Foster Adaptivity and Balance in Learning with Noisy Labels

Mengmeng Sheng[1], Zeren Sun[1](✉), Tao Chen[1], Shuchao Pang[1], Yucheng Wang[2], and Yazhou Yao[1](✉)

[1] Nanjing University of Science and Technology, Nanjing, China
{shengmengmemg, zerens, taochen, pangshuchao, yazhou.yao}@njust.edu.cn
[2] Horizon Robotics, Beijing, China
yucheng.wang@horizon.cc

In this supplementary material, we provide additional evaluations of our proposed SED. We further investigate the effect of the detailed composition of the SCS and SCR in our SED. Additional experiments are conducted and reported according to the following themes:

- **Local and Global Threshold in SCS.** Considering class balance during sample selection, our SCS adaptively adjusts the threshold in an epoch-wise and class-wise manner to enable effective clean sample identification. Table 1 shows the results of using SCS with and without local and global thresholds.
- **Truncated Normal Distribution in SCR.** Considering the confidence of the corrected label of noisy samples, we propose a dynamic truncated normal distribution as a sample re-weighting function to mitigate the biased label correction. Table 1 also shows results of re-weighting and non-reweighting experimental setups.
- **Comparison with existing sample selection methods.** To provide a more comprehensive visualization of the adaptive and class-balanced sample selection process of our SED, we conducted additional experimental analyses. Fig.1 illustrates the selection thresholds and precision of different sample selection methods (*i.e.*, Small-loss, GMM and ours SED) for each class on CIFAR100N-Sym-50%.
- **Effect of Hyper-parameters in EMA.** To dynamically reflect the learning performance changes of the model, we use the exponential moving average (EMA) to dynamically update self-adaptive thresholds in SCS, truncated normal distribution parameters in SCR, and the mean-teacher model (θ^*). The results of the ablation experiments are presented in Table 2 and Table 3.
- **Extend Experiment Results.** To further verify the robustness of our SED as the training processes, Fig. 2 and Fig. 3 record the test accuracy of the training process under different noise conditions on CIFAR100N and CIFAR80N. Fig. 4, Fig. 5, and Fig. 6 present some sample selection results of our SED on real-world datasets.

1 Futher Analysis

In this section, we investigate the impact of the detailed composition of the SCS and SCR in our SED. As outlined, our proposed SED selects and re-weights

Table 1: Effect of each component in test accuracy (%) on CIFAR100N (Sym-50%).

Model	Test Accuracy
Standard	34.10
Standard+SCS w/o local threshold	53.36
Standard+SCS w/o global threshold	55.64
Standard+SCS w/o EMA	54.72
Standard+SCS	58.21
Standard+SCR	54.42
Standard+CR	55.51
Standard+SCS+SCR w/o re-weighting	59.75
Standard+SCS+SCR w/o EMA	60.08
Standard+SCS+SCR	60.43
Standard+SCS+CR	59.28
Standard+SCR+CR	55.98
Standard+SCS+SCR+CR	62.65

samples based on class-specific thresholds that are calculated in a data-driven manner. This helps to promote self-adaptivity and balance in both sample selection and re-weighting. Specifically, we mainly explore the effect of local thresholds in our SCS and the dynamic truncated normal distribution in our SCR during the training process. Besides, we employ the exponential moving average (EMA) to further refine these thresholds to alleviate unstable training. Thus, we further explore the influence of the hyper-parameter (*i.e.*, m and α) of the exponential moving average in SCS, SCR, and the mean-teacher model θ^*.

1.1 Local and Global Thresholds in SCS.

Our proposed SCS adaptively adjusts the threshold in an epoch-wise and class-wise manner to enable effective clean sample identification. As shown in Table 1, we provide the result of using SCS without local and global thresholds. The performance drops by 4.85% and 2.57% accordingly. This proves that our local threshold design is crucial for improving the robustness of the model. By considering class balance during sample selection, our SCS can significantly alleviate the learning bias of the model.

Furthermore, our SCS employs the exponential moving average (EMA) [6] to further refine the global and local threshold. This helps to alleviate the issue of unstable training caused by large perturbation of the averaged predicted probability. As indicated in Table 1, it is noticeable that the utilization of EMA has resulted in a significant increase (*i.e.*, 3.49%) in model performance. This clearly demonstrates the advantageous impact of EMA on our SCS.

1.2 Truncated Normal Distribution in SCR.

Considering the confidence of the corrected label of noisy samples, we propose a dynamic truncated normal distribution as a sample re-weighting function to mitigate the biased label correction. Samples with higher correction confidence

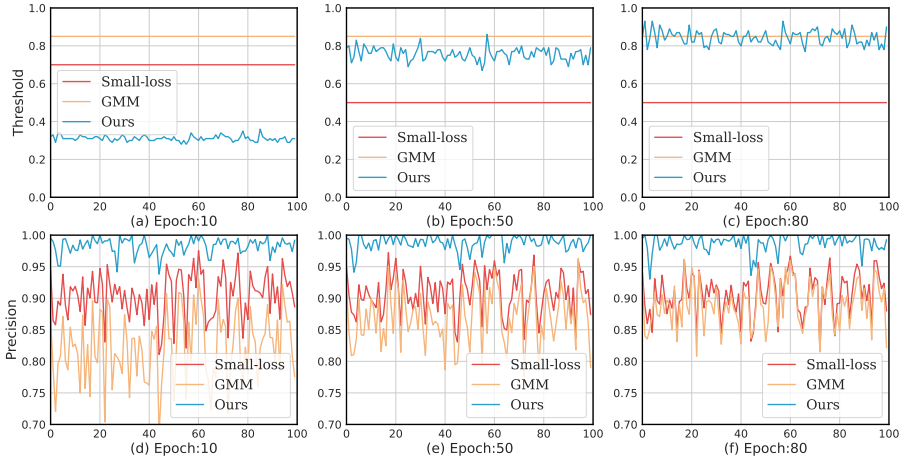

Fig. 1: The selection threshold (a-c) and precision (c-d) of different sample selection methods for each class on CIFAR100N-Sym-50% *vs.* epochs.

are less likely to be incorrectly labeled compared to those with lower confidence, thus being assigned larger weights. We further perform an ablation study to explore the effectiveness of re-weighting corrected labels based on their confidence. As shown in Table 1, re-weighting corrected labels improves the performance by 0.68% compared to the without-reweighting case.

In our proposed SCR, we employ a dynamic truncated normal distribution, whose mean and variance values are μ_t and σ_t at the t-th epoch, to assign weights for different samples. Moreover, to enable class-balanced re-weighting and promote training stability, we propose to estimate $\mu_t(c)$ and $\sigma_t^2(c)$ for each class c based on their historical estimations using EMA. Table 1 shows that employing SCR with EMA achieves an additional 0.35% performance gain compared to SCR without EMA.

1.3 Comparison with existing sample selection methods.

Due to memorization effect [1], prior sample selection methods tend to regard samples with small losses as clean ones. Some other methods utilize Gaussian Mixture Model (GMM) to partition losses, as seen in DivideMix [2]. However, these methods often require proper prior knowledge (*e.g.*, a pre-defined drop rate or threshold) to achieve effective sample selection. To promote self-adaptivity and class balance in sample selection, we propose to integrate global and local thresholds for each category when distinguishing between clean and noisy data in our SED. Fig. 1 illustrates the selection thresholds and precision of different sample selection methods (*i.e.*, Small-loss, GMM and ours SED) for each class on CIFAR100N-Sym-50%.

As shown in Fig. 1 (a-c), thresholds in our SED are class-dependent and dynamically adjust with epoch progression compared to existing approaches.

Table 2: Effects of hyper-parameter m for SCS and SCR in test accuracy (%) on CIFAR100N (noise rate and noise type are 0.5 and symmetric, respectively).

m **for SCS and SCR**	α **for the mean-teacher model**	**results**
0.85	0.85	57.82
0.90	0.85	57.92
0.95	0.85	59.41
0.99	0.85	62.63
0.999	0.85	61.60

Table 3: Effects of hyper-parameter α for the mean-teacher model on CIFAR100N (noise rate and noise type are 0.5 and symmetric, respectively).

m **for SCS and SCR**	α **for the mean-teacher model**	**results**
0.85	0.90	57.73
0.85	0.95	58.04
0.85	0.99	57.46
0.85	0.999	57.86

Fig. 1 (d-f) presents a detailed comparison of the precision between our method and existing sample selection methods. Clearly, our SED demonstrates a more balanced selection precision across categories. Furthermore, upon comparing the selection precision of GMM across different epochs, selection results of our SED exhibit greater stability.

1.4 Effect of Hyper-parameters in EMA.

Considering that the model is inevitable to fit some noisy samples in the later stage of training, we resort to the exponential moving average (EMA) to achieve more reliable sample selection, label correction, and sample re-weighting. EMA introduces the model's results in the historical iteration to increase the stability of the training process.

To dynamically reflect the learning performance changes of the model, we use EMA to update our global and local thresholds in SCS dynamically and the truncated normal distribution parameters in SCR. Besides, our mean-teacher model (θ^*) is also updated in an EMA manner. The exponential moving average strategy requires a factor (*i.e.*, m for SCS and SCR, α for the mean-teacher model) to balance the weight of past and current results, thus ensuring the stability of the training process. As shown in Table 1, employing SCS with EMA and employing SCR with EMA both achieve nontrivial performance gains.

Table 2 shows the test accuracy of our SED under different m for SCS and SCR with a fixed α (*i.e.*, 0.85) on CIFAR100N-Sym50%. It can be observed that the best performance is achieved when $m = 0.99$. Table 3 further shows the test accuracy of our SED under different α for the mean-teacher model with a fixed

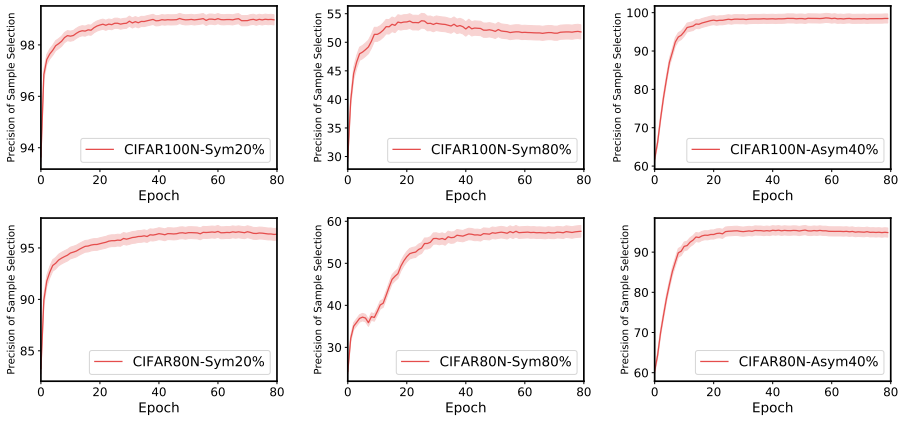

Fig. 2: The overall precision of sample selection (%) *vs.* epochs on CIFAR100N and CIFAR80N. Experiments are conducted under various noise conditions ("Sym" and "Asym" denote the symmetric and asymmetric label noise, respectively).

m (*i.e.*, 0.85) on CIFAR100N-Sym50%. As we observe from Table 3, the varying alpha values have only a minor effect on the performance of our SED. The best performance is obtained when $\alpha = 0.95$.

2 Extend Experiment Results

Due to the memorization effect [1] (*i.e.*, models tend to fit clean and simple samples first and then gradually memorize noisy ones), the network optimization based on the cross-entropy loss usually leads to an ill-suited solution. During our experiments, CE refers to the conventional training baseline that utilizes the entire noisy dataset with the cross-entropy loss. As shown in Fig. (3) (c), the test accuracy of CE first increases to a certain level and then decreases due to overfitting. Fig. (3) (c) has already shown that robust methods for noisy labels like our SED are able to mitigate the impact of overfitting noisy samples in the later stage of training. The test accuracy of these robust methods increases monotonously as the training continues.

Fig. 3 further verifies the robustness of our SED as the training processes. Fig. 3 records the test accuracy of the training process under different noise conditions (*i.e.*, "Sym" and "Asym" denote the symmetric and asymmetric label noise, respectively. 20%, 40%, and 80% denote the noise rate.) on CIFAR100N and CIFAR80N. It can be observed that the test accuracy continues to grow under various noise conditions. This explicitly proves that our method can prevent the model from overfitting the noisy samples, thereby enhancing its robustness.

To further verify the effectiveness of our proposed SED in practical scenarios, we conduct experiments on three real-world noisy datasets (*i.e.*, Web-Aircraft, Web-Car, and Web-Bird [3]), whose training images are crawled from web image

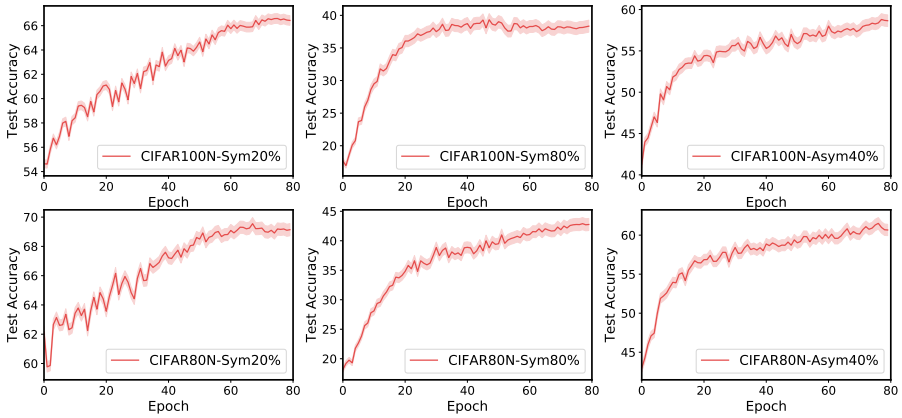

Fig. 3: The test naccuracy (%) *vs.* epochs on CIFAR100N and CIFAR80N. Experiments are conducted under various noise conditions ("Sym" and "Asym" denote the symmetric and asymmetric label noise, respectively).

search engines. Fig. 4, Fig. 5, and Fig. 6 show the comparison of the sample selection results between our SED and two state-of-the-art (SOTA) methods (*i.e.*, JoCoR [5] and Co-LDL [4]) on these three real-world datasets. These sample selection methods primarily seek to split samples into two subsets: a noisy subset and a clean subset. The samples marked with purple represent the incorrectly selected samples. The samples marked with green are the correctly selected samples. If the sample selection is correctly performed in all three methods, the corresponding mark is displayed in black. Fig. 4, Fig. 5, and Fig. 6 visually illustrate the excellent performance of our SED.

References

1. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A.C., Bengio, Y., Lacoste-Julien, S.: A closer look at memorization in deep networks. In: Int. Conf. Mach. Learn. pp. 233–242 (2017)
2. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: Int. Conf. Learn. Represent. (2020)
3. Sun, Z., Hua, X.S., Yao, Y., Wei, X.S., Hu, G., Zhang, J.: Crssc: salvage reusable samples from noisy data for robust learning. In: ACM Int. Conf. Multimedia. pp. 92–101 (2020)
4. Sun, Z., Liu, H., Wang, Q., Zhou, T., Wu, Q., Tang, Z.: Co-ldl: A co-training-based label distribution learning method for tackling label noise. IEEE Trans. Multimedia pp. 1093–1104 (2022)
5. Wei, H., Feng, L., Chen, X., An, B.: Combating noisy labels by agreement: A joint training method with co-regularization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13723–13732 (2020)
6. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., Tang, Z.: Jo-src: A contrastive approach for combating noisy labels. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5192–5201 (2021)

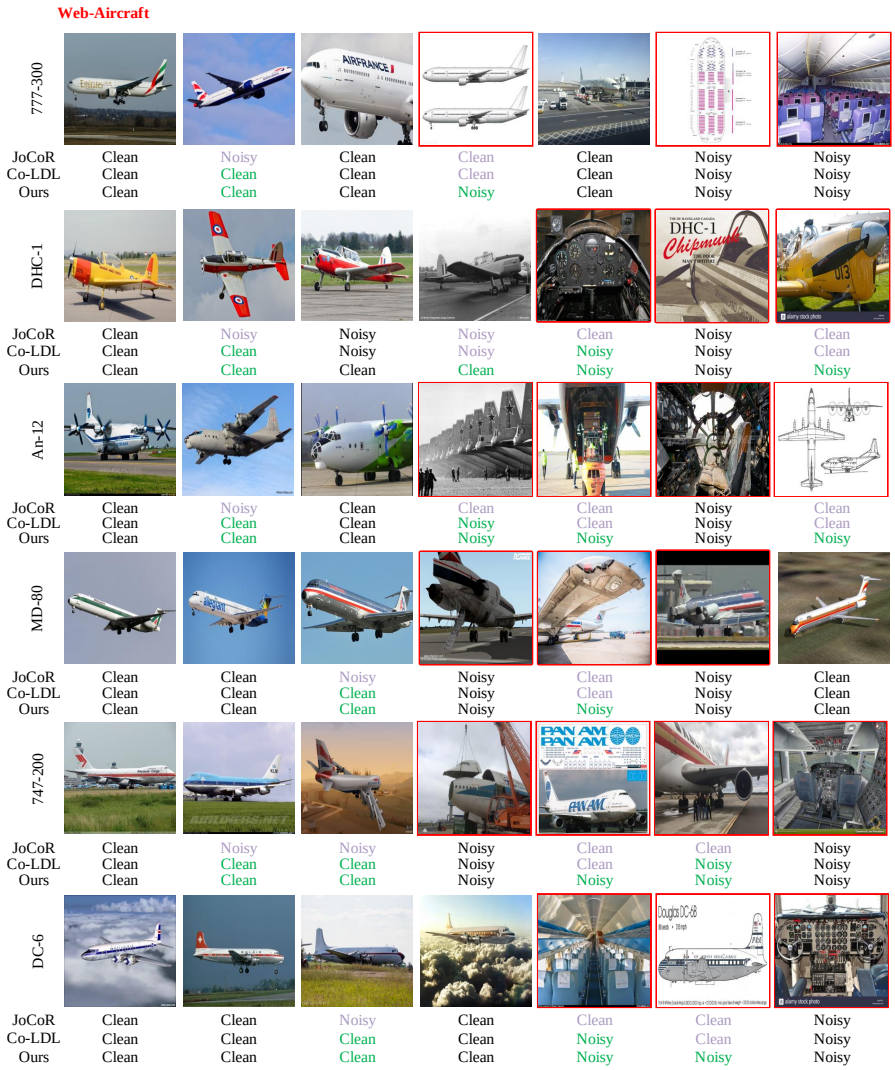

Fig. 4: The comparison of sample selection results with SOTA approaches on Web-Aircraft. Each row shows seven samples that are randomly selected from the same fine-grained category. Images in the red box indicate noisy samples.

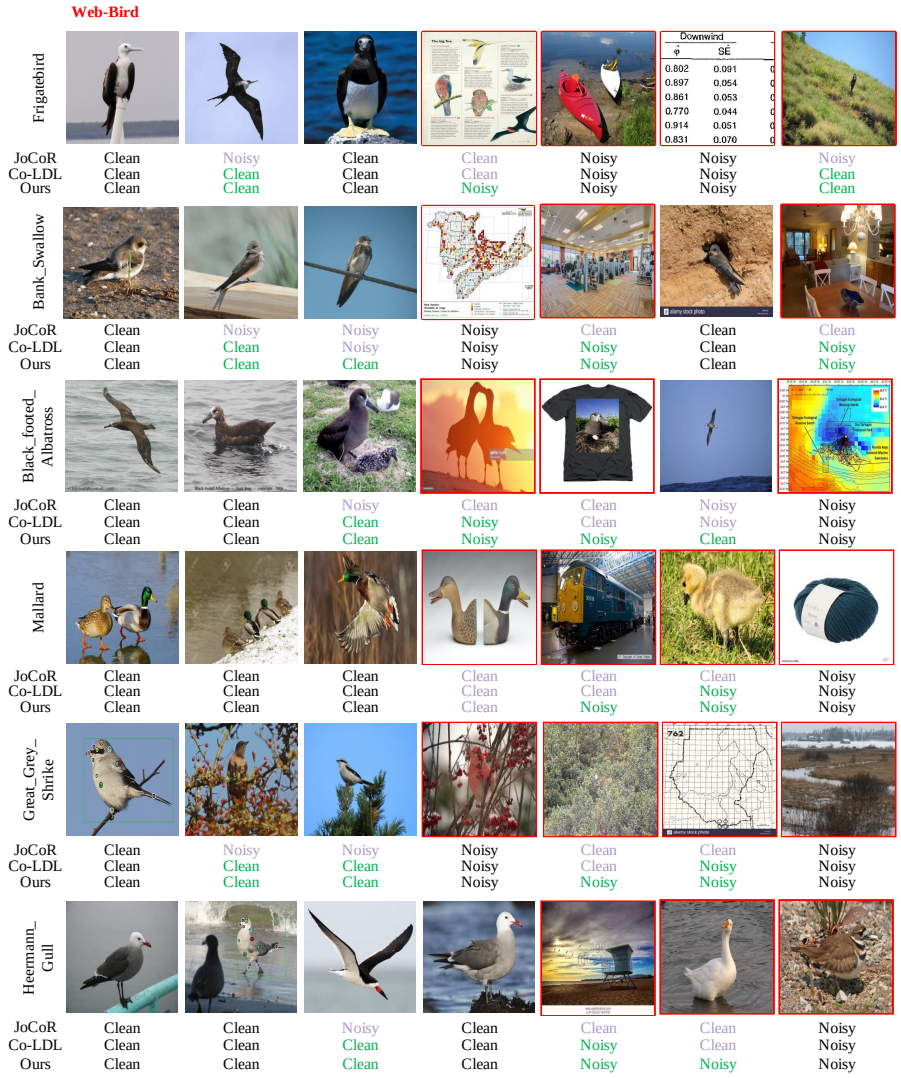

Fig. 5: The comparison of sample selection results with SOTA approaches on Web-Bird. Each row shows seven samples that are randomly selected from the same fine-grained category. Images in the red box indicate noisy samples.

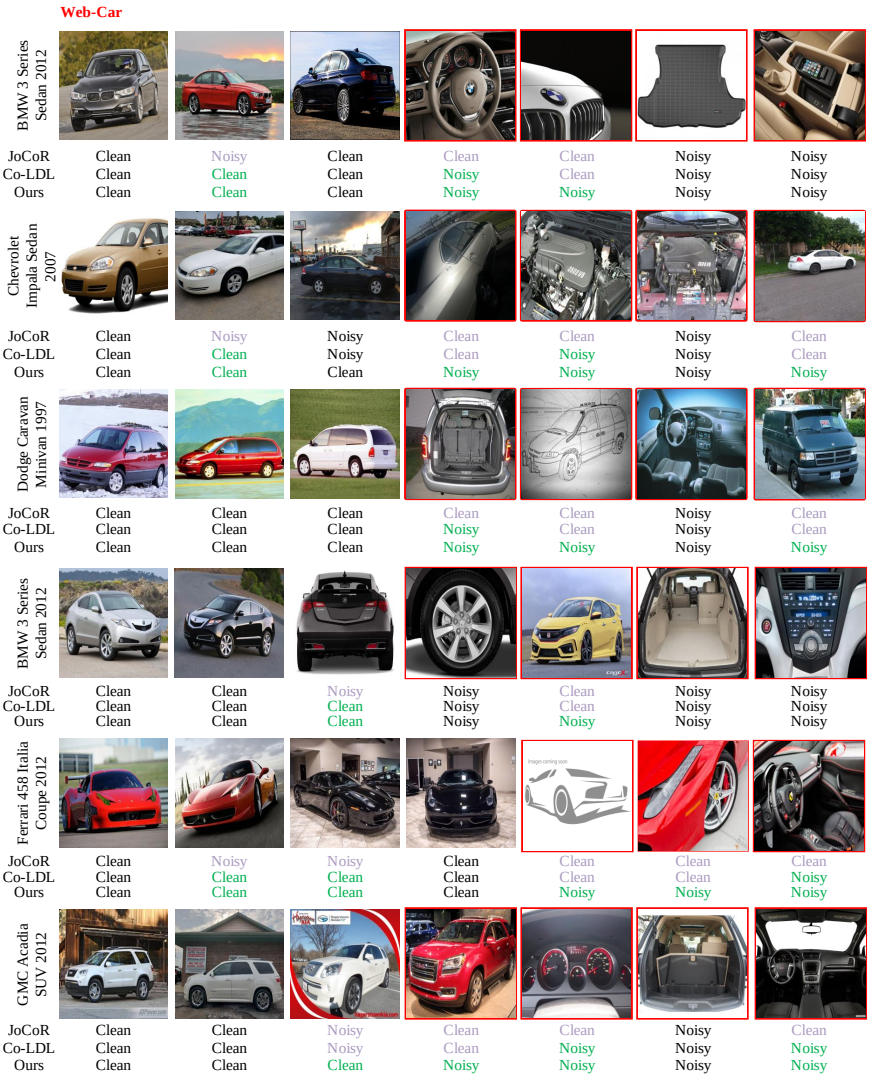

Fig. 6: The comparison of sample selection results with SOTA approaches on Web-car. Each row shows seven samples that are randomly selected from the same fine-grained category. Images in the red box indicate noisy samples.