CRISP：连续视频目标实例分割的对比残差注入和语义提示

Baichen Liu , Qi Lyu , Xudong Wang , Jiahua Dong ,
Lianqing Liu *Senior Member, IEEE*, Zhi Han *Member, IEEE*

***Abstract*—视频持续实例分割要求既有吸收新对象类别的灵活性，又有保留先前学习到的类别的稳定性，同时在帧间保持时间一致性。在这项工作中，我们引入了对比残差注入和语义提示（CRISP），这是一个针对在持续视频实例分割中解决实例层面、类别层面和任务层面混淆的早期尝试。对于实例层面的学习，我们对实例追踪进行建模并构建实例相关性损失，该损失强调与先前查询空间的相关性，同时加强当前任务查询的特异性。对于类别层面的学习，我们构建了一个自适应残差语义提示（ARSP）学习框架，该框架通过类别文本生成一个可学习的语义残差提示池，并使用一个调节性的查询-提示匹配机制来建立当前任务查询与语义残差提示之间的映射关系。同时，引入了一种基于对比学习的语义一致性损失，以在增量训练期间保持对象查询与残差提示之间的语义一致性。对于任务层面的学习，为了确保查询空间内跨任务级别的相关性，我们引入了一种简洁而强大的增量提示初始化策略。在 YouTube-VIS-2019 和 YouTube-VIS-2021 数据集上的大量实验表明，CRISP 在长期持续视频实例分割任务中明显优于现有的持续分割方法，避免了灾难性遗忘，并有效改善了分割和分类性能。代码可在 https://github.com/01upup10/CRISP 获得。**

***Index Terms*—Continual video instance segmentation, continual learning, prompt tuning, residual semantic prompt, contrastive learning.**

I. 介绍

深度学习 学习方法在各种任务中表现优异，例如图像分类 [1]–[7] 、物体检测 [8]–[10] 和图像分割 [11]–[15] 。图像实例分割 [16]–[18] 在分割任务中起着关键作用，因为它统一了物体检测和像素级掩码预测，从而实现了对图像中每个单独对象的精确划分。近年来，图像实例分割的方法已经从多阶段、基于级联的流水线发展到端到端的实时框架，并最终发展为统一的、注意力导向的模型，提供了更高的准确性和更快的推理速度。像 Mask2Former [19] 这样的通用架构已将注意力导向的框架扩展到在一个统一的模型中支持全景、语义和实例图像分割。此外，适用于视频实例分割（VIS）的 Mask2Former 也在不修改架构的情况下实现了最先进的性能 [20] 。尽管取得了这些进展，但在学习新知识的过程中，与人类等智能系统能够将先前学到的知识和技能应用于新环境不同的是，深度网络在学习新信息时可能会遇到旧类别知识被覆盖的问题，这被称为灾难性遗忘 [21]–[23] 。

持续学习（CL）[24]–[27] 研究如何构建一个模型，使其能够从数据和任务中持续学习新技能，同时避免遗忘先前任务。通过使模型能够增量学习，CL 方法 [28]–[30] 允许深度神经网络适应新任务，同时保留之前学到的知识。在持续图像分割任务中，许多方法已取得显著进展。基于知识蒸馏的持续分割方法 [13], [31], [32] 通过从旧模型传递知识来减轻灾难性遗忘，同时伪标签允许新模型使用先前学习到的类别标签进行训练。然而，基于知识蒸馏的方法由于需要双网络传递和超参数微调而增加了计算开销，这使得训练变得复杂。此外，随着类别数量的增加，保持高效且可扩展的蒸馏过程变得困难，知识传递可能难以有效捕捉新的任务特征 [33] 。基于提示调优的方法 [33]–[38] 通常设计一个共享的键值对池来管理每轮学习的提示。这些键的学习标准通常涉及弱监督，将选定的键拉近到对应的查询特征，同时附加正交约束以促进多样化。这类方法的主要问题是，由于键空间被不断更新且无法访问过去的查询或任务，选择机制本身容易发生灾难性遗忘，可能导致查询和键之间的不匹配 [39] 。

持续视频实例分割（CVIS）引入了比静态图像分割更多的独特复杂性。首先，当模型必须在学习新类别的同时保留对旧类别的时空推理能力时，会出现时间遗忘。如果不能在特征（例如，运动模式或物体轨迹）中保持时间一致性，就会导致跟踪性能随着时间的推移而退化。其次，增量更新可能会扭曲潜在的实例嵌入，而这些嵌入对于保持一致的跨帧关联至关重要。与实例身份是帧局部的图像任务不同，视频模型依赖于时间上稳定的嵌入来连接跨帧的实例。即便是持续学习过程中细微的参数变动，也可能在跟踪中传播错误。第三，旧类别缺乏标注加剧了背景模糊。在动态视频场景中，未标注的旧类别实例（例如，在新的“巴士”训练阶段之前学习的“汽车”类别）有被误分类为正在演变的背景的一部分的风险，特别是当运动线索重叠时。最后，视频数据（包括空间和时间）的高维性对基于复习的传统持续学习策略提出了挑战。存储过去类别的代表性样本在计算上变得不可行，而简单的帧采样往往无法捕捉复习所需的重要时间上下文。

由于目前没有用于 CVIS 的方法，我们将最先进的持续图像实例分割方法 ECLIPSE [33] 适应到视频领域，并评估其性能。ECLIPSE 是一个免蒸馏的持续图像分割框架，它利用冻结的主干网络和轻量级的提示机制来平衡稳定性和可塑性。通过冻结基础模型的所有参数，并在引入新类别时仅迭代地微调一小部分提示嵌入，ECLIPSE 在某种程度上自然地缓解了灾难性遗忘，同时保留了先前的知识。然而，当我们将 ECLIPSE 应用于 CVIS 任务时，如图 1 所示，ECLIPSE 在三个层面上表现出语义混淆：实例级、类别级和任务级。在第一个序列中，实例 [0] 是一个人，而实例 [1] 和 [2] 是摩托车。随着视频的进行，我们观察到 ECLIPSE 错误地将实例 [1] 和 [2] 合并，它们是两个在道

This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFB4707700, the National Natural Science Foundation of China under Grant U23A20343, and the Doctoral Research Startup Fund of the Natural Science Foundation of Liaoning Province under Grant 2025-BS-0193. (Corresponding author: Zhi Han).

Baichen Liu, Qi Lyu, Xudong Wang, Lianqing Liu, and Zhi Han are with the State Key Laboratory of Robotics and Intelligent Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: liubaichen@sia.cn; lvqi@sia.cn; wangxudong@sia.cn; lqliu@sia.cn; hanzhi@sia.cn;). Qi Lyu and Xudong Wang are also with the University of Chinese Academy of Sciences, Beijing 100049, China.

Jiahua Dong is with the Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates. (e-mail: dongjiahua1995@gmail.com).

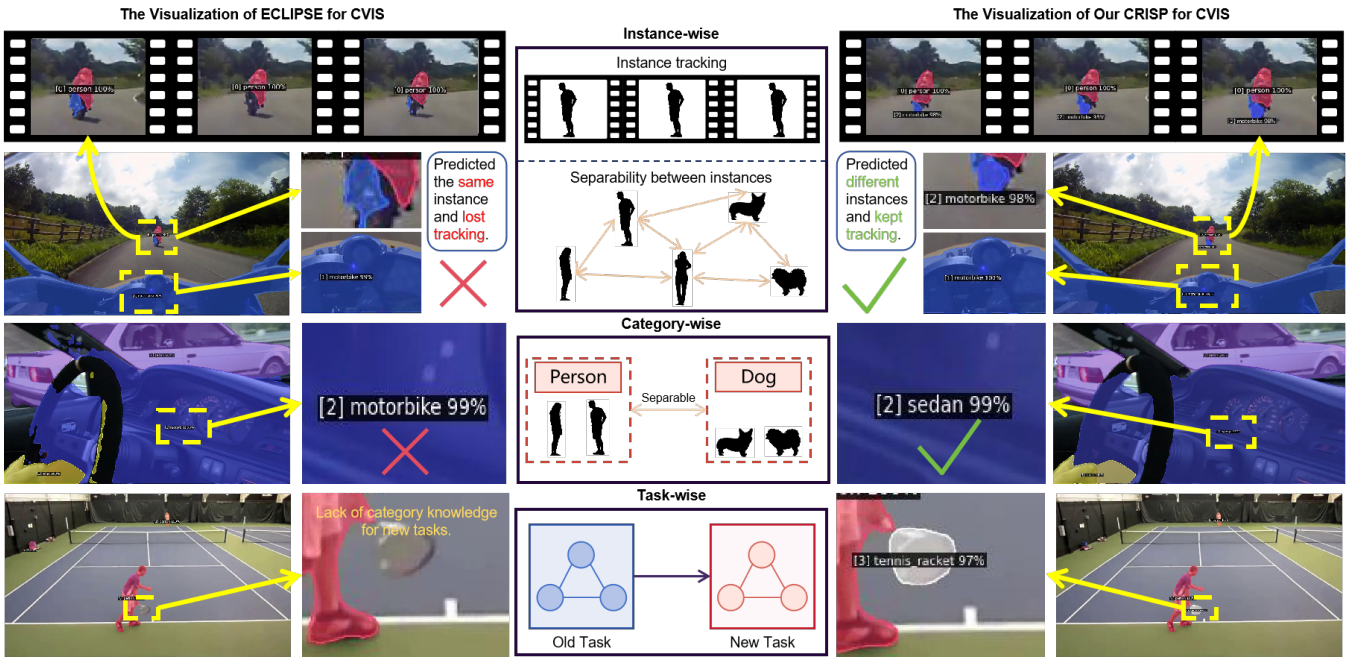

Fig. 1. ECLIPSE [33] 和我们的 CRISP 方法在 CVIS 任务上的比较。左侧，ECLIPSE 在实例层次、类别层次和任务层次上表现出语义混淆，使得难以准确区分不同实例，导致同类别实例之间的误分类，以及在学习新类别时的频繁遗漏。右侧，我们的 CRISP 框架在 CVIS 任务上有效地解决了这些问题。

路上保持一定距离的摩托车，成为一个单一的实例。我们将这种类型的错误称为“实例级”混淆。在第二个序列中，实例 [0] 和 [2] 是轿车，而实例 [1] 是一只手。我们观察到 ECLIPSE 有时会错误地将实例 [2] 分类为摩托车。我们定义这种同一视频和类别的实例被错误分类到不同类别的错误为“类别级”混淆。在第三个序列中，场景中包含旧任务的“人”类别和当前任务的“网球拍”类别。在持续学习过程中，为新任务引入的可学习查询的数量少于最初使用的数量。为了保留旧类别的知识，ECLIPSE 通过平均池化和复制旧类查询来初始化增量查询。然而，这种初始化方案进一步削弱了其学习新类别的能力，导致 ECLIPSE 无法检测到当前任务的“网球拍”实例 [3]。我们将这种由于可学习变量数量限制仍然与旧任务绑定并受到影响而导致的限制称为“任务级”混淆。

为了进一步分析语义混淆背后的原因，我们从图 1 的第二序列的最左侧图像中提取了三个实例的 150 个查询，并使用 t 分布随机邻嵌入 (t-SNE) [40] 对它们进行投影。此外，我们报告了实例之间成对的欧氏距离。如图 2a 所示，轿车和摩托车的嵌入紧密相邻，解释了为什么 ECLIPSE 会混淆这些类别。此外，在持续学习过程中，我们分析了图 2c 中增量查询与旧类查询之间的协方差矩阵，发现增量查询彼此高度相似。我们假设这个问题源于 ECLIPSE 为增量查询的初始化方案，该方案通过对旧类别查询进行平均池化和复制来生成新类别查询，从而导致语义漂移。

基于对三个层次语义混淆的上述调查，在本文中，我们提出了一种名为对比残差注入和语义提示（CRISP）的方法，以解决 CVIS 任务中的挑战。首先，对于实例级学习，考虑到实例之间的特异性，我们对实例跟踪进行建模并构建实例相关性损失，该损失强调与先验查询空间的相关性，同时加强当前任务查询的特异性。其次，对于类别级学习，我们构建了一个自适应残差语义提示（ARSP）学习框架，该框架构建了一个由类别文本生成的可学习残差语义提示池，并使用一个可调的查询提示匹配器来建立当前任务查询与残差语义提示之间的映射关系。ARSP 模块将相同实例的嵌入拉近，并将不同类别的嵌入拉远。图 1 的第二行右侧显示我们的方法正确标记了两个轿车实例，图 2b 显示轿车嵌入（实例 [0] 和 [2]）形成了一个与“手”嵌入很好区分的紧密聚类，证实了我们的方法显著提高了类内紧凑性和类间可分性。同时，引入了一种基于对比学习的语义一致性损失，以保持增量训练过程中，目标查询和残余提示之间的语义一致性。第三，对于任务层面学习，为了确保查询空间内任务间的相关性，我们引入了一种简洁而强大的增量提示初始化策略。我们将增量学习过程中的提示视为查询向量，对旧任务查询的集合应用主成分分析 [41], [42] (PCA)，以识别其最重要的方向。如图 2d 所示，这种基于 PCA 的方案成功地去关联了新的增量查询，而在图 1 的第三行右侧，我们的 CRISP 方法正确地标记了新任务的“tennis_racket”类别。

总之，为了解决 CVIS 中的实例混淆、类别混淆和任务混淆，我们的 CRISP 方法是一种早期应用于 CVIS 任务的尝试，有效地缓解了灾难性遗忘和语义混淆的问题。对 YouTube-VIS-2019 [43] 和 YouTube-VIS-2021 [44] 的实验结果表明，我们提出的 CRISP 方法在 CVIS 中达到了最先进的性能。此外，随着连续学习步骤的增加，CRISP 相比其他连续学习方法显示出了显著的改进。

本文的贡献如下：

- 我们提出的 CRISP 是应用于 CVIS 任务的早期尝试，取得了最先进的结果，特别是在连续学习步骤次数增加的情况下。
- 对于实例间的混淆，我们提出了一种实例关联损失，该损失显式地建模实例跟踪，以增强实例之间的特异性。
- 针对类别混淆，我们提出了一种一级残差语义注入模块 ARSP，通过将残差语义提示注入自注意力层，有

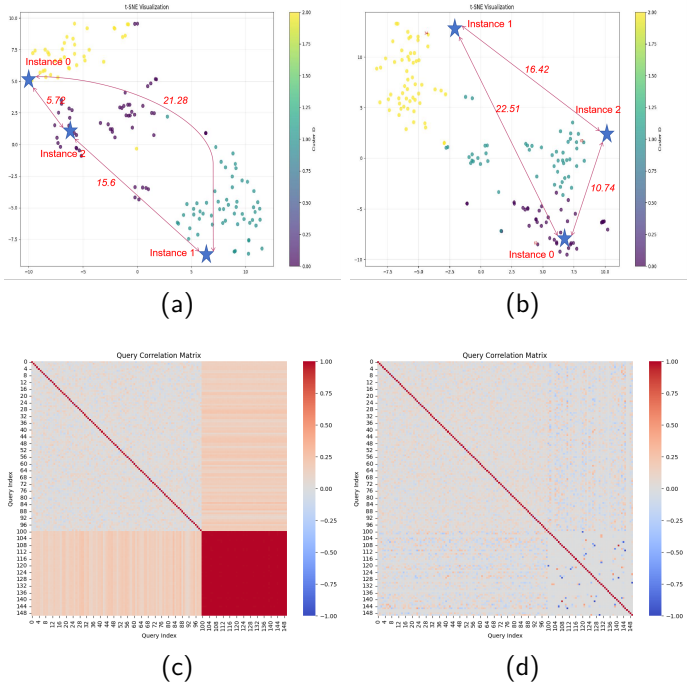

Fig. 2. 对比分析在 ECLIPSE [33] 和我们的 CRISP 框架下的查询嵌入及其相关性。(a) 使用 ECLIPSE 对三个实例的 150 个查询进行 t-SNE 可视化，显示轿车和摩托车之间的重叠簇，解释了类别混淆。(b) 在 CRISP 下对相同查询进行 t-SNE 可视化，展示各个实例的紧密且分开的簇。(c) ECLIPSE 中增量和旧类查询的协方差矩阵，揭示了容易导致语义漂移的高查询间相关性。(d) CRISP 的协方差矩阵，显示出有效的去相关和稳定的嵌入关系，以防止漂移。

效提高模型的分割性能。此外，我们基于对比学习提出了一种语义一致性损失，以增强相似样本的一致性。

- 针对任务方面的混淆，我们提出了 PCA 引导初始化，该方法有效地保持了先前查询空间中的相关性，同时确保当前任务查询空间的特异性。

II. 相关工作

实例分割。早期的图像实例分割方法如 MNC [16] 展示了在级联流水线中结合检测和掩码预测的可行性。随后，FCIS [17] 推进了这项工作，通过支持实例掩码与分类的端到端训练。Mask R-CNN [45] 的引入在一个两阶段框架内统一了目标检测和掩码生成，设定了准确性和灵活性的新标准。为了满足实时需求，SOLO [18] 提出了一个基于位置的单阶段方法，提供速度与精度。最近，基于 transformer 的架构如 DETR [46] 将图像实例分割重新定义为一个直接的集合预测问题，而 Mask2Former [19] 这样的通用框架通过整合注意力机制来扩展这一范式，适用于全景、实例和语义分割任务。

VIS 是一项任务，旨在同时检测、分割和跟踪视频序列中的实例对象。ISTR [47] 是第一个以 Transformer 为基础的端到端 VIS 框架，将 VIS 任务视为直接的端到端并行序列解码问题。IFC [48] 由一个 CNN 骨干网和 Transformer 编码-解码层组成，通过有效编码输入片段中的上下文，利用简洁的记忆标记作为传递信息以及总结每帧场景的手段，减少了帧间信息传递的开销。MTN [49] 采用两个教师网络指导学生网络学习旧知识和新知识。具体而言，前一个教师网络监督当前学生网络以保留先前知识，当前教师网络监督当前学生网络以适应新类别。然而，在解决灾难性遗忘和背景转移的问题上，CVIS 任务的挑战仍然是一个前沿研究难题。

连续图像分割。大多数连续图像分割方法通常采用蒸馏策略，例如知识蒸馏 [13], [31], [32], [50] 和伪标签 [51], [52]。知识蒸馏通过从旧模型传递知识来缓解灾难性遗忘，而伪标签允许新模型使用以前学习类别的标签进行训练。这些方法在连续图像分割任务中取得了良好效果。然而，现有为静态图像设计的连续分割方法，忽视了跨帧错位实例关联的复合错误，这在视频任务中会加剧遗忘。此外，在以图像为中心的框架中，并未解决在增量更新过程中运动和外观表示的渐进退化，即时间特征漂移。关键的是，能够帮助区分背景和未注释旧类别实例的运动线索在当前连续学习范式中仍未得到充分利用。

在持续学习中基于提示和基于查询的方法。基于提示的方法通过学习一小组可插入的模型提示，而不是直接修改编码器参数，具备强大的能力以防止持续学习中的灾难性遗忘。另一方面，基于查询的方法利用一组固定的可学习对象查询来探测由冻结的或部分微调的编码器产生的视觉特征。每个查询负责通过 Transformer 架构（例如，DETR [46] 、可变形 DETR [53] 、Mask2Former [19] ）中的交叉注意机制关注并提取特定实例级的表示，比如对象掩码或边界框。通过将实例提取与编码器更新解耦，基于查询的设计在增量任务中保持稳定的特征关联，有助于保持对象实例的时间和空间一致性。因此，在持续分割任务中，基于查询和基于提示的方法是互补的：查询驱动模型视觉输入的实例级推理，而提示注入特定任务或特定类别的语义，以指导学习过程并在增量步骤中缓解遗忘。

首次尝试在持续学习中使用提示微调的是 Learning-to-Prompt (L2P) [38] ，其使用一个跨所有任务的共享提示池，并使用输入图像作为查询从池中选择最合适的提示。DualPrompt [37] 引入了一般和任务特定提示的层次结构，采用前缀微调而不是传统的提示微调，其中前缀提示被

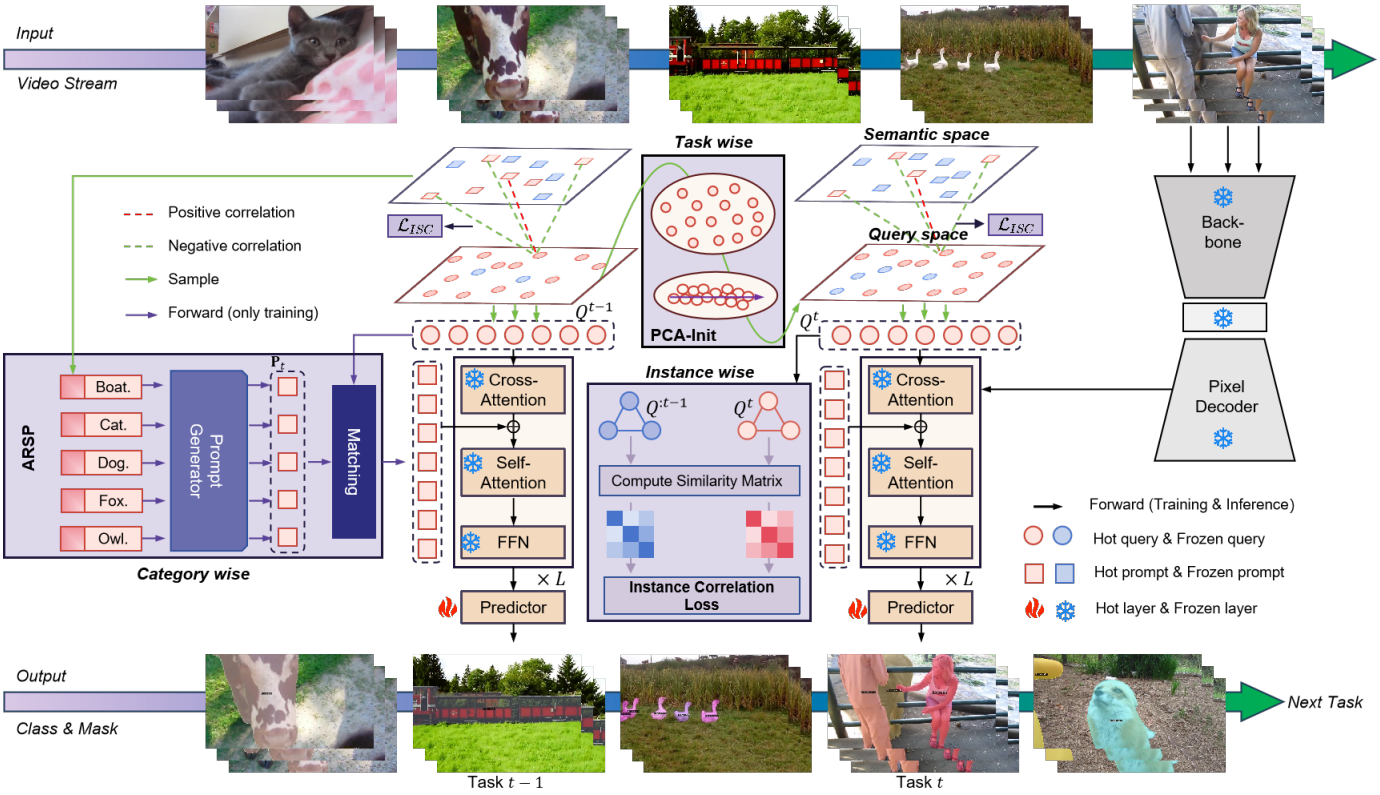

Fig. 3. 我们提出的用于 CVIS 的 CRISP 架构。包含类别相关的实例语义一致性损失（第 ?? 节）的 ARSP 模块（第 III-C 节），任务相关的 PCA 引导初始化（第 ?? 节），以及实例相关的实例关联损失（?? ）。值得注意的是，我们的方法在训练期间只注入残差语义提示，而且旧任务的查询和提示将会被冻结。

添加到 MSA 层的键和值中，而不是直接添加到输入标记。CODA-Prompt [35] 进一步引入了端到端提示机制和正交软约束，以促进提示的独立性。STAR-Prompt [39] 提出了利用 CLIP 模型的稳定类原型和查询图像作为键来检索提示的两级适应机制，增强提示选择策略的稳定性。CoMFormer [54] 引入了第一个以查询为中心的持续全景分割方法，结合知识蒸馏和伪标签缓解灾难性遗忘。在此基础上，CoMasTRe [12] 保留了蒸馏目标，但在持续学习管道中将掩码生成与类别预测解耦。类似地，BalConpas [55] 通过汇集特征级蒸馏与代表性样本的重放缓冲来对抗遗忘，使模型能够在不损害已经学习的知识的情况下获取新类别。

III. 方法论

A. 问题设置

CVIS 的目标是学习一个能够在视频序列中分割和跟踪对象实例的模型，同时随着多次学习阶段的进行，逐步适应新的类别。令 $\mathcal{V}$ 表示输入视频空间，其中每个视频 $v \in \mathcal{V}$ 由一系列时间帧 $\{x_1, \ldots, x_T\}$ 组成。在每个学习步骤 t 中，引入一组新的类别 $\mathcal{C}_t$ ，并且模型需要在所有观察到的类别 $\mathcal{C}_{1:t}$ 中对帧之间的实例进行分割和关联。在步骤 t 中，模型接收包含仅对 $\mathcal{C}_t$ 进行注释的视频的训练数据集 $\mathcal{D}_t$ 。每个注释由帧 x_i 中的 k^{th} 对象的实例掩码 $m_{i,k}^t \in \{0,1\}^{H \times W}$ 和类别标签 $c_{i,k}^t \in \mathcal{C}_t$ 组成，其中 $H \times W$ 是空间分辨率。关键是，之前类别 $\mathcal{C}_{1:t-1}$ 和未来类别 $\mathcal{C}_{t+1:T}$ 的注释是完全不存在的，甚至没有被合并到背景类中。模型必须预测 $\mathcal{C}_{1:t}$ 的每帧实例掩码和类别标签，通过在帧之间关联相同的实例来保持时间一致性，并在学习 $\mathcal{C}_t$ 的同时保持对 $\mathcal{C}_{1:t-1}$ 的性能，以避免灾难性遗忘。

B. 概述

首先，由于任务之间固有的视觉相似性，各类别之间的混淆是不可避免的。为缓解这一问题，我们提出了一种自适应残差语义提示（ARSP）学习框架，并辅以实例语义一致性损失以增强判别特征的学习。其次，尽管通过冻结大多数参数的提示微调部分缓解了灾难性遗忘，但仍然存在任务混淆的问题。为解决这一限制，我们引入了 PCA 引导的初始化，这是一种新颖的方法，允许新任务查询有效继承先前任务的知识，从而提高任务区分度。第三，为应对实例混淆，我们提出了一种基于实例追踪建模的实例关联损失，该损失加强了跨任务实例表示的一致性。

C. 自适应残差语义提示训练

ARSP 是一种一级提示调优策略，用于指导模型在训练期间学习语义信号，以解决类别混淆问题。如图 3 所示，ARSP 集成了三个核心组件：残差语义提示生成器、自适应查询提示匹配机制和分层注入架构。

残差语义提示生成器。遵循 CoOp [56] 范式，我们构建了一个残差语义提示池 $\boldsymbol{P}_t = [\boldsymbol{p}_0|\boldsymbol{p}_1|\boldsymbol{p}_2|...|\boldsymbol{p}_{c_t}] \in \mathbb{R}^{c_t \times d}$ ，其中包含 c_t 个特定类别的原型，这些原型位于 d 维空间中，其中 c_t 对应任务 t 的总类别数。一个提示生成器 $\mathcal{G}_\theta$ ，实现在 CLIP [57] 文本编码器上，将可学习的符号序列 $\mathcal{X} \in \mathbb{R}^{c_t \times d}$ 通过 $\mathcal{G}_\theta$ 映射到残差语义提示 $\boldsymbol{P}_t$ 。

自适应查询提示匹配。我们通过一种匹配机制获得与每个查询 $q^l \in \mathbb{R}^d$ 对应的残差语义提示。匹配过程可以表示为：

$$\begin{aligned} \boldsymbol{a} &= (a_1, a_2, \cdots, a_{c_t})^T, \\ \text{s.t. } a_i &= \arg\max_{j=1}^{c_t} S_{i,j} \quad \text{for} \quad i = 1, 2, \ldots, N_q^t, \\ S_{i,j} &= \frac{\boldsymbol{Q}_t \boldsymbol{P}_t^T}{\| \boldsymbol{Q}_t \|_F \| \boldsymbol{P}_t^T \|_F}, \end{aligned} \tag{1}$$

其中 N_q^t 表示任务 t 中的查询数量，$\boldsymbol{Q}_t \in \mathbb{R}^{N_q^t \times d}$ 表示任务 t 具有维度 d 的对象查询，$\|\cdot\|$ 表示 Frobenius 范数。

层次注入架构。在视频 [20] 架构中建立在 Mask2Former 的基础上，我们通过一种改进的自注意力操作将残差提示注入到多尺度特征解码中：

$$\boldsymbol{O}^l = Softmax(\frac{\boldsymbol{Q}^l (\boldsymbol{K}^l)^T}{\sqrt{d_k}})(\boldsymbol{V}^l + \boldsymbol{P}_m^l), \tag{2}$$

，其中 $\boldsymbol{O}^l$ 代表自注意力 l 层的输出，$\boldsymbol{Q}^l$ 、$\boldsymbol{K}^l$ 和 $\boldsymbol{V}^l$ 分别表示 Transformer 解码器层 l 的查询、键和值，$\boldsymbol{P}_m^l$ 表示与 $\boldsymbol{Q}^l$ 对应的匹配残差语义提示。

在推理过程中，我们不注入残差语义提示，以提高推理速度。因此，为了在持续训练过程中保持对象查询和残差提示之间的语义一致性，我们基于 [58] 中提出的对比学习框架制定了一个语义一致性损失，其将语义空间与查询空间对齐。给定对象查询 $\boldsymbol{Q}_t \in \mathbb{R}^{N_q^t \times d}$ 和当前任务的残差提示 $\boldsymbol{P}_t \in \mathbb{R}^{c_t \times d}$ ，通过计算余弦相似度矩阵 $\boldsymbol{S} \in \mathbb{R}^{N_q^t \times c_t}$ ，实例语义一致性损失表示为：其中 a_i 表示通过将方程 1 应用于对象查询 $q^l \in Q^l$ 得出的索引，而 $\mathbb{I}_{ij}$ 表示为：

$$\mathbb{I}_{ij} = \begin{cases} 0 & \text{if } j = a_i \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

虽然 ECLIPSE 通过均值池化和复制操作为新类别 $\mathcal{C}^t$ 生成查询，在图像分割任务中表现出色，但其在直接适应视频分割任务时表现出明显的局限性。具体来说，由于如图 2c 所示生成的查询之间的高一致性，这一特性显著影响了长期视频实例分割场景中任务之间的可分性，导致任务间的混淆。为了缓解训练期间由于查询一致性引起的任务混淆，我们提出了一种简洁有效的提示初始化策略。首先，应用主成分分析 [42] (PCA) 提取旧任务查询的主成分。随后，我们从对应最大特征值的前 c_t 个中采样主成分向量。最后，这些采样的潜在表示与先前的查询流形对齐，以形成当前任务的初始查询。初始化过程在算法 1 中有详细介绍。

在 CVIS 任务中，我们将跨帧实例跟踪的挑战以及同一帧中不同实例之间的混淆归类为实例级混淆。视频中的 Mask2Former 引入了对象查询，通过用一系列查询表示不同的实例来建模实例之间的关联。假设模型总共有 N 个查询，它理论上可以跟踪 N 个实例。因此，这些查询应该是成对可分的。为此，我们将实例跟踪建模为:

$$min \parallel \hat{\boldsymbol{Q}} \hat{\boldsymbol{Q}}^T - \boldsymbol{I} \parallel_F, \tag{4}$$

，其中 $\hat{\boldsymbol{Q}}$ 代表标准化查询，$\boldsymbol{I}$ 是身份矩阵。

考虑到同一类别中的不同实例应具有某种相似性，因此查询并不是完全正交的。因此，我们利用 $\boldsymbol{Q}$ 和 $\boldsymbol{Q}^T$ 的内积矩阵表征查询之间的相关性，并将其建模为:

$$min \parallel \hat{\boldsymbol{Q}} \hat{\boldsymbol{Q}}^T - \hat{\boldsymbol{Q}}_0 \hat{\boldsymbol{Q}}_0^T \parallel_F . \tag{5}$$

Algorithm 1 PCA 引导初始化。

Input: queries of old tasks $\boldsymbol{Q}_o \in \mathbb{R}^{N_q^{:t-1} \times d}$, the length of queries from previous tasks $N_q^{:t-1}$, total categories in old tasks $c_{:t-1}$, total categories in current task c_t .

Output: initialized queries for current task $\boldsymbol{Q}_t$.

$\boldsymbol{Q}' \leftarrow PCA(\boldsymbol{Q}_o) \in \mathbb{R}^{c_{:t-1} \times d}$;

$\mathcal{N} \leftarrow [\parallel q'_1 \parallel_2, \parallel q'_2 \parallel_2, \parallel q'_3 \parallel_2, \cdots, \parallel q'_{c_{:t-1}} \parallel_2]$;

Select indices of top- c_t norms:

$\mathcal{I} \leftarrow argsort(-\mathcal{N})[: c_t]$;

$\boldsymbol{Q}_t \leftarrow \boldsymbol{Q}'[\mathcal{I}, :] \in \mathbb{R}^{c_t \times d}$;

Align $\boldsymbol{Q}_t$:

$a_{ori} \leftarrow \frac{1}{N_q^{:t-1}} \sum_{i=1}^{N_q^{:t-1}} \parallel \boldsymbol{Q}_o[i, :] \parallel_2$;

$a_{pca} \leftarrow \frac{1}{c_{:t-1}} \sum_{i=1}^{c_{:t-1}} \mathcal{N}$;

$\boldsymbol{Q}_t \leftarrow \frac{a_{ori}}{a_{pca}} \boldsymbol{Q}_t$;

return $\boldsymbol{Q}_t$.

TABLE I
CVIS 在 YouTube-VIS-2019 数据集上的结果。上面显示的是 20-4 情景的结果，下面显示的是 20-2 情景的结果。最佳结果以红色突出显示，次优结果以蓝色突出显示。以下表格使用相同的标记。

Model	mAP	AP50	AP75	AR1	AR10	FR
FT	6.53	11.39	6.96	11.59	11.99	21.19
MiB [59]	9.01	17.75	7.29	14.03	24.34	17.78
CoMFormer [54]	12.76	24.43	11.08	16.85	26.22	13.6
ECLIPSE [33]	25.03	39.83	27.59	32.81	40.48	2.23
CoMBO [60]	15.8	29.40	16.6	25.54	28.55	16.05
CRISP	28.1	43.3	31.98	38.04	44.71	1.93
FT	3.72	6.61	4.27	9.67	10.13	15.56
MiB [59]	0.94	2.07	0.91	3.58	5.86	16.26
CoMFormer [54]	5.26	9.01	5.32	9.57	12.01	11.4
ECLIPSE [33]	19.09	29.05	21.81	26.06	31.92	3.21
CoMBO [60]	4.37	8.17	4.25	9.59	10.71	15.02
CRISP	25.14	37.46	28.17	32.87	39.13	1.13

基于方程 5 ，我们提出了一种实例相关性损失函数，表示为：

$$\mathcal{L}_{IC} = MSE(\hat{\boldsymbol{Q}}_t \hat{\boldsymbol{Q}}_t^T, \hat{\boldsymbol{Q}}_0 \hat{\boldsymbol{Q}}_0^T), \tag{6}$$

其中 $\hat{\boldsymbol{Q}}_t$ 表示任务 t 的归一化查询，$\hat{\boldsymbol{Q}}_0$ 表示初始任务的归一化查询，$MSE(\cdot)$ 表示均方误差损失。

D. 总损失

我们遵循 [19] 来定义分割损失函数 $\mathcal{L}_{Seg}$ ，该损失函数由二元交叉熵损失和 dice 损失组成的掩码损失以及由交叉熵定义的分类损失组成。总损失函数可以表示为：

$$\mathcal{L}_{tol} = \mathcal{L}_{Seg} + \lambda_{ISC} \mathcal{L}_{ISC} + \lambda_{IC} \mathcal{L}_{IC}. \tag{7}$$

在我们的方法中，我们设置 $\lambda_{ISC} = \lambda_{IC} = 3$ 。我们还在每个 transformer 解码器层 ($l \in \{1, \ldots, L\}$) 添加辅助实例相关损失 $\mathcal{L}_{IC}^l$ 。

IV. 实验

A. 实验设置

数据集：我们在 YouTube-VIS-2019 [43] 和 YouTube-VIS-2021 [44] 上评估了我们的方法。YouTube-VIS-2019 包含 2,883 个高分辨率的 YouTube 视频，包括 2,238 个训练视频，302 个验证视频和 343 个测试视频。标注包括 40

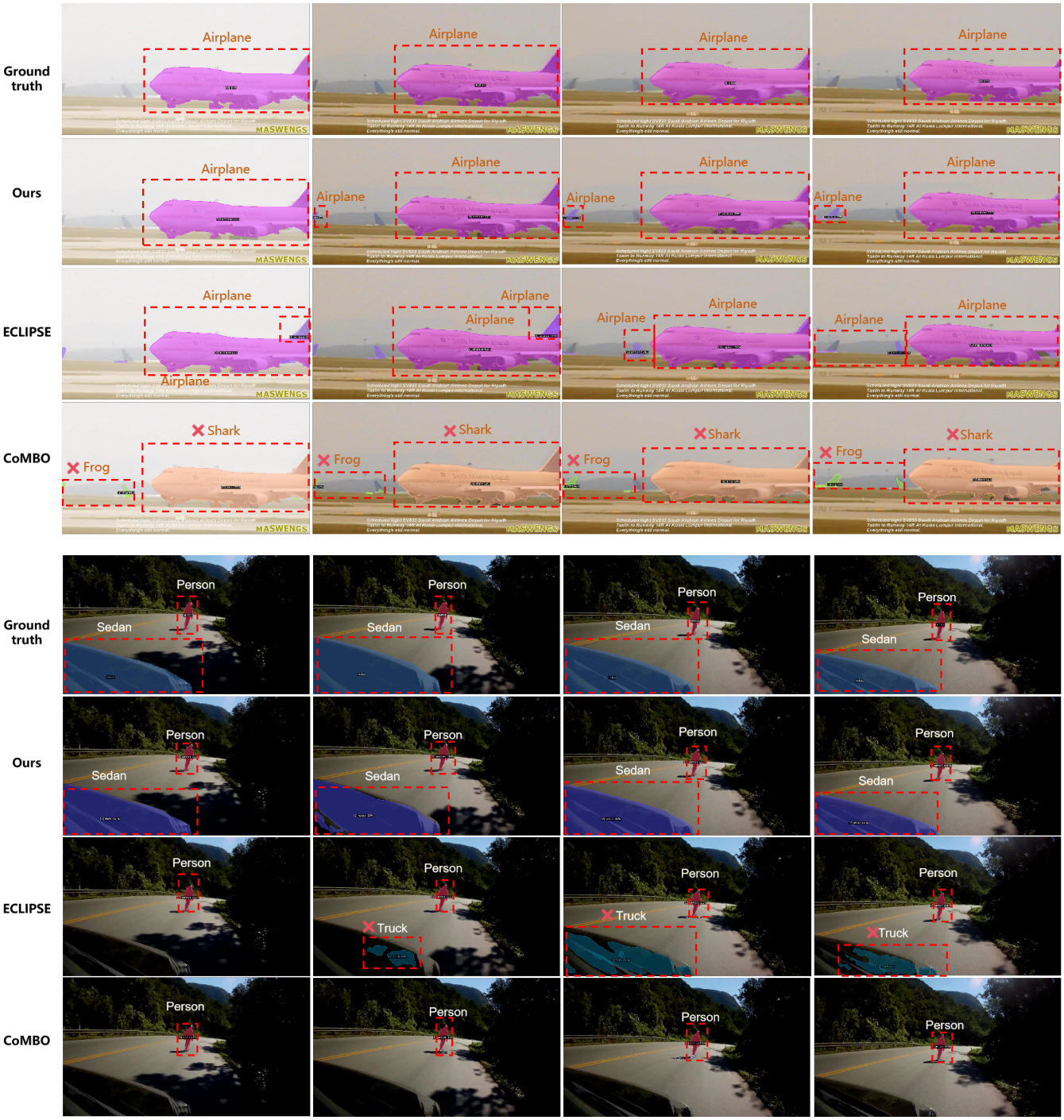

Fig. 4. 与其他方法的比较分析。在上面的例子中，我们的方法能够准确地分割出人和轿车，同时保持一致的实例跟踪。在下面的例子中，它正确识别出了飞机，并且还分割出了一个在真实数据中没有标注的实例。

Fig. 5. 通过我们的 CRISP 对几个具有挑战性样本的可视化结果。在第一组视频帧中，我们的 CRISP 成功实现了对两个外观相似且重叠的猿准确分割。第二组帧展示了我们的 CRISP 对高速移动汽车的有效跟踪能力。最终实验结果证明，CRISP 即使在目标人物暂时消失和重新出现时仍能保持稳定的分割性能。

TABLE II
CVIS 在 YouTube-VIS-2021 数据集上的结果。上面显示的是 20-5 场景的结果，下面显示的是 10-10 场景的结果。

Model	mAP	AP50	AP75	AR1	AR10	FR
FT	4.08	6.90	4.52	6.43	7.36	32.63
MiB [59]	5.82	10.87	4.49	9.52	15.96	25.44
CoMFormer [54]	14.49	28.01	14.46	23.01	30.15	10.58
ECLIPSE [33]	22.26	33.39	24.74	24.81	31.03	5.03
CoMBO [60]	15.35	25.18	16.40	24.01	29.39	19.88
CRISP	26.13	39.13	28.82	28.84	37.49	4.38
FT	5.12	10.24	4.54	7.97	9.34	42.99
MiB [59]	8.14	13.03	9.06	10.11	12.25	24.95
CoMFormer [54]	12.83	23.35	13.19	16.92	26.14	23.67
ECLIPSE [33]	19.04	30.94	19.92	25.4	32.06	4.38
CoMBO [60]	20.29	33.94	21.58	26.19	33.64	19.75
CRISP	20.05	31.82	21.54	25.26	32.71	2.73

个常见的对象类别，如人类、动物和车辆。YouTube-VIS-2021 总共有 3,859 个高分辨率的 YouTube 视频，其中包含 2,985 个训练视频，421 个验证视频和 453 个测试视频。标签基于 YouTube-VIS-2019 的标签进行了更新，合并了鹰和猫头鹰为鸟类，将猿归入猴子，去掉了手，并添加了新的类别，包括飞盘、松鼠和鲸鱼。

指标：我们报告几个标准的指标来对两个数据集进行全面评估 [43], [44] 。平均精度（AP）（精度-召回曲线下的面积），最终任务结果报告 mAP（平均平均精度）、AP50 和 AP75，而每个任务的分析包括 APs、APm 和 APl。平均召回率（AR）（视频中固定分割实例下的最大召回率），报告 AR1 和 AR10；以及持续学习评估。遗忘率（FR）量化在新任务学习过程中遗忘之前任务知识的程度。FR 的计算如下：

$$FR = \frac{1}{N_c}\sum_{t=1}^{T-1}\frac{1}{T-t}\sum_{c_t=1}^{C_t}\frac{\mathbb{I}(A_{t,c}-A_{T,c})_{\geq 0}(A_{t,c}-A_{T,c})}{A_{t,c}}, \quad (8)$$

其中 N_c 表示类别的数量，C_t 表示任务 t 中的类别数量，$A_{t,c}$ 和 $A_{T,c}$ 分别表示模型首次学习到类别 c 和最终任务时的 mAP。$\mathbb{I}(A_{t,c} - A_{T,c})_{\geq 0}$ 可以表示为：

$$\mathbb{I}_{ij} = \begin{cases} 0 & \text{if } A_{t,c} - A_{T,c} \geq 0 \\ 1 & \text{otherwise.} \end{cases} \quad (9)$$

持续学习协议：我们采用先前任务中的增量方法，并使用符号 $N_{ini} - N_{inc}$ 来表征场景，其中 N_{ini} 表示初始类别数，N_{inc} 表示每个增量阶段增加的类别数。为了全面评估，我们为每个数据集设计了两个具有挑战性的协议。对于 YouTube-VIS-2019，我们设置了 20-4，首先学习 20 个基础类别，然后每步添加 4 个新类别（N_t =6），以及 20-2，学习 20 个基础类别，然后每步添加 2 个新类别（N_t =11）。对于 YouTube-VIS-2021，我们设置了 20-5，首先学习 20

TABLE III
20-4 在 YouTube-VIS-2019 数据集上的结果。表格中的每个单元格分别代表 APs、APM 和 APL。这一设置在表 III - VI 中始终一致应用。

Model	Step0	Step1	Step2	Step3	Step4	Step5
FT	12.20/17.64/21.07	9.34/15.86/19.49	6.01/10.39/12.23	6.31/9391/15.17	3.96/5.03/10.25	4.03/4.25/10.81
MiB [59]	10.82/16.63/22.15	9.9/13.54/20.43	6.58/17.52/21.12	6.48/10.74/21.13	5.12/12.52/19.71	3.4/10.83/19.02
CoMFormer [54]	10.34/16.42/19.57	9.26/17.65/22.09	9.76/17.61/21.83	8.84/16.26/24.36	7.07/13.23/24.87	8.32/13.5/20.24
ECLIPSE [33]	11.96 / 22.03 /27.86	9.75/19.92/26.63	10.28/ 20.89 /28.18	10.88 / 22.88 /32.19	10.88 / 22.77 / 32.99	12.76 / 24.96 / 35.65
CoMBO [60]	13.16 / 22.51 / 28.72	12.34 / 22.20 / 29.87	10.67 /19.78/ 30.49	10.56/18.97/ 33.66	8.62/18.94/29.10	6.44/14.5/26.16
CRISP	11.96 /21.96/ 29.03	11.38 / 21.78 / 28.16	12.32 / 22.89 / 29.61	13.6 / 24.07 / 33.5	13.5 / 27.7 / 35.51	15.74 / 30.02 / 39.65

TABLE IV
20-2 结果在 YouTube-VIS-2019 数据集上。

Model	Step0	Step1	Step2	Step3	Step4	Step5	Step6	Step7	Step8	Step9	Step10
FT	12.2 17.64 21.07	12.21 15.42 17.19	3.88 11.53 10.95	3.71 9.51 10.73	5.44 8.8 10.18	3.38 8.66 11.29	6.71 3.95 10.26	5.79 2.35 6.71	2.29 3.52 4.56	6.34 3.09 6.24	1.11 2.85 7.22
MiB [59]	10.82 16.63 22.15	11.78 14.13 17.57	10.09 12.97 18.81	7.13 10.93 18.77	7.36 10.21 16.23	7.68 9.16 16.28	4.62 8.79 14.97	1.12 2.61 9.26	1.32 2.18 6.09	3.51 2.09 5.43	0.37 3.1 3.94
CoMFormer [54]	8.53 13.78 17.81	7.73 11.95 17.38	9.00 12.63 19.85	6.93 11.65 19.13	6.29 7.36 17.35	3.95 .7.61 15.30	5.69 5.01 15.90	4.43 4.28 10.94	1.96 3.81 11.52	2.79 6.25 11.26	2.51 4.77 9.52
ECLIPSE [33]	12.42 22.52 27.58	11.45 21.1 25.89	10.92 21.8 26.77	9.6 20.44 25.84	9.51 19.55 26.49	9.51 19.4 27.91	10.78 17.98 29.26	10.03 17.58 28.81	10.02 17.57 27.5	10.64 16.8 26.82	10.58 17.28 28.58
CoMBO [60]	10.77 21.10 28.06	11.08 22.01 29.72	7.88 20.51 26.08	8.65 18.51 24.70	4.79 15.40 22.60	7.41 16.32 22.58	6.31 9.00 17.64	4.50 5.85 16.67	3.56 5.38 13.52	4.26 2.92 10.21	2.91 3.26 8.29
CRISP	13.53 22.99 29.33	12.67 20.94 28.68	12.71 21.57 29.31	12.71 22.18 29.60	13.09 22.18 29.91	13.09 22.31 31.95	15.52 22.06 34.34	15.51 22.06 34.40	15.51 23.86 34.45	15.51 24.01 34.51	14.50 24.59 34.44

TABLE V
在 YouTube-VIS-2021 数据集上的 20-5 结果。

Model	Step0	Step1	Step2	Step3	Step4
FT	8.91/18.53/22.21	6.21/14.05/20.31	0.93/2.75/4.66	0.99/5.12/3.24	2.42/6.4/5.55
MiB [59]	8.95/17.64/22.74	4.78/15.57/18.91	5.76/12.97/18.98	1.79/13.52/14.65	2.72/16.68/12.51
CoMFormer [54]	6.51/12.67/20.96	6.93/15.25/25.93	6.41/16.9/25.74	7.08/18.69/23.82	8.34/16.68/26.03
ECLIPSE [33]	10.07/22.76/31.62	8.74 /20.9/31.41	9.43/23.23/35.07	9.27 / 25.84 / 33.58	9.84/ 29.87 / 35.26
CoMBO [60]	10.16 / 23.06 / 32.11	8.07/ 21.26 / 32.32	10.25 / 27.2 / 38.29	8.06/22.33/31.60	10.06 /18.22/28.65
CRISP	11.04 / 24.78 / 31.95	10.2 / 22.47 / 32.26	10.8 / 24.71 / 36.18	10.97 / 27.82 / 35.26	12.88 / 31.35 / 39.42

个基础类别，然后每步添加 5 个新类别（N_t =5），以及 10-10，学习 10 个基础类别，然后每步添加 10 个新类别（N_t =4）。

训练设置：我们为 CVIS 开发了基于 Mask2Former 框架且以 ResNet-50 为骨干网的 CRISP 方法。我们根据数据集和任务的特点调整训练参数。对于 YouTube-VIS-2019，初始任务以学习率 0.0001 和批量大小 16 训练 1500 次迭代。在增量学习期间，每个类别经过 150 次迭代微调。对于 YouTube-VIS-2021 的 20-5 任务，初始阶段使用相同的基本参数，而每个类别的增量迭代次数增加到 300 。对于 10-10 任务，初始迭代次数减少到 750 ，其他参数保持不变。

我们将我们的方法与 MiB [59] 、CoMFormer [54] 、ECLIPSE [33] 、CoMBO [60] 以及 Fine-tune (FT) 进行比较。FT 代表在没有任何持续学习方法的情况下训练模型，它展示了基本性能。正如表 I 和表 II 中总结的那样，我们在三个场景中实现了最先进的表现，并在一个场景中获得了第二名。

定性比较：图 4 展示了我们的方法与其他方法的定性分析结果。在上面的例子中，我们的方法准确地分割了人物和轿车，同时在整个视频序列中保持一致的实例 ID。在下面的例子中，它成功地检测到飞机，甚至识别出了一个在真实标注中缺失的额外实例，显示出了其对不完整标注的鲁棒性。图 5 展示了由我们的 CRISP 框架处理的一些具

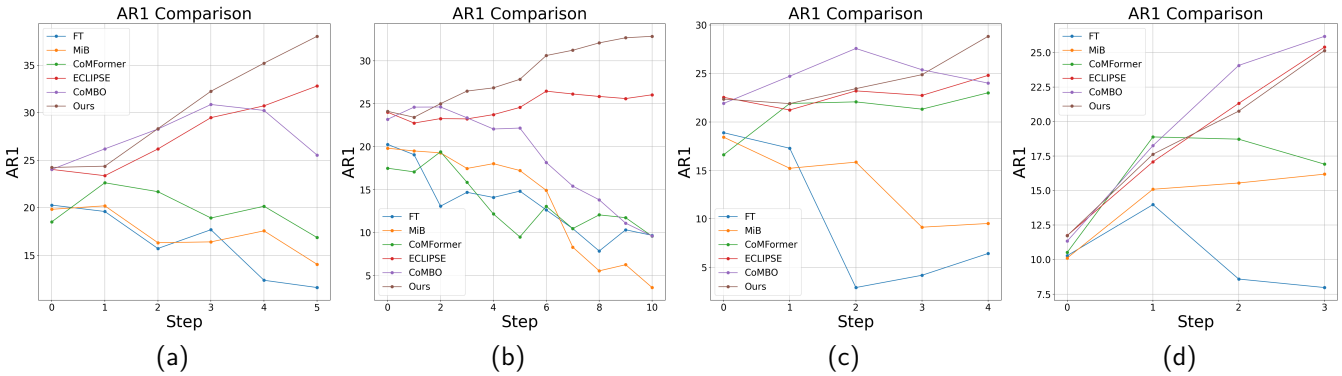

Fig. 6. 在 YouTube-VIS-2019 数据集上的 20-4、20-2 场景以及在 YouTube-VIS-2021 上的 20-5、10-10 场景中的每一步 AR1 结果。

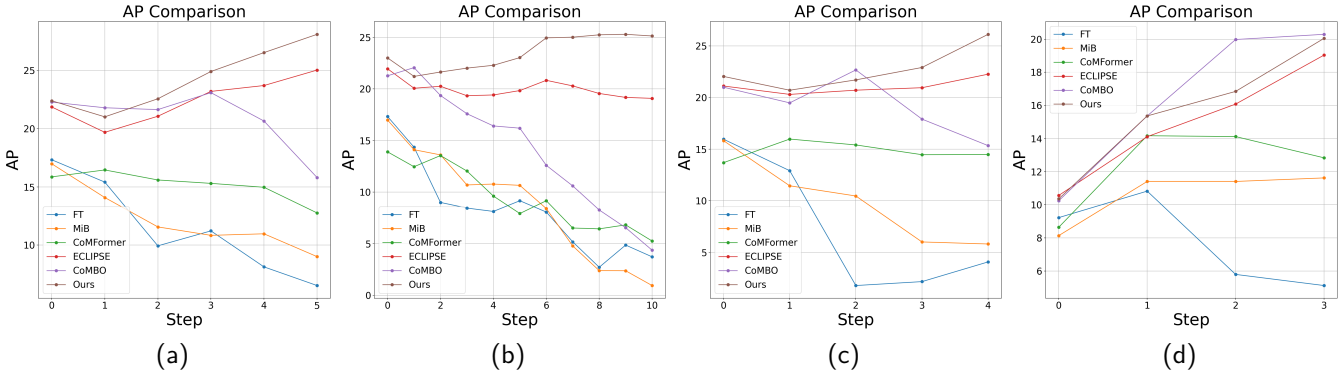

Fig. 7. 在 YouTube-VIS-2019 数据集上的 20-4、20-2 场景以及在 YouTube-VIS-2021 数据集上的 20-5、10-10 场景的每个步骤的 mAP 结果。

TABLE VI
10-10 结果在 YouTube-VIS-2021 数据集上。

Model	Step0	Step1	Step2	Step3
FT	5.96 10.28 13.67	6.10 10.32 16.46	3.6 6.31 11.26	3.39 8.32 6.67
MiB [59]	6.1 17.92 12.35	6.53 14.46 19.61	6.31 16.08 23.76	5.75 22.12 23.68
CoMFormer [54]	5.64 10.67 11.90	8.44 15.65 21.96	6.48 18.32 25.55	6.41 19.22 24.56
ECLIPSE [33]	6.59 12.18 16.12	6.04 16.54 23.00	6.03 20.09 30.86	7.10 25.37 35.47
CoMBO [60]	5.62 11.32 15.26	6.24 17.91 24.72	7.43 23.5 36.46	8.39 27.55 32.76
CRISP	6.67 12.02 15.78	8.3 18.86 24.57	8.99 22.6 30.66	9.6 28.43 34.2

有挑战性的案例的可视化结果，包括重叠的相似实例、动态场景，以及实例像素消失后再次出现的场景。在具有高度相似性和实例重叠的场景中，特征相似性往往导致分割中实例混淆。然而，CRISP 始终如一地为这些实例保留了准确的分割和分类。在动态场景中，实例形状和外观的持续变化可能会妨碍可靠的跟踪，但我们的方法在帧之间保持车的稳定跟踪。即使在实例部分消失然后再次出现的情况下，CRISP 也显示出强大的鲁棒性。值得注意的是，它不仅为视频中的实例预测出准确的掩码和类别，即使某些实例在真实标注中缺失，也在整个视频序列中提供了卓越的跟踪性能。

定量结果：表 III - VI 展示了任务层级的 APs, APm 和 APl，这些指标分别评估分割小型、中型和大型物体的能力。通过在实例层级和任务层级共同建模，我们的方法在整个增量学习过程中始终优于基于蒸馏的方法。在短期增量场景（10-10）中，我们观察到我们的方法在 APs 等指标上表现优于 CoMBO [60]，但在最终任务上的整体 mAP 上略有落后。我们假设这种差异源于标注的蒙版边界与真实的物体轮廓之间的不一致。

图 6 和图 7 展示了所有增量步骤中的 AR1 和 mAP 结果。在长期增量情境下，我们的方法达到了最先进的性能。相反，在短期增量设置下，CoMBO [60] 显示了更强的结果。我们将此归因于短期情境下数据量增加和任务间隔缩短，这促进了持续适应并减少了遗忘先前学习任务的风险。

B. 消融研究

如表 VII 所示，我们在 YouTube-VIS-2019 数据集的 20–4 划分下进行了消融研究，以量化每个提出组件的贡献。去除 PCA 引导的初始化（PI）会导致明显的性能下降，验证了其通过更好地对齐查询初始化来缓解任务混淆的作用。仅使用实例相关性（IC）损失能改善与跟踪相关的指标，但当与 PI 结合时，其益处明显放大（第 4–5 列与第

TABLE VII
对所提组件的消融研究。ARSP: 自适应残差语义提示，ISC: 实例语义一致性，PI: PCA 引导初始化，IC: 实例相关性损失。

ARSP	ISC	PI	IC	20-4 on YouTube-VIS-2019 (6 steps) mAP	AP50	AP75	AR1	AR10	FR
				25.03	39.83	27.59	32.81	40.48	2.23
		✓		26.3	40.81	29.15	34.71	41.99	2.11
			✓	25.69	40.43	28.35	33.40	39.47	2.05
✓				26.42	40.86	29.48	33.37	39.64	2.06
✓		✓	✓	27.64	42.69	31.24	34.84	42.18	2.05
✓	✓			27.83	42.7	30.97	34.92	41.62	2.04
✓	✓	✓		27.5	42.41	30.30	36.68	43.35	2.27
✓	✓	✓	✓	28.1	43.3	31.98	38.04	44.71	1.93

TABLE VIII
关于提出组件的逐任务消融研究。ARSP：自适应残差语义提示，ISC：实例语义一致性，PI：PCA 引导初始化，IC：实例关联损失。

ARSP	ISC	PI	IC	mAP results of task-wise (6 steps) Step0	Step1	Step2	Step3	Step4	Step5
				21.86	19.69	21.07	23.2	23.71	25.03
		✓		21.86	20.08	21.68	23.51	24.65	26.3
			✓	21.86	19.74	21.1	23.45	23.96	25.69
✓				22.21	20.4	21.78	24.53	25.11	26.42
✓		✓	✓	22.21	20.37	21.81	24.53	25.87	27.65
✓	✓			22.4	20.82	22.47	25.71	26.55	27.83
✓	✓	✓		22.4	21.04	22.47	24.5	25.88	27.5
✓	✓	✓	✓	22.4	21.02	22.57	24.91	26.53	28.1

6–9 列），突出显示任务间相关性保护与实例级别识别间的协同作用。自适应残余语义提示（ARSP）通过自适应查询提示对齐进一步提高分割精度，而实例语义一致性（ISC）损失增强了类内紧凑性和类间可分性，在各项指标上实现了一致性提升。值得注意的是，整合了 ARSP、ISC、PI 和 IC 的完整 CRISP 模型实现了最高的整体 mAP（28.1）和最佳的每类性能，展示了所有组件的互补性。

至关重要的是，图 VIII 从任务角度加强了这些发现。当通过每个学习步骤的 mAP 进行评估时，PI 在新任务中始终缓解性能退化，IC 损失稳定了跨时间环境的实例跟踪，PI 和 IC 的联合应用保留了新任务的适应性和旧任务的保持能力。这种方案在整体和任务的双重确认中，提供了所提出模块的稳健性和普适性具有强有力的证据。

V. 结论

在这项工作中，我们提出了 CRISP，这是一个对持续视频实例分割任务的早期尝试。对于类别方面，我们通过 ARSP 向模型注入语义信息以指导模型在训练过程中学习语义信号。同时，通过一个实例语义一致性损失，模型不断加强查询空间和语义空间之间的对齐。对于任务方面，我们采用 PCA 引导初始化，在进行初始化过程时保留任务之间的内在连接。对于实例方面，我们引入了实例相关性损失，它强调与先前查询空间的相关性，同时增强当前任务查询的特异性。在 YouTube-VIS-2019 和 YouTube-VIS-2021 数据集上进行的广泛实验表明，CRISP 在长期的持续视频实例分割任务上达到了最先进的性能，有效避免了灾难性遗忘，并展示了其持续学习能力。目前，我们的方法仅在持续视频实例分割任务上进行了测试。未来的工作将探索更多的持续视频分割场景，例如持续视频全景分割。

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[2] N. Guo, K. Gu, J. Qiao, and J. Bi, "Improved deep cnns based on nonlinear hybrid attention module for image classification," *Neural Networks*, vol. 140, pp. 158–166, 2021.

[3] N. Rodríguez-Barroso, E. Martínez-Cámara, M. V. Luzón, and F. Herrera, "Backdoor attacks-resilient aggregation based on robust filtering of outliers in federated learning for image classification," *Knowledge-Based Systems*, vol. 245, p. 108588, 2022.

[4] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1277–1286.

[5] L. Lin, P. Yan, X. Xu, S. Yang, K. Zeng, and G. Li, "Structured attention network for referring image segmentation," *IEEE Transactions on Multimedia*, vol. 24, pp. 1922–1932, 2021.

[6] X. Zhang, Z. Xiao, J. Ma, X. Wu, J. Zhao, S. Zhang, R. Li, Y. Pan, and J. Liu, "Adaptive dual-axis style-based recalibration network with class-wise statistics loss for imbalanced medical image classification," *IEEE Transactions on Image Processing*, vol. 34, pp. 2081–2096, 2025.

[7] Y. Rong, T. Lin, H. Chen, Z. Fan, and X. Chen, "Searching discriminative regions for convolutional neural networks in fundus image classification with genetic algorithms," *IEEE Transactions on Image Processing*, vol. 33, pp. 5949–5958, 2024.

[8] Q. Wang, Z. Huang, H. Fan, S. Fu, and Y. Tang, "Unsupervised person re-identification based on adaptive information supplementation and foreground enhancement," *IET Image Processing*, vol. 18, no. 14, pp. 4680–4694, 2024.

[9] W. Ren, J. Luo, W. Jiang, L. Qu, Z. Han, J. Tian, and H. Liu, "Learning self- and cross-triplet context clues for human-object interaction detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 10, pp. 9760–9773, 2024.

[10] Y. Wu, J. Li, Y. Yuan, A. K. Qin, Q.-G. Miao, and M.-G. Gong, "Commonality autoencoder: Learning common features for change detection from heterogeneous images," *IEEE transactions on neural networks and learning systems*, vol. 33, no. 9, pp. 4257–4270, 2021.

[11] W. Cong, Y. Cong, and Y. Ren, "Lightweight class incremental semantic segmentation without catastrophic forgetting," *IEEE Transactions on Image Processing*, pp. 1–1, 2025.

[12] Y. Gong, S. Yu, X. Wang, and J. Xiao, "Continual segmentation with disentangled objectness learning and class recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3848–3857.

[13] C.-B. Zhang, J.-W. Xiao, X. Liu, Y.-C. Chen, and M.-M. Cheng, "Representation compensation networks for continual semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7053–7064.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[15] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[16] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3150–3158.

[17] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2359–2367.

[18] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *European conference on computer vision*. Springer, 2020, pp. 649–665.

[19] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[20] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, "Mask2former for video instance segmentation," 2021. [Online]. Available: https://arxiv.org/abs/2112.10764

[21] R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[22] A. Robins, "Catastrophic forgetting, rehearsal and pseudorehearsal," *Connection Science*, vol. 7, no. 2, pp. 123–146, 1995.

[23] S. Thrun, "Lifelong learning algorithms," in *Learning to learn*. Springer, 1998, pp. 181–209.

[24] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[25] P. Singh, V. K. Verma, P. Mazumder, L. Carin, and P. Rai, "Calibrating cnns for lifelong learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 579–15 590, 2020.

[26] J. Dong, L. Wang, Z. Fang, G. Sun, S. Xu, X. Wang, and Q. Zhu, "Federated class-incremental learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[27] J. Dong, H. Li, Y. Cong, G. Sun, Y. Zhang, and L. Van Gool, "No one left behind: Real-world federated class-incremental learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2054–2070, 2024.

[28] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Task-specific normalization for continual learning of blind image quality models," *IEEE Transactions on Image Processing*, vol. 33, pp. 1898–1910, 2024.

[29] K. Thandiackal, L. Piccinelli, R. Gupta, P. Pati, and O. Goksel, "Multi-scale feature alignment for continual learning of unlabeled domains," *IEEE Transactions on Medical Imaging*, vol. 43, no. 7, pp. 2599–2609, 2024.

[30] S. Yan, J. Xie, and X. He, "Der: Dynamically expandable representation for class incremental learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3014–3023.

[31] F. Cermelli, M. Mancini, S. R. Bulo, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9233–9242.

[32] A. Douillard, Y. Chen, A. Dapogny, and M. Cord, "Plop: Learning without forgetting for continual semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4040–4050.

[33] B. Kim, J. Yu, and S. J. Hwang, "Eclipse: Efficient continual learning in panoptic segmentation with visual prompt tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3346–3356.

[34] H. Chen, Z. Wu, X. Han, M. Jia, and Y.-G. Jiang, "Promptfusion: Decoupling stability and plasticity for continual learning," in *European Conference on Computer Vision*. Springer, 2024, pp. 196–212.

[35] J. S. Smith, L. Karlinsky, V. Gutta, P. Cascante-Bonilla, D. Kim, A. Arbelle, R. Panda, R. Feris, and Z. Kira, "Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 909–11 919.

[36] R. Wang, X. Duan, G. Kang, J. Liu, S. Lin, S. Xu, J. Lü, and B. Zhang, "Attriclip: A non-incremental learner for incremental knowledge learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3654–3663.

[37] Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy *et al.*, "Dualprompt: Complementary prompting for rehearsal-free continual learning," in *European conference on computer vision*. Springer, 2022, pp. 631–648.

[38] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 139–149.

[39] M. Menabue, E. Frascaroli, M. Boschini, E. Sangineto, L. Bonicelli, A. Porrello, and S. Calderara, "Semantic residual prompts for continual learning," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–18.

[40] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[41] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[42] I. Jolliffe, "Principal component analysis," in *International encyclopedia of statistical science*. Springer, 2011, pp. 1094–1096.

[43] L. Yang, Y. Fan, and N. Xu, "Video instance segmentation," in *International Journal of Computer Vision,*, 2019.

[44] L. Yang, Y. Fan, Y. Fu, and N. Xu, "The 3rd large-scale video object segmentation challenge - video instance segmentation track," Jun. 2021.

[45] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[46] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[47] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8741–8750.

[48] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, "Video instance segmentation using inter-frame communication transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 352–13 363, 2021.

[49] Y. Gu, C. Deng, and K. Wei, "Class-incremental instance segmentation via multi-teacher networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1478–1486.

[50] H. Zhao, F. Yang, X. Fu, and X. Li, "Rbc: Rectifying the biased context in continual semantic segmentation," in *European Conference on Computer Vision*. Springer, 2022, pp. 55–72.

[51] S. Cha, Y. Yoo, T. Moon *et al.*, "Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning," *Advances in neural information processing systems*, vol. 34, pp. 10 919–10 930, 2021.

[52] M. H. Phan, S. L. Phung, L. Tran-Thanh, A. Bouzerdoum *et al.*, "Class similarity weighted knowledge distillation for continual semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 866–16 875.

[53] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[54] F. Cermelli, M. Cord, and A. Douillard, "Comformer: Continual learning in semantic and panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3010–3020.

[55] J. Chen, R. Cong, Y. Luo, H. H. S. Ip, and S. Kwong, "Strike a balance in continual panoptic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 126–142.

[56] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[58] S. Lee, J. Seo, K. Han, M. Choi, and S. Im, "Context-aware video instance segmentation," *arXiv preprint arXiv:2407.03010*, 2024.

[59] F. Cermelli, M. Mancini, S. Rota Bulò, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9230–9239.

[60] K. Fang, A. Zhang, G. Gao, J. Jiao, C. H. Liu, and Y. Wei, "Combo: Conflict mitigation via branched optimization for class

incremental segmentation," *arXiv preprint arXiv:2504.04156*, 2025.