Supplementary Materials: Enhancing Robustness in Learning with Noisy Labels: An Asymmetric Co-Training Approach

Anonymous Authors

In this supplementary material, we offer additional evaluations of our proposed Asymmetric Co-Training (ACT) approach. We conduct further analysis of hyper-parameters and qualitative analyses to demonstrate the effectiveness of our method. Additionally, we compare the robustness of sample selection between our ACT and existing Symmetric Co-Training (SCT) methods across various noise settings. Additional experiments are conducted and reported according to the following themes:

- **Qualitative analysis on real-world datasets.** To enhance the visualization of our method's performance, we present qualitative analyses on real-world datasets (*i.e.*, Web-Aircraft, Web-Bird, Web-Car, and Food101N). As depicted in Fig. 1 and Fig. 2, we showcase several visualization results of clean and noisy samples selected by our sample selection methods across various fine-grained categories.
- **Sensitivity of Hyper-parameters.** To achieve optimal performance of our proposed method, we study the sensitivity of hyper-parameters (*i.e.*, τ and λ) utilized in our ACT. τ is adjusted to control the extent of label memorization of the NTM for mining more valuable clean samples. Meanwhile, λ mainly governs the weight of the regularization term loss in Eq. (11). We pressent the model performance under different ρ and τ settings in Fig. 3 and Fig. 4 across various noisy datasets (*i.e.*, CIFAR100N and CIFAR80N) and different noise settings (*i.e.*, Sym-20% and Sym-80% and Asym-40%).
- **Extended comparison with existing SCT methods.** To illustrate the superior label noise mitigation effects of our proposed ACT compared to existing SCT methods (*i.e.*, Decoupling, Co-teaching, Co-teaching+, and JoCoR), we conduct extended comparisons using precision, recall, and F1 score metrics. Figs. 5-10 present the extend comparison results with the four metrics under different noise conditions on CIFAR100N and CIFAR80N.

1 QUALITATIVE ANALYSIS

In this work, we introduce a novel asymmetric co-training approach to alleviate the harmful effects of noisy labels. Specifically, we introduce two criteria (*i.e.*, Criteria 1 and 2) and formulate the asymmetric sample selection and mining strategy based on the relationship between model predictions and given labels in ACT. In order to further visualize the performance of our proposed ACT, we provide qualitative analysis of these real-world datasets (*i.e.*, Web-Aircraft, Web-Bird, Web-Car, and Food-101N). As shown in Fig. 1, we provide several visualization results of clean and noisy samples selected by our asymmetric sample selection methods on three fine-grained categories (*i.e.*, *Yak-42*, *Carolina Wren*, and *Tesla Model S Sedan 2012*) from Web-Aircraft, Web-Bird, and Web-Car. Besides, Fig. 2 shows other visualization results on other three fine-grained categories (*i.e.*, *Apple Pie*, *Pizza*, and *Sushi*) from Food101N.

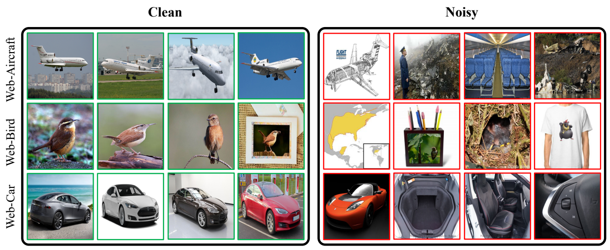


Figure 1: Some visualization results of clean and noisy samples selected by our ACT on Web-Aircraft, Web-Bird, and Web-Car. The corresponding fine-grained class names are *Yak-42*, *Carolina Wren*, and *Tesla Model S Sedan 2012*.

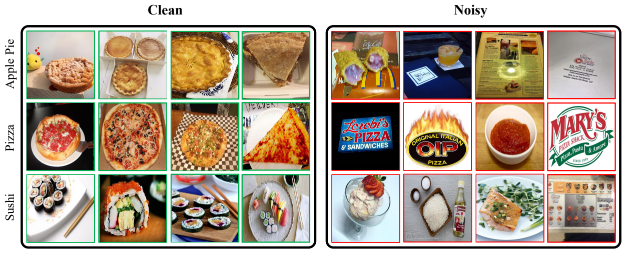
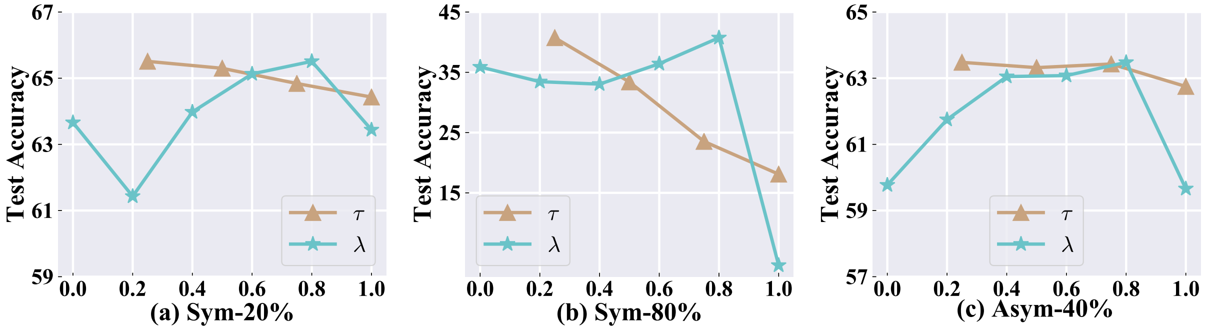


Figure 2: Some visualization results of clean and noisy samples selected by our ACT on Food101N. The corresponding fine-grained class names are *Apple Pie*, *Pizza*, and *Sushi*.

It is evident that our proposed asymmetric sample selection can effectively distinguish between clean and noisy samples. In particular, our ACT method proves effective in addressing the intricate and diverse noise scenarios present in real-world noisy datasets (*e.g.*, the open-set noise). Furthermore, our ACT does not necessitate prior knowledge like noise rates, thereby enhancing its practicality in real-world scenarios.

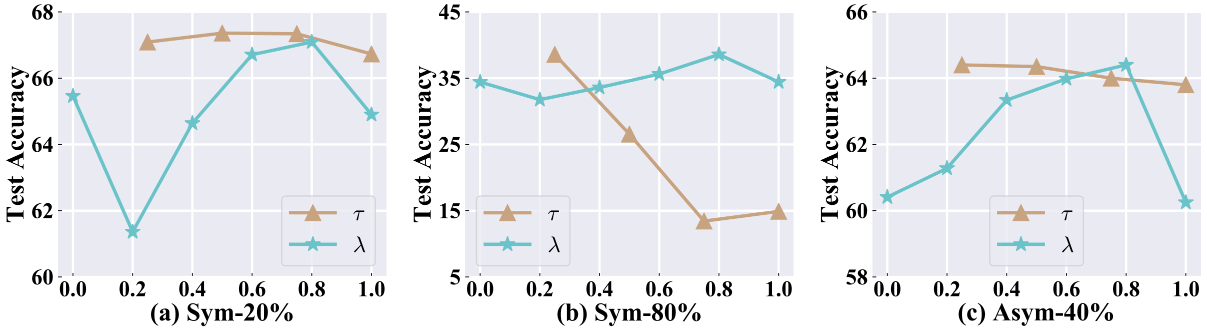
2 SENSITIVITY OF HYPER-PARAMETERS

In this section, we study the sensitivity of hyper-parameters (*i.e.*, τ and λ) utilized in our ACT. In our ACT framework, we introduce two novel criteria to select and mine clean samples more precisely. Criterion 2 indicates that before the NTM suffers from label memorization, more valuable clean samples can be mined for the RTM. To ensure the precision of the clean subset participating in the robust training of the RTM, we regulate the clean sample mining process using τ. τ serves as a threshold to measure the extent of label memorization for the NTM. It is evident that τ remains robust across a range of values from 0.0 to 1.0. Furthermore, λ primarily controls the weight of the regularization loss in Eq. (11). We present



(a) Sym-20% (b) Sym-80% (c) Asym-40%

Figure 3: Effects of the hyper-parameters (*i.e.*, λ and τ) on CIFAR100N with various noise settings (*i.e.*, Sym-20% (a), Sym-80% (b), and Asym-40% (c)).



(a) Sym-20% (b) Sym-80% (c) Asym-40%

Figure 4: Effects of the hyper-parameters (*i.e.*, λ and τ) on CIFAR80N with various noise settings (*i.e.*, Sym-20% (a), Sym-80% (b), and Asym-40% (c)).

the model performance under different τ and λ settings in Fig. 3 and Fig. 4 across various noise settings (*i.e.*, Sym-20%, Sym-80%, and Asym-40%) on CIFAR100N and CIFAR80N.

We can observe that a properly selected τ and λ can boost the model performance further. When τ is 0.8, and λ is 0.2, our ACT achieves the highest performance on the test set on synthetic noisy CIFAR100N and CIFAR80N across nearly all noise settings. It is noteworthy that in the Sym-80% noise settings, the higher noise rate results in heightened sensitivity and volatility in performance concerning hyper-parameters.

3 EXTENDED COMPARISON RESULTS

As discussed in the main paper, the SCT strategy finds wide application in existing sample selection methods [1–7]. SCT methods typically involve the simultaneous training of two networks with identical architectures but distinct weight initialization. Previous SCT methods have explored both agreement-based [1, 5] and disagreement-based [2, 7] sample selection strategies for addressing noisy labels. However, the information gains associated with SCT are significantly limited due to the divergence in capabilities between the paired networks primarily stemming from distinct initialization. This limitation reduces their accuracy in selecting clean samples, consequently leading to deteriorated model performance.

To further showcase the effectiveness of our ACT, we additionally evaluate the performance of our asymmetric sample selection and mining approach by comparing sample identification results with classic SCT methods (*i.e.*, Decoupling, Co-teaching, Co-teaching+, and JoCoR). This evaluation is conducted using precision, recall, and F1 score metrics on CIFAR100N and CIFAR80N datasets across different noise settings (as shown in Figs. 5-10). It is evident that the selection precision of our ACT surpasses that of other SCT methods by a significant margin. Although the recall of our ACT initially starts at a relatively lower level, reflecting the trade-off made to ensure the reliability of selected clean samples, it eventually exceeds that of its SCT counterparts. Furthermore, both the F1 score and test accuracy of our ACT consistently outperform those of existing SCT methods. Remarkably, our ACT maintains optimal performance even under the most challenging noise settings, such as CIFAR100N-Sym-80% and CIFAR80N-Sym-80%.

REFERENCES

[1] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*. 8536–8546.

[2] Eran Malach and Shai Shalev-Shwartz. 2017. Decoupling "when to update" from "how to update". In *Advances in Neural Information Processing Systems*. 960–970.

[3] Zeren Sun, Huafeng Liu, Qiong Wang, Tianfei Zhou, Qi Wu, and Zhenmin Tang. 2022. Co-LDL: A Co-Training-Based Label Distribution Learning Method for Tackling Label Noise. *IEEE Transactions on Multimedia* (2022), 1093–1104.

[4] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z. Li. 2021. Co-learning: Learning from Noisy Labels with Self-supervision. In *ACM International Conference on Multimedia*. 1405–1413.

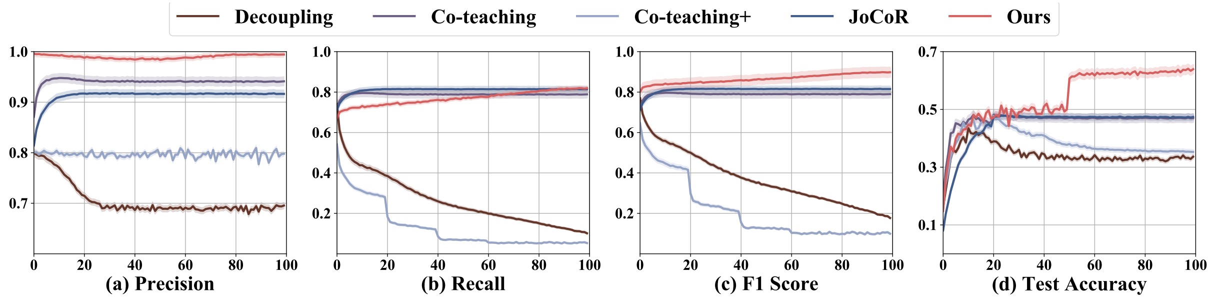


Figure 5: The comparison between SOTA methods and our ACT on precision, recall, F1 score, and test accuracy *vs.* epochs. Experiments are conducted on CIFAR100N with Sym-20%.

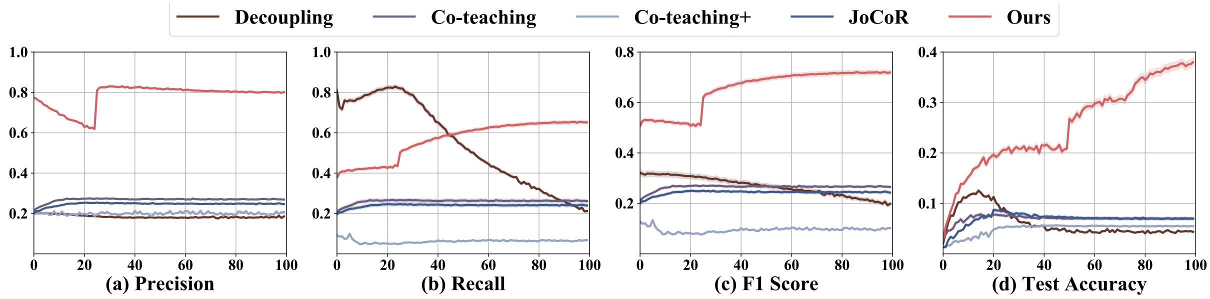


Figure 6: The comparison between SOTA methods and our ACT on precision, recall, F1 score, and test accuracy *vs.* epochs. Experiments are conducted on CIFAR100N with Sym-80%.

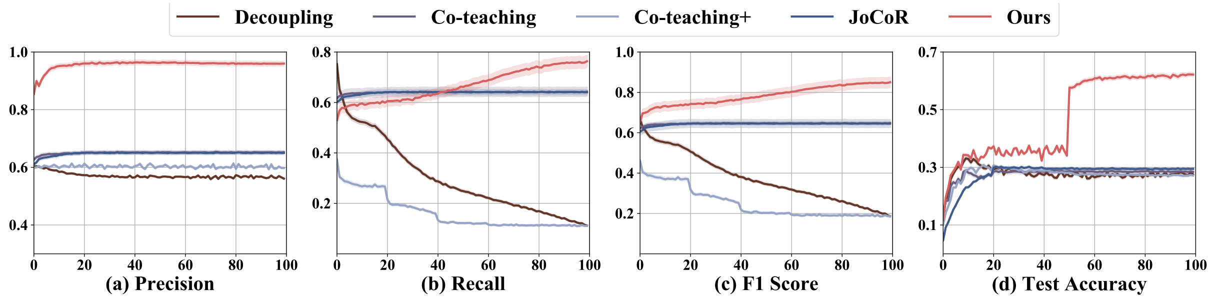
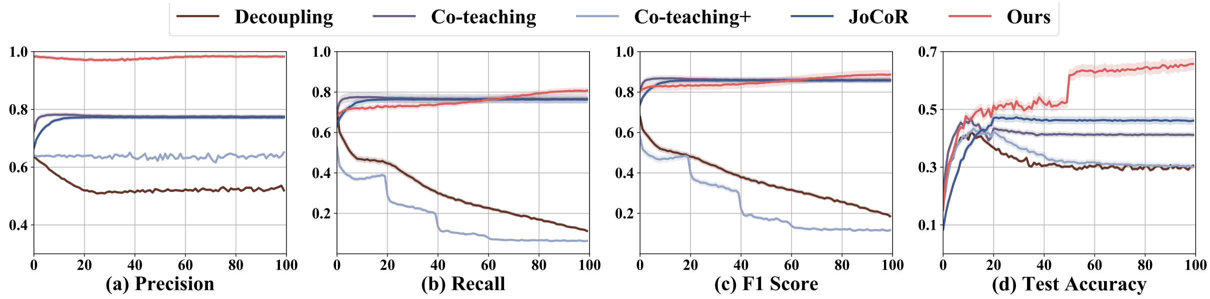


Figure 7: The comparison between SOTA methods and our ACT on precision, recall, F1 score, and test accuracy *vs.* epochs. Experiments are conducted on CIFAR100N with Asym-40%.

[5] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularization. In *IEEE Conference on Computer Vision and Pattern Recognition.* 13723–13732.

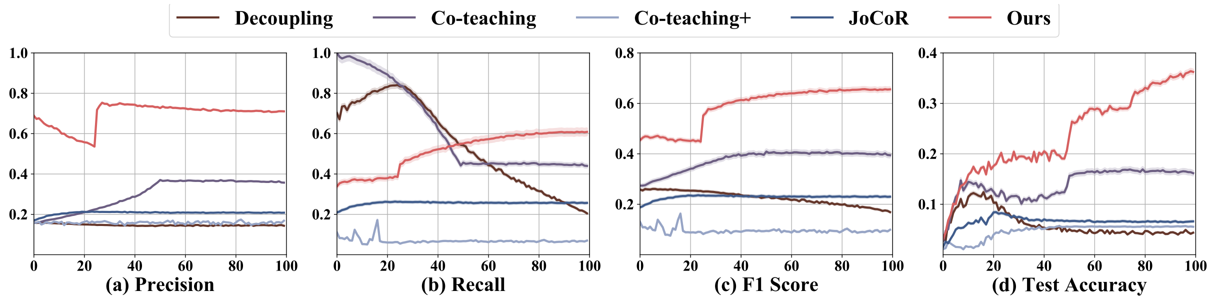
[6] Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. 2021. Jo-SRC: A Contrastive Approach for Combating Noisy Labels. In *IEEE Conference on Computer Vision and Pattern Recognition.* 5192–5201.

[7] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. 2019. How does Disagreement Help Generalization against Label Corruption?. In *International Conference on Machine Learning.* 7164–7173.



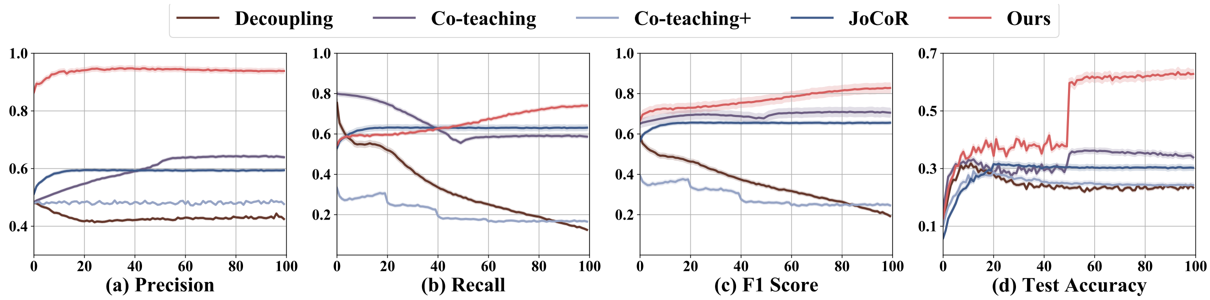
(a) Precision (b) Recall (c) F1 Score (d) Test Accuracy

Figure 8: The comparison between SOTA methods and our ACT on precision, recall, F1 score, and test accuracy *vs.* epochs. Experiments are conducted on CIFAR80N with Sym-20%.



(a) Precision (b) Recall (c) F1 Score (d) Test Accuracy

Figure 9: The comparison between SOTA methods and our ACT on precision, recall, F1 score, and test accuracy *vs.* epochs. Experiments are conducted on CIFAR80N with Sym-80%.



(a) Precision (b) Recall (c) F1 Score (d) Test Accuracy

Figure 10: The comparison between SOTA methods and our ACT on precision, recall, F1 score, and test accuracy *vs.* epochs. Experiments are conducted on CIFAR80N with Asym-40%.