Medical seminar

Wentao Lei

2021/12/1

Paper List

- CVPR2020: FocalMix: Semi-Supervised Learning for 3D Medical Image Detection
- NIPS2021: FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling

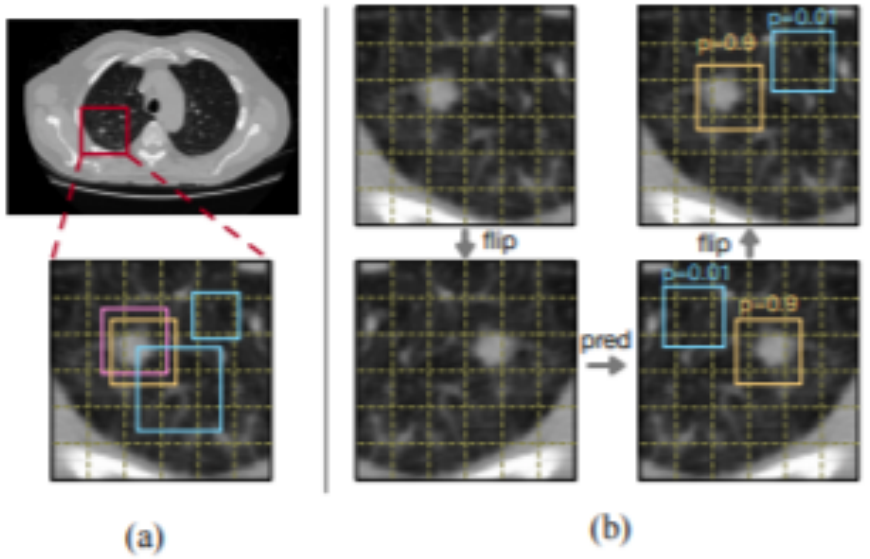
FocalMix: Semi-Supervised Learning for 3D Medical Image Detection

Dong Wang[1*] Yuan Zhang[2*] Kexin Zhang[2,3†] Liwei Wang[1,2]
[1]Center for Data Science, Peking University
[2]Key Laboratory of Machine Perception, MOE, School of EECS, Peking University
[3]Yizhun Medical AI Co., Ltd
{wangdongcis,yuan.z,zhangkexin,wanglw}@pku.edu.cn

Background

Object Detection in 3D Medical Images

- 图片被分为多个网格来对应真实的groundtruth
- 每个网格对应一个特征的检测区域
- 如果网格和对应groundtruth的iou超过一定阈值则为睁眼本，否则为负样本



Methods

Focal Loss

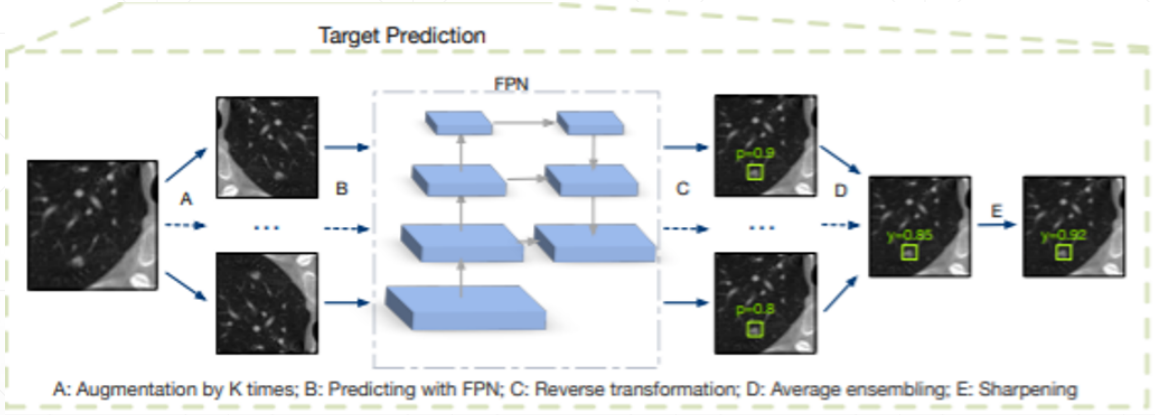
- γ 是focal loss的参数，这里设为2.
- 用于降低置信度过高的样本在训练中的权重
- 正样本和负样本的权重不同

$$SFL(p) = [\alpha_0 + y(\alpha_1 - \alpha_0)] \cdot |y - p|^{\gamma} \cdot CE(y, p),$$

Methods

Prediction

- 对每个未标注样本做k次数据增强
- 对每个增强后的样本进行一次预测
- 平均每次预测的结果并且进行锐化的操作



$$\bar{y} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{p}_{\mathrm{Model}}(\hat{u}_k;\theta). \quad \mathrm{Sharpen}(\bar{y}, T)_i = \bar{y}_i^{\frac{1}{T}} \Big/ \sum_{j=1}^{L} \bar{y}_j^{\frac{1}{T}},$$

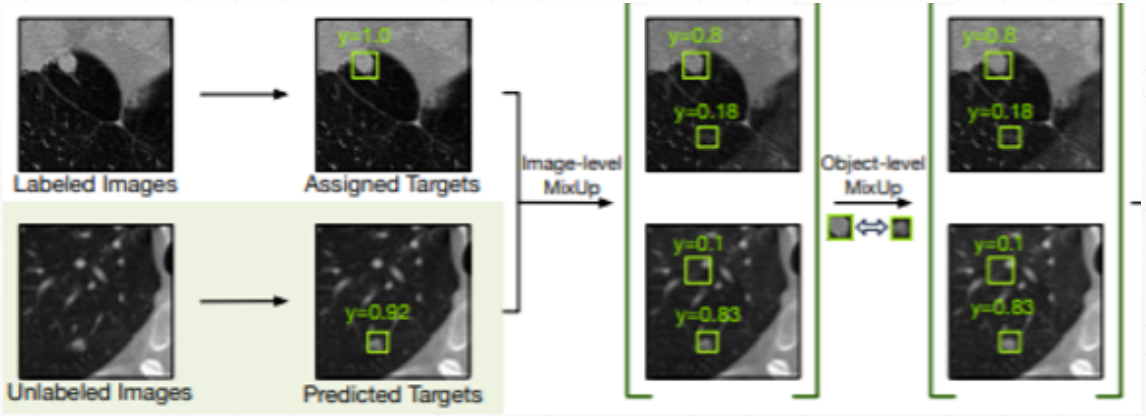
Methods

Mix-up

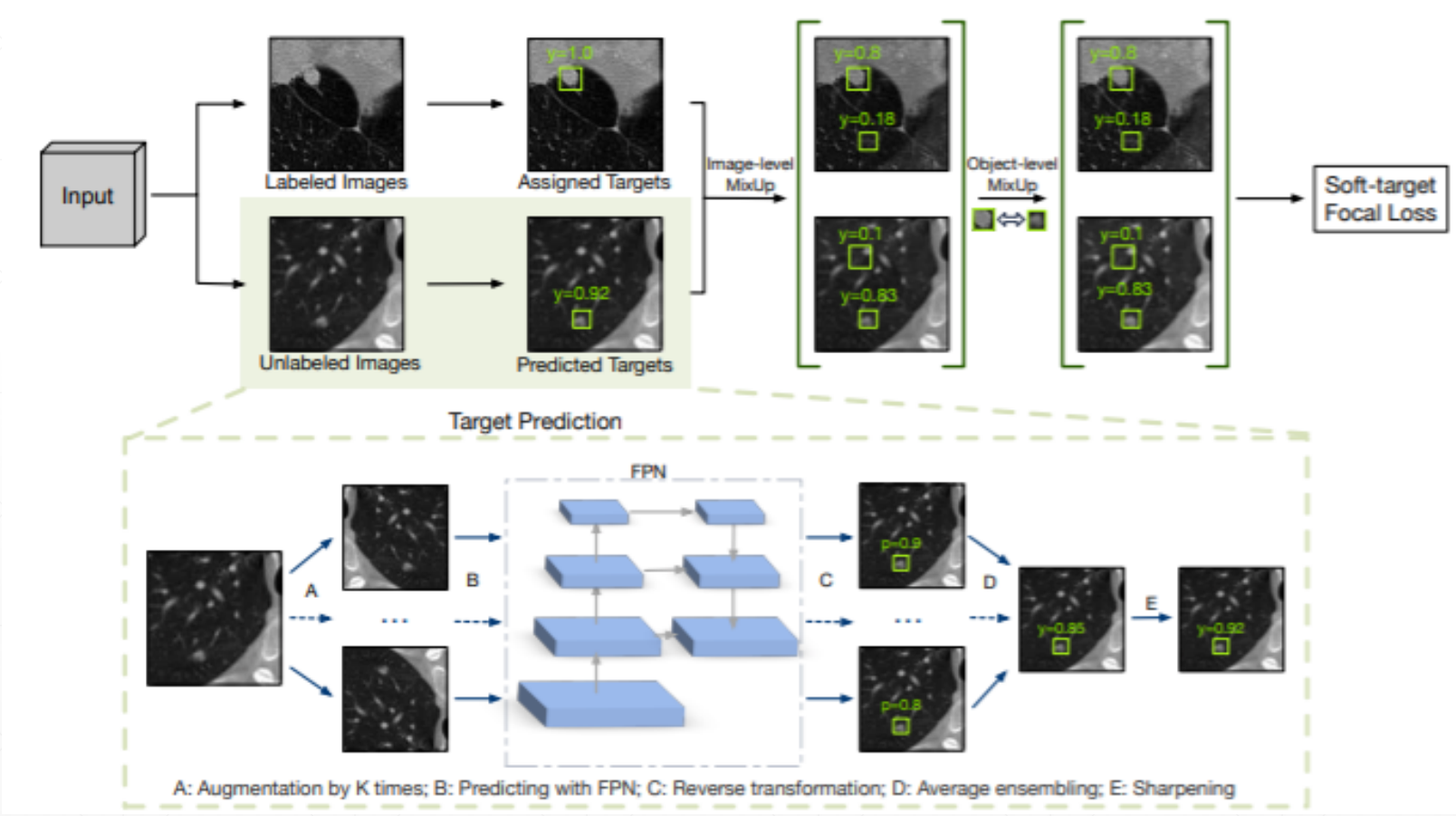
- x为输入的图片，y为对应的标签
- 进行了图片和目标两个层级的mix-up

$$\hat{x} = \tilde{\lambda}x + (1 - \tilde{\lambda})x', \tag{12}$$

$$\hat{y}_i = \tilde{\lambda}y_i + (1 - \tilde{\lambda})y'_i, \forall i. \tag{13}$$



Framework


Input
Labeled Images
Assigned Targets
Unlabeled Images
Predicted Targets
Image-level MixUp
Object-level MixUp
Soft-target Focal Loss
Target Prediction
FPN
A
B
C
D
E
A: Augmentation by K times; B: Predicting with FPN; C: Reverse transformation; D: Average ensembling; E: Sharpening

Experiment

- LUNA16是LIDC-IDRI数据集的高质量子集。共888个胸部CT扫描，1186个标注大于3毫米的结节。所有的注释都得到至少3(4个)放射科医师的同意。其他混淆的结节和非结节则标记为“无关发现”，在评估时既不计入假阳性，也不计入真阳性。

- NLST（国家肺筛查试验）最初是为了比较胸部CT和胸部X射线检查肺癌的有效性而建立的。NLST数据集中大约有75000次CT扫描，这些扫描具有参与者的特征、扫描测试结果、诊断程序等。由于结节位置等注释在该数据集中不可用，仅在第4.4节所述的选择过程后将其用作额外的未标记数据集。

Experiment

Labeled	Unlabeled	Recall(%) @ FPs							CPM(%)	Improv.
		0.125	0.25	0.5	1	2	4	8		
25	-	46.7	54.0	60.6	68.6	74.4	79.1	82.4	66.6	**11.5 (17.3%)**
25	400	57.6	64.5	74.6	80.5	87.0	90.1	92.1	**78.1**	
50	-	57.2	65.7	71.4	77.9	82.6	85.6	87.2	75.4	**6.6 (8.8%)**
50	400	64.1	71.0	78.7	85.2	89.3	92.3	93.5	**82.0**	
100	-	64.9	73.8	79.7	85.2	89.0	92.3	94.5	82.8	**4.4 (5.3%)**
100	400	73.4	80.9	84.8	88.6	92.3	94.7	96.1	**87.2**	

Experiment

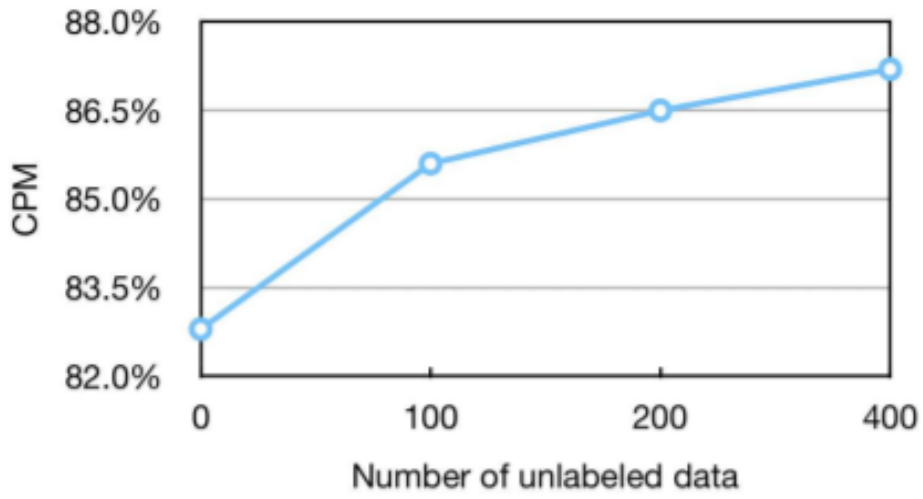


Figure 3: **Performance with different amounts of unlabeled data on LUNA16.** We use 100 labeled images.

Experiment

(a) Loss function.

Loss Function	CPM(%)
Supervised	82.8
SFL w/o soft α, β	Fail
SFL w/o soft α	84.4
SFL w/o soft β	83.7
SFL	**85.2**

(b) Augmentation times (K).

K	CPM(%)
1	85.9
2	86.3
4	**87.2**
8	87.1

(c) MixUp method.

MixUp Level		CPM(%)
Image	Object	
-	-	85.2
✓	-	86.7
✓	✓	**87.2**

Table 3: **Ablation study.** Models are trained with 100 labeled scans and 400 unlabeled ones. *Fail* denotes a divergent result.

FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling

Bowen Zhang[*]
Tokyo Institute of Technology
bowen.z.ab@m.titech.ac.jp

Yidong Wang[*]
Tokyo Institute of Technology
wang.y.ca@m.titech.ac.jp

Wenxin Hou
Microsoft
wenxinhou@microsoft.com

Hao Wu
Tokyo Institute of Technology
wu.h.aj@m.titech.ac.jp

Jindong Wang[†]
Microsoft Research Asia
jindwang@microsoft.com

Manabu Okumura[†]
Tokyo Institute of Technology
oku@pi.titech.ac.jp

Takahiro Shinozaki[†]
Tokyo Institute of Technology
shinot@ict.e.titech.ac.jp

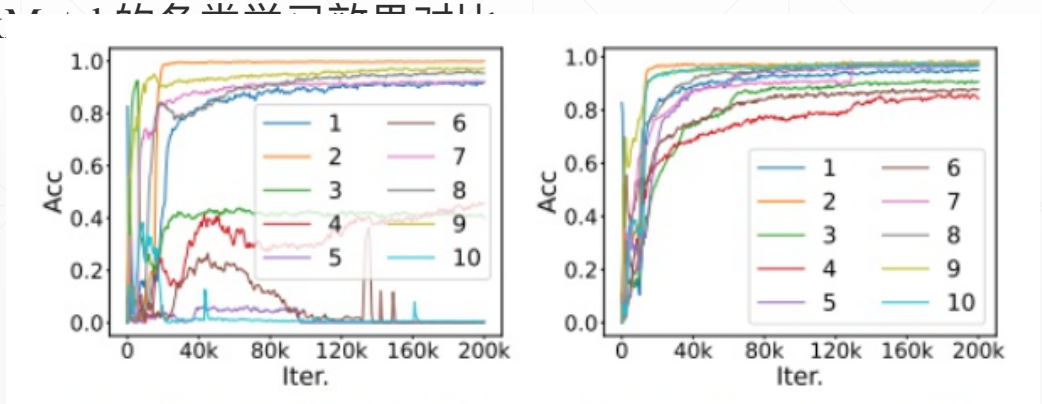
Background

置信度阈值

- 半监督学习(Semi-supervised Learning, SSL)一直受到研究者广泛的关注，因为它能高效地利用大量的未标注数据去提升模型性能。其中伪标签(Pseudo Labeling, PL)是一个很重要的技术。然而，随着模型训练而产生的伪标签往往伴随着大量错误标注，很多算法因此设定了一个高而固定的阈值，来选取那些置信度高的伪标签去计算无监督损失。高阈值可以有效地降低确认偏差(confirmation bias)，过滤有噪数据，因此目前最先进的半监督算法如UDA和FixMatch都用到了这样的阈值。

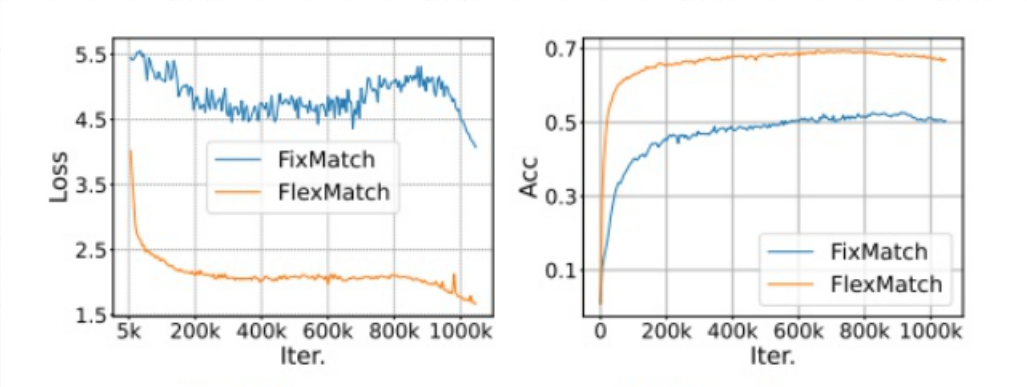
Background

- 然而本文作者提出这种固定的高阈值存在一定问题。
- 第一，对于分类任务而言，不同的类别的学习难度是不同的，模型在某一时刻对各类的学习情况也是不同的。学的比较好的类，或是简单的类，置信度自然会比较高，就更容易被固定阈值选取。而那些困难的类别，或是当下学的不是很好的类，由于置信度会偏低，就不容易被选到。这样就会导致模型有点“偏科”，比如一个孩子数学学得很好，家长又天天给他看数学书，于是他的数学分就越来越高。而语文本身学的就差，又很少去看语文书，导致语文分数一直提升不上来。表现到模型上就是：对困难类别的拟合不会很好，导致困难类别的最终精度不会很高。如图是FixMatch和Flex



Background

第二，在训练的起步阶段，受随机初始化影响，模型很可能把数据都**盲目地预测到一个类**里面去并且信心很高。如果一个batch中，只选出了这样错误的高信心伪标签，就会把模型往一个错误的方向优化。同时，即便一些样本的预测是正确的，由于开始阶段普遍置信度偏低，导致每个batch的**数据利用率不高**（大部分被过滤掉了），也会导致收敛很慢。如图是FixMatch和FlexMatch的收敛速度对比。



Motivation

为了解决第一个问题，作者引入了课程学习的思想，把单独的固定阈值转化成了**逐类的动态阈值，根据类别难度给每个类不同的阈值，且这些阈值可以随着模型的学习情况进行实时调整。**

针对第二个问题，作者引入了阈值的warm-up。其思想是，前期由于置信度不是很可靠，我们并不完全根据置信度来选样本，而是让**所有类的阈值逐渐从0开始上升**，给所有样本一个被学习的机会，等模型逐渐稳定获得辨识能力后再恢复到设计的动态阈值，其思想类似学习率的warm-up，因此叫threshold warm-up。

Methods

一个最简单的想法是通过类别准确率(class-wise accuracy)来确定。即：降低准确率更低的类的阈值，给这些类的数据更多被学习的机会，以让模型更好地拟合这些类。而对于准确率已经很高的类，就保持高阈值，以确保最终的精度。这是一个很理想的方法，但是却存在一些问题。首先，这种方式需要一个额外的有标签的验证集来评价各类的准确率，这在半监督学习下是一笔昂贵的开销，因为我们的标记数据已经很少了（本文实验中在最少的情况下每类只用了4个标记数据）。其次， 这种方式需要引入大量的额外计算，因为要想实时调整动态阈值，需要在每一步迭代后都做一个额外的前向传播来计算类别准确率。这会大幅降低算法速度。而CPL用了一种巧妙且简单的方法，使得既不需要额外验证集，也

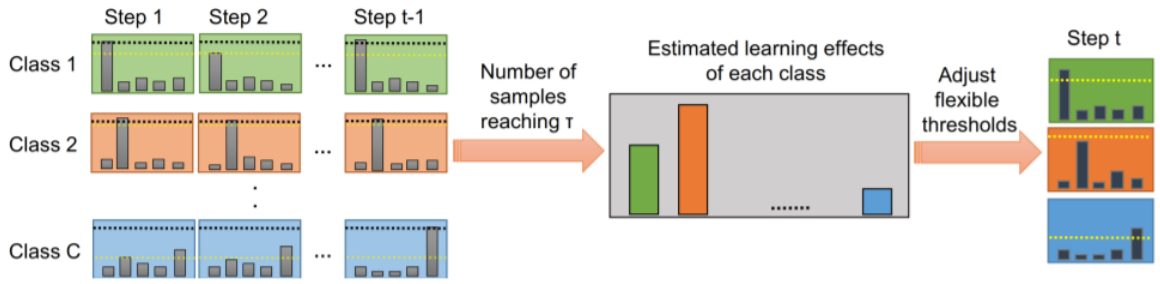


Figure 1: Illustration of Curriculum Pseudo Label (CPL). The estimated learning effects of each class are decided by the number of unlabeled data samples falling into this class and above the fixed threshold. They are then used to adjust the flexible thresholds to let pass the optimal unlabeled data. Note that the estimated learning effects do not always grow – they may also decrease if the predictions of the unlabeled data fall into other classes in later iterations.

Methods

Step1：学习效果预估。其实就是在所有样本中对高于固定阈值且属于某一类别的样本的一个计数。

$$\sigma_t(c) = \sum_{n=1}^{N} 1(\max(p_{m,t}(y|u_n)) > \tau) \cdot 1(\arg\max(p_{m,t}(y|u_n) = c).$$

Step2：归一化。 由于预估学习效果是对样本的一个计数，他的大小会随数据集包含样本数而变，因此需要对其进行归一化使其范围在0到1之间。注意这里归一化分母不是所有类的统计的求和，而是取所有类预估学习效果中的最大值。

$$\beta_t(c) = \frac{\sigma_t(c)}{\max_c \sigma_t}$$

Methods

Step3：确定阈值， 这里的公式其实已经可以作为最终的动态阈值了，然而作者又提出了两个tricks。

$$\mathcal{T}_t(c) = \beta_t(c) \cdot \tau.$$

阈值预热。 如前文所述，文中引入了阈值预热来解决前期高确认偏差的问题。

$$\beta_t(c) = \frac{\sigma_t(c)}{\max(\max_c \sigma_t, N - \sum_c \sigma_t)}$$

非线性映射。 相比于公式那样的直接scale固定阈值，非线性映射使得阈值的调整可以更加自由，你可以设计任意形状的函数来实现从“归一化预估学习效果 [公式] ”到“最终动态阈值 ”的映射

$$\mathcal{T}_t(c) = \mathcal{M}(\beta_t(c)) \cdot \tau.$$

Experiment

Dataset	CIFAR-10			CIFAR-100			STL-10			SVHN	
Label Amount	40	250	4000	400	2500	10000	40	250	1000	40	1000
PL	69.51±4.55	41.02±3.56	13.15±1.84	86.10±1.50	58.00±0.38	36.48±0.13	74.48±1.48	55.63±5.38	31.80±0.29	60.32±2.46	9.56±0.25
Flex-PL	**65.41**±1.35	**36.37**±1.57	**10.82**±0.04	**74.85**±1.53	**44.15**±0.19	**29.13**±0.26	**69.26**±0.60	**41.28**±0.46	**24.63**±0.14	**36.90**±1.19	**8.64**±0.08
UDA	7.33±2.03	5.11±0.07	4.20±0.12	44.99±2.28	27.59±0.24	22.09±0.19	37.31±3.03	12.07±1.50	6.65±0.25	4.40±2.31	**1.93**±0.01
Flex-UDA	**5.33**±0.13	**5.05**±0.02	**4.07**±0.06	**33.64**±0.92	**24.34**±0.20	**20.07**±0.13	**12.84**±2.60	**8.05**±0.21	**5.77**±0.08	**3.78**±1.67	1.97±0.06
FixMatch	6.78±0.50	4.95±0.07	4.09±0.02	46.76±0.79	28.15±0.81	22.47±0.66	35.42±6.43	10.49±1.03	6.20±0.20	**4.36**±2.16	**1.97**±0.03
FlexMatch	**4.99**±0.16	**4.80**±0.06	**3.95**±0.03	**32.44**±1.99	**23.85**±0.23	**19.92**±0.06	**10.87**±1.15	**7.71**±0.14	**5.56**±0.22	5.36±2.38	2.86±0.91
Fully-Supervised	4.45± 0.12			19.07± 0.18			-			2.14± 0.02	

谢谢观看！