U2PL：Semi-Supervised Semantic Segmentation **U**sing **U**nreliable **P**seudo-**L**abels

U2PL: 使用不可靠伪标签的半监督语义分割

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, Xinyi Le
1Shanghai Jiao Tong University
2 The Chinese University of Hong Kong
3 SenseTime Research

目 录

摘 要

半监督任务的关键在于充分利用无标签数据。一种常见的做法（Self-Training）是选择高度自信的预测作为pseudo ground-truth，但是这样的做法会导致一个问题，大多数像素在训练中因为是unreliable的所以没有被使用。

「Every Pixel Matters」

即使预测是模糊的，本文认为每一个像素都有其价值。直觉上，不可靠的预测可能会在预测概率最高的类中起到混淆的效果，然而它对应该不属于剩余类的像素有置信度。因此，这样的像素可以被视为那些最没有可能的类别的**负样本**。基于这样的认识，本文开发了一个pipeline来充分利用未标记的数据，具体地说，通过预测熵来分离可靠像素和不可靠像素，将每个不可靠像素推到由负样本组成的分类队列中，并设法用所有候选像素来训练模型。考虑到训练不断迭代，预测变得越来越精确，我们自适应地调整可靠-不可靠划分的阈值。在各种benchmarks和datasets上的实验结果表明，本文方法很优于最先进的替代方法。

摘 要

发现问题：（1）过往为标签方案会从熵/置信度等角度出发，丢弃低质量伪标签样本，减少错误

（2）信息对re-training的影响 符合直觉，但是没有充分利用无标签像素

「Every Pixel Matters」

问题转化：（1）证明低质量标签的价值

（2）怎么筛选低质量样本

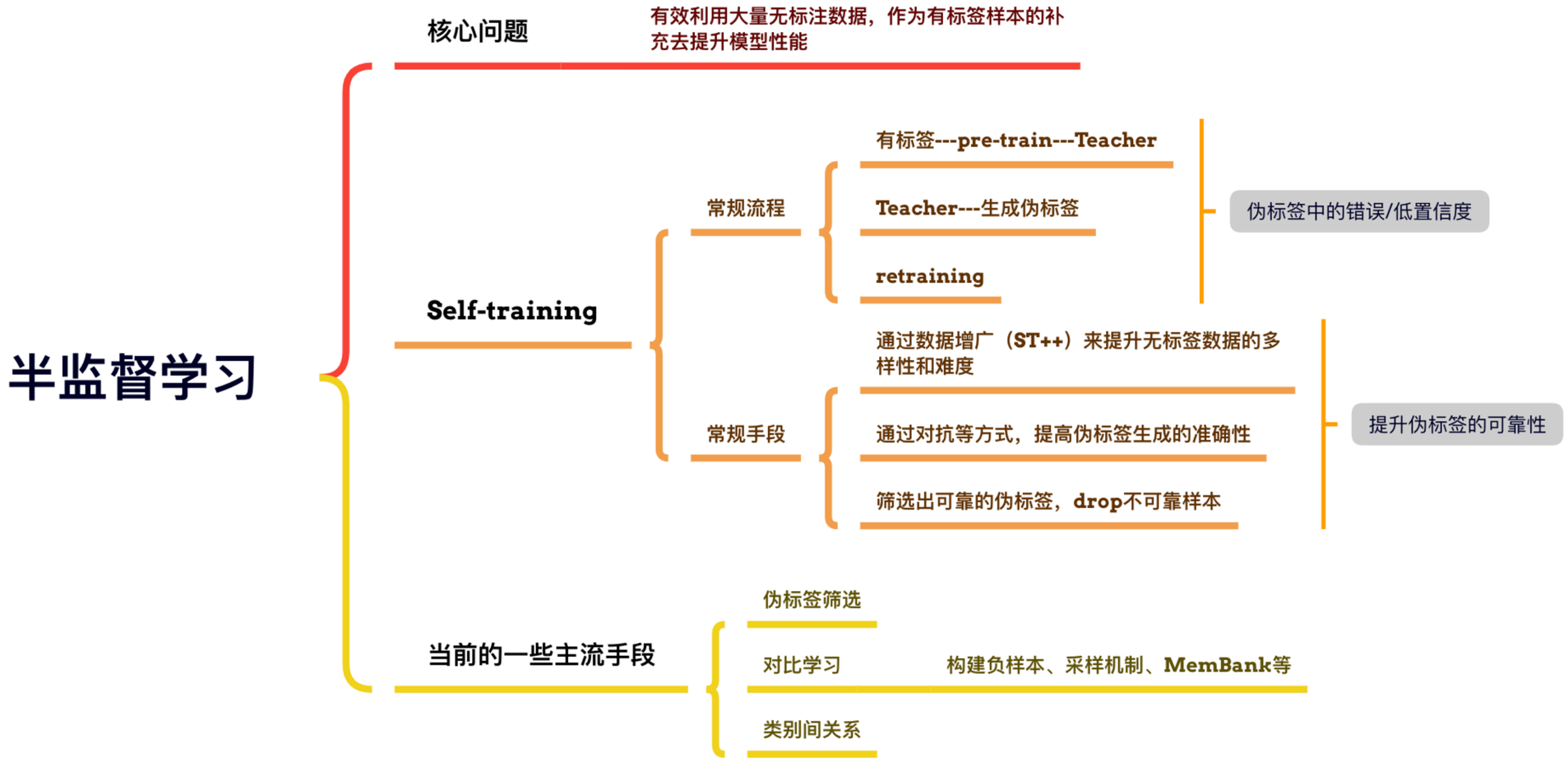
（3）怎么用低质量样本

充分利用无标签数据带来模型精度的提升

总 结

本文提出了一个半监督语义切分框架U2PL，该框架通过将不可靠的伪标签加入到训练中，其性能优于许多现有的最新方法，这表明我们的框架为半监督学习研究提供了一个新的有希望的范例。本文的消融实验证明了这项工作的洞察力是非常可靠的。定性结果为其有效性提供了直观的证明，尤其是在语义对象之间的边界或其他歧义区域上具有更好的性能。缺点：耗时。

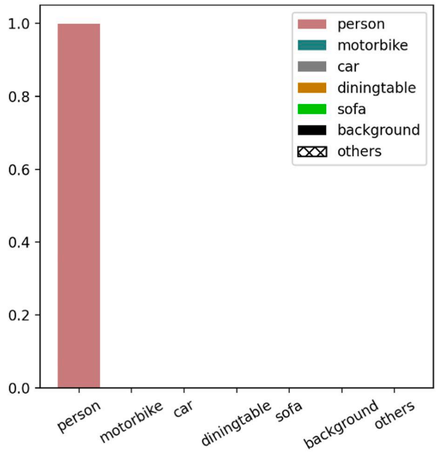
背景介绍


半监督学习
核心问题
有效利用大量无标注数据，作为有标签样本的补充去提升模型性能
Self-training
常规流程
有标签---pre-train---Teacher
Teacher---生成伪标签
retraining
伪标签中的错误/低置信度
常规手段
通过数据增广（ST++）来提升无标签数据的多样性和难度
通过对抗等方式，提高伪标签生成的准确性
筛选出可靠的伪标签，drop不可靠样本
提升伪标签的可靠性
当前的一些主流手段
伪标签筛选
对比学习
构建负样本、采样机制、MemBank等
类别间关系

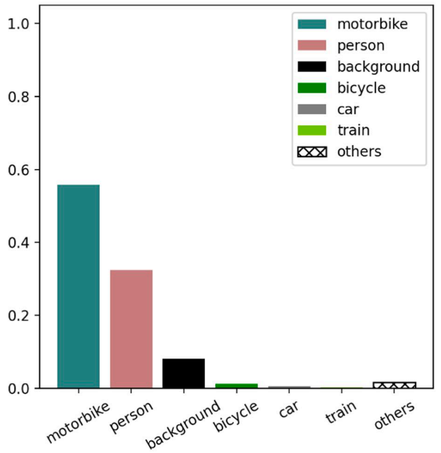
背景介绍

(a) Input image with entropy map.

(b) Pseudo-label after filtering.



(c) Reliable prediction (yellow cross).



(d) Unreliable prediction (white cross).

问题定义

$$\begin{cases} D^l = \{(x_i, y_i)\}_{i=1}^{M} \\ D^u = \{(u_i)\}_{i=1}^{N} \\ N \gg M \end{cases}$$

(l: labeled, u: unlabeled)

$$\mathscr{L} = \mathscr{L}_s + \lambda_u \mathscr{L}_u + \lambda_c \mathscr{L}_c$$

(u: unsupervised, s: supervised, c: contrastive loss)

$$\mathcal{L}_s = \frac{1}{|\mathcal{B}_l|} \sum_{(\mathbf{x}_i^l, \mathbf{y}_i^l) \in \mathcal{B}_l} \ell_{ce}(f \circ h(\mathbf{x}_i^l; \theta), \mathbf{y}_i^l),$$

$$\mathcal{L}_u = \frac{1}{|\mathcal{B}_u|} \sum_{\mathbf{x}_i^u \in \mathcal{B}_u} \ell_{ce}(f \circ h(\mathbf{x}_i^u; \theta), \hat{\mathbf{y}}_i^u),$$

$$\mathcal{L}_c = -\frac{1}{C \times M} \sum_{c=0}^{C-1} \sum_{i=1}^{M} \log \left[\frac{e^{\langle \mathbf{z}_{ci}, \mathbf{z}_{ci}^+ \rangle / \tau}}{e^{\langle \mathbf{z}_{ci}, \mathbf{z}_{ci}^+ \rangle / \tau} + \sum_{j=1}^{N} e^{\langle \mathbf{z}_{ci}, \mathbf{z}_{cij}^- \rangle / \tau}} \right]$$

朴素的SELF-TRAINING方案

1. 【监督学习】在有标签图像D^l上完全训练得到一个初始的教师模型T,

2. 【伪标签生成】用教师模型T在所有的无标签图像D^u上预测one-hot伪标签，得到伪标签集$\hat{D}^u = \left\{(u_i, T(u_i))\right\}_{i=1}^{N}$

3. 【Re-training】混合有标签图像和无标签图像及其伪标签$D_l \cup \hat{D}^u$，在其上重新训练一个学生模型S，用于最终的测试

$$\mathscr{L}^u_{plain} = H(T(x), S(\mathscr{A}^w(x))$$

PSEUDO-LABELING

1. 【监督学习】在有标签图像D^l上完全训练得到一个初始的教师模型T,

2. 【伪标签生成】用教师模型T在所有的无标签图像D^u上预测one-hot伪标签，得到伪标签集
$\hat{D}^u = \left\{(u_i, T(u_i))\right\}_{i=1}^{N}$

3. 【Re-training】混合有标签图像和无标签图像及其伪标签$D_l \cup \hat{D}^u$，在其上重新训练一个学生模型S，用于最终的测试

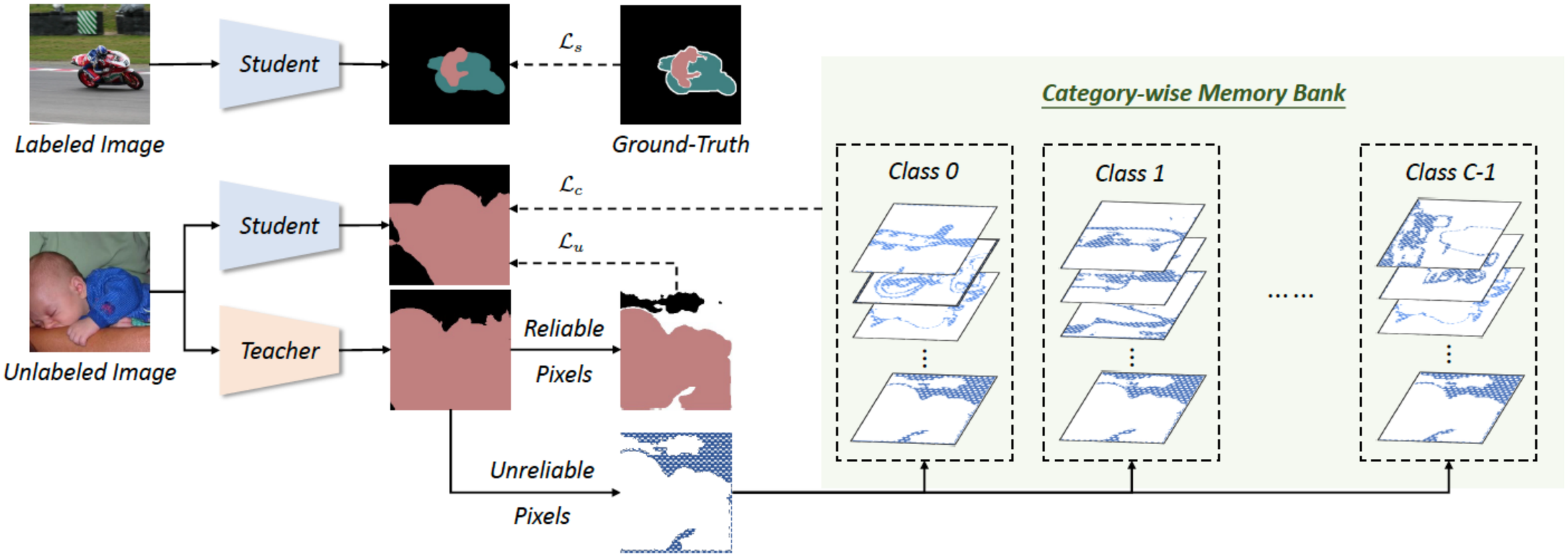
$$\mathscr{H}(p_{ij}) = -\sum_{c=0}^{C-1} p_{ij}(c) log p_{ij}(c)$$

$$\hat{y}_{ij}^u = \begin{cases} \arg\min_c p_{ij}(c)), if\ \mathscr{H} < \gamma_t \\ ignore,\ otherwise \end{cases}$$

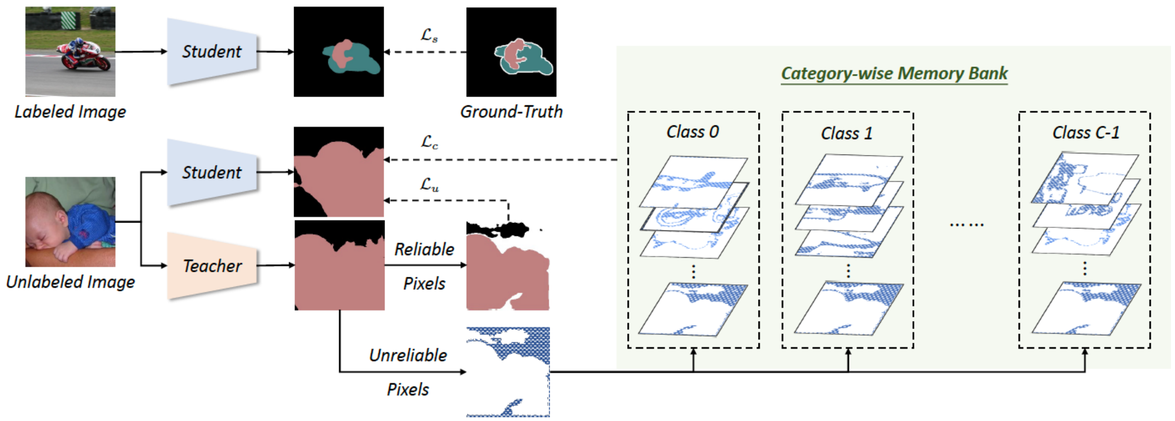
$$\alpha_t = \alpha_0(1 - \frac{t}{total\ epoch})$$

$$\gamma_t = np.percentile(H.flatten(), 100 * (1 - \alpha_t))$$

使用不可靠的伪标签U2PL



使用不可靠的伪标签U2PL



1. Anchor Pixels

2. Positive Samples for each Anchor

3. Negative Samples for each Anchor

使用不可靠的伪标签U2PL

1. Anchor Pixels

2. Positive Samples for each Anchor

3. Negative Samples for each Anchor

$$\mathcal{A}_c^l = \left\{ z_{ij} \,|\, y_{ij} = c,\ p_{ij}(c) > \delta_p \right\} \qquad (\delta_p = 0.3)$$

$$\mathcal{A}_c^u = \left\{ z_{ij} \,|\, \hat{y}_{ij} = c,\ p_{ij}(c) > \delta_p \right\} \qquad (\delta_p = 0.3)$$

$$\mathcal{A}_c = \mathcal{A}_c^l \bigcup \mathcal{A}_c^u$$

使用不可靠的伪标签U2PL

1. Anchor Pixels

2. Positive Samples for each Anchor

3. Negative Samples for each Anchor

$$z_c^{\dagger} = \frac{1}{|\mathcal{A}_c|} \sum_{z_c \in \mathcal{A}_c} z_c$$

使用不可靠的伪标签U2PL

1. Anchor Pixels

2. Positive Samples for each Anchor

3. Negative Samples for each Anchor

对于类别c，一个合格的negative sample应该具备以下特征

1. 不属于类别c

2. 难以与其属于的类别和类别c区分开

使用不可靠的伪标签U2PL

1. Anchor Pixels

2. Positive Samples for each Anchor

3. Negative Samples for each Anchor

$$n_{ij}(c) = \begin{cases} n_{ij}^{l}(c), & if\ image\ i\ is\ labeled \\ n_{ij}^{u}(c), & otherwise \end{cases}$$

$$n_{ij}^{l}(c) = 1[y_{ij} \neq c] \cdot 1[0 \leq \mathcal{O}_{ij}(c) \leq 3]$$

$$n_{ij}^{u}(c) = 1[\mathcal{H}(p_{ij}) > \gamma_t] \cdot 1[r_l \leq \mathcal{O}_{ij}(c) \leq 20]$$

对于类别c，一个合格的negative sample应该具备以下特征

1. 不属于类别c

2. 难以与其属于的类别和类别c区分开

$$\mathcal{N}_c = \left\{ z_{ij} \,|\, n_{ij} = c \right\}$$

使用不可靠的伪标签U2PL

Algorithm 1: Using Unreliable Pseudo-Labels

Initialize $\mathcal{L} \leftarrow 0$;
Sample labeled images $\mathcal{B}_l$ and unlabeled images $\mathcal{B}_u$;
for $\mathbf{x}_i \in \mathcal{B}_l \cup \mathcal{B}_u$ **do**
    Get probabilities: $\mathbf{p}_i \leftarrow f \circ h(\mathbf{x}_i; \theta_t)$;
    Get representations: $\mathbf{z}_i \leftarrow g \circ h(\mathbf{x}_i; \theta_s)$;
    for $c \leftarrow 0$ **to** $C - 1$ **do**
        Get anchors $\mathcal{A}_c$ based on Eq. (11);
        Sample M anchors: $\mathcal{B}_A \leftarrow \texttt{sample}\ (\mathcal{A}_c)$;
        Get negatives $\mathcal{N}_c$ based on Eq. (16);
        Push $\mathcal{N}_c$ into memory bank $\mathcal{Q}_c$;
        Pop oldest ones out of $\mathcal{Q}_c$ if necessary;
        Sample N negatives: $\mathcal{B}_N \leftarrow \texttt{sample}\ (\mathcal{Q}_c)$;
        Get $\mathbf{z}^+$ based on Eq. (12);
        $\mathcal{L} \leftarrow \mathcal{L} + \ell(\mathcal{B}_A, \mathcal{B}_N, \mathbf{z}^+)$ based on Eq. (4);
    end
end
Output: contrastive loss $\mathcal{L}_c \leftarrow \frac{1}{|\mathcal{B}| \times C} \mathcal{L}$

实验内容

Dataset： Pascal VOC 2012 && Cityscapes && SBD && PseudoSeg

网络结构：

- Backbone: ResNet-101
- Decoder: DeepLab v3+

Implementation Details

For the training on the blender and classic PASCAL VOC 2012 dataset, we use stochastic gradient descent **(SGD) optimizer** with **initial learning rate 0:001, weight decay as 0:0001, crop size as 513 x 513, batch size as 16 and training epochs as 80**. For the training on the Cityscapes dataset, we also use stochastic gradient descent **(SGD) optimizer** with **initial learning rate 0:01, weight decay as 0:0005, crop size as 769x769, batch size as 16 and training epochs as 200**. In all experiments, the decoder's learning rate is ten times that of the backbone. We use the poly scheduling to decay the learning rate during the training process.

方案对比

Table 1. Comparison with state-of-the-art methods on *classic* **PASCAL VOC 2012** `val` set under different partition protocols. The labeled images are selected from the original VOC `train` set, which consists of 1,464 samples in total. The fractions denote the percentage of labeled data used for training, followed by the actual number of images. All the images from SBD [18] are regarded as unlabeled data. "SupOnly" stands for supervised training without using any unlabeled data. † means we reproduce the approach.

Method	1/16 (92)	1/8 (183)	1/4 (366)	1/2 (732)	Full (1464)
SupOnly	45.77	54.92	65.88	71.69	72.50
MT† [38]	51.72	58.93	63.86	69.51	70.96
CutMix† [15]	52.16	63.47	69.46	73.73	76.54
PseudoSeg [50]	57.60	65.50	69.14	72.41	73.23
PC^2Seg [48]	57.00	66.28	69.78	73.05	74.15
U^2PL (w/ CutMix)	**67.98** (+15.82)	**69.15** (+5.68)	**73.66** (+4.20)	**76.16** (+2.43)	**79.49** (+2.95)

方案对比

Table 2. Comparison with state-of-the-art methods on *blender* **PASCAL VOC 2012** `val` set under different partition protocols. All labeled images are selected from the augmented VOC `train` set, which consists of 10,582 samples in total. "SupOnly" stands for supervised training without using any unlabeled data. † means we reproduce the approach.

Method	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
SupOnly	67.87	71.55	75.80	77.13
MT† [38]	70.51	71.53	73.02	76.58
CutMix† [15]	71.66	75.51	77.33	78.21
CCT [33]	71.86	73.68	76.51	77.40
GCT [22]	70.90	73.29	76.66	77.98
CPS [9]	74.48	76.44	77.68	78.64
AEL [21]	77.20	77.57	78.06	80.29
U^{2}PL (w/ CutMix)	**77.21** (+5.55)	**79.01** (+3.50)	**79.30** (+1.97)	**80.50** (+2.29)

Table 3. Comparison with state-of-the-art methods on **Cityscapes** `val` set under different partition protocols. All labeled images are selected from the Cityscapes `train` set, which consists of 2,975 samples in total. "SupOnly" stands for supervised training without using any unlabeled data. † means we reproduce the approach.

Method	1/16 (186)	1/8 (372)	1/4 (744)	1/2 (1488)
SupOnly	65.74	72.53	74.43	77.83
MT† [38]	69.03	72.06	74.20	78.15
CutMix† [15]	67.06	71.83	76.36	78.25
CCT [33]	69.32	74.12	75.99	78.10
GCT [22]	66.75	72.66	76.11	78.34
CPS† [9]	69.78	74.31	74.58	76.81
AEL† [21]	74.45	75.55	77.48	79.01
U^{2}PL (w/ CutMix)	***70.30*** (+3.24)	***74.37*** (+2.54)	***76.47*** (+0.11)	***79.05*** (+0.80)
U^{2}PL (w/ AEL)	**74.90** (+0.45)	**76.48** (+0.93)	**78.51** (+1.03)	**79.12** (+0.11)

消融实验

不可靠伪标签的有效性

Table 4. **Ablation study on using pseudo pixels with different reliability**, which is measured by the entropy of pixel-wise prediction (see Sec. 3.3). "Unreliable" denotes selecting negative candidates from pixels with top 20% highest entropy scores. "Reliable" denotes the bottom 20% counterpart. "All" denotes sampling regardless of entropy.

	Unreliable	Reliable	All
1/8 (1323)	**79.01**	77.30	77.40
1/4 (2646)	**79.30**	77.35	77.57

Table A2. **Ablation study on using pseudo pixels with different reliability**, which is measured by the entropy of pixel-wise prediction. "Unreliable" denotes selecting negative candidates from pixels with top 20% highest entropy scores. "Reliable" denotes the bottom 20% counterpart. "All" denotes sampling regardless of entropy. We prove this effectiveness under 1/2 and 1/4 partition protocol on Cityscapes `val` set.

	Unreliable	Reliable	All
1/2 (1488)	**79.05**	77.19	76.96
1/4 (744)	**76.47**	75.16	74.51

消融实验

- 对比学习不是必须的

Table A4. Using unreliable pseudo-labels based on binary classification on **PASCAL VOC 2012** `val` set under different splits.

Method	1/16 (662)	1/8 (1323)	1/4 (2646)	1/2 (5291)
SupOnly	67.87	71.55	75.80	77.13
MT [38]	70.51	71.53	73.02	76.58
U^{2}PL (w/ $\mathcal{L}_c$)	**77.21**	**79.01**	79.30	**80.50**
U^{2}PL (w/ $\mathcal{L}_b$)	75.36	76.62	**79.64**	79.80

汇报完毕 肯请指正

杨润一 2022年4月15日