Active Self-Paced Learning for Cost-Effective and Progressive Face Identification

Liang L, Wang K, Meng D, et al. [J]. **TPAMI**, 2017

目录 »

CONTENT

01

脸部识别分类

脸部识别分类

Dataset	# images	# persons	# images/person
CACD	56,138	500	79~306
CASIA-WebFace-Sub	181,901	925	100~804

将属于同一个人的不同光照，表情，脸部附件等情况下的照片归为一类，初始状态仅有小部分图片有标记。

特征提取：预训练好的深度卷积神经网络

多分类器：One-vs-All Linear SVM

02

整体框架

整体框架

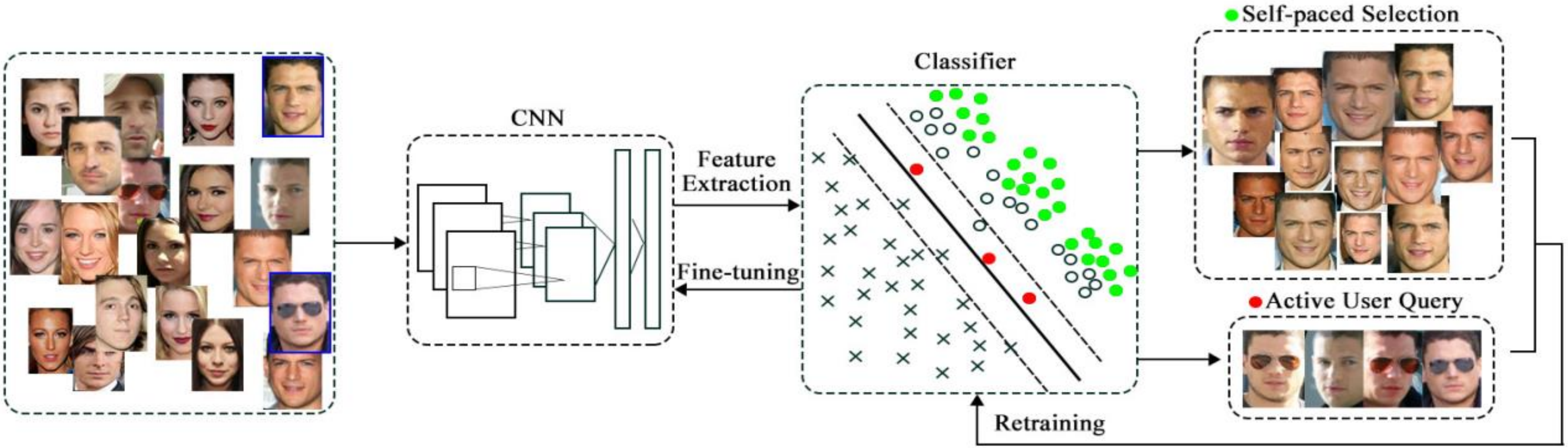


Fig. 2. Illustration of our proposed cost-effective framework. The pipeline includes stages of CNN and model initialization; classifier updating; high-confidence sample labeling by the SPL, low-confidence sample annotating by AL and CNN fine-tuning, where the arrows represent the workflow. The images highlighted by blue in the left panel represent the initially selected samples.

整体框架

Initialization

随机选择少量样本提取特征并标记

Classifier updating

第一轮仅使用少量标记样本训练初始分类器，之后通过SPL选择的高置信度的伪标记样本和AL查询的低置信度样本更新模型

High-confidence sample pseudo-labeling

根据SPL的权值（v）来对未标记样本进行排序（置信度高的权重大）并根据当前分类器的hinge loss来给样本标记

Low-confidence sample annotating

基于AL的Uncertainty策略对最不确定的K个进行查询

CNN fine-tuning

将SPL和AL选中的样本对CNN进行增量训练微调，使得其后轮的迭代中提取的特征更好

03

目标函数

目标函数

假设我们有共n张照片，拍自不同的m人，我们共有m个分类器（OVA SVM）来对每个样本进行分类训练数据 $\boldsymbol{D} = \{\boldsymbol{x}_i\}_{i=1}^{n} \subset \boldsymbol{R}^d$，对应的标记为 $\boldsymbol{y}_i = \{\boldsymbol{y}_i^{(j)} \in \{-\mathbf{1}, \mathbf{1}\}\}_{j=1}^{m}$，$\boldsymbol{y}_i^{(j)}$是样本$\boldsymbol{x}_i$是否属于第j个类的标记

$$\min_{\{\boxed{\mathbf{w},\mathbf{b}},\mathbf{v},\mathbf{y}_i \in \{-1,1\}^m, i \notin \Omega^{\boldsymbol{\lambda}}\}} \sum_{j=1}^{m} \frac{1}{2} \|\mathbf{w}^{(j)}\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right) \quad s.t. \quad \mathbf{v} \in \boldsymbol{\Psi}^{\boldsymbol{\lambda}}, \tag{2}$$

SVM模型参数

AL查询的样本

SPL的样本权值

$\mathbf{v} = \{[v_1^{(j)}, v_2^{(j)}, \cdots, v_n^{(j)}]^T\}_{j=1}^{m}$ denotes the weight variables reflecting the training samples' importance,

目标函数

ASPL

$$\min_{\{\mathbf{w},\mathbf{b},\mathbf{v},\mathbf{y}_i\in\{-1,1\}^m, i\notin\Omega^{\boldsymbol{\lambda}}\}} \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{w}^{(j)}\|_2^2 + \tag{2}$$

$$\underbrace{C \cdot L\left(\mathbf{w}^{(j)}, b^{(j)}, \mathcal{D}, \mathbf{y}^{(j)}, \mathbf{v}^{(j)}\right)} + f\left(\mathbf{v}^{(j)}; \lambda_j\right)$$

$$s.t. \quad \boxed{\mathbf{v} \in \mathbf{\Psi}^{\boldsymbol{\lambda}},}$$

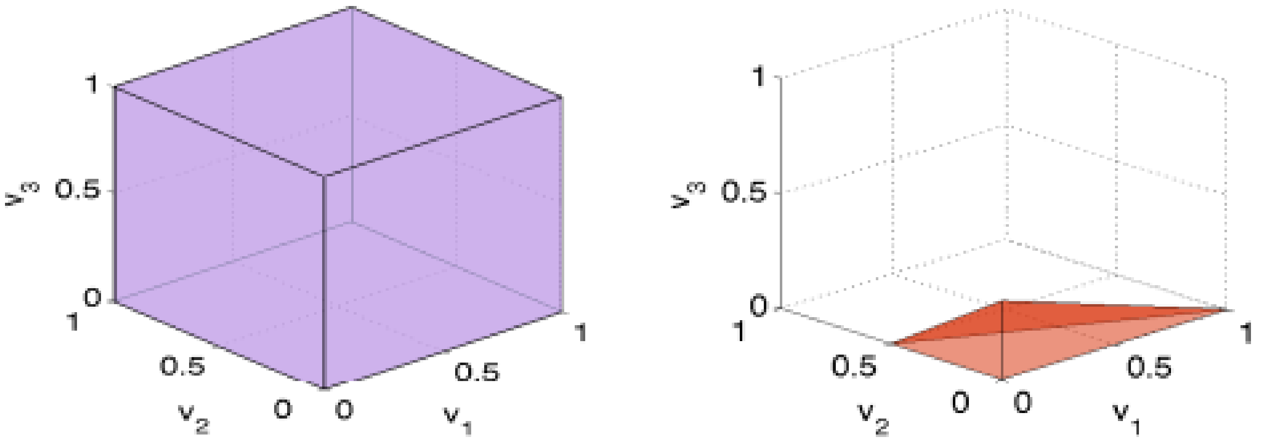
$$= \sum_{i=1}^{n} v_i^{(j)} \left(1 - y_i^{(j)}(\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)})\right)_+$$

$$\text{s.t.} \sum_{j=1}^{m} |y_i^{(j)} + 1| \leq 2, y_i^{(j)} \in \{-1, 1\}, i \notin \Omega^{\lambda},$$

SPL

$$(\mathbf{w}_{t+1}, \mathbf{v}_{t+1}) = \underset{\mathbf{w}\in\mathbb{R}^d, \mathbf{v}\in\{0,1\}^n}{\operatorname{argmin}} \left(r(\mathbf{w}) + \sum_{i=1}^{n} v_i f(\mathbf{x}_i, \mathbf{y}_i; \mathbf{w}) - \frac{1}{K}\sum_{i=1}^{n} v_i \right)$$

Without considering the learning objective, we can see that v1 tends to be learned earlier than v2 and v3. This is because if we uniformly sample sufficient points in the feasible region of the coordinate (v1, v2, v3), the expecte



(a) SPL (b) SPCL

Figure 1: Comparison of feasible regions in SPL and SPCL.

$$\min_{\mathbf{w},\mathbf{v}\in[0,1]^n} \mathbb{E}(\mathbf{w},\mathbf{v};\lambda,\Psi)=\sum_{i=1}^{n} v_i L(y_i,g(\mathbf{x}_i,\mathbf{w}))+f(\mathbf{v};\lambda) \quad (3)$$
$$\text{s.t. } \mathbf{v}\in\Psi$$

Jiang L, Meng D, Zhao Q, et al. Self-paced Curriculum Learning[C]// AAAI Conference on Artificial Intelligence. 2015.

- $\Psi_i^{\boldsymbol{\lambda}} = [0,1]$ is for the pseudo-labeled sample, i.e., $i \notin \Omega^{\boldsymbol{\lambda}}$. Then, its importance weights with respect to all the classifiers $\{v_i^{(j)}\}_{j=1}^m$ need to be learned in the SPL optimization.
- $\Psi_i^{\boldsymbol{\lambda}} = \{1\}$ is for the sample annotated by the AL process, i.e., $\exists j \quad s.t. \quad i \in \Omega^{\lambda_j}$. Thus, its importance weights are deterministically set during the model training, i.e., $v_i^{(j)} = 1$.

Classifier Updating

$$\min_{\{\mathbf{w},\mathbf{b},\mathbf{v},\mathbf{y}_i \in \{-1,1\}^m, i \notin \Omega^{\boldsymbol{\lambda}}\}} \sum_{j=1}^{m} \frac{1}{2} \|\mathbf{w}^{(j)}\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right) \quad s.t. \quad \mathbf{v} \in \boldsymbol{\Psi}^{\boldsymbol{\lambda}}, \tag{2}$$

固定$\{\{x_i\}_{i=1}^n, v, \{y_i\}_{i=1}^n, \Psi^\lambda\}$，原公式退化为：

$$\min_{\mathbf{w},\mathbf{b}} \sum_{j=1}^{m} \frac{1}{2} \|\mathbf{w}^{(j)}\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right)$$

这是一个标准的带权SVM优化问题

High-confidence Sample Labeling:

$$\min_{\{\mathbf{w},\mathbf{b},\mathbf{v},\mathbf{y}_i\in\{-1,1\}^m, i\notin\Omega^{\boldsymbol{\lambda}}\}} \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{w}^{(j)}\|_2^2 + C\sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right) \quad s.t. \quad \mathbf{v}\in\boldsymbol{\Psi}^{\boldsymbol{\lambda}}, \tag{2}$$

固定$\{\boldsymbol{w}, \boldsymbol{b}, \{x_i\}_{i=1}^{n}, \{y_i\}_{i=1}^{n}, \boldsymbol{\Psi}^{\lambda}\}$，可通过以下目标来优化v（未标记样本）：

$$\min_{\mathbf{v}\in[0,1]} \sum_{j=1}^{m} C\sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right), \quad s.t. \quad \mathbf{v}\in\boldsymbol{\Psi}^{\boldsymbol{\lambda}}. \tag{5}$$

这是一个凸优化问题，可通过内点法，梯度法等解决

High-confidence Sample Labeling:

$$\min_{\{\mathbf{w},\mathbf{b},\mathbf{v},\mathbf{y}_i\in\{-1,1\}^m,i\notin\Omega^{\boldsymbol{\lambda}}\}} \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{w}^{(j)}\|_2^2 + C\sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right) \quad s.t. \quad \mathbf{v} \in \boldsymbol{\Psi}^{\boldsymbol{\lambda}}, \tag{2}$$

固定$\{\boldsymbol{w}, \boldsymbol{b}, \{x_i\}_{i=1}^{n}, \boldsymbol{v}, \boldsymbol{\Psi}^{\lambda}\}$，可通过以下目标来优化未标记样本的 $\boldsymbol{y}_i$：

$$\min_{\mathbf{y}_i\in\{-1,1\}^m,i\in\mathcal{S}} \sum_{i=1}^{n}\sum_{j=1}^{m} v_i^{(j)} \ell_{ij}$$

$$\text{s.t.}, \sum_{j=1}^{m} |y_i^{(j)} + 1| \leq 2.$$

Low-confidence Sample Annotating

1.Annotated Sample Verifying:
对已有标记的样本再取最不确定的5个通过AL进行二次确认（抗噪声）

2.Low-confidence Definition:
当m个分类器中有多个分类器的输出为1时，认为该样本是难以区分的，以这个为标准来选择AL查询的样本
We employed this simple strategy just due to its intuitive rationality and efficiency.

Low-confidence Sample Annotating

将AL查到的样本更新，并加入课程Psi中

- $\Psi_i^{\lambda} = [0, 1]$ is for the pseudo-labeled sample,
- $\Psi_i^{\lambda} = \{1\}$ is for the sample annotated by the AL

For those annotated samples, the corresponding $\Psi_i^{\lambda} = \{1\}$ with expectation value 1 over the whole set, while for others $\Psi_i^{\lambda} = [0, 1]$ with expectation value 1/2. Thus the more informative samples still have a larger expectation than the others.

Feature Representation Updating

将SPL和AL选中的样本对CNN进行微调，使得之后提取的特征更有效，对所有的层的学习率固定为0.001

Pace Parameter lamda Updating

$$\lambda_j^t = \begin{cases} \lambda^0, & t = 0 \\ \lambda_j^{(t-1)} + \alpha * \eta_j^t, & 1 \le t \le \tau \\ \lambda_j^{(t-1)}, & t > \tau, \end{cases} \tag{10}$$

we empirically set $\{\lambda^0, \alpha\} = \{0.2, 0.08\}$.

τ is set as 12

Algorithm 1: The sketch of ASPL framework

Input: Input dataset $\{\mathbf{x}_i\}_{i=1}^{n}$

Output: Model parameters $\mathbf{w}$, $\mathbf{b}$

1: Use pre-trained CNN to extract feature representations of $\{\mathbf{x}_i\}_{i=1}^{n}$. Initialize multiple annotated samples into the curriculum $\boldsymbol{\Psi}^{\boldsymbol{\lambda}}$ and corresponding $\{\mathbf{y}_i\}_{i=1}^{n}$ and $\mathbf{v}$. Set an initial pace parameter $\boldsymbol{\lambda} = \{\lambda^0\}^m$.

while not converged **do**

2: Update $\mathbf{w}, \mathbf{b}$ by one-vs-all SVM

3: Update $\mathbf{v}$ by the SPL via Eqn. (7)

4: Pseudo-label high-confidence samples $\{\mathbf{y}_i\}_{i\in S}$ by the reranking via Eqn. (8)

5: Update the unclear class set U

6: Verify the annotated samples by AL.

7: Update low-confidence samples $\{\mathbf{y}_i, \Psi_i^{\lambda}\}_{i\in\phi}$ by the AL

if u unseen classes have labeled,

Handle u new classes via the steps in Sect. 4.1

Go to the step 2

end if

8: In every T iterations:

- Update $\{\mathbf{x}_i\}_{i=1}^{n}$ through fine-tuning CNN
- Update $\boldsymbol{\lambda}$ according to Eqn. (10)

9: **end while**

10: **return** $\mathbf{w}, \mathbf{b}$;

04

实验验证

Dataset

TABLE 1
The summarization of datasets we used.

Dataset	# images	# persons	# images/person
CACD	56,138	500	79~306
CASIA-WebFace-Sub	181,901	925	100~804

CACD:
它包含一批从互联网收集的2000名名人的图像，这些名人年龄，姿势，照明和遮挡都不一样。 Chen等人手动注释了200名名人中的一小部分[47]。为了更好地证明实验效果，我们额外标记了300个名人来增加该子集，并获得一组总共56138张图像。

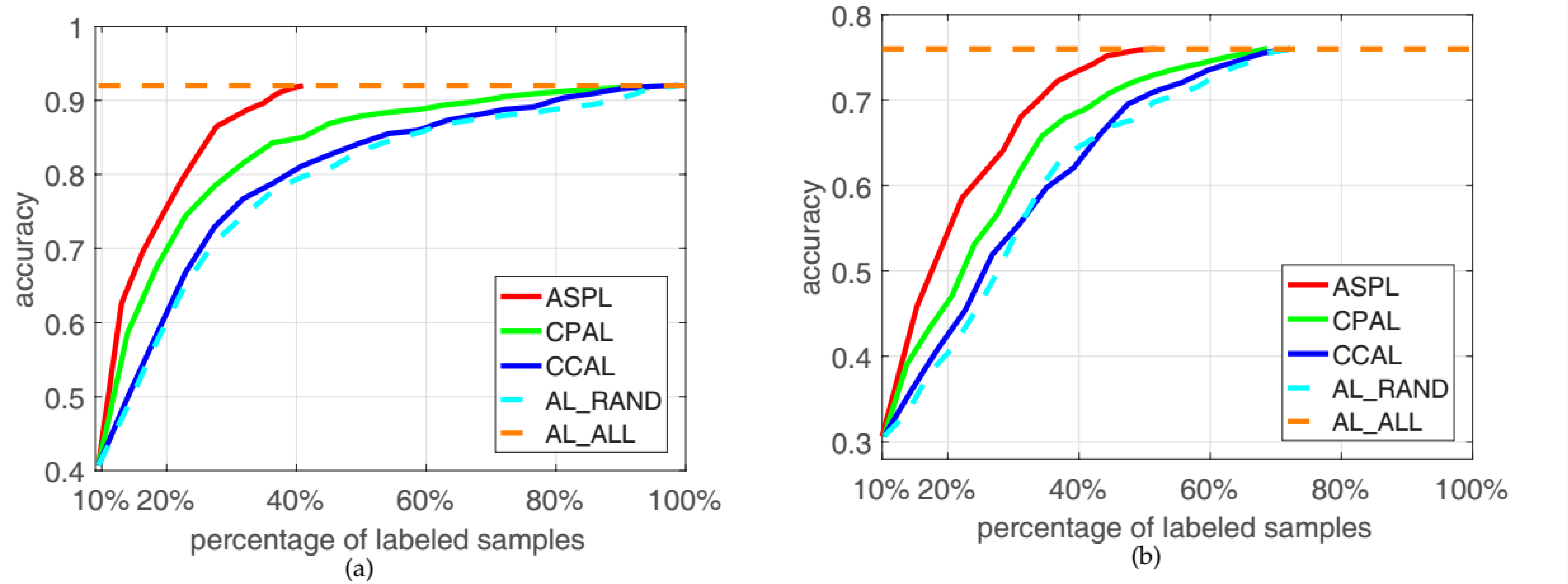
CASIA-WebFace dataset:
拥有10,575个人，494,414个图像。 CASIA-WebFace是非常具有挑战性的，其图像都是从互联网收集的，具有不同视角和不同场景下的光照。 每个人的样本数量从3到804不等，这是非常不平衡的。 对于非常少的样本（低于100）的人，不能进行实验分析。 因此，我们丢弃少于100个样本的人像，以形成CASIA-WebFace-Sub数据集。 CASIA-WebFace-Sub数据集共有181,901张图片，925人。

Experiment setting

做了5次不同的初始化的实验并报告其平均结果，对比模型：

i）CPAL（Convex Programming based Active Learning）：基于预测不确定性和样本多样性，在每轮迭代中选择一批样本；
ii）CCAL（Confidence-based Active Learning via SVMs）：仅选择**一个**具有最低预测置信度的样本；
iii）AL_RAND：在训练阶段随机选择未标记的样本进行标注。该方法丢弃了所有主动学习技术，可以被认为是下界
iv）AL_ALL：所有未标记的样本都被标注用于训练分类器。这种方法可以被认为是上限（分类器可以实现的最佳性能）。

Experiment



SPL选择的高置信度伪标记样本能显著降低人工标注

Fig. 3. Results on (a) CACD and (b) CASIA-WebFace-Sub datasets. The vertical axes represent the recognition accuracy and the horizontal axes represent the percentage of annotated samples of the whole set.

各组成部分功能分析:

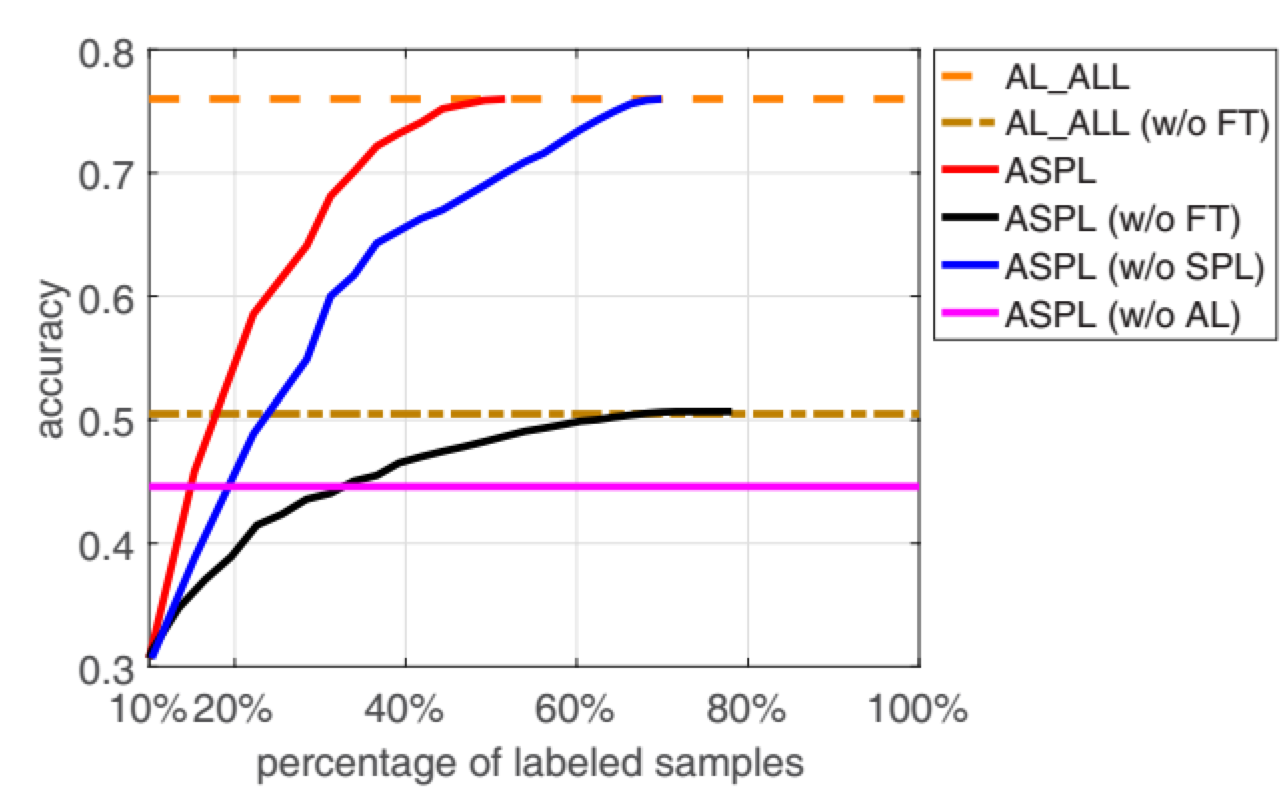
i）ASPL（w/o FT）：禁用CNN微调

ii）ASPL（w/o SPL）：丢弃高信度样本伪标签;

iii）ASPL（w/o AL）：忽略主动标注的低置信度样本;

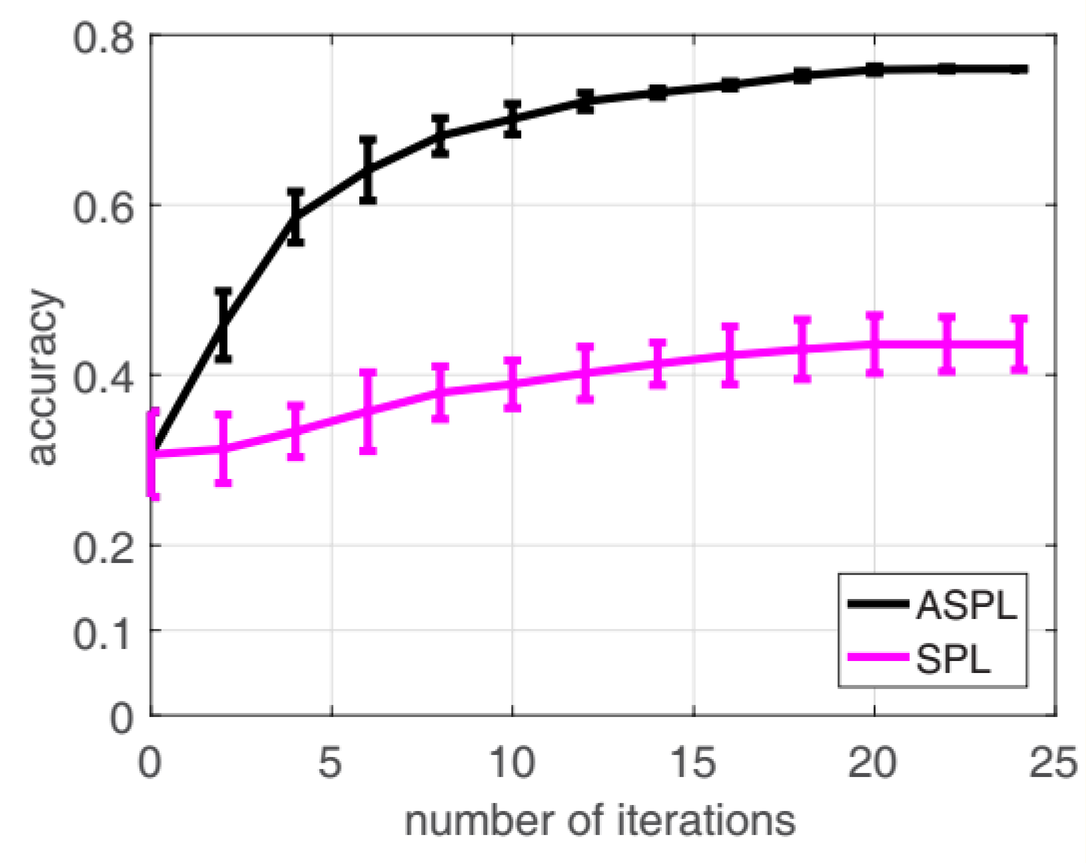
iv）AL_ALL：所有未标记样本都有标记但禁用CNN微调

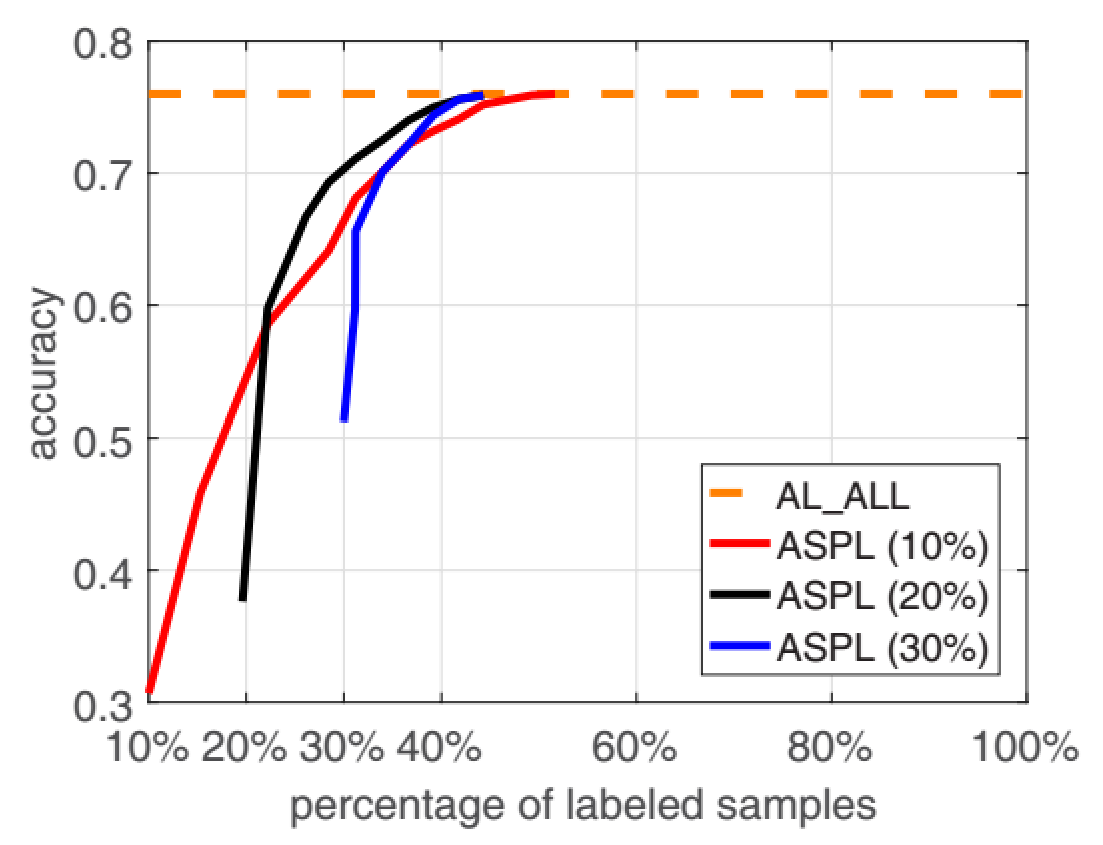
v）AL_ALL（w/o FT）：所有未标记样本都有标记但禁用CNN微调



初始化的影响

SPL的效果会强烈地受到初始化的影响，通过加入AL可以减轻这部分影响。





在不同的初始样本，ASPL在后续迭代能达到稳定的结果，而SPL仍然变化很大。 该结果表明AL组件在处理较差的初始化时有效。

尽管一开始表现不佳，但是在训练过程中，模型的表现会被AL带向正确的方向。 也就是模型对初始样本的多样性和数量不敏感。

相同的迭代轮数所标记的样本数

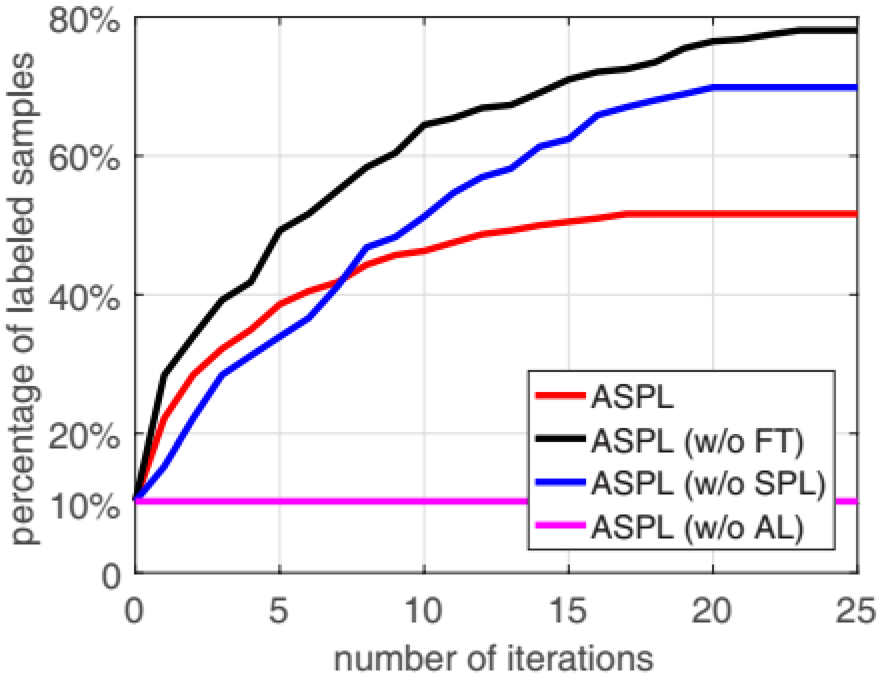


Fig. 7. The comparison of different number of initial samples and the further required annotation ported of the AL process on the CASIA-WebFace-Sub dataset.

当训练迭代次数小时，ASPL模型需要相对较多的标注。随着训练的继续，由于模型逐渐成熟，所以需要标注的数量开始减少。该观察结果表明，当分类器在后期学习阶段变得可靠时，用户标注的负担将减轻。

伪标记性能

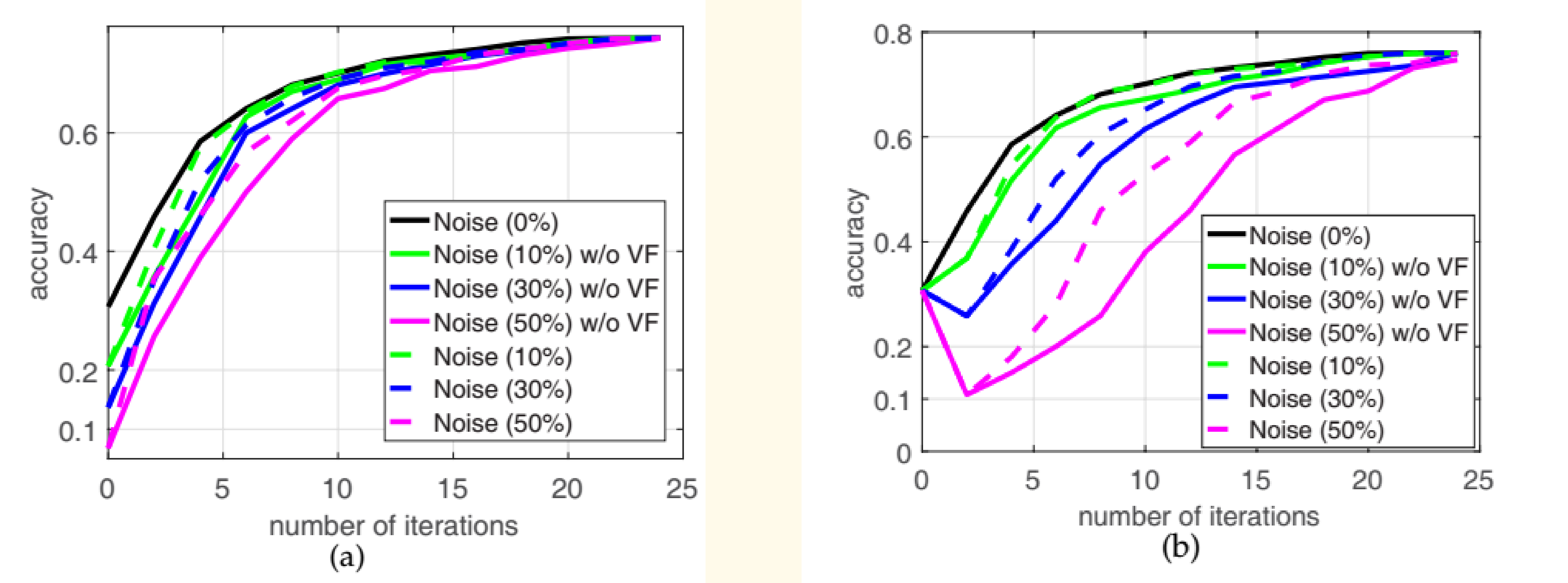
TABLE 3
The error rates of the pseudo-labels assigned by SPL on high-confidence samples.

# iteration	5	10	15	20	25
ASPL (w/o FT)	8.2%	6.9%	5.1%	5.0%	4.9%
ASPL	4.5%	4.1%	3.4%	3.3%	3.3%

随着用户标注的增加，ASPL可以自动为未标记样本分配更可靠的伪标签。

（i）Ex-1：a（a = 0%; 10%; 30%; 50%）噪声图像被添加到每个人的初始样本中。

（ii）Ex-2：初始无噪声，但b（b = 0%; 10%; 30%; 50%）在AL过程中故意加入噪声。



ASPL可以以较快的速度从噪声图像中恢复。这证明了所提出的标注样本验证的有效性

Fig. 8. Robust analysis of ASPL under two types of noisy samples. (a) Using different number of noisy samples as the initial annotation. (b) Adding different number of noisy samples at the 10-th step (denoted by the black spots).

Why effective?

对AL来说，SPL能在模型成熟后大大降低人工标注的负担

对SPL来说，通过AL用户标注低置信度样本，使得SPL可用于较少的初始样本且能够减轻初始化的影响

重心在如何降低标注cost,而不是性能

Thank you