, pages 3386–3394, 2017. 3, 6

[85] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *IEEE ICCV*, pages 4481–4490, 2017. 1, 2, 3, 6

[86] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. *IJCV*, 127(3):282–301, 2019. 6

[87] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2

[88] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *IEEE CVPR*, pages 3899–3908, 2016. 2

[89] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *IEEE CVPR*, pages 9481–9490, 2019. 2, 6

[90] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *IEEE CVPR*, pages 136–145, 2017. 6

[91] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J. Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *IEEE ICCV*, 2019. 6

[92] Wenguan Wang, Jianbing Shen, Xuelong Li, and Fatih Porikli. Robust video object cosegmentation. *IEEE TIP*, 24(10):3137–3148, 2015. 2

[93] Wenguan Wang, Jianbing Shen, Xiankai Lu, Steven CH Hoi, and Haibin Ling. Paying attention to video object pattern understanding. *IEEE TPAMI*, 2020. 2

[94] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *IEEE CVPR*, pages 3395–3402, 2015. 2

[95] Wenguan Wang, Jianbing Shen, Fatih Porikli, and Ruigang Yang. Semi-supervised video object segmentation with super-trajectories. *IEEE TPAMI*, 41(4):985–998, 2018. 1

[96] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1):38–49, 2017. 3, 6

[97] Wenguan Wang, Jianbing Shen, Jianwen Xie, and Fatih Porikli. Super-trajectory for video segmentation. In *IEEE ICCV*, pages 1671–1679, 2017. 2

[98] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, 40(1):20–33, 2017. 3

[99] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *IEEE CVPR*, pages 3064–3074, 2019. 2, 6

[100] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *IEEE ICCV*, 2021. 9

[101] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE CVPR*, 2021. 1

[102] Zheng Wang, Xinyu Yan, Yahong Han, and Meijun Sun. Ranking video salient object detection. In *ACM MM*, pages 873–881, 2019. 3

[103] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020. 4

[104] Peisong Wen, Ruolin Yang, Qianqian Xu, Chen Qian, Qingming Huang, Runmin Cong, and Jianlou Si. DMVOS: Discriminative matching for real-time video object segmentation. In *ACM MM*, 2020. 1

[105] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *J EXP PSYCHOL HUMAN*, 15(3):419, 1989. 2

[106] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *IEEE ICCV*, pages 7264–7273, 2019. 4

[107] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *IEEE CVPR*, pages 7376–7385, 2018. 2, 6

[108] Huaxin Xiao, Bingyi Kang, Yu Liu, Maojun Zhang, and Jiashi Feng. Online meta adaptation for fast video object segmentation. *IEEE TPAMI*, 42(5):1205–1217, 2019. 2

[109] Chenliang Xu, Caiming Xiong, and Jason J Corso. Streaming hierarchical video segmentation. In *ECCV*, pages 626–639, 2012. 2

[110] Mingzhu Xu, Bing Liu, Ping Fu, Junbao Li, and Yu Hen Hu. Video saliency detection via graph clustering with motion energy and spatiotemporal objectness. *IEEE TMM*, 21(11):2790–2805, 2019. 6

[111] Mingzhu Xu, Bing Liu, Ping Fu, Junbao Li, Yu Hen Hu, and Shou Feng. Video salient object detection

via robust seeds extraction and multi-graphs manifold propagation. *IEEE TCSVT*, 2019. 3, 6

[112] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018. 1, 2

[113] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *IEEE ICCV*, pages 7284–7293, 2019. 3, 6, 7

[114] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *IEEE ICCV*, pages 5188–5197, 2019. 1

[115] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *IEEE ICCV*, pages 931–940, 2019. 1, 6, 7

[116] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. 6

[117] Lihi Zelnik-Manor and Michal Irani. Event-based analysis of video. In *IEEE CVPR*, volume 2, pages II–II, 2001. 1

[118] Kaihua Zhang, Long Wang, Dong Liu, Bo Liu, Qingshan Liu, and Zhu Li. Dual temporal memory network for efficient video object segmentation. In *ACM MM*, 2020. 1

[119] Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *IEEE CVPR*, pages 6949–6958, 2020. 1

[120] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, pages 269–284, 2018. 4

[121] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE CVPR*, pages 2881–2890, 2017. 5

[122] Jia Zheng, Weixin Luo, and Zhixin Piao. Cascaded convlstms using semantically-coherent data synthesis for video object segmentation. *IEEE Access*, 7:132120–132129, 2019. 2

[123] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE TIP*, 29:8326–8338, 2020. 1

[124] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, 2020. 1, 2, 6, 7

[125] Xiaofei Zhou, Zhi Liu, Chen Gong, and Wei Liu. Improving video saliency detection via localized estimation and spatiotemporal refinement. *IEEE TMM*, 20(11):2993–3007, 2018. 3

[126] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. Kaleido-bert: Vision-language pre-training on fashion domain. In *IEEE CVPR*, pages 12647–12657, 2021. 9