

基于辅助知识的图像/视频分割算法研究

--四建楼 博士报告

--2021.11.30

0、个人信息简介

1、研究背景及主要内容

2、借助像素间判别信息的实时视频目标分割算法

3、借助帧间时空信息的快速视频语义分割算法

4、借助辅助数据集的弱样本图像语义分割算法

5、在站期间其他工作概述

基本信息

博士后姓名	出生日期	性别	民族
四建楼	1989.12.27	男	汉

教育背景

学校	专业	时间	学历
北京邮电大学	信息与通信工程	2012.09-2018.06	博士研究生
吉林大学	通信工程	2008.09-2012.06	本科生

工作背景

公司或机构	职位	时间
北京市商汤科技开发有限公司	高级研究员	2018.06-至今
中科院自动化所&北京商汤	博士后	2019.01-至今
南洋理工大学	助理研究员	2016.11-2017.11

研究背景

图像/视频分割

分割精度：类别错误、细节精度等
分割时空一致性：边界与区域一致、帧间一致等

全监督分割

1. 像素级标注难度大，过渡区域真值定义模糊
2. 已有数据集规模有限，不同数据集有可能存在标签冲突
3. 待标注数据数量巨大，像素级标注经济及人力成本高昂

主要内容

有限标注数据，挖掘辅助知识，优化分割任务(同时尽量遵循低延迟原则)

原始知识

语义分割：像素语义类别
目标分割：模板像素语义类别
实例分割：物体像素实例信息及语义类别
全景分割：物体像素实例及语义类别、事物像素语义类别

分割数据集

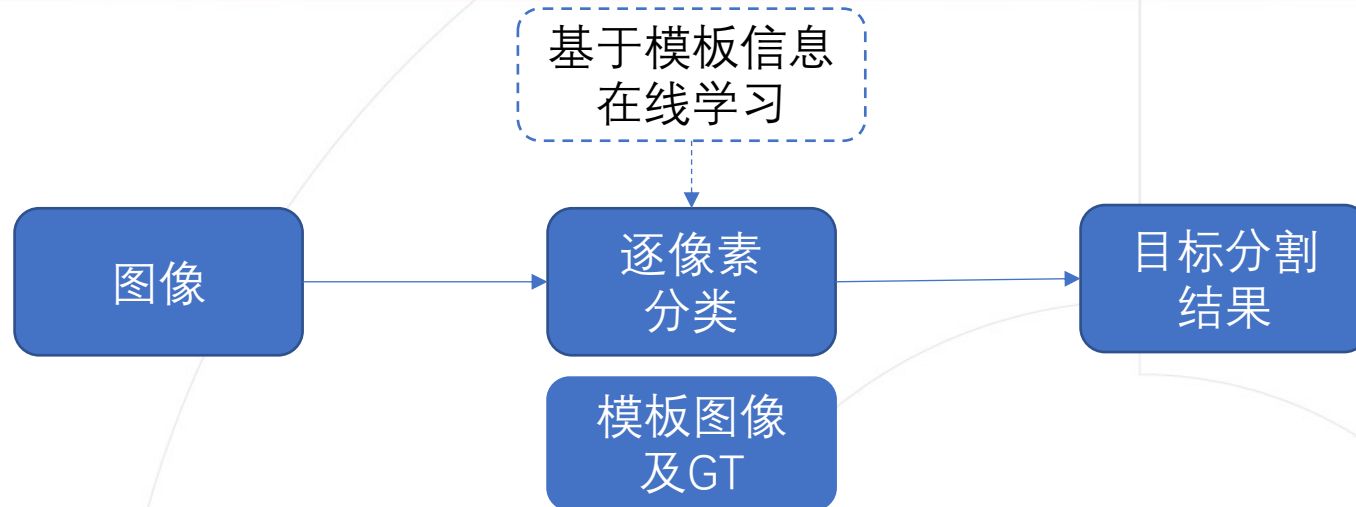
辅助知识

像素间：像素间语义相关性、位置相关性等
图像/样本间：图像间相似关系、时域关系等
数据集/任务间：无监督/半监督/弱监督、零样本/少样本、跨数据集/多任务等

借助像素间判别信息的实时视频目标分割算法

背景及动机

基于分类器学习VOS



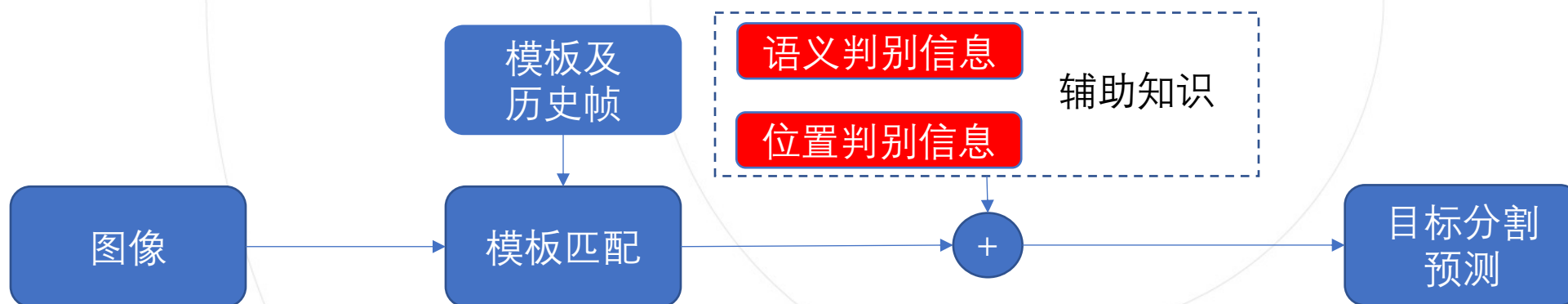
- 1. 基于目标分类
- 2. 分割精度高
- 3. 高耗时

基于模板匹配学习VOS



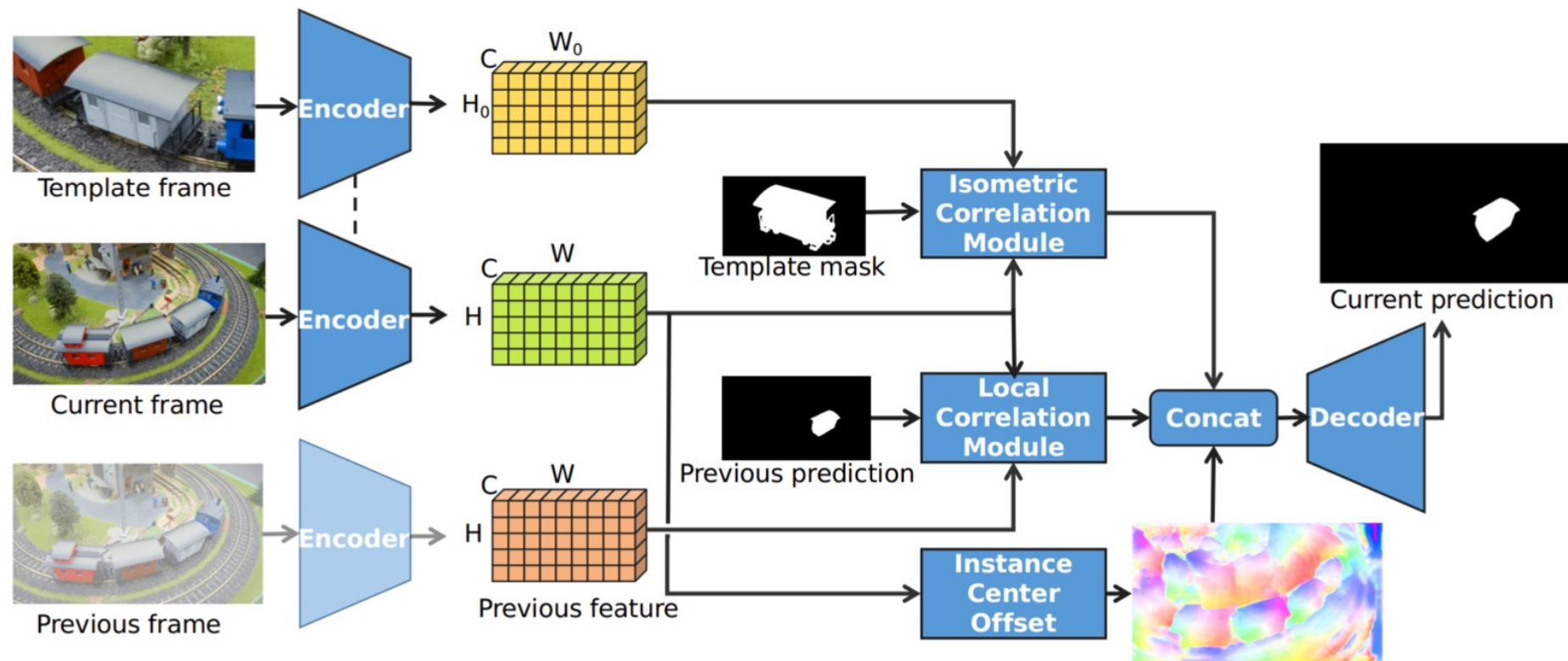
- 1. 基于目标匹配
- 2. 丢失原图判别信息
- 3. 低耗时

我们方法



算法介绍

框架及流程



- 1) 基本结构模板匹配：低延迟
- 2) 引入等距相关性学习模块 (Isometric Correlation Module)：增强模板匹配相似度图的类别辨识度
- 3) 引入局部相关性学习模块 (Local Correlation Module)：挖掘帧间匹配相似度关系
- 4) 引入实例中心偏移预测模块 (Instance Center Offset)：增强特征的抗背景目标干扰能力

1. 等距相关性计算模块

1. 假设图像特征为 \mathbf{F} ，匹配相似度图为 \mathbf{Z} ；我们的目标是在相似度图空间中引入特征空间的像素间距离关系，从而保留语义判别信息

2. 易证：只需对模板特征 \mathbf{F}_t 进行正交化即可；

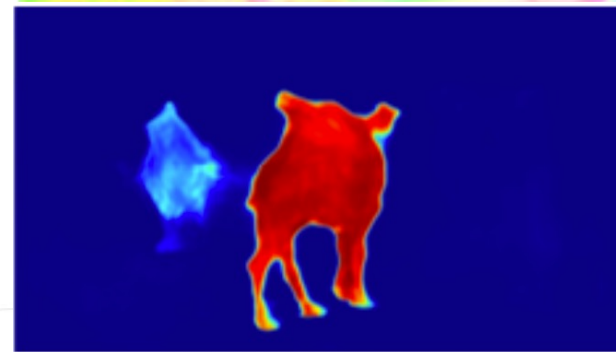
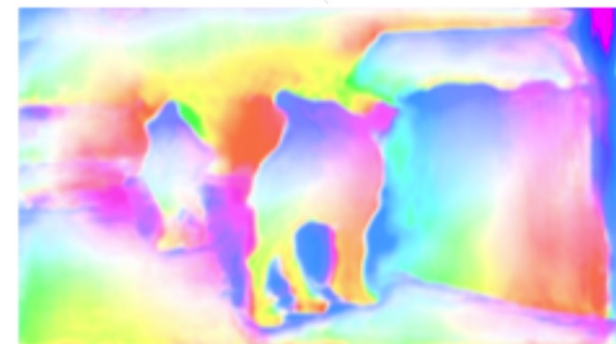
$$\begin{aligned} \cos(\mathbf{Z}^m, \mathbf{Z}^n) &= \cos(\mathbf{Y}\mathbf{F}_t^m, \mathbf{Y}\mathbf{F}_t^n) \\ &= \frac{(\mathbf{F}_t^m)^T \mathbf{Y}^T \mathbf{Y} \mathbf{F}_t^n}{\sqrt{(\mathbf{F}_t^m)^T \mathbf{Y}^T \mathbf{Y} \mathbf{F}_t^m} \sqrt{(\mathbf{F}_t^n)^T \mathbf{Y}^T \mathbf{Y} \mathbf{F}_t^n}} \\ &= \frac{(\mathbf{F}_t^m)^T \mathbf{F}_t^n}{\sqrt{(\mathbf{F}_t^m)^T \mathbf{F}_t^m} \sqrt{(\mathbf{F}_t^n)^T \mathbf{F}_t^n}} \\ &= \cos(\mathbf{F}_t^m, \mathbf{F}_t^n), \end{aligned}$$

3. 使用SVB实现正交化；每段视频只需对模板进行

2. 位置中心偏移预测

1. 基于图像特征，预测实例中心及逐像素的位置偏移，从而保留位置判别信息

2. 推断阶段，假设实例运动缓慢；可用上一帧中心及当前偏移对像素聚类

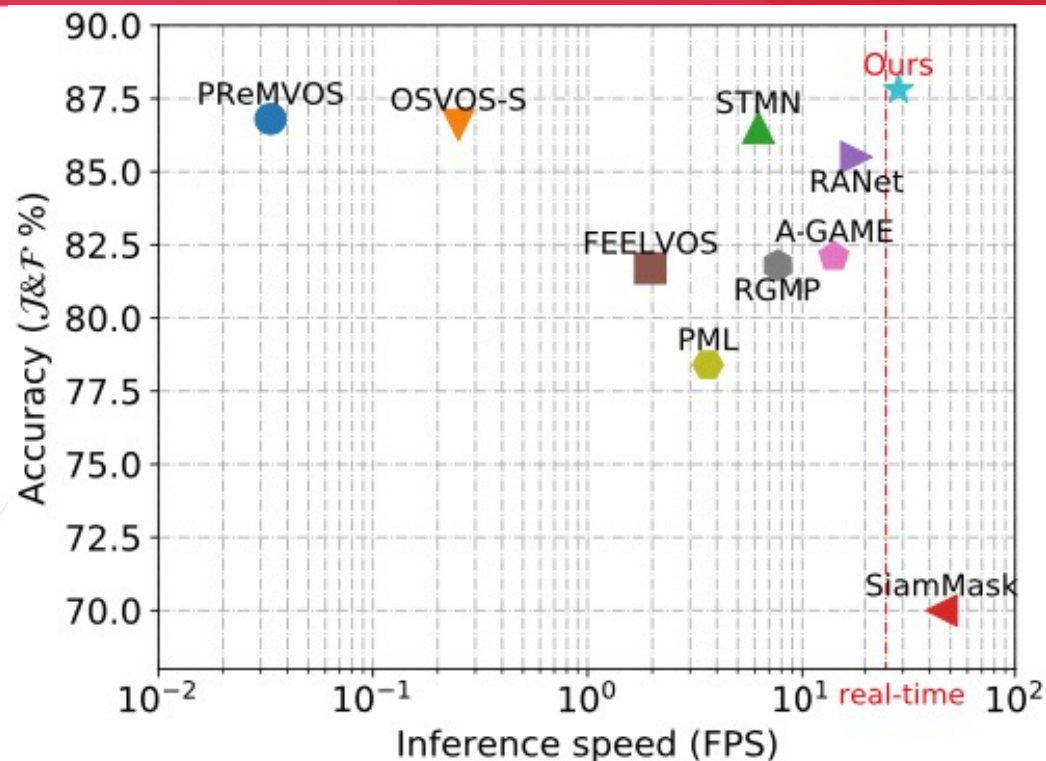


借助像素间判别信息的实时视频目标分割算法

实验结果

	OL	$J & F$	J_{mean}	F_{mean}	speed
MaskTrack ^[123]	✓	77.6	79.7	75.4	-
OSVOS ^[65]	✓	80.2	79.8	80.6	4.0s
CINM ^[64]	✓	84.2	83.4	85.0	-
DyeNet ^[113]	✓	-	86.2	-	2.33s
OnAVOS ^[67]	✓	85.5	86.1	84.9	-
OSVOS-S ^[66]	✓	86.6	85.6	87.5	4.0s
PreMVOS ^[115]	✓	86.8	84.9	88.6	>30s
RANet+ ^[95]	✓	87.1	86.6	87.6	4.0s
PLM ^[106]		66.4	70.2	62.5	280ms
VPN ^[129]		67.9	70.2	65.5	630ms
SiamMask ^[107]		69.8	71.7	67.8	22ms
CTN ^[118]		71.4	73.5	69.3	30ms
PML ^[68]		78.4	77.4	79.3	275ms
VideoMatch ^[69]		-	81.0	-	320ms
A-GAME ^[109]		81.5	82.2	81.9	140ms
FEELVOS ^[71]		81.7	81.1	82.2	510ms
RGMP ^[116]		81.8	81.5	82.0	130ms
STCNN ^[120]		83.8	83.8	83.8	3.9s
DTN ^[130]		83.6	83.7	83.5	70ms
CRN ^[108]		85.1	84.4	85.7	730ms
RANet ^[95]		85.5	85.4	85.5	55ms
STM ^[70]		86.5	84.8	88.1	160ms
Ours(baseline)		84.1	85.1	83.0	26ms
Ours		87.8	87.5	88.0	35ms

单目标VOS
数据集
DAVIS-
2016上的
对比实验及
消融实验



消融实验

	LC	CO	SVB	$J & F$	J_{mean}	F_{mean}	speed
1				84.1	85.1	83.0	26ms
2	✓			85.0	85.5	84.6	28ms
3	✓	✓		86.2	86.3	86.0	35ms
4	✓		✓	86.5	87.0	86.1	28ms
5	✓	✓	✓	87.8	87.5	88.0	35ms

借助像素间判别信息的实时视频目标分割算法

实验结果

DAVIS-2016
上可视化对比



实验结果

多目标VOS数据集DAVIS-2017上的对比实验

	OL	$\mathcal{J} \& \mathcal{F}$	\mathcal{J}_{mean}	\mathcal{F}_{mean}
OSVOS ^[65]	✓	60.3	56.6	63.9
OnAVOS ^[67]	✓	67.9	64.5	71.2
OSVOS-S ^[66]	✓	68.0	64.7	71.3
CINM ^[64]	✓	70.7	67.2	74.0
DyeNet ^[113]	✓	74.1	-	-
PRemVOS ^[115]	✓	77.8	73.9	81.7
SiamMask ^[107]		56.4	54.3	58.5
STCNN ^[120]		61.7	58.7	64.6
VideoMatch ^[69]		62.4	56.5	68.2
RANet ^[95]		65.7	63.2	68.2
AGSS-VOS ^[131]		66.6	63.4	69.8
RGMP ^[116]		66.7	64.8	68.6
FEELVOS ^[71]		69.1	65.9	72.3
STM ^[70]		71.6	69.2	74.0
Ours		72.9	70.9	74.9

背景及动机

基于图像训练的VSS

图像训练

逐帧预测

- 1. 模型易轻量化处理
- 2. 丢失时域一致性信息

借助视频训练的VSS

视频帧序列

光流提取

序列建模

视频分割预测

- 1. 光流或序列建模, 引入高额计算或误差
- 2. 利用时域一致性信息

我们的方法

视频帧序列

稀疏时域
Transformer

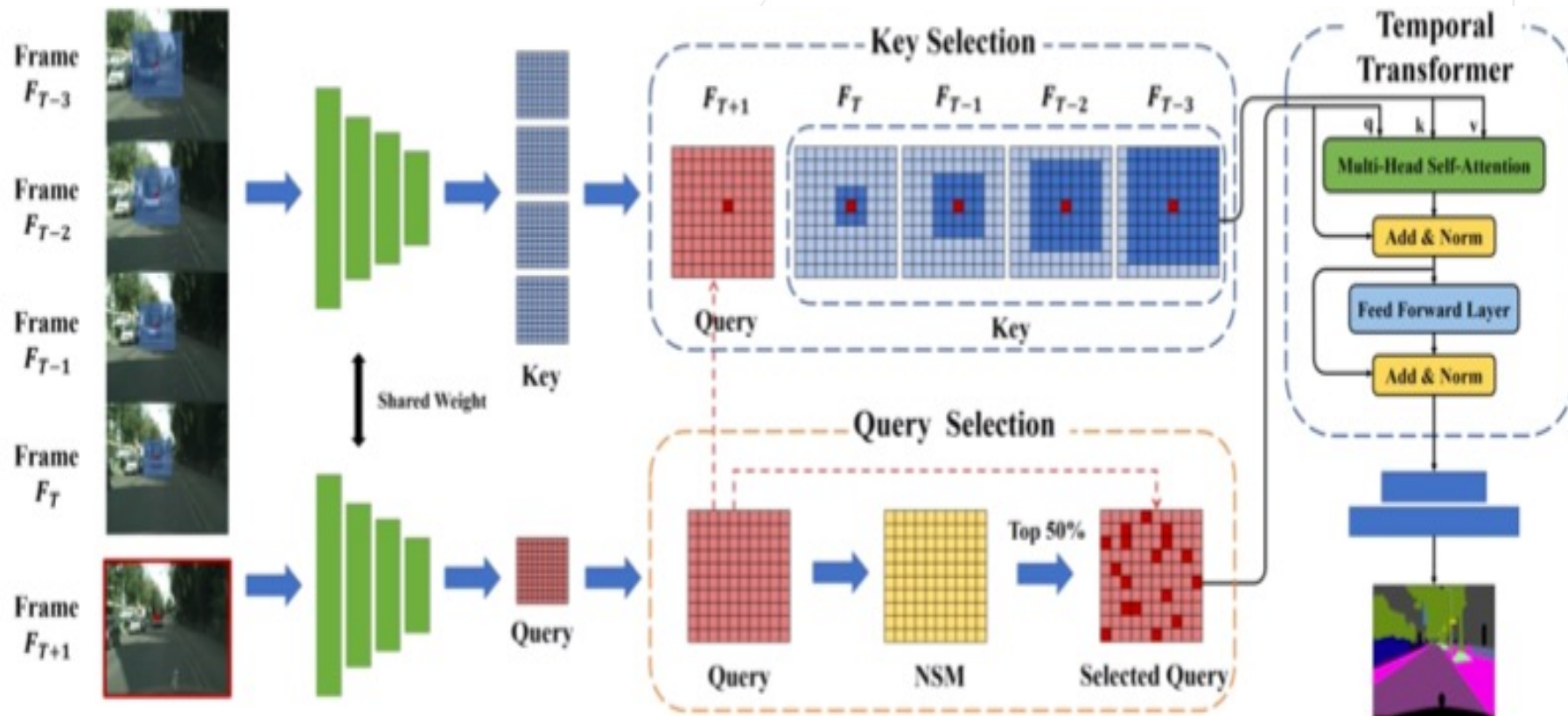
视频分割预测

- 1. 低延迟
- 2. 半监督

高效地帧间辅助知识挖掘

算法介绍

框架及流程



- 1) 利用时域Transformer提取帧间关系
- 2) 引入Query特征选择筛选出难分区域，同时降低计算量
- 3) 引入Key特征选择聚焦高相关区域，同时降低计算量

算法介绍

细节说明

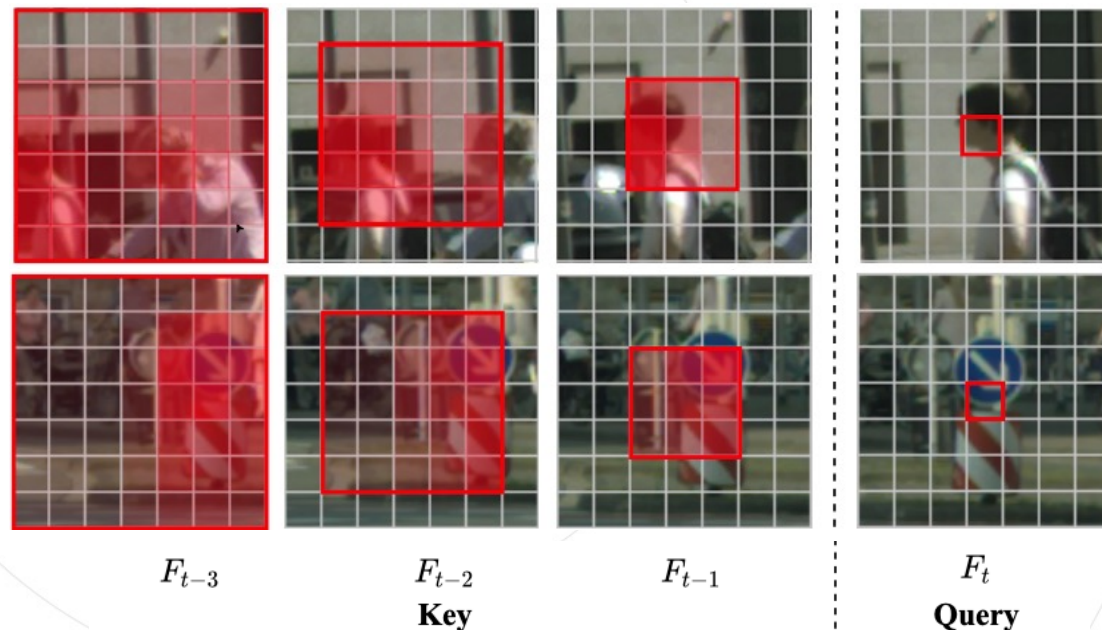
1. Query特征选择

1. 提出**NSM指标**，评估Query每个像素的复杂程度
 - 相似度与均匀分布的KL散度：中心区域附近是否复杂
 - 余弦相似度：中心区域与邻域是否相似
2. Query特征选择过程
 - 计算Query区域的逐像素NSM指标
 - 只保留Top 50%区域



2. Key特征选择

1. Key特征选择基本原则：
 - 距离当前帧越远，Key特征搜索半径越大
2. Key特征选择过程：
 - 根据与当前帧的时域距离，确定Key搜索半径
 - 根据Query中心位置，搜集所有Key特征



实验结果

Cityscapes和Camvid上高精度算法对比

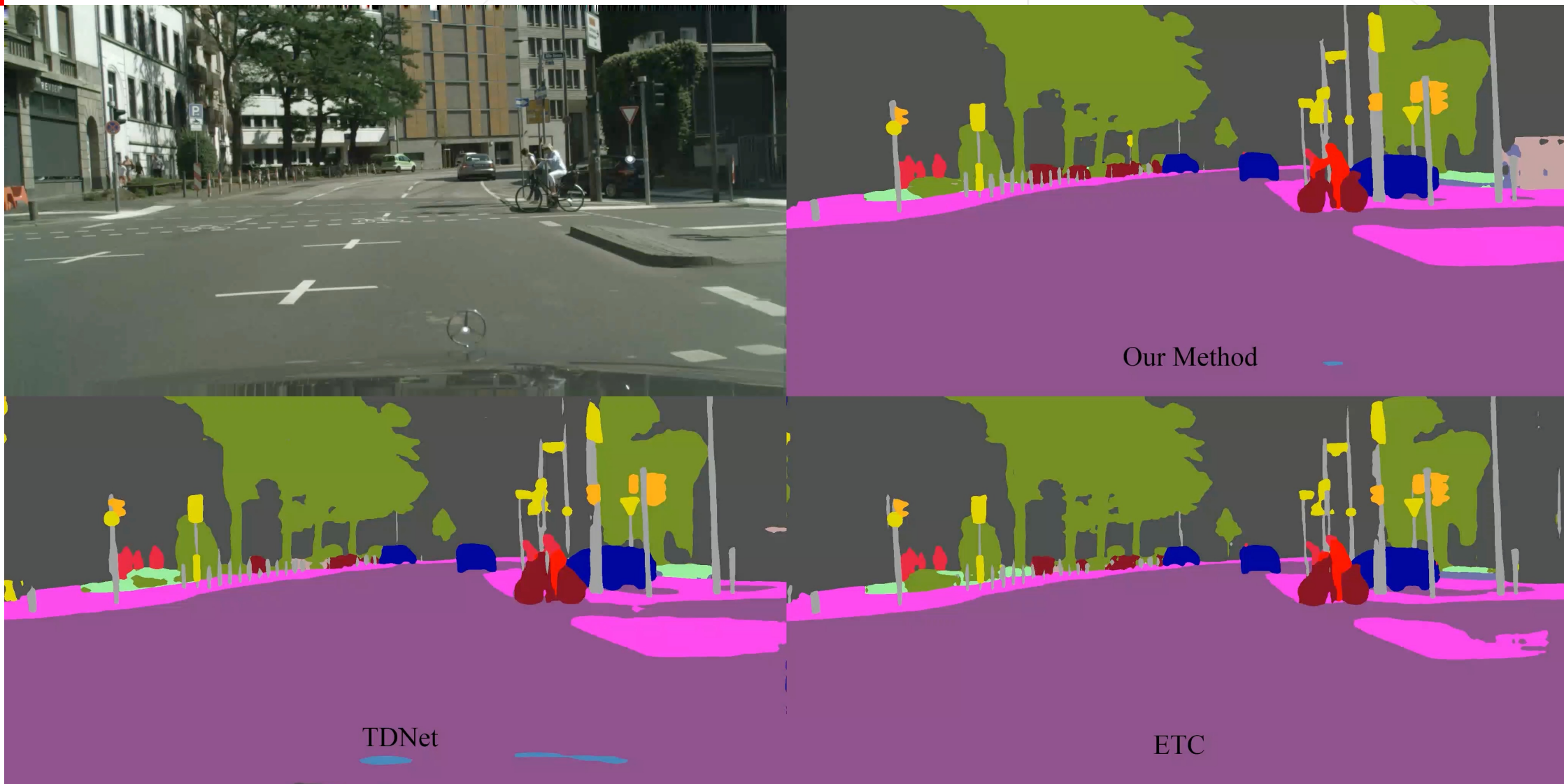
Method	Backbone	Cityscapes			Camvid		
		mIoU(%)↑	TC(%)↑	fps(frame/s)↑	mIoU(%)↑	TC(%)↑	fps(frame/s)↑
NetWarp ^[4]	ResNet101	80.6	-	0.3	67.1	-	2.8
DFP ^[139]	ResNet101	68.7	71.4	9.7	-	-	-
GRFP ^[136]	ResNet101	69.4	-	3.2	66.1	-	4.4
LVS ^[161]	ResNet101	76.8	-	5.9	-	-	-
Accel ^[133]	ResNet101/18	72.1	70.3	3.6	66.7	-	7.6
PSPNet18 ^[28]	ResNet18	75.5	68.5	10.8	71.0	-	24.4
PSPNet50 ^[28]	ResNet50	78.1	-	4.2	74.7	-	8.5
PSPNet101 ^[28]	ResNet101	79.4	69.7	2.1	77.6	77.1	4.1
TDNet-PSP18 ^[132]	ResNet18	76.8	70.4	11.8	72.6	73.2	25.2
TDNet-PSP50 ^[132]	ResNet101	79.9	71.1	5.6	76.0	77.4	11.1
ETC-PSP18 ^[135]	ResNet18	73.1	70.6	10.8	75.2	77.3	24.4
ETC-PSP101 ^[135]	ResNet101	79.5	71.7	2.1	79.4	78.6	4.1
STT-PSP18	ResNet18	77.3	73.0	11.5	76.1	81.4	24.7
STT-PSP101	ResNet101	82.5	73.9	2.2	80.2	82.3	4.2

Cityscapes上高性能算法对比

Method	Backbone	mIoU(%)↑	TC(%)↑	fps(frame/s)↑
DVSNet ^[137]	ResNet18	63.2	-	30.3
ICNet ^[152]	ResNet50	67.7	-	50.0
LadderNet ^[148]	DenseNet121	72.8	-	30.3
SwiftNet ^[150]	ResNet18	75.4	-	43.5
BiSeNet18 ^[158]	ResNet18	73.8	-	50.0
BiSeNet34 ^[158]	ResNet34	76.0	-	37.0
TDNet-BiSe18 ^[132]	ResNet18	75.0	70.2	47.6
TDNet-BiSe34 ^[132]	ResNet34	76.4	71.1	38.5
ETC-Mobi ^[135]	MobileNetV2	73.9	69.9	20.8
STT-BiSe18	ResNet18	75.8	71.4	44.2
STT-BiSe34	ResNet34	77.3	72.0	33.8

实验结果

可视化对比



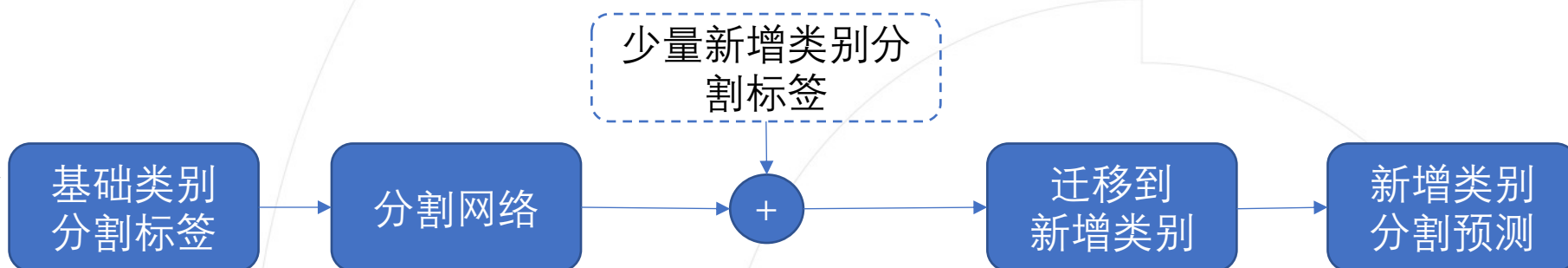
背景及动机

弱监督语义分割



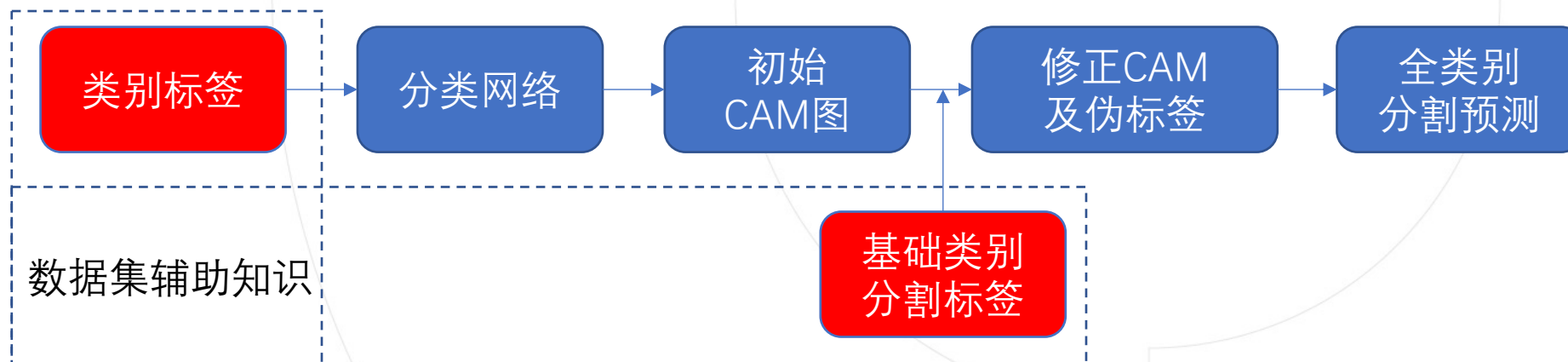
- 1. 数据易获取
- 2. 精度较差

零/小样本语义分割



- 新增类别需要分割标签，否则迁移效果受限

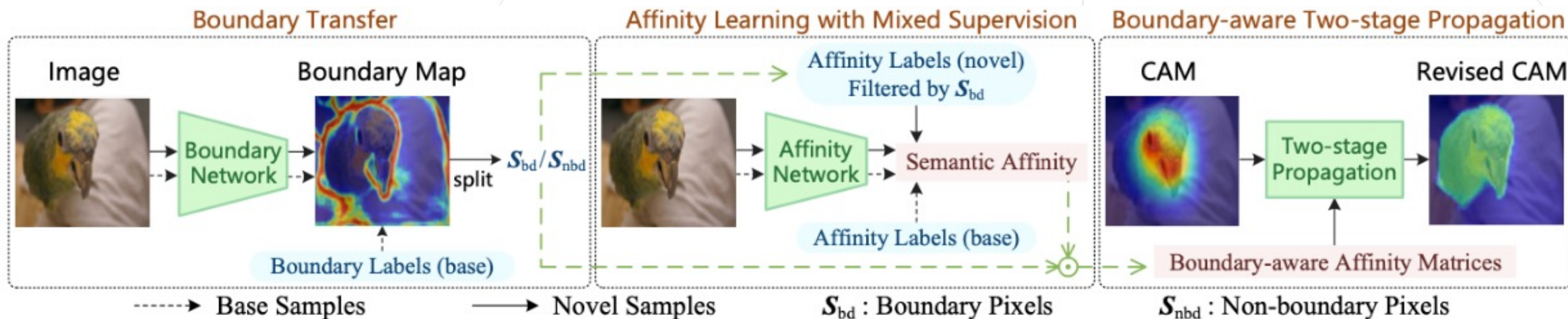
弱样本语义分割



- 1. 新增类别只需分类标签
- 2. 高效利用基础类别分割标签

算法介绍 框架及流程

基于语义亲密度和边界迁移的响应扩展(Response Expansion by Transferring semantic Affinity and Boundary, RETAB)



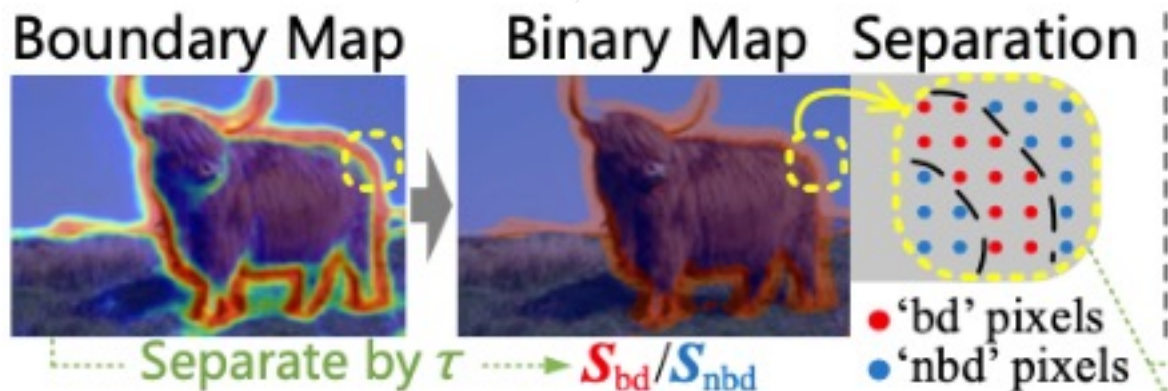
RETAB 算法可以配合任意WSSS 基线模型使用，使用过程分成以下三个步骤:

1. 训练分类模型得到CAM 作为初始响应图;
2. 基于RETAB 进行响应传播，并得到新增类别的伪标签;
3. 联合使用基础类别的稠密标注及新增类别的伪标签以全监督方式训练语义分割网络。

算法介绍

细节说明

边界敏感的两阶段信息传播过程



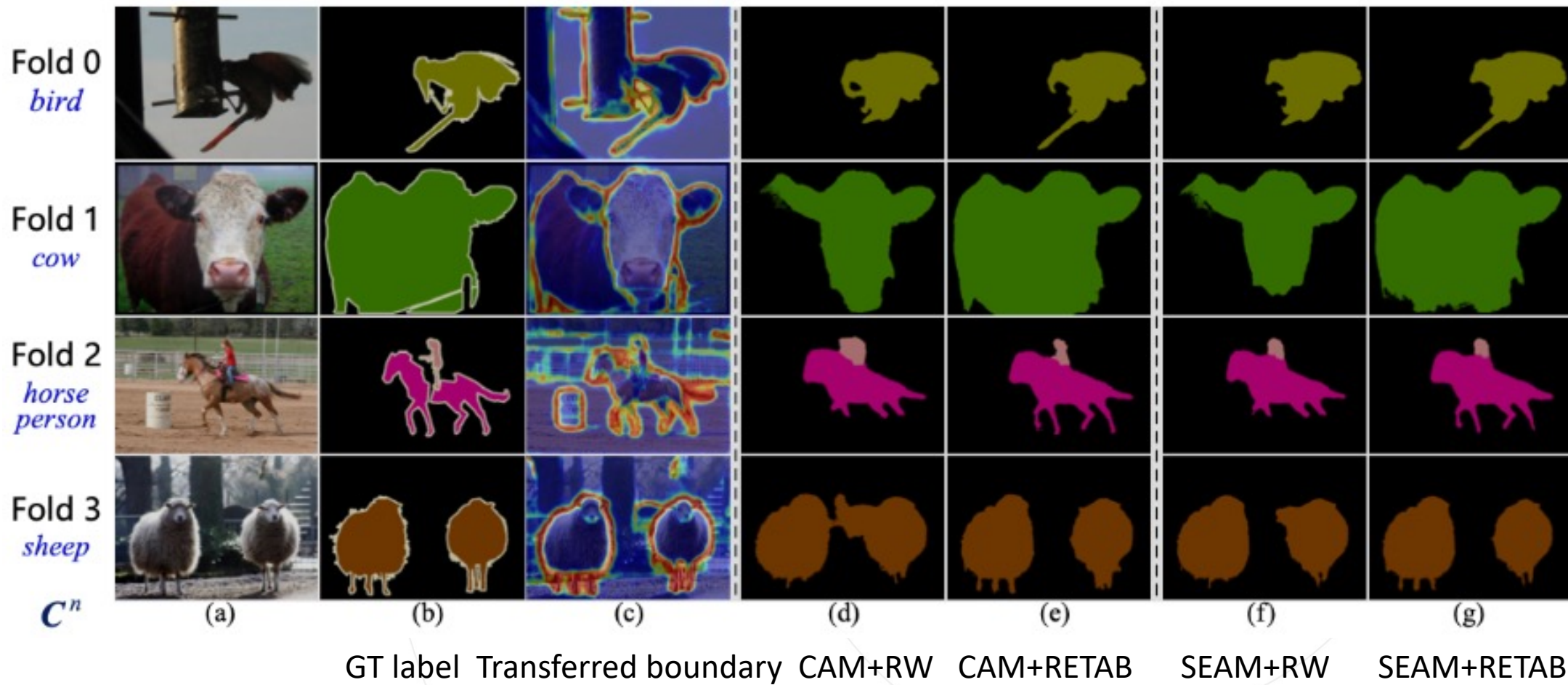
实验结果

PASCAL VOC 2012 测试集上的语义分割性能对比

Method	fold 0			fold 1			fold2			fold3		
	\mathcal{C}	\mathcal{C}^b	\mathcal{C}^n	\mathcal{C}	\mathcal{C}^b	\mathcal{C}^n	\mathcal{C}	\mathcal{C}^b	\mathcal{C}^n	\mathcal{C}	\mathcal{C}^b	\mathcal{C}^n
MCOF ^[205]	61.2	62.6	56.7	61.2	59.7	66.0	61.2	60.6	63.3	61.2	63.6	53.4
SSDD ^[214]	65.5	67.6	58.8	65.5	64.5	68.7	65.5	63.5	72.1	65.5	68.0	57.7
BES ^[188]	66.6	68.8	59.6	66.6	64.9	71.9	66.6	64.7	72.6	66.6	69.3	57.8
SvM ^[206]	66.7	67.5	64.2	66.7	65.8	69.7	66.7	65.6	70.6	66.7	69.6	57.7
CAM+RW ^[193]	63.7	65.4	58.1	63.7	63.7	63.8	63.7	61.4	71.0	63.7	65.8	56.8
CAM+RW(SEGGT)	73.8	78.5	58.6	74.8	76.5	69.5	73.7	74.4	71.5	73.9	79.2	56.9
CAM+RW(AFFGT+SEGGT)	75.2	78.7	64.0	75.3	76.5	71.5	74.6	75.2	72.7	74.1	79.3	57.5
CAM+RETAB	76.3	78.8	68.0	76.0	76.1	75.9	75.4	75.4	75.6	74.8	79.2	60.8
SEAM+RW ^[201]	65.7	-	-	65.7	-	-	65.7	-	-	65.7	-	-
SEAM+RW(SEGGT)	74.0	78.7	59.1	74.5	75.6	71.1	73.5	73.3	74.0	73.7	78.1	59.6
SEAM+RW(AFFGT+SEGGT)	74.9	78.9	62.1	75.2	76.4	71.4	74.3	74.3	74.3	74.2	78.7	59.8
SEAM+RETAB	75.5	78.9	64.6	76.0	76.6	74.0	75.1	75.0	75.6	74.8	79.0	61.5
Fully Oracle	77.9	79.1	74.4	77.9	77.7	78.7	77.9	76.4	82.9	77.9	79.6	72.6

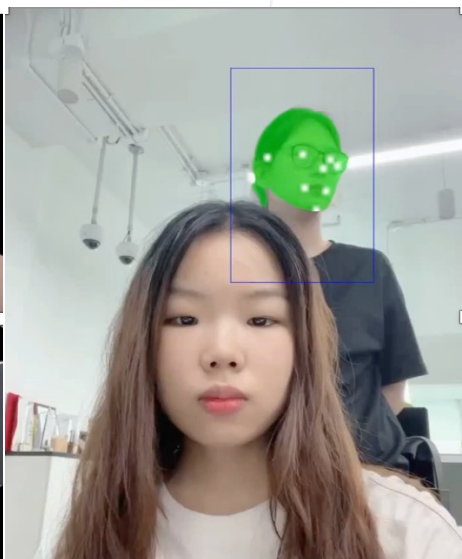
实验结果

PASCAL VOC 2012 测试集上的可视化对比



分割与编辑+Depth

- 移动端实时视频目标分割：人像/人头/头发/人脸/衣物/皮肤/天空等
- 移动端实时视频实例分割：人头/人脸等
- 移动端轻量目标解析：人体等
- 移动端轻量目标抠像：人体/头发/天空
- 移动端轻量单目深度估计：室内/室外等
-



虚拟试穿+手势

- 移动端实时手势识别
- 移动端实时虚拟美甲
- 移动端实时虚拟试鞋
- 移动端实时虚拟试表
-

