研究生学习工作周报

院　　系	人工智能学院
专　　业	电子信息
姓　　名	佘依函
学　　号	231226006052
导　　师	周静 张俊驰
周报日期	2023 年 10 月 28 日

摘要

1. Two-shot Video Object Segmentation 论文阅读。

目录

第一章 学习工作总结

1.1 Two-shot Video Object Segmentation 论文阅读

本周对 Two-shot Video Object Segmentation 论文阅读进行阅读，通过阅读 Tco-shot VOS 的网络架构部分了解其可以应用于大部分的 VOS 项目中同时其使用了 Youtube-VOS 的数据集，和 R2VOS 使用的数据集相同，故可以尝试将此训练方法应用于 R2VOS 中。附论文的阅读笔记

Two-shot Video Object Segmentation

Meta Data

Title	Two-shot Video Object Segmentation
Journal	
Authors	Kun Yan; Xiao Li; Fangyun Wei; Jinglu Wang; Chenbin Zhang; Ping Wang; Yan Lu
Pub. date	2023-03-21
期刊标签	
DOI	10.48550/arXiv.2303.12078
附件	Yan et al_2023_Two-shot Video Object Segmentation.pdf

研究背景 & 基础 & 解决方法

研究背景

密集注释的视频上训练昂贵耗时

基础

基本思想是在训练期间为未标记的帧生成伪标签，并结合标记数据和伪标记数据来优化模型。

解决方法

1.以半监督的方式在稀疏注释的视频上预训练 VOS 模型，第一帧始终是有标签的帧。

2.采用预训练的 VOS 模型为所有未标记的帧生成伪标签，随后将其存储在伪标签库中。

3.标记数据和伪标记数据上重新训练 VOS 模型，对第一帧没有任何限制。

“Previous Methods (Densely Annotated Videos)” (Yan 等, 2023, p. 1)

研究内容

其基本思想是在训练过程中为未标记的帧生成可信的伪标签，并结合标记数据和伪标记数据来优化模型。具体来说，STCN 将随机选择的三元组标记帧作为输入，但监督仅应用于最后两个 - VOS 需要第一帧的注释作为参考来分割后续帧中出现的感兴趣的目标。

“Our contributions can be summarized as follows” (Yan 等, 2023, p. 2)

1.首次证明了 two-shot VOS 的可行性：不使用未标记的数据，每个视频的两个标记帧也可以训练出 VOS 模型。

2.提出了一种简单而有效的训练范例，以利用未标记帧中存在的丰富信息。

3.和 full 的评分相差不多，具有竞争性。

研究过程

“Video object segmentation.” (Yan 等, 2023, p. 2)

online：需要实时对网络进行调整

offline：无需调整，需要通过 mask 来分割目标。基于匹配的方法通常采用存储体来存储帧集合的特征，然后采用特征匹配来分割查询帧。

“Semi-supervised learning” (Yan 等, 2023, p. 3)

半监督学习是利用少量标记数据和大量未标记数据来提高模型性能的有效方法。

一致性：基于一致性的方法强制不同扰动的预测之间的一致性，例如模型扰动、数据增强和对抗性扰动。

伪标签：基于伪标签的方法为未标记的数据生成 onehot 伪标签。然后结合标记数据和伪标记数据对模型进行优化。

方法

“Preliminary” (Yan 等, 2023, p. 3)

一般来说，随着训练的进行，要跳过的最大帧数逐渐从 0 增加到 K。

在本文设定的情况下，只能访问两个标记帧。为了减少由不可靠的伪标签引起的错误传播，我们在第一阶段训练中采用 STCN 作为我们的基础模型，因为它只需要三组帧作为输入。

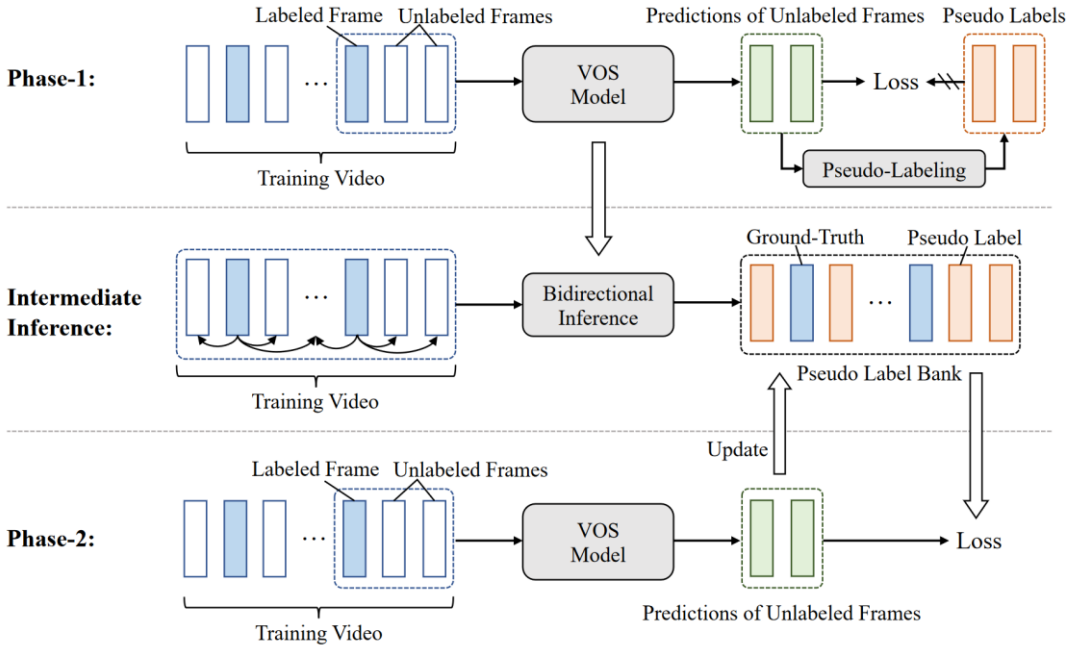
“STCN” (Yan 等, 2023, p. 3)

给定一个训练视频，STCN 首先对三组帧进行采样作为输入。

然后，它根据第一帧的 ground-truth 预测第二帧的 mask，并根据前一帧的预测以及第一帧的 ground-truth 预测第三帧的 mask。

STCN 的目标函数是标准分割损失，应用于两个预测中的每一个。

“Problem formulation and overview” (Yan 等, 2023, p. 3)



“Phase-1 training” (Yan 等, 2023, p. 3)

基础模型：

STCN

接受三帧输入

两个标记帧 + 一个伪标记帧

最后两帧可以没有标记，在实现中，最后两帧有 0.5 的概率均未标记，并且有 0.5 的概率有一帧被标记。双镜头 VOS 的训练与全集 VOS 相同，只是训练三元组由带有真

实标签的标记帧和带有伪标签的未标记帧组成。

具体来说，给定一个随机采样的三元组，其中最后两帧由 N1 个标记帧和 N2 个未标记帧组成（N1 = 1，N2 = 1 或 N1 = 0，N2 = 2），总体损失 L 是　监督损失 LS 和无监督损失 LU 的总和，分别影响标记和未标记的帧。 LS 是标准分割损失，可以表示为：

$$\mathcal{L}_S = \frac{1}{HWN_1} \sum_{n=1}^{N_1} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{H}(\boldsymbol{Y}_n^{(i,j)}, \boldsymbol{P}_n^{(i,j)})$$

H：高度

W：宽度

$ \mathcal{H}$：交叉熵函数

$\boldsymbol{P}_n^{(i, j)}$：第 n 个标记帧中像素 (i, j) 的预测

$\boldsymbol{Y}_n^{(i, j)}$：表示相应的真实值。

$$\mathcal{L}_U = \frac{1}{HWN_2} \sum_{n=1}^{N_2} \sum_{i=1}^{H}$$

1：指示函数，用于过滤掉最大置信度低于预定义阈值 τ1 的预测。

$\hat{\boldsymbol{Y}}_n^{(i, j)}$ ： = argmax(P (i,j) n) 表示对应的 one-shot 伪标签。

“τ1 = 0.9” (Yan 等, 2023, p. 4)

“Phase-2 training” (Yan 等, 2023, p. 5)

为了充分利用未标记数据，在第二阶段训练解除了对参考帧的限制，允许其为标记帧或者伪标记帧。

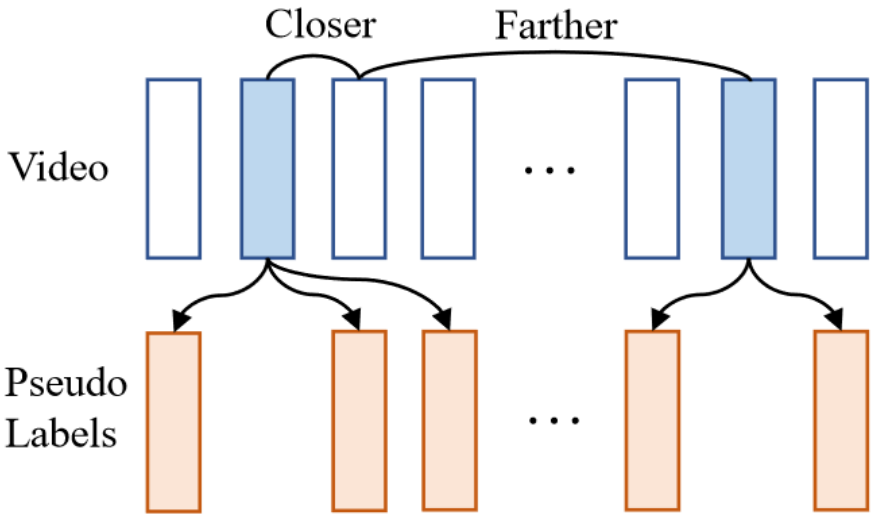
设置基础：第一阶段所训练的 VOS 模型为所有未标记帧生成的伪标签。

伪标记数据被存储在伪标签库中，在构建训练三元组时提供有效的访问，其中参考

帧被选择作为伪标记帧。

“Intermediate inference and pseudo-label bank.” (Yan 等, 2023, p. 5)

在开始第二阶段训练之前执行中间推理。



建立一个伪标签库来存储与未标记帧相关的伪标签

“Training.” (Yan 等, 2023, p. 5)

第 2 阶段的训练过程与第 1 阶段的训练过程相同，只是参考（或第一个）帧可以是带标签的帧，也可以是带有来自伪标签库的伪标签的未标签帧。

“Update pseudo-label bank.” (Yan 等, 2023, p. 5)

一旦预测 P (i,j) 满足 max(P (i,j)) ≥ τ2 的条件，其中 τ2 表示预先定义的阈值，伪标签库中相应的伪标签将更新为 ˆ Y (i, j) = argmax(P (i,j))。我们默认设置 τ2 = 0.99。

“Experiments” (Yan 等, 2023, p. 5)

“Datasets.” (Yan 等, 2023, p. 5)

DAVIS 2016/2017 和 YouTube-VOS 2018/2019

在 two-shot 设置中，我们随机选择每个视频的两个标记帧作为标记数据，而其余帧作为未标记数据。

仅使用 YouTube-VOS 和 DAVIS 的 7.3% 和 2.9% 标记数据。

“Evaluation metric.” (Yan 等, 2023, p. 5)

对于 DAVIS 数据集采用标准指标：区域相似度 J 、轮廓精度 F 及其平均 J &F。

对于 YouTube-VOS 数据集，报告了已见和未见类别的 J 和 F，以及它们的平均得分 G。

“Implementation details” (Yan 等, 2023, p. 5)

PyTorch

Phase-1-training:

采用在具有合成变形的静态图像数据集上进行预训练的 STCN

随机跳帧中的参数 K 随着课程学习进度从 5 逐渐增加到 25。

阈值 τ1 设置为 0.9。

分别探索 STCN、RDE-VOS 和 XMem。阈值 τ2 设置为 0.99。

“Main results” (Yan 等, 2023, p. 6)

进行比较：

(1)他对应版本在全套上训练；

(2)在两次数据集上进行训练，而不使用未标记的数据；

(3)其他经过全套训练的强基线。

结论：

（1）每个视频两个标记帧几乎足以训练一个令人满意的 VOS 模型 - 即使未使用未标记的数据。2-shot STCN 在 YouTube-VOS 2018 基准测试中已经达到 80.8% 的分数，仅比全套 STCN 的 83.0% 分数低 2.2%。

(2) 通过使用 YouTube-VOS 和 DAVIS 基准的 7.3% 和 2.9% 标记数据，我们的方法取得了与全套训练的对应方法相当的结果，并且大大优于原生 2-shot 对应方法。

2-shot STCN 在 DAVIS 2017/YouTubeVOS 2019 上达到 85.1%/82.7%，比简单的 2-shot STCN 高 +4.1%/+2.1%，同时低 -0.1%/-0.0%比全套 STCN。

“Ablation study” (Yan 等, 2023, p. 7)

所有消融研究均通过应用我们的 STCN 方法在 Youtube-VOS 2019 上进行。

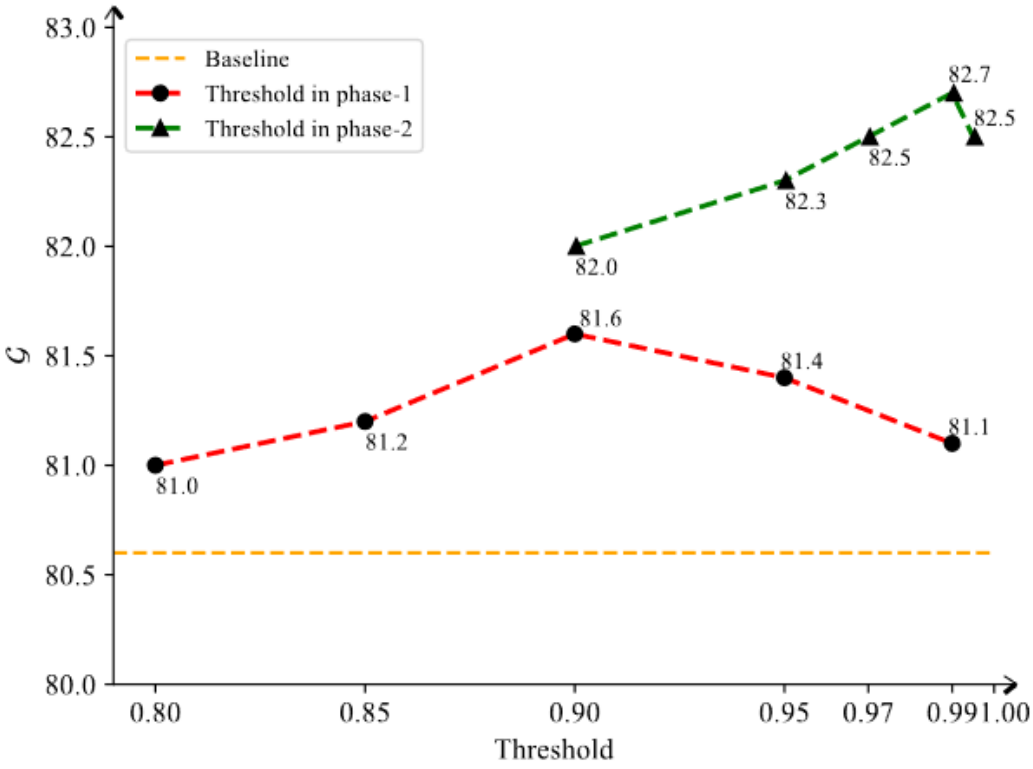
“Effects of each phase.” (Yan 等, 2023, p. 7)

Components	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Baseline	80.6	79.5	83.8	75.7	83.4
+phase-1	$81.6_{+1.0}$	79.3	83.5	77.7	86.0
+phase-2	$82.7_{+1.1}$	80.9	85.1	78.3	86.6

“Thresholds of pseudo-labeling.” (Yan 等, 2023, p. 7)

有两个超参数 τ1 和 τ2 分别控制阶段 1 和 2 中的伪标记。

通过改变 τ1 和 τ2 显示了两条精度曲线。使用较高的阈值可以保证生成的伪标签的质量，但产生的伪数据量较少，反之亦然。我们在第 2 阶段训练中采用更高的阈值，因为第 2 阶段的预测比第 1 阶段的预测更准确。



“Bidirectional inference.” (Yan 等, 2023, p. 7)

原因：（1）一些无标签帧在单向推理中与伪标签没有关联； (2)双向推理减轻了错误传播问题。

“Visualization of feature space.” (Yan 等, 2023, p. 8)

配备我们方法的 2-shot STCN 和全套 STCN 都显示出更紧凑的集群。

图 1.1 论文阅读笔记

第二章 第九周总结及第十周学习计划

序号	上周任务	完成情况及备注	本周任务
1	Two-shot Video Object Segmentation	仔细阅读了框架和部署实施部分	复现 Two-shot
2	R2VOS	本周代码阅读进度缓慢	将 Two-shot 应用于 R2VOS
3			SpVOS: Efficient Video Object Segmentation with Triple Sparse Convolution
备注：			

参考

1. https://doi.org/10.48550/arXiv.2303.12078