MART: Mechanism-disentanglement Anchor-Routed Training for Learning with Open-World Noisy Data

Supplementary Material

A. Experimental Setup

A.1. Task Settings

We consider three types of tasks in our experiments (Refer Table A.1). Specifically, we verify the effectiveness of our proposed method in the situation of Learning with Open-world Noisy Data (LOND) and Learning with Closed-world Noisy Data (LCND), as well as Learning with Real-world Noisy Data (LRND).

Setup	IND Noise in D_{train}	OOD in D_{train}	OOD in D_{test}
LOND	✓	✓	✓
LCND	✓	✗	✗
LRND	✓	✓	✗

Table A.1. Three types of tasks in our experiments.

A.2. Datasets

Synthetic Noise Datasets. We evaluate our method on the synthetic noise datasets CIFAR100N and CIFAR80N, which are both constructed based on CIFAR100 [22]. CIFAR100 contains 60,000 RGB images, of which 50,000 are used for training and 10,000 for testing. Following prior work, we generate a closed-set noise dataset CIFAR100N, and an open-set noise dataset CIFAR80N. Specifically, when constructing CIFAR80N, the last 20 classes of CIFAR100 are treated as out-of-distribution samples. We adopt two common noise types: symmetric noise and asymmetric noise, with noise ratio $n \in (0, 1)$. Symmetric noise randomly replaces the original label with any other class; asymmetric noise is closer to real-world situations and usually replaces the label with a semantically similar class.

Real-world Noise Datasets. We also conduct experiments on real-world noise datasets, including Web-Aircraft, Web-Bird and Web-Car [53]. The training images in these datasets are obtained from web image search and therefore inevitably contain label noise. Compared with artificially constructed noise datasets, real data in practical applications is more complex and closer to reality, containing multiple types of label noise such as symmetric noise, asymmetric noise, and open-set noise.

A.3. Experimental Settings

We adopt a 7-layer CNN as the backbone on sythetic noise datasets. The optimizer is SGD with momentum 0.9, and training runs for 300 epochs: the first 50 epochs perform a warm-up phase with a fixed learning rate of 0.001; the subsequent formal training epochs decrease the learning rate from 0.02 using cosine annealing to enhance robustness. The batch size is 128. On real-world data, we employ an ImageNet-pretrained ResNet50 [12] as the backbone, again using SGD with momentum 0.9 for 120 epochs with 10 epochs warm-up phase. The initial learning rate is 0.005, and cosine annealing is used for learning rate decay.

Baselines. On CIFAR100N and CIFAR80N, we compare the proposed method with various current SOTA approaches, including: Decoupling [34], Co-teaching [11], Co-teaching+ [70], JoCoR [57], DivideMix [25], Jo-SRC [67], Co-LDL [54], UNICON [20], SOP[30], AGCE [77], DISC [27], ANL [68], NPN [45], ACT [47], SED [46], DRS [37] and CA2C [48]. We also provide a "Standard" baseline, i.e., conventional training on the full noisy dataset. On real-world datasets such as Web-Aircraft, Web-Bird, and Web-Car, we further include methods such as PENCIL [69], AFM [40], PLC [74], WarPI [49], CoDis [63], and VRI [50] for comparison.

B. Hyperparameter Sensitivity Analysis

We conduct univariate sensitivity experiments on CIFAR80N (Sym-20%) to evaluate the stability of key designs. Figure B.1 (left) analyzes the gate temperature (τ) of mechanism disentanglement (controlling the "hardness" of the invariant and variant subspaces): when $\tau \in \{0.1, 0.5, 1.0, 2.0\}$, the accuracies are 70.9%, 71.5%, 71.2%, and 71.1% respectively. Too small (0.1) makes the gate overly sharp and tends to "lock" discriminative dimensions early; too large (2.0) causes mixing of the two subspaces and insufficient disentanglement. The range 0.5–1.0 is the most stable (fluctuation $< 0.4\%$), and we set 0.5 in the main experiments. Figure B.1 (right) analyzes the class-contrast gate threshold (δ) of neighborhood routing. When $\delta = 0.01/0.05/0.1/0.5$, the accuracies are about 71.23%/71.48%/71.02%/70.90%. A too-small threshold lets through neighbors that are "close but wrong-class"; a too-large threshold wrongly eliminates hard samples and weakens the aggregation signal. $\delta = 0.05$ is optimal, and the curve is smooth within 0.01–0.10 (less than 0.5 percentage points). Overall, our method shows good robustness to core hyperparameters. This is consistent with the method's motivation—the mechanism disentanglement requires a "neither too hard nor too soft" disentangling

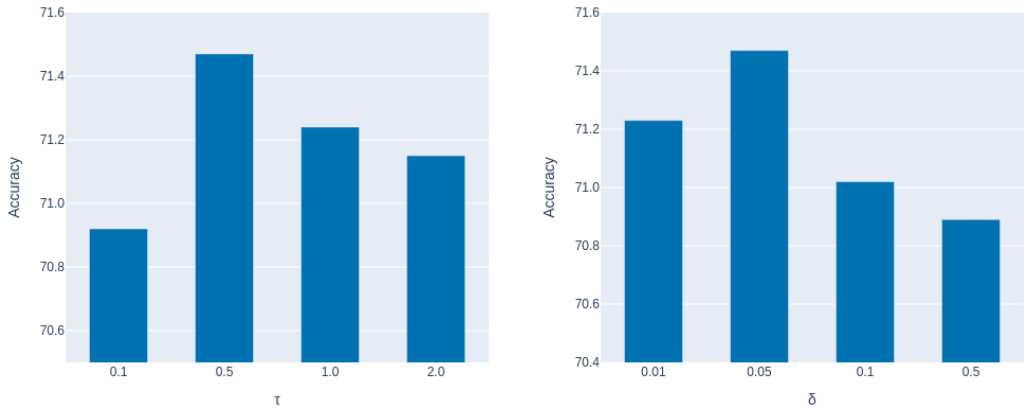


Figure B.1. Hyperparameter Sensitivity Analysis. (left) Analyzes the gate temperature (τ) of mechanism disentanglement. (right) Analyzes the class-contrast gate threshold (δ) of neighborhood routing.

strength to ensure a causally invariant discriminative subspace, while neighborhood routing relies on a "neither too loose nor too strict" class-contrast threshold to purify neighborhood evidence without losing hard positives.

C. Additional Analysis

C.1. Open-world Mixed Distribution

To systematically evaluate robustness under varying sources and intensities of open-set contamination, we fix 50% symmetric closed-set noise on CIFAR100N and inject unseen samples from TinyImageNet and Places-365 into the training set as OOD, with injection sizes of 10k and 20k respectively. The OOD samples are mislabeled as known classes during training, with no additional annotations or priors introduced; all methods use the same backbone, and we report test accuracy on known classes.

OOD Dataset	# OOD	SED	DRS	CA2C	Ours
TinyImageNet	10k	61.63	62.33	62.82	**63.45**
	20k	60.94	63.25	63.69	**65.13**
Places-365	10k	61.78	63.92	64.31	**65.87**
	20k	61.20	62.58	62.87	**63.85**

Table C.2. Performance Comparison under Open-world Mixed Distribution Noise.

As shown in Table C.2, our method obtain the best performance in all four settings: TinyImageNet-10k/20k reach 63.45% and 65.13%, and Places-365-10k/20k reach 65.87% and 63.85%. Among them, TinyImageNet-20k increases over strong baselines by 1.4–1.9 percentage points, while on Places-365-10k it leads by 1.56–4.09 percentage points. This indicates that under dense contamination and mixed distributions, our training paradigm can still stably refine the supervisory signal and maintain high recognition accuracy. These gains mainly stem from two factors: mechanism disentanglement performs discrimination, selection and rejection only in the invariant subspace while constraining contrast to the variant side; and neighborhood routing implements two-level gated routing via class retrieval anchors. They together stabilize pseudo labels and sample weights, thereby exhibiting consistent advantages across different OOD types and scales.

C.2. Performance of OOD Detection

We further evaluate the rejection ability of our method. Under the three noise settings of CIFAR80N (Sym-20%, Sym-80% and Asym-40%), we follow the same backbone and training procedure as in the main experiments and employ the FPR95 metric to measure the model's OOD detection performance. FPR95 is the False Positive Rate (FPR) when the recall achieve 95% (the lower the better).

Noise / Method	SED	DRS	CA2C	Ours
Sym-20%	87.65	86.79	86.34	**85.41**
Sym-80%	92.56	92.13	92.73	**91.59**
Asym-40%	89.02	88.09	88.35	**87.02**

Table C.3. Performance of OOD Detection.

As shown in Table C.3, our method reaches 85.41%, 91.59% and 87.02% under the three noise types and is all superior to the compared methods. Taking Sym-80% as an example, we reduce the false acceptance rate by 0.54–1.14 percentage points relative to strong baselines; we also maintain stable advantages under Sym-20% and Asym-40%. They indicate that under high-recall requirements we can

reject unknown classes at lower cost. It is worth noting that the lead under the three noise forms is consistent with the improvement trend of classification accuracy discussed earlier, showing that our method achieves consistent benefits on both "recognition–rejection" ends and is insensitive to changes in noise intensity. The performance gains mainly come from mechanism disentanglement restricting class-related operations to the invariant subspace and isolating contrastive consistency to the variant side; neighborhood routing implements dual gating via class retrieval anchors and together with the negative-channel margin what explicitly removes OOD, and they jointly improve OOD detection capability.

C.3. Sample Selection Performance

We further evaluate sample selection precision: the proportion of truly clean samples within the set judged as "clean samples". The measurement follows the main experiment: we adopt a unified backbone and training schedule, and compute precision for sample selection according to the ground-truth labels. The evaluation covers CIFAR100N and CIFAR80N, using the challenging Sym-80% for noise type and ratio. Table C.4 reports the results: our method achieves the highest precision on both open-set and closed-set noise, reaching 56.54% (CIFAR100N) and 61.32% (CIFAR80N).

Dataset / Method	SED	DRS	CA2C	Ours
CIFAR100N	52.35	54.41	50.82	**56.54**
CIFAR80N	57.81	59.93	56.05	**61.32**

Table C.4. Sample Selection Performance (Precision of Sample Selection).

Compared with strong baselines, the improvements on CIFAR100N over SED,DRS and CA2C are 4.19%, 2.13% and 5.72% respectively; on CIFAR80N the gains are 3.51%, 1.39% and 5.27%. Higher selection precision means that the samples entering the supervised branch and pseudo-label updating are cleaner, thereby reducing the cumulative amplification of erroneous supervision, yielding smoother convergence of the training curve and synchronized gains in subsequent classification and rejection; this is more evident under the more difficult open-set noise. Unlike previous experiments, the improvements here mainly lie in the conservativeness and balance of the selection mechanism itself: mechanism disentanglement constrains discriminative evidence to the invariant subspace to prevent noise infiltration; neighborhood routing employs class anchors to suppress "near but wrong-class" neighbors, while two-stage aggregation reduces early confirmation bias and label jitter. They result in fewer misselected samples and better cross-class balance, thus precision rises more stably over training.

C.4. Training Time and Computing Cost

We add an ablation that reports both the incremental performance gains and the corresponding computational overhead when modules are introduced progressively (Table C.5 and Table C.6) highlighting not having substantial overhead with the proposed components.

Method / Noise	Training Time (s)
Baseline	46.5
w/ Mechanism Disentanglement	47.6
w/ Dual Gating	47.9
w/ Two-stage Aggregation	48.6
MART	49.8

Table C.5. Training time.

Method / Noise	Computing Cost
Baseline	6.9
w/ Mechanism Disentanglement	7.3
w/ Dual Gating	7.0
w/ Two-stage Aggregation	7.2
MART	7.6

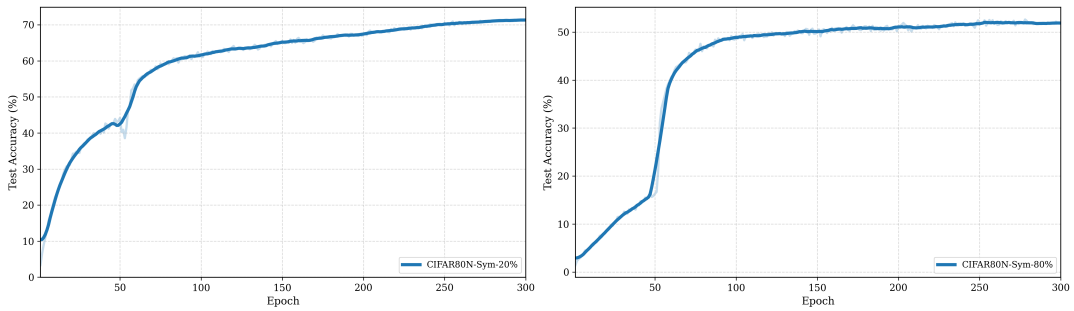
Table C.6. Computing cost (GB).

C.5. Training Dynamics

We provide additional learning curves under Sym-20% and Sym-80% noise settings (Figure C.2). Both the test accuracy and the sample selection precision improve throughout training and eventually stabilize. Notably, under the extreme Sym-80% noise, the selection precision continues to increase and converges, indicating that the sample-identification feedback loop does not amplify confirmation bias; instead, it progressively purifies the supervision signal and leads to smooth convergence.

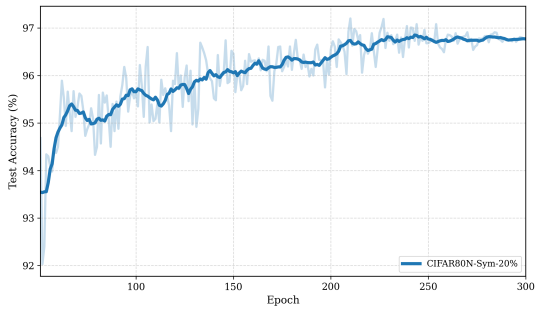
D. Attention Visualization

To further analyze whether the mechanism disentanglement is effective, we add attention visualizations under Sym-20% (Figure D.3). We observe a clearer structure and weaker entanglement in the invariant subspace representation, which is consistent with the intended division of labor: making decisions and correcting labels in the invariant subspace, while applying stability regularization in the variant subspace. We emphasize that this serves as **functional** evidence of disentanglement (i.e., the subspaces assume different roles and yield benefits), rather than a claim of strict identification of underlying causal factors.

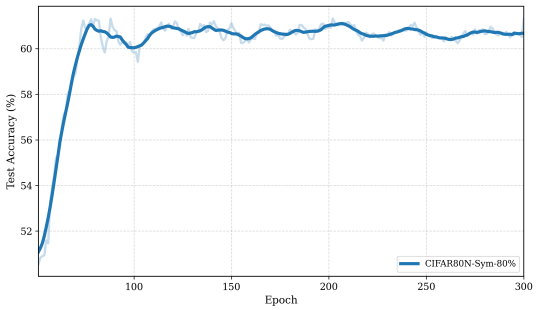


(a) Test Accuracy-Sym-20%

(b) Test Accuracy-Sym-80%



(c) Sample Selection-Sym-20%



(d) Sample Selection-Sym-80%

Figure C.2. Test accuracy and sample selection precision vs. epochs.

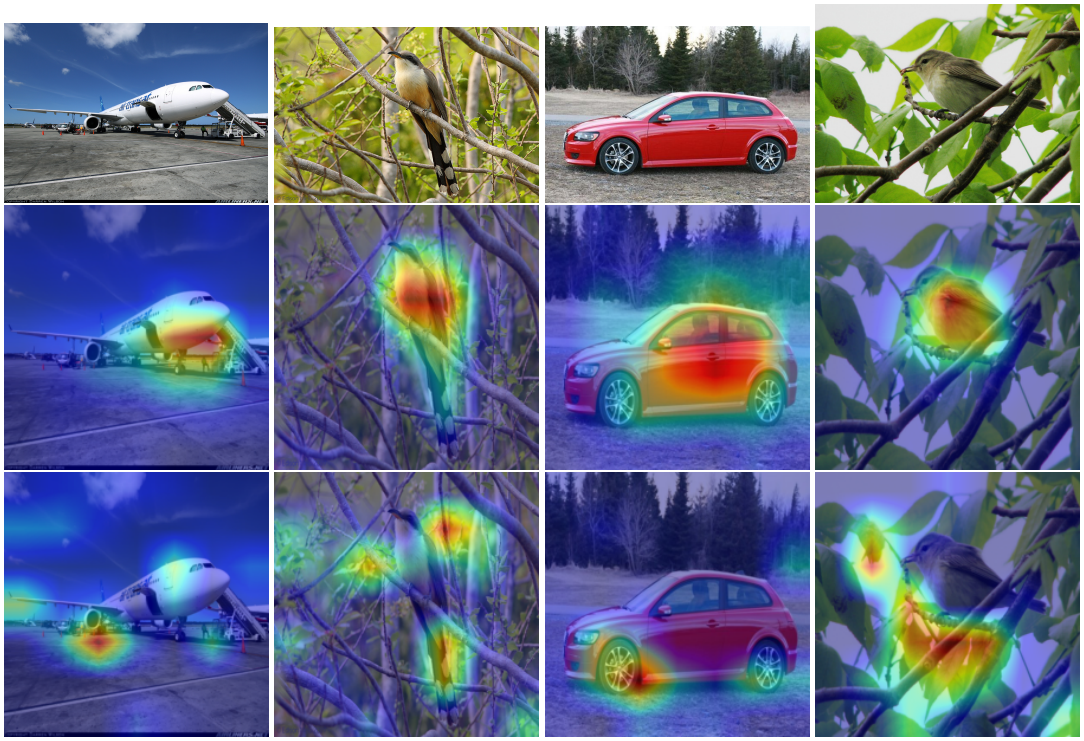


Figure D.3. Attention visualization for invariant space and variant space.