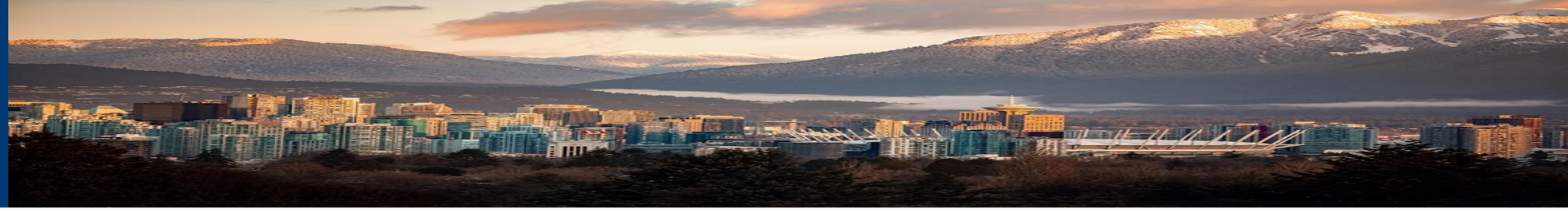
The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA



Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

Yue Duan[1], Zhen Zhao[2], Lei Qi[3], Luping Zhou[2] , Lei Wang[4], and Yinghuan Shi[1]

[1] *Nanjing University, China* [2] *University of Sydney, Australia* [3] *Southeast University, China* [4] *University of Wollongong, Australia*

Arxiv: https://arxiv.org/abs/2312.12237

Github: https://github.com/NJUyued/SoC4SS-FGVC

Homepage: https://njuyued.github.io/

南京大学 NANJING UNIVERSITY

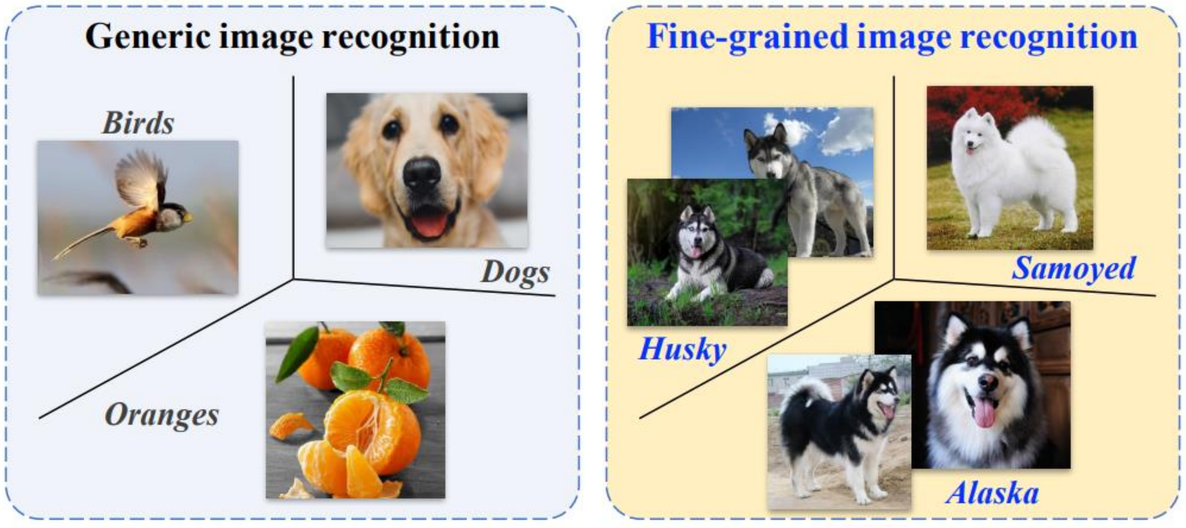
The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

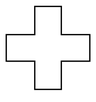
Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

NANJING UNIVERSITY

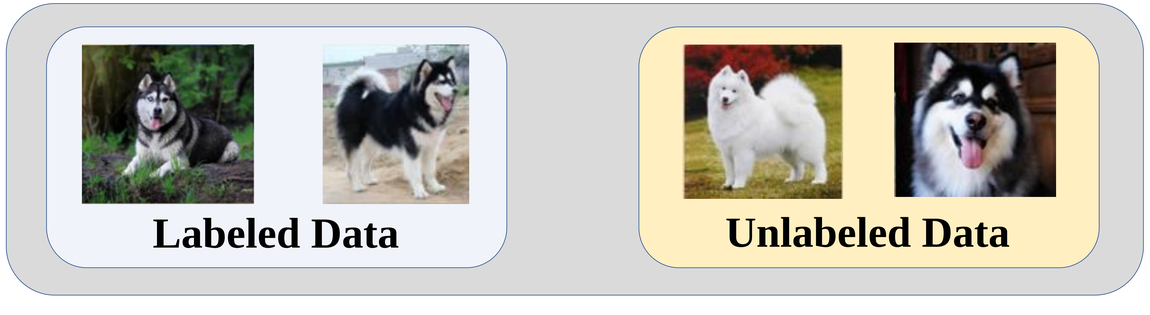
Introduction

Generic image recognition V.S. Fine-grained image recognition (FGVC)





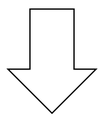
Semi-supervised Learning



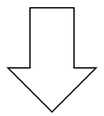
Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

Introduction

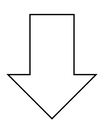
Semi-supervised learning (SSL) aims to leverage a pool of unlabeled data to alleviate the dependence of deep models on labeled data.



However, current SSL approaches achieve promising performance with clean and ordinary data, *but become unstuck against the indiscernible unlabeled data*.



A typical and worth discussing example is the *semi-supervised fine-grained visual classification (SS-FGVC)* [1].



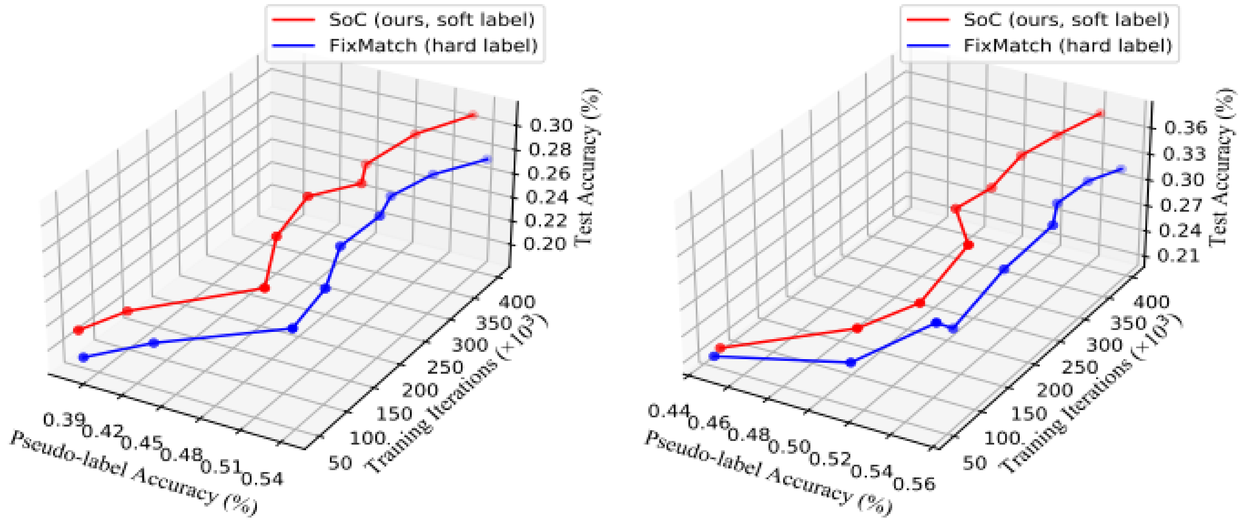
Fine-grained data may severely affect the quality of pseudo-labels and consequently pull down the model performance.

[1] Su, J.-C.; Cheng, Z.; and Maji, S. 2021. A realistic evaluation of semi-supervised learning for fine-grained classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

南京大學 NANJING UNIVERSITY



(a) Semi-Aves (b) Semi-Fungi

Motivation

The bad effects of incorrect hard label on the model have overridden its low entropy advantage while *soft label can benefit the model because it could still provide useful information although it is wrong.*

⇩

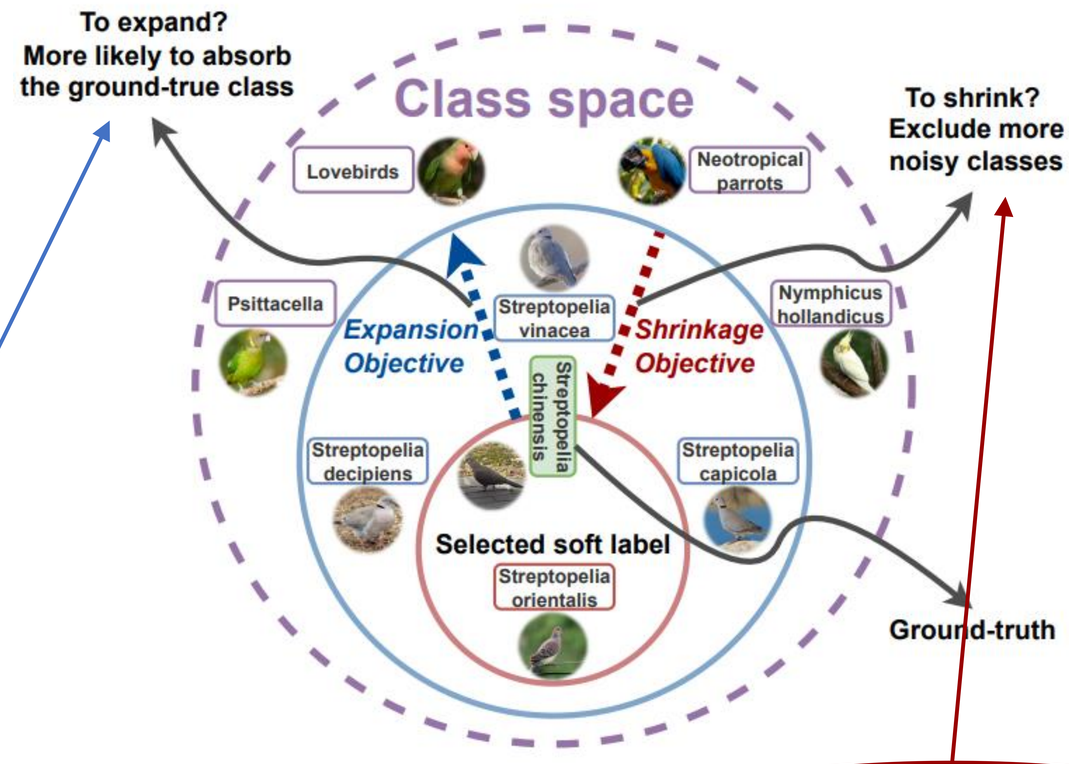
Reasonably utilizing soft labels will enable the model to be more composed when dealing with fine-grained data that is difficult to distinguish.

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

南京大學 NANJING UNIVERSITY

A Coupled Optimization Goal for SS-FGVC



Objective 1 (Expansion Objective). *Encourage the pseudo-label to ~~contain the ground-truth~~ class as much as possible:*

$$\max_{\theta} \mathbb{E}_{\mathbf{x}_i^{\mathtt{ulb}} \in \mathcal{D}^{\mathtt{ulb}}} \left[\mathbb{1}(y_i^\star \in \{c \mid g_{i,(c)} = 1\}) p_{i,(y_i^\star)} \right], \quad (5)$$

where $y_i^\star \in \mathcal{Y}$ is the ground-truth label of $\mathbf{x}_i^{\mathtt{ulb}}$. For simplicity, we denote $p_{i,(y_i^\star)} \mathbb{1}(y_i^\star \in \{c \mid g_{i,(c)} = 1\})$ as $z_i^{\mathtt{obj1}}$.

Objective 2 (Shrinkage Objective). *Encourage the pseudo-label to ~~contain as few~~ classes as possible:*

$$\min_{\theta} \mathbb{E}_{\mathbf{x}_i^{\mathtt{ulb}} \in \mathcal{D}^{\mathtt{ulb}}} \sum_{c=1}^{K} g_{i,(c)}, \quad (6)$$

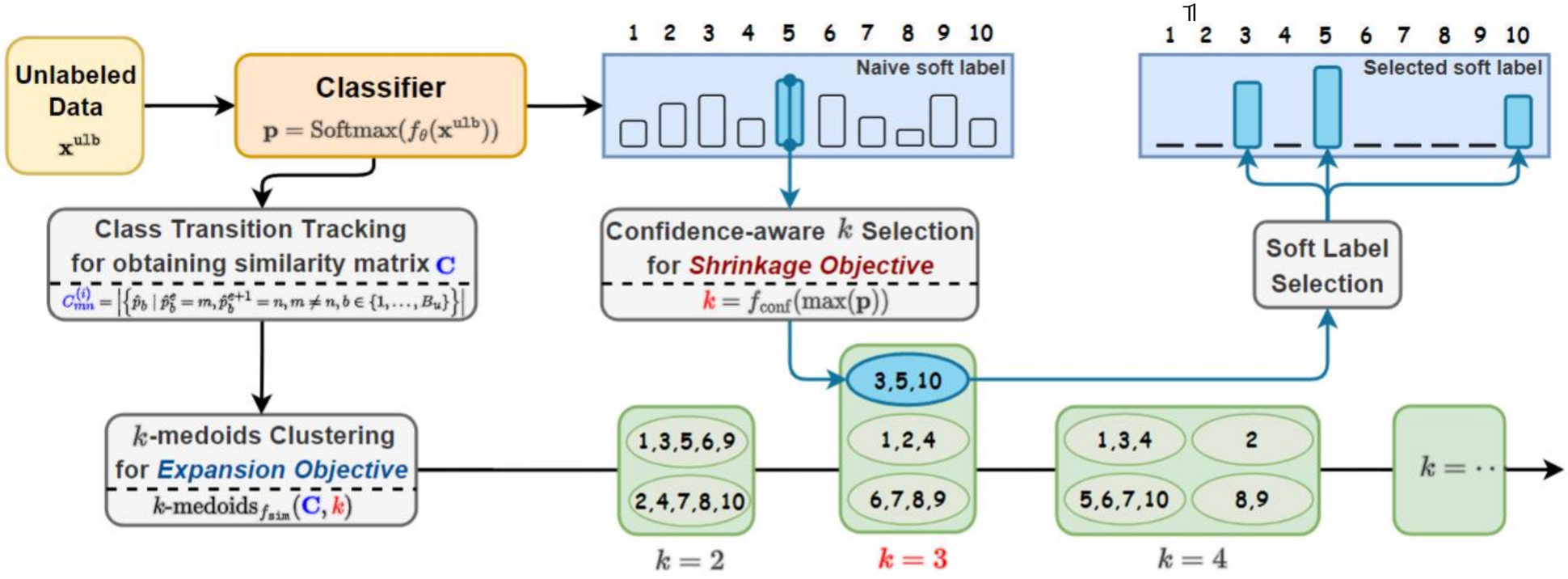
where $\sum_{c=1}^{K} g_{i,(c)}$ is denoted as $z_i^{\mathtt{obj2}}$ for simplicity.

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

南京大學 NANJING UNIVERSITY

Overview of Method



- For **Expansion Objective**, use class-level similarity obtained through class transfer tracking to acquire **candidate class clusters**.

- For **Shrinkage Objective**, select a more appropriate **clustering granularity** based on prediction confidence.

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

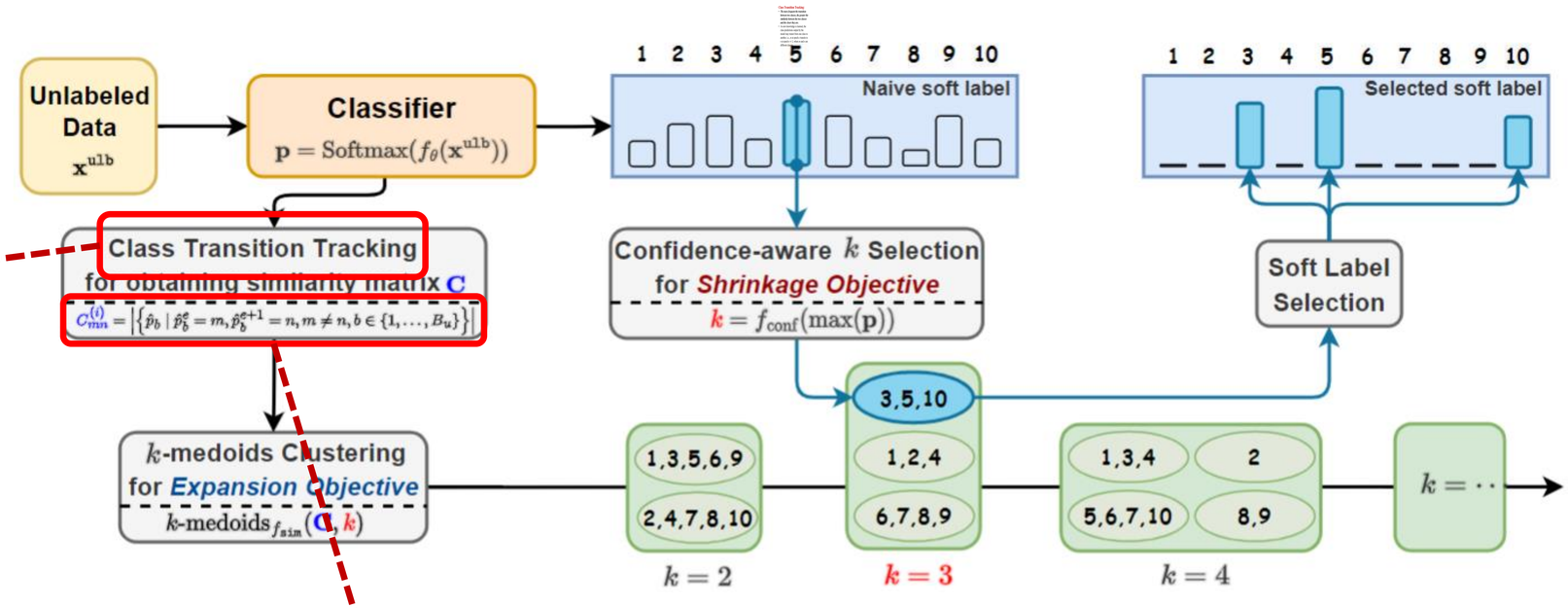
NANJING UNIVERSITY

Method

- For **Expansion Objective**, use class-level similarity obtained through class transfer tracking to acquire **candidate class clusters**.

Class Transition Tracking:

- ***The more frequent the transition between two classes, the greater the similarity between the two classes and the closer they are.***
- As new knowledge is learned, the class predictions output by the model may transit from one class to another, i.e., *m* at epoch *e* transits to *n* at epoch $e+1$, where *m* and *n* are different classes.



$$C_{mn}^{(i)} = \left|\left\{\hat{p}_b \mid \hat{p}_b^e = m, \hat{p}_b^{e+1} = n, m \neq n, b \in \{1, \ldots, B_u\}\right\}\right|$$

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

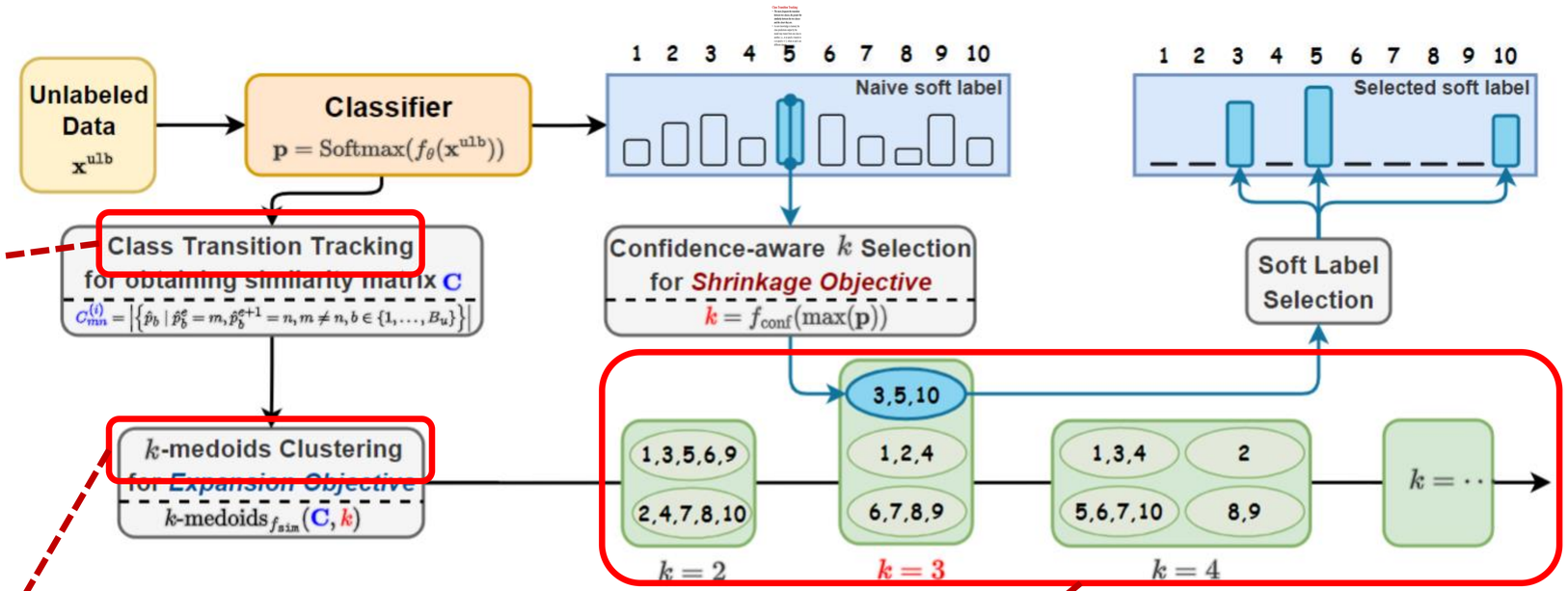
南京大學 NANJING UNIVERSITY

Method

- For **Expansion Objective**, use class-level similarity obtained through class transfer tracking to acquire **candidate class clusters**.

Class Transition Tracking:

- ***The more frequent the transition between two classes, the greater the similarity between the two classes and the closer they are.***
- As new knowledge is learned, the class predictions output by the model may transit from one class to another, i.e., *m* at epoch *e* transits to *n* at epoch $e+1$, where *m* and *n* are different classes.



Clustering using K-medoids

$$\mathcal{S} = k\text{-medoids}_{f_{\mathrm{sim}}}(\mathbf{C}, k)$$

Obtain candidate clusters

$$\mathcal{C}_i = S_s, s = \underset{c \in \{1,\ldots,k\}}{\arg\max}\ \mathbb{1}(\hat{p}_i \in S_c)$$

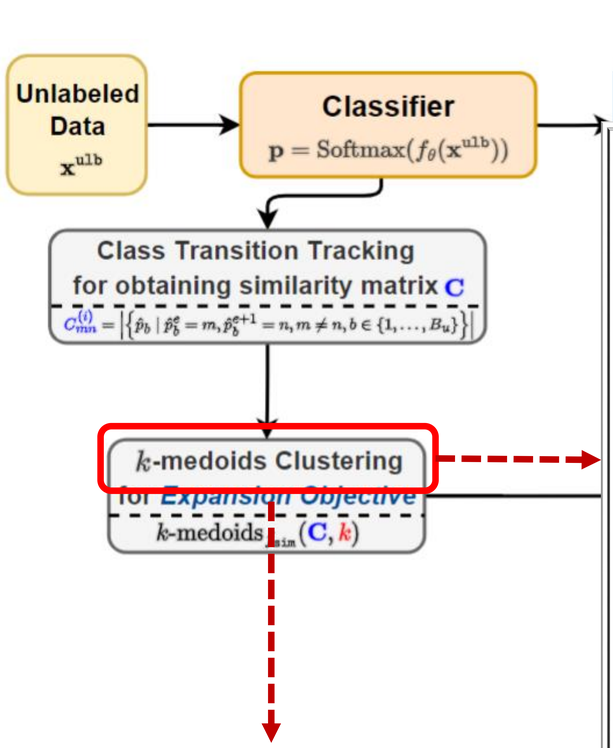
The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

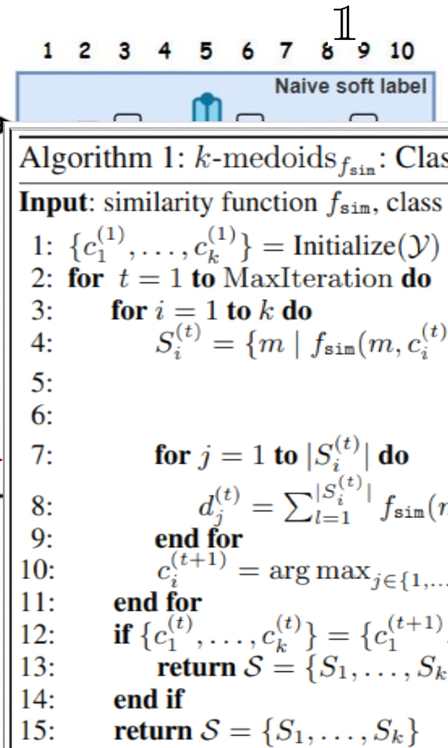
NANJING UNIVERSITY

Method

- For **Expansion Objective**, use class-level similarity obtained through class transfer tracking to acquire **candidate class clusters**.



Clustering using K-medoids



Algorithm 1: k-medoids$_{f_{sim}}$: Class Transition Tracking based k-medoids Clustering

Input: similarity function f_{sim}, class transition matrix $\mathbf{C}$, cluster number k

$\{c_1^{(1)}, \ldots, c_k^{(1)}\} = \mathrm{Initialize}(\mathcal{Y})$ ▷ Randomly select centroids over the class space
for $t = 1$ **to** MaxIteration **do**
  for $i = 1$ **to** k **do**
    $S_i^{(t)} = \{m \mid f_{sim}(m, c_i^{(t)}) \geq f_{sim}(m, c_j^{(t)}), 1 \leq j \leq k\}$ ▷ Assignment step
    ▷ $f_{sim}(m, n) = \frac{C_{mn}+C_{nm}}{2}$ is the similarity between class m and n
    ▷ The larger the $f_{sim}(m, n)$, the more similar the classes m and n are
  for $j = 1$ **to** $|S_i^{(t)}|$ **do**
    $d_j^{(t)} = \sum_{l=1}^{|S_i^{(t)}|} f_{sim}(m_j, m_l)$ ▷ $m_j, m_l \in S_i^{(t)}$
  end for
  $c_i^{(t+1)} = \arg\max_{j \in \{1,\ldots,k\}}(d_j^{(t)})$ ▷ Update medoids step
end for
if $\{c_1^{(t)}, \ldots, c_k^{(t)}\} = \{c_1^{(t+1)}, \ldots, c_k^{(t+1)}\}$ **then**
  return $\mathcal{S} = \{S_1, \ldots, S_k\}$
end if
return $\mathcal{S} = \{S_1, \ldots, S_k\}$
end for

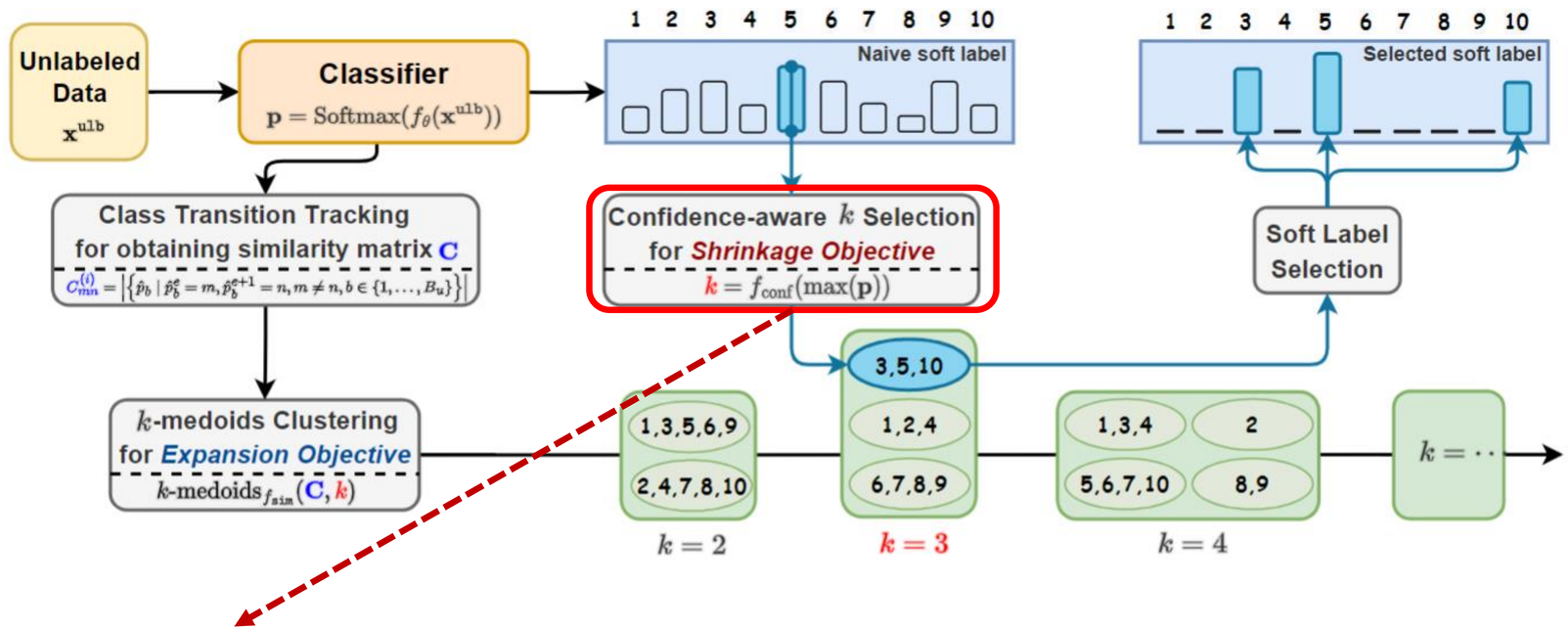
$$\mathcal{S} = k\text{-medoids}_{f_{sim}}(\mathbf{C}, k)$$

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

南京大学 NANJING UNIVERSITY

Method

- For **Shrinkage Objective**, select a more appropriate **clustering granularity** based on prediction confidence.



The number of clusters is selected based on confidence, i.e., the granularity of the clustering.

$$k_i = f_{\mathrm{conf}}(\max(\mathbf{p}_i)),$$

In this implementation, a simple and effective linear function is used.

$$k_i = \lceil (\frac{\max(\mathbf{p}_i)}{\alpha} + \frac{2}{K}) \times K - \frac{1}{2} \rceil,$$

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

Method

- **Putting It All Together**

$$\mathcal{L}_{\text{sup}} = \frac{1}{B}\sum_{n=1}^{B} H(\mathbf{y}_i, f_\theta(\text{Aug}_\text{w}(\mathbf{x}_i^{\text{ulb}}))),$$

Step 1. The soft pseudo-label for weakly-augmented $\mathbf{x}_i^{\text{ulb}}$ is computed as $\mathbf{p}_i^{\text{w}} = \text{Softmax}(f_\theta(\text{Aug}_\text{w}(\mathbf{x}_i^{\text{ulb}})))$. With obtained $\mathbf{p}_i^{\text{w}}$, the k_i^{w} for clustering is selected by Eq. (13).
Step 2. Class transition matrix $\mathbf{C}$ is calculated by Eq. (7) in the current iteration. By Eq. (8), the CTT based k-medoids clustering is performed with k_i^{w} and $\mathbf{C}$ to obtain $\mathcal{C}_i^{\text{w}}$.
Step 3. $\mathbf{g}_i^{\text{w}}$ is computed by Eq. (3) with $\mathcal{C}_i^{\text{w}}$ and then the selected pseudo-label $\tilde{\mathbf{p}}_i^{\text{w}}$ is computed by Eq. (4) with $\mathbf{g}_i^{\text{w}}$.

$$\mathcal{L}_{\text{cos}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} H(\tilde{\mathbf{p}}_i^{\text{w}}, f_\theta(\text{Aug}_\text{s}(\mathbf{x}_i^{\text{ulb}}))).$$

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cos}}\mathcal{L}_{\text{cos}},$$

Algorithm 2: SoC: **So**ft Label Selection with Confidence-Aware **C**lustering based on Class Transition Tracking

Input: class number: K, labeled data set: $\mathcal{D}^{\text{lb}} = \{(\mathbf{x}_i^{\text{lb}}, \mathbf{y}_i)\}_{i=1}^{N}$, unlabeled data set: $\mathcal{D}^{\text{ulb}} = \{\mathbf{x}_i^{\text{ulb}}\}_{i=1}^{M}$, model: f_θ, weak and strong augmentation: Aug_w and Aug_s, class prediction bank: $\{l_i\}_{i=1}^{M}$, class tracking matrices: $\{\mathbf{C}^{(i)}\}_{i=1}^{N_b}$, CTT based k-medoids Clustering: $k\text{-medoids}_{f_{\text{sim}}}$, mapping function from confidence to k: f_{conf}

for $t = 1$ **to** MaxIteration **do**
  Sample labeled data batch $\{(\mathbf{x}_i^{\text{lb}}, \mathbf{y}_i)\}_{i=1}^{B} \subset \mathcal{D}^{\text{lb}}$
  Sample unlabeled data batch $\{\mathbf{x}_i^{\text{ulb}}\}_{i=1}^{\mu B} \subset \mathcal{D}^{\text{ulb}}$
  $\mathcal{L}_{\text{sup}} = \frac{1}{B}\sum_{i=1}^{B} H(\mathbf{y}_i, f_\theta(\mathbf{x}_i^{\text{lb}}))$ ▷ Compute the supervised loss
  for $i = 1$ **to** μB **do**
    $d = \text{Index}(\mathbf{x}_i^{\text{ulb}})$ ▷ Obtain the index of $\mathbf{x}_i^{\text{ulb}}$
    $\mathbf{p}_i^{\text{w}} = \text{Softmax}(f_\theta(\text{Aug}_\text{w}(\mathbf{x}_i^{\text{ulb}})))$ ▷ Compute soft pseudo-label for weakly-augmented $\mathbf{x}_i^{\text{ulb}}$
    $\hat{p}_i = \arg\max(\mathbf{p}_i^{\text{w}})$ ▷ Compute class prediction
    if $l_d \neq \hat{p}_i$ **then**
      $C_{l_d\hat{p}_i}^{(n)} = C_{l_d\hat{p}_i}^{(n)} + 1$ ▷ Perform class transition tracking $l_d = \hat{p}_i$
    end if
    $k_i = f_{\text{conf}}(\max(\mathbf{p}_i^{\text{w}}))$ ▷ Select k for CTT based k-medoids Clustering
    $\{S_1, \ldots, S_k\} = k\text{-medoids}_{f_{\text{sim}}}\left(\text{Average}(\{\mathbf{C}^{(i)}\}_{i=1}^{N_b}), k_i\right)$ ▷ See Algorithm 1
    $\mathcal{C}_i = S_s, s = \arg\max_{c\in\{1,\ldots,k\}} \mathbb{1}(\hat{p}_i \in S_c)$
    $\mathbf{g}_i = (\mathbb{1}(1 \in \mathcal{C}_i), \ldots, \mathbb{1}(K \in \mathcal{C}_i))$ ▷ Compute label selection indicator $\mathbf{g}_i$
    $\tilde{\mathbf{p}}_i = \text{Normalize}(\mathbf{g}_i \circ \mathbf{p}_i)$ ▷ Compute selected soft label
  end for
  $\mathcal{L}_{\text{cos}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} H(\tilde{\mathbf{p}}_i^{\text{w}}, f_\theta(\text{Aug}_\text{s}(\mathbf{x}_i^{\text{ulb}})))$ ▷ Compute the consistency loss
  return $\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cos}}\mathcal{L}_{\text{cos}}$ ▷ Optimize total loss
end for

The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

NANJING UNIVERSITY

Theoretical Support

The **Shrinkage Objective** is equivalent to the **entropy minimization** objective (see the full version of this paper [2]).

In SoC, for $\mathbf{x}_i^{\text{slb}}$, we obtain the set of candidate classes $\mathcal{C}_i$, the prediction probabilities $\mathbf{p}_i$ and the selected soft label $\tilde{\mathbf{p}}_i$. First, we resort the the vector sequence $\mathbf{p}_i = (p_{i,(1)}, \ldots, p_{i,(K)})$ to $(p_{i,(a_1)}, \ldots, p_{i,(a_{|\mathcal{C}_i|})}, p_{i,(b_1)}, \ldots, p_{i,(b_{|\mathcal{Y}\setminus\mathcal{C}_i|})})$, where K is the number of classes, $a_{j_a} \in \mathcal{C}_i$ and $b_{j_b} \in \mathcal{Y} \setminus \mathcal{C}_i$, *i.e.*, we put the probabilities of unselected classes at the back end of the vector sequence. Meanwhile, the subsequences $(p_{i,(a_1)}, \ldots, p_{i,(a_{|\mathcal{C}_i|})})$ and $(p_{i,(b_1)}, \ldots, p_{i,(b_{|\mathcal{Y}\setminus\mathcal{C}_i|})})$ in $\mathbf{p}_i$ are sorted by value in descending order, *i.e.*, $\max(p_{i,(a_{j_a})}) - p_{i,(a_1)} - p_{i,(1)}$ and $\max(p_{i,(b_{j_b})}) - p_{i,(b_1)} - p_{i,(|\mathcal{C}_i|+1)}$. Then, denoting $d - \sum_{c=|\mathcal{C}_i|+1}^{K} p_{i,(c)}$, we prove a equivalent form of Lemma 1:

$$
\begin{aligned}
-\sum_{c=1}^{K} p_{i,(c)} \log p_{i,(c)} \geq & -\sum_{c=1}^{|\mathcal{C}_i|} \frac{p_{i,(c)}}{1-d} \log \frac{p_{i,(c)}}{1-d} \\
& - -\sum_{c=1}^{|\mathcal{C}_i|} \frac{p_{i,(c)}}{1-d} \log p_{i,(c)} + \sum_{c=1}^{|\mathcal{C}_i|} \frac{p_{i,(c)}}{1-d} \log(1-d) \\
& - -\sum_{c=1}^{|\mathcal{C}_i|} \frac{p_{i,(c)}}{1-d} \log p_{i,(c)} + \log(1-d). \qquad (18)
\end{aligned}
$$

Rewriting Eq. (18) we obtain its equivalent:

$$
\begin{aligned}
\sum_{c=1}^{|\mathcal{C}_i|} \frac{p_{i,(c)}}{1-d} \log p_{i,(c)} - \sum_{c=1}^{|\mathcal{C}_i|} p_{i,(c)} \log p_{i,(c)} \geq & \sum_{c=|\mathcal{C}_i|+1}^{K} p_{i,(c)} \log p_{i,(c)} + \log(1-d) \\
(\frac{d}{1-d}) \sum_{c=1}^{|\mathcal{C}_i|} p_{i,(c)} \log p_{i,(c)} \geq & \sum_{c=|\mathcal{C}_i|+1}^{K} p_{i,(c)} \log p_{i,(c)} + \log(1-d). \qquad (19)
\end{aligned}
$$

When $d - 0$, Eq. (19) obviously holds. Given $d - \sum_{c=|\mathcal{C}_i|+1}^{K} p_{i,(c)} > 0$, according to the property of $h(x) - x \log x$, when only one entry $p_{i,(z)}$ of $(p_{i,(|\mathcal{C}_i|+1)}, \ldots, p_{i,(K)})$ has a non-zero value, $\sum_{c=|\mathcal{C}_i|+1}^{K} p_{i,(c)} \log p_{i,(c)}$ is maximized, *i.e.*, the term on the right side of the inequality Eq. (19) obtains the maximum. Given $\max(p_{i,(|\mathcal{C}_i|+1)}, \ldots, p_{i,(K)}) - p_{i,(|\mathcal{C}_i|+1)}$, the non-zero entry $p_{i,(z)}$ is obviously $p_{i,(|\mathcal{C}_i|+1)}$. In this case, we have $d - p_{i,(|\mathcal{C}_i|+1)}$ and $\sum_{c=|\mathcal{C}_i|+1}^{K} p_{i,(c)} \log p_{i,(c)} - p_{i,(|\mathcal{C}_i|+1)} \log p_{i,(|\mathcal{C}_i|+1)}$. Thus, we rewrite Eq. (19) as

$$
\begin{aligned}
(\frac{p_{i,(|\mathcal{C}_i|+1)}}{1-p_{i,(|\mathcal{C}_i|+1)}}) \sum_{c=1}^{|\mathcal{C}_i|} p_{i,(c)} \log p_{i,(c)} \geq & \, p_{i,(|\mathcal{C}_i|+1)} \log p_{i,(|\mathcal{C}_i|+1)} + \log(1-p_{i,(|\mathcal{C}_i|+1)}) \\
(\frac{p_{i,(|\mathcal{C}_i|+1)}}{1-p_{i,(|\mathcal{C}_i|+1)}})(p_{i,(1)} \log p_{i,(1)} + \sum_{c=2}^{|\mathcal{C}_i|} p_{i,(c)} \log p_{i,(c)}) \geq & \, p_{i,(|\mathcal{C}_i|+1)} \log p_{i,(|\mathcal{C}_i|+1)} + \log(1-p_{i,(|\mathcal{C}_i|+1)}). \qquad (20)
\end{aligned}
$$

It's worth noting that we have $\arg\max(\mathbf{p}_i) \in \mathcal{C}_i$ by Eq. (9), therefore we obtain $\max(\mathbf{p}_i) \in \{p_{i,(1)}, \ldots, p_{i,(|\mathcal{C}_i|)}\}$, *i.e.*, $p_{i,(1)} \geq p_{i,(|\mathcal{C}_i|+1)}$. In other words, there is no constraint on the value of $\sum_{c=2}^{|\mathcal{C}_i|} p_{i,(c)} \log p_{i,(c)}$ when $p_{i,(1)}$ and $p_{i,(|\mathcal{C}_i|+1)}$ are given. Thus, by Jensen Inequality we know that when $p_{i,(2)} - p_{i,(3)} - \cdots - p_{i,(|\mathcal{C}_i|)} - \frac{1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)}}{|\mathcal{C}_i|-1}$, $\sum_{c=2}^{|\mathcal{C}_i|} p_{i,(c)} \log p_{i,(c)}$ obtains the minimum, *i.e.*, the left side of the inequality Eq. (20) obtains the minimum with the given $p_{i,(1)}$ and $p_{i,(|\mathcal{C}_i|+1)}$ where $p_{i,(1)} \geq p_{i,(|\mathcal{C}_i|+1)}$. In this case, we can rewrite Eq. (20) as

$$
\begin{aligned}
\frac{p_{i,(|\mathcal{C}_i|+1)}}{1-p_{i,(|\mathcal{C}_i|+1)}} p_{i,(1)} \log p_{i,(1)} + \frac{p_{i,(|\mathcal{C}_i|+1)}(|\mathcal{C}_i|-1)}{1-p_{i,(|\mathcal{C}_i|+1)}} \times \frac{1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)}}{|\mathcal{C}_i|-1} \log \frac{1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)}}{|\mathcal{C}_i|-1} & \\
\geq p_{i,(|\mathcal{C}_i|+1)} \log p_{i,(|\mathcal{C}_i|+1)} + \log(1-p_{i,(|\mathcal{C}_i|+1)}). \qquad (21) &
\end{aligned}
$$

By Eq. (21), we obtain

$$
\begin{aligned}
& \frac{p_{i,(|\mathcal{C}_i|+1)}}{1-p_{i,(|\mathcal{C}_i|+1)}} p_{i,(1)} \log p_{i,(1)} + \frac{p_{i,(|\mathcal{C}_i|+1)}(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)})}{1-p_{i,(|\mathcal{C}_i|+1)}} |\log(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)}) - \log(|\mathcal{C}_i|-1)| \\
& \geq p_{i,(|\mathcal{C}_i|+1)} \log p_{i,(|\mathcal{C}_i|+1)} + \log(1-p_{i,(|\mathcal{C}_i|+1)}) \\
& \frac{p_{i,(|\mathcal{C}_i|+1)}}{1-p_{i,(|\mathcal{C}_i|+1)}} p_{i,(1)} \log p_{i,(1)} + \frac{p_{i,(|\mathcal{C}_i|+1)}(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)})}{1-p_{i,(|\mathcal{C}_i|+1)}} \log(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)}) - \\
& p_{i,(|\mathcal{C}_i|+1)} \log p_{i,(|\mathcal{C}_i|+1)} - \log(1-p_{i,(|\mathcal{C}_i|+1)}) \geq \frac{p_{i,(|\mathcal{C}_i|+1)}(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)})}{1-p_{i,(|\mathcal{C}_i|+1)}} \log(|\mathcal{C}_i|-1) \\
& \frac{p_{i,(1)}}{(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)})} \log p_{i,(1)} + \log(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)}) - \\
& \frac{(1-p_{i,(|\mathcal{C}_i|+1)})(p_{i,(|\mathcal{C}_i|+1)} \log p_{i,(|\mathcal{C}_i|+1)} + \log(1-p_{i,(|\mathcal{C}_i|+1)}))}{p_{i,(|\mathcal{C}_i|+1)}(1-p_{i,(1)}-p_{i,(|\mathcal{C}_i|+1)})} \geq \log(|\mathcal{C}_i|-1). \qquad (22)
\end{aligned}
$$

Letting $x - p_{i,(1)}$ and $y - p_{i,(|\mathcal{C}_i|+1)}$, we define the function

$$
f(x,y) - \frac{x}{1-x-y} \log x + \log(1-x-y) - \frac{(1-y)(y \log y + \log(1-y))}{y(1-x-y)}. \qquad (23)
$$

Then we can obtain

$$
\frac{\partial f(x,y)}{\partial x} - -\frac{(y-1)\,(y \log(x) - \log(1-y) - y \log(y))}{y\,(1-x-y)^2} \geq 0. \qquad (24)
$$

Given $x \geq y$ (due to $p_{i,(1)} \geq p_{i,(|\mathcal{C}_i|+1)}$), $f(x,y)$ is minimized when $x - y$. Plugging $x - y$ into $f(x,y)$, we obtain

$$
f(y) - \frac{y}{1-2y} \log y + \log(1-2y) - \frac{(1-y)(y \log y + \log(1-y))}{y(1-2y)} \qquad (25)
$$

and the first derivative of $f(y)$:

$$
f'(y) - \frac{(2y^2-4y+1)\log(1-y)}{(1-2y)^2 y^2}. \qquad (26)
$$

Solving $f'(y) - 0$ we obtain $y - 1 - \frac{\sqrt{2}}{2}$. It is easy to verify that $f'(y)$ exists for all y such that $0 < y < \frac{1}{2}$ and the sign of $f'(y)$ changes from negative to positive. Thus, $f(y)$ has a minimum at $y - 1 - \frac{\sqrt{2}}{2}$, *i.e.*,

$$
\min f(y) - f(1 - \frac{\sqrt{2}}{2}) - \frac{1}{4}\left(\sqrt{2}\log(16) + \log(64) - 4\log\left(1 - \frac{\sqrt{2}}{2}\right) + 4\log(\sqrt{2}-1)\right) \approx 2.36655. \qquad (27)
$$

Obviously, we have $\log 10 < f(1 - \frac{\sqrt{2}}{2}) < \log 11$. By Eq. (27), we know that Eq. (22) holds when $|\mathcal{C}_i| \leq 11$, *i.e.*, $z_i^{\text{obj2}} \leq 11$ (by $z_i^{\text{obj2}} - \sum_{c=1}^{K} g_{i,(c)}$ and Eq. (3)). It is worth noting that the strict requirement for z_i^{obj2} is to ensure that Eq. (22) holds in the worst case, where the class distribution of *selected classes* excluding $\arg\max(p_i)$ is "uniform distribution" (*i.e.*, for any $a, b \in \mathcal{C}_i \wedge a, b \neq \arg\max(p_i)$, $p_{i,(a)} - p_{i,(b)}$) and the class distribution of *unselected classes* is "one-hot distribution" (*i.e.*, $|\{p_{i,(z)} \mid p_{i,(z)} \neq 0 \wedge z \in \mathcal{Y} \setminus \mathcal{C}_i\}| - 1$). In the vast majority of cases, any z_i^{obj2} will make Eq. (22). Meanwhile, given an apposite α, thanks to Confidence-Aware k Selection in Sec. 3.3, $z_i^{\text{obj2}} \leq 11$ for the worst case also can be guaranteed after a few training iterations because the model becomes more confident on $\mathbf{x}_i^{\text{slb}}$ (implying a smaller z_i^{obj2}).

In fact, the mentioned extreme situation is almost impossible to occur in reality. For specific, with training, the prediction distribution outputted by the model will tend to be non-uniform unless the model learns nothing. Thus, the requirement for z_i^{obj2} will be greatly relaxed. To verify this point, we show an intuitive example: even if the class distribution of selected classes is uniform distribution (*i.e.*, for any $a, b \in \mathcal{C}_i$, $p_{i,(a)} - p_{i,(b)}$), Lemma 1 holds no matter what value z_i^{obj2} takes.

Lemma 1. *Given $\tilde{\mathbf{p}}_i$ in SoC (implying $\mathcal{C}_i \subsetneqq \mathcal{Y}$ and $\hat{p}_i \in \mathcal{C}_i$), we show that the entropy of $\tilde{\mathbf{p}}_i$ is smaller than that of $\mathbf{p}_i$:*

$$
\mathcal{H}(\tilde{\mathbf{p}}_i) \leq \mathcal{H}(\mathbf{p}_i), \qquad (11)
$$

where $\mathcal{H}(\cdot)$ refers to the entropy.

Proof for Theorem 1.. Given $\mathcal{C}_i^{(2)} \subsetneqq \mathcal{C}_i^{(1)}$, we can treat $\tilde{\mathbf{p}}_i^{(1)}$ as $\mathbf{p}_i$ and $\tilde{\mathbf{p}}_i^{(2)}$ as $\tilde{\mathbf{p}}_i$ in Lemma 1, *i.e.*, treat the previously obtained $\tilde{\mathbf{p}}_i$ as the naive soft pseudo-label. Thus, $\mathcal{H}(\tilde{\mathbf{p}}_i^{(2)}) \leq \mathcal{H}(\tilde{\mathbf{p}}_i^{(1)})$ holds. For any $\mathcal{C}_i^{(j)} \subsetneqq \mathcal{C}_i^{(j-1)}$, we can repeat the above proof to obtain $\mathcal{H}(\tilde{\mathbf{p}}_i^{(j)}) \leq \mathcal{H}(\tilde{\mathbf{p}}_i^{(j-1)})$. Thus, we can obtain $\mathcal{H}(\tilde{\mathbf{p}}_i^{(m)}) \leq \cdots \leq \mathcal{H}(\tilde{\mathbf{p}}_i^{(1)})$. □

Theorem 1. *In SoC, minimizing $|\mathcal{C}_i^{(1)}|$ to $|\mathcal{C}_i^{(m)}|$: $\mathcal{C}_i^{(m)} \subsetneqq \ldots \mathcal{C}_i^{(2)} \subsetneqq \mathcal{C}_i^{(1)}$ (i.e., $\sum_{c=1}^{K} \mathbf{g}_i^{(m)} < \cdots < \sum_{c=1}^{K} \mathbf{g}_i^{(2)} < \sum_{c=1}^{K} \mathbf{g}_i^{(1)}$), we show that the entropy of $\mathbf{p}_i$ is minimizing:*

$$
\mathcal{H}(\tilde{\mathbf{p}}_i^{(m)}) \leq \mathcal{H}(\tilde{\mathbf{p}}_i^{(m-1)}) \leq \ldots \mathcal{H}(\tilde{\mathbf{p}}_i^{(2)}) \leq \mathcal{H}(\tilde{\mathbf{p}}_i^{(1)}), \qquad (10)
$$

where $\mathcal{H}(\cdot)$ refers to the entropy.

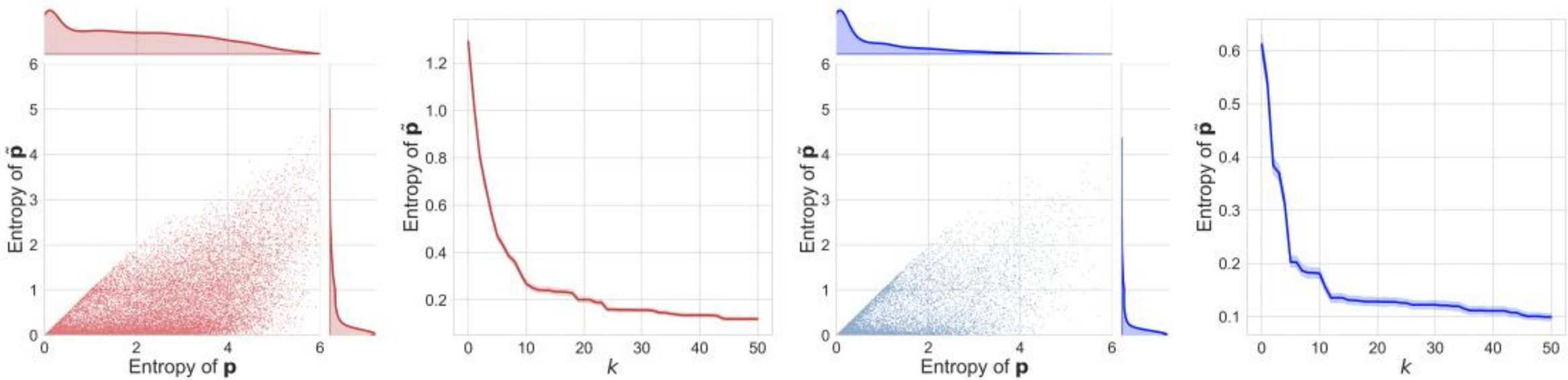
[2] https://arxiv.org/abs/2312.12237

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

南京大学 NANJING UNIVERSITY

Experimental Support

The **Shrinkage Objective** is equivalent to the **entropy minimization** objective (see the full version of this paper [2]).



(a) Semi-Aves (b) Semi-Aves (c) Semi-Fungi (d) Semi-Fungi

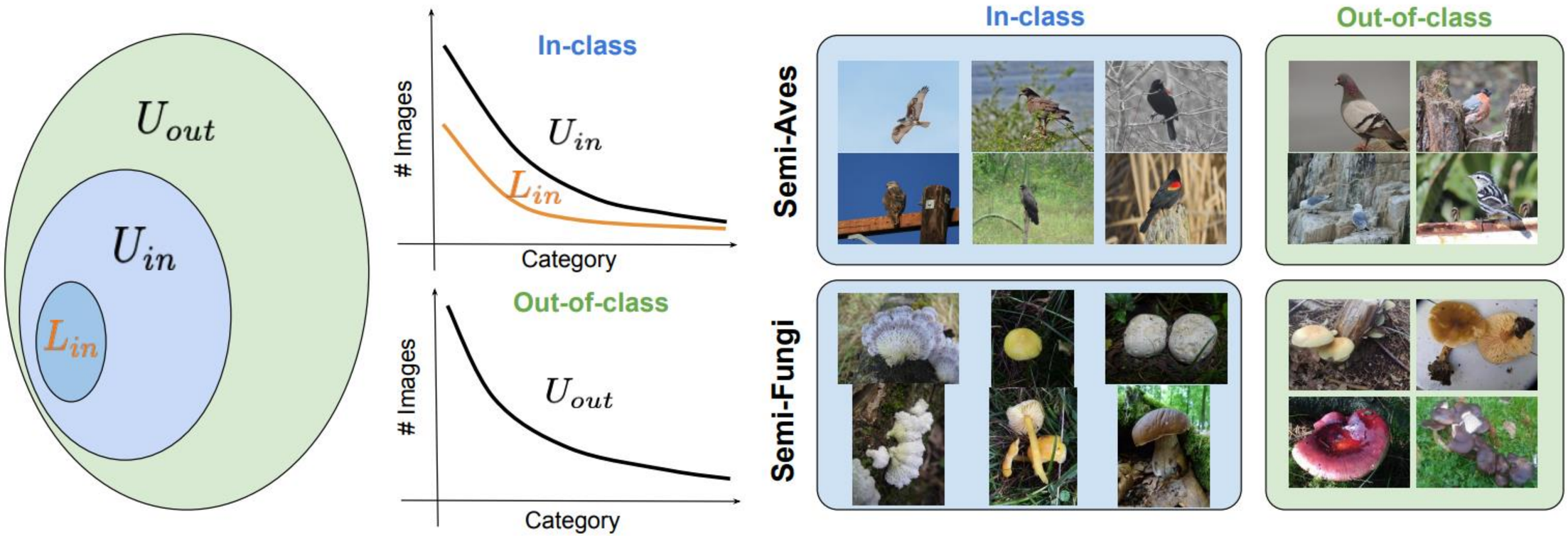
- The entropy of the pseudo-labels produced by our method is smaller than that of the original pseudo-labels.
- **The larger *k* is, the smaller the entropy is.**

[2] https://arxiv.org/abs/2312.12237

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

Experiments

- **Datasets： Semi-Aves (200 classes) and Semi-Fungi (200 classes) [1]**



- **Semi-Aves** is divided into the training set and validation set with a total of 5,959 labeled images and 26,640 unlabeled images, and the test set with 8,000 images.
- **Semi-Fungi** is divided into the training set and validation set with a total of 4,141 labeled images and 13,166 unlabeled images, and the test set with 4,000 images.

[1] A realistic evaluation of semi-supervised learning for fine-grained classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

Experiments

- **SS-FGVC results on Semi-Aves and Semi-Fungi**

Dataset	Pseudo-label	Method	Year	from scratch Top1	from scratch Top5	from ImageNet Top1	from ImageNet Top5	from iNat Top1	from iNat Top5
Semi-Aves	—	Supervised oracle	—	57.4±0.3	79.2±0.1	68.5±1.4	88.5±0.4	69.9±0.5	89.8±0.7
		MoCo (He et al. 2020)	CVPR' 20	28.2±0.3	53.0±0.1	52.7±0.1	78.7±0.2	68.6±0.1	87.7±0.1
	Hard label	Pseudo-Label (Lee et al. 2013)	ICML' 13	16.7±0.2	36.5±0.8	54.4±0.3	78.8±0.3	65.8±0.2	86.5±0.2
		Curriculum Pseudo-Label (Cascante-Bonilla et al. 2021)	AAAI' 21	20.5±0.5	41.7±0.5	53.4±0.8	78.3±0.5	69.1±0.3	87.8±0.1
		FixMatch (Sohn et al. 2020)	NIPS' 20	28.1±0.1	51.8±0.6	*57.4±0.8*	78.5±0.5	70.2±0.6	87.0±0.1
		FlexMatch (Zhang et al. 2021)*	NIPS' 21	27.3±0.5	49.7±0.8	53.4±0.2	77.9±0.3	67.6±0.5	87.0±0.2
		MoCo + FlexMatch*	NIPS' 21	35.0±1.2	58.5±1.0	53.4±0.4	77.0±0.2	68.9±0.3	87.7±0.2
	Soft label	KD-Self-Training (Su, Cheng, and Maji 2021)	CVPR' 21	22.4±0.4	44.1±0.1	55.5±0.1	79.8±0.1	67.7±0.2	87.5±0.2
		MoCo + KD-Self-Training (Su, Cheng, and Maji 2021)	CVPR' 21	31.9±0.1	56.8±0.1	55.9±0.2	80.3±0.1	70.1±0.2	88.1±0.1
		SimMatch (Zheng et al. 2022)*	CVPR' 22	24.8±0.5	48.1±0.6	53.3±0.5	77.9±0.8	65.4±0.2	86.9±0.3
		MoCo + SimMatch*	CVPR' 22	32.9±0.4	57.9±0.3	53.7±0.2	78.8±0.5	65.7±0.3	87.1±0.2
		SoC	Ours	*31.3±0.8 (↑11.4%)*	*55.3±0.7 (↑6.8%)*	57.8±0.5 (↑0.7%)	*80.8±0.5 (↑1.3%)*	*71.3±0.3 (↑1.6%)*	*88.8±0.2 (↑1.1%)*
		MoCo + SoC	Ours	**39.3±0.2 (↑12.3%)**	**62.4±0.4 (↑6.7%)**	**58.0±0.4 (↑3.8%)**	**81.7±0.4 (↑1.7%)**	**70.8±0.4 (↑1.0%)**	**88.9±0.5 (↑0.9%)**
Semi-Fungi	—	Supervised oracle	—	60.2±0.8	83.3±0.9	73.3±0.1	92.5±0.3	73.8±0.3	92.4±0.3
		MoCo (He et al. 2020)	CVPR' 20	33.6±0.2	59.4±0.3	55.2±0.2	82.9±0.2	52.5±0.4	79.5±0.2
	Hard label	Pseudo-Label (Lee et al. 2013)	ICML' 13	19.4±0.4	43.2±1.5	51.5±1.2	81.2±0.2	49.5±0.4	78.5±0.2
		Curriculum Pseudo-Label (Cascante-Bonilla et al. 2021)	AAAI' 21	31.4±0.6	55.0±0.6	53.7±0.2	80.2±0.1	53.3±0.5	80.0±0.5
		FixMatch (Sohn et al. 2020)	NIPS' 20	32.2±1.0	57.0±1.2	56.3±0.5	80.4±0.5	58.7±0.7	81.7±0.2
		FlexMatch (Zhang et al. 2021)*	NIPS' 21	36.0±0.9	59.9±1.1	59.6±0.5	82.4±0.5	60.1±0.6	82.2±0.5
		MoCo + FlexMatch*	NIPS' 21	44.2±0.6	67.0±0.8	59.9±0.8	82.8±0.7	61.4±0.6	83.2±0.4
	Soft label	KD-Self-Training (Su, Cheng, and Maji 2021)	CVPR' 21	32.7±0.2	56.9±0.2	56.9±0.3	81.7±0.2	55.7±0.3	82.3±0.2
		MoCo + KD-Self-Training (Su, Cheng, and Maji 2021)	CVPR' 21	39.4±0.3	64.4±0.5	58.2±0.5	84.4±0.2	55.2±0.5	82.9±0.2
		SimMatch (Zheng et al. 2022)*	CVPR' 22	36.5±0.9	61.7±1.0	56.6±0.4	81.8±0.6	56.7±0.3	80.9±0.4
		MoCo + SimMatch*	CVPR' 22	42.2±0.5	67.0±0.4	56.5±0.2	82.5±0.3	57.4±0.2	81.3±0.4
		SoC	Ours	*39.4±2.3 (↑7.9%)*	*62.5±1.1 (↑1.3%)*	*61.4±0.4 (↑3.0%)*	*83.9±0.6 (↑1.8%)*	*62.4±0.2 (↑3.8%)*	*85.1±0.2 (↑3.5%)*
		MoCo + SoC	Ours	**47.2±0.5 (↑6.8%)**	**71.3±0.2 (↑6.4%)**	**61.9±0.3 (↑3.3%)**	**85.8±0.2 (↑3.6%)**	**62.5±0.4 (↑1.8%)**	**84.7±0.2 (↑1.8%)**

- The models are trained from scratch, or ImageNet/iNat pre-trained or the model initialized with MoCo [3] learning on the unlabeled data.

[3] He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

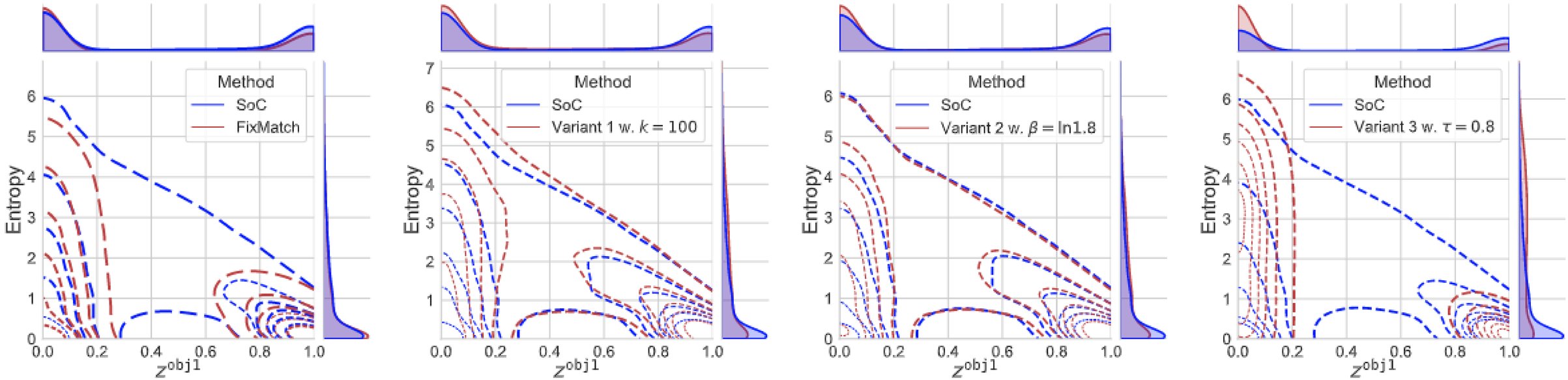
The 38th Annual AAAI Conference on Artificial Intelligence
FEBRUARY 20-27, 2024 | VANCOUVER, CANADA

Roll With the Punches: Expansion and Shrinkage of Soft Label Selection for Semi-supervised Fine-Grained Learning

南京大學 NANJING UNIVERSITY

Experiments

- **Ablation studies: comparison with different SoC variants**



(a) SoC vs. FixMatch (b) SoC vs. Variant 1 (c) SoC vs. Variant 2 (d) SoC vs. Variant 3

- **SoC vs. FixMatch**: Highlighting the performance advantage of soft label.
- **Variant 1**: Fixing the number of clusters.
- **Variant 2**: Adjusting the number of clusters non-linearly based on confidence.
- **Variant 3**: Filtering pseudo-labels with a fixed threshold.

Thanks!

Arxiv: https://arxiv.org/abs/2312.12237

Github: https://github.com/NJUyued/SoC4SS-FGVC

Homepage: https://njuyued.github.io/