Figure 6: **Example qualitative results from PASCAL VOC** 2012. (a) input, (b) ground truth, (c) supervised only, (d) ours (w/o CutMix Aug.), and (e) ours (w/ CutMix Aug.). All the approaches use DeepLabv3+ with ResNet-101 as the segmentation network.

[CVPR 2021] CPS: 基于交叉伪监督的半监督语义分割

Charles

在这篇文章，我们将解读一下我们发表在CVPR 2021的工作CPS: Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision. 我们提出的半监督语义分割算法，在Cityscapes数据集中，使用额外3000张无标注的图像，可以在val set达到82.4% mIoU （单尺度测试）。

Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision

Xiaokang Chen[1*] Yuhui Yuan[2] Gang Zeng[1] Jingdong Wang[2]
[1]Key Laboratory of Machine Perception (MOE), Peking University [2]Microsoft Research Asia

作者单位：北京大学，微软亚洲研究院

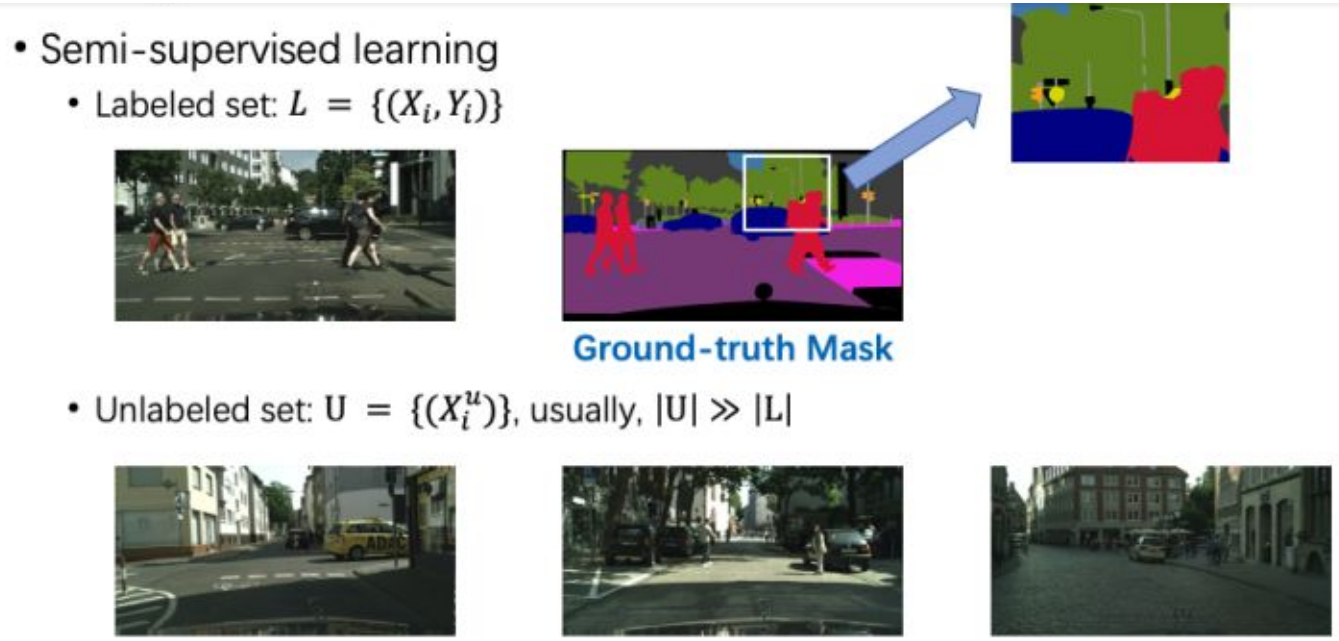
代码：git.io/CPS.

论文：arxiv.org/abs/2106.0122...

在这篇论文中，我们为半监督语义分割任务设计了一种**非常简洁而又性能很好**的算法：cross pseudo supervision (CPS)。训练时，我们使用两个相同结构，但是不同初始化的网络，添加约束使得**两个网络对同一样本的输出是相似的**。具体来说，

相关工作

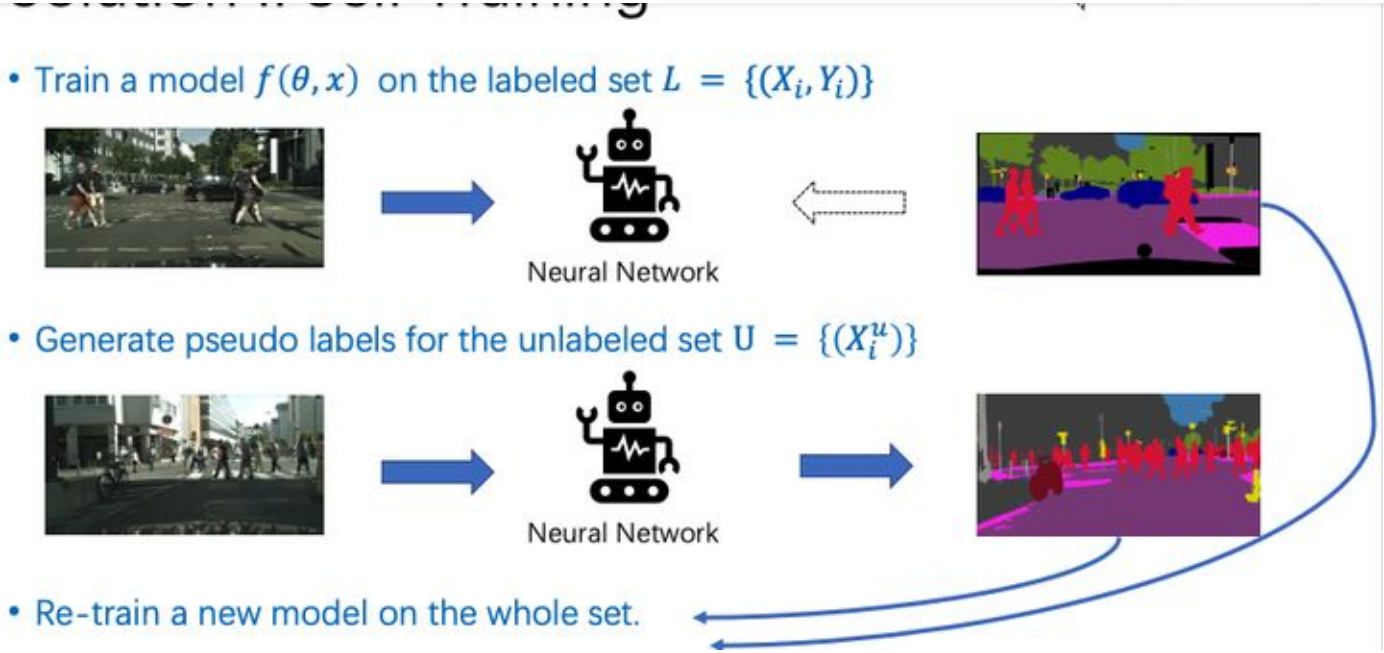
在最开始，我们先来回顾一下半监督语义分割任务的相关工作。不同于图像分类任务，数据的标注对于语义分割任务来说是比较困难而且成本高昂的。我们需要为图像的每一个像素标注一个标签，包括一些特别细节的物体，比如下图中的电线杆 (Pole)。但是，我们可以很轻松的获得RGB数据，比如摄像头拍照。那么，如何利用大量的无标注数据去提高模型的性能，成为半监督语义分割领域研究的问题。



我们将半监督分割的工作总结为两种：self-training和consistency learning。一般来说，self-training是离线处理的过程，而consistency learning是在线处理的。

（1）Self-training

Self-training主要分为3步。第一步，我们在有标签数据上训练一个模型。第二步，我们用预训练好的模型，为无标签数据集生成伪标签。第三步，使用有标注数据集的真值标签，和无标注数据集的伪标签，重新训练一个模型。



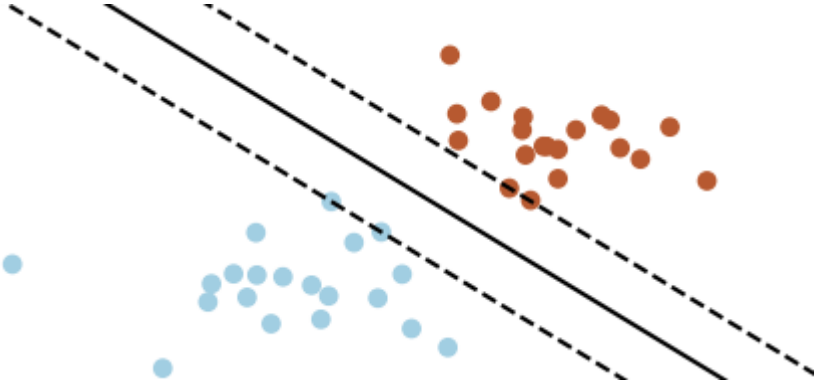
（2）Consistency learning

Consistency learning的核心idea是：**鼓励模型对经过不同变换的同一样本有相似的输出**。这里“变换”包括高斯噪声、随机旋转、颜色的改变等等。

Consistency learning基于两个假设：smoothness assumption 和 cluster assumption。

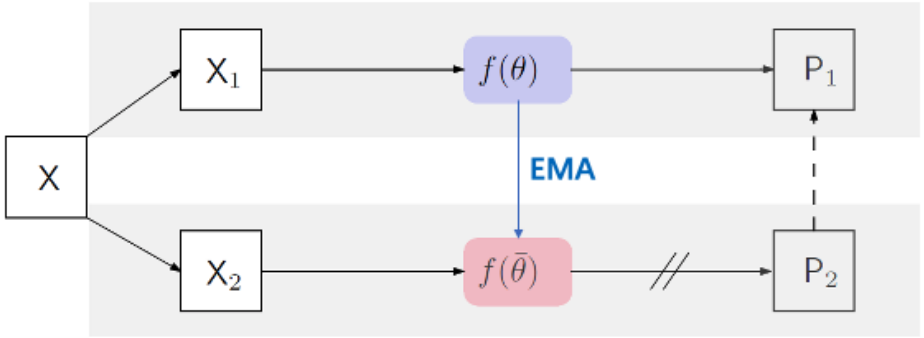
- **Smoothness assumption**: samples close to each other are likely to have the same label.
- **Cluster assumption**: Decision boundary should l
  distribution.

分布密度低的区域。怎么理解这个“密度低”？我们知道，类别与类别之间的区域，样本是比较稀疏的，那么一个好的决策边界应该尽可能处于这种样本稀疏的区域，这样才能更好地区分不同类别的样本。例如下图中有三条黑线，代表三个决策边界，实线的分类效果明显好于另外两条虚线，这就是处于低密度区域的决策边界。



那么，**consistency learning是如何提高模型效果的呢**？在consistency learning中，我们通过对一个样本进行扰动（添加噪声等等），即改变了它在feature space中的位置。但我们希望模型对于改变之后的样本，预测出同样的类别。这个就会导致，在模型输出的特征空间中，同类别样本的特征靠的更近，而不同类别的特征离的更远。只有这样，扰动之后才不会让当前样本超出这个类别的覆盖范围。这也就导致学习出一个更加compact的特征编码。

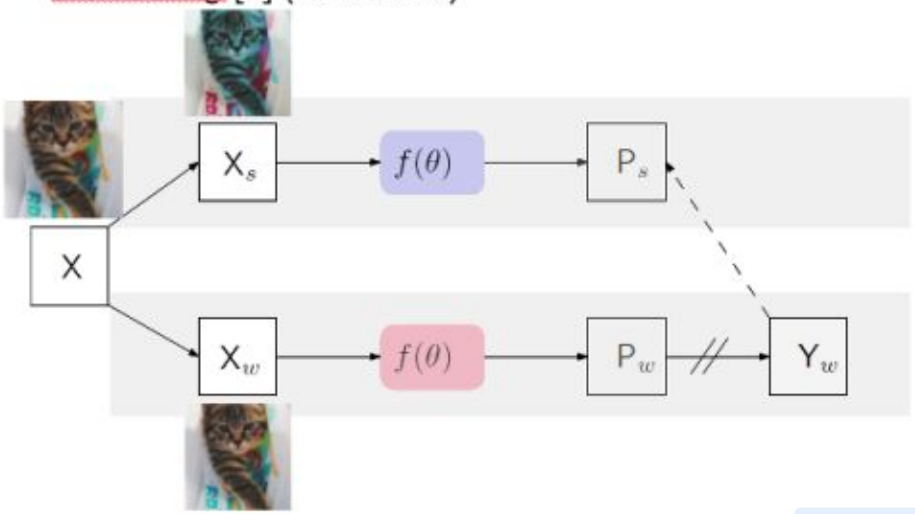
当前，Consistency learning主要有三类做法：mean teacher，CPC，PseudoSeg。



- X → X1/X2: gaussian noise
- $f(\bar{\theta})$ is the EMA of $f(\theta)$
- //: stop gradient
- P1/P2: confidence vector

Mean teacher是17年提出的模型。给定一个输入图像X，添加不同的高斯噪声后得到X1和X2。我们将X1输入网络f(θ)中，得到预测P1；我们对f(θ)计算EMA，得到另一个网络，然后将X2输入这个EMA模型，得到另一个输出P2。最后，我们用P2作为P1的目标，用MSE loss约束。

- PseudoSeg [1] (ICLR 2021)



- Xw: weak augmentation
  - crop, resize, flip
- Xs: strong augmentation
  - weak aug + strong color jitter
- //: stop gradient
- Ps/Pw: confidence vectors
- Yw: pseudo labels

像输入同一个网络f(θ)，得到两个不同的输出。因为“弱增强”下训练更加稳定，他们用“弱增强”后的图像作为target。

- CPC: Cross probability consistency [1] (ECCV 2020)

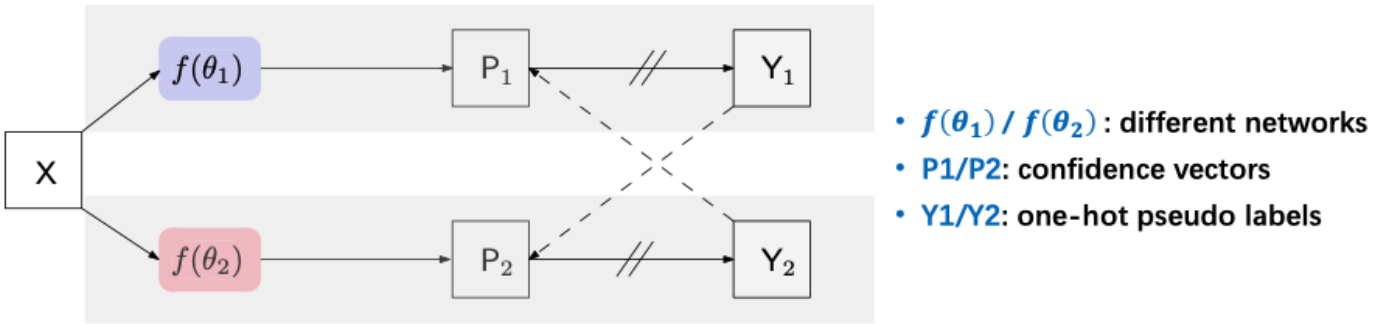
CPC是发表在ECCV 2020的工作（Guided Collaborative Training for Pixel-wise Semi-Supervised Learning）的**简化版本**。在这里，我只保留了他们的核心结构。他们将同一图像输入两个不同网络，然后约束两个网络的输出是相似的。这种方法虽然简单，但是效果很不错。

Motivation

从上面的介绍我们可以简单总结一下：

大家近年来都focus在consistency learning上，而忽略了self-training。实际上，我们实验发现，self-training在数据量不那么小的时候，性能非常的强。那么我们很自然的就想到，为什么不把这两种方法结合起来呢？于是就有了我们提出的CPS：cross pseudo supervision。

Cross Pseudo Supervision (CPS)



我们可以看到，CPS的设计非常的简洁。训练时，我们使用两个网络f(θ1) 和 f(θ2)。这样对于同一个输入图像X，我们可以有两个不同的输出P1和P2。我们通过argmax操作得到对应的one-hot标签Y1和Y2。类似于self-training中的操作，我们将这两个伪标签作为监督信号。举例来说，我们用Y2作为P1的监督，Y1作为P2的监督，并用cross entro

特定的初始化，没准CPS的效果会更好~

在测试的时候，我们只使用其中一个网络进行inference，所以**不增加任何测试/部署时候的开销**。

实验部分

（1）Low data setting。

首先是有标签数据比较少的情况。

我们的方法在VOC和Cityscapes两个数据集的几种不同的数据量情况下都达到了SOTA。表格中1/16, 1/4等表示用原始训练集的 1/16, 1/4作为labeled set，剩余的 15/16, 3/4作为unlabeled set。

Method	ResNet-50				ResNet-101			
	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
MT [32]	66.77	70.78	73.22	75.41	70.59	73.20	76.62	77.61
CCT [27]	65.22	70.87	73.43	74.75	67.94	73.00	76.17	77.56
CutMix-Seg [11]	68.90	70.70	72.46	74.49	72.56	72.69	74.25	75.89
GCT [17]	64.05	70.47	73.45	75.20	69.77	73.30	75.25	77.14
Ours (w/o CutMix Aug.)	68.21	73.20	74.24	75.91	72.18	75.83	77.55	78.64
Ours (w/ CutMix Aug.)	**71.98**	**73.67**	**74.90**	**76.15**	**74.48**	**76.44**	**77.68**	**78.64**

Table 2: **Comparison with state-of-the-arts** on the Cityscapes val set under different partition protocols. All the methods are based on DeepLabv3+.

Method	ResNet-50				ResNet-101			
	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
MT [32]	66.14	72.03	74.47	77.43	68.08	73.71	76.53	78.59
CCT [27]	66.35	72.46	75.68	76.78	69.64	74.48	76.35	78.29
GCT [17]	65.81	71.33	75.30	77.09	66.90	72.96	76.45	78.58
Ours (w/o CutMix Aug.)	69.79	74.39	76.85	78.64	70.50	75.71	77.41	80.08
Ours (w/ CutMix Aug.)	**74.47**	**76.61**	**77.83**	**78.77**	**74.72**	**77.62**	**79.21**	**80.21**

在跟PseudoSeg的对比中，和他们同样的数据划分list，我们也超越了他们的性能：

doSeg [44]. The results of all the other methods are from [44].

Method	#(labeled samples)			
	732	366	183	92
AdvSemSeg [13]	65.27	59.97	47.58	39.69
CCT [27]	62.10	58.80	47.60	33.10
MT [32]	69.16	63.01	55.81	48.70
GCT [17]	70.67	64.71	54.98	46.04
VAT [26]	63.34	56.88	49.35	36.92
CutMix-Seg [11]	69.84	68.36	63.20	55.58
PseudoSeg [44]	72.41	69.14	65.50	57.60
Ours (w/ CutMix Aug.)	**75.88**	**71.71**	**67.42**	**64.07**

这是我们的方法跟self-training进行比较的结果。可以看到，我们的方法由于鼓励模型学习一个更加compact的特征编码，显著地优于self-training。

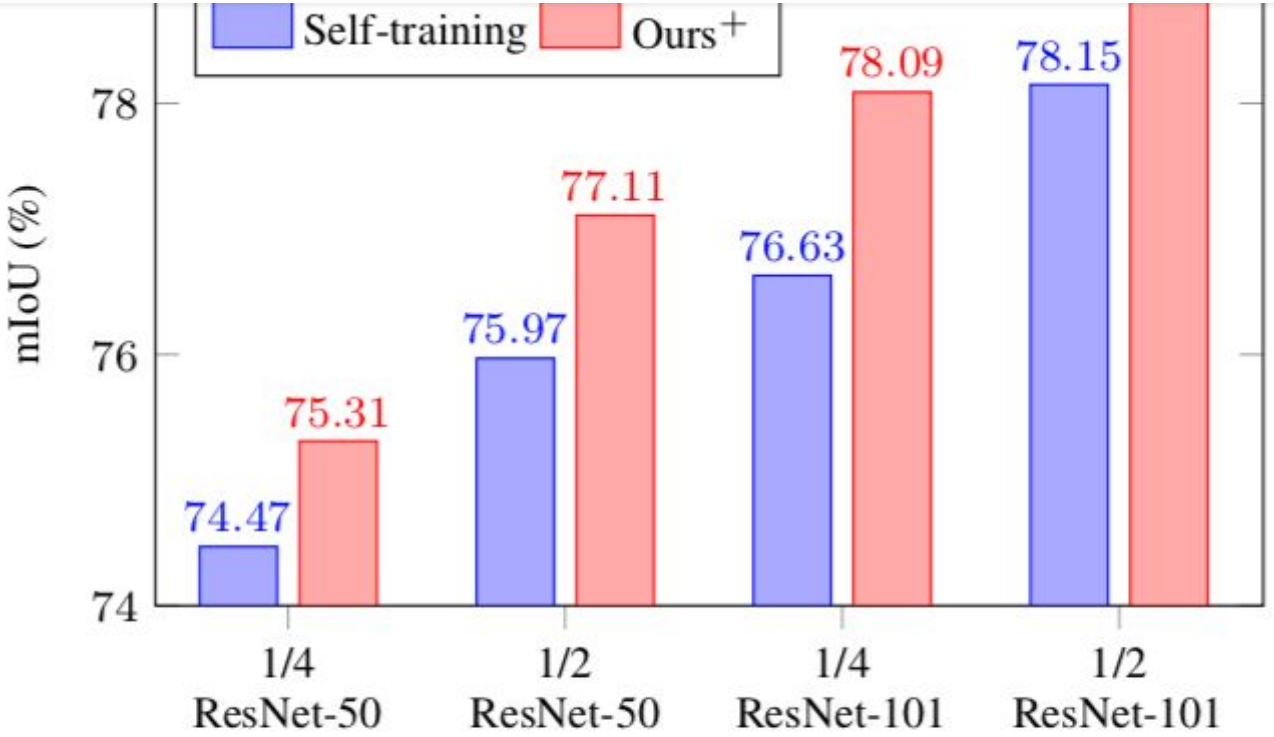
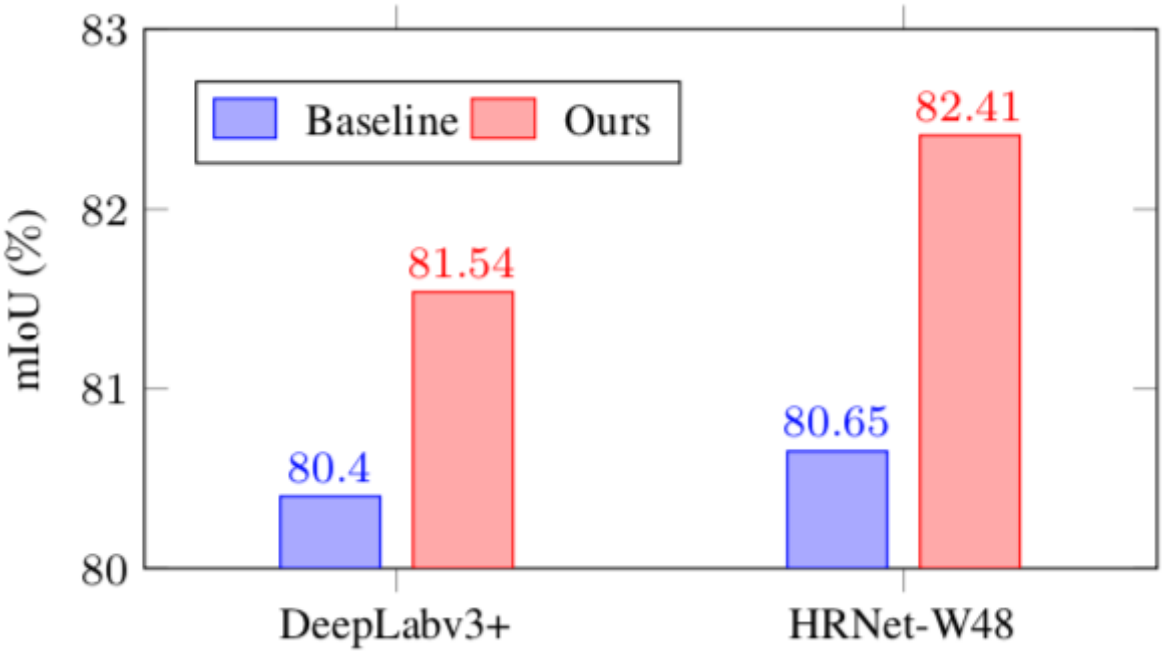


Figure 5: **Comparison with self-training** on PASCAL VOC 2012. The self-training approach is a two-stage approach which takes more training epochs. For a fair comparison, we train our approach with more training epochs (denoted by 'Ours$^+$) so that their epochs are comparable. The CutMix augmentation is not used.

我们还在数据量比较多的情况下进行了实验。在Cityscapes数据集，我们拿训练集的全部图片（大约3000张）作为labeled set，并从coarse set中随机采样3000张RGB图片作为unlabeled set。我们在两个模型进行了实验：DeepLabv3+和HRNet-W48。可以看到，我们的半监督算法可以在非常强的baseline上显著提高性能，最终HRNet-W48在验证集上可以达到单尺度测试下82.4%的mIoU。



（1）分割预测的定量结果。

我们在PASCAL VOC数据集上可视化了一些分割的预测结果。（c）列是仅使用labeled data进行训练的结果，（d）（e）列是我们的预测，（b）列是真值标签。可以看出，由于标注数据很少，（c）的结果不能准确识别物体的语义和边界，而我们CPS可以很好地处理这些问题。

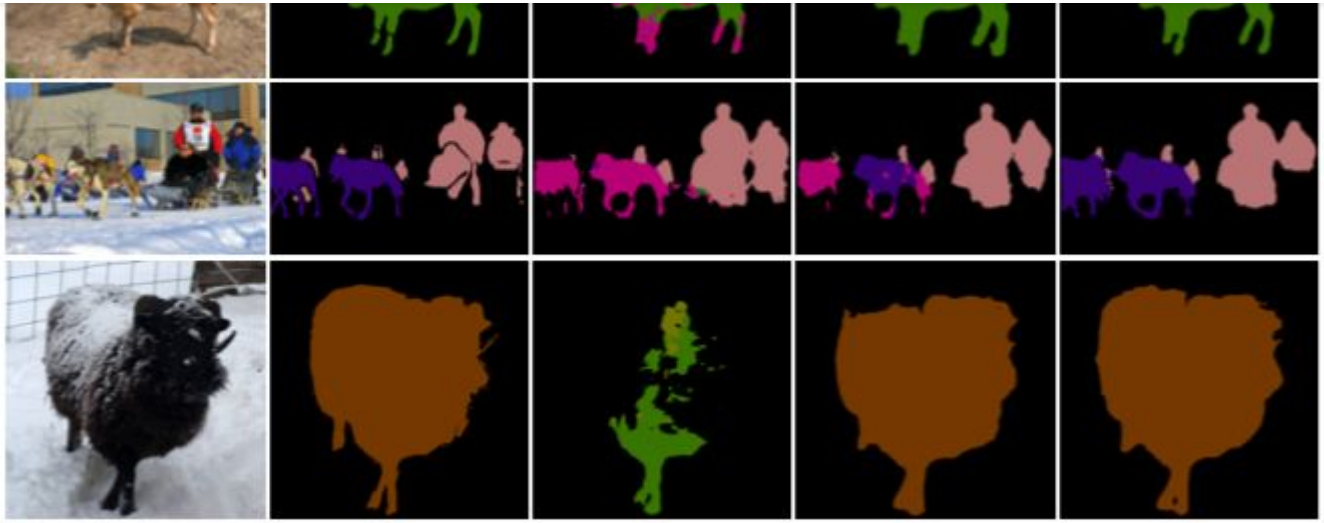


Figure 6: **Example qualitative results from PASCAL VOC** 2012. (a) input, (b) ground truth, (c) supervised only, (d) ours (w/o CutMix Aug.), and (e) ours (w/ CutMix Aug.). All the approaches use DeepLabv3+ with ResNet-101 as the segmentation network.

（2）两个网络的性质分析。

我们在PASCAL VOC上可视化了双路网络的预测的标签的重合情况。我们可以看到，训练初期，overlap较小，通过约束一致性，可以防止单个网络往错误的方向去优化。随着训练进行，overlap逐渐增大，说明两个网络的预测都变得更加准确。

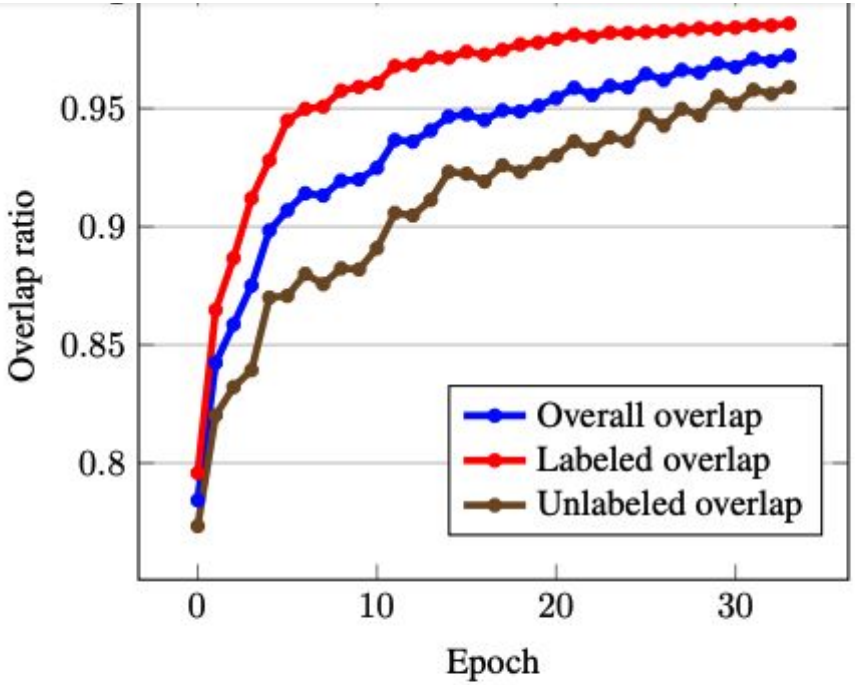


Figure 7: **Prediction overlap of the two networks on PASCAL VOC 2012 under the 1/8 partition.** We use DeepLabv3+ with ResNet-50 as the segmentation network. We only calculate the overlap ratio in the object region, and the pixels belong to the 'background' class are ignored.

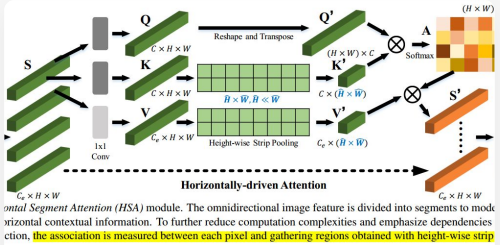
编辑于 06-05

semantic segmentation 计算机视觉 CVPR

文章被以下专栏收录

- **charles的论文解读**
  介绍下自己的/别人的论文。
- **计算机视觉论文速递**
  欢迎关注微信公众号：CVer
- **CVer计算机视觉**
  CVer：一个专注于分享计算机视觉的平台

推荐阅读



【attention系列】捕捉全方位的上下文信息（CVPR 2021）

树梢的风　　发表于计算机视觉...

CVPR2019 Decoders 对于语义分割的重要性

今天为大家推荐一篇CVPR2019 关于语义分割的文章 Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation, 该文章提…

you62580

线上分享|申发力 分割：FCN和Gu

| 极视角线上分享 第 分割是对图像像素 计算机视觉领域的 究问题，在自动驾 和三维重建等需要 领域具有重要作用

极市平台

写下你的评论...

charlesnini 06-09

感觉这篇和也是前几天刚放出来的另一个工作Robust Mutual Learning for Semi-supervised Semantic Segmentation很像，不过没有用mean-teacher架构和对伪标签噪声的refine，更简洁一些。

1

Charles (作者) 回复 charlesnini 06-09

哈哈是的，我们的核心思路非常的像，只不过我们做的早了些~

赞

Charles (作者) 回复 charlesnini 06-09

还有一篇文章，是半监督pose的，CVPR 2021截稿之后放在Arxiv，跟我们的核心idea也几乎一样，算是同期工作，（但我们之前都不知道还有别的组在研究这种做法，也是很巧了），等于在不同的领域也论证了CPS的作用~

arxiv.org/pdf/2011.1249...

赞

展开其他 2 条回复

Eli WU 06-12

分享一下我们今年MICCAI刚被接收的半监督工作，做3D器官分割，核心思想和CPS殊途同归，Semi-supervised Left Atrium Segmentation

赞

这个和带噪声训练的co-teaching方法有很大的相似性。有一个问题，两个模型随机初始化，一开始的时候性能都很差，这个时候互相监督是不是没什么作用？而且我似乎没看到已有的标签的监督。没看原文，所以有些地方不是很清楚，如果有理解的不对的地方还请指出。

赞

Charles (作者) 回复 姜饼哥 06-07

Hi, 有标签数据也是有监督的，这个在paper里有写具体的公式。关于训练初期的问题，GitHub也有人问过类似的，我在这引用一下在GitHub的回答~
question about paper · Issue #3 · charlesCXK/TorchSemiSeg

赞

姜饼哥 回复 Charles (作者) 06-09

不好意思我有点看不懂，在那个issue的回复中，两个model预测都为wrong的情况下，为何有可能有利于模型的训练？比如说简单的前景背景二类分割，两个模型全预测为背景，那么互相作为target，岂不是互相肯定了错误的判断，往错误的方向越走越远了？

赞

查看全部 8 条回复

Jingbo 06-05

赞

德普 06-05

赞

不错 结果简洁有力😄

赞