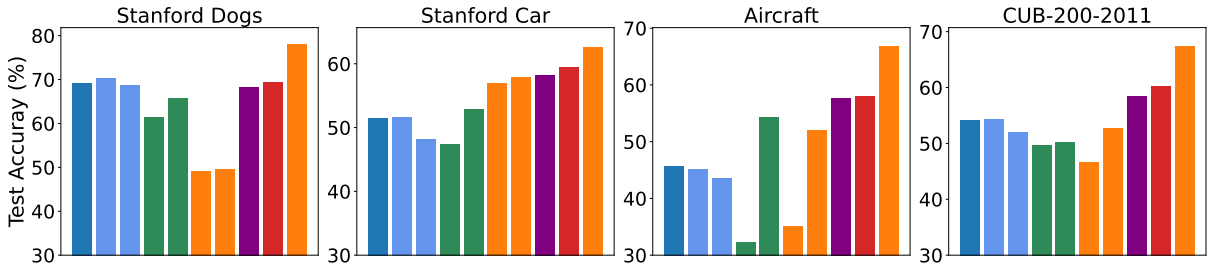
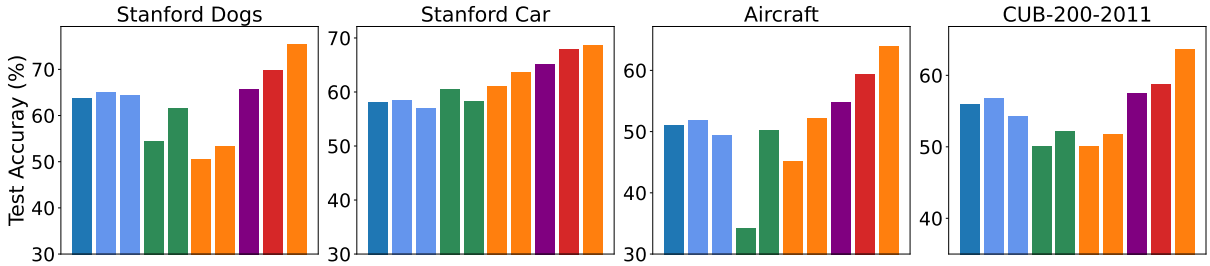
Appendix for Fine-Grained Classification with Noisy Labels



(a) *40% Symmetric noisy labels*



(b) *30% Asymmetric noisy labels*

Figure 1. Ten tested methods (left → right): cross-entropy, label smooth, confidence penalty, GCE, SYM, Co-teaching, JoCoR, MW-Net, MLC, DivideMix. Methods with same color belong to same LNL robust strategy. The **x-axis** denotes their performance on typical LNL task while the performance increases gradually from left to right.

Table 1. An ablation study on 40% symmetric noisy labels. The performance of Co-teaching can be improved by several robust techniques and gets close to the performance of DivideMix (the SOTA).

Co-teaching	Mixup	Pseudo-label	Conf. reg.	EMA	Stanford Dogs	CUB-200-2011
✓					49.15 (48.92)	46.57 (46.22)
✓	✓				62.79 (60.10)	54.04 (53.09)
✓	✓	✓			72.44 (71.97)	65.77 (63.91)
✓	✓	✓	✓		75.21 (73.94)	66.47 (65.73)
✓	✓	✓	✓	✓	77.84 (77.41)	**67.64** (67.20)
DivideMix					**77.93** (76.28)	67.35 (66.96)

A. A prior study

In this section, we conduct a preliminary investigation to evaluate the performance of current LNL on LNL-FG. We adpot pre-trained ResNet-18 as the backbone and set varying noisy conditions. Fig. 2 in Introduction, Figure 1 and Table 1 exhibit the qualitative results. Our finds are divided into two parts,

- **Not all investigated algorithms can achieve significant performance for LNL-FG as they achieved in LNL, demonstrating the difficulty of fine-grained noisy settings.** In Stanford Dogs and CUB-200-2011, Cross-entropy, a non-robust method, attains competitive generalization performance while outperforming more than half methods. The insufficient robustness of these methods empirically demonstrates that LNL-FG poses a more challenging noisy condition for model learning and has not attracted much attention.
- **The generalization performance of LNL methods heavily relies on techniques that can mitigate overfitting on noisy labels.** In Table 1, we select co-teaching, a method with poor performance on LNL-FG, and add

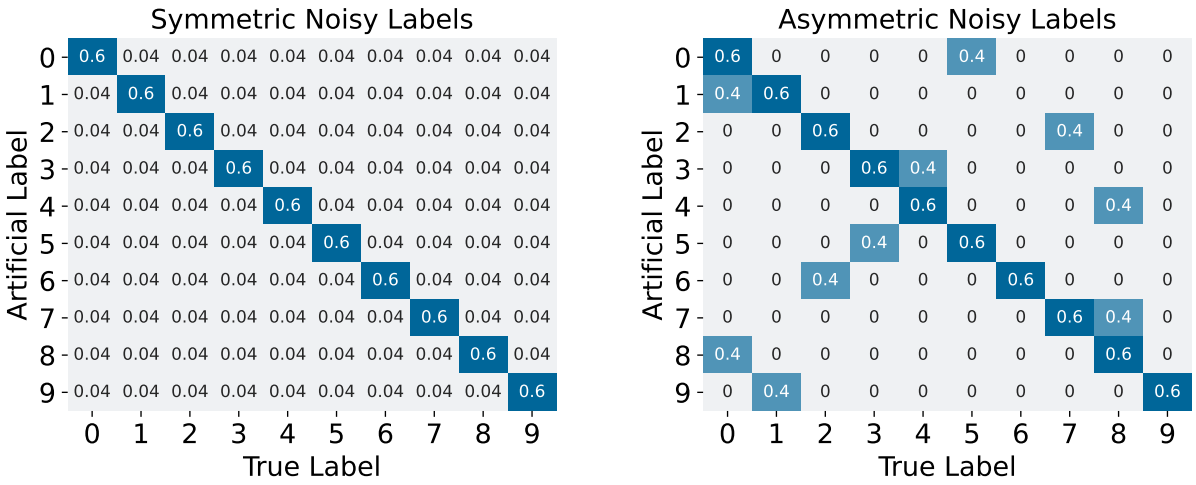


Figure 2. **Illustration** of symmetric and asymmetric noise transition matrix. There is a 10-classes classification task and the noise ratio is set as $r = 0.4$.

four robust techniques step by step. Each technique improves the performance of the basic method. Combining with Mixup, pseudo-label, confidence regularization, and EMA model, top-1 testing accuracy of Co-teaching on the clean test set of *Stanford Dogs* is improved by 13.64%, 9.65%, and 2.77%, respectively. However, integrating these existing techniques into the training process is difficult or requires customized adaptations, inspiring us to design a general method which can be applied to current LNL methods for improving their performance on LNL-FG.

B. More implementation details

B.1. Noise transition matrix

We give an illustration of two types of the transition matrix in Figure 2.

B.2. Settings of benchmarks

In this work, we select four fine-grained datasets, two generic datasets, and one open-world noisy set to verify the effectiveness of our method. The detailed information of these benchmarks is shown in Table 2. For validation and hyper-parameter adjustment, we reserve 10% clean training samples and construct noisy benchmark on the rest samples. Besides, we adopt weak augmentation strategies for comparison, including randomly cropping from 255×255 to 224×224 and horizontally flipping.

Table 2. Statistic information of the benchmarks and relevant settings.

Datasets	# Train	# Test	# Classes	# Size	# Features	Model	# Warmup	# Epochs
Fine-grained set						(pre-trained)		
Aircraft [9]	6667	3333	100	224	512	ResNet-18	10	100
CUB-200-2011 [12]	5994	5794	200	224	512	ResNet-18	10	100
Stanford-Cars [5]	8114	8441	196	224	512	ResNet-18	10	100
Stanford-Dogs [4]	12000	8580	120	224	512	ResNet-18	5	100
Generic set								
CIFAR-10 [6]	50000	10000	10	32	512	PreAct ResNet-18	10	100
CIFAR-100 [6]	50000	10000	100	32	512	PreAct ResNet-18	20	100
Real-world set						(pre-trained)		
Food-101N [2]	55000	25000	101	224	2048	ResNet-50	5	50
Clothing-1M [17]	1000000	10000	14	224	2048	ResNet-50	1	15

B.3. Settings of comparison methods

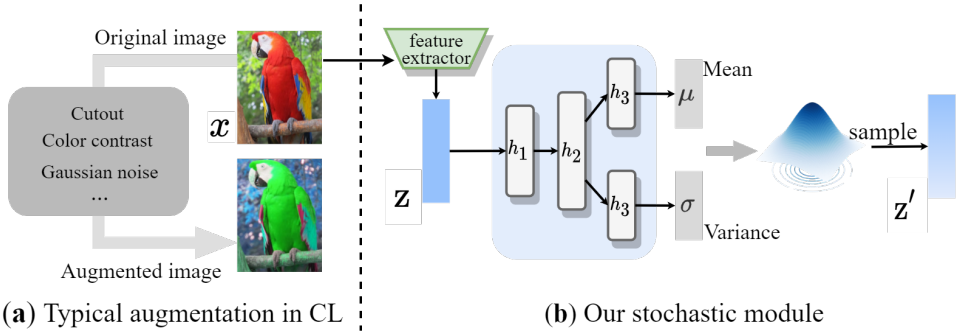
We compare our proposal with cross-entropy loss function and the following baselines:

- **Label smooth** [8], which reassigns the sample label from a hard version to a soft version like $\{0, 0, 1\} \rightarrow \{0.05, 0.05, 0.9\}$. This method confronts the effects of noisy labels by mitigating over-confidence of the model on the given label.
- **Confidence penalty** [10], which stems from the motivation of penalizing low entropy output distributions. It connects a maximum entropy based confidence penalty to label smoothing through the direction of the KL divergence.
- **GCE** [18], which analyzes the robustness of MAE and the poor performance. Then, the author presents a theoretically grounded set of noise-robust loss functions that can be seen as a generalization of MAE and CCE.
- **SYM** [13], which obeys the paradigm of the symmetric loss function that ensembles CE and reversed CE. The latter is demonstrated as a robust loss function.
- **Co-teaching** [3], which ensembles two branches for alternatively selecting samples with small losses and feeds them to another network training. *Co-training* strategy alleviates the error accumulation of the selection to some degree.
- **JoCoR** [14], which leverages the framework of Co-teaching and further designs a KL term for consistent output of two networks. It explores the lower bound of small loss and prompts accurate selection.
- **MW-Net** [11], which designs a meta-network for generating the sample weight via learning a function from loss to weight. The meta-weight is inserted into the training of the classification network by bi-level strategy.
- **MLC** [19], which also designs a meta-network for label correction. It learns from the original label and feature embeddings and outputs the corrected label.
- **DivideMix** [7], which belongs to a hybrid approach that bases on *sample selection* and ensembles co-training, pseudo-labeling, and Mixup. It attains state-of-the-art performance on LNL.

For fair comparisons, we keep the same hyper-parameters as they reported in their published versions, where some settings are marginally adjusted, and we report them in table 3. In addition, we adjust the selection process in Co-teaching [3] and JoCoR [14] when combines with our algorithm. Since our algorithm changes the original noise ratio in the training set, we replace the pre-estimation of the noise ratio with the dynamic strategy (i.e., GMM fits the losses among all samples).

Table 3. Detailed settings of compared methods in experiments.

Method	Settings
SYM [13]	SYM $= \alpha \times$ CE $+ \beta \times$ RCE, where $\alpha = 0.1, \beta = 1$
Label Smooth [8]	smooth coefficient $\lambda = 0.1$
MW-Net [11]	extra clean sample number $N = 5\times$ category number
MLC [19]	extra clean sample number $N = 5\times$ category number



(**a**) Typical augmentation in CL (**b**) Our stochastic module

Figure 3. **Illustration** of stochastic module. Compared to typical augmentation strategies in contrastive learning, we replace it with a stochastic module. Original feature embedding $\mathbf{z}_i$ is input into a stochastic network. Then, the augmented embedding $\mathbf{z}_i'$ can be sampled from the generated distribution. We consider that the property of stochastic provides more complex feature transformation than typical augmentation in images space, as well as avoiding manually defined augmentation strategies for different datasets.

C. More results and analysis

C.1. Stochastic module *vs.* Deep VIB

In this paper, we build a stochastic module for learnable feature transformation, which is constructed as a three-layer MLP structure as shown in Figure 3 (**b**). The architecture of stochastic module is similar to *deep VIB* [1] and we give the difference between these two methods as follows.

The main difference is that we incorporate a stochastic embedding module into the contrastive learning (CL) framework, which can sample a feature vector for CL and avoid sophisticated data augmentation in typical CL. But *Deep VIB* conducts Monte Carlo estimation of the expectation of the conditional prediction distribution for supervised learning. *VIB* has been proved as a better regularizer compared with other forms (e.g., confidence penalty & label smoothing) and can naturally benefit LNL in the supervised learning framework.

C.2. MLP structure of stochastic module

We actually have tried different MLP architecture settings in the following experiments. Table 4 exhibits the comparison results with five structures. It can be seen that varying MLP settings do not remarkably affect the final results. Therefore, we prefer to adopt the simple yet effective one, *i.e.*, $\{h_1, h_2, h_3\} = \{512, 2048, 512\}$. Compared to the the backbone whose params is around 11.9 M, the learnable params of this module is only 0.06 M, which do not cause complex computation.

Table 4. Test accuracy (%) of *CE + SNSCL* with different MLP architecture on 40% symmetric noisy labels. The average best score among three times are reported.

Architecture $\{h_1, \cdots, h_n\}$	Stanford Dogs	CUB-200-2011	Aircraft	Stanford Cars
512 - 1024 - 512	74.79	68.79	70.30	76.44
512 - 2048 - 512	**75.27**	68.83	**70.48**	**76.72**
512 - 4096 - 512	75.01	**69.09**	70.19	76.51
512 - 1024 - 1024 - 512	74.96	68.66	69.84	75.90
512 - 2048 - 2048 - 512	75.04	69.00	69.07	75.61

C.3. Results on LNL task

Table 5. Comparisons with test acc. (%) on **generic** classification task. The solid results denote the improvement of our method SNSCL. The average results among three times are reported.

	CIFAR-10		CIFAR-100	
	Symm. 40%	Asym. 40%	Symm. 40%	Asym. 40%
Peer Loss [11]	84.29 / **92.21**	85.18 / **91.59**	50.53 / **69.82**	50.17 / **68.90**
JoCoR [14]	85.44 / **92.70**	83.91 / **91.41**	55.97 / **71.44**	50.97 / **69.89**
CDR [16]	86.13 / **93.83**	85.79 / **92.08**	60.18 / **71.95**	59.49 / **71.57**
SFT [15]	89.54 / **94.59**	89.93 / **94.27**	69.72 / **74.52**	69.29 / **73.19**
DivideMix [7]	94.80 / **95.92**	93.40 / **94.90**	74.92 / **76.04**	72.10 / **75.16**

C.4. Visualization

Figure 4 demonstrates that SNSCL improves the representation ability of the feature extractor under noisy CIFAR-10 & 100 and achieves more distinguishable class representation.

C.5. Robust learning curves

Figure 5 shows the robust learning curves of our algorithm under all noise conditions.

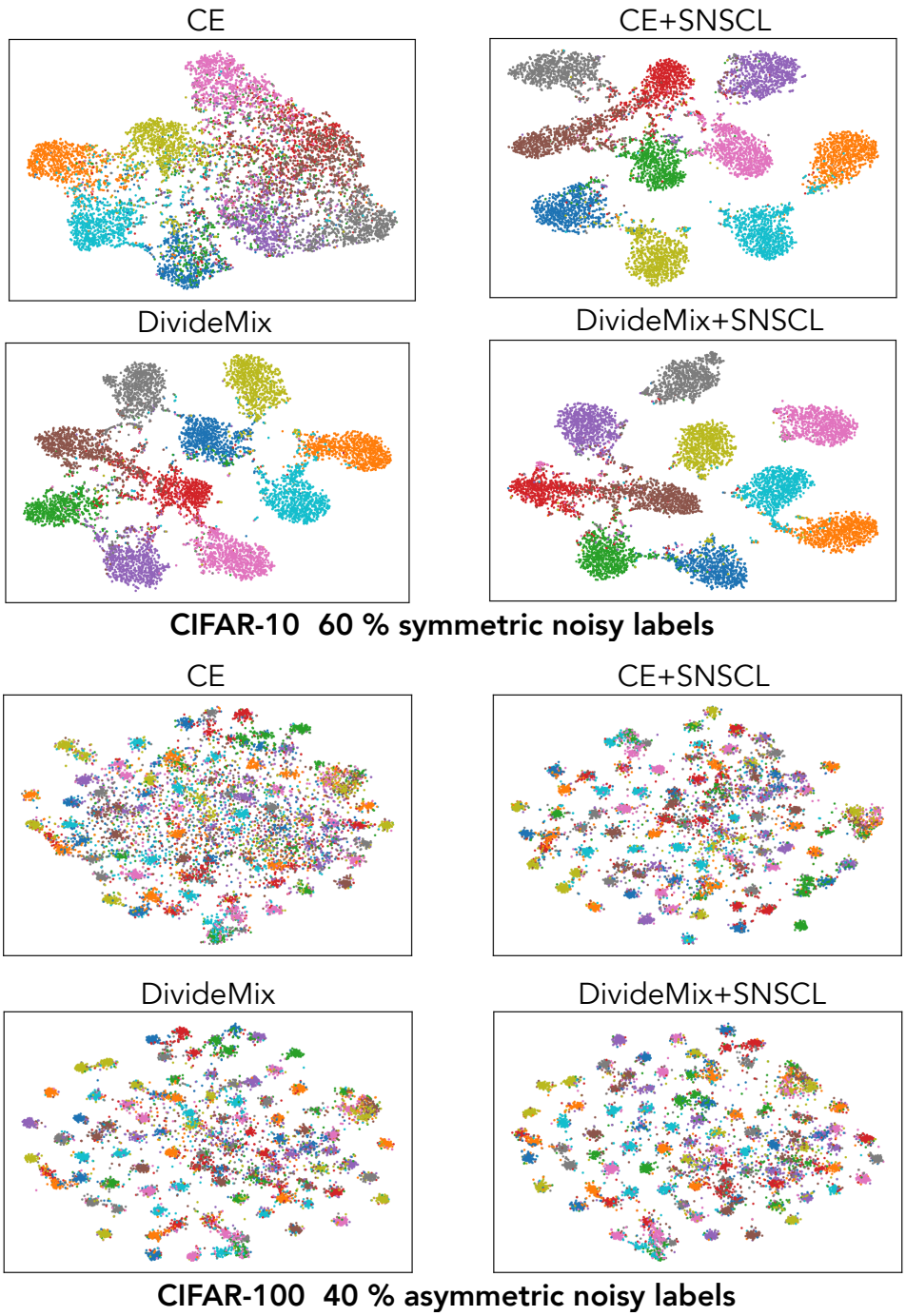
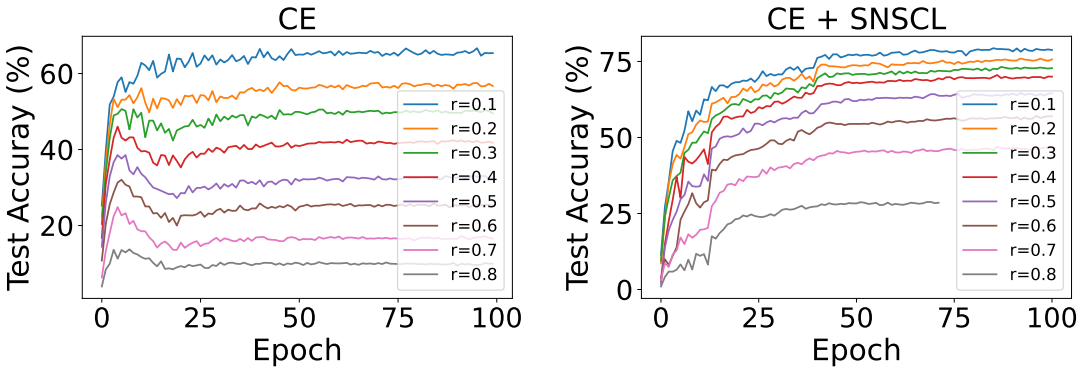
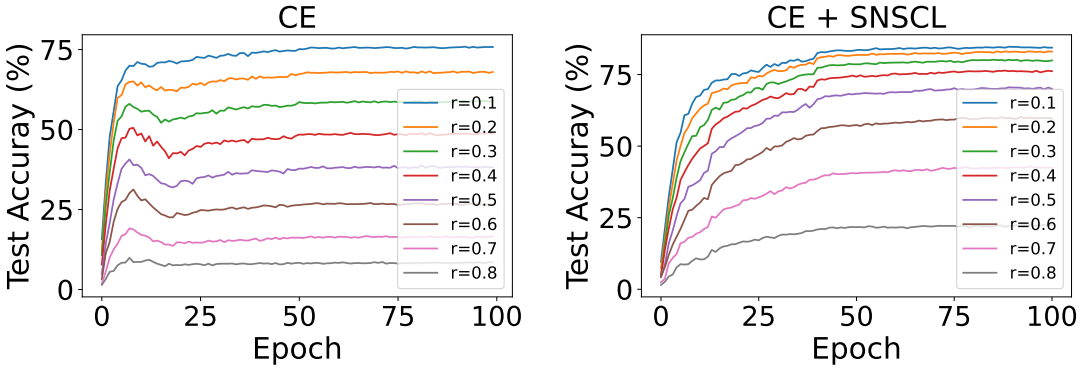


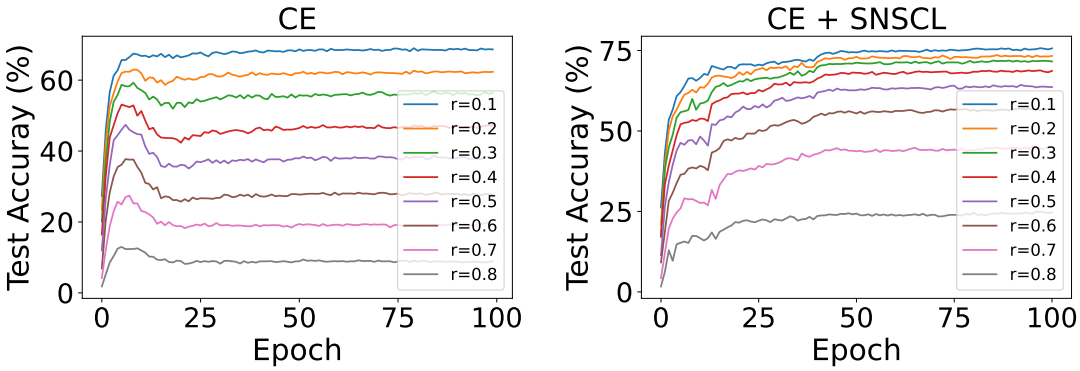
Figure 4. t-SNE visualization for feature representation.



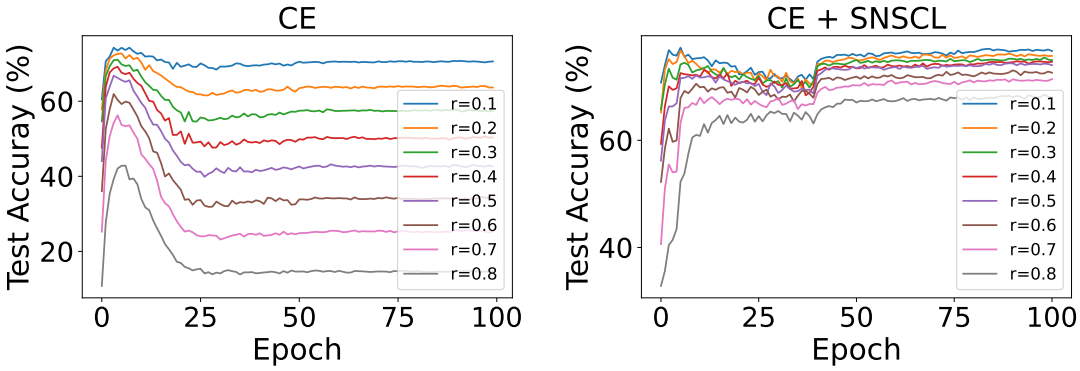
(a) **Stanford Car**, *Symmetric label noise, ResNet-18*



(b) **Aircraft**, *Symmetric label noise, ResNet-18*



(c) **CUB-200-2011**, *Symmetric label noise, ResNet-18*



(d) **Stanford Dogs**, *Symmetric label noise, ResNet-18*

Figure 5. **Comparisons with training curves as noise ratio increases**. We detailedly plot more training results about test accuracy (%) vs. noise ratio r, where there is symmetric noise and $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Under all noise ratio, SNSCL both remarkably improve the performance of the baseline.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. 4

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 2

[3] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018. 3

[4] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *CVPR workshop*, 2011. 2

[5] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013. 2

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Citeseer, 2009. 2

[7] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2020. 3, 4

[8] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, 2020. 3

[9] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. In *arXiv preprint arXiv:1306.5151*, 2013. 2

[10] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR*, 2017. 3

[11] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 3, 4

[12] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. In *California Institute of Technology*, 2011. 2

[13] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019. 3

[14] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 2020. 3, 4

[15] Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *ECCV*, 2022. 4

[16] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*, 2021. 4

[17] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015. 2

[18] Zhilu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *NeurIPS*, 2018. 3

[19] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *AAAI*, 2021. 3