视频场景解析综述：进展、挑战与展望

谢国欢, 南开大学软件学院, 中国
SYED ARIFF SYED HESHAM, 南洋理工大学电气与电子工程学院 & 新加坡科技研究局资讯通信研究院, 新加坡
李冰, 电子科技大学信息与通信工程学院, 中国
郭文雅 and 孙国磊, 南开大学计算机学院, 中国
程明明 and 刘云*, 南开大学计算机学院，VCIP, 中国, 南开大学前沿交叉学科研究院, 中国, and 南开国际先进研究院（深圳福田), 中国

视频场景解析（Video Scene Parsing, VSP）研究像素级的视频理解：需要对每一帧中的每个像素进行分割，为每个区域赋予语义类别，并使每个目标的身份在时间上保持一致。本综述回顾了 VSP 在五项任务上的最新进展，涵盖视频语义分割（VSS）、视频实例分割（VIS）、视频全景分割（VPS）、视频跟踪与分割（VTS）以及开放词汇视频分割（OVVS）。我们将相关文献组织为一条架构演进主线：从手工设计的运动与外观线索，到全卷积、基于注意力和基于查询的设计，再到近期基于基础模型的方法；并梳理了各类方法如何建模时序上下文、保持目标身份，以及如何在精度与效率之间取得平衡。随后，我们比较了构成当前评测体系的数据集、评价指标与基准趋势。除了对方法进行归纳整理之外，我们着重阐述贯穿整个领域的设计权衡与反复出现的失效模式，即时序闪烁、遮挡导致的身份切换、长尾类别，以及标注—模型容量—延迟之间的张力。最后，我们展望了迈向鲁棒、高效且面向开放世界的 VSP 系统的未来研究方向。相关代码已公开。

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Computer vision tasks**; **Scene understanding**.

关键词: 视频场景解析，视频分割，语义分割，跟踪与分割，开放词汇分割

ACM Reference Format:
谢国欢, Syed Ariff Syed Hesham, 李冰, 郭文雅, 孙国磊, 程明明, and 刘云. 2026. 视频场景解析综述：进展、挑战与展望. *ACM Comput. Surv.* XX, XX, Article XX (August 2026), 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 引言

视频场景解析（VSP）是计算机视觉中的一个基础问题，它根据具体任务的不同，为视频像素分配语义标签、实例掩码以及持久的身份标识。VSP 涵盖视频语义分割（Video Semantic Segmentation, VSS）、视频实例分割（Video Instance Segmentation, VIS）、视频全景分割（Video Panoptic Segmentation, VPS）、视频跟踪与分割（Video Tracking & Segmentation, VTS）以及开放词汇视频分割（Open-Vocabulary Video Segmentation, OVVS）。通过将静态图像分割 [150]

*通讯作者：刘云（电子邮箱：liuyun@nankai.edu.cn）

Authors' Contact Information: 谢国欢, 南开大学软件学院, 天津, 中国, xieguohuan@mail.nankai.edu.cn; Syed Ariff Syed Hesham, 南洋理工大学电气与电子工程学院 & 新加坡科技研究局资讯通信研究院, 新加坡, syedhesh002@e.ntu.edu.sg; 李冰, 电子科技大学信息与通信工程学院, 成都, 中国, windof47@gmail.com; 郭文雅, guowenya@dbis.nankai.edu.cn; 孙国磊, guolei.sun@nankai.edu.cn, 南开大学计算机学院, 天津, 中国; 程明明, cmm@nankai.edu.cn; 刘云, liuyun@nankai.edu.cn, 南开大学计算机学院，VCIP, 天津, 中国 and 南开大学前沿交叉学科研究院, 天津, 中国 and 南开国际先进研究院（深圳福田）, 深圳, 中国.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1557-7341/2026/8-ARTXX
https://doi.org/XXXXXXX.XXXXXXX

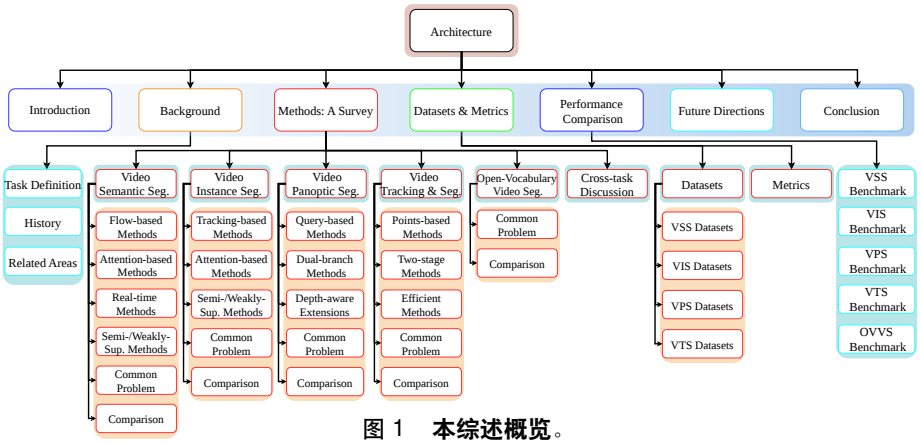


图 1　本综述概览。

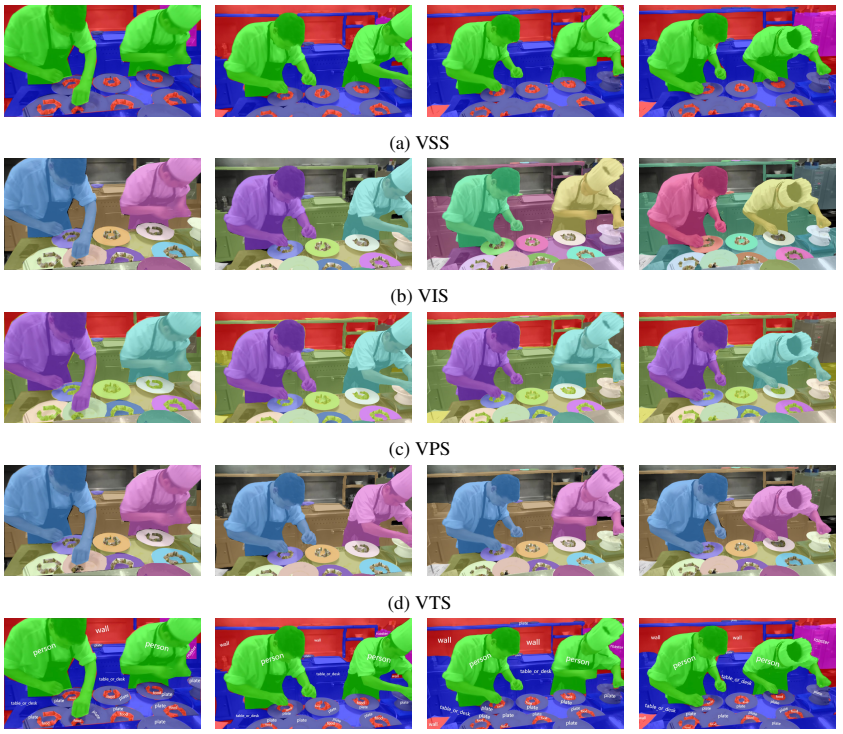
与动态场景理解 [233] 联系起来，VSP 既是一个重要的研究问题，也是一个实用的感知模块。在研究层面，它引出了关于时序一致性 [46, 67, 79]、时空特征提取 [84, 148] 以及动态目标鲁棒跟踪 [13] 的一系列问题。在应用层面，无论是自动驾驶、智能监控、机器人还是视频编辑，可靠的 VSP 都是下游安全决策的前提。

支撑 VSP 的技术体系已经历了三次更迭。早期系统依赖手工设计的描述子，包括颜色直方图、纹理和光流 [43, 81, 146]，并结合经典的学习方法，如聚类 [211]、基于图的方法 [51]、支持向量机（SVM）[137]、随机森林 [5] 以及马尔可夫随机场或条件随机场 [68, 101]。这些流程奠定了重要的基础，但可扩展性较差，并且需要大量面向特定领域的特征工程。深度学习打破了这一瓶颈：全卷积网络（Fully Convolutional Networks, FCNs）[20, 35, 150, 225] 能够学习用于稠密像素级预测的层次化特征，基于 FCN 的方法 [65, 133, 164, 224, 234] 很快成为 VSP 领域的主流框架。随后，Transformer [165] 开启了第三个时代：自注意力能够捕获长程依赖与全局上下文，其视觉领域的变体 [15, 22, 37, 39, 52, 107, 108, 120, 187, 201, 216] 通过建模超越 CNN 局部感受野限制的空间与时间交互，重塑了图像与视频分割。

这些方法论上的转变也拓宽了 VSP 需要解决的问题范围。视频跟踪与分割（VTS）的要求不止于逐帧分割，还需在帧间保持一致的身份 [28, 168]，这就要求在遮挡、突发运动和复杂交互下实现鲁棒的关联，而这对拥挤场景中的多目标跟踪和高级视频编辑流程至关重要。另一个前沿方向——开放词汇视频分割（OVVS）——将 CLIP 式的视觉—语言先验 [144] 与分割相结合，以摆脱经典 VSS 的封闭标签集。近期的开放词汇方法 [24, 57, 58, 192, 218] 能够分割预定义类别之外的目标，并以零样本方式泛化到无约束视频中普遍存在的稀有或新出现的实体。

在此背景下，现有综述仍存在三处结构性空白。**其一，覆盖范围碎片化**。Wang 等人 [176] 的早期综述以视频目标分割（Video Object Segmentation, VOS）为核心，仅将 VSS、VIS 和 VPS 作为相邻主题讨论，因为在其撰写时 VTS 和 OVVS 尚处于萌芽阶段。Li 等人 [86] 则仅聚焦于 Transformer 范式，对长期支撑该领域的卷积方法论着墨不多。其他综述要么从图像分割出发进行推广 [48, 71]，要么孤立地讨论开放词汇识别 [185, 229]，因此尚无一篇综述覆盖完整的 VSP 任务族。**其二，分析深度有限**。现有综述大多沿用罗列方法的写法，列出模型与精度数据，却很少考察相互竞争的方法谱系如何随任务设定、数据规模和计算预算的变化而演变。它们也未能充分分析四种时序建模机制——光流扭曲（flow warping）、注意力、基于查询的集合预测以及时空记忆——如何在精度、延迟、内存占用与长时域建模之间进行权衡。**其三，缺乏跨任务的综合分析**。VSS、VIS、VPS、VTS 与 OVVS 共享时序一致性、身份保持、长尾识别和标注效率等核心挑战，但大多数综述将它们分开讨论，而非从共同的跨任务维度加以审视。本综述旨在从跨任务视角对这些问题进行统一梳理。

具体而言，本综述有三点贡献。**第一，统一的任务覆盖范围**。我们将 VSP 视为由五项任务构成的一个有机整体，即 VSS、VIS、VPS、VTS 和 OVVS。§3 给出针对各任务的方法分类



(a) VSS

(b) VIS

(c) VPS

(d) VTS

(e) OVVS

图 2　VSP 各任务示意（样例来自 VIPSeg [121]）。VSS：相同颜色表示相同语义类别（不区分实例）。VIS/VPS：不同实例使用不同颜色（VPS 额外包含背景语义）。VTS：同一实例的颜色在帧间保持不变。OVVS：预定义标签集之外的新类别，并在每个分割区域上叠加显示类别标签。

体系，例如基于光流、基于注意力、实时以及半监督/弱监督的 VSS；§3.6 则将针对各任务的分析整合为统一的跨任务分析框架，涵盖基于 CNN、基于 Transformer、基于基础模型以及高标注效率学习等范式。**第二，基于共同维度的跨任务分析**。每个子任务的末尾都设有**共性问题**与**方法比较**两段，将各方法映射到一组权衡维度上，包括运动先验保真度、精度—延迟平衡、监督方式、身份稳定性与长尾鲁棒性。§3.6 进一步将共性问题归纳为四个维度——时序一致性、身份保持、长尾/开放世界识别和标注效率——并统一置于“标注—容量—延迟三难困境”之下。该节还归纳了典型的失效模式，使本综述从方法清单转变为一个诊断框架。**第三，性能演进的实证依据**。§3 与 §5 将静态基准表格与四幅性能演进图相结合，分别基于 Cityscapes [33]、VSPW [122]、YouTube-VIS [205] 和 KITTI-MOTS [168]，为跨任务的性能趋势提供可复现的视角。

图 1 展示了本综述的组织结构：从问题定义与背景出发，依次介绍方法、数据集与评测，直至未来方向，并在方法、资源、基准和跨任务讨论各部分中保持五项 VSP 任务的对齐。最后的展望将开放世界、统一、多模态、高效以及基于大语言模型的 VSP 与前文分析的失效模式和发展趋势联系起来。

2　背景

2.1　任务定义

五项 VSP 任务共享同一个基本框架，区别仅在于输出标签需要携带哪些信息。令 $V = \{I_t\}_{t=1}^{T}$ 和 Y 分别表示包含 T 帧的输入视频与输出分割空间。基于深度学习的 VSP 模型旨在寻找一

个从视频帧到分割标签的最优映射 $f^{*}: V \to Y$；图 2 以 $T=4$ 帧为例，可视化了 VSS、VIS、VPS、VTS 和 OVVS。

视频语义分割（VSS）。 VSS 是最基础的情形：它为每一帧生成类别图，与图像语义分割的区别仅在于额外要求连续帧之间具有时序一致性。给定由 T 帧 RGB 图像组成的片段 $V=\{I_t\}_{t=1}^{T}\in\mathbb{R}^{T\times H\times W\times 3}$，VSS 寻求一个映射 f^{VSS}，为每一帧返回逐像素的类别图 Y_t：

$$f^{\mathrm{VSS}}:\ V\ \longmapsto\ Y^{\mathrm{VSS}}=\{Y_t\}_{t=1}^{T},\qquad Y_t\in\mathcal{C}^{H\times W}, \tag{1}$$

其中 $\mathcal{C}$ 是固定的封闭标签集（例如 Cityscapes 的 19 个类别）。令 $\mathcal{T}_{t-1\to t}(\cdot)$ 表示借助光流或学习得到的对应关系对前一帧预测进行的传播，$\mathcal{O}_t$ 表示在第 $t-1$ 帧与第 t 帧中均可见的像素，则时序不一致率定义为

$$\Omega(Y^{\mathrm{VSS}})\ =\ \frac{1}{T-1}\sum_{t=2}^{T}\frac{1}{|\mathcal{O}_t|}\sum_{p\in\mathcal{O}_t}1\big[\,Y_t(p)\ \neq\ \big(\mathcal{T}_{t-1\to t}Y_{t-1}\big)(p)\,\big]\,. \tag{2}$$

此处 Ω 是仅针对 $\mathcal{O}_t$ 定义的软优化目标，因为在遮挡边界处，以及像素位置 p 上发生目标出现与消失时，标签发生变化是合理的。由于 $\mathcal{T}_{t-1\to t}$ 只是一个估计，我们期望 Ω 取一个较小的非零值。在数据集层面，Ω 以 mVC_n 的形式报告（§4.2）。值得注意的是，Y_t 只携带类别标签而不携带实例身份，这正是 VSS 与下文 VIS 的区别所在。

视频实例分割（VIS）。 VIS 在 VSS 的基础上增加了显式的实例身份维度，输出贯穿整个片段的 N 条实例时空管（tube，即跨帧掩码序列）：

$$f^{\mathrm{VIS}}:\ V\ \longmapsto\ Y^{\mathrm{VIS}}=\big\{(M_i,\,c_i,\,\mathrm{id}_i)\big\}_{i=1}^{N},\qquad M_i\in\{0,1\}^{T\times H\times W}, \tag{3}$$

其中 M_i 是第 i 个实例的时空二值掩码（即“时空管”），$c_i\in\mathcal{C}$ 为其类别，id_i 是在 T 帧范围内有效的唯一标识符。因此，VIS 的时序约束关注的是身份一致性，而不仅仅是类别一致性：相邻帧之间 id_i 的改变、漏检的实例或错误的类别分配，视发生变化的具体部分而定，都可能造成身份错误或分割错误。

视频全景分割（VPS）。 VPS 将实例与语义融合在一起，把 VSS 的逐像素类别图与 VIS 的实例时空管合并为一个互不重叠的划分，并区分 K 个 *stuff* 区域（不可数的区域类别，例如道路、天空）与 N 条 *thing* 时空管（可数的物体实例）：

$$f^{\mathrm{VPS}}:\ V\ \longmapsto\ Y^{\mathrm{VPS}}=\underbrace{\{(R_k,\,c_k^{\mathrm{stuff}})\}_{k=1}^{K}}_{\text{stuff 区域，无身份}}\ \cup\ \underbrace{\{(M_i,\,c_i^{\mathrm{thing}},\,\mathrm{id}_i)\}_{i=1}^{N}}_{\text{thing 时空管，携带身份}}, \tag{4}$$

其中 $R_k\subset\{1,\dots,T\}\times\{1,\dots,H\}\times\{1,\dots,W\}$ 是语义上连续的 stuff 区域，thing 分支则沿用 VIS 的时空管形式。两个分支在每个时空位置上互斥，因此每个体素恰好获得一个（类别，身份）对，标准的 VPQ 指标在两者上汇总时空管级的 IoU。由于额外要求 stuff 在帧间保持连续，并与 thing 的身份联合评测，VPS 严格难于单独的 VSS 或 VIS。

视频跟踪与分割（VTS）。 VTS 继承了 VIS 的时空管输出 $\mathcal{Y}^{\mathrm{VTS}}=\{(M_i,c_i,\mathrm{id}_i)\}_{i=1}^{N}$，但所假设的时间跨度 T 远长于 VIS 评测所用的数秒片段，从而将关注点从逐帧分割转向长时程的身份保持：

$$f^{\mathrm{VTS}}:\ V\ \longmapsto\ Y^{\mathrm{VTS}},\qquad \mathrm{id}_i^{(t)}=\mathrm{id}_i^{(t')}\ \ \forall\,(t,t')\in\{1,\dots,T\}^2\ \ \text{且}\ \ |t-t'|\to\infty, \tag{5}$$

即使目标经历遮挡或重新进入视野，其身份也必须保持不变。因此，评测指标从瞬时的 AP 转向累积的 MOTSA、sMOTSA 和 IDS，典型的失效模式则变为外观漂移、大幅度相机运动以及遮挡后重现。

开放词汇视频分割（OVVS）。 OVVS 与上述四项任务的区别在于标签空间：封闭集 $\mathcal{C}$ 被替换为由文本编码器 $\phi(\cdot)$ 张成的开放语义流形，从而保证对于任意测试时词表 $\mathcal{V}$（可能包含训练中未见过的自然语言名词），都能借助视觉—语言相似度完成类别分配。令 $\psi(\cdot)$ 为将视觉特征映射到与文本嵌入相同空间的视觉编码器，实例 i 在第 t 帧中的掩码嵌入为

$\psi_i^{(t)} = \psi\left(M_i^{(t)} \odot I_t\right)$。类别分配即为如下的余弦相似度决策：

$$f^{\mathrm{OVVS}} : (V, \mathcal{V}) \longmapsto \left\{(M_i, c_i, \mathrm{id}_i)\right\}_{i=1}^{N}, \qquad s_i^{(t)}(v) = \frac{\left\langle \psi_i^{(t)}, \phi(v) \right\rangle}{\left\|\psi_i^{(t)}\right\| \left\|\phi(v)\right\|}, \tag{6}$$
$$c_i = \arg\max_{v \in \mathcal{V}} \frac{1}{|\mathcal{F}_i|} \sum_{t \in \mathcal{F}_i} s_i^{(t)}(v),$$

其中 $\mathcal{F}_i$ 是实例 i 保持可见的帧的集合，因此每条轨迹只被赋予一个标签 id_i。只有当 $\mathcal{V}$ 是扁平结构时，该标签才是唯一的：一旦标签集中出现上位词，就会有多个 v 同时正确，而 arg max 只是按惯例选取一个最佳匹配，并非由任务本身的性质所决定。这一处替换便将任意宿主任务（VSS/VIS/VPS/VTS）的封闭集分类器转变为视觉—语言对齐，但同时带来了三个视频特有的难点：*(i)* 粒度歧义，同一像素可能匹配词表中的多个层级（例如“交通工具”/“汽车”/“轿车”），而封闭词汇协议会在数据集层面预先固定这些类别层级；*(ii)* 时序对齐漂移，由于 ϕ 和 ψ 通常在静态图像上预训练，同一目标在不同帧中可能因姿态、光照或运动模糊的变化而投影到不同的文本邻域，从而产生封闭词汇 VSP 中不可能出现的“语义闪烁”；*(iii)* 长尾与零样本泛化相互耦合，已见类别上的精度与未见类别上的连贯掩码相互争夺表征容量，这解释了 OVVS 与全监督 VSP 之间长期存在的差距。

表 1 汇总了上述五种形式化定义，从输出结构、时序约束、标签空间与主要挑战几个方面对各任务进行了对比。沿这些维度来看，输出结构从逐帧类别图演进为携带身份的时空管以及 thing/stuff 混合划分；时序约束从标签一致性收紧为长时程的身份持续性；标签空间则从固定集合开放为无界的词表。

2.2 发展历程

图像分割研究始于早期的边界检测方法 [38]，其中许多方法后来被迁移到视频上。最初的 VSP 尝试依据强度相似性对像素进行分组 [17, 51, 127, 131, 191]，并未进行显式的时空建模。第二波工作引入了光流 [6, 104, 124, 157]、图 [51, 75] 和图割 [134] 先验，以追求第一波方法所欠缺的时序稳定性。第一个真正的转折点出现在深度 CNN 时期：端到端的 FCN [20, 35, 150, 225] 取代了手工设计的运动与外观线索。由此形成了“逐帧 FCN + 事后平滑”的范式，以可解释的先验换取了学习得到的层次化特征，但也继承了逐帧闪烁的问题，随后只能借助光流 [72] 或 CRF [214] 推理进行事后修补。在这一 CNN 范式的基础上，又出现了另外三个转折点，各自催生出一个分支。第二个转折点加入了显式的实例身份 [205]，使关联与遮挡成为首要问题，而不再是后处理中的附带问题。第三个转折点将约束的关键从精度转向标注成本与延迟，推动了高标注效率的自监督/无监督学习 [19, 62, 82] 以及实时架构 [94, 132, 133, 151, 197] 的发展。第四个转折点以基于 Transformer 的注意力和基于查询的

表 1 五项 VSP 任务的比较：从输出结构、时序约束、标签空间与主要挑战四个方面进行对比（参见 §2.1 中的形式化定义）。

任务	输出结构	时序约束	标签空间	主要挑战
VSS	逐帧类别图 $\{Y_t\}_{t=1}^{T}$（无身份）	跨帧标签一致性（mVC_n）	封闭，thing + stuff	时序闪烁与延迟之间的权衡
VIS	thing 时空管 $\{(M_i, c_i, \mathrm{id}_i)\}$	片段级身份一致性	封闭，仅 thing	遮挡下的身份切换
VPS	stuff 区域 ∪ thing 时空管（互不重叠）	类别 + 身份的联合一致性	封闭，thing + stuff	thing/stuff 平衡，长尾
VTS	长视频流上的 thing 时空管	长时程身份持续性	封闭，仅 thing	遮挡—重现后的重识别
OVVS	宿主任务的输出，类别由文本匹配给出	宿主任务约束 + 稳定的视觉—语言对齐	开放词汇 $\mathcal{V}$	粒度歧义、语义闪烁、零样本泛化

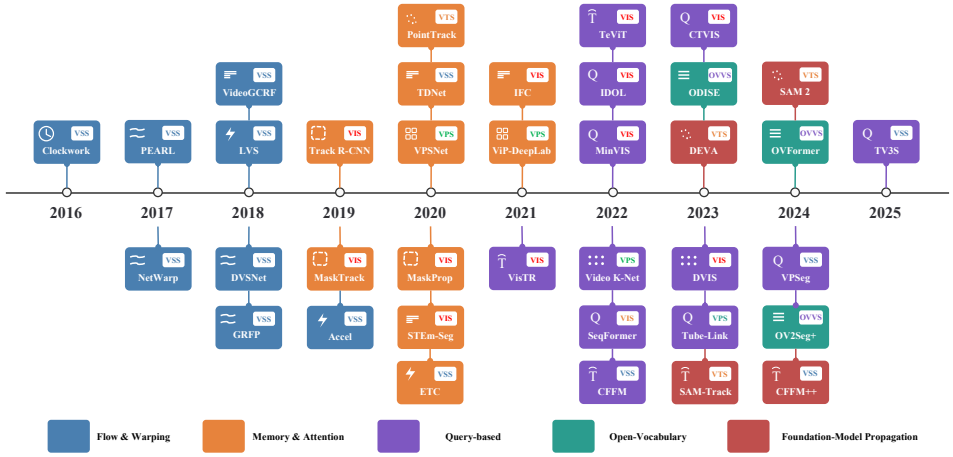


图 3 视频场景解析的架构演进（2016–2025）。代表性方法按五个范式带组织，并按目标任务（VSS/VIS/VPS/VTS/OVVS）着色；所有方法也均出现在 §3 的各任务汇总表中。

集合预测 [159, 161] 取代了手工设计的时序融合，将检测、分割与关联统一起来，代价是二次方的计算复杂度和对大规模训练数据的依赖。

我们在 图 3 中明确展示了这一架构演进：按时间顺序划分为五个范式，由上述四个转折点分隔。每一次转变都解决了前一范式的主要局限，同时又引入了新的局限，§3.6 中的跨任务分析将对这一规律加以阐明。

2.3 相关研究领域

VSP 并非孤立存在，而是大量借鉴了三个相邻领域，每个领域都只解决同一基础问题的某个特定方面：图像分割提供了逐像素分割的概念，视频目标分割提供了时序建模机制，视频目标检测则提供了以目标为中心的特征增强方法。下文依次梳理各领域的贡献，并说明 VSP 与这些领域的区别。

图像语义分割。 图像语义分割自始至终推动着 VSP 的进步 [1, 2, 41, 113, 172, 189, 224]。早期的 VSS 方法 [115, 153] 只是将图像分割器逐帧应用于视频。后续工作学会了利用时空一致性来同时提升精度与效率，但图像分割始终是其方法论基础。这种依赖关系十分深入：一个不带任何分割专用解码器的朴素 ViT 已能作为具有竞争力的图像分割器 [74]。VidEoMT [126] 进一步将这一思想推广到视频，表明“骨干网络即分割器”的原则跨越了与 VSP 相关的图像—视频边界依然成立。

*与 VSP 的区别。*二者在三个方面存在区别：*(i)* 时序约束，图像分割只要求逐帧像素正确，而 VSP 还需额外保证时序一致性（mVC、clip-AP、VPQ、IDS 及其零样本变体）；*(ii)* 输出空间，从 $H\times W$ 扩展为 $T\times H\times W$ 的时空体，VIS/VPS/VTS 还需进一步保持跨帧身份；*(iii)* 失效模式，图像分割器在类别边界处出错，而 VSP 还会在实例出现/消失、身份切换以及时序上碎片化的稀有类别预测上出错。

视频目标分割。 为分割问题加入时间维度，便得到了与 VSP 最为接近的领域——视频目标分割（VOS）。VOS 为 VSP 提供了许多时空构建模块，可分为三条谱系，每一条都在 VSP 中留下了清晰的印记。*(i)* 在线微调（OSVOS [12]、PReMVOS [112]）利用首帧标注将通用掩码网络适配到特定实例；PReMVOS 更进一步，将候选生成、精修与基于重识别（ReID）的合并融为一体，形成了 VIS 和 VTS 沿用至今的“先分割、后跟踪”标准流程。*(ii)* 掩码传播与匹配（MaskTrack [135]、RVOS [166]）将分割视为跨帧扭曲或循环预测，表明稠密掩码无需逐帧重新检测即可传播。*(iii)* 时空记忆网络（STM [128]、Episodic Graph Memory [110]、Global Context VOS [93]、ERMN [190]、Collaborative VOS [208]、STCN [30]、XMem [29]、Cutie [27]）

表 2 所综述 VSS 方法的主要特点汇总。

方法	发表	核心架构	技术特点	训练数据集
Clockwork [151]	ECCV-W'2016	FCN-8s	自适应的逐层时钟调度	NYUDv2 [154]/Cityscapes [33]
NetWarp [46]	ICCV'2017	FCN + 光流扭曲	基于光流的表征扭曲	CamVid [11]/Cityscapes [33]
PEARL [72]	ICCV'2017	FCN + GAN	预测式特征学习	CamVid [11]/Cityscapes [33]
VideoGCRF [16]	CVPR'2018	FCN+CRF	高斯 CRF	DAVIS [136, 139]/CamVid [11]
EUVS [65]	ECCV'2018	贝叶斯 CNN	光流引导的特征聚合	CamVid [11]
LVS [94]	CVPR'2018	FCN	关键帧选择	CamVid [11]/Cityscapes [33]
GRFP [125]	CVPR'2018	FCN + GRU	时序特征聚合	CamVid [11]/Cityscapes [33]
DVSNet [197]	CVPR'2018	FCN+ 强化学习	关键帧选择	Cityscapes [33]
Accel [67]	CVPR'2019	FCN	关键帧选择	KITTI [49]
SSeg [234]	CVPR'2019	FCN	弱监督学习	CamVid [11]/Cityscapes [33]
Naive-Student [19]	ECCV'2020	FCN+ 知识蒸馏	半监督学习	Cityscapes [33]
EFC [36]	AAAI'2020	FCN	时序特征聚合	CamVid [11]/Cityscapes [33]
TDNet [63]	CVPR'2020	记忆网络	基于注意力的特征聚合	CamVid [11]/Cityscapes [33]/NYUDv2 [154]
ETC [106]	ECCV'2020	FCN + 知识蒸馏	知识蒸馏	CamVid [11]/Cityscapes [33]
DepthMix [62]	CVPR'2021	FCN	半监督学习	Cityscapes [33]
CFFM [159]	CVPR'2022	Transformer	静态与运动上下文学习	Cityscapes [33]/VSPW [122]
MRCFA [161]	ECCV'2022	Transformer	对跨帧亲和度施加注意力	Cityscapes [33]/VSPW [122]
CIRKD [204]	CVPR'2022	知识蒸馏	跨图像关系	CamVid [11]/Cityscapes [33]/Pascal VOC [40]
IFR [235]	CVPR'2022	FCN	半监督学习	CamVid [11]/Cityscapes [33]
MVNet [69]	CVPR'2023	Transformer	多光谱输入	MVSeg [69]
SSLTM [82]	CVPR'2023	FCN + Transformer	半监督学习	Cityscapes [33]/VSPW [122]
MPVSS [180]	NIPS'2023	FCN + Transformer	关键帧选择	Cityscapes [33]/VSPW [122]
BLO [203]	CVPR'2023	Transformer	关键帧选择	Cityscapes [33]/ADE20K [227]/Pascal VOC [40]
VPSeg [54]	CVPR'2024	Transformer	消失点	ACDC [149]/Cityscapes [33]
TDC [236]	CVPR'2024	FCN	半监督学习	CamVid [11]/Cityscapes [33]
CFFM++ [160]	PAMI'2024	Transformer	全局时序上下文学习	CamVid [11]/Cityscapes [33]
TV3S [61]	CVPR'2025	Transformer + Mamba	状态空间	VSPW [122]/Cityscapes [33]

将时序建模重新表述为对外部特征库的读写操作，其中 XMem 借鉴 Atkinson–Shiffrin 记忆模型设计的三级记忆结构，使 VOS 能以恒定的内存开销处理分钟级视频。

*与 VSP 的区别。*尽管共享时序机制，VOS 与 VSP 在三个方面存在区别：*(i)* 问题设定，VOS 与类别无关，且以指定实例为处理对象，目标由首帧掩码（半监督）或显著性线索（无监督）指定，类别从不参与其中；而 VSP 需要对每个像素分类（VSS）、发现并标注每个实例（VIS）、统一 thing/stuff（VPS）或保持轨迹身份（VTS）；*(ii)* 监督方式，VOS（DAVIS、YouTube-VOS）属于单样本/少样本设定，而 VSP（Cityscapes-VSS、VSPW、YouTube-VIS、VIPSeg、KITTI-MOTS）属于稠密的全监督设定；*(iii)* 失效模式，VOS 的难点在于外观漂移以及对已知目标的重识别，而 VSP 还需面对类别混淆、实例出现/消失以及 stuff 闪烁。因此，VOS 对 VSP 的贡献在于架构层面（记忆读写、查询—键匹配、掩码传播），而非语义层面。

视频目标检测。 VOS 在时间上跟随一个指定的目标，而视频目标检测（Video Object Detection, VOD）则将问题拓宽为逐帧定位并分类所有感兴趣的目标。这类检测器将时序线索注入图像级检测流程 [116, 158, 173, 219, 231, 232]，因此与实例级 VSP 共享时序一致性、运动模糊和遮挡等核心挑战。这些共享的机制也为 VSP 奠定了有益的基础，因为场景动态与目标间交互正是 VSP 的核心所在。

*与 VSP 的区别。*VOD 与 VIS 共享时序构建模块（特征聚合、跨帧关联、运动补偿），但在三个方面存在区别：*(i)* 几何粒度，VOD 预测边界框（4 个标量），而 VSP 预测像素掩码（$H \times W$），在边界精度、细长结构和遮挡方面的要求高出一个数量级；*(ii)* 语义覆盖，VOD 只处理固定词表中的 *thing* 类别，而 VSP 涵盖稠密语义（VSS）、VPQ 评测下的 thing+stuff（VPS）、仅含 thing 的稠密掩码（VIS/VTS）或开放词汇识别（OVVS）；*(iii)* 身份保持，VOD 只关联边界框，而 VTS 还额外要求掩码形状在时间上保持连贯，从而排除了“边界框一致但掩码闪烁”的失效情形。

3 方法综述

3.1 视频语义分割

VSS 将图像语义分割扩展到视频，在视频中，朴素的逐帧推理会忽略时间上的连续性 [71]。我们按时间顺序和功能将相关方法组织为四个方向：用于早期时序对齐的基于光流的传播、

用于学习跨帧上下文的基于注意力的聚合、用于控制延迟的实时设计，以及用于降低标注成本的半监督/弱监督学习。表 2 汇总了所综述 VSS 方法的主要特点。

基于光流的方法。 光流提供了一种几何上可解释的运动先验：它在亮度恒定假设下估计逐像素的稠密位移，用于跨帧传递信息 [65]。两条技术路线以不同方式使用光流，而这种差异决定了各自的失效方式。*(i)* 特征传播将光流作为外部扭曲算子作用于已计算好的特征 [46, 72, 197]，因此光流误差以输入损坏的形式进入，分割器无法加以纠正。*(ii)* 显式时空建模则将运动纳入推理过程本身，因而能够削弱自身误差的影响：例如在遮挡下对传播进行门控 [125]、联合优化光流与分割 [36]，或借助 CRF、多分辨率精修以及特征级聚合 [16, 67, 106]。两条路线共享同一个上限，即精度受限于光流质量；路线 (ii) 仅能使性能退化更为平缓。

基于注意力的方法。 VSS 中的注意力机制主要充当时空特征聚合器，其发展可沿三条功能主线来理解，而非若干互不相关的类别。*(i)* 时序蒸馏与多尺度融合将计算预算投入到时间轴上，但方式各异：TDNet [63] 将表征拆分到一个轻量的时序分支上；CFFM 系列 [159, 160] 先分离静态上下文与运动上下文，再扩大时间感受野；MRCFA [161] 则作用于跨帧亲和度而非特征本身。*(ii)* 知识迁移与跨视角迁移保留了聚合器，但改变了输入给它的内容 [69, 82, 204]。*(iii)* 关键帧引导与几何先验引入显式先验，以减少注意力需要学习的内容 [54, 180]。总体趋势是从以计算换取时序上下文，转向从结构中获取时序上下文。

实时方法。 实时 VSS 降低延迟的主要途径是在时间上摊销计算，而非缩小骨干网络。两种调度机制占据主导：缓存—跳过（cache-and-skip），即网络的不同部分以不同速率更新 [151]；以及传播—刷新（propagate-and-refresh），即由自适应规则决定何时重新计算、何时进行扭曲传播 [94, 197]。后续的注意力与掩码传播流程将这两种机制作为附加组件继承了下来 [63, 106, 159, 180]，因此它们带来的加速与精度提升在机制上基本相互独立。TV3S [61] 则改变了成本模型：它通过 Mamba 骨干网络逐块传播状态，使时序上下文以线性时间累积。

定量权衡。表 8 与 图 4(a) 共同表明，实时 VSS 遵循两条主要的帕累托路线：轻量骨干网络配合关键帧剪枝可最大化速度，而基于注意力的特征复用则牺牲部分吞吐量以换取更高的精度。这将针对严格延迟进行优化的方法，与追求更均衡的精度—效率权衡的方法区分开来。

半监督/弱监督方法。 当带标注的视频稀缺时，这类方法利用时序连贯性与时空相关性，从无标注或稀疏标注的片段中获取替代监督 [234–236]。三种做法反复出现：基于模型自身伪标签的自训练 [19]；利用几何线索构造更难的无标注样本的数据增强 [62]；以及在长的无标注片段上施加注意力 [82]。

伪标签噪声与训练不稳定性。共用的自训练循环带有两个系统性风险，而视频场景会使其进一步恶化。首先是伪标签噪声的时序放大：单帧错误会在相邻帧之间传播并被反复自我确认，对稀有类别和运动边界的影响最为严重。Naive-Student [19] 直接展示了这一点——其在长尾类别上的提升远小于头部类别；DepthMix [62] 则报告了在缺乏可靠深度先验时出现的边界漂移。其次是训练动态不稳定：伪标签会随着教师模型的更新而漂移，因此师生联合优化容易出现特征坍塌与类别坍塌。常用的缓解手段（置信度阈值、EMA 教师、一致性正则化 [235, 236]）本身也带来了超参数敏感性。因此，半监督/弱监督 VSS 中的标注效率是有代价的：它将统计风险从标注环节转移到了训练环节，并与全监督上界之间留下一定的 mIoU 差距——这一差距随标注预算的增加而缩小，但随类别长尾程度的加剧而扩大。

共性问题。 VSS 面临三个相互耦合的主要问题：时序一致性、边界闪烁以及精度—延迟权衡。与逐帧分割不同，VSS 必须使像素标签在时间上保持稳定，因为朴素的逐帧推理会在目标边界以及小目标或快速运动区域上产生可见的闪烁，而这正是 VSPW [122] 上的视频一致性指标 mVC_n 所惩罚的。两种主流的补救手段方向相反。光流扭曲 [46, 67] 在几何上可解释，但在大位移和遮挡下会失效；跨帧注意力 [63, 159] 更为灵活，但其开销随时空体积呈二次方增长。这使时序稳定性直接与精度—延迟权衡挂钩，因为用于恢复实时性能的关键帧调度和特征复用技巧同样会削弱跨帧连贯性——这正是 §3.6 中三难困境的一个具体实例。

表 3 所综述 VIS 方法的主要特点汇总。

方法	发表	核心架构	技术特点	训练数据集
MaskTrack R-CNN [205]	ICCV'2019	Mask R-CNN	基于检测的跟踪	YouTube-VIS [205]
STEm-Seg [4]	ECCV'2020	FCN	时空嵌入	DAVIS [139]/YouTube-VIS [205]/KITTI-MOTS [168]
MaskProp [9]	CVPR'2020	Mask R-CNN	实例特征传播	YouTube-VIS [205]
MVAE [96]	CVPR'2020	Mask R-CNN+VAE	变分推断	YouTube-VIS [205]/KITTI-MOTS [168]
MOTSNet [140]	CVPR'2020	Mask R-CNN	无监督学习	KITTI-MOTS [168]/BDD100K [212]
SemiTrack [44]	CVPR'2021	SOLO	半监督学习	Cityscapes [33]/YouTube-VIS [205]
CompFeat [45]	AAAI'2021	Mask R-CNN	时空特征对齐	YouTube-VIS [205]
IFC [66]	NIPS'2021	FCN + Transformer	帧间通信	YouTube-VIS [205]
Propose-Reduce [97]	ICCV'2021	Mask R-CNN	先提议后约简	DAVIS [139]/YouTube-VIS [205]
SG-Net [102]	CVPR'2021	FCOS	单阶段分割	YouTube-VIS [205]
fIRN [103]	CVPR'2021	Mask R-CNN	弱监督学习	Cityscapes [33]/YouTube-VIS [205]
VisTR [178]	CVPR'2021	Transformer	基于 Transformer 的分割	YouTube-VIS [205]
TraDeS [183]	CVPR'2021	FCN	基于检测的跟踪	YouTube-VIS [205]/KITTI-MOTS [168]
CrossVIS [206]	ICCV'2021	FCN	动态卷积	YouTube-VIS [205]/OVIS [141]
TPR [87]	PAMI'2022	FCN	时序金字塔路由	YouTube-VIS [205]
VISOLO [59]	CVPR'2022	Transformer	基于网格的时空聚合	COCO [99]/YouTube-VIS [205]
MinVIS [64]	NIPS'2022	Transformer	基于查询的跟踪	YouTube-VIS [205]/OVIS [141]
STC [70]	ECCV'2022	FCN	对比学习	YouTube-VIS [205]/OVIS [141]
TrackFormer [119]	CVPR'2022	FCN + Transformer	跟踪查询	MOT17 [123]/MOTS20 [168]
SeqFormer [184]	ECCV'2022	Transformer	序列 Transformer	YouTube-VIS [205]
IDOL [186]	ECCV'2022	Transformer	对比学习	YouTube-VIS [205]
TeViT [207]	CVPR'2022	Transformer	金字塔特征交互	YouTube-VIS [205]/OVIS [141]
GenVIS [60]	CVPR'2023	Transformer	基于查询的训练流程	YouTube-VIS [205]/OVIS [141]
VideoCutLER [177]	CVPR'2023	DINO + FCN	无监督学习	DAVIS [139]/YouTube-VIS [205]
CTVIS [210]	ICCV'2023	Transformer	对比学习	YouTube-VIS [205]/OVIS [141]
DVIS [221]	ICCV'2023	Transformer	解耦策略	YouTube-VIS [205]/OVIS [141]
OVFormer [42]	ECCV'2024	Transformer	对比学习	LV-VIS [170]/YouTube-VIS [205]/OVIS [141]
OV2Seg+ [169]	IJCV'2024	Transformer	开放词汇学习	LVIS [56]/LV-VIS [170]/YouTube-VIS [205]…

方法比较。 三种权衡构成了 VSS 的设计空间。在运动先验维度上，光流 [36, 46, 67] 的几何显式性（可解释且计算廉价，但受限于光流质量，在大位移或光照变化下较为脆弱）与注意力 [63, 159, 161] 的数据驱动灵活性相互对立；后者摒弃了这一先验，代价是 $\mathcal{O}((THW)^2)$ 的开销，除非借助稀疏化或 TV3S [61] 等状态空间替代方案加以缓解。在保真度—延迟维度上，起决定作用的与其说是骨干网络，不如说是计算如何在时间上摊销，因此关键帧调度 [94, 180, 197, 203] 和特征蒸馏 [106, 204] 能够使采用计算开销较大骨干网络的方法也能实现实时推理。监督维度与前两者正交：半监督/弱监督方案 [19, 62, 82, 236] 解决的是标注瓶颈而非架构问题，其代价是一定的 mIoU 差距，且该差距随类别不平衡和伪标签噪声的加剧而扩大。

3.2 视频实例分割

VIS 由文献 [205] 提出，它统一了逐帧的实例检测、分割与跨帧跟踪。我们沿三条经常相互交叠的主线对方法进行归类：显式关联检测结果的基于跟踪的方法、利用查询或记忆机制进行集合预测的基于注意力的方法，以及降低视频掩码标注成本的半监督/弱监督方法。下文的顺序遵循该领域从显式关联到基于查询的时序建模的发展脉络，同时需注意个别方法可能同时属于多条主线。表 3 汇总了所综述 VIS 方法的主要特点。

基于跟踪的方法。 这类方法将“逐帧检测/分割 + 时序关联”的范式从图像推广到视频 [9, 45, 87, 96, 97, 102, 206]。它逐步扩大定义关联的单元，每一次扩大都消除一类错误，同时暴露出下一类错误。*(i)* 显式逐帧匹配——借助跟踪分支 [168, 205] 或回归得到的运动偏移 [183]——具有可解释性，但会继承每一次漏检。*(ii)* 片段级隐式设计通过在时空体上进行预测来摒弃匹配 [4, 66]，避免了上述失效模式，但时间感受野在训练时便已固定。*(iii)* 显式连贯性方法在此基础上添加一致性目标 [70]，但只要遮挡的持续时间超过片段长度，身份切换就依然存在。

基于注意力的方法。 在 VIS 中，注意力将检测、分割与关联统一为单一的集合预测目标，因此各方法的主要区别在于如何施加时序一致性，可分为三组。*(i)* 集合预测的统一化将整个片段作为预测单元：VisTR [178] 让查询在整个时空体上进行注意力计算；SeqFormer [184] 进一步指出，时序聚合必须与逐帧的空间注意力解耦，否则掩码保真度会下降；TeViT [207] 则将这一思路移植到 Transformer 金字塔上。*(ii)* 解耦流程将时序建模移出训练过程，在推理时借助查询相似度 [64] 或专门的关联流 [221] 恢复身份；这种做法代价低廉，却把一致性交

表 4 所综述 VPS 方法的主要特点汇总。

方法	发表	核心架构	技术特点	训练数据集
VPSNet [76]	CVPR'2020	Mask R-CNN	时空特征对齐	VIPER-VPS [76]/Cityscapes-VPS [76]
ViP-DeepLab [143]	CVPR'2021	FCN	深度感知的全景分割	Cityscapes-VPS [76]
SiamTrack [181]	CVPR'2021	孪生 FCN	有监督对比学习	VIPER-VPS [76]/Cityscapes-VPS [76]
PVPS [118]	ECCV'2022	FCN	环视全景 VPS	WOD: PVPS [118]
Video K-Net [91]	CVPR'2022	Transformer	基于查询的学习	KITTI [49]/Cityscapes [33]/VIPSeg [121]
Clip-PanoFCN [121]	CVPR'2022	FCN	两阶段分割	VIPSeg [121]
TubeLink [90]	ICCV'2023	Transformer	基于查询的学习	VIPSeg [121]
PolyphonicFormer [213]	ECCV'2022	Transformer	基于查询的学习	KITTI [49]/Cityscapes-VPS [76]
ODISE [193]	CVPR'2023	UNet + Transformer	开放词汇全景特征	ADE20K [227]/COCO [14]

给了一个不受分割损失直接约束的组件。*(iii)* 记忆与对比稳定化重新将一致性内化到模型之中 [59, 60]，其中 CTVIS [210] 的对比查询库弥补了 (ii) 所留下的缺口。

半监督/弱监督方法。 与 VSS 中的对应方法类似，半监督/弱监督 VIS 将时序连贯性用作无需额外标注的监督信号。沿着这条脉络，监督信号逐步减弱：无标注视频配合有标注图像 [44]、仅依赖运动与时序一致性 [103]、自动生成的 MOTS 标注 [140]，直至一个从逐帧掩码合成训练视频的完全无监督流程 [177]。其精度仍落后于有监督的上限，瓶颈在于伪标签漂移与遮挡歧义，即 §3.6 中所形式化的三难困境的"标注"一边。

共性问题。 VIS 面临三种相互交织的主要失效模式：长时跟踪、身份切换与遮挡。它们共同导致了在长片段或遮挡片段上的剩余差距，这在 OVIS [141] 上表现得最为明显。*(i)* 身份切换：当两个同类实例在图像平面上交叉时，原始的查询相似度关联（如 MinVIS）往往会在随后的帧中发生错误绑定，此时帧级 mAP 几乎不受影响，而身份感知的指标却急剧下降；CTVIS [210]（对比查询记忆）与 DVIS [221]（显式关联头）是最直接的应对措施，但两者都引入了额外的超参数（对比温度、关联阈值），而这些超参数的跨数据集稳定性本身仍是一个开放问题。*(ii)* 长时跟踪：YouTube-VIS 片段上的性能已被基于查询的模型推至饱和，但在 OVIS 上精度大幅下降——OVIS 的视频长度比 YouTube-VIS 片段长一个数量级，而片段级注意力的时间感受野受限于训练片段的长度。GenVIS [60]（多片段训练）与 SeqFormer [184] 从两个角度应对这一问题，但都未能解决推理时对持久实例状态的根本需求。*(iii)* 遮挡：OVIS 专为暴露遮挡问题而构建，其重度遮挡子集上的精度远低于轻度遮挡子集。主要的失效方式是：实例在遮挡下消失，重新出现时被赋予了新的身份，于是遮挡问题退化为身份切换问题。这三种失效有一个共同的根源：缺乏持久、显式的逐实例记忆，因此时序一致性只能在注意力窗口内自发涌现，§3.6 将其形式化为三难困境的"容量"一边。

方法比较。 在架构维度上，方法经历了从显式的基于检测的跟踪级联 [183, 205]，到无提议的时空体方法 [4, 66]，再到完全统一的查询式 Transformer [64, 178, 210, 221] 的迁移，以归纳偏置换取表征的通用性。级联方法模块化、代价低，但会传播检测—分割—关联各阶段的误差，并在遮挡下表现不佳；查询式 Transformer 则将三个子任务吸收进单一的集合预测目标，代价是繁重的预训练、二次方复杂度的时空注意力，以及在长片段上不稳定的优化。近期结果中的决定性因素已不再是原始的模型容量，而是如何施加时序一致性：对比记忆（CTVIS [210]）、任务解耦（DVIS [221]）或隐式的查询连贯性（MinVIS [64]）。正交的监督维度针对的是标注成本而非架构 [44, 103, 177]，其在 OVIS [141] 上与有监督最优方法之间持续存在的差距表明，今后的关键约束是遮挡鲁棒性，而非模型容量。

3.3 视频全景分割

VPS [76] 将 VSS 与 VIS 融合为一项任务：每个像素都需获得语义标签，同时前景实例需被区分出来并随时间跟踪。

我们沿两条相互正交的轴来组织 VPS 方法。架构轴对比双分支 CNN 与基于查询的 Transformer，追溯该领域如何从分离的 stuff/thing 流走向统一的集合预测。任务设定轴则将标准的 2D VPS 与深度感知的 3D VPS 区分开来——在后者中，深度作为输入或输出信号出现，而并不标志着一个新的架构族。两条轴互不重叠，因为深度感知 VPS 既可以基于 CNN 实现（如 ViP-DeepLab [143]），也可以基于查询式 Transformer 实现（如 PolyphonicFormer [213]）。

因此，我们并不把深度感知 VPS 视为与基于查询和双分支方法并列的第三种范式，而是将其视为对任务设定的扩展。表 4 汇总了所综述 VPS 方法的主要特点。

基于查询的方法。 基于查询的 VPS 将分割与跟踪重新表述为查询交互问题，从而去除了手工设计的跟踪启发式规则 [181]。Video K-Net [91] 在 K-Net 的基础上引入一组可学习的卷积核，联合编码语义类别与实例身份；这些卷积核与像素特征交互，并隐式地保持时序一致性。Tube-Link [90] 将视频切分为短片段，用 Transformer 为每个片段生成时空管掩码，并通过查询间注意力将它们链接起来，从而无需外部跟踪器。PolyphonicFormer [213] 在查询设计中增加了深度分支，利用全景分割与深度估计之间的相互促进。

推动向查询式方法转变的因素。 VPS 从双分支 CNN 向查询式 Transformer 的迁移由三个因素推动。*(i)* 统一的 *thing/stuff* 表征：双分支设计（VPSNet [76]、PVPS [118]）训练两条流，再通过启发式融合将其拼接起来，由此带来边界错位和身份漂移，而查询式设计正是为消除这些问题而提出的。查询式转向由 Video Mask2Former 与 Video K-Net [91] 所巩固：前者是 Mask2Former [26] 的视频扩展，以共享的 Transformer 解码器将视频分割重新表述为掩码分类；后者用同一组卷积核重新表示 thing 与 stuff，使融合成为处理流程内部的操作。*(ii)* 关联的隐式化：两种在 VIS 中得到验证的方案——图像级训练配合推理时的查询相似度关联（MinVIS [64]），以及检测/分割/关联的显式解耦（DVIS [221]）——已被移植到 VPS 中，将关联从一个架构组件降格为一种推理策略，并降低了训练成本。*(iii)* 时序稳定性：单纯的查询相似度关联在长片段或遮挡片段下会失效（参见 §3 中关于 VIS 长时跟踪的讨论）；CAVIS [83] 通过对比查询记忆和显式关联头将 CTVIS 的方案引入 VPS，因此 VPS 如今也遵循了此前在 VIS 中观察到的“统一—解耦—再稳定”的演进路径。

双分支方法。 双分支方法将全景预测解耦为 stuff 分割分支与 thing 实例分支，再将两路输出融合为统一的全景图。PVPS [118] 将这一设计扩展到 VPS：它将五个同步相机的画面拼接成 220° 的视场，并利用已知的相机参数将 2D 像素投影到共享的 3D 空间中，从而在布局复杂、遮挡频繁的广角场景中实现一致的分割。

深度感知扩展。 与上述基于查询和双分支的范式不同，这一类别并非额外的架构范式，而是 VPS 任务设定在 3D 上的扩展：深度被加入输入/输出通道，而底层架构既可以基于 CNN，也可以基于查询式 Transformer，因此这条轴与范式轴相互正交。ViP-DeepLab [143] 将该任务视为单目深度估计与语义及实例标注的联合问题，从而恢复出带有逐点类别和身份的 3D 点云。PolyphonicFormer [213] 则采用基于查询的形式，在共享解码器中让深度头与全景头相互促进。

共性问题。 VPS 面临三个相互耦合的主要问题：thing/stuff 平衡、长尾覆盖以及实时性瓶颈。VPS 区别于 VSS 和 VIS 之处，并不在于它叠加了这两项任务，而在于它必须在一个互不重叠的视频级输出中协调 thing 实例与 stuff 区域。表 11 中的基准趋势指向三个持续存在的瓶颈。*(i) thing/stuff* 不平衡：基于查询的方法在实例级的 thing 与稠密的 stuff 区域之间共享表征容量，因此在身份感知的实例推理与稳定的背景解析之间取得平衡仍然困难。双分支设计缓解了这种竞争，却带来了自身的边界融合误差。*(ii)* 长尾类别覆盖：与 VIPSeg [121] 等更广泛的开放场景设定相比，规模较小的城市场景基准上的精度提升更为稳定，这表明词表扩展与数据不平衡正日益限制着进展。*(iii)* 实时 *VPS*：延迟作为一个评测维度仍然报告不足、尚不成熟，因为在线 VPS 必须在保持 thing 身份的同时输出稠密的全景预测。这使得精度—延迟权衡比 VSS 中更加尖锐，并将 VPS 直接与 §3.6 中的跨任务瓶颈联系了起来。

方法比较。 该领域从双分支 CNN [76, 118] 转向基于查询的 Transformer [90, 91]，其驱动力与其说是骨干网络的容量，不如说是集合预测在表征上的经济性：同一组可学习的卷积核或查询即可同时承载语义类别与实例身份，从而消除了双分支设计中脆弱的 thing/stuff 融合启发式规则，并使跨帧关联成为查询持久性的副产物。深度感知更准确地应理解为一种任务扩展（即在 CNN 骨干网络 [143] 或查询式 Transformer [213] 上实现的、以 3D 为基础的 VPS 变体），而非一个并列的方法类别。这些进展并非没有代价：查询注意力的复杂度为 $\mathcal{O}(N_q \cdot T \cdot HW)$，实时 VPS 仍是一个开放问题；而 Cityscapes-VPS 上近乎饱和的结果与

表 5 所综述 VTS 方法的主要特点汇总。

方法	发表	核心架构	技术特点	训练数据集
Track R-CNN [168]	CVPR'2019	Mask R-CNN	基于检测的跟踪	KITTI MOTS [168]/MOTSChallenge [168]
PointTrack [198]	ECCV'2020	PointTrack	基于点的跟踪	APOLLO MOTS [198]
MaskProp [9]	CVPR'2020	Mask R-CNN	实例特征传播	YouTube-VIS [205]
ASB [32]	ICCV'2021	分配空间	Top-k + 动态规划	KITTI MOTS [168]/MOTSChallenge [168]
MPNTrackSeg [10]	IJCV'2022	MPN + GNN	基于图	MOT20 [34]/KITTI [49]/HiEve [100]
MITS [196]	ICCV'2023	Transformer	跟踪与分割联合建模	YouTube-VOS [195]/DAVIS [136]
SAMTrack [31]	arXiv'2023	SAM + DeAOT	多模态交互	DAVIS [136, 139]
SAM-PT [145]	arXiv'2023	SAM + PT	基于点的跟踪	DAVIS [136]/YouTube-VOS [195]/BDD100K [212]
DEVA [28]	ICCV'2023	SAM	分割与跟踪解耦	VIPSeg [121]/YouTube-VIS [205]
SAM 2 [147]	arXiv'2024	SAM	提示式分割	SA-V [147]

VIPSeg [121]（124 个类别，长尾分布）上低得多的分数之间的反差表明，关键约束已从架构的表达能力转向数据稀缺条件下的长尾覆盖与稳定的 thing/stuff 平衡。

3.4 视频跟踪与分割

VTS 随时间联合进行目标检测、跟踪与像素级分割，可应用于自动驾驶、监控和行为识别。我们按照实例在时间上的表示方式对方法进行组织（基于点的方法、两阶段基于提议的方法，以及高效分配/解耦的方法），并在下文的"方法比较"中沿两条相互正交的轴——*耦合方式*（紧耦合 *vs.* 解耦）与*表示方式*（稠密 *vs.* 稀疏）——重新审视它们。表 5 汇总了所综述 VTS 方法的主要特点。

基于点的方法。 基于点的方法用稀疏的 2D 点云替代稠密的卷积特征，以更好地区分前景与背景。PointTrack [198] 在四种模态（偏移、颜色、类别、位置）上将前景及其周围区域都表示为点云，用于联合分割与跟踪。SAM-PT [145] 则在帧间传播正负点，并通过 SAM [78] 迭代地精修掩码，同时剔除被遮挡的点、补充新出现的可见点。MITS [196] 进一步强化定位，通过自注意力将分割掩码与边界框耦合起来，实现统一的跟踪。这三种方法表明，点级表示能以轻量的代价达到有竞争力的精度。

两阶段方法。 两阶段方法先生成区域提议，再进行分类、分割和跨帧关联，从而使定位与精修相互解耦。Track R-CNN [168] 确立了这一模式：它在 ResNet-101 骨干网络上加入 3D 卷积，并增加一个跨帧匹配嵌入的关联头。MaskProp [9] 沿同一路线扩展 Mask R-CNN，加入一个掩码传播分支，使实例特定的特征随时间传递。MPNTrackSeg [10] 在流量守恒约束下将关联步骤重新表述为图上的消息传递，对关联边进行分类，并基于 CNN 嵌入预测掩码。SAM-Track [31] 拓宽了适用场景，同时支持交互模式（SAM [78] + Grounding DINO [105]）与自主模式，以抑制无约束场景中的身份漂移。

高效方法。 高效方法通过将关联重新表述为优化问题或拆分处理流程来降低关联的成本。ASB [32] 走的是优化路线，它将跟踪重新表述为分配空间搜索：在融合 IoU、外观和距离的代价函数下，用 Hungarian–Murty 算法找出帧间的 top-k 匹配，并以动态规划完成全局优化以及跨越检测中断的长时重关联。DEVA [28] 走的是解耦路线，它将分割与跟踪分离，使任务特定的分割器逐帧运行，再由双向传播模块随时间将结果链接起来，并能够跨任务泛化。

共性问题。 在 VTS 中，关键约束是时长：在一段长视频流中始终保持一个实例的完整性，比解析任何单个片段都更难。VTS 与 VIS 共享相同的输出结构 $\{(M_i, c_i, \mathrm{id}_i)\}$（§2.1），但二者考验的时序范围不同。VIS 通常在短片段上评分，因此依赖片段级集合预测和基于查询的注意力；而 VTS 面向更长的视频流，身份连续性、遮挡恢复和帧间关联在其中占据主导。评价指标的侧重点也随之变化：VIS 主要报告 AP，而 VTS 突出 sMOTSA、MOTSA 和 IDS 等跟踪感知的指标。正因如此，VTS 方法往往偏好轻量化表示、显式关联、图匹配或基于传播的设计，而非完全照搬 VIS 的架构。因此，VIS 与 VTS 之间的分化与其说源于输出格式，不如说源于时间跨度、评测协议与部署场景。

方法比较。 VTS 的核心问题不是分割什么，而是分割与跟踪如何耦合。在耦合轴上，紧耦合的联合框架 [168, 196] 端到端地学习嵌入与关联，但一旦长时间遮挡扰乱了嵌入流形，就会变得脆弱。解耦流程 [28, 31, 145] 采取相反的策略：将逐帧分割交给 SAM [78]，把时序链

接留给一个轻量的传播器，因而继承了开放词汇的通用性，但在长时间间隔下的重识别上仍有困难。在表示轴上，基于点的编码 [145, 198] 以稠密的像素覆盖换取一种稀疏且对形变鲁棒的代理表示；而分配空间 [32] 与图消息传递 [10] 的形式化方法则以额外的推理开销为代价，将跟踪重新表述为组合优化问题。尽管已问世六年，PointTrack [198] 在 KITTI-MOTS 上仍具竞争力（表 12），这表明在 VTS 中，精度的主要驱动因素是精心的表示设计，而非原始的模型容量或基础模型的规模。

3.5 开放词汇视频分割

OVVS 融合了语义分割、时序跟踪与基于语言的识别，并借助视觉—语言模型突破固定词表的限制。三个挑战仍未解决：语言歧义、时序对齐漂移以及稀有类别泛化。

OVVS 将开放词汇识别嫁接到某个宿主 VSP 任务之上。（*VIS*） OVFormer [42] 通过统一的嵌入对齐与半在线推理，将视觉—语言特征与实例查询对齐；OV2Seg+ [169] 将通用提议生成器与记忆引导的跟踪器以及 VLM 驱动的分类器结合起来；CLIP-VIS [230] 通过类别无关的掩码生成和时序 top-K 查询匹配来适配冻结的 CLIP 骨干网络。（*VPS*） ODISE [193] 利用文本到图像扩散模型的特征进行图像级的开放词汇全景分割。我们将其列为一个能为视频流程提供开放词汇全景识别能力的组件，而非原生的时序 VPS 模型。（*VSS*） OV2VSS [88] 加入了时空融合模块、随机帧增强以及专门的视频—文本编码器。这三类方法遵循同一套蓝图：把识别交给冻结的视觉—语言骨干网络，只保留一个轻量的时序或对齐适配器用于训练。

共性问题。 OVVS 的关键约束在于识别，而非掩码生成。由于当前所有流程都把分类交给冻结的图像级视觉—语言骨干网络 [144]，只训练轻量的适配器，因此三种失效模式反复出现。*(i)* *粒度歧义*：同一个分割区域可以合理地匹配多个词表层级（“交通工具” / “汽车” / “轿车”），因此所报告的分数在一定程度上取决于测试词表的措辞（§2.1）。*(ii)* *语义闪烁*：由于视觉—语言嵌入是在静态图像上训练的，同一目标可能随姿态或光照的变化在不同的文本邻域间漂移，从而违背了封闭集 VSP 视为理所当然的类别一致性约束。*(iii)* *长尾与零样本泛化相互耦合*：即便骨干网络相当，在 VSPW（表 9）和 YouTube-VIS（表 10）上与有监督方法之间的巨大差距依然存在，这表明瓶颈在于开放词汇识别，而非时序建模（参见 §3.6）。

方法比较。 当前 OVVS 设计的主要区别在于视觉—语言边界位于何处。适配器式方法（CLIP-VIS [230]、OV2VSS [88]）保持 VLM 冻结，保留了零样本通用性，但限制了语言对掩码的塑造程度；对齐式方法（OVFormer [42]）端到端地训练嵌入对齐，以向已见类别偏移的风险换取基类上的精度；解耦流程（OV2Seg+ [169]、ODISE [193]）则将类别无关的掩码生成与 VLM 分类分离开来。值得注意的是，三条路线在 YouTube-VIS 上的 AP 彼此相差仅数个点（表 10），却都落后于有监督方法 20 个点以上，这表明关键约束在于冻结的图像级 VLM 本身，而非某种特定的适配器设计。

3.6 跨任务讨论

尽管 VSS、VIS、VPS、VTS 和 OVVS 通常被视为各有其指标、数据集与架构的独立任务，但每项任务都必须回答同样的三个问题：一个像素表示什么、它属于哪个目标，以及这一归属如何随时间保持。各任务的区别主要在于把哪个问题放在首位。这一共同基础孕育了若干横跨各任务的挑战，没有任何一项任务能够单独解决它们。

3.6.1 **各任务的挑战—方案—范式图谱。** 在讨论共性主题之前，我们先用统一的“挑战—方案—范式”结构解读每项任务。**VSS** 关注逐像素的时序稳定性（mVC_n），从光流扭曲 [46] 发展到注意力 [159]，再到状态空间建模 [61]。**VIS** 关注跨帧身份，从基于检测的跟踪 [205] 发展到集合预测 [64, 221]，再到对比查询记忆 [83, 210]。**VPS** 追求均衡且互不重叠的 thing/stuff 输出，方法涵盖双分支融合 [76, 118]、查询统一 [90, 91] 以及深度感知的 3D 扩展 [143, 213]。**VTS** 关注长时程身份，方法包括联合检测与跟踪 [168, 196]、解耦的 SAM 式传播 [28, 145]，以及点云或分配空间嵌入 [32, 198]。**OVVS** 以开放集泛化为目标，方法包括 CLIP 对齐 [144]、

视觉—语言适配器 [42, 170] 以及无需掩码的文本驱动分割 [193]。纵观各任务，反复出现的规律是：显式先验让位于隐式的集合或查询预测，随后又走向显式的记忆或对比式稳定化。

3.6.2 **时序一致性**。时序一致性是所有 VSP 任务共同面临的挑战，只是在各任务中的表现形式不同：VSS 要求运动下的逐像素标签保持稳定（VSPW [122] 上的 mVC_n），VIS 和 VPS 要求身份稳定，VTS 则要求长时程的稳定性。各方法依赖三种机制：显式的光流扭曲 [46, 67]、隐式的跨帧注意力或记忆 [63, 210]，以及预训练查询内在的连贯性 [64, 221]。没有一种机制能够全面胜出：光流擅长短程，注意力擅长中程，而在强大基础模型的支撑下，查询连贯性擅长长程。因此，近期的最优方法往往将三者结合使用，时序一致性最好被理解为一种多尺度的性质。

3.6.3 **遮挡下的身份保持**。每当一个目标消失后又重新出现，所有携带身份的任务（VIS、VPS、VTS 以及实例级 OVVS）都必须在早已变化的背景中重新识别它。这是实践中最显眼的失效：KITTI-MOTS [168] 上的身份切换、YouTube-VIS 与 OVIS 之间的差距 [141]，以及长 VIPSeg [121] 序列上 VPQ 的急剧衰减，都可以追溯到同一个弱点。应对途径有两条：基于对比记忆的关联，以内存占用换取重识别的鲁棒性 [186, 210]；以及基于 SAM 的传播 [78]，将身份交给一个与任务无关的链接器 [28, 145]。两者都未能彻底解决长时间遮挡问题，而短片段与长片段基准之间的差距仍是最可靠的诊断依据。

3.6.4 **长尾与开放世界泛化**。随着数据集从少数几个城市场景类别（Cityscapes [33]、KITTI [49]）扩展到大词表语料（VSPW [122]、VIPSeg [121]、LV-VIS [170]），五项任务都出现了长尾现象：稀有类别被系统性地分割不足，而未见类别则完全超出了封闭词汇模型的能力范围。OVVS 明确采用了开放世界假设，但其 mIoU 和 AP（表 9、表 10）仍大幅落后于有监督方法，这说明开放词汇识别目前受限于泛化能力，而非词表规模。共同的瓶颈在于：以掩码训练的视觉表征与视觉—语言基础模型 [144] 所蕴含的更广泛语义先验之间的对齐。因此，任何一项任务上的进展（例如 OVFormer [42]、ODISE [193]）都应能迁移到其他任务，这也支持建立一个统一的跨任务研究议程。

3.6.5 **标注—容量—延迟三难困境**。每项 VSP 任务都受制于标注成本、模型容量与延迟三者之间的权衡。像素精确且时间一致的标注，其成本比图像掩码高出一个数量级，因此高标注效率学习在 VSP 中反复出现：VSS 采用自训练/一致性约束 [19, 62, 82]，VIS 采用伪标签传播/无监督合成 [44, 103, 177]，VTS 和 OVVS 则将基础模型 [78, 144] 用作零样本/少样本监督。然而，容量的提升（更宽的骨干网络、更长的预训练、更重的查询注意力）是以延迟为代价的（参见 表 8），与自动驾驶和具身感知的需求不相容。这三个维度难以独立优化：高数据利用效率的学习需要足够的容量来吸收带噪声的监督，而实时部署又需要在不损失时序连贯性的前提下对这些容量进行蒸馏。对三难困境进行统一处理，是当前 VSP 中尚未得到充分探索的研究方向。

3.6.6 **典型失效模式**。少数几种具体的失效反复出现，而每一种都指向一个更深层的局限。*(i)* 长时间遮挡：基于查询的 VIS Transformer（MinVIS [64]、DVIS [221]）在 OVIS [141] 上的 AP 比 YouTube-VIS 低 10 个点以上；一旦遮挡解除前的间隔超过几秒，基于 SAM 的传播 [28, 145] 就会丢失身份，这表明身份被编码为一种短时程的性质。*(ii)* 快速运动与运动模糊：基于光流的 VSS [36, 46, 67] 和基于检测的跟踪式 VIS [183, 205] 依赖帧间对应关系，而这种对应在大位移和模糊下会崩溃，导致重影和轨迹断裂；注意力聚合只能缓解症状，而无法消除病因。*(iii)* 拥挤场景中的身份切换：在 VIPSeg [121] 和 BDD100K [212] 上，解耦的跟踪器只能依赖区分度不足的外观嵌入，导致 thing-VPQ 停滞在 stuff-VPQ 之下。对比嵌入 [186, 210] 提高了上限，但剩余误差仍与人群密度相关。*(iv)* 稀有与细粒度类别：VSPW [122]（124 个类别）和 LV-VIS [170]（1,196 个类别，其中 553 个作为新类别留出）都呈现严重的长尾类别分布，稀有类别在这些数据集中被系统性地分割不足，因为封闭词汇训练从未内化稀有类别的语义先验，而当前的 OVVS 方法 [42, 170, 193] 只部分弥补了这一差距。

表 6 视频场景解析数据集统计。

数据集	年份	发表	视频数	训练/验证/测试/开发	标注类型	用途	类别数
CamVid [11]	2009	PRL	4	（帧：467/100/233/-）	VSS	城市	11
NYUDv2 [154]	2012	ECCV	518	（帧：795/654/-/-）	VSS	室内	40
Cityscapes [33]	2016	CVPR	5,000	2,975/500/1,525	VSS	城市	19
ACDC [149]	2021	ICCV	-	（帧：1600/406/2000）	VSS	城市	19
VSPW [122]	2021	CVPR	3,536	2,806/343/387/-	VSS	通用	124
MVSeg [69]	2023	CVPR	738	452/84/202	VSS	城市	26
YouTube-VIS [205]	2019	ICCV	3,859	2,985/421/453/-	VIS	通用	40
KITTI MOTS [168]	2019	CVPR	21	12/9/-/-	VIS	城市	2
MOTSChallenge [168]	2019	CVPR	4	-/-/-/-	VIS	城市	1
OVIS [141]	2021	IJCV	901	607/140/154/-	VIS	通用	25
BDD100K [212]	2020	ECCV	100,000	7,000/1,000/2,000/-	VSS, VIS	驾驶	40 (VSS), 8 (VIS)
Cityscapes-VPS [76]	2020	CVPR	500	400/100/-/-	VPS	城市	19
VIPER-VPS [76]	2020	CVPR	124	（帧：134K/50K/70K/-）	VPS	城市	23
KITTI-STEP [179]	2021	NeurIPS	50	12/9/29	VPS	城市	19
MOTChallenge-STEP [179]	2021	NeurIPS	4	2/2/-/-	VPS	城市	7
VIPSeg [121]	2022	CVPR	3536	2806/343/387	VPS	通用	124
WOD: PVPS [118]	2022	ECCV	2860	2800/20/40	VPS	城市	28
APOLLO MOTS [198]	2020	ECCV	-	（帧：3.4k/2.3k/5.7k/-）	VIS	城市	-
HiEve [100]	2023	IJCV	32	19/0/13/-	VTS	人体	14
DAVIS16 [136]	2016	CVPR	50	30/20/-/-	VOS（目标级）	通用	-
DAVIS17 [139]	2017	-	150	60/30/30/30	VOS（实例级）	通用	-

3.6.7 **跨任务定量综合**。将表 8–表 12 并列阅读，可以发现三个规律。架构趋同：除 VTS 外，每项任务的精度领先者都采用 Transformer 解码器，且均出现于 2022–2024 年（VPSeg 与 TubeFormer 分别为 82.5 与 63.2 mIoU，CTVIS 为 55.1 AP，PolyphonicFormer 与 TubeLink 分别为 65.4 VPQ 与 49.4 STQ），带来的收益并不均匀（VSPW/Cityscapes mIoU 分别为 Δ+25.7/+7.4，VIS AP 为 +24.8，Cityscapes-VPS/VIPSeg 为 +8.0/+17.9）；VTS 是例外，其中一种点云嵌入方法相对其基于检测的跟踪基线仍带来 +9.3 的 sMOTSA 提升，并使 IDS 降至基线的约四分之一。效率断崖：在这些各任务的精度领先者中，只有 PointTrack 能够实时运行（22.2 FPS），而 VSS/VIS/VPS 的领先者要么只有个位数 FPS，要么未报告 FPS；BLO（30.8 FPS）等面向速度的设计比精度前沿低数个 mIoU，因此三难困境中的“延迟”一边在各任务中都构成约束。标注天花板：开放词汇方法在 VSPW 上比有监督方法低 46 mIoU，在 YouTube-VIS 上低 20–23 AP，这 20–46 个点的差距将瓶颈定位于有标注的监督，而非模型容量。

关于跨时代比较的说明。这些差值横跨 2016–2025 年，骨干网络从 ResNet-18 到 Swin-L 不等，预训练和训练策略也各不相同，因此模型容量、数据与训练方案都与我们所归因的架构变化相互混杂；与 FPS 数值一样，它们反映的是范式层面的变化，而非某一架构的独立贡献。

因此，VSS、VIS、VPS、VTS 和 OVVS 之间的界限在很大程度上是基准设计的产物。其背后的问题是共通的，而进展越来越多地来自以这一共同基础为出发点的方法。

4 数据集与评价指标

我们概述常用的 VSP 数据集（§4.1）与评价指标（§4.2）。

4.1 数据集

我们按子任务组织具有代表性的 VSP 数据集，每个数据集给出一行简介，完整的数值见 表 6。

4.1.1 VSS **数据集**。

- **CamVid [11]**：4 段驾驶视频 / 800 个已标注帧（467/100/233）；首个带有语义类别标签的视频数据集。

- **NYUDv2 [154]**：1,449 帧由 Kinect 采集的稠密标注 RGB-D 室内图像；提供类别与实例 ID，用于室内分割。
- **Cityscapes [33]**：来自 50 座城市的双目驾驶序列 / 5,000 张精细标注图像和 20,000 张粗略标注图像（均提供像素级标注）/ 30 个类别；经典的城市场景 VSS 基准。
- **ACDC [149]**：4,006 张恶劣天气图像（雾/夜间/雨/雪），并配有正常条件下的对应图像；面向鲁棒性评测。
- **VSPW [122]**：3,536 段视频 / 251,632 帧 / 124 个类别 / 15 FPS / 200+ 种场景；首个大规模、多场景、稠密标注的 VSS 基准。
- **MVSeg [69]**：738 段配对的 RGB—热红外视频 / 3,545 帧昼夜城市场景图像 / 26 个类别；面向多光谱 VSS。

4.1.2 VIS 数据集。

- **YouTube-VIS [205]**：3,859 段视频（2,985/421/453 划分）/ 40 个类别 / 131K 个掩码；由 YouTube-VOS 扩展而来，是经典的 VIS 基准。
- **KITTI MOTS [168]**：21 段驾驶视频 / 8,008 帧 / 26,899 个汽车标注 + 11,420 个行人标注；用于联合检测—跟踪—分割。
- **MOTSChallenge [168]**：4 段视频 / 2,862 帧 / 26,894 个行人掩码；仅含行人的拥挤场景跟踪基准。
- **OVIS [141]**：901 段视频（$\bar{t} \approx 12.8$ 秒）/ 25 个类别 / 296K 个掩码 / 5,223 个实例；面向严重遮挡。

4.1.3 VPS 数据集。

- **Cityscapes-VPS [76]**：2,400/300/300 帧 / 每个 30 帧片段标注 6 帧 / 19 个类别；城市场景 VPS 基准。
- **VIPER-VPS [76]**：254K 帧分辨率为 1080×1920 的合成 GTA-V 驾驶画面 / 10 个 thing 类 + 13 个 stuff 类；大规模的合成 VPS 数据源。
- **KITTI-STEP [179]**：50 段视频 / 18,181 帧 / 2 个 thing 类 + 17 个 stuff 类 / 126K 个掩码；遵循 STEP 协议的城市场景基准。
- **MOTChallenge-STEP [179]**：4 段视频 / 2,075 帧 / 1 个 thing 类 + 6 个 stuff 类；行人 STEP 基准。
- **VIPSeg [121]**：3,536 段视频 / 84,750 帧 / 232 种场景 / 58 个 thing 类 + 66 个 stuff 类 / 926K 个掩码；规模最大的开放场景 VPS 基准。
- **WOD:PVPS [118]**：2,860 段视频 / 100K 帧 / 8 个跟踪类别 + 28 个语义类别；多相机、时间一致的全景标注。

4.1.4 VTS 数据集。

- **APOLLO MOTS [198]**：11.4K 帧 ApolloScape 图像 / 平均每帧 5.65 辆汽车（密度为 KITTI MOTS 的 2×，拥挤汽车数量为其 2.5×）；以汽车为中心，用于 2D/3D 跟踪。
- **HiEve [100]**：1M 个姿态标注 / 56K 个复杂事件中的动作实例 / 轨迹平均长度 480+ 帧；以人为中心。
- **DAVIS16 [136]**：50 个序列（30/20）/ 3,455 帧 / 每个片段一个目标；经典的短片段 VOS 式基准。
- **DAVIS17 [139]**：150 个序列 / 10,459 帧 / 376 个目标；包含干扰物、遮挡和快速运动的多目标片段。

4.2 评价指标

我们回顾 VSP 中最常用的十三种评价指标，按照从原始重叠度到时序、效率和身份感知类指标的顺序展开，最后给出各任务的指标报告建议。

交并比（IoU）。 IoU 是本节其余指标所依托的基本模板，即预测掩码与其真值之间的 Jaccard 重叠度：

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{7}$$

其中 TP/FP/FN 分别为真正例、假正例和假负例。IoU 同时惩罚过分割与分割不足。

平均交并比（mIoU）。 对 C 个类别的 IoU 取平均，可避免高频类别淹没稀有类别：

$$\mathrm{mIoU} = \frac{1}{C}\sum_{i=1}^{C} \mathrm{IoU}_i. \tag{8}$$

正是这种类别平衡性，使 mIoU 成为 VSS 的首要指标。

视频一致性（VC）。 VC 检验标签在 n 个连续帧中是否保持稳定。令 S_i/S_i' 分别为第 i 帧的真值/预测标签，C 为片段长度，则

$$\mathrm{VC}_n = \frac{1}{C-n+1}\sum_{i=1}^{C-n+1} \frac{\left|\left(\bigcap_{j=0}^{n-1} S_{i+j}\right) \cap \left(\bigcap_{j=0}^{n-1} S'_{i+j}\right)\right|}{\left|\bigcap_{j=0}^{n-1} S_{i+j}\right|}, \tag{9}$$

数据集级别的平均值为 $\mathrm{mVC}_n = \frac{1}{N}\sum_{k=1}^{N} \mathrm{VC}_n^{(k)}$。$\mathrm{mVC}_n$ 越高，表示时序闪烁越少。

每秒帧数（FPS）。 FPS 报告原始吞吐量，$\mathrm{FPS} = 1/\bar{t}$，其中 $\bar{t} = \frac{1}{N}\sum_{i=1}^{N} t_i$。由于它是一个平均值，因此会掩盖尾部延迟。

最大延迟（Max Latency）。 最大延迟捕捉了 FPS 所掩盖的信息，即最慢单帧的耗时 $\max_i\{t_i\}$。对于自动驾驶或 AR 等硬实时场景，这一指标才是关键，因为一次延迟尖峰就足以打断感知—控制回路。

平均精度（AP）。 AP 之于 VIS，正如 mIoU 之于 VSS：它将图像指标推广到视频，并在 $|T|$ 个 IoU 阈值上取平均：

$$\mathrm{AP} = \frac{1}{|T|}\sum_{t\in T} \mathrm{AP}_t. \tag{10}$$

在多个阈值下进行评估，可以同时考察掩码的精确程度与检测结果的置信度，这使 AP 成为 YouTube-VIS 和 OVIS 上 VIS 的主要指标。

平均召回率（AR）。 当模型对每段视频返回其前 K 个实例时所能达到的最佳召回率：

$$\mathrm{AR} = \frac{1}{|V|}\sum_{v\in V} \frac{\mathrm{TP}_v(K)}{\mathrm{GT}_v}. \tag{11}$$

AR_{10} 与 AP 配合使用，作为召回率的上限参考。

分割与跟踪质量（STQ）。 STQ 将视频跟踪拆分为两个问题：实例是否保持了其身份（关联质量，AQ），以及其像素是否被正确标注（类别级分割质量，SQ）：

$$\mathrm{AQ} = \frac{1}{|G|}\sum_{z_g\in G} \frac{1}{|gid(z_g)|} \sum_{z_f\,:\, z_f\cap z_g\neq\emptyset} \mathrm{TPA}(z_f, z_g)\times \mathrm{IoU}_{id}(z_f, z_g), \tag{12}$$

$$\mathrm{SQ} = \frac{1}{|C|}\sum_{c\in C} \frac{|f_{\mathrm{sem}}(c)\cap g_{\mathrm{sem}}(c)|}{|f_{\mathrm{sem}}(c)\cup g_{\mathrm{sem}}(c)|}, \qquad \mathrm{STQ} = \sqrt{\mathrm{AQ}\times\mathrm{SQ}}, \tag{13}$$

几何平均将两者融合为一个分数，而单独查看 AQ 则能揭示帧级指标所遗漏的身份错误。

多目标跟踪与分割精度（MOTSP）。 MOTSP 仅评价掩码形状，即在已匹配对 $\mathcal{M}_t$ 上的平均 IoU：

$$\mathrm{MOTSP} = \frac{\sum_t \sum_{(i,j)\in\mathcal{M}_t} \mathrm{IoU}(i,j)}{\sum_t |\mathcal{M}_t|}. \tag{14}$$

表 7　各任务的指标选择指南：主要指标、推荐的诊断指标，以及该诊断指标可揭示的失效模式。

任务	首要指标	诊断指标	可揭示的失效模式
VSS	mIoU	mVC_8、mVC_{16}；FPS	时序闪烁
VIS	AP	AR_{10}；STQ-AQ	遮挡下的身份切换
VPS	VPQ、STQ	$\mathrm{VPQ}_{\mathrm{thing}}/\mathrm{VPQ}_{\mathrm{stuff}}$；$\mathrm{PQ}=\mathrm{SQ}\times\mathrm{RQ}$	thing/stuff 不平衡；区分边界误差与识别误差
VTS	sMOTSA	IDS；MOTSA；MOTSP	遮挡下的身份漂移
OVVS	宿主任务分数	新类别与基类分数的对比	零样本泛化与已见类别拟合的对比

由于只在已匹配的对上取平均，它能在不受任何身份错误影响的情况下反映边界质量。

身份切换（IDS）。 IDS 统计原始的身份交换次数，即其所匹配的预测与前一帧相比携带了不同 ID 的真值掩码的数量：

$$\mathrm{IDS}=\left|\{m\in M\mid c^{-1}(m)\neq\emptyset\ \wedge\ pred(m)\neq\emptyset\ \wedge\ id_{c^{-1}(m)}\neq id_{c^{-1}(pred(m))}\}\right|. \tag{15}$$

下文的 MOTSA 与 sMOTSA 正是将这种错误折算进了单一的准确率数值中。

多目标跟踪与分割准确率（MOTSA）。 MOTSA 综合考虑正确匹配、假正例与身份切换：

$$\mathrm{MOTSA}=\frac{\sum_t(|TP_t|-|FP_t|-|IDS_t|)}{\sum_t|GT_t|}. \tag{16}$$

软 MOTSA（sMOTSA）。 这是一种以 IoU 加权的 MOTSA，它按掩码 IoU 对 TP 进行软化，而非将其整体计数：

$$\mathrm{sMOTSA}=\frac{\sum_t(\sum_{(i,j)\in\mathcal{M}_t}\mathrm{IoU}(i,j)-|\mathrm{FP}_t|-|\mathrm{IDS}_t|)}{\sum_t|\mathrm{GT}_t|}, \tag{17}$$

这样一来，一个松散但正确的掩码不再与一个紧致的掩码得到相同的分数。

视频全景质量（VPQ）。 VPQ 将全景质量推广到时间维度，在 k 帧窗口上进行评分。对于每个片段，它计算每一对真值轨迹与预测轨迹之间的 IoU，IoU> 0.5 时计为一个 TP，FP/FN 照常处理。于是

$$\mathrm{VPQ}_k=\frac{1}{N_{classes}}\sum_c\frac{\sum_{(u,\hat{u})\in T_P^c}\mathrm{IoU}(u,\hat{u})}{|T_P^c|+\frac{1}{2}|F_P^c|+\frac{1}{2}|F_N^c|}. \tag{18}$$

当 $k=0$ 时，它退化为图像级的 PQ。当 $k>0$ 时，窗口内的任何身份切换都会产生一个 FP/FN，因此 VPQ 具备了帧级 mIoU 所缺乏的显式时序敏感性。

分解形式及其对 thing/stuff 的诊断价值。 PQ 可分解为分割质量（SQ）与识别质量（RQ）的乘积：

$$\mathrm{PQ}=\underbrace{\frac{\sum_{(u,\hat{u})\in T_P}\mathrm{IoU}(u,\hat{u})}{|T_P|}}_{\mathrm{SQ}}\times\underbrace{\frac{|T_P|}{|T_P|+\frac{1}{2}|F_P|+\frac{1}{2}|F_N|}}_{\mathrm{RQ}}, \tag{19}$$

其中较低的 SQ 意味着掩码边界质量差，较低的 RQ 意味着识别/匹配失败。因此，VPS 基准分别在 thing 与 stuff 上报告 VPQ，并以 $\mathrm{VPQ}_{\mathrm{thing}}-\mathrm{VPQ}_{\mathrm{stuff}}$ 量化 thing 与 stuff 之间的不平衡（§3），这是 mIoU（VSS）和 AP（VIS）都无法提供的诊断信息。

各任务的指标选择指南。 每项任务都将一个主要指标与一个诊断分数配对使用，从而使模型能力与失效模式都能得到体现。表 7 汇总了五项任务的推荐搭配。没有任何一个主要指标能够独立反映模型表现：一个时序不稳定或频繁发生身份切换的模型依然可能获得很高的 mIoU、AP 或 VPQ，而且若干指标只有结合其测量协议才能得到正确的解读。

表 8 VSS 方法在 Cityscapes 验证集 [33] 上的定量结果（mIoU 与 FPS）。

方法	骨干网络	mIoU	FPS
DVSNet [197]	ResNet-101	70.4	19.8
LVS [94]	ResNet-101	76.8	5.8
GRFP(5) [125]	ResNet-101	69.4	3.3
Accel [67]	ResNet-101	75.5	1.1
TDNet [63]	ResNet-50	79.9	5.6
ETC [106]	ResNet-18	73.1	9.5
DepthMix [62]	ResNet-101	71.2	1.9
CIRKD [204]	ResNet-18	74.7	-
MRCFA [161]	MiT-B1	75.1	21.5
CFFM [159]	MiT-B1	75.1	23.6
MPVSS [180]	Swin-L	81.6	7.2
BLO [203]	TopFormer	74.7	**30.8**
SSLTM [82]	ResNet-50	79.7	-
VPSeg [54]	MiT-B3	**82.5**	-
CFFM++ [160]	MiT-B1	78.7	20.4
TV3S [61]	MiT-B1	75.6	-

表 9 VSS 在 VSPW 验证集 [122] 上的基准结果。以 mIoU、mVC 和 FPS 报告结果。灰色方法使用开放词汇监督。

方法	骨干网络	mIoU	mVC_8	mVC_{16}	FPS
CFFM [159]	MiT-B5	49.3	90.8	87.1	11.3
CFFM++ [160]	MiT-B5	50.1	90.8	87.4	10.4
MRCFA [161]	MiT-B5	49.9	90.9	87.4	–
TCB [122]	ResNet-101	37.5	87.0	82.1	10.0
Video K-Net [91]	ResNet-101	38.0	87.2	82.3	–
MPVSS [180]	ResNet-101	38.8	84.8	79.6	–
TV3S [61]	MiT-B5	49.8	91.7	**88.7**	–
TubeFormer [77]	Axial-ResNet	**63.2**	**92.1**	87.9	–
OV2VSS [88]	ResNet-101	17.2	-	-	–

5 性能比较

我们在标准基准上比较所综述的 VSP 方法，重点关注评测协议较为成熟的 VSS、VIS 和 VPS。

5.1 VSS 基准

实验设置。 在 Cityscapes [33] 上，我们报告 mIoU 和 FPS。对于 表 8，FPS 数值取自文献 [176] 中的统一基准测试（单块 NVIDIA RTX 2080 Ti），在没有公开实现时，少数条目借用原论文的数值。对于 表 9（VSPW），FPS 数值遵循 CFFM++ [160] 的设定，在单块 RTX 6000（24G）上以 480 × 853 的分辨率测得。标记为“–”的条目表示在这些协议下未报告。

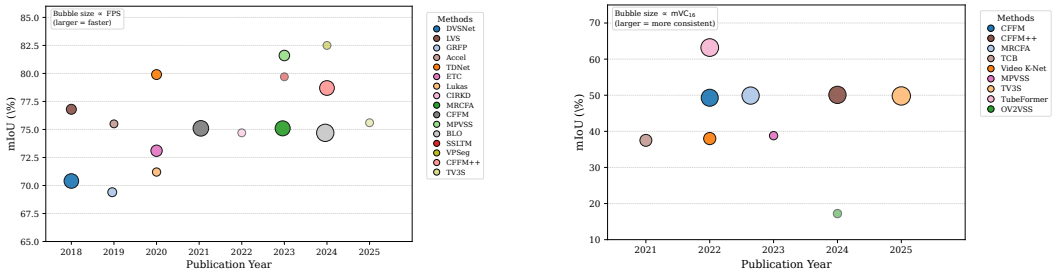
5.1.1 **结果**。表 8 和 表 9 表明，VSS 的性能提升在不同约束下来自不同的机制。在 Cityscapes 上，注入强几何先验或场景先验的方法（如 VPSeg [54]）往往能提高峰值精度，因为城市布局相对结构化。相比之下，BLO [203] 和 CFFM 系列 [159, 160] 在效率方面表现出色，因为它们避免了为每一帧重新计算稠密特征——要么剪除冗余的关键帧，要么复用时间上稳定的特征。这解释了为什么最佳精度与最佳速度并非由同一种设计取得：一种利用更强的空间先验，另一种则利用时间冗余。在 VSPW 上，TubeFormer [77] 受益于片段级的时空管建模，这与该数据集视频长且多样的特点直接契合；而 TV3S [61] 则通过状态空间传播提升时序一致性。OV2VSS [88] 的剩余差距表明，开放词汇 VSS 目前更多地受限于类别识别，而非时序平滑。

趋势分析。 图 4 将同样的机制层面差异沿时间展开。在 Cityscapes 上，更强的逐帧解析能力与城市场景先验有助于提高性能，而最大的效率提升来自削减时间冗余的设计，而非扩大骨干网络。VSPW 则将压力转向长程一致性与场景多样性，这有利于基于时空管和基于状态空间的方法，因为它们能在更长的片段上保留信息。开放词汇方法的数据点始终与有监督方法的轨迹相分离，这表明当语义分类器无法可靠地识别未见或稀有类别时，仅仅增加时序上下文是不够的。

5.2 VIS 基准

实验设置。 我们遵循文献 [205]，在 YouTube-VIS 验证集上报告基于时空 IoU 的 AP 和 AR。在 表 10 中，FPS 数值来自各方法的原始论文，或相同设置下的第三方比较。我们用上标标注硬件差异：† V100，‡ RTX 2080 Ti，§ GTX 1080 Ti，¶ A100。比较 FPS 时应考虑到这一点。

5.2.1 **结果**。 表 10 表明，近期 VIS 的进展与其说来自更大的骨干网络，不如说来自更好的实例级时序推理。最强的有监督方法都采用基于查询的形式，其中可学习的查询为每个目标提供一个稳定的“槽位”，从而联合解决分割与关联问题。CTVIS [210] 进一步推进了这



(a) Cityscapes 验证集（气泡大小 ∝ FPS）。 (b) VSPW 验证集（气泡大小 ∝ mVC_{16}）。

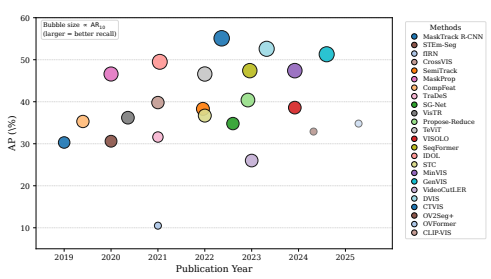
图 4 VSS 方法在 Cityscapes 与 VSPW 上的性能演进。颜色表示方法；(a) 中气泡大小 = FPS，(b) 中气泡大小 = mVC_{16}。

表 10 VIS 方法在 YouTube-VIS 验证集 [205] 上的精度与效率评估。灰色方法使用开放词汇监督。FPS 的上标（†/‡/§/¶）表示所用的 GPU，图例见本基准的“实验设置”段落。没有上标的 FPS 条目来自未披露 GPU 型号的论文。

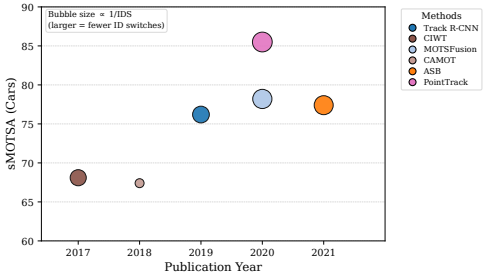
方法	骨干网络	AP50	AP75	AR1	AR10	AP	FPS
MaskTrack R-CNN [205]	ResNet-50	51.1	32.6	31.0	35.5	30.3	20.0†
STEm-Seg [4]	ResNet-50	50.7	33.5	31.6	37.1	30.6	7.0†
fIRN [103]	ResNet-50	27.2	6.2	12.3	13.6	10.5	–
CrossVIS [206]	ResNet-50	56.8	38.9	35.6	40.7	39.8	39.8†
SemiTrack [44]	ResNet-50	61.1	39.8	36.9	44.5	38.3	–
MaskProp [9]	STSN-ResNeXt-101-64x4d	-	51.2	44.0	52.6	46.6	–
CompFeat [45]	ResNet-50	56.0	38.6	33.1	40.3	35.3	–
TraDeS [183]	DLA-34	52.6	32.8	30.1	29.1	31.6	–
SG-Net [102]	ResNet-50	56.1	36.8	35.8	40.8	34.8	–
VisTR [178]	ResNet-50	59.8	36.9	37.2	42.4	36.2	69.9†
Propose-Reduce [97]	ResNet-50	63.0	43.8	41.1	49.7	40.4	–
TeViT [207]	MsgShifT	71.3	51.6	44.9	54.3	46.6	68.9†
VISOLO [59]	ResNet-50	56.3	43.7	35.7	42.5	38.6	40.0‡
SeqFormer [184]	ResNet-50	69.8	51.8	45.5	54.8	47.4	72.3†
IDOL [186]	ResNet-50	74.0	52.9	47.7	58.7	49.5	30.6
STC [70]	ResNet-50	57.2	38.6	36.9	44.5	36.7	40.3‡
MinVIS [64]	ResNet-50	69.0	52.1	45.7	55.7	47.4	–
GenVIS [60]	ResNet-50	72.0	57.8	49.5	60.0	51.3	35.6¶
VideoCutLER [177]	ResNet-50	50.7	24.2	-	42.4	26.0	–
DVIS [221]	ResNet-50	76.5	58.2	47.4	60.4	52.6	18.7§
CTVIS [210]	ResNet-50	**78.2**	**59.1**	**51.9**	**63.2**	**55.1**	–
OV2Seg+ [169]	ResNet-50	-	-	-	-	32.9	–
OVFormer [42]	ResNet-50	-	-	-	-	34.8	–
CLIP-VIS [230]	ResNet-50	-	-	-	-	32.2	–

一方向，使查询表征在帧间更具判别性，直接针对遮挡下的身份切换问题。AR 上的差距仍然表明，检索出多个合理的实例要比把正确的长时关联排在首位更容易，因此 VIS 对查询之间的竞争和身份歧义依然敏感。开放词汇方法即使骨干网络相当也依然落后，这表明其主要弱点在于对未见类别的语义定位（grounding），而不仅仅是掩码生成。

趋势分析。 图 5a 阐明了架构上的转变。早期的基于检测的跟踪系统依赖逐帧提议，因此在关联之前就已继承了检测误差。后来基于查询的浪潮之所以带来提升，是因为它将实例视为贯穿整个片段的持久集合元素，使掩码与身份能够一同被精修。这解释了精度与召回率为何同时提高：更好的查询既减少了重复或碎片化的轨迹，又找回了更多的目标。位置更低的开放词汇方法簇标志着另一种瓶颈：一旦引入以语言为条件的类别分配，即使时序关联很强，识别的不确定性也会成为限制因素。



(a) YouTube-VIS 验证集：气泡大小 $\propto \mathrm{AR}_{10}$。



(b) KITTI-MOTS 汽车类验证集：气泡大小 $\propto 1/\mathrm{IDS}$。

图 5　VIS 与 VTS 方法的性能演进。左：YouTube-VIS 上的 VIS，气泡 $= \mathrm{AR}_{10}$。右：KITTI-MOTS 汽车类上的 VTS，气泡 $= 1/\mathrm{IDS}$（越大表示越稳定）。

表 11　VPS 验证集上的基准结果（VPQ、STQ、FPS）。

方法	骨干网络	Cityscapes-VPS [76] (VPQ)	KITTI-STEP [179] (STQ)	VIPSeg [121] (STQ)	FPS
VPSNet [76]	ResNet-50	57.4	-	-	1.3
VIP-Deeplab [143]	ResNet-50	63.1	-	-	~2.5
SiamTrack [181]	ResNet-50	57.8	-	-	–
Clip-PanoFCN [121]	ResNet-50	-	-	31.5	–
PolyphonicFormer [213]	ResNet-50	**65.4**	-	-	–
TubeFormer [77]	Axial-ResNet	-	70.0	-	–
SLOT-VPS [228]	Swin-L	63.7	-	-	–
TubeLink [90]	Swin-B	-	72.2	**49.4**	–
Video K-Net [91]	Swin-B	62.2	**73.0**	46.3	4.9

5.3　VPS 基准

实验设置。　我们在 Cityscapes-VPS、KITTI-STEP 和 VIPSeg 上报告 VPQ 和 STQ。表 11 中的 FPS 数值汇集了原始论文与跨方法比较的结果：VPSNet-FuseTrack 取自文献 [209]（Cityscapes-VPS 验证集），ViP-DeepLab 按文献 [138] 的估计约为每帧 >400 ms（≈2.5 FPS），Video K-Net 取自文献 [28]。这些工作所用的硬件各不相同（例如 V100 与 Titan X），因此 FPS 的绝对值仅具参考意义，不可严格比较。

5.3.1　**结果**。表 11 表明，VPS 的性能与数据集结构高度耦合。PolyphonicFormer [213] 在 Cityscapes-VPS 上表现良好，因为其统一的查询设计与深度感知推理契合结构化的驾驶场景；而 Video K-Net [91] 和 TubeLink [90] 在更有利于时空管级或基于卷积核的时序分组方法的基准上表现更好。没有哪种方法能在所有基准上领先，这说明 VPS 方法仍围绕 thing 跟踪、stuff 解析与时序聚合之间的特定平衡进行优化。当基准从紧凑的城市标签集转向更广泛的开放场景词表时，关键难点便从全景融合转向长尾识别与稳定的 thing—stuff 分配。延迟仍只是一个次要的报告维度，这说明该领域优先追求统一表征的质量，其次才是实时部署。

趋势分析。　VPS 的趋势分为两种情形。紧凑的驾驶基准进步很快，因为其场景布局、类别集合和运动模式都足够规则，更强的查询融合或深度感知融合便能减少大部分误差。VIPSeg 仍然更难，因为通用视频暴露出长尾类别、多样的场景以及更难预测的 thing—stuff 边界。这意味着未来 VPS 的提升可能不再主要依靠又一个全景融合头，而更多地来自更好的类别先验、高数据利用效率的长尾学习，以及能够在不降低 stuff 一致性的前提下保持 thing 身份稳定的时序机制。

5.4　VTS 基准

实验设置。　我们在表 12 中报告 KITTI-MOTS 验证集上的 sMOTSA、MOTSA 和 IDS。Track R-CNN、MOTSFusion 和 PointTrack 的 FPS 数值由 PointTrack [198] 在统一的单 GPU 协议下

表 12 VTS 方法在 KITTI-MOTS 验证集 [168] 上的评估。

方法	检测器	汽车			行人			FPS
		sMOTSA	MOTSA	IDS	sMOTSA	MOTSA	IDS	
Track R-CNN [168]	Track R-CNN	76.2	87.8	93	46.8	65.1	78	2.0
CIWT [129]	Track R-CNN	68.1	79.4	106	42.9	61.0	42	–
MOTSFusion [111]	Track R-CNN	78.2	90.0	36	50.1	68.0	34	1.2
CAMOT [130]	Track R-CNN	67.4	78.6	220	39.5	57.6	131	–
ASB [32]	Track R-CNN	77.4	89.6	41	48.9	66.7	28	–
PointTrack [198]	Track R-CNN	**85.5**	**94.9**	**22**	**62.4**	**77.3**	**19**	**22.2**

报告的单帧耗时换算而来（见其表 2，在 KITTI-MOTS 验证集上端到端地完成检测 + 分割 + 跟踪）；原论文未披露 GPU 型号，因此 FPS 的绝对值仅具参考意义。

5.4.1 **结果**。 表 12 显示，PointTrack [198] 最清晰地说明了在 VTS 中，身份表示可能比检测器的选择更为重要。由于所比较的方法基本共享同一个检测来源，性能差距更应归因于各方法如何表示和匹配实例。PointTrack 基于点的嵌入比 ROI 级外观特征更直接地保留了目标的几何与形变线索，因而在运动和遮挡下关联更加稳定。汽车与行人之间的剩余差距指向另一种限制：小型、带关节且频繁被遮挡的目标仍然对强大的关联设计构成挑战。因此，VTS 的进展由身份稳定的表示所驱动，而不仅仅取决于掩码质量。

趋势分析。 图 5b 表明，只有当关联表示具有足够的判别性时，分割质量与跟踪稳定性才会同步提升。早期的匹配启发式方法可以生成尚可接受的掩码，却仍会造成身份漂移，因为它们优化的更多是帧级重叠，而非长时程的身份连续性。PointTrack 通过在几何感知的空间中嵌入实例改变了这一趋势，降低了对脆弱的外观匹配的依赖。这表明未来 VTS 的提升应聚焦于在遮挡、形变和重新进入视野时依然保持稳定的表示，而不仅仅是改进逐帧分割器。

5.5 OVVS 基准

表 9 与 表 10 揭示了一个一致的语义瓶颈。有监督方法拥有强大的掩码解码器与时序关联模块，而开放词汇变体还需额外地将像素、目标身份与训练中稀有或未见过的语言类别对齐。这一额外的语义定位步骤解释了为何即使骨干网络相当，VSS 和 VIS 上的差距依然存在。因此，问题并不只是分割或跟踪能力较弱，而在于难以将图像—语言语义迁移为时间上连贯的视频掩码。缩小这一差距可能需要更好的视频—语言监督、稀有类别校准，以及能在帧间保持类别身份稳定的机制，这也印证了 §3.6 和 §6 中讨论的开放世界方向。

6 未来方向

开放世界视频场景解析。 *瓶颈。*封闭词汇 VSP 在分布内已趋于饱和，但在稀有或未见类别上性能崩溃：图 5a 显示 OV2Seg+/OVFormer/CLIP-VIS 落后于有监督最优方法 20–23 AP，表 9 显示 OV2VSS 在 VSPW 上落后于 TubeFormer 约 46 mIoU。*技术挑战。*必须将像素级的掩码质量（定位）与开放集识别（长尾、依赖先验）解耦。基于 CLIP 的分类头继承了图像级的偏差，在小目标或时序上不一致的目标上性能下降；而随着类别体系的扩展，持续学习又面临遗忘基类的风险。*候选方向。*(i) 将掩码提议路由至冻结的 VLM 进行分类的解耦架构；(ii) 借助掩码令牌回放实现内存高效的持续分割；(iii) 利用互联网规模的图像—文本对自举新类别的监督，并通过跨帧掩码一致性进行时序对齐。近期代表性工作：[55, 73, 162, 163, 175, 185, 199]。

面向分割的视觉推理。 *瓶颈。*正如我们的失效模式分析（§3.6.6）所示，最优的 VSP 方法在长时间遮挡、拥挤场景中的身份切换以及稀有细粒度类别上仍会失败，而这些失败的根源不在于像素层面的能力，而在于缺乏跨越时间与空间的显式关系推理。*技术挑战。*为分割赋予推理能力，需要一种同时支持快速的稠密逐像素预测（低于 100 ms）与慢速结构化关系推理（目标恒常性、因果交互）的表示，而这两种机制在历史上一直互不相容。*候选方向。*(i) 将快速掩码解码器与慢速场景图推理器配对的双流设计；(ii) 借助目标恒常性先验推测被

遮挡掩码的物理信息驱动动力学建模；(iii) 以语言为中介的推理，即由 VLM 生成文本形式的场景描述，再由下游掩码网络进行定位。近期代表性工作：[7, 50, 200, 226]。

高效视频理解。 *瓶颈。*如图 4 所示，精度与 FPS 的提升沿相对独立的方向演进：精度最高的方法（TubeFormer、MRCFA）只有个位数 FPS，而实时方法（DVSNet、MPVSS）则牺牲精度以换取更高的吞吐量，没有一种方法能在两个维度上同时占优。*技术挑战。*三个相互耦合的因素限定了这一前沿：(i) 逐帧注意力在空间令牌数上的复杂度为 $O(N^2)$；(ii) 片段级时序注意力会将其再乘以时间窗口长度；(iii) 在边缘设备上，决定实际运行时间的是内存带宽而非 FLOPs，因此单纯削减 FLOPs 未必能转化为实际加速。*候选方向。*(i) 利用跨帧冗余的稀疏/线性注意力；(ii) 以运动显著性为条件的动态令牌剪枝；(iii) 在 K 帧内复用重型特征的“关键帧 + 传播”混合方案；(iv) 联合优化目标加速器延迟的硬件感知神经架构搜索（NAS）。近期代表性工作：[47, 98, 117, 182, 194, 217]。

其他方向概述。 其余四个方向受前文证据的约束较弱。*统一的视频场景解析*有望将四个任务特定的方法族合而为一，前提是能够调和异构的监督与评价指标，而不让任何单一目标占据主导 [3, 21, 89, 155, 202, 215, 220, 222]。*面向分割的多模态融合*针对仅靠 RGB 会失效的情形，如伪装、低光照以及外观模糊的边界，其难点在于空间、时间与语义上的错位 [8, 95, 109, 156, 167, 174, 223, 237]。*生成式分割*有望在遮挡下表示多个合理的掩码而非唯一一个，前提是随机采样能够与实时确定性相调和 [23, 53, 85, 142, 170]。*基于大语言模型的分割*提供了指令级的控制，但仍以图像为中心且代价高昂，因此开放问题在于如何将缓慢变化的语言规划映射到快速变化的掩码上 [18, 25, 80, 92, 114, 152, 171, 188]。

7 伦理与社会影响

由于 §1 将智能监控列为一个应用领域，VSP 在伦理上的风险值得明确说明。该领域所优化的两种能力抬高了风险：一个能在长视频流中保持身份的模型（§2.1）在设计上就是一个重识别系统；而像素精确的掩码对体形、步态和助残器具的暴露远比边界框更为清晰。由此引出三个问题。*(i) 知情同意*：推动进展的开放场景语料（VSPW [122]、VIPSeg [121]、YouTube-VIS [205]）由网络视频构建，视频中的人物并未表示同意。*(ii) 公平性*：§3.6 中的长尾在语义上并非中性，被分割不足的稀有情形包括轮椅、婴儿车以及非典型体型，而这些情形在汇总分数中被平均掉了。*(iii) 双重用途*：封闭标签集原本是一种偶然的保护措施，而 OVVS 在设计上将其移除，因为自由形式的查询使基于属性条件的搜索成为一项默认能力。有用的缓解措施包括：限制所存储身份嵌入的生命周期、在设备端进行推理，以及为发布的数据与模型检查点提供数据说明书（datasheet）。我们之所以提出这些问题，是因为在本文所指出的瓶颈上取得进展，同样也会扩展监控能力。

8 结论

本综述回顾了 VSP 从手工设计特征到深度卷积、基于注意力以及 Transformer 架构的演进，这些架构在分割精度与时序连贯性上带来了持续的提升。若干问题仍然悬而未决（具有明确理论依据的时空融合、长时程一致性，以及超越封闭词汇的识别），而且尽管数据集与评价指标日趋完善，距离工业部署仍存在明显差距。未来的进展将取决于效率、对真实世界动态的鲁棒性以及对未见概念的适应能力。我们希望本文的梳理以及上文提出的研究方向，能为视频场景解析的后续研究提供有益的参考。

致谢

本工作部分得到国家自然科学基金（No. 62576176、62225604）、天津市科技重大专项（No. 25ZXRGGX00120）以及中央高校基本科研业务费专项资金（南开大学，No. 070-63253235、070-63261010）的资助。计算资源由南开大学超算中心（NKSC）提供。

参考文献

[1] Jiwoon Ahn and Suha Kwak. 2018. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 4981–4990.

[2] Nikita Araslanov and Stefan Roth. 2020. Single-stage semantic segmentation from image labels. In *IEEE Conf. Comput. Vis. Pattern Recog.* 4253–4262.

[3] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. 2023. BURST: A benchmark for unifying object recognition, segmentation and tracking in video. In *IEEE Winter Conf. App. Comput. Vis.* 1674–1683.

[4] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. 2020. STEm-Seg: Spatio-temporal embeddings for instance segmentation in videos. In *Eur. Conf. Comput. Vis.* 158–177.

[5] Vijay Badrinarayanan, Ignas Budvytis, and Roberto Cipolla. 2013. Semi-Supervised Video Segmentation Using Tree Structured Graphical Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 11 (2013), 2751–2764.

[6] Xue Bai, Jue Wang, and Guillermo Sapiro. 2010. Dynamic color flow: A motion-adaptive color model for object segmentation in video. In *Eur. Conf. Comput. Vis.* 617–630.

[7] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. 2024. One token to seg them all: Language instructed reasoning segmentation in videos. In *Annu. Conf. Neur. Inform. Process. Syst.* 6833–6859.

[8] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work?. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1539–1550.

[9] Gedas Bertasius and Lorenzo Torresani. 2020. Classifying, segmenting, and tracking object instances in video with mask propagation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 9739–9748.

[10] Guillem Brasó, Orcun Cetintas, and Laura Leal-Taixé. 2022. Multi-object tracking and segmentation via neural message passing. *Int. J. Comput. Vis.* 130, 12 (2022), 3035–3053.

[11] Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. 2008. Segmentation and Recognition Using Structure from Motion Point Clouds. In *Eur. Conf. Comput. Vis.* 44–57.

[12] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. 2017. One-shot video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 221–230.

[13] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. 2018. The 2018 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1803.00557* (2018).

[14] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. COCO-Stuff: Thing and stuff classes in context. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1209–1218.

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *Eur. Conf. Comput. Vis.* 213–229.

[16] Siddhartha Chandra, Camille Couprie, and Iasonas Kokkinos. 2018. Deep Spatio-Temporal Random Fields for Efficient Video Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 8915–8924.

[17] Jason Chang, Donglai Wei, and John W Fisher. 2013. A video representation using temporal superpixels. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2051–2058.

[18] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. 2024. ShareGPT4Video: Improving video understanding and generation with better captions. In *Annu. Conf. Neur. Inform. Process. Syst.* 19472–19495.

[19] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D. Collins, Ekin Dogus Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. 2020. Naive-Student: Leveraging Semi-Supervised Learning in Video Sequences for Urban Scene Segmentation. In *Eur. Conf. Comput. Vis.* 695–714.

[20] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2018), 834–848.

[21] Mu Chen, Liulei Li, Wenguan Wang, Ruijie Quan, and Yi Yang. 2024. General and Task-Oriented Video Segmentation. In *Eur. Conf. Comput. Vis.* 72–92.

[22] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. 2022. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *IEEE Conf. Comput. Vis. Pattern Recog.* 16144–16155.

[23] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. 2023. A generalist framework for panoptic segmentation of images and videos. In *Int. Conf. Comput. Vis.* 909–919.

[24] Xi Chen, Shuang Li, Ser-Nam Lim, Antonio Torralba, and Hengshuang Zhao. 2023. Open-vocabulary panoptic segmentation with embedding modulation. In *Int. Conf. Comput. Vis.* 1141–1150.

[25] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE Conf. Comput. Vis. Pattern Recog.* 24185–24198.

[26] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1290–1299.

[27] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. 2024. Putting the object back into video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3151–3161.
[28] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. 2023. Tracking anything with decoupled video segmentation. In *Int. Conf. Comput. Vis.* 1316–1326.
[29] Ho Kei Cheng and Alexander G Schwing. 2022. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Eur. Conf. Comput. Vis.* 640–658.
[30] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. 2021. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Annu. Conf. Neur. Inform. Process. Syst.* 34 (2021), 11781–11794.
[31] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558* (2023).
[32] Anwesa Choudhuri, Girish Chowdhary, and Alexander G Schwing. 2021. Assignment-space-based multi-object tracking and segmentation. In *Int. Conf. Comput. Vis.* 13598–13607.
[33] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3213–3223.
[34] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. 2020. MOT20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003* (2020).
[35] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. 2018. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2393–2402.
[36] Mingyu Ding, Zhe Wang, Bolei Zhou, Jianping Shi, Zhiwu Lu, and Ping Luo. 2020. Every Frame Counts: Joint Learning of Video Segmentation and Optical Flow. In *AAAI Conf. Artif. Intell.* 10713–10720.
[37] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. 2024. UniRepLKNet: A universal perception large-kernel ConvNet for audio video point cloud time-series and image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 5513–5524.
[38] PHILOSO EPHY DO CT OR OF. 1961. *Machine perception of three-dimensional, so lids*. Ph. D. Dissertation. PhD thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY.
[39] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Int. Conf. Learn. Represent.* 1–12.
[40] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* 111, 1 (2015), 98–136.
[41] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. 2020. CIAN: Cross-image affinity net for weakly supervised semantic segmentation. In *AAAI Conf. Artif. Intell.* 10762–10769.
[42] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. 2024. Unified embedding alignment for open-vocabulary video instance segmentation. In *Eur. Conf. Comput. Vis.* 225–241.
[43] Sándor Fazekas, Tomer Amiaz, Dmitry Chetverikov, and Nahum Kiryati. 2009. Dynamic texture detection based on motion analysis. *Int. J. Comput. Vis.* 82 (2009), 48–63.
[44] Yang Fu, Sifei Liu, Umar Iqbal, Shalini De Mello, Humphrey Shi, and Jan Kautz. 2021. Learning to Track Instances without Video Annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.* 8676–8685.
[45] Yang Fu, Linjie Yang, Ding Liu, Thomas S Huang, and Humphrey Shi. 2021. CompFeat: Comprehensive feature aggregation for video instance segmentation. In *AAAI Conf. Artif. Intell.* 1361–1369.
[46] Raghudeep Gadde, V. Jampani, and Peter Gehler. 2017. Semantic Video CNNs Through Representation Warping. In *Int. Conf. Comput. Vis.* 4463–4472.
[47] Roland Gao. 2023. Rethinking dilated convolution for real-time semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 4675–4684.
[48] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Pablo Martinez-Gonzalez, and Jose Garcia-Rodriguez. 2018. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing* 70 (2018), 41–65.
[49] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3354–3361.
[50] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. 2025. The Devil is in Temporal Token: High Quality Video Reasoning Segmentation. *arXiv preprint arXiv:2501.08549* (2025).
[51] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. 2010. Efficient hierarchical graph-based video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2141–2148.
[52] Jiaqi Gu, Hyoukjun Kwon, Dilin Wang, Wei Ye, Meng Li, Yu-Hsin Chen, Liangzhen Lai, Vikas Chandra, and David Z Pan. 2022. Multi-scale high-resolution vision transformer for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 12094–12103.
[53] Zhangxuan Gu, Haoxing Chen, and Zhuoer Xu. 2024. DiffusionInst: Diffusion model for instance segmentation. In *IEEE Int. Conf. Acoust. Speech SP*. 2730–2734.

[54] Diandian Guo, Deng-Ping Fan, Tongyu Lu, Christos Sakaridis, and Luc Van Gool. 2024. Vanishing-Point-Guided Video Semantic Segmentation of Driving Scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3544–3553.
[55] Pinxue Guo, Zixu Zhao, Jianxiong Gao, Chongruo Wu, Tong He, Zheng Zhang, Tianjun Xiao, and Wenqiang Zhang. 2024. VideoSAM: Open-World Video Segmentation. *arXiv preprint arXiv:2410.08781* (2024).
[56] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 5356–5364.
[57] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. 2023. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Int. Conf. Comput. Vis.* 1086–1096.
[58] Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, et al. 2023. Global knowledge calibration for fast open-vocabulary segmentation. In *Int. Conf. Comput. Vis.* 797–807.
[59] Su Ho Han, Sukjun Hwang, Seoung Wug Oh, Yeonchool Park, Hyunwoo Kim, Min-Jung Kim, and Seon Joo Kim. 2022. VISOLO: Grid-based space-time aggregation for efficient online video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2896–2905.
[60] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. 2023. A generalized framework for video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 14623–14632.
[61] Syed Ariff Syed Hesham, Yun Liu, Guolei Sun, Henghui Ding, Jing Yang, Ender Konukoglu, Xue Geng, and Xudong Jiang. 2025. Exploiting Temporal State Space Sharing for Video Semantic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 24211–24221.
[62] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Köring, Suman Saha, and Luc Van Gool. 2020. Three Ways to Improve Semantic Segmentation with Self-Supervised Depth Estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 11125–11135.
[63] Ping Hu, Fabian Caba Heilbron, Oliver Wang, Zhe L. Lin, Stan Sclaroff, and Federico Perazzi. 2020. Temporally Distributed Networks for Fast Video Semantic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 8815–8824.
[64] De-An Huang, Zhiding Yu, and Anima Anandkumar. 2022. MinVIS: A minimal video instance segmentation framework without video-based training. In *Annu. Conf. Neur. Inform. Process. Syst.* 31265–31277.
[65] Po-Yu Huang, Wan-Ting Hsu, Chun-Yueh Chiu, Ting-Fan Wu, and Min Sun. 2018. Efficient uncertainty estimation for semantic segmentation in videos. In *Eur. Conf. Comput. Vis.* 520–535.
[66] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. 2021. Video instance segmentation using inter-frame communication transformers. In *Annu. Conf. Neur. Inform. Process. Syst.* 13352–13363.
[67] Samvit Jain, Xin Wang, and Joseph E. Gonzalez. 2018. Accel: A Corrective Fusion Network for Efficient Semantic Segmentation on Video. In *IEEE Conf. Comput. Vis. Pattern Recog.* 8858–8867.
[68] Won-Dong Jang and Chang-Su Kim. 2016. Streaming Video Segmentation via Short-Term Hierarchical Segmentation and Frame-by-Frame Markov Random Field Optimization. In *Eur. Conf. Comput. Vis.* 599–615.
[69] Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan Loddon Yuille, and Li Cheng. 2023. Multispectral Video Semantic Segmentation: A Benchmark Dataset and Baseline. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1094–1104.
[70] Zhengkai Jiang, Zhangxuan Gu, Jinlong Peng, Hang Zhou, Liang Liu, Yabiao Wang, Ying Tai, Chengjie Wang, and Liqing Zhang. 2022. STC: Spatio-temporal contrastive learning for video instance segmentation. In *Eur. Conf. Comput. Vis.* 539–556.
[71] Licheng Jiao, Ruohan Zhang, Fang Liu, Shuyuan Yang, Biao Hou, Lingling Li, and Xu Tang. 2021. New generation deep learning for video object detection: A survey. *IEEE Trans. Neur. Net. Learn. Syst.* 33, 8 (2021), 3195–3215.
[72] Xiaojie Jin, Xin Li, Huaxin Xiao, Xiaohui Shen, Zhe Lin, Jimei Yang, Yunpeng Chen, Jian Dong, Luoqi Liu, Zequn Jie, et al. 2017. Video scene parsing with predictive feature learning. In *Int. Conf. Comput. Vis.* 5580–5588.
[73] Tarun Kalluri, Weiyao Wang, Heng Wang, Manmohan Chandraker, Lorenzo Torresani, and Du Tran. 2024. Open-world instance segmentation: Top-down learning with bottom-up supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2693–2703.
[74] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan De Geus. 2025. Your vit is secretly an image segmentation model. In *IEEE Conf. Comput. Vis. Pattern Recog.* 25303–25313.
[75] Anna Khoreva, Fabio Galasso, Matthias Hein, and Bernt Schiele. 2015. Classifier based graph construction for video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 951–960.
[76] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In-So Kweon. 2020. Video Panoptic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 9856–9865.
[77] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. 2022. Tubeformer-DeepLab: Video mask transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.* 13914–13924.
[78] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Int. Conf. Comput. Vis.* 4015–4026.

[79] Abhijit Kundu, Vibhav Vineet, and Vladlen Koltun. 2016. Feature space optimization for semantic video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3168–3175.
[80] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. LISA: Reasoning segmentation via large language model. In *IEEE Conf. Comput. Vis. Pattern Recog.* 9579–9589.
[81] CF Lam and Moon-Chuen Lee. 1998. Video segmentation using color difference histogram. In *IAPR International Workshop on Multimedia Information Analysis and Retrieval*. 159–174.
[82] Jiangwei Lao, Weixiang Hong, Xin Guo, Yingying Zhang, Jian Wang, Jingdong Chen, and Wei Chu. 2023. Simultaneously Short- and Long-Term Temporal Modeling for Semi-Supervised Video Semantic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 14763–14772.
[83] Seunghun Lee, Jiwan Seo, Kiljoon Han, Minwoo Choi, and Sunghoon Im. 2025. CAVIS: Context-Aware Video Instance Segmentation. In *Int. Conf. Comput. Vis.* 4507–4517.
[84] Shih-Po Lee, Si-Cun Chen, and Wen-Hsiao Peng. 2021. GSVNet: Guided spatially-varying convolution for fast semantic segmentation on video. In *IEEE Int. Conf. Multimedia Expo*. 1–6.
[85] Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. 2023. MomentDiff: Generative video moment retrieval from random to real. In *Annu. Conf. Neur. Inform. Process. Syst.* 65948–65966.
[86] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. 2024. Transformer-Based Visual Segmentation: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 12 (2024), 10138–10163.
[87] Xiangtai Li, Hao He, Yibo Yang, Henghui Ding, Kuiyuan Yang, Guangliang Cheng, Yunhai Tong, and Dacheng Tao. 2022. Improving video instance segmentation via temporal pyramid routing. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 5 (2022), 6594–6601.
[88] Xinhao Li, Yun Liu, Guolei Sun, Min Wu, Le Zhang, and Ce Zhu. 2025. Towards Open-Vocabulary Video Semantic Segmentation. *IEEE Trans. Multimedia* (2025).
[89] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. 2024. OMG-Seg: Is one model good enough for all segmentation?. In *IEEE Conf. Comput. Vis. Pattern Recog.* 27948–27959.
[90] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. 2023. TubeLink: A flexible cross tube framework for universal video segmentation. In *Int. Conf. Comput. Vis.* 13923–13933.
[91] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. 2022. Video K-Net: A simple, strong, and unified baseline for video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 18847–18857.
[92] Yifan Li, Zhixin Lai, Wentao Bao, Zhen Tan, Anh Dao, Kewei Sui, Jiayi Shen, Dong Liu, Huan Liu, and Yu Kong. 2025. Visual Large Language Models for Generalized and Specialized Applications. *arXiv preprint arXiv:2501.02765* (2025).
[93] Yu Li, Zhuoran Shen, and Ying Shan. 2020. Fast video object segmentation using the global context module. In *Eur. Conf. Comput. Vis.* 735–750.
[94] Yule Li, Jianping Shi, and Dahua Lin. 2018. Low-latency video semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 5997–6005.
[95] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *IEEE Conf. Comput. Vis. Pattern Recog.* 26763–26773.
[96] Chung-Ching Lin, Ying Hung, Rogerio Feris, and Linglin He. 2020. Video instance segmentation tracking with a modified VAE architecture. In *IEEE Conf. Comput. Vis. Pattern Recog.* 13147–13157.
[97] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. 2021. Video instance segmentation with a propose-reduce paradigm. In *Int. Conf. Comput. Vis.* 1739–1748.
[98] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal shift module for efficient video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.* 7083–7093.
[99] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.* 740–755.
[100] Weiyao Lin, Huabin Liu, Shizhan Liu, Yuxi Li, Rui Qian, Tao Wang, Ning Xu, Hongkai Xiong, Guo-Jun Qi, and Nicu Sebe. 2020. Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint arXiv:2005.04490* (2020).
[101] Buyu Liu and Xuming He. 2015. Multiclass semantic video segmentation with object-level active inference. In *IEEE Conf. Comput. Vis. Pattern Recog.* 4286–4294.
[102] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. 2021. SG-Net: Spatial granularity network for one-stage video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 9816–9825.
[103] Qing Liu, Vignesh Ramanathan, Dhruv Kumar Mahajan, Alan Loddon Yuille, and Zhenheng Yang. 2021. Weakly Supervised Instance Segmentation for Videos with Temporal Mask Consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.* 13963–13973.

[104] Shuaicheng Liu, Lu Yuan, Ping Tan, and Jian Sun. 2014. Steadyflow: Spatially smooth optical flow for video stabilization. In *IEEE Conf. Comput. Vis. Pattern Recog.* 4209–4216.
[105] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *Eur. Conf. Comput. Vis.* 38–55.
[106] Yifan Liu, Chunhua Shen, Changqian Yu, and Jingdong Wang. 2020. Efficient semantic video segmentation with per-frame inference. In *Eur. Conf. Comput. Vis.* 352–368.
[107] Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, and Luc Van Gool. 2024. Vision transformers with hierarchical attention. *Machine Intelligence Research* 21, 4 (2024), 670–683.
[108] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.* 12009–12019.
[109] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2024. Unified-IO 2: Scaling autoregressive multimodal models with vision language audio and action. In *IEEE Conf. Comput. Vis. Pattern Recog.* 26439–26455.
[110] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. 2020. Video object segmentation with episodic graph memory networks. In *Eur. Conf. Comput. Vis.* 661–679.
[111] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. 2020. Track to reconstruct and reconstruct to track. *IEEE Trans. Robot. Autom. Let.* 5, 2 (2020), 1803–1810.
[112] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. 2018. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conf. Comput. Vis.* 565–580.
[113] Ping Luo, Guangrun Wang, Liang Lin, and Xiaogang Wang. 2017. Deep dual learning for semantic image segmentation. In *Int. Conf. Comput. Vis.* 2718–2726.
[114] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *Eur. Conf. Comput. Vis.* 417–435.
[115] Behrooz Mahasseni, Sinisa Todorovic, and Alan Fern. 2017. Budget-aware deep semantic video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1029–1038.
[116] Muhammad Nawfal Meeran, Bhanu Pratyush Mantha, et al. 2024. SAM-PM: Enhancing video camouflaged object detection using spatio-temporal attention. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1857–1866.
[117] Sachin Mehta and Mohammad Rastegari. 2021. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178* (2021).
[118] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Han Yan, Siyuan Qiao, Yukun Zhu, Liang-Chieh Chen, Henrik Kretzschmar, and Dragomir Anguelov. 2022. Waymo Open Dataset: Panoramic Video Panoptic Segmentation. In *Eur. Conf. Comput. Vis.* 53–72.
[119] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. 2022. TrackFormer: Multi-object tracking with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.* 8844–8854.
[120] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. 2022. AdaViT: Adaptive vision transformers for efficient image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 12309–12318.
[121] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. 2022. Large-scale Video Panoptic Segmentation in the Wild: A Benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.* 21001–21011.
[122] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. 2021. VSPW: A Large-scale Dataset for Video Scene Parsing in the Wild. In *IEEE Conf. Comput. Vis. Pattern Recog.* 4133–4143.
[123] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. 2016. MOT16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831* (2016).
[124] Manjunath Narayana, Allen Hanson, and Erik Learned-Miller. 2013. Coherent motion segmentation in moving camera videos using optical flow orientations. In *Int. Conf. Comput. Vis.* 1577–1584.
[125] David Nilsson and Cristian Sminchisescu. 2018. Semantic video segmentation by gated recurrent flow propagation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 6819–6828.
[126] Narges Norouzi, Idil Esen Zulfikar, Niccolo Cavagnero, Tommie Kerssies, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. 2026. VidEoMT: Your ViT is secretly also a video segmentation model. *arXiv preprint arXiv:2602.17807* (2026).
[127] Peter Ochs, Jitendra Malik, and Thomas Brox. 2013. Segmentation of moving objects by long term video analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 6 (2013), 1187–1200.
[128] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. 2019. Video object segmentation using space-time memory networks. In *Int. Conf. Comput. Vis.* 9226–9235.
[129] Aljoša Osep, Wolfgang Mehner, Markus Mathias, and Bastian Leibe. 2017. Combined image-and world-space tracking in traffic scenes. In *IEEE Int. Conf. Robot. Autom.* 1988–1995.
[130] Aljoša Ošep, Wolfgang Mehner, Paul Voigtlaender, and Bastian Leibe. 2018. Track, then decide: Category-agnostic vision-based multi-object tracking. In *IEEE Int. Conf. Robot. Autom.* 3494–3501.

[131] Anestis Papazoglou and Vittorio Ferrari. 2013. Fast object segmentation in unconstrained video. In *Int. Conf. Comput. Vis.* 1777–1784.
[132] Hyojin Park, Alan Yessenbayev, Tushar Singhal, Navin Kumar Adhikari, Yizhe Zhang, Shubhankar Mangesh Borse, Hong Cai, Nilesh Prasad Pandey, Fei Yin, Frank Mayer, et al. 2022. Real-time, accurate, and consistent video semantic segmentation via unsupervised adaptation and cross-unit deployment on mobile device. In *IEEE Conf. Comput. Vis. Pattern Recog.* 21431–21438.
[133] Matthieu Paul, Christoph Mayer, Luc Van Gool, and Radu Timofte. 2020. Efficient video semantic segmentation with labels propagation and refinement. In *IEEE Winter Conf. App. Comput. Vis.* 2873–2882.
[134] Yi Peng, Li Chen, Fang-Xin Ou-Yang, Wei Chen, and Jun-Hai Yong. 2014. JF-Cut: A parallel graph cut approach for large-scale image and video. *IEEE Trans. Image Process.* 24, 2 (2014), 655–666.
[135] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. 2017. Learning video object segmentation from static images. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2663–2672.
[136] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 724–732.
[137] Federico Perazzi, Oliver Wang, Markus H. Gross, and Alexander Sorkine-Hornung. 2015. Fully Connected Object Proposals for Video Segmentation. In *Int. Conf. Comput. Vis.* 3227–3234.
[138] Andra Petrovai and Sergiu Nedevschi. 2022. Time-Space Transformers for Video Panoptic Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 925–934.
[139] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 DAVIS challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).
[140] Lorenzo Porzi, Markus Hofinger, Idoia Ruiz, Joan Serrat, Samuel Rota Bulò, and Peter Kontschieder. 2019. Learning Multi-Object Tracking and Segmentation From Automatic Annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.* 6845–6854.
[141] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge J. Belongie, Alan Loddon Yuille, Philip H. S. Torr, and Song Bai. 2022. Occluded Video Instance Segmentation: A Benchmark. *Int. J. Comput. Vis.* 130, 8 (2022), 2022–2039.
[142] Lu Qi, Lehan Yang, Weidong Guo, Yu Xu, Bo Du, Varun Jampani, and Ming-Hsuan Yang. 2024. UniGS: Unified representation for image generation and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 6305–6315.
[143] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Loddon Yuille, and Liang-Chieh Chen. 2020. ViP-DeepLab: Learning Visual Perception with Depth-aware Video Panoptic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3996–4007.
[144] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.* 8748–8763.
[145] Frano Rajič, Lei Ke, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu. 2023. Segment anything meets point tracking. *arXiv preprint arXiv:2307.01197* (2023).
[146] Hatem A Rashwan, Mahmoud A Mohamed, Miguel Angel García, Bärbel Mertsching, and Domenec Puig. 2013. Illumination robust optical flow model based on histogram of oriented gradients. In *German Conf. Pattern Recog.* 354–363.
[147] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).
[148] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2015), 1137–1149.
[149] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. 2021. ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding. In *Int. Conf. Comput. Vis.* 10745–10755.
[150] Evan Shelhamer, Jonathan Long, and Trevor Darrell. 2017. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 640–651.
[151] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell. 2016. Clockwork convnets for video semantic segmentation. In *Eur. Conf. Comput. Vis.* 852–868.
[152] Hengcan Shi, Son Duy Dao, and Jianfei Cai. 2025. LLMFormer: Large language model for open-vocabulary semantic segmentation. *Int. J. Comput. Vis.* 133, 2 (2025), 742–759.
[153] Mennatullah Siam, Sepehr Valipour, Martin Jagersand, and Nilanjan Ray. 2017. Convolutional gated recurrent networks for video segmentation. In *Int. Conf. Image Process.* 3090–3094.
[154] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *Eur. Conf. Comput. Vis.* 746–760.
[155] Yiran Song, Qianyu Zhou, Xiangtai Li, Deng-Ping Fan, Xuequan Lu, and Lizhuang Ma. 2024. BA-SAM: Scalable bias-mode attention mask for segment anything model. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3162–3173.

[156] Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. 2022. CLiMB: A continual learning benchmark for vision-and-language tasks. In *Annu. Conf. Neur. Inform. Process. Syst.* 29440–29453.
[157] Deqing Sun, Erik B Sudderth, and Michael J Black. 2012. Layered segmentation and optical flow estimation over time. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1768–1775.
[158] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. 2021. Mamba: Multi-level aggregation via memory bank for video object detection. In *AAAI Conf. Artif. Intell.* 2620–2627.
[159] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc van Gool. 2022. Coarse-to-Fine Feature Mining for Video Semantic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3116–3127.
[160] Guolei Sun, Yun Liu, Henghui Ding, Min Wu, and Luc Van Gool. 2024. Learning Local and Global Temporal Contexts for Video Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 10 (2024), 6919–6934.
[161] Guolei Sun, Yun Liu, Hao Tang, Ajad Chhatkuli, Le Zhang, and Luc Van Gool. 2022. Mining relations among cross-frame affinities for video semantic segmentation. In *Eur. Conf. Comput. Vis.* 522–539.
[162] Lv Tang, Peng-Tao Jiang, Haoke Xiao, and Bo Li. 2025. Towards Training-Free Open-World Segmentation via Image Prompt Foundation Models. *Int. J. Comput. Vis.* 133, 1 (2025), 1–15.
[163] Omkar Thawakar, Sanath Narayan, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. 2025. Video instance segmentation in an open-world. *Int. J. Comput. Vis.* 133, 1 (2025), 398–409.
[164] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. 2016. Video segmentation via object flow. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3899–3908.
[165] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Annu. Conf. Neur. Inform. Process. Syst.* 6000–6010.
[166] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. 2019. RVOS: End-to-end recurrent network for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 5277–5286.
[167] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. 2023. Connecting vision and language with video localized narratives. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2461–2471.
[168] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. 2019. MOTS: Multi-object tracking and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 7942–7951.
[169] Haochen Wang, Cilin Yan, Keyan Chen, Xiaolong Jiang, Xu Tang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. 2024. OV-VIS: Open-vocabulary video instance segmentation. *Int. J. Comput. Vis.* 132, 11 (2024), 5048–5065.
[170] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. 2023. Towards open-vocabulary video instance segmentation. In *Int. Conf. Comput. Vis.* 4057–4066.
[171] Junchi Wang and Lei Ke. 2024. LLM-Seg: Bridging image segmentation and large language model reasoning. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1765–1774.
[172] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. 2019. PANet: Few-shot image semantic segmentation with prototype alignment. In *Int. Conf. Comput. Vis.* 9197–9206.
[173] Shiyao Wang, Yucong Zhou, Junjie Yan, and Zhidong Deng. 2018. Fully motion-aware network for video object detection. In *Eur. Conf. Comput. Vis.* 542–557.
[174] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. In *Annu. Conf. Neur. Inform. Process. Syst.* 61501–61513.
[175] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. 2021. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Int. Conf. Comput. Vis.* 10776–10785.
[176] Wenguan Wang, Tianfei Zhou, Fatih Murat Porikli, David J. Crandall, and Luc Van Gool. 2022. A Survey on Deep Learning Technique for Video Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 6 (2022), 7099–7122.
[177] Xudong Wang, Ishan Misra, Ziyun Zeng, Rohit Girdhar, and Trevor Darrell. 2023. VideoCutLER: Surprisingly Simple Unsupervised Video Instance Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 22755–22764.
[178] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. 2021. End-to-end video instance segmentation with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.* 8741–8750.
[179] Mark Weber, Jun Xie, Maxwell D. Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, B. Leibe, Daniel Cremers, Aljosa Osep, Laura Leal-Taixé, and Liang-Chieh Chen. 2021. STEP: Segmenting and Tracking Every Pixel. In *NeurIPS Datasets and Benchmarks*. 1–13.
[180] Yuetian Weng, Mingfei Han, Haoyu He, Mingjie Li, Lina Yao, Xiaojun Chang, and Bohan Zhuang. 2023. Mask propagation for efficient video semantic segmentation. In *Annu. Conf. Neur. Inform. Process. Syst.* 7170–7183.
[181] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. 2021. Learning to Associate Every Segment for Video Panoptic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2705–2714.

[182] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. MeMViT: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 13587–13597.
[183] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. 2021. Track to Detect and Segment: An Online Multi-Object Tracker. In *IEEE Conf. Comput. Vis. Pattern Recog.* 12347–12356.
[184] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. 2022. SeqFormer: Sequential transformer for video instance segmentation. In *Eur. Conf. Comput. Vis.* 553–569.
[185] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. 2024. Towards open vocabulary learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 7 (2024), 5092–5113.
[186] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. 2022. In defense of online models for video instance segmentation. In *Eur. Conf. Comput. Vis.* 588–605.
[187] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. 2022. P2T: Pyramid pooling transformer for scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 11 (2022), 12760–12771.
[188] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. 2024. GSVA: Generalized segmentation via multimodal large language models. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3858–3869.
[189] Guohuan Xie, Xin He, Dingying Fan, Le Zhang, Ming-Ming Cheng, and Yun Liu. 2026. Make It Up: Fake Images, Real Gains in Generalized Few-shot Semantic Segmentation. *arXiv preprint arXiv:2603.27206*.
[190] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. 2021. Efficient regional memory network for video object segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1286–1295.
[191] Chenliang Xu and Jason J Corso. 2012. Evaluation of super-voxel methods for early video processing. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1202–1209.
[192] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. 2023. Learning open-vocabulary semantic segmentation models from natural language supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2935–2944.
[193] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2955–2966.
[194] Jiacong Xu, Zixiang Xiong, and Shankar P Bhattacharyya. 2023. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In *IEEE Conf. Comput. Vis. Pattern Recog.* 19529–19539.
[195] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. 2018. YouTube-VOS: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327* (2018).
[196] Yuanyou Xu, Zongxin Yang, and Yi Yang. 2023. Integrating boxes and masks: A multi-object framework for unified visual tracking and segmentation. In *Int. Conf. Comput. Vis.* 9738–9751.
[197] Yu-Syuan Xu, Tsu-Jui Fu, Hsuan-Kung Yang, and Chun-Yi Lee. 2018. Dynamic video segmentation network. In *IEEE Conf. Comput. Vis. Pattern Recog.* 6556–6565.
[198] Zhenbo Xu, Wei Zhang, Xiao Tan, Wei Yang, Huan Huang, Shilei Wen, Errui Ding, and Liusheng Huang. 2020. Segment as points for efficient online multi-object tracking and segmentation. In *Eur. Conf. Comput. Vis.* 264–281.
[199] Zihui Xue, Kumar Ashutosh, and Kristen Grauman. 2024. Learning object state changes in videos: An open-world perspective. In *IEEE Conf. Comput. Vis. Pattern Recog.* 18493–18503.
[200] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. 2024. VISA: Reasoning video object segmentation via large language models. In *Eur. Conf. Comput. Vis.* 98–115.
[201] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. 2022. Multiview transformers for video recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3333–3343.
[202] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. 2024. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *AAAI Conf. Artif. Intell.* 6449–6457.
[203] Changdi Yang, Pu Zhao, Yanyu Li, Wei Niu, Jiexiong Guan, Hao Tang, Minghai Qin, Bin Ren, Xue Lin, and Yanzhi Wang. 2023. Pruning parameterization with bi-level optimization for efficient semantic segmentation on the edge. In *IEEE Conf. Comput. Vis. Pattern Recog.* 15402–15412.
[204] Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yong Xu, and Qian Zhang. 2022. Cross-Image Relational Knowledge Distillation for Semantic Segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 12309–12318.
[205] Linjie Yang, Yuchen Fan, and Ning Xu. 2019. Video Instance Segmentation. In *Int. Conf. Comput. Vis.* 5187–5196.
[206] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. 2021. Crossover learning for fast online video instance segmentation. In *Int. Conf. Comput. Vis.* 8043–8052.
[207] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. 2022. Temporally efficient vision transformer for video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2885–2895.
[208] Zongxin Yang, Yunchao Wei, and Yi Yang. 2020. Collaborative video object segmentation by foreground-background integration. In *European Conference on Computer Vision*. 332–348.

[209] Weicai Ye, Xinyue Lan, Ge Su, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. 2022. HybridTracker: A Lightweight and Joint Tracking Model with Pixel and Instance for Video Panoptic Segmentation. *arXiv preprint arXiv:2203.01217* (2022).
[210] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. 2023. CTVIS: Consistent training for online video instance segmentation. In *Int. Conf. Comput. Vis.* 899–908.
[211] Chen-Ping Yu, Hieu M. Le, Gregory J. Zelinsky, and Dimitris Samaras. 2015. Efficient Video Segmentation Using Parametric Graph Partitioning. In *Int. Conf. Comput. Vis.* 3155–3163.
[212] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2636–2645.
[213] Haobo Yuan, Xiangtai Li, Yibo Yang, Guangliang Cheng, Jing Zhang, Yunhai Tong, Lefei Zhang, and Dacheng Tao. 2022. PolyphonicFormer: Unified query learning for depth-aware video panoptic segmentation. In *Eur. Conf. Comput. Vis.* 582–599.
[214] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. 2022. Neural window fully-connected CRFs for monocular depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3916–3925.
[215] Xu Yuan, Li Zhou, Zenghui Sun, Zikun Zhou, and Jinsong Lan. 2025. Instruction-guided Multi-Granularity Segmentation and Captioning with Large Multimodal Model. In *AAAI Conf. Artif. Intell.* 9725–9733.
[216] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.* 12104–12113.
[217] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. 2023. Faster segment anything: Towards lightweight SAM for mobile applications. *arXiv preprint arXiv:2306.14289* (2023).
[218] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. 2023. A simple framework for open-vocabulary segmentation and detection. In *Int. Conf. Comput. Vis.* 1020–1031.
[219] Miao Zhang, Jie Liu, Yifei Wang, Yongri Piao, Shunyu Yao, Wei Ji, Jingjing Li, Huchuan Lu, and Zhongxuan Luo. 2021. Dynamic context-sensitive filtering network for video salient object detection. In *Int. Conf. Comput. Vis.* 1553–1563.
[220] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. 2024. OMG-LLaVA: Bridging image-level, object-level, pixel-level reasoning and understanding. In *Annu. Conf. Neur. Inform. Process. Syst.* 71737–71767.
[221] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. 2023. Dvis: Decoupled video instance segmentation framework. In *Int. Conf. Comput. Vis.* 1282–1291.
[222] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. 2025. DVIS++: Improved decoupled framework for universal video segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* (2025).
[223] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. 2023. Meta-Transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802* (2023).
[224] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. 2018. ICNet for real-time semantic segmentation on high-resolution images. In *Eur. Conf. Comput. Vis.* 405–420.
[225] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2881–2890.
[226] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. 2024. ViLLa: Video Reasoning Segmentation with Large Language Model. *arXiv preprint arXiv:2407.14500* (2024).
[227] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.* 127, 3 (2019), 302–321.
[228] Yi Zhou, Hui Zhang, Hana Lee, Shuyang Sun, Pingjun Li, Yangguang Zhu, ByungIn Yoo, Xiaojuan Qi, and Jae-Joon Han. 2022. Slot-VPS: Object-centric representation learning for video panoptic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3093–3103.
[229] Chaoyang Zhu and Long Chen. 2024. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Trans. Pattern Anal. Mach. Intell.* (2024).
[230] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. 2025. CLIP-VIS: Adapting CLIP for open-vocabulary video instance segmentation. *IEEE Trans. Circ. Syst. Video Technol.* 35, 2 (2025), 1098–1110.
[231] Xizhou Zhu, Jifeng Dai, Lu Yuan, and Yichen Wei. 2018. Towards high performance video object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.* 7210–7218.
[232] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Int. Conf. Comput. Vis.* 408–417.
[233] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Deep feature flow for video recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 2349–2358.
[234] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. 2019. Improving semantic segmentation via video propagation and label relaxation. In *IEEE Conf. Comput. Vis. Pattern Recog.*

8856–8865.
[235] Jiafan Zhuang, Zilei Wang, and Yuan Gao. 2022. Semi-supervised video semantic segmentation with inter-frame feature reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3263–3271.
[236] Jiafan Zhuang, Zilei Wang, Yixin Zhang, and Zhun Fan. 2024. Infer from what you have seen before: Temporally-dependent classifier for semi-supervised video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* 3575–3584.
[237] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. 2023. Generalized decoding for pixel, image, and language. In *IEEE Conf. Comput. Vis. Pattern Recog.* 15116–15127.