Learning from Noisy Labels with Decoupled Meta Label Purifier Supplementary Material

Yuanpeng Tu[1*] Boshen Zhang[2*] Yuxi Li[2*] Liang Liu[2] Jian Li[2]
Yabiao Wang[2] Chengjie Wang[2,3†] Cai Rong Zhao[1†]
[1]Dept. of Electronic and Information Engineering, Tongji Univeristy, Shanghai
[2]YouTu Lab, Tencent, Shanghai, [3]Shanghai Jiao Tong University
{2030809, zhaocairong}@tongji.edu.cn
{boshenzhang, yukiyxli, leoneliu, swordli, caseywang, jasoncjwang}@tencent.com

1. Algorithm Details

In this appendix, we provide material that could not be included in the main manuscript due to space constraints. First, Section 1 presents the detailed algorithm of the proposed method. Section 2 provides insights about how IPC works from the perspective of label aggregation. Finally, Section 4 presents additional experimental results of DMLP.

Algorithm 1 delineates the proposed DMLP in detail, where the equation labels are consistent with the main text of the paper. Firstly, the feature extractor $G(\cdot;\theta_G)$ is pretrained via a self-supervised learning stage. IPC (Line 5-9) and EAC (Line 12-16) are two mutually reinforcing label correcting processes of DMLP. After m iterations of optimization, the purified labels ${Y_t}^*$ are obtained. Finally, the well-trained classifier $C(\cdot;w_c)$ can directly infer on test data or re-train a LNL framework with ${Y_t}^*$ to further boost performance.

2. Detailed Interpretation of IPC

In this section we further discuss how our IPC utilizes the decoupled feature representation to model the risk resulted from label noises from the perspective of label aggregation. In the main part of our paper, the IPC process seeks to predict the labels of validation samples via its optimally estimated linear estimator as

$$y'_{v,i}(Y_t) = Y_t^T F_t (F_t^T F_t)^{-1} \mathbf{f}_{v,i}. \tag{1}$$

Since the representation features are normalized after self-supervised training, thus $(F_t^T F_t)^{-1}$ can be interpreted as the inversed covariance matrix of distribution from training data, and the estimated covariance matrix maps the feature of validation data into the observation space of training data

$$\mathbf{f}'_{v,i} = (F_t^T F_t)^{-1} \mathbf{f}_{v,i}. \tag{2}$$

With the definition of Eq. (2), the term of matrix multiplication in Eq. (1) can be interpreted as a similarity matrix

$$\alpha = F_t \left[(F_t^T F_t)^{-1} \mathbf{f}_{v,i} \right] = F_t \mathbf{f}'_{v,i}, \tag{3}$$

each entry of vector $\alpha \in \mathbb{R}^b$ represents the similarity between $f'_{v,i}$ and each training sample of F_t. Finally, the output prediction on validation data can be regarded as the attentive aggregation over all labels in a training batch

$$y'_{v,i}(Y_t) = Y_t^T \alpha. \tag{4}$$

Ideally, since the self-supervised training process is trained without noisy labels in a contrastive manner, the feature distribution intrinsically forms a cluster-like manifold, i.e. samples with the same semantic label are closer in feature space

$$\mathbf{f}_{t,j}^T \mathbf{f}'_{v,i} > \mathbf{f}_{t,k}^T \mathbf{f}'_{v,i} \quad \textbf{when} \quad y_{v,i} = y_{t,j}, y_{v,i} \neq y_{t,k}. \tag{5}$$

Consequently, when a training sample in F_t is more similar to the validation sample, its semantic label will contribute more to prediction $y'_{v,i}(Y_t)$ and vice versa. Therefore, when penalizing on the discrepancy between $y'_{v,i}(Y_t)$ and $y_{v,i}$, we put more penalty on training samples with similar feature distribution but different labels from $(x_{v,i}, y_{v,i})$, which is more likely to be noisy samples.

3. Derivation of Optimal Regression

In Eq. (4) of the manuscript, we take the Frobenius norm as objective for regression, thus we have

$$\begin{aligned}\mathcal{L} &= \|\sigma(\alpha Y_t) - F_t w\|^2 + \lambda \|w\|^2 \\ &= \mathbf{Tr}([\sigma(\alpha Y_t) - F_t w]^\top [\sigma(\alpha Y_t) - F_t w]) + \lambda \|w\| \\ &= \mathbf{Tr}(w^\top F_t^\top F_t w) - 2\mathbf{Tr}(w^\top F_t^\top \sigma(\alpha Y_t)) + \lambda \|w\|^2 + C\end{aligned}$$

where C is a constant unrelated to w, $\mathbf{Tr}(\cdot)$ denotes the Trace of matrix. With the property $\frac{\partial \mathbf{Tr}(A^\top B)}{\partial A} = B$, we can obtain

Algorithm 1 The workflow of DMLP.

Input: Noisy training set D_t, clean validation set D_v, feature extractor $G(\cdot;\theta_G)$, classifier $C(\cdot;w_c)$, batch size b, max iterations m, period for regular label substitution T.

Procedure:

Self-supervised training for $G(\cdot;\theta_G)$
Generate features **f** by Eq. (1)
for $i = 1$ to m **do**
  /***IPC** starts*/
  $\{F_t, Y_t\} \leftarrow SampleMiniBatch(\mathbf{f}, \mathrm{D}_t, b)$
  Calculate closed-form solution $w^*(Y_t)$ by Eq. (6)
  Predict validation set labels y'_v by Eq. (7)
  Calculate label purification loss $\mathcal{L}_{\mathrm{val}}(Y_t)$ by Eq. (8).
  Update training labels Y_t in backward process.

  /***EAC** starts*/
  Calculate loss for the classifier $C(\cdot;w_c)$ by Eq. (10)
  Update classifier parameter w_c in backward process.
  if $i = nT$ **then**
    Update training labels Y_t by Eq. (11)
  end if
end for

Output: The purified labels Y_t^*.

the necessary condition of optimal w^* via $\frac{\partial \mathcal{L}}{\partial w} = 0$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w} &= 2F_t^\top F_t w - 2F_t^\top \sigma(\alpha Y_t) + 2\lambda w \\ &= 2(F_t^\top F_t + \lambda I)w - 2F_t^\top \sigma(\alpha Y_t) = \mathbf{0}\end{aligned}$$

Hence we have the closed-form optimal regression parameter $w^*(Y_t)$

$$\begin{aligned}w^*(Y_t) &= (F_t^\top F_t + \lambda I)^{-1} F_t^\top \sigma(\alpha Y_t) \\ w^*(Y_t)^\top \mathbf{f}_{v,i} &= \sigma(\alpha Y_t)^\top F_t \left({F_t}^\top F_t + \lambda I\right)^{-1} \mathbf{f}_{v,i}\end{aligned}$$

4. Experimental Details

4.1. Experimental Settings

For the CIFAR10/100 dataset, most parameters of the SimCLR [3] algorithm are set as suggested in the original implementation. The classifier $C(\cdot;w_c)$ is trained with a learning rate of 0.005 and batch size of 500, while the batch size for the label purifier in IPC is set to 7,000. The classifier and the purifier are both trained with the Adam optimizer. For DMLP-Mix, all the hyper-parameters of DivideMix are set as the authors suggested. As for the real-world Clothing1M dataset, we follow the original protocol to split the training and testing sets. The given validation set is utilized for meta-learning. The batch sizes for the classifier and the label purifier in IPC are set to 500 and 10,000 respectively. The learning rate for the former is set to 0.03 while for the latter is 0.02. Similar to the setting on CIFAR-10/100, the Adam optimizer is also adopted.

4.2. Details of Compared Methods

As discussed in manuscript, DMLP is compared with most recent relevant works. Specifically, the competing works can be coarsely categorized into two group, noisy sample detection and label correction. The former usually identifying and reducing the importance of suspicious false-labeled samples during training, either by directly selecting the clean samples out of training set (Co-teaching [9], Co-teaching+ [21], Iterative-CV [2], Sel-CL+ [14], ELR+ [15], C2D [24], DivideMix [11], REED [22], MOIT+ [16]) or adjusting the soft weight of each training sample (RRL [13], M-correction [1], GCE [8], CDR [18]). The latter aims to correct the corrupted labels and augment the training data. Typical paradigms are correction-by-prediction, i.e., utilizing the prediction of deep model to correct labels, including Joint-Optim [17], PENCIL [20], Self-Learning [10]. Others resort to a small set of clean validation set with meta-learning training strategies, i.e., Meta-Learning [12], MLC [25], MSLC [6], Meta-Cleaner [7], Meta-Weight [5], FaMUS [19], MSLG [4], Zhang, et al. [23].

4.3. More Applications of DMLP

As mentioned in the manuscript, DMLP can be applied to work collaboratively with the existing LNL framework to boost performance. To further verify the effectiveness of DMLP, we also plot the accuracy curve of the proposed DMLP and its baseline methods in Fig. 4-9. It is observed that all the applications of DMLP perform consistently better over their corresponding baselines throughout the training process, especially under high-level noise cases.

4.4. Details of Experimental Result

• **Detailed Comparison with Coupled Methods.** In the manuscript we compare DMLP against coupled meta label correction methods MLC [25] and MSLC [6] with same self-supervised pretrained weights. Here we provide more detailed comparison results with the original implemented MLC and MSLC in terms of corrected label accuracy and backbone quality (revealed by linear evaluation accuracy). As shown in Fig. 1, by simplifying the complex coupled meta-learning process into individual representation learning and non-nested meta label purification, DMLP can achieve superior performance to these methods in the sense of purified label accuracy and meanwhile obtain representations of better quality, which further verifies our empirical findings of Fig. 1 in the manuscript.

• **Detailed Comparison with Decoupled Baselines.** The proposed non-nested meta label purifier plays a crucial role

Table 1. Investigation on the influence of meta-purification on CIFAR-10.

Method		Noisy ratio 20%	50%	80%	90%
SimCLR-DivideMix	Best	92.2	91.2	92.1	85.7
	Last	82.8	81.3	77.0	10.9
DMLP-DivideMix	Best	**96.3**	**95.8**	**94.5**	**94.3**
	Last	**96.2**	**95.6**	**94.3**	**94.0**

in DMLP. To fairly verify its superiority to existing label correction methods, we train MLC and MSLC in a decoupled way where their backbone is fixed with SimCLR self-supervised weights as in DMLP. As shown in Table 2, decoupled training scheme can largely boosts their performance, which is also in line with our empirical findings in Fig. 1 of the manuscript. Besides, it can be obviously observed that DMLP-Naive shows great advantage over the decoupled MLC and MSLC across all the settings especially under high noise, demonstrating the effectiveness of the non-nested meta label purifier.

• **Effect of the Meta-purification on DMLP-DivideMix.** We further investigate the effect of the non-nested meta label purifier on the afterward LNL framework in DMLP-DivideMix setting. To do this, we initialize a model with SimCLR and apply DivideMix to train the model as our baseline. According to the results in Table 1, the accuracy suffers from a significant performance drop when removing the non-nested meta label purifier from our pipeline (i.e, SimCLR-DivideMix), especially for severe label noise cases, this can be attributed to the DNN inevitably gradually memorizes the noisy labels when updating the backbone and classifier simultaneously. In contrast, in DMLP-DivideMix, labels purified by the meta-learner yield higher accuracy, guiding the afterward LNL framework to learn more robust and discriminative decision boundaries.

• **Detailed Comparison on the label accuracy.** In addition to the comparison of 50% and 90% noise between MLC, MSLC and DMLP on the CIFAR-10 in the manuscript, we also visualize the label accuracy curves for 20% and 90% noise. As shown in Fig. 2, DMLP consistently shows great superiority over MLC and MSLC throughout the training process. Moreover, Fig. 3 shows corrected label accuracy curve of DMLP under symmetric and asymmetric noise settings on CIFAR-10, which demonstrates that high quality labels can be generated by DMLP across all noisy settings.

• **Experimental results with statistical significance.** Besides the results shown in the main text, we provide detailed results with Standard Deviations (STD) to show the statistical significance of proposed method on CIFAR-10/100. Which are shown in Table 4.

4.5. EAC as Classifier

Besides DMLP-Naive, we can also take the well-trained linear classifier $C(\cdot; w_c)$ in the non-nested meta label purifier for the test set prediction, this is termed as DMLP-EAC. As shown in Table. 3, though DMLP-EAC is only an individual linear classifier, it can also perform well especially under high noisy settings on CIFAR-10/100. Moreover, DMLP-EAC can already outperform most of state-of-the-art LNL methods by a considerable margin and achieve comparable performance to DMLP-Naive on the Clothing1M dataset, which further demonstrates that DMLP is more suitable to tackle with real-world noise. Finally, Table. 3 shows the comparison between DMLP-IPC and REED (no stage-3) [22], which simply trains a linear classifier on well-established representations without extra operations. Though REED (no stage-3) achieves overall good results, DMLP-IPC can still obtain consistent performance gains over this baseline under all noise settings, especially on the high noise level.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *Proc. International Conference on Machine Learning (ICML)*, 2019.

[2] Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *Proc. International Conference on Machine Learning (ICML)*, 2019.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020.

[4] Algan et al. Meta soft label generation for noisy labels. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7142–7148, 2021.

[5] Shu et al. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019.

[6] Wu et al. Learning to purify noisy labels via meta soft label corrector. In *AAAI*, 2021.

[7] Zhang et al. Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition. In *CVPR*, 2019.

[8] Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2703–2708, 2021.

[9] 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 *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[10] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.

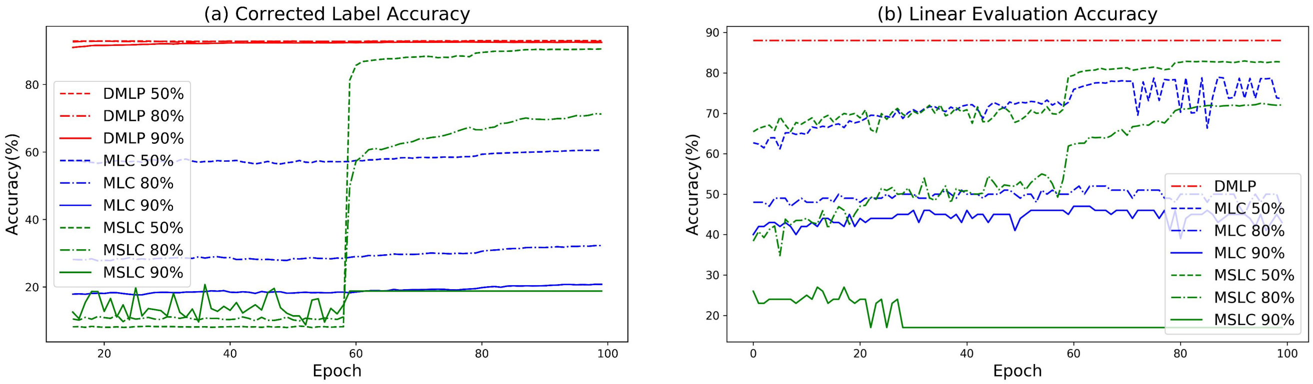

Figure 1. Comparison with two state-of-the-art coupled optimization based meta label correction methods MLC [25] and MSLC [6] on corrected label accuracy and linear evaluation accuracy.

Table 2. Comparison with MLC and MSLC on CIFAR-10/100. "†" denotes training with fixed self-supervised pretrained ResNet-18. "*" denotes training with self-supervised pretrained ResNet-18.

Dataset Method	Noise ratio	CIFAR-10 20%	 50%	 80%	 90%	CIFAR-100 20%	 50%	 80%	 90%
MLC* [25]	Best	91.8	86.2	77.6	72.9	62.2	53.8	46.5	39.6
	Last	91.6	85.9	77.5	72.6	61.6	53.0	46.2	39.2
MSLC* [6]	Best	92.0	87.7	78.0	67.8	70.8	64.1	36.4	19.8
	Last	92.0	87.5	77.9	67.3	70.2	63.8	34.3	18.7
MLC† [25]	Best	92.0	90.2	89.0	88.9	65.9	59.4	54.4	54.2
	Last	91.6	89.4	88.5	88.1	65.2	59.2	54.1	54.0
MSLC† [6]	Best	92.1	90.4	87.3	84.7	71.7	64.7	53.3	46.8
	Last	92.0	90.0	87.2	84.2	71.6	64.4	53.0	46.4
DMLP-Naive	Best	**94.7**	**94.2**	**93.5**	**92.8**	**72.7**	**68.0**	**63.5**	**61.3**
	Last	**94.2**	**94.0**	**93.2**	**92.0**	**72.3**	**67.4**	**63.2**	**60.9**

Table 3. Comparison between DMLP-EAC, DMLP-Naive and DMLP-DivideMix on CIFAR-10/100 and Clothing1M datasets. "‡" denotes reproduced results.

Dataset Method	Noise ratio	CIFAR-10 20%	 50%	 80%	 90%	CIFAR-100 20%	 50%	 80%	 90%	Clothing1M
REED(no stage-3)‡	Best	89.1	88.6	87.3	85.1	62.6	61.5	58.4	53.2	46.05
	Last	88.9	88.5	87.0	84.9	62.5	61.4	58.2	52.9	45.81
DMLP-EAC	Best	91.3	91.0	90.3	89.3	65.7	64.1	60.2	57.6	77.31
	Last	91.2	90.7	90.2	89.2	65.5	63.8	60.1	57.5	77.31
DMLP-Naive	Best	94.7	94.2	93.5	92.8	72.7	68.0	63.5	61.3	77.77
	Last	94.2	94.0	93.2	92.0	72.3	67.4	63.2	60.9	77.70
DMLP-DivideMix	Best	**96.3**	**95.8**	**94.5**	**94.3**	**79.9**	**76.8**	**68.6**	**65.8**	**78.23**
	Last	**96.2**	**95.6**	**94.3**	**94.0**	**79.4**	**76.1**	**68.5**	**65.4**	**78.23**

[11] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[12] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Learning to learn from noisy labeled data. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

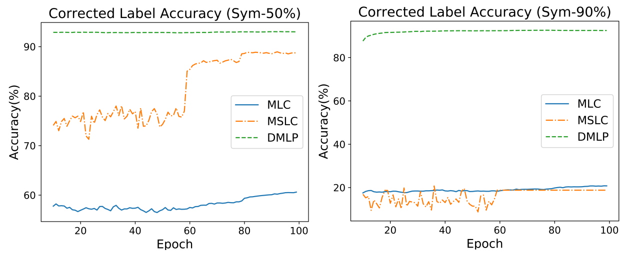

Figure 2. Comparison of corrected label accuracy curve under symmetric-50% (left), symmetric-90% (middle) noise settings on CIFAR-10.

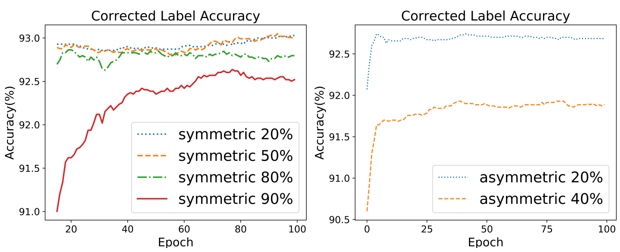

Figure 3. Corrected label accuracy curve of DMLP under symmetric (left), asymmetric (middle) noise settings on CIFAR-10.

Table 4. Comparison on CIFAR-10/100 with symmetric noise.

Dataset	CIFAR-10			
Method	20%	50%	80%	90%
DMLP-Naive	94.28±0.10	94.02±0.21	93.31±0.19	92.16±0.20
DMLP-DivideMix	**96.20±0.11**	**95.63±0.13**	**94.22±0.14**	**93.97±0.22**
Dataset	CIFAR-100			
Method	20%	50%	80%	90%
DMLP-Naive	72.39±0.08	67.60±0.26	63.17±0.14	61.09±0.20
DMLP-DivideMix	**79.31±0.21**	**76.11±0.10**	**68.42±0.12**	**65.55±0.23**

[13] Junnan Li, Caiming Xiong, and Steven CH Hoi. Learning from noisy data with robust representation learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 9485–9494, 2021.

[14] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. *arXiv preprint arXiv:2203.04181*, 2022.

[15] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[16] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6606–6615, 2021.

[17] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] 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 *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[19] Youjiang Xu, Linchao Zhu, Lu Jiang, and Yi Yang. Faster meta update strategy for noise-robust deep learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 144–153, 2021.

[20] Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *Proc. International Conference on Machine Learning (ICML)*, 2019.

[22] Hui Zhang and Quanming Yao. Decoupling representation and classifier for noisy label learning. *arXiv preprint arXiv:2011.08145*, 2020.

[23] Z. Zhang, H. Zhang, S. Arik, H. Lee, and T. Pfister. Distilling effective supervision from severe label noise. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1657–1667, 2022.

[25] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais. Meta label correction for noisy label learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2021.

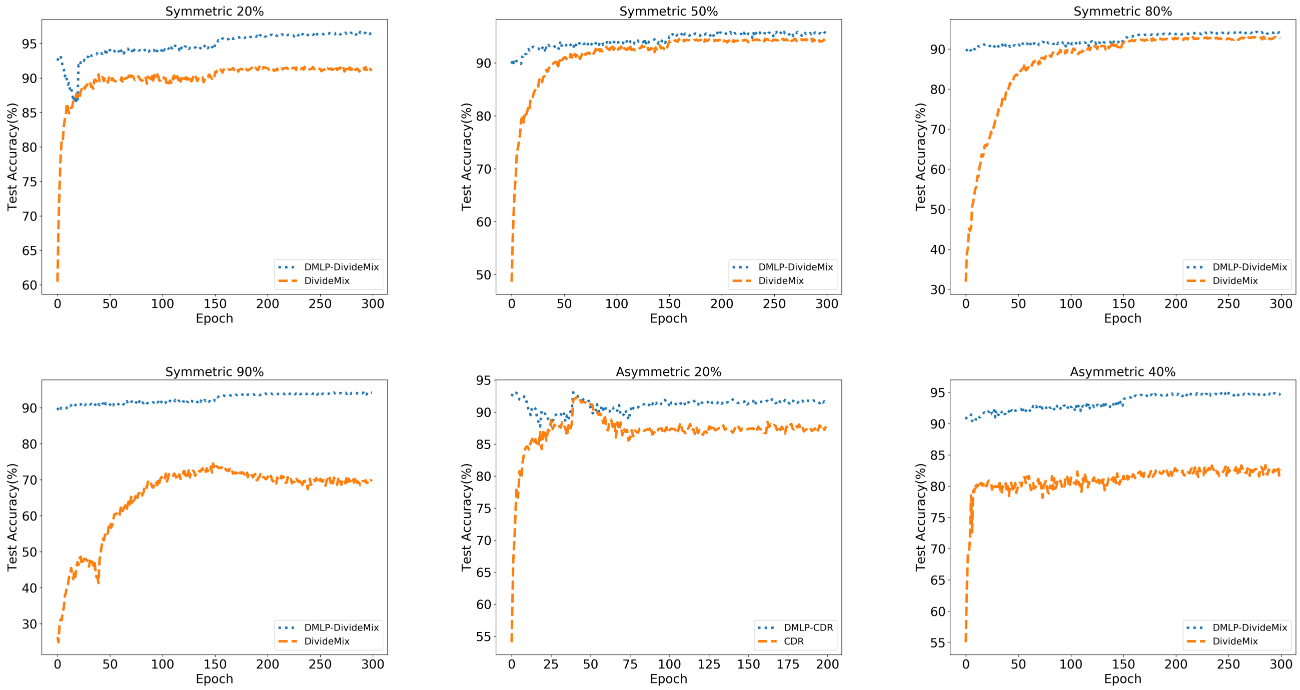

Figure 4. Accuracy curve of DMLP-DivideMix and DivideMix on CIFAR-10 under different noise settings.

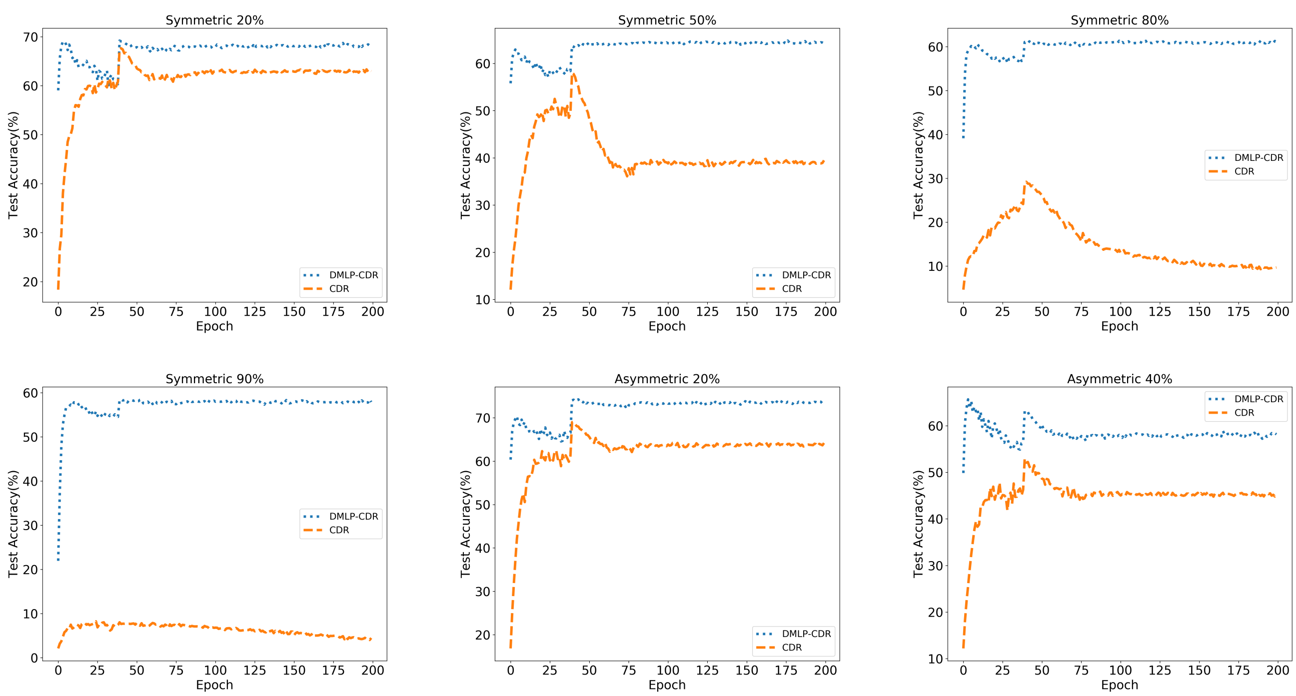

Figure 5. Accuracy curve of DMLP-DivideMix and DivideMix on CIFAR-100 under different noise settings.

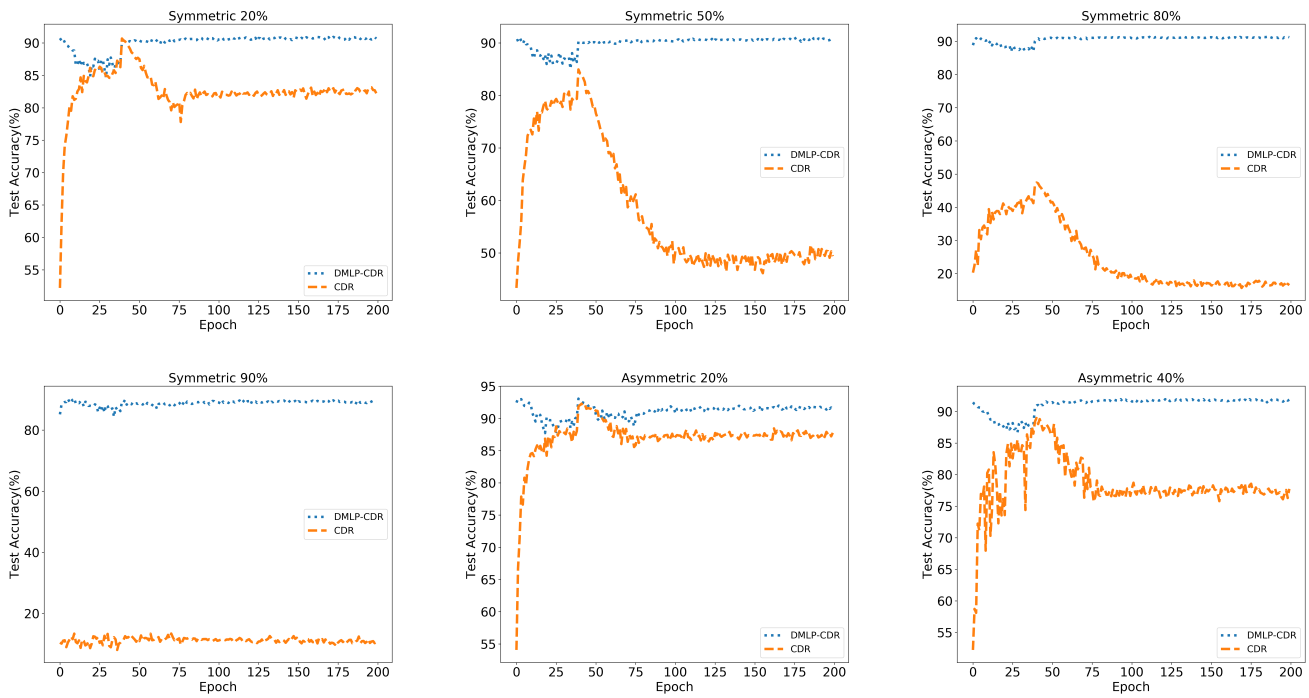

Figure 6. Accuracy curve of DMLP-CDR and CDR on CIFAR-10 under different noise settings.

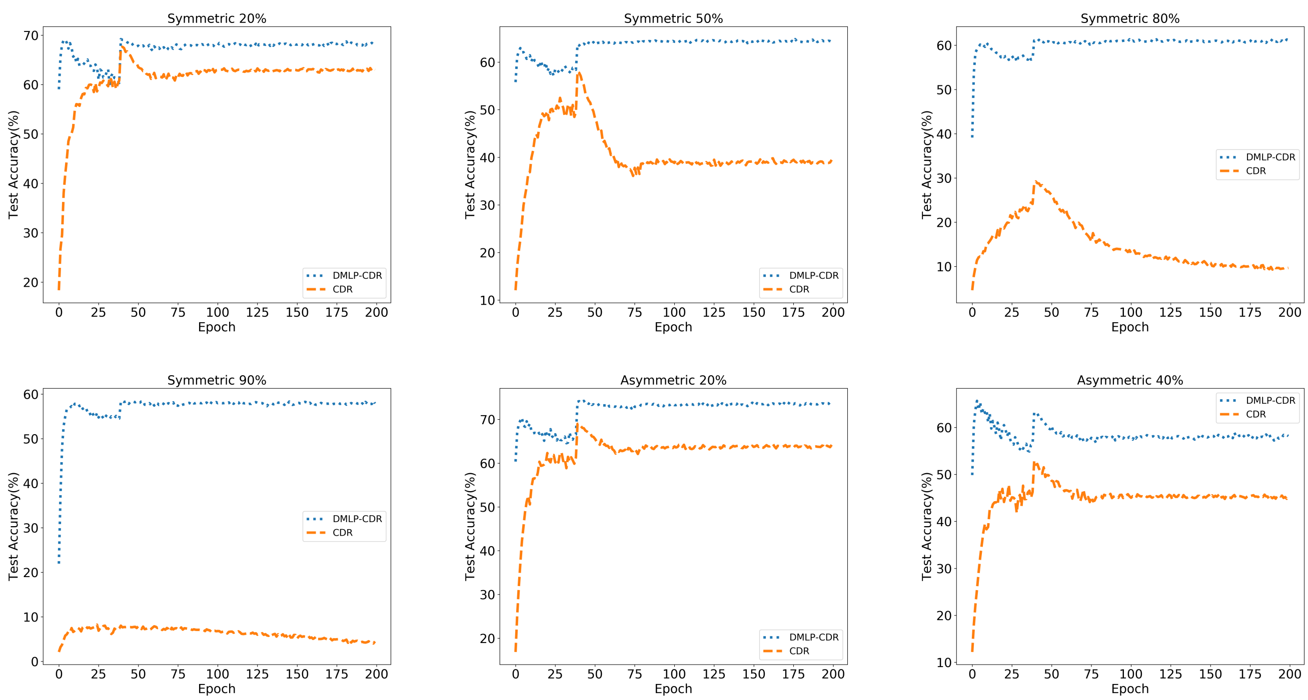

Figure 7. Accuracy curve of of DMLP-CDR and CDR on CIFAR-100 under different noise settings.

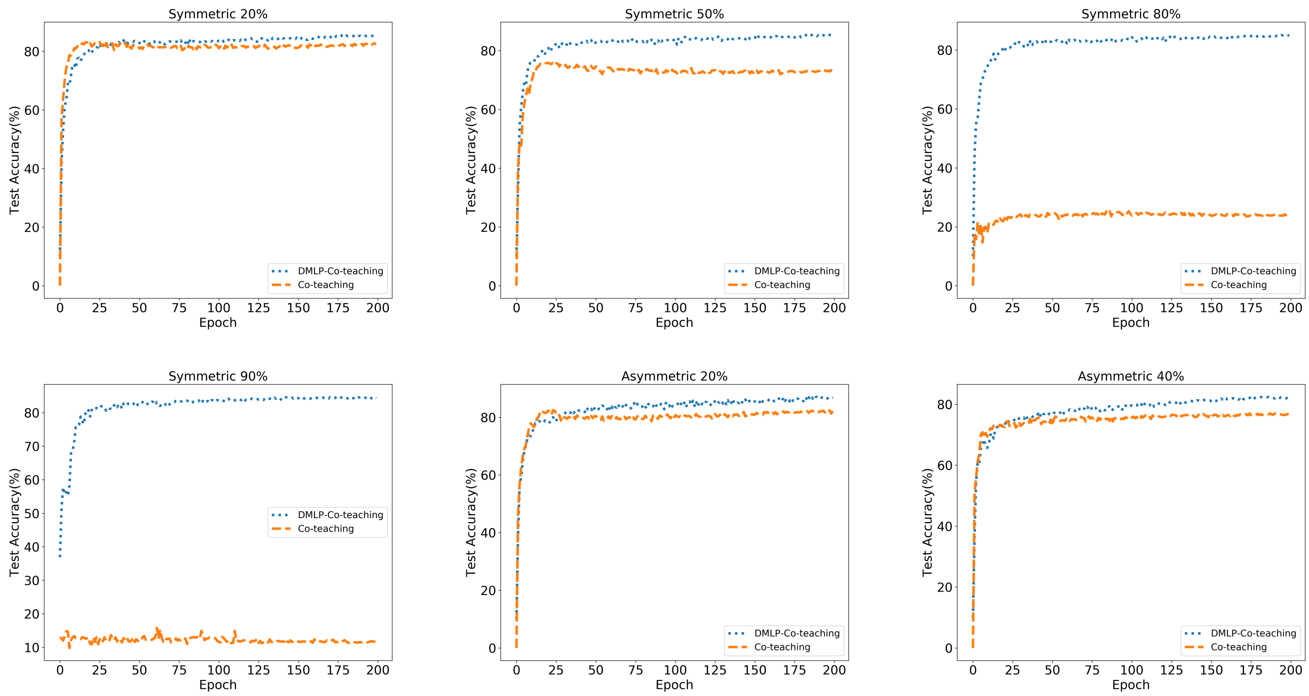

Figure 8. Accuracy curve of DMLP-Co-teaching and Co-teaching on CIFAR-10 under different noise settings.

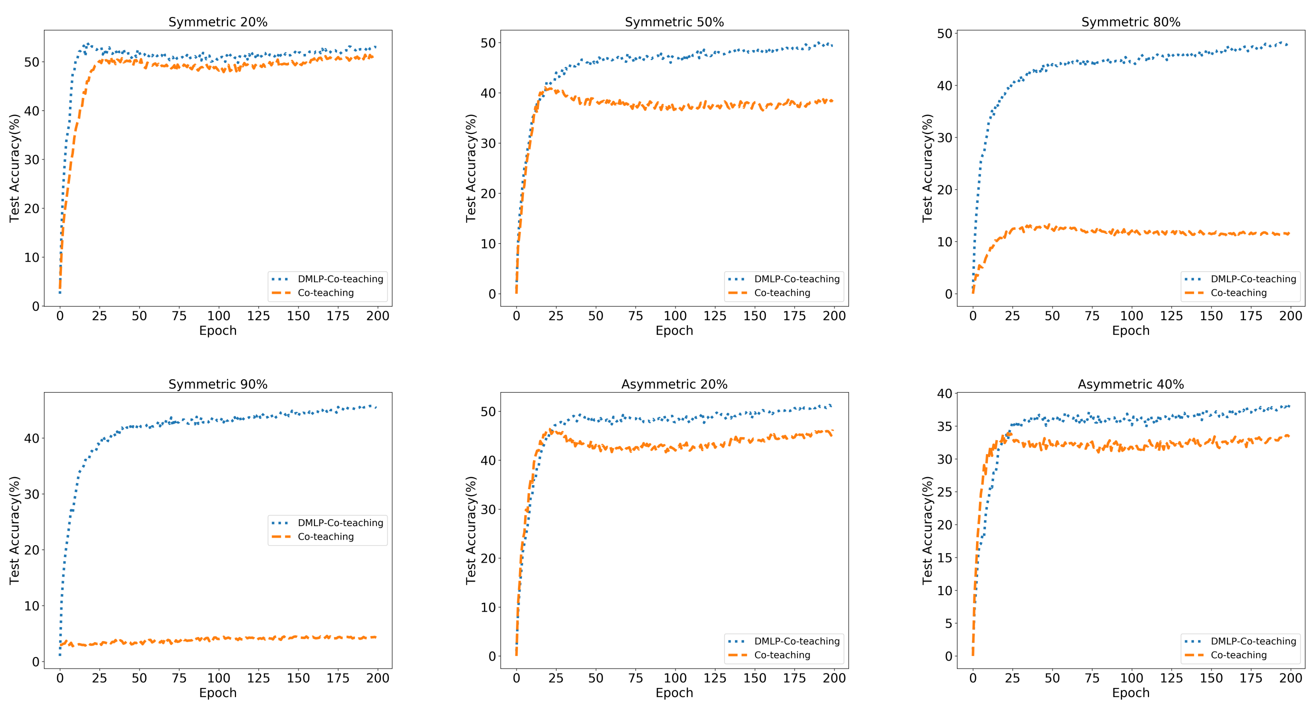

Figure 9. Accuracy curve of of DMLP-Co-teaching and Co-teaching on CIFAR-100 under different noise settings.