模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

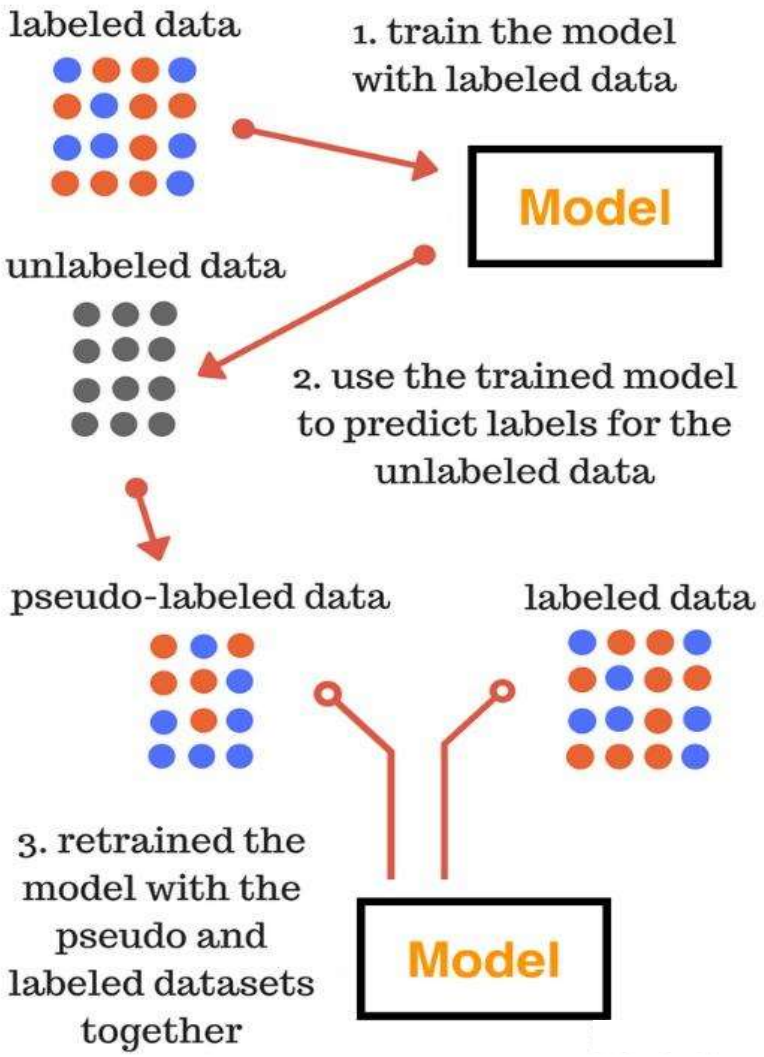
InPL: Pseudo-labeling the Inliers First for Imbalanced Semi-supervised Learning

Zhuoran Yu **Yin Li** **Yong Jae Lee**
University of Wisconsin-Madison
{zhuoran.yu, yin.li}@wisc.edu
yongjaelee@cs.wisc.edu

ICLR 2023

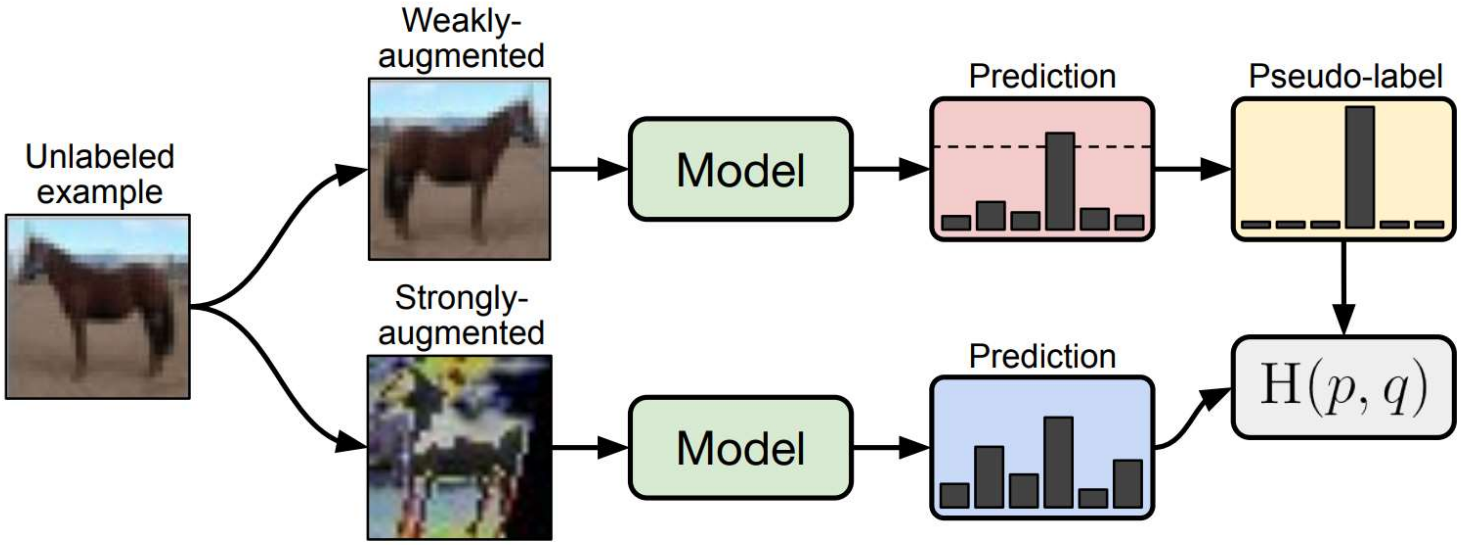
Background

Semi-supervised Learning (SSL)



Related work

FixMatch (2020)



$$\mathcal{L}_u = \frac{1}{B_u} \sum_{b=1}^{B_u} \mathbb{1}[\max_i(p(y_i|\omega(\mathbf{x_b}))) \geq \tau_c] \, \mathcal{H}(\hat{p}(\mathbf{y}|\omega(\mathbf{x_b})), p(\mathbf{y}|\Omega(\mathbf{x_b})))$$

Energy Score

The essence of the energy-based model (EBM) is to build a function $E(x)$: $\mathbb{R}^D \to \mathbb{R}$ that maps each point x of an input space to a single, non-probabilistic scalar called the energy. A collection of energy values could be turned into a probability density $p(x)$ through the Gibbs distribution:

$$p(y \mid \mathbf{x}) = \frac{e^{-E(\mathbf{x},y)/T}}{\int_{y'} e^{-E(\mathbf{x},y')/T}} = \frac{e^{-E(\mathbf{x},y)/T}}{e^{-E(\mathbf{x})/T}}$$

The Helmholtz free energy $E(x)$ of a given data point $x \in \mathbb{R}^D$ can be expressed as the negative of the log partition function:

$$E(\mathbf{x}) = -T \cdot \log \int_{y'} e^{-E(\mathbf{x},y')/T}$$

Related work

The energy-based model has an inherent connection with modern machine learning, especially discriminative models. To see this, we consider a discriminative neural classifier $f(x)$: $\mathbb{R}^D \rightarrow \mathbb{R}^K$, which maps an input $x \in \mathbb{R}^D$ to K real-valued numbers known as logits. These logits are used to derive a categorical distribution using the softmax function:

$$p(y \mid \mathbf{x}) = \frac{e^{f_y(\mathbf{x})/T}}{\sum_i^K e^{f_i(\mathbf{x})/T}}$$

we can define an energy for a given input (x, y) as $E(x, y) = -f_y(x)$. More importantly, without changing the parameterization of the neural network $f(x)$, we can express the free energy function $E(x; f)$ over $x \in \mathbb{R}^D$ in terms of the denominator of the softmax activation:

$$E(\mathbf{x}; f) = -T \cdot \log \sum_i^K e^{f_i(\mathbf{x})/T}$$

Related work

Class-Imbalanced Semi-Supervised Learning.

DARP (2020) refine the pseudo-labels through convex optimization targeted specifically for the imbalanced scenario.

CReST (2021) achieve class-rebalancing by pseudo-labeling unlabeled samples with frequency that is inversely proportional to the class frequency.

ABC (2021) introduce an auxiliary classifier that is trained with class-balanced sampling

Adsh (2022) extend the idea of adaptive thresholding to the long-tailed scenario.

DASO (2022) use a similarity-based classifier to complement pseudo-labeling.

Related work

Out-of-Distribution Detection

DS3L (2020) optimizes a meta network to selectively use unlabeled data;

OpenMatch (2021) trains an outlier detector with soft consistency regularization.

Motivation

Current research status

State-of-the-art imbalanced SSL methods are build upon the pseudo-labeling and consistency regularization frameworks by augmenting them with additional modules that tackle specific imbalanced issues. Critically, these methods still rely on confidence-based thresholding for pseudo-labeling, in which only the unlabeled samples whose predicted class confidence surpasses a very high threshold (e.g., 0.95) are pseudo-labeled for training.

Faces two major drawbacks

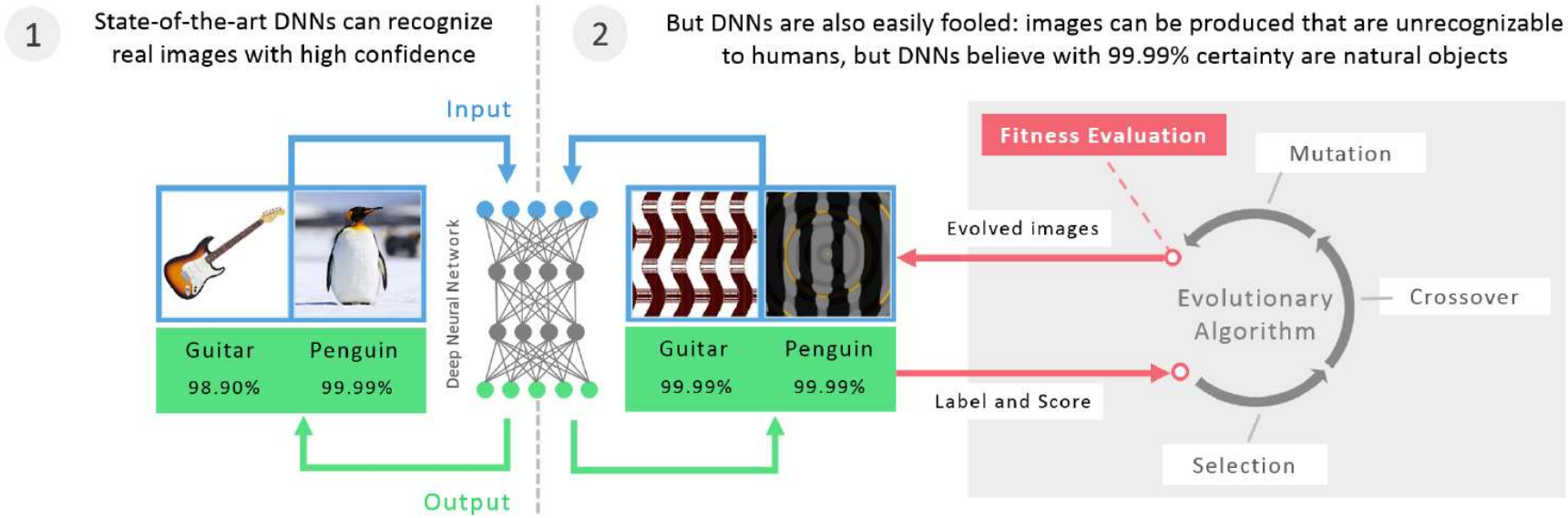
1. applying a high confidence threshold yields significantly lower recall of pseudo-labels for minority classes, resulting in an exacerbation of class imbalance. Lowering the threshold can improve the recall for tail classes but at the cost of reduced precision for other classes.

2. prior studies show that softmax-based confidence scores in deep networks can be arbitrarily high on even out-of-distribution samples.

Motivation

Example

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images
CVPR 2015



Energy Score vs. Softmax Score

$$p(x) = \frac{e^{-E(x,f)/T}}{\int_x e^{-E(x,f)/T}}$$

$$\log p(x) = -E(x,f)/T - \overbrace{\log Z}^{\text{contant for all } x}$$

The energy score and the negative logarithmic likelihood of the data points are linearly aligned, while low energy means high likelihood, which means it is more likely to be ID data, and vice versa, it is more likely to be OOD data.

Motivation

Energy Score vs. Softmax Score

For an ID data, its negative logarithmic likelihood expectation is smaller. However, the higher the confidence level of this classification, the better, and the two conflict.

$$\max_y p(y|x) = \frac{e^{f^{max}(x)}}{\sum_i e^{f_i(x)}} = \frac{1}{\sum_i e^{f_i(x)-f^{max}(x)}}$$

$$\log \max_y p(y|x) = E(x, f(x) - f^{max}(x)) = E(x, f) + f^{max}(x)$$

$$\log \max_y p(y|x) = -\log p(x) + f^{max}(x) - \log Z$$

$$E(\mathbf{x}; f) = -T \cdot \log \sum_i^K e^{f_i(\mathbf{x})/T}$$

$$\log p(x) = -E(x, f)/T - \overbrace{\log Z}^{\text{contant for all } x}$$

Method

The training of pseudo-labeling SSL methods for image classification involves two loss terms: the supervised loss $\mathcal{L}_s$ computed on human-labeled data and the unsupervised loss $\mathcal{L}_u$ computed on unlabeled data.

The final loss at each training iteration is computed by $\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_u$ with λ as a hyperparameter to balance the loss terms. The model parameters are updated with this loss after each iteration.

$$\mathcal{L}_s = \frac{1}{B_s}\sum_{b=1}^{B_s}\mathcal{H}(\mathbf{y_b}, p(\mathbf{y}|\omega(\mathbf{x_b})))$$

Method

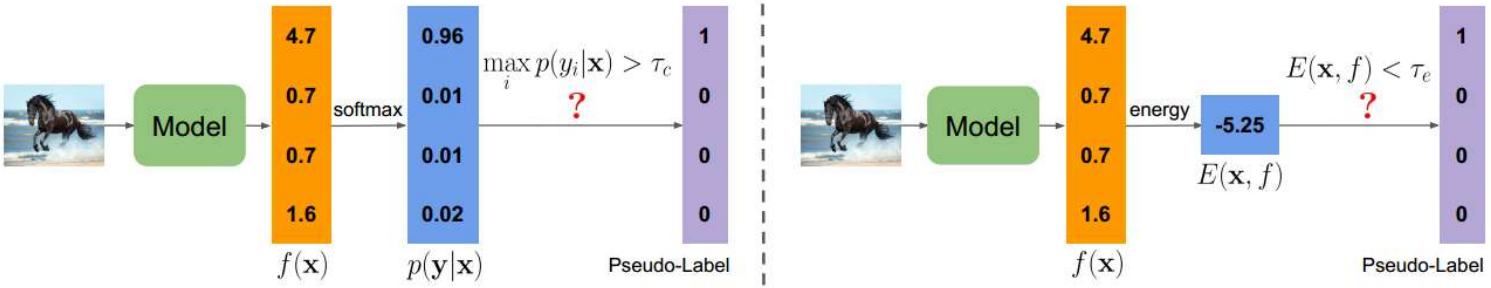


Figure 2: Overview of Confidence-based Pseudo-Labeling vs. Inlier Pseudo-Labeling.

$$\mathcal{L}_u = \frac{1}{B_u}\sum_{b=1}^{B_u} \mathbb{1}[\max_i(p(y_i|\omega(\mathbf{x_b}))) \geq \tau_c]\, \mathcal{H}(\hat{p}(\mathbf{y}|\omega(\mathbf{x_b})), p(\mathbf{y}|\Omega(\mathbf{x_b})))$$

$$\mathcal{L}_u = \frac{1}{B_u}\sum_{b=1}^{B_u} \mathbb{1}[E(\omega(\mathbf{x_b}), f(\omega(\mathbf{x_b}))) < \tau_e]\, \mathcal{H}(\hat{p}(\mathbf{y}|\omega(\mathbf{x_b})), p(\mathbf{y}|\Omega(\mathbf{x_b})))$$

Method

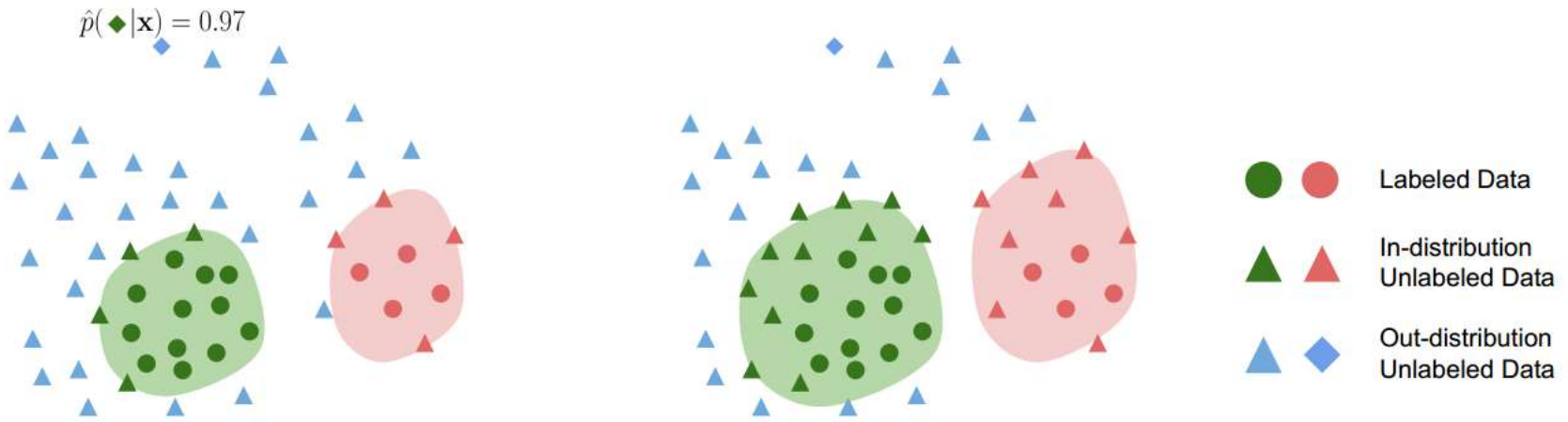
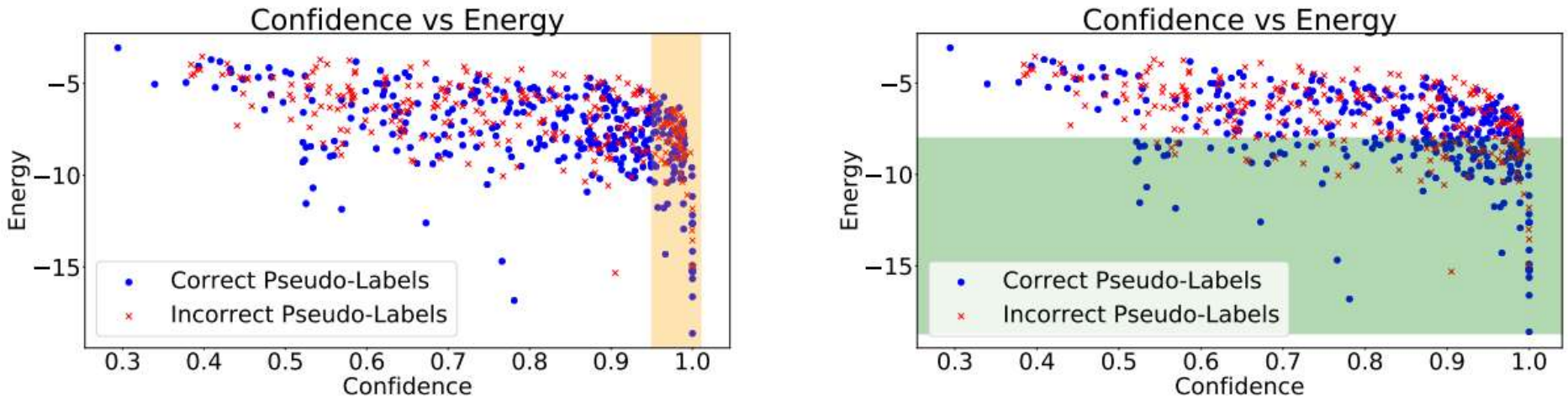


Figure 1: We illustrate the idea of InPL with a toy example with one head class (green) and one tail class (red). (a) At the beginning of training, only a few unlabeled samples are close enough to the training distribution formed by the initial labeled data. Note that with a confidence-based approach, the diamond unlabeled sample would be added as a pseudo-label for the green class since the model's confidence for it is very high (0.97). Our InPL instead ignores it since its energy score is too high and is thus considered out-of-distribution at this stage. (b) As training progresses, the training distribution is evolved by both the initial labeled data and the pseudo-labeled "in-distribution" unlabeled data, and more unlabeled data can be included in training. In this example, with our approach InPL, the diamond sample would eventually be pseudo-labeled as the red class.

Method



(a) Confidence-based Pseudo-labeling

(b) Inlier Pseudo-labeling

Figure 3: **Visualization: confidence vs energy score**: The shaded region shows the unlabeled samples that are pseudo-labeled. Inlier Pseudo-Labeling can produce correct pseudo-labels for many low-confident unlabeled samples, increasing recall while filtering out many false positives.

ENERGY-BASED PSEUDO-LABELING VS. CONFIDENCE-BASED PSEUDO LABELING

	CIFAR10-LT			CIFAR100-LT	
	$\gamma = 50$	$\gamma = 100$	$\gamma = 200$	$\gamma = 50$	$\gamma = 100$
UDA (Xie et al., 2020a)	80.21±0.49	72.19±1.51	63.32±1.67	46.79±0.76	41.47±0.97
FixMatch (Sohn et al., 2020)	80.84±0.20	72.95±1.32	63.25±0.13	46.99±0.37	41.49±0.38
FixMatch-UPS (Rizve et al., 2021)	81.75±0.56	73.17±1.63	64.38±0.56	-	-
FixMatch-InPL w/o AML (ours)	**83.36**±0.38	**76.05**±0.84	**66.47**±1.06	**48.03**±0.31	**42.53**±0.68
FixMatch-Debias + AML (Wang et al., 2022)	83.53±0.67	76.92±1.72	67.70±0.44	**50.24**±0.46	44.12±0.81
FixMatch-InPL(ours)	**83.92**±0.52	**77.44**±1.17	**68.47**±1.15	49.96±0.36	**44.33**±0.61
OpenMatch (Saito et al., 2021)	81.01±0.45	73.15±1.03	63.22±1.86	46.92±0.28	40.76±0.81
FixMatch-D3SL (Guo et al., 2020)	81.20±0.33	72.71±2.32	65.09±1.72	46.83±0.45	41.22±0.39

Table 1: **Top-1 accuracy of FixMatch variants on CIFAR 10-LT/100-LT**. For CIFAR10-LT and CIFAR100-LT, we use 10% and 30% data as labeled sets, respectively. We use Wide ResNet-28-2 (Zagoruyko & Komodakis, 2016) for CIFAR 10-LT and WRN-28-8 for CIFAR 100-LT. All methods are trained with the default FixMatch training schedule (Sohn et al., 2020). Results are reported with the mean and standard deviation over 3 different runs.

These results show the efficacy of InPL over standard confidence-based pseudo-labeling for imbalanced SSL.

Experiment

ParNeC 模式识别与神经计算研究组 PAttern Recognition and NEural Computing

COMPARISON TO STATE-OF-THE-ART IMBALANCED SSL APPROACHES

Dataset	CIFAR10-LT			CIFAR100-LT
Imbalance Ratio	$\gamma = 100$	$\gamma = 150$	$\gamma = 200$	$\gamma = 20$
FixMatch (Sohn et al., 2020)	72.3±0.33 / 53.8±0.63	68.5±0.60 / 45.8±1.15	66.3±0.49 / 42.4±0.94	51.0±0.20 / 32.8±0.41
w/ DARP+cRT (Kim et al., 2020)	78.1±0.89 / 66.6±1.55	73.2±0.85 / 57.1±1.13	-	54.7±0.46 / 41.2±0.42
w/ CReST+ (Wei et al., 2021)	76.6±0.46 / 61.4±0.85	70.0±0.82 / 49.4±1.52	-	51.6±0.29 / 36.4±0.46
w/ ABC (Lee et al., 2021)	81.1±0.82 / 72.0±1.77	77.1±0.46 / 64.4±0.92	73.9±1.18 / 58.1±2.72	56.3±0.19 / 43.4±0.42
w/ ABC-InPL (ours)	**82.9**±0.60 / **76.4**±1.49	**79.7**±0.71 / **70.8**±1.43	**76.4**±1.09 / **63.7**±2.03	**57.7**±0.33 / **46.4**±0.26
RemixMatch (Berthelot et al., 2020)	73.7±0.39 / 55.9±0.87	69.9±0.23 / 48.4±0.60	68.2±0.37 / 45.4±0.70	54.0±0.29 / 37.1±0.37
w/ DARP+cRT (Kim et al., 2020)	78.5±0.61 / 66.4±1.69	73.9±0.59 / 57.4±1.45	-	55.1±0.45 / 43.6±0.58
w/ CReST+ (Wei et al., 2021)	75.7±0.34 / 59.6±0.76	71.3±0.77 / 50.8±1.56	-	54.6±0.48 / 38.1±0.69
w/ ABC (Lee et al., 2021)	82.4±0.45 / 75.7±1.18	80.6±0.66 / 72.1±1.51	**78.8**±0.27 / 69.9±0.99	57.6±0.26 / 46.7±0.50
w/ ABC-InPL(ours)	**83.6**±0.45 / **81.7**±0.97	**81.3**±0.83 / **76.8**±0.88	**78.8**±0.75 / **74.5**±1.47	**58.4**±0.25 / **48.9**±0.36

	CIFAR10-LT				CIFAR100-LT			
	$\gamma = 100$		$\gamma = 150$		$\gamma = 10$		$\gamma = 20$	
	$N_1 = 500$ $M_1 = 4000$	$N_1 = 1500$ $M_1 = 3000$	$N_1 = 500$ $M_1 = 4000$	$N_1 = 1500$ $M_1 = 3000$	$N_1 = 50$ $M_1 = 400$	$N_1 = 150$ $M_1 = 300$	$N_1 = 50$ $M_1 = 400$	$N_1 = 150$ $M_1 = 300$
FixMatch† (Sohn et al., 2020)	68.5±0.94	71.5±0.31	62.9±0.36	72.4±1.03	-	-	-	-
w/ Adsh† (Cui et al., 2019)	76.3±0.86	78.1±0.42	67.5±0.45	73.7±0.34	-	-	-	-
FixMatch (Sohn et al., 2020)	67.8±1.13	77.5±1.32	62.9±0.36	72.4±1.03	45.2±0.55	56.5±0.06	40.0±0.96	50.7±0.25
w/ DARP (Kim et al., 2020)	74.5±0.78	77.8±0.63	67.2±0.32	73.6±0.73	49.4±0.20	58.1±0.44	43.4±0.87	52.2±0.66
w/ CREST+ (Wei et al., 2021)	76.3±0.86	78.1±0.42	67.5±0.45	73.7±0.34	44.5±0.94	57.4±0.18	40.1±1.28	50.1±0.21
w/ DASO (Oh et al., 2022)	76.0±0.37	79.1±0.75	70.1±1.81	75.1±0.77	49.8±0.24	59.2±0.35	43.6±0.09	52.9±0.42
w/ ABC (Lee et al., 2021)	78.9±0.82	83.8±0.36	66.5±0.78	80.1±0.45	47.5±0.18	59.1±0.21	41.6±0.83	53.7±0.55
w/ ABC-DASO (Oh et al., 2022)	80.1±1.16	83.4±0.31	70.6±0.80	80.4±0.56	50.2±0.62	60.0±0.32	44.5±0.25	**55.3**±0.53
w/ ABC-InPL (Ours)	**81.4**±0.76	**84.4**±0.20	**77.5**±1.57	**80.9**±0.82	**51.8**±1.09	**61.0**±0.32	**44.6**±1.24	55.1±0.51

ABC-InPL consistently outperforms the confidence-based counterparts on CIFAR10-LT and CIFAR100-LT under the ABC framework.

These results demonstrate that InPL achieves strong performance on imbalanced data across different frameworks and evaluation settings.

RESULTS ON IMAGENET

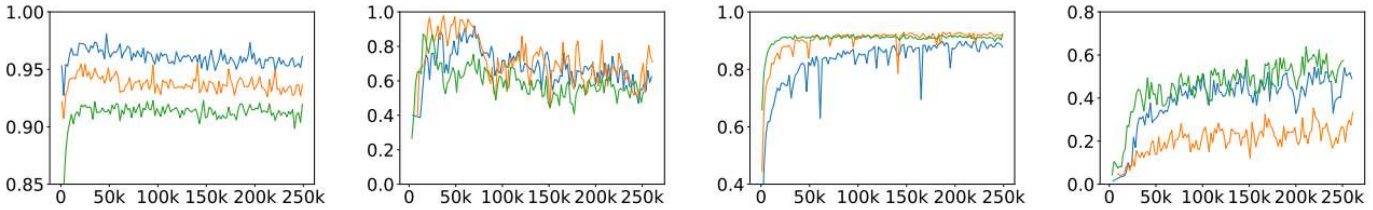
	ImageNet-127 (Imbalanced)	ImageNet (Balanced)
FixMatch	51.96	56.34
FixMatch-InPL (ours)	**54.82**	**57.92**

Table 4: Results on ImageNet-127 and ImageNet. We use sample 10% data as the labeled set for ImageNet-127 and use 100 labels per class for ImageNet. Our approach outperforms the confidence-based counterpart in FixMatch on both datasets.

InPL outperforms the vanilla FixMatch with confidence-based pseudo-labeling by a large margin, which further demonstrates the efficacy of our approach on large-scaled datasets.

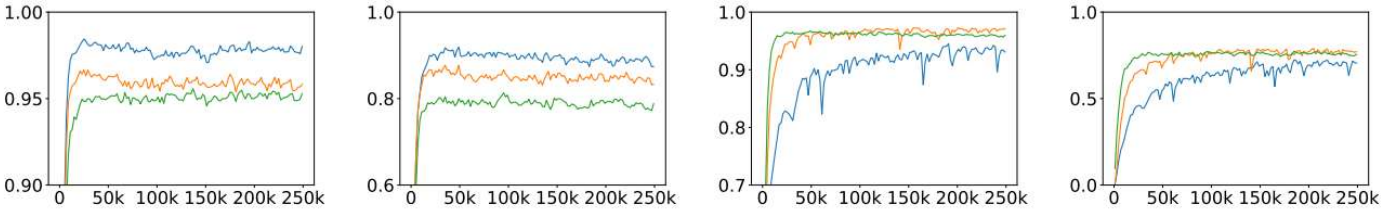
This shows that the energy-based pseudo-labeling also has potential to become a general solution to SSL problems.

PSEUDOLABEL PRECISION AND RECALL ANALYSIS



(a) Precision: Overall (b) Precision: Tail (c) Recall: Overall (d) Recall: Tail

Figure 4: **Precision-Recall Analysis**: We compare pseudo-label precision and recall between InPL and FixMatch. Orange and green curves denote FixMatch with threshold 0.95 and 0.6 respectively. InPL is shown in blue, which achieves improved recall for tail classes and better overall precision.



(a) Precision: Head (b) Precision: Body (c) Recall: Head (d) Recall: Body

Figure B: **Precision-Recall Analysis on Head and Body Classes**: Orange and green curves denote FixMatch with threshold 0.95 and 0.6 respectively. InPL is denoted by blue curves. InPL consistently achieves higher pseudo-label precision with slightly lower recall compared with the confidence-based pseudo-labeling baselines.

Compared with FixMatch, InPL achieves higher precision for overall, head, and body pseudo-labels. Importantly, it doubles FixMatch's recall of tail pseudo-labels without hurting the precision much. This shows that InPL predicts more true positives for the tail classes and also becomes less biased to the head classes.

Thanks