Supplementary Material For: Self-Tuning for Data-Efficient Deep Learning

Ximei Wang [*1] **Jinghan Gao** [*1] **Mingsheng Long** [1] **Jianmin Wang** [1]

A. Original Results for Sensitivity Analysis

In Section 5.7 of the main paper, we report the performance of Self-Tuning on *Standard Cars* on various values of feature size L and queue size D. With a 3D plot, it vividly shows the robustness of Self-Tuning to different values of L and D. However, the detailed numbers of Self-Tuning are not accessible via a figure. To this end, we report the original results here in Table 1 and Table 2 respectively. Due to the constraint of computing resources, we did not conduct experiments with $L > 1028$ and $D > 32$, though there is a high probability that better results would be achieved with larger values of L and D.

Table 1. Classfication accuracy on *Stanford Cars* with 15% labels.

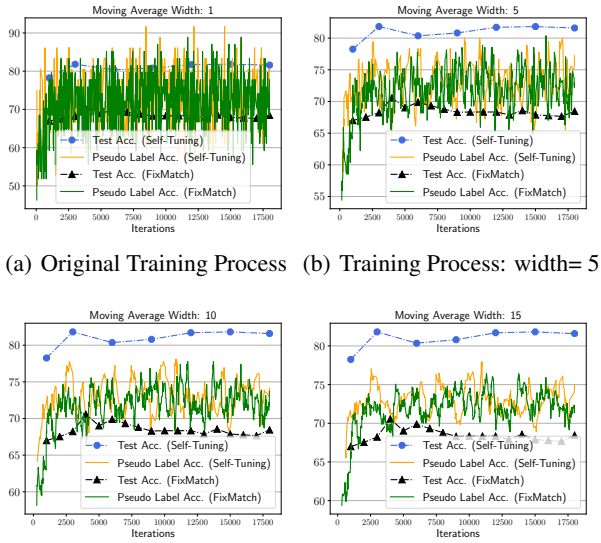
Test Acc. L \ D	8	16	24	32
128	69.87	70.45	70.60	70.8
256	70.23	70.67	70.70	70.98
512	70.61	70.87	71.25	72.04
1028	71.79	71.99	72.21	72.50

Table 2. Classfication accuracy on *Stanford Cars* with 30% labels.

Test Acc. L \ D	8	16	24	32
128	81.43	82.01	82.16	82.36
256	81.66	82.23	82.26	82.54
512	82.06	82.43	82.81	83.01
1028	82.79	83.27	83.18	83.58

B. Analysis on Why Self-Tuning Works

In Section 5.8 of the main paper, Figure 7(a) shows that Self-Tuning has a larger improvement over the accuracy of pseudo-labels than FixMatch, given an identical pre-trained model with approximate pseudo-label accuracy. Due to the space limit, more details of Figure 7(a) are missing in the main paper, which will be covered here. For Figure 7(a), the test accuracy is calculated *on the whole test set* while the pseudo-label accuracy is calculated *on each minibatch of the unlabeled data*, resulting in a smooth test accuracy and an unsteady pseudo-label accuracy. The pseudo-label accuracy in the main paper is the original training process. For clear comparison, we further provide training processes smoothed by a moving average method with different moving average widths are shown in Figure 1 here. These figures reveal that even pseudo-label accuracy is shaking across different training iterations, Self-Tuning has steady performance, indicating that the proposed pseudo group contrast (PGC) mechanism successfully mitigates the reliance on pseudo-labels and boosts the tolerance to false-labels.



(a) Original Training Process (b) Training Process: width= 5

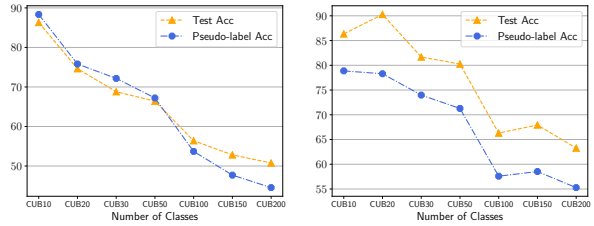
(c) Training Process: width= 10 (d) Training Process: width= 15

Figure 1. Comparisons between Self-Tuning with FixMatch on pseudo-label accuracy and test accuracy.

Different from Figure 7(a) that focuses on the initial pseudo-label accuracy with an identical pre-trained model, Figure 7(b) shows the between the gap between test ac-

[*]Equal contribution [1]School of Software, BNRist, Tsinghua. E-mail: Ximei Wang (wxm17@mails.tsinghua.edu.cn). Correspondence to: Mingsheng Long <mingsheng@tsinghua.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.



(a) Analysis on FixMatch (b) Analysis on Self-Tuning

Figure 2. Comparisons between Self-Tuning with FixMatch on pseudo-label accuracy and test accuracy.

curacy with pseudo-label accuracy *after the model converges*. Figure Figure 7(b) in the main paper reveals that Self-Tuning has a larger gap between test accuracy with the accuracy of pseudo-labels than that of FixMatch. This observation is consistently held under different sizes of label space. Detailed numbers of test accuracy and pseudo-label accuracy when the label space enlarges from 10 (*CUB10*) to 200 (*CUB200*) are shown in Figure 2(a) and Figure 2(b). Since the pseudo-label accuracy is recorded after the model converges, a higher pseudo-label accuracy for Self-Tuning than FixMatch is seen, though they are initialized with an identical pre-trained model with approximate pseudo-label accuracies for a fair comparison.

Further, as mentioned in the main paper, by unifying the exploration of labeled and unlabeled data and the transfer of a pre-trained model, Self-Tuning escapes from the dilemma of just developing TL or SSL methods. As analyzed in the main paper, this unified form of learning labeled and unlabeled data is much better than a sequential form.

C. When Self-Tuning Works and Fails

We empirically found that Self-Tuning works in most downstream datasets except a target dataset with long-tailed distributed labeled data, such as *Caltech* and *SUN-397*. Dominated by head-classes, it is hard for a model to distinguish tailed-classes from other classes, even with the proposed Self-Tuning method since *a cross-entropy loss on labeled data is included*. Further, we also found that this conclusion is consistently held with previous methods including L2-SP, DELTA, BSS and Co-Tuning. The performance on these long-tailed datasets was not reported by previous methods. Maybe a simple combination with long-tailed learning methods, e.g. resampling or reweighting, can alleviate this problem and we leave it as future work.