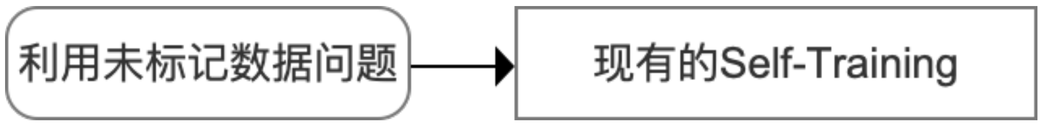
ST++: Make Self-training Work Better for Semi-supervised Semantic Segmentation

ST++：让自我训练更好地用于半监督语义分割

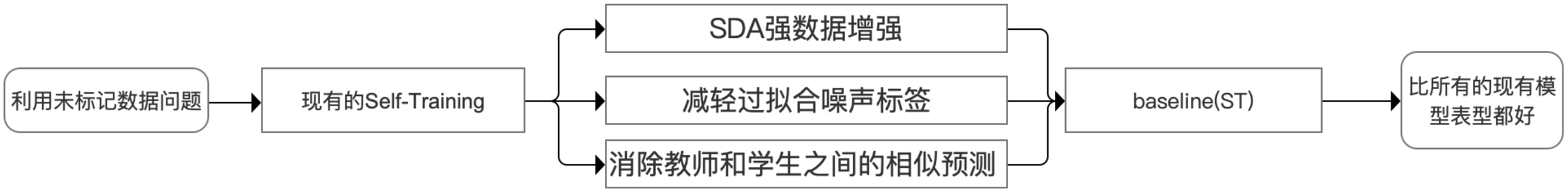
Lihe Yang1 Wei Zhuo3 Lei Qi4,1 Yinghuan Shi1,2* Yang Gao1
1State Key Laboratory for Novel Software Technology, Nanjing University
2National Institute of Healthcare Data Science, Nanjing University
3Tencent 4Southeast University

目 录

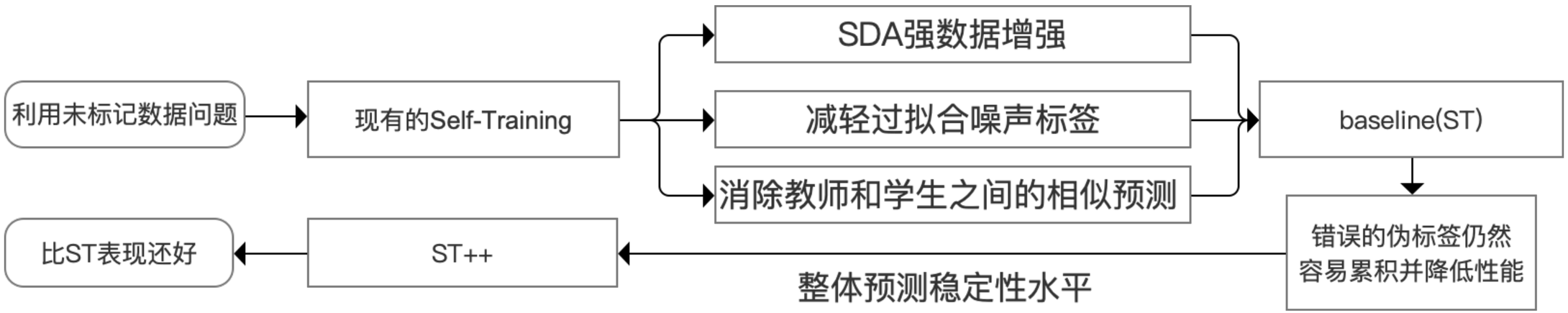
摘 要



摘 要


利用未标记数据问题
现有的Self-Training
SDA强数据增强
减轻过拟合噪声标签
消除教师和学生之间的相似预测
baseline(ST)
比所有的现有模型表型都好

摘 要

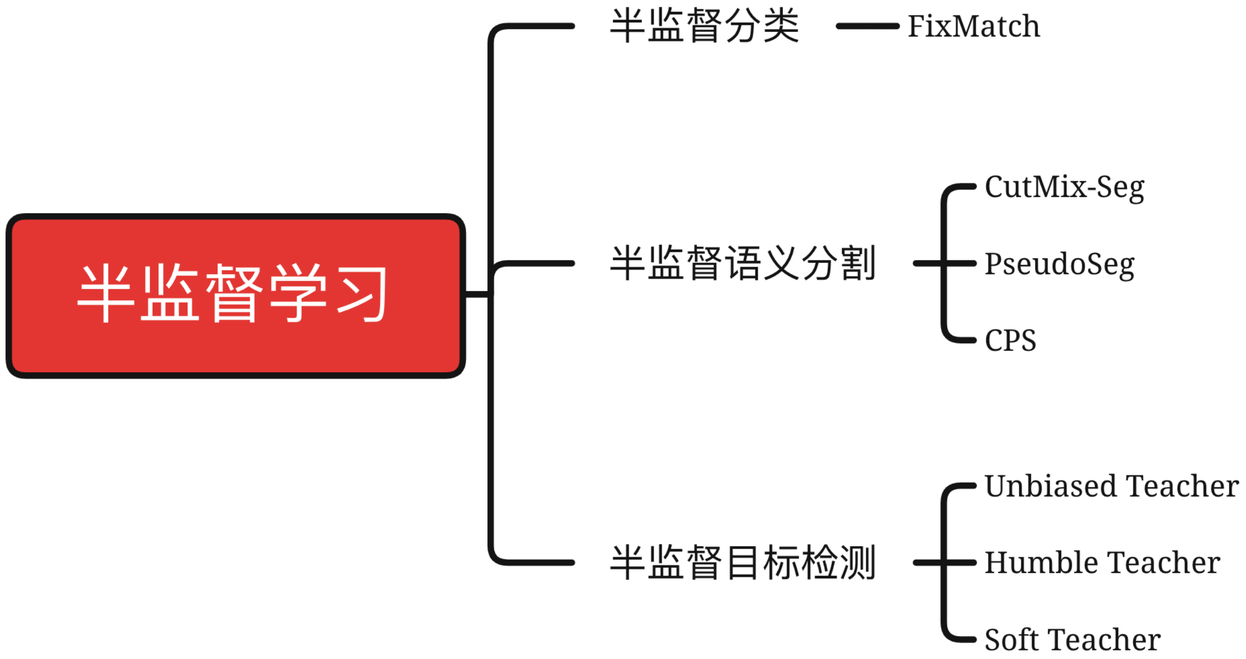

利用未标记数据问题
现有的Self-Training
SDA强数据增强
减轻过拟合噪声标签
消除教师和学生之间的相似预测
baseline(ST)
错误的伪标签仍然
容易累积并降低性能
整体预测稳定性水平
ST++
比ST表现还好

总 结

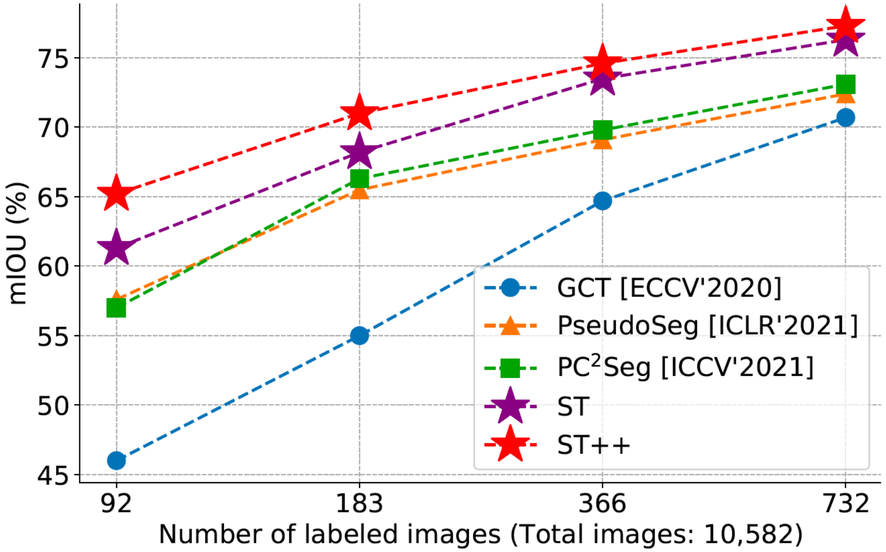
在这项工作中，我们首先通过对未标记图像引入强数据增强、减轻过拟合噪声标签以及消除教师和学生之间的相似预测，构建了一个用于半监督语义分割的baseline(ST)。此外，本文还提出了一种先进的框架(ST++)，以进一步利用未标记的图像。

通过在各种基准测试和设置中进行大量实验，我们的ST和ST++框架都大大优于以前的方法。基于这些启发性的结果，我们进一步详细检验了每个组成部分的有效性，并提供了一些实证分析。我们希望这个简单而有效的框架能成为这一领域未来工作的有力基础或竞争对手。

背景介绍



- 近来较多的半监督学习工作都是基于端到端的框架来做的，学生模型不断学习教师模型产生的伪标签。
- 但是仍然在训练过程中伪标签会误导学生模型的学习
- 耗时耗算力



背景介绍

两种改进方案

- 在学习无标签图像时，在其上施加**强数据增广**，增加学习的难度，学得额外的信息，并缓解对错误伪标签的**过拟合**，例如颜色抖动、灰度和模糊，其中**简单的颜色抖动**最有效
- 由易至难、从可靠标签到不可靠标签，**渐进式**地利用无标签图像及其伪标签。其中，我们提出基于第一阶段训练过程中**伪标签的稳定性**来选取**可靠的图像，而非像素。**

Self Training的一般步骤

- 有监督预训练
- 生成伪标签
- 重新训练

问题定义

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

(l: labeled, u: unlabeled)

$\mathscr{L} = \mathscr{L}^s + \lambda \mathscr{L}^u$ (u: unsupervised, s: supervised)

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

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

ST：注入SDA在无标签数据集

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

1. 在重新训练阶段对无标签图像进行强数据增广来学习

2. 学生模型是在强增广的图像上学习的，可以在教师模型的基础上学得更加丰富的表征

ST：注入SDA在无标签数据集

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。

1. **Entropy minimization**
2. **Consistency regularization**

ST：注入SDA在无标签数据集

Algorithm 1: ST Pseudocode

Input: Labeled training set $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^M$,
Unlabeled training set $\mathcal{D}^u = \{u_i\}_{i=1}^N$,
Weak/strong augmentations $\mathcal{A}^w/\mathcal{A}^s$,
Teacher/student model T/S

Output: Fully trained student model S

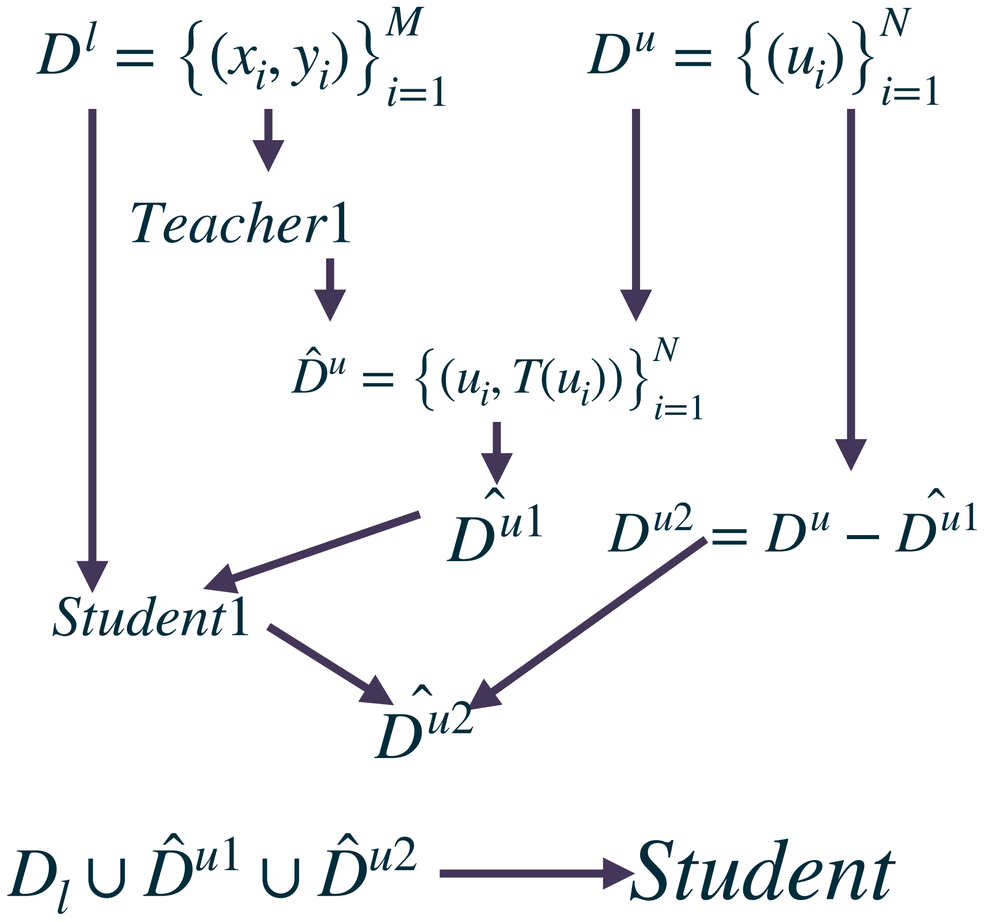
Train T on $\mathcal{D}^l$ with cross-entropy loss $\mathcal{L}_{ce}$
Obtain pseudo labeled $\hat{\mathcal{D}}^u = \{(u_i, T(u_i))\}_{i=1}^N$
Over-sample $\mathcal{D}^l$ to around the size of $\hat{\mathcal{D}}^u$
for *minibatch* $\{(x_k, y_k)\}_{k=1}^B \subset (\mathcal{D}^l \cup \hat{\mathcal{D}}^u)$ **do**
  for $k \in \{1, \ldots, B\}$ **do**
    if $x_k \in \mathcal{D}^u$ **then**
      $x_k, y_k \leftarrow \mathcal{A}^s(\mathcal{A}^w((x_k, y_k))$
    else
      $x_k, y_k \leftarrow \mathcal{A}^w(x_k, y_k)$
    $\hat{y}_k = S(x_k)$
  Update S to minimize $\mathcal{L}_{ce}$ of $\{(\hat{y}_k, y_k)\}_{k=1}^B$
return S

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

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

其中，$\mathscr{A}^s$是使用SDA来对输入进行优化

ST++：根据伪标签的稳定性来选取可靠的图像



可靠的无标签图像的选择策略

$$s_i = \sum_{j=1}^{K-1} meanIOU(M_{ij}, M_{ik})$$

s_i —>stability score，反映了u_i的可靠性和稳定性，M_{ij}表示在第j个checkpoint上的u_i预测出的伪标签。

$$meanIOU = \frac{1}{K-1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum p_{ij} + \sum p_{ji} - p_{ii}}$$

其中p_{ii}是true positive，p_{ij}和p_{ji}分别是false positive和false negative

ST++：根据伪标签的稳定性来选取可靠的图像

Algorithm 2: ST++ Pseudocode

Input: Same as Algorithm 1

Output: Same as Algorithm 1

Train T on $\mathcal{D}^l$ and save K checkpoints $\{T_j\}_{j=1}^K$

for $u_i \in \mathcal{D}^u$ **do**

  for $T_j \in \{T_j\}_{j=1}^K$ **do**

    Pseudo mask $M_{ij} = T_j(u_i)$

  Compute s_i with Equation 4 and $\{M_{ij}\}_{j=1}^K$

Select R highest scored images to compose $\mathcal{D}^{u_1}$

$\mathcal{D}^{u_2} = \mathcal{D}^u - \mathcal{D}^{u_1}$

$\mathcal{D}^{u_1} = \{(u_k, T(u_k))\}_{u_k \in \mathcal{D}^{u_1}}$

Train S on $(\mathcal{D}^l \cup \mathcal{D}^{u_1})$ with ST re-training

$\mathcal{D}^{u_2} = \{(u_k, S(u_k))\}_{u_k \in \mathcal{D}^{u_2}}$

Re-initialize S

Train S on $(\mathcal{D}^l \cup \mathcal{D}^{u_1} \cup \mathcal{D}^{u_2})$ with ST re-training

return S

Algorithm 1: ST Pseudocode

Input: Labeled training set $\mathcal{D}^l = \{(x_i, y_i)\}_{i=1}^M$,
Unlabeled training set $\mathcal{D}^u = \{u_i\}_{i=1}^N$,
Weak/strong augmentations $\mathcal{A}^w/\mathcal{A}^s$,
Teacher/student model T/S

Output: Fully trained student model S

Train T on $\mathcal{D}^l$ with cross-entropy loss $\mathcal{L}_{ce}$

Obtain pseudo labeled $\hat{\mathcal{D}}^u = \{(u_i, T(u_i))\}_{i=1}^N$

Over-sample $\mathcal{D}^l$ to around the size of $\hat{\mathcal{D}}^u$

for *minibatch* $\{(x_k, y_k)\}_{k=1}^B \subset (\mathcal{D}^l \cup \hat{\mathcal{D}}^u)$ **do**

  for $k \in \{1, \ldots, B\}$ **do**

    if $x_k \in \mathcal{D}^u$ **then**

      $x_k, y_k \leftarrow \mathcal{A}^s(\mathcal{A}^w((x_k, y_k))$

    else

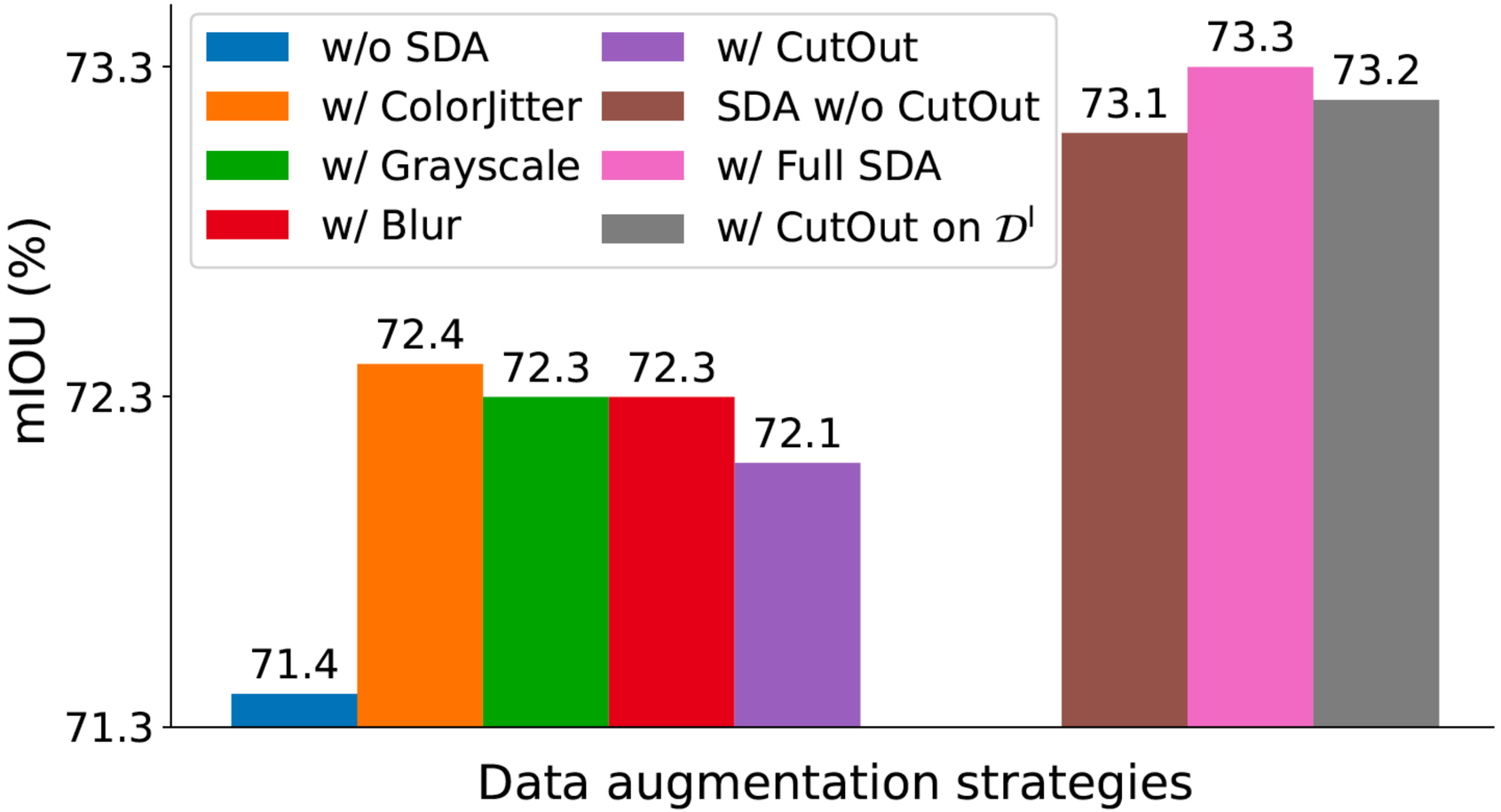
      $x_k, y_k \leftarrow \mathcal{A}^w(x_k, y_k)$

    $\hat{y}_k = S(x_k)$

  Update S to minimize $\mathcal{L}_{ce}$ of $\{(\hat{y}_k, y_k)\}_{k=1}^B$

return S

ST++：根据伪标签的稳定性来选取可靠的图像



实验内容

- **Dataset： Pascal VOC 2012 && Cityscapes两个数据集**
- **网络结构：网络结构主要如下，除 DeepLab v2 为 MS-CoCo Pretrain 外，其余均为 ImageNet Pretrain**
  - Backbone: ResNet-50, ResNet-101
  - Decoder: PSPNet, DeepLab v3+, DeepLab v2
- **Implementation Details**

Batch Size: 16 → 2 * 8 V100 (Pascal) / 4 * 4 V100 (Cityscapes)
Learning Rate:
Poly:0.001 (Pascal) / 0.004 (Cityscapes)
The LR of the segmentation head is 10 times larger than that of the backbone
Epochs: 80 (Pascal) / 240 (Cityscapes)
WDA (Weak Data Augmentation)
Random flip
Rand Size (Multiscale): [0.5, 2]
Rand Crop: 321x321(Pascal) / 721x721(Cityscapes)
SDA (strong data augmentation) on unlabeled data
Color Jitter
Gray Scale
Blur
CutOut with random values filled
TTA (Test Time Augmentation): 5 scales and horizontal flipping

方案对比

Network	Method	1/16 (662)	1/8 (1323)	1/4 (2645)	Network	Method	1/16 (662)	1/8 (1323)	1/4 (2645)
PSPNet ResNet-50	SupOnly	63.8	67.2	69.6	DeepLabv3+ ResNet-50	SupOnly	64.8	68.3	70.5
	CCT [41]	62.2	68.8	71.2		ECS [37]	-	70.2	72.6
	DCC [31]	67.1	71.3	72.5		DCC [31]	70.1	72.4	74.0
	ST	***69.1***	***73.0***	***73.2***		**ST**	***71.6***	***73.3***	***75.0***
	ST++	**69.9**	**73.2**	**73.4**		**ST++**	**72.6**	**74.4**	**75.4**
DeepLabv2 ResNet-101	SupOnly	64.3	67.6	69.5	DeepLabv3+ ResNet-101	SupOnly	66.3	70.6	73.1
	AdvSSL [26]	62.6	68.4	69.9		S4GAN [38]	69.1	72.4	74.5
	S4L [57]	61.8	67.2	68.4		GCT [28]	67.2	72.5	75.1
	GCT [28]	65.2	70.6	71.5		DCC [31]	72.4	74.6	76.3
	ST	***68.6***	***71.6***	***72.5***		**ST**	***72.9***	***75.7***	***76.4***
	ST++	**69.3**	**72.0**	**72.8**		**ST++**	**74.5**	**76.3**	**76.6**

方案对比

Method	ResNet-50 / ResNet-101		
	1/16 (662)	1/8 (1323)	1/4 (2645)
SupOnly	64.0 / 68.4	69.0 / 73.3	71.7 / 74.7
CCT [41]	65.2 / 67.9	70.9 / 73.0	73.4 / 76.2
CutMix-Seg [18]	68.9 / 72.6	70.7 / 72.7	72.5 / 74.3
GCT [28]	64.1 / 69.8	70.5 / 73.3	73.5 / 75.3
CPS [12]	68.2 / 72.2	73.2 / 75.8	74.2 / 77.6
CPS† [12]	72.0 / ***74.5***	73.7 / 76.4	74.9 / ***77.7***
ST	***72.2*** / 74.0	***74.8*** / ***76.9***	***75.5*** / 77.6
ST++	**73.2 / 74.7**	**75.5 / 77.9**	**76.0 / 77.9**

Method	# Labeled images (Total: 10582)				
	92	183	366	732	1464
SupOnly	50.7	59.1	65.0	70.6	74.1
GCT [28]	46.0	55.0	64.7	70.7	-
CutMix-Seg [18]	55.6	63.2	68.4	69.8	-
PseudoSeg [66]	57.6	65.5	69.1	72.4	73.2
CPS [12]	***64.1***	67.4	71.7	75.9	-
PC^2Seg [61]	57.0	66.3	69.8	73.1	74.2
ST	61.3	***68.2***	***73.5***	***76.3***	**78.9**
ST++	**65.2**	**71.0**	**74.6**	**77.3**	**79.1**

Fully-supervised setting (10582 images): 78.2

方案对比

Method	# Labeled images (Total: 10582)				
	92	183	366	732	1464
SupOnly	50.7	59.1	65.0	70.6	74.1
GCT [28]	46.0	55.0	64.7	70.7	-
CutMix-Seg [18]	55.6	63.2	68.4	69.8	-
PseudoSeg [66]	57.6	65.5	69.1	72.4	73.2
CPS [12]	***64.1***	67.4	71.7	75.9	-
PC^2Seg [61]	57.0	66.3	69.8	73.1	74.2
ST	61.3	***68.2***	***73.5***	***76.3***	**78.9**
ST++	**65.2**	**71.0**	**74.6**	**77.3**	**79.1**
Fully-supervised setting (10582 images): 78.2					

Method	1/30 (100)	1/8 (372)	1/4 (744)
DeepLabv3+, ResNet-101			
DMT [17]	54.8	63.0	-
CutMix-Seg [18]	55.7	65.8	68.3
ClassMix [40]	-	61.4	63.6
PseudoSeg [66]	61.0	69.8	72.4
DeepLabv3+, ResNet-50			
SupOnly	55.1	65.8	68.4
DCC [31]	-	69.7	72.7
ST	***60.9***	***71.6***	***73.4***
ST++	**61.4**	**72.7**	**73.8**

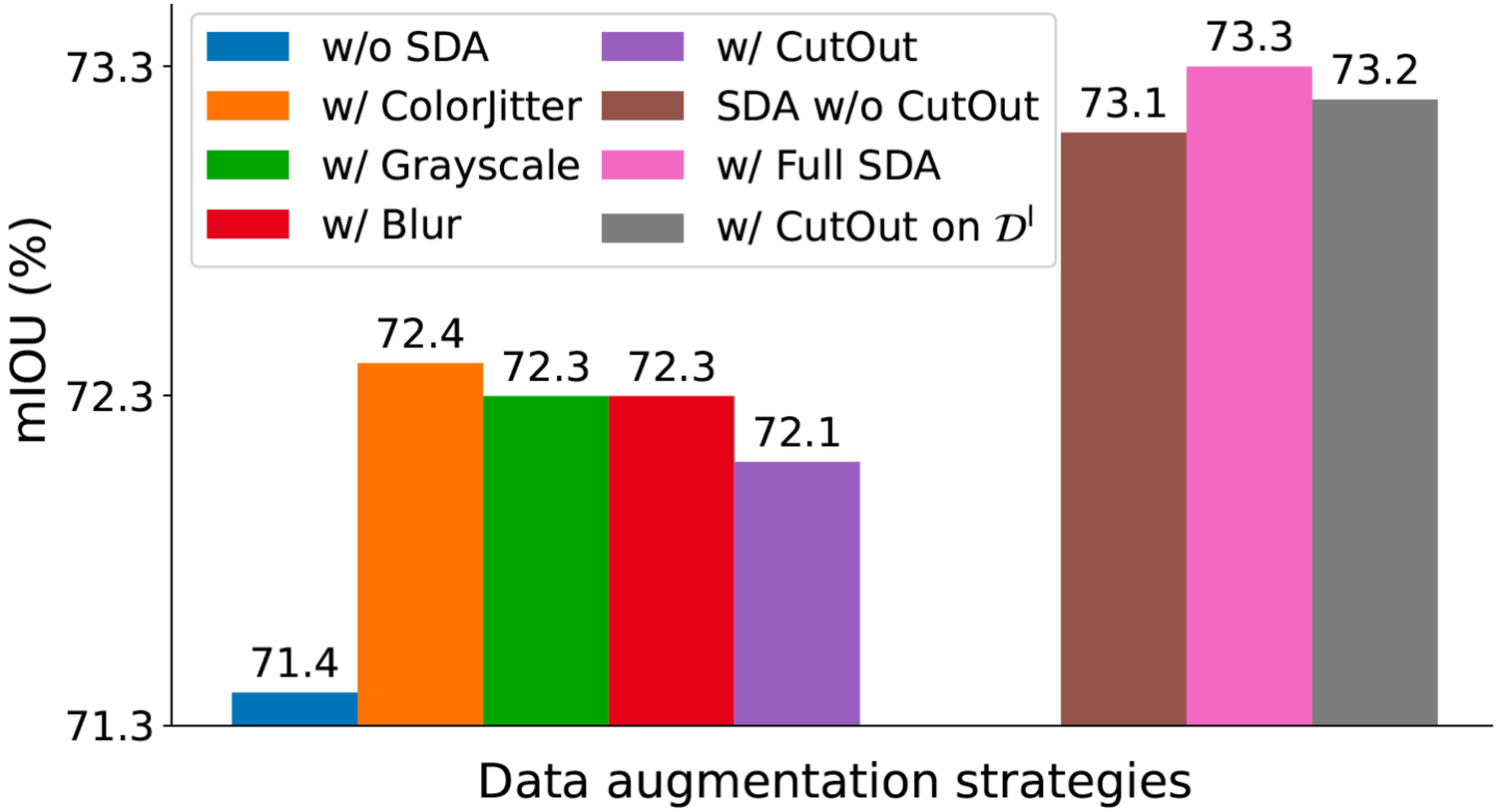
消融实验

强数据增广的意义

Apply SDA on		1/16	1/8	1/4
labeled data	unlabeled data	(662)	(1323)	(2645)
		70.9	71.4	73.5
✓	✓	71.0	73.0	74.3
	✓	**71.6**	**73.3**	**75.0**

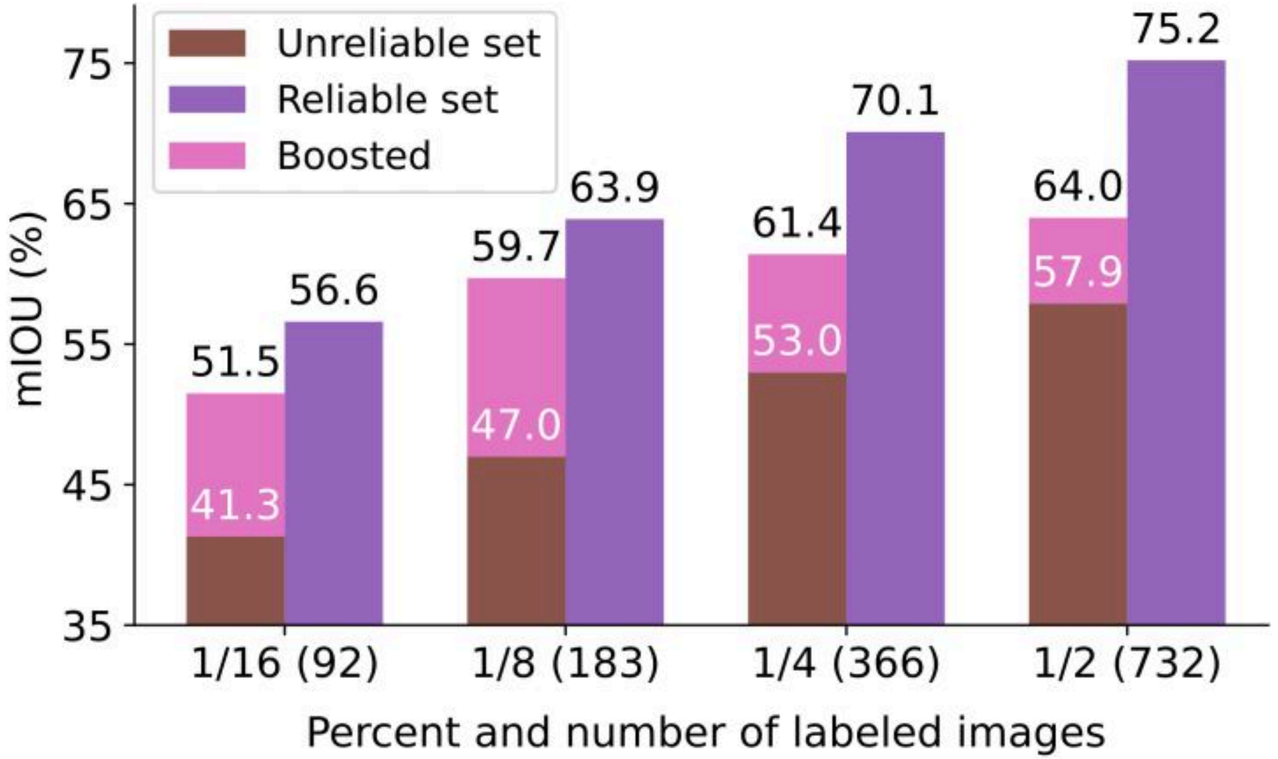
消融实验

不同SDA的意义



消融实验

ST++中选取出的可靠样本和不可靠样本的伪标签质量对比



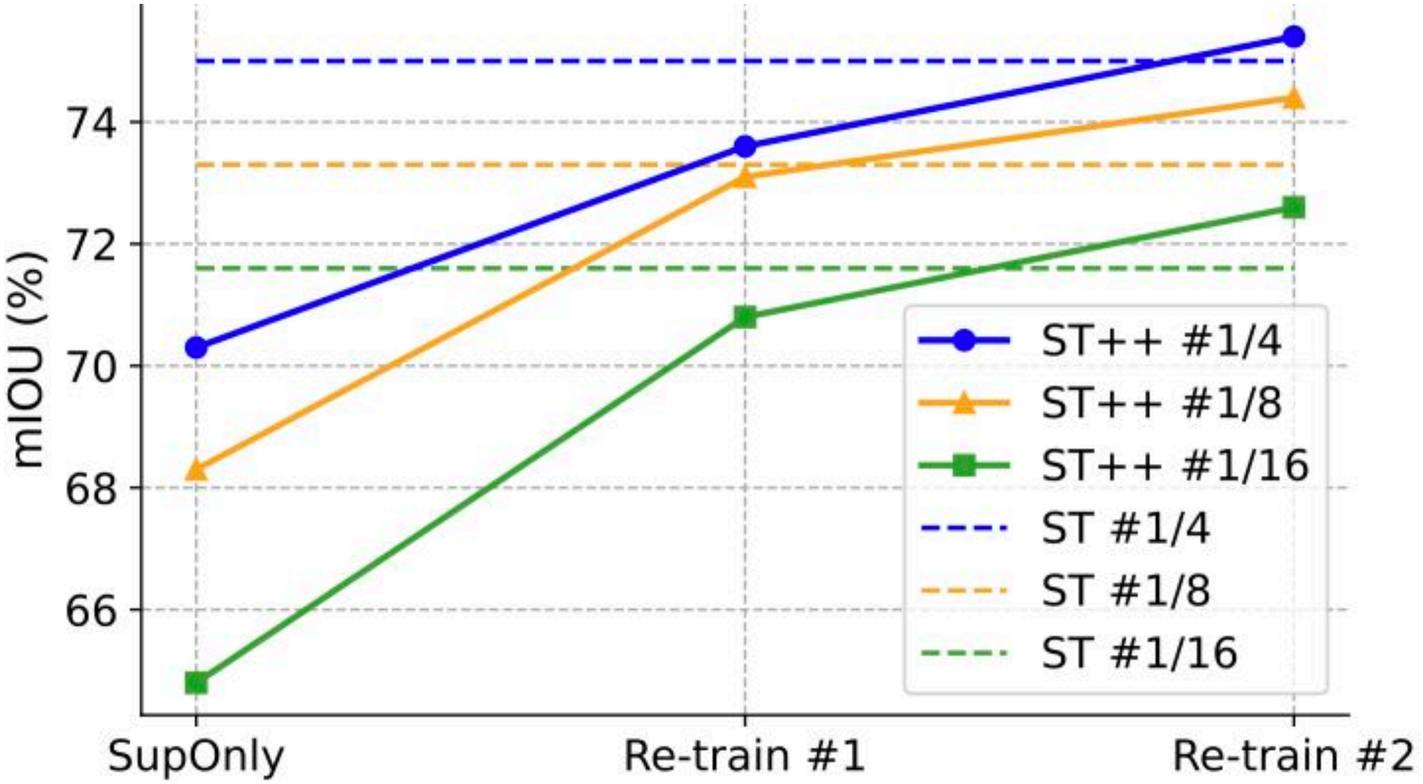
消融实验

- ST++的提升是否仅仅受益于多阶段的策略?

Method	1/16 (662)	1/8 (1323)	1/4 (2645)
One-stage re-training (our ST)	71.6	73.3	75.0
Random two-stage re-training	71.3	73.9	74.7
Selective re-training (our ST++)	**72.6**	**74.4**	**75.4**

消融实验

ST++中两阶段训练的performance



消融实验

ST++中可靠图像的选取比例

Proportion of reliable images	25%	50% (default)	75%
mIOU (%)	74.0	74.4	74.5

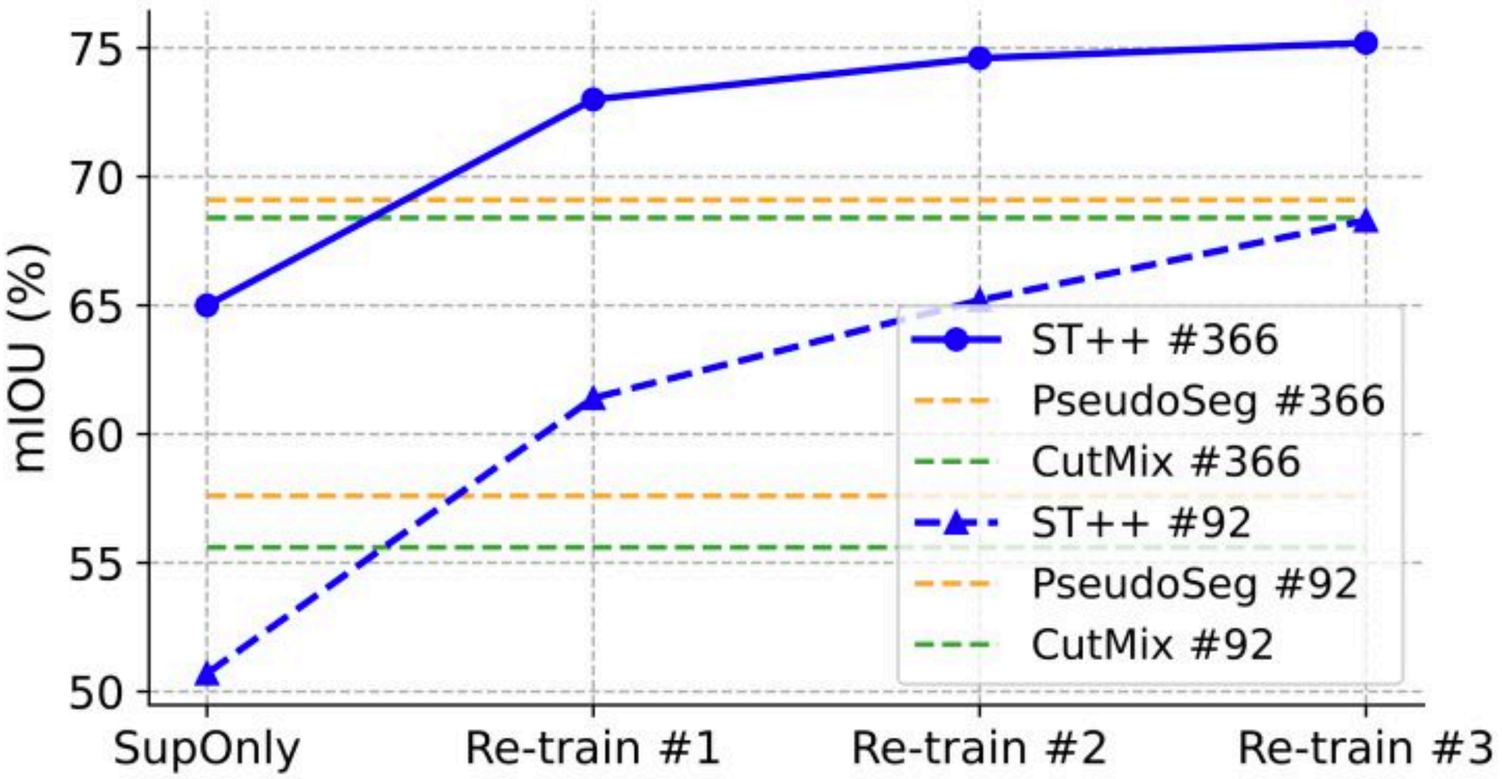
消融实验

ST++中图像级和像素级选择策略的比较

Method	1/16 (662)	1/8 (1323)	1/4 (2645)
our ST (w/o iterative re-train)	71.6	73.3	75.0
Image-level re-train (our ST++)	**72.6**	**74.4**	**75.4**
Pixel-level re-train phase #1	71.3	73.5	74.9
Pixel-level re-train phase #2	71.3	73.8	74.7

消融实验

ST++能否继续迭代训练提升?



一些名词

- 知识蒸馏
- SDA 强数据增广
- Self-training
- 一致性正则化 Consistency Regularization
- 熵最小化 Entropy Minimization
- meanIOU
- 四种强数据增广策略
  - colorjitter, blur, grayscale, Cutout

汇报完毕 肯请指正

杨润一 2022年4月15日