Leveraging Dissimilarity Invariance as a Robust Anchor for Learning with Noisy Labels

Wenxiao Fan, Kan Li†

School of Computer Science, Beijing Institute of Technology
{wenxiaofan, likan}@bit.edu.cn

Abstract

Deep learning models excel in visual recognition but suffer severe performance drops when training labels are corrupted by noise. Under label noise prior work cannot learn accurate similarities and thus misguide the learning process. In this paper, we uncover a complementary and novel phenomenon, Dissimilarity Invariance, whereby semantic dissimilarity between unrelated samples remains stable despite label noise. Leveraging this insight, we propose NegScale, a plug-and-play framework that shifts focus from fragile similarity to robust dissimilarity. NegScale integrates: (1) Structured Negative Orthogonality Penalty (SNOP), enforcing subspace orthogonality for unrelated samples; and (2) Dissimilarity-Calibrated Similarity Adjustment (DCSA), suppressing spurious similarity using dissimilarity anchors. We also give theoretical analysis that proves Dissimilarity Invariance and the effectiveness of NegScale. Empirical results demonstrate that NegScale consistently outperforms state-of-the-art baselines, establishing new benchmarks on CIFAR with synthetic noise and real-world datasets.

Introduction

Deep learning models have achieved remarkable success across a wide range of visual recognition tasks (Bochkovskiy, Wang, and Liao 2020; Marriott, Romdhani, and Chen 2021), but their performance degrades sharply when training labels are corrupted by noise. One underlying cause of this brittleness lies in how noisy labels distort the semantic structure of the data, especially the learned similarities between samples (Chen et al. 2023; Fan and Li 2025). In practice, we find that similarity are inherently fragile under label noise (see Fig. 1). When samples presumed to belong to the same or semantically related classes are corrupted by incorrect labels, their pairwise similarities become severely distorted. Aligning the model according to these erroneous affinities not only degrades performance but also prevents the model from learning meaningful relationships.

In contrast, we identify an underexplored phenomenon, named **Dissimilarity Invariance**, in which semantic similarity between unrelated samples (negative pairs)—referred to as dissimilarity—remains remarkably stable even as label noise increases. Specifically, representations of semantically unrelated classes maintain consistent dissimilarity levels across a wide range of noise rates. This observation suggests that, while label noise can severely distort similarity, it has far less impact on dissimilarity. As a result, models can reliably learn and leverage these robust dissimilarity patterns to improve performance under noisy supervision.

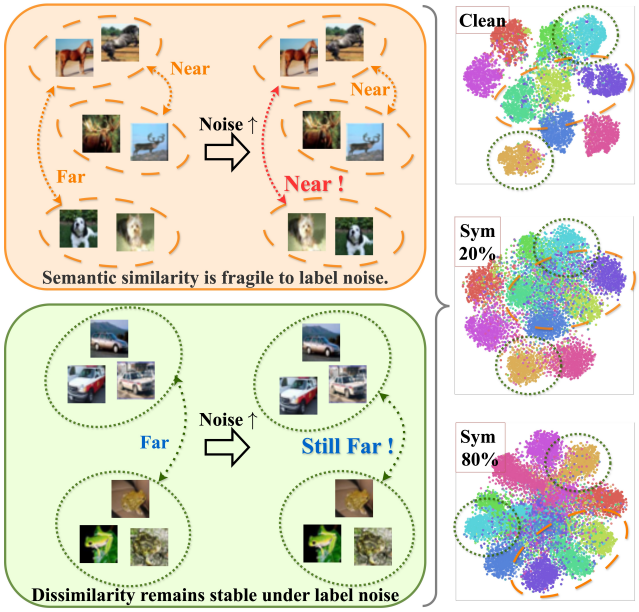
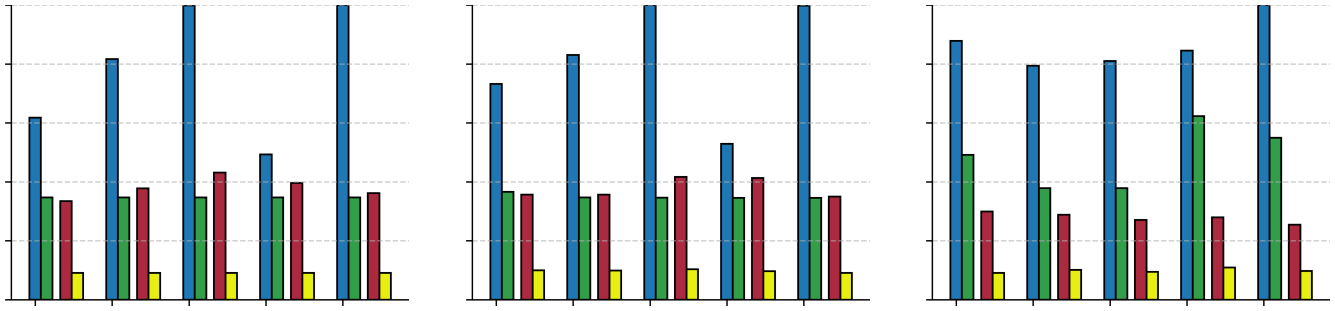


Figure 1: Illustration of Dissimilarity Invariance. Semantic similarity (e.g., horse vs. dog, as highlighted in the orange region above) is vulnerable to label noise, exhibiting significant variations under corruption. In contrast, inter-class dissimilarity (e.g., automobile vs. frog, shown in the green region below) remains largely stable.

Motivated by this observation, we shift focus from fragile similarity toward reliable dissimilarity as the robust anchor. We present Negative Scale (NegScale), a novel, plug-and-play framework designed to robustly model dissimilarity and suppress spurious similarity under label noise. NegScale comprises two modules: Structured Negative Orthogonality Penalty (SNOP), which enforces orthogonality among negative pairs within each minibatch to ensure that genuine



(a) Clean-Clean pairs ($y_i = \tilde{y}_i, y_j = \tilde{y}_j$) (b) Clean-Noisy pairs ($y_i = \tilde{y}_i, y_j \neq \tilde{y}_j$) (c) Cross-noise pairs ($y_i = \tilde{y}_i = \tilde{y}_j \neq y_j$)

Figure 2: Key observations of dissimilarity invariance under synthetic noises with CIFAR-10. Small / Large pLCA refers to semantically related and semantically unrelated sample pairs, respectively. Small / Large Clean indicates the similarity of the corresponding pairs when trained under clean labels. The y-axis represents the normalized similarity. The x-axis shows the noise rates, from left to right: [Sym 20%, Sym 50%, Sym 80%, Ins 20%, Ins 40%]. The different colors correspond to the sets of Small pLCA, Small Clean, Large pLCA, and Large Clean.

negative pairs occupy distinct subspaces; and Dissimilarity-Calibrated Similarity Adjustment (DCSA), which leverages dissimilarity anchors to identify and down-weight misleading similarity, thereby preventing the model from internalizing false associations. We further provide a theoretical analysis to explain the underlying cause of Dissimilarity Invariance and to justify the effectiveness of NegScale. Requiring no external data or pre-training, NegScale can be seamlessly integrated into existing training pipelines. Empirical evaluations on both synthetic and real-world noisy-label benchmarks confirm that our dissimilarity-centric approach consistently outperforms SOTA methods. In a nutshell, our contributions are as follows:

- We observe a key phenomenon, Dissimilarity Invariance, wherein the similarity between semantically unrelated samples remains notably more robust under label noise.
- We propose a novel framework, called NegScale, which enables the model to capture more accurate dissimilarity relationships while simultaneously suppressing spurious similarity via dissimilarity anchors, thereby enhancing its robustness.
- Experimental results show that our method advances state-of-the-art results on CIFAR with synthetic label noise, as well as on real-world noisy datasets.

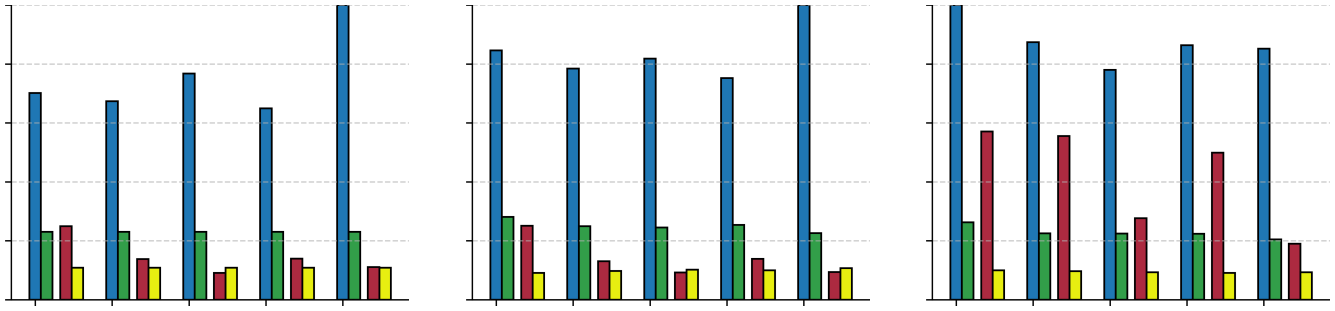
Releated Work

Numerous research have recently addressed the issue of learning from noisy labels. Depending on how they approach handling noisy datasets, we categorize the current algorithms into three main groups and one subcategory.

Loss Correction. In order to reduce the negative impact of noisy labels, loss correction methods adjust the loss of all training samples before updating the parameters of the model (Arazo et al. 2019; Li et al. 2017b; Song, Kim, and Lee 2019). The estimated noise transition matrix or other approaches are used by the loss correction methods to modify the loss of all training samples, which is then used to update the network parameters . However, the noise transition matrix's parameters are highly challenging to estimate, and comparable adjustments made to all samples invariably suffer from the accumulated incorrect correction. This can have an enormous impact on the model's final performance (Jiang et al. 2018; Han et al. 2018; Song et al. 2020). *Sample Selection.* To avoid the false correction, many studies use the sample selection to improve the performance of the model (Chen et al. 2019; Song, Kim, and Lee 2019; Han et al. 2018). ANNE (Cordeiro and Carneiro 2025) uses integrates loss-based sampling with the feature-based sampling methods FINE and Adaptive KNN. However, this family of methods discard a very large number of samples and select only a portion of samples for learning, which are easy to result in knowledge waste. *Semi-supervised Learning.* In terms of the problems with the previous methods such as false correction and knowledge waste, Semi-supervised Learning is proposed and has achieved excellent results in recent years (Song et al. 2020; Chen et al. 2023; Fan and Li 2025). The core idea of Semi-supervised Learning is treating the possibly noisy samples as unlabeled, whereas the rest samples as labeled. However, this family of methods requires careful setting of hyperparameters, which indirectly increases the complexity of the calculation.

Methodology

Preliminary. Let $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$ denote a training set with noisy labels, where $x_i \in \mathcal{X}$ is the input sample and $\tilde{y}_i \in \mathcal{Y} = \{1, \ldots, C\}$ is its (possibly corrupted) label, the ground-truth label is y_i, N is the number of samples, C is the number of classes and $\mathcal{B}$ is the batch. We denote the feature extractor by $f(\cdot;\theta)$ and the classifier head by $g_c(\cdot)$. The full model prediction is given by $g_c(f(x_i))$. The noise rate is τ. We define the cosine similarity between two feature vectors $f_i = f(x_i)$ and $f_j = f(x_j)$ as: $s_{ij} = \frac{f_i^\top f_j}{\|f_i\| \cdot \|f_j\|}$. Our goal is to train a robust model that learns clean feature representations even under such label corruption.



(a) Clean-Clean pairs ($y_i = \tilde{y}_i, y_j = \tilde{y}_j$) (b) Clean-Noisy pairs ($y_i = \tilde{y}_i, y_j \neq \tilde{y}_j$) (c) Cross-noise pairs ($y_i = \tilde{y}_i = \tilde{y}_j \neq y_j$)

Figure 3: Key observations of dissimilarity invariance under various noise types of real world noisy dataset CIFAR-10N. Small / Large pLCA refers to semantically related and semantically unrelated sample pairs, respectively. Small / Large Clean indicates the similarity of the corresponding pairs when trained under clean labels. The x-axis shows the noise types, from left to right: [Aggre, Rand1, Rand2, Rand3, Worst]. The legend is consistent with Fig. 2.

Noise Type	**Acc**	**Same Class**	**Small pLCA**	**Large pLCA**
Clean	95.72	0.94	0.48	0.44
Sym 20%	87.32	0.80	0.54	0.48
Sym 50%	80.04	0.74	0.55	0.48
Sym 80%	56.89	0.63	0.58	0.49
Aggre 8%	92.16	0.89	0.68	0.49
Rand1 17%	89.13	0.82	0.66	0.45
Rand2 17%	89.14	0.84	0.67	0.43
Rand3 17%	89.39	0.83	0.65	0.45
Worst 40%	83.31	0.82	0.73	0.44
Mean	-	0.80	0.60	0.46
Δ With Clean	-	-0.14	0.12	0.03

Table 1: Accuracy and similarity variations of ResNet-18 trained with Cross-Entropy on CIFAR-10 under various noise. Line 3-7 correspond to synthetic noise settings, while Line 8-12 represent real-world noise from CIFAR-10N.

How to Determine the Semantic Relatedness between Two Samples and Identify Negative Pairs? We use the pair Lowest Common Ancestor (pLCA) distance (Bertinetto et al. 2020; Shi et al. 2024), which measures the semantic distance between two classes based on class taxonomy, as demonstrated in Eq. (1). A lower pLCA score indicates a closer semantic relationship between two classes $(y, \hat{y})$.

$$D_{pLCA}(y, \hat{y}) \stackrel{\text{def}}{=} |h(y) - h(N_L(y, \hat{y}))| + |h(\hat{y}) - h(N_L(y, \hat{y}))| \quad (1)$$

where h represents the tree depth of a node and N_L is the lowest common ancestor class node for the label within the hierarchy. The predefined taxonomic hierarchy of pLCA distance is shown in the Appendix. We can leverage the pairwise label pLCA distance to determine the semantic relatedness between two samples. Samples are considered semantically related if their pLCA distance is below threshold $\eta_{\min}$, and semantically unrelated if exceeding threshold $\eta_{\max}$, $\eta_{\min} \leq \eta_{\max}$. Within each training epoch, we first freeze the model parameters to compute inter-sample similarities and pLCA distances, get and normalize the set of semantically related samples $\mathcal{P}_{\text{sim}} = \{(i,j) | 0 < D_{pLCA}(\tilde{y}_i, \tilde{y}_j) \leq \eta_{\min}\}$ and unrelated samples $\mathcal{P}_{\text{neg}} = \{(i,j) | D_{pLCA}(\tilde{y}_i, \tilde{y}_j) > \eta_{\max}\}$, then unfreeze the parameters to perform gradient-based optimization of the loss function.

Key Observations of Dissimilarity Invariance

While LNL methods can extract robust features under noisy labels, they often neglect structural relationships among samples, causing learned similarities to become fragile and susceptible to error amplification. To investigate this issue, we examine sample-pair similarities across different noise types. Our analysis reveals a clear contrast: similarities between semantically related samples (small pLCA) vary significantly as noise increases, whereas those between unrelated samples (large pLCA)—which we refer to as dissimilarity—remain remarkably stable. Negative pairs consistently maintain low cosine similarity with minimal variance across noise rates (Tab. 1). We term this effect **Dissimilarity Invariance**, emphasizing that label noise primarily distorts positive relations while leaving dissimilar pairs largely unaffected.

To further investigate Dissimilarity Invariance, we categorize sample pairs based on the degree of label corruption and analyze how their similarity deviates, as shown in Figs. 2 and 3. Specifically, we divide each sample pair (i, j) into three categories: (1) Clean-Clean pairs: both samples have clean labels, as $y_i = \tilde{y}_i, y_j = \tilde{y}_j$; (2) Clean-Noisy pairs: one sample has a noisy label, as $y_i = \tilde{y}_i, y_j \neq \tilde{y}_j$; (3) Cross-noise pairs: one sample has a noisy label, and that noisy label coincides with the clean label of the other sample, as $y_i = \tilde{y}_i = \tilde{y}_j, y_j \neq \tilde{y}_j$. Figs. 2 and 3 further support three key conclusions: (1) Dissimilarity Invariance holds universally, regardless of whether the samples are affected by noise, as shown in Figs. 2(a), 2(b), 3(a) and 3(b). (2) The effect of Dissimilarity Invariance is more pronounced under real-world noise compared to synthetic noise. (3) Mislabeled samples significantly disrupt similarity within semantically

Dataset	CIFAR-10					CIFAR-100				
Noise Type	**Sym**			**Pair**	**Ins**	**Sym**			**Pair**	**Ins**
Method / Noise Rate	**20%**	**50%**	**80%**	**40%**	**40%**	**20%**	**50%**	**80%**	**40%**	**40%**
Co-teaching (Han et al. 2018)	88.2	50.7	21.1	55.3	59.5	58.5	33.0	5.8	39.2	40.7
DivideMix (Li, Socher, and Hoi 2020)	95.7	94.4	92.9	92.1	95.1	76.9	74.2	59.6	52.3	76.1
Co-learning (Tan et al. 2021)	91.8	79.3	37.0	66.3	78.9	70.3	63.9	38.9	49.1	62.9
SELC+ (Lu and He 2022)	94.9	87.2	78.6	88.1	84.2	76.4	62.4	37.2	45.2	44.3
RoLR (Chen et al. 2023)	96.4	95.7	94.2	92.8	93.7	78.6	74.6	66.2	76.1	77.2
RankMatch (Zhang et al. 2023)	96.4	95.4	94.2	94.4	93.8	79.3	77.6	67.2	75.8	76.5
CrossSplit (Kim et al. 2023)	96.9	96.3	95.4	96.0	95.8	79.9	75.7	64.6	76.8	79.2
DMLP (Naive) (Tu et al. 2023)	94.2	94.0	93.2	93.9	93.2	72.3	70.1	63.2	71.8	72.2
DMLP (DivideMix) (Tu et al. 2023)	96.2	95.6	94.3	95.0	95.4	79.4	76.1	68.5	76.4	78.9
CCL (Fan and Li 2025)	97.0	96.5	94.6	96.1	96.2	79.5	77.4	70.3	77.2	80.0
ANNE (Cordeiro and Carneiro 2025)	96.9	96.2	95.3	95.7	96.2	80.4	78.1	**73.0**	66.4	78.4
RoLR + NegScale (Ours)	**97.2**	**96.6**	**95.6**	**96.3**	**96.5**	**80.9**	**78.7**	70.8	**77.8**	**80.4**

Table 2: Comparison with state-of-the-art methods on CIFAR-10/100 datasets under various types of noise. The results of other methods are from the published results of corresponding papers. The best results are indicated in bold.

related pairs, but their impact on negative pairs remains notably smaller, as illustrated in Figs. 2(c) and 3(c).

NegScale: Learning with Stable Dissimilarity

Building on our core observation that dissimilarity between negative pairs remains stable even under label noise, we derive two key insights: (1) accurate dissimilarity can still be learned despite noisy supervision; and (2) such reliable dissimilarity can be leveraged to rectify corrupted similarity signals. Guided by these insights, we propose NegScale, a robust learning framework composed of two complementary modules: SNOP, which enforces structured orthogonality among negative pairs, and DCSA, which calibrates noisy similarities using dissimilarity-aware adjustments.

Structured Negative Orthogonality Penalty. SNOP enforces orthogonality among negative pairs at the batch level, encouraging their separation in the representation space. The underlying idea is that dissimilar samples should lie in orthogonal directions, reflecting semantic independence—a principle shown to improve robustness in prior work (Guo et al. 2022; Yuan and Yang 2022). We detail the formulation of SNOP below. First, we define the difference vector set as:

$$F_{\text{neg}} = \left\{ \frac{f_i - f_j}{\|f_i - f_j\|_2} \,|\, (i,j) \in \mathcal{P}_{\text{neg}} \right\} \tag{2}$$

where each column corresponds to a normalized difference between a dissimilar pair. Intuitively, if all these directions are orthogonal, then $F_{\text{neg}}F_{\text{neg}}^\top$ should approximate the identity matrix I. So we define the global orthogonality loss as:

$$\mathcal{L}_{\text{global}} = \|F_{\text{neg}}F_{\text{neg}}^\top - I\|_F^2 \tag{3}$$

While $F_{\text{neg}}F_{\text{neg}}^\top$ captures global repulsion structure, it may overlook local misalignments caused by particularly confusing or uncertain samples. To address this, we introduce a confidence-weighted local orthogonality loss that directly penalizes inner product alignment between negative pairs:

$$\mathcal{L}_{\text{local}} = \sum_{(i,j)\in\mathcal{P}_{\text{neg}}} w_{ij} \cdot (f_i^\top f_j)^2 \tag{4}$$

where the weight w_{ij} is defined based on per-sample confidence as:

$$w_{ij} = (1 - c_i) \cdot (1 - c_j) \tag{5}$$

where c_i is the confidence of sample i's label prediction, computed as the softmax probability of its predicted class. This design assigns higher penalties to pairs involving low-confidence (potentially noisy) samples, enforcing stronger dissimilarity. In contrast, high-confidence pairs are deemed more reliable and penalized less. The full SNOP loss is a combination of the global and local penalties:

$$\mathcal{L}_{\text{SNOP}} = \|F_{\text{neg}}F_{\text{neg}}^\top - I\|_F^2 + \sum_{(i,j)\in\mathcal{P}_{\text{neg}}} w_{ij} \cdot (f_i^\top f_j)^2 \tag{6}$$

Dissimilarity-Calibrated Similarity Adjustment. Under label noise, many such pairs are mislabeled, leading to spurious similarity (Figs. 2 and 3). To address this, DCSA imposes soft upper bounds on similarity using dissimilarity, preventing mismatched pairs from becoming overly similar. For each pair $(i,j) \in \mathcal{P}_{\text{sim}}$, we assess their proximity to dissimilar samples by computing a adaptive calibration factor:

$$\delta_{ij} = \frac{1}{2} \max_{(p,q)\in\mathcal{P}_{\text{neg}}} (f_i^\top f_p + f_j^\top f_q), \tag{7}$$

This reflects how much i and j are entangled with unrelated samples. A large δ_{ij} suggests that the pair may be semantically ambiguous or mislabeled, and thus their similarity should be restricted. To prevent the model from learning overly confident similarity on such risky pairs, we define the DCSA loss as:

$$\mathcal{L}_{\text{DCSA}} = \sum_{(i,j)\in\mathcal{P}_{\text{sim}}} \left[\max(0, s_{ij} - (1 - \delta_{ij}))\right]^2 \tag{8}$$

Eq. (8) penalizes positive pairs whose similarity exceeds a soft upper bound $(1 - \delta_{ij})$, which adapts based on their proximity to negative samples. DCSA acts as a dissimilarity-aware gate that prevents overfitting to noisy positives.

Final Objective. The total training loss is composed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{org}} + \lambda \mathcal{L}_{\text{SNOP}} + \mu \mathcal{L}_{\text{DCSA}} \tag{9}$$

The coefficients λ and μ balance dissimilarity enforcement and noisy similarity suppression. As our method requires no additional information, it can serve as a plug-in by replacing $\mathcal{L}_{\text{org}}$ with the desired loss in existing methods.

Theoretical Analysis

Why Negative-Pair Similarity Remains Invariant. We show that, under *symmetric* label noise and vanilla cross-entropy SGD, the expected variation in the similarity of negative pairs is smaller than that of semantically related pairs via first-order change in cosine similarity.

Let $f_i(t) \in \mathbb{R}^d$ be the ℓ_2-normalized feature of sample x_i at SGD iteration t. The linear classifier is $W = [w_1, \ldots, w_C] \in \mathbb{R}^{d \times C}$; each logit is $p_i^{(k)} = w_k^\top f_i, \quad k = 1, \ldots, C$. From the nature of semantic similarity, we know that semantically related pairs $(i, j) \in \mathcal{P}_{\text{sim}}$ tend to have a higher overlap in their logits ($z_i^{(k)} \approx z_j^{(k)}$ for most k), while semantically unrelated pairs $(i', j') \in \mathcal{P}_{\text{neg}}$ exhibit minimal overlap. In other words, the Kendall tau distance (Kendall 1938) between the logits of semantically similar pairs $\pi(i, j)$ is typically smaller than that of negative pairs $\pi(i', j')$. A single (unnormalized) gradient step on f_i is

$$\begin{aligned} f_i(t+1) &= f_i(t) - \gamma\, g_i \\ g_i = \nabla_{f_i} \mathcal{L}_i &= \sum_{k=1}^{C} p_i^{(k)}\, w_k \;-\; w_{\tilde{y}_i} \end{aligned} \tag{10}$$

where γ is the learning rate. After renormalization $\|f_i(t+1)\| = 1$, the first-order change in feature is $\Delta f_i = -\gamma\, g_i + O(\gamma^2)$. For a pair (i, j) with $y_i \neq y_j$, define $s_{ij}(t) = f_i(t)^\top f_j(t)$. After one gradient step (dropping $O(\gamma^2)$ and renormalization effects), the updated similarity is

$$\begin{aligned} s_{ij}(t+1) &= f_i(t+1)^\top f_j(t+1) \\ &= \big(f_i - \gamma g_i\big)^\top \big(f_j - \gamma g_j\big) \\ &= s_{ij} - \gamma \langle g_i,\, f_j \rangle - \gamma \langle f_i,\, g_j \rangle + O(\gamma^2), \end{aligned} \tag{11}$$

so the first-order increment is

$$|\Delta s_{ij}| = |s_{ij}(t+1) - s_{ij}(t)| \approx |\gamma \Big[\langle g_i,\, f_j \rangle + \langle f_i,\, g_j \rangle \Big]| \tag{12}$$

Without loss of generality, we first examine $\gamma \langle g_i,\, f_j \rangle$:

$$\begin{aligned} \gamma \langle g_i,\, f_j \rangle &= \gamma \big[(\sum_{k=1}^{C} p_i^{(k)}\, w_k \;-\; w_{\tilde{y}_i}) \cdot f_j \big] \\ &= \gamma \big[\sum_{k=1}^{C} p_i^{(k)}\, w_k f_j \;-\; w_{\tilde{y}_i} f_j \big] \\ &= \gamma \sum_{k=1}^{C} p_i^{(k)} p_j^{(k)} - p_j^{(\tilde{y}_i)} \approx \gamma \sum_{k=1}^{C} p_i^{(k)} p_j^{(k)} \end{aligned} \tag{13}$$

We drop $p_j^{(\tilde{y}_i)}$ because, for a model with reasonable discriminative ability, this value tends to be small when $y_i \neq y_j$, and can therefore be safely ignored. A similar conclusion also holds for $\gamma \langle f_i,\, g_j \rangle$. So for $(i, j) \in \mathcal{P}_{\text{sim}}$ and $(i', j') \in \mathcal{P}_{\text{neg}}$:

$$\begin{aligned} |\Delta s_{ij}| - |\Delta s_{i'j'}| &= |2\gamma \Big[\sum_{k=1}^{C} p_i^{(k)} p_j^{(k)} - \sum_{k=1}^{C} p_{i'}^{(k)} p_{j'}^{(k)} \Big]| \\ &= |2\gamma \sum_{k=1}^{C} \Big[p_i^{(k)} p_j^{(k)} - p_{i'}^{(k)} p_{j'}^{(k)} \Big]| \end{aligned} \tag{14}$$

From the benigning, we know that the Kendall tau distance of (i, j) is greater than that of (i', j'). According to the rearrangement inequality (Cvetkovski 2012), this implies that Eq. (14) > 0. This result indicates that, during each update, the similarity changes more significantly for semantically similar pairs than for semantically dissimilar ones, directly supporting the validity of Dissimilarity Invariance.

Why NegScale is Effective. We theoretically show that NegScale mitigate noise-induced feature perturbations and margin erosion. Proofs appear in the appendix.

Lemma 1 (Feature Perturbation Bound under NegScale) *Let f_i and $\tilde{f}_i$ be the feature representations learned under clean and noisy labels respectively. Suppose the feature extractor f is trained with NegScale. Assume each of $\mathcal{L}_{SNOP}$ and $\mathcal{L}_{DCSA}$ is minimized to at most $\delta_{SNOP}, \delta_{DCSA}$ respectively, and that the gradient norm of cross-entropy loss on noisy labels is bounded by $\|\nabla \mathcal{L}_{CE}\| \leq G$. Then the feature perturbation due to noisy labels is upper bounded as:*

$$\|\tilde{f}_i - f_i\| \leq \frac{G \cdot \tau}{\lambda \cdot \delta_{SNOP} + \mu \cdot \delta_{DCSA}} \tag{15}$$

Remark 1 *This lemma quantifies how much the learned feature $\tilde{f}_i$ deviates from its clean counterpart f_i under noisy supervision. The bound highlights that this deviation increases linearly with the noise rate τ but is effectively suppressed by the regularization strengths λ and μ. As a result, incorporating SNOP and DCSA reduces representation instability caused by corrupted labels.*

Following prior works on generalization and robustness under label noise (Bartlett, Foster, and Telgarsky 2017; Huh and Rebeschini 2024), we assume that a linear classifier $W = [W_1, \ldots, W_C] \in \mathbb{R}^{d \times C}$ predicts class labels via:

$$\hat{y}_i = \arg\max_c W_c^\top f_i \tag{16}$$

and the classifier achieves a clean margin $\gamma_0 > 0$, i.e.,

$$W_{y_i}^\top f_i - \max_{c \neq y_i} W_c^\top f_i \geq \gamma_0. \tag{17}$$

Theorem 1 (Classification Error Bound under NegScale) *Let $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be a normalized feature extractor, and let $\mathcal{F} = \{f_i = f_\theta(x_i)\}$ denote the feature representations learned on clean labels. Let $\tilde{\mathcal{F}} = \{\tilde{f}_i\}$ denote the feature representations learned under uniform label noise with rate $\tau < \frac{1}{2}$. The generalization error under label noise using NegScale is bounded by:*

$$|\mathbb{E}\left[\mathbb{1}(\hat{y}_i^{clean} \neq y_i)\right] - \mathbb{E}\left[\mathbb{1}(\hat{y}_i^{noisy} \neq y_i)\right]| \;\leq\; \frac{2\epsilon}{\gamma_0} \tag{18}$$

where $\epsilon = (G \cdot \tau)/(\lambda \cdot \delta_{SNOP} + \mu \cdot \delta_{DCSA})$.

Dataset		CIFAR-10N					CIFAR-100N
Method	Noise Type Noise Rate	Aggre 9.0%	Rand1 17.2%	Rand2 18.12%	Rand3 17.64%	Worst 40.2%	Fine 40.2%
Co-teaching (Han et al. 2018)		89.9	87.8	87.2	87.4	62.3	40.5
JoCoR (Wei et al. 2020)		90.6	88.8	88.5	88.1	66.7	40.1
DivideMix (Li, Socher, and Hoi 2020)		93.2	92.8	92.6	93.1	89.2	55.2
Co-learning (Tan et al. 2021)		92.4	91.3	91.2	91.4	81.0	47.9
RoLR (Chen et al. 2023)		95.4	94.9	94.7	95.2	92.3	62.3
RankMatch (Zhang et al. 2023)		95.6	94.8	95.1	95.3	92.8	65.2
CCL(Fan and Li 2025)		96.4	96.0	95.8	96.1	93.1	65.5
ANNE (Cordeiro and Carneiro 2025)		96.2	95.7	95.5	95.9	93.0	66.0
RoLR + NegScale (Ours)		**96.6**	**96.2**	**96.0**	**96.4**	**93.5**	**66.3**

Table 3: Comparison with state-of-the-art methods on CIFAR-N. The results are from (Wei et al. 2023) and our replication.

NCT	RoLR	DISC	CCL	ANNE	Ours
84.1	88.5	87.1	89.7	88.2	**90.7**

Table 4: Comparison with other methods on Animal-10N. The results of other methods are from (Cordeiro and Carneiro 2025).

Remark 2 *Theorem 1 shows that, under small feature perturbations ϵ—which are effectively controlled by NegScale—the increase in generalization error due to label noise is linearly bounded in ϵ/γ. Thus, robustness to label noise is directly linked to the preservation of dissimilarity constraints during training.*

Experiments

Datasets. To verify the effectiveness of our method, we perform our method on classification tasks with six benchmarks: CIFAR-10 (Krizhevsky, Hinton et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), CIFAR-10N (Wei et al. 2022), CIFAR-100N (Wei et al. 2022), Animal-10N (Song, Kim, and Lee 2019) and WebVision (Li et al. 2017a). The last four benchmarks are real-world noisy datasets.

Implementation Details. We conduct experiments using noise types Symmetric (Sym), Asymmetric (Pair), and Instance-dependent (Ins) noise for evaluation. All reported results are averaged over the last 10 training epochs. For methods lacking available results in original papers, we reimplemented and reproduced them under the same evaluation protocol. The weak and strong data augmentations used follow the settings in (Chen et al. 2023). For hyperparameters, we set $\lambda = \mu = 1, \eta_{\max} = 5, \eta_{\min} = 3$ as default. Since our method is designed as a plug-in, we combine it with RoLR as the default setting throughout experiments. Further implementation details and descriptions of Noise Injections are provided in the Appendix.

Experimental Results

Results on CIFAR with Synthetic Noise. Tab. 2 shows that our plug-in method, when combined with RoLR as the base, consistently outperforms state-of-the-art approaches across all noise levels on both CIFAR-10 and CIFAR-100 under various synthetic noise settings. In comparison to sample selection methods such as RankMatch (Zhang et al. 2023), our approach achieves a notable performance gain of 3.6% (70.8% vs. 67.2%) on CIFAR-100 with 80% noise. We also surpass DMLP (Tu et al. 2023) and ANNE (Cordeiro and Carneiro 2025), nearly across all noise settings, with especially significant improvements on the more challenging CIFAR-10 and CIFAR-100 under heavy noise.

Method	WebVision		ILSVRC12	
	top-1	top-5	top-1	top-5
DSOS	77.8	92.0	74.4	90.8
DivideMix	77.3	91.6	75.2	90.8
UNICON	77.6	93.4	75.3	93.7
RoLR	81.8	94.1	75.5	93.8
RankMatch	79.9	93.6	77.4	94.3
CCL	82.3	94.6	78.2	94.9
ANNE	82.0	94.0	76.8	92.7
RoLR + NegScale	**83.1**	**94.8**	**79.4**	**95.0**

Table 5: Comparison with state-of-the-art methods on (mini) WebVision dataset. Numbers denote top-1 (top-5) accuracy (%) on the WebVision and ImageNet ILSVRC12 validation set. The results of other methods are from (Zhang et al. 2023; Cordeiro and Carneiro 2025).

Results on Real-world Datasets. Tab. 3, Tab. 4, and Tab. 5 present results on CIFAR0N, Animal-10N, and WebVision, respectively. Our method consistently outperforms all competing approaches across these real-world noisy datasets, demonstrating strong robustness and generalizability. Notably, when compared to UNICON (Karim et al. 2022)—a hybrid method that integrates semi-supervised learning with contrastive learning—our method surpasses SOTA by over 3% in top-1 accuracy on both the mini-WebVision and ILSVRC12 validation sets, while also matching the best top-5 accuracy on these benchmarks. These results highlight the effectiveness of NegScale in real-world noisy scenarios.

Results of Plug-in with Various Method. We validate our method's plug-and-play capability by integrating it with DivideMix, RankMatch, and ANNE. As Table Tab. 6 shows, incorporating our dissimilarity-based regularization consis-

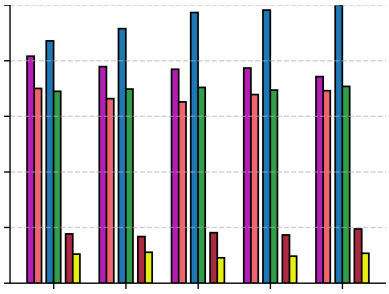


Figure 4: Results on similarity variation after training with NegScale with CIFAR-10 under 50% symmetric noise. For better visualization, similarity values of *Same Class* and *Same Clean* pairs are divided by 2 to avoid distortion due to their larger magnitudes. The legend is consistent with Fig. 2.

Dataset	CIFAR-10		CIFAR-100	
Noise ratio	**50%**	**80%**	**50%**	**80%**
DivideMix	94.4	92.9	74.2	59.6
+ NegScale	**96.5**	**94.6**	**77.4**	**70.3**
RankMatch	95.4	94.2	77.6	67.2
+ NegScale	**96.6**	**95.4**	**78.1**	**70.8**
ANNE	96.2	95.3	78.1	73.0
+ NegScale	**96.6**	**96.2**	**78.7**	**73.5**

Table 6: Test accuracy of NegScale combined with other methods on CIFAR-10/CIFAR-100 with symmetric noise.

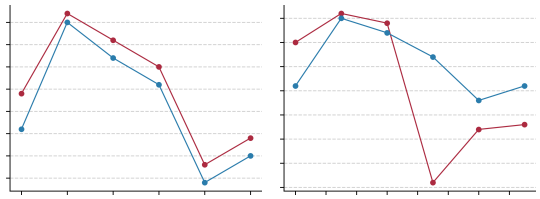
tently boosts test accuracy on CIFAR-10 and CIFAR-100 under symmetric noise. These results demonstrate its broad applicability and seamless integration with existing noisy-label learning frameworks.

Results on Similarity. Fig. 4 illustrates the changes in inter-sample similarity after training with NegScale. We observe that our method effectively preserves the dissimilarity between semantically unrelated samples, while also mitigating the undesired increase in similarity among semantically related but noisy pairs. This helps reduce the risk of learning spurious semantic correlations introduced by label noise. Notably, NegScale also increases the similarity between same-class pairs—unlike the degradation observed in Tab. 1—which we attribute to the model learning more robust representations through dissimilarity constraints. This suggests that emphasizing reliable dissimilarity can indirectly facilitate better alignment of truly related samples.

Effects of Components of NegScale. We remove the corresponding components to study the effects of each component of our method, such as SNOP and DCSA, and compare them with the full NegScale framework. As shown in Tab. 7, removing SNOP or DCSA leads to a significant drop in performance, especially under high noise ratios. This indicates that both components are crucial for the effectiveness of our method. Additionally, we observe that random selection of negative pairs, which does not leverage the semantic dissimilarity structure, results in a substantial performance degradation, further confirming the importance of structured dissimilarity learning. These results validate our design choices and highlight the effectiveness of NegScale in enhancing robustness against label noise.

Dataset	CIFAR-10		CIFAR-100	
Noise ratio	**50%**	**80%**	**50%**	**80%**
NegScale	96.6	95.6	78.7	70.8
w/o SNOP	95.1	94.3	76.7	67.2
w/o DSCA	95.3	94.8	77.0	68.1
w Random Selection	94.1	92.9	74.1	65.2

Table 7: Ablation study results of test accuracy (%) on CIFAR-10 and CIFAR-100 with symmetric noise.



(a) λ and μ (b) $\eta_{\min}$ and $\eta_{\max}$

Figure 5: Ablation study on hyperparameters with CIFAR-10 under different noises (Sym 0.5, Ins 0.4).

Sensitivity Analysis. Our framework introduces four hyperparameters—λ and μ in the total loss (Eq. (9)) and $\eta_{\min}, \eta_{\max}$ for negative-pair selection—and we evaluate their impact via sensitivity analysis (see Figs. 5(a) and 5(b)). In Fig. 5(a), we evaluate different combinations, (0.1, 0.1), (1, 1), (5, 5), (10, 10), (2, 0.5), and (0.5, 2) (from left to right on the x-axis). We find that setting $\lambda = \mu = 1$ yields the best accuracy, demonstrating that a balanced SNOP-DCSA regularization is essential, whereas skewed weights impair robustness. In Fig. 5(b), we evaluate different combinations, (3,3), (3,5), (5,5), (1,7), (1,3), (4,6), (from left to right on the x-axis). Similarly, both overly strict and overly loose thresholds $\eta_{\min}, \eta_{\max}$ lead to degraded performance, highlighting the necessity of carefully calibrating the dissimilarity range to select informative negative pairs.

Conclusion

In this paper, we identify and formalize the phenomenon of *Dissimilarity Invariance*, where semantic dissimilarity between unrelated samples remains notably stable even under severe label noise. Motivated by this observation, we propose **NegScale**, a plug-in framework that explicitly exploits dissimilarity as a robust inductive signal for learning under noisy supervision. NegScale consists of two complementary modules: SNOP, which imposes structured orthogonality among negative pairs to enforce local dissimilarity constraints, and DCSA, which calibrates similarity learning by referencing stable dissimilarity across the feature space. Extensive experiments on both synthetic and real-world noisy datasets validate the effectiveness of our approach, showing consistent improvements over state-of-the-art baselines.

Acknowledgments

This work is supported by Beijing Natural Science Foundation (No.4222037, L181010).

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2019. Unsupervised Label Noise Modeling and Loss Correction. In Chaudhuri, K.; and Salakhutdinov, R., eds., *ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 312–321. PMLR.

Bartlett, P. L.; Foster, D. J.; and Telgarsky, M. 2017. Spectrally-normalized margin bounds for neural networks. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 6240–6249.

Bertinetto, L.; Müller, R.; Tertikas, K.; Samangooei, S.; and Lord, N. A. 2020. Making Better Mistakes: Leveraging Class Hierarchies With Deep Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 12503–12512. Computer Vision Foundation / IEEE.

Bochkovskiy, A.; Wang, C.; and Liao, H. M. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *CoRR*, abs/2004.10934.

Chen, M.; Cheng, H.; Du, Y.; Xu, M.; Jiang, W.; and Wang, C. 2023. Two Wrongs Don't Make a Right: Combating Confirmation Bias in Learning with Label Noise. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 14765–14773. AAAI Press.

Chen, P.; Liao, B.; Chen, G.; and Zhang, S. 2019. Understanding and Utilizing Deep Neural Networks Trained with Noisy Labels. In Chaudhuri, K.; and Salakhutdinov, R., eds., *ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 1062–1070. PMLR.

Cordeiro, F. R.; and Carneiro, G. 2025. ANNE: Adaptive Nearest Neighbours and Eigenvector-based sample selection for robust learning with noisy labels. *Pattern Recognit.*, 159: 111132.

Cvetkovski, Z. 2012. *The Rearrangement Inequality*, 61–67. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-23792-8.

Fan, W.; and Li, K. 2025. Combating Semantic Contamination in Learning with Label Noise. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 2870–2878. AAAI Press.

Guo, K.; Zhou, K.; Hu, X.; Li, Y.; Chang, Y.; and Wang, X. 2022. Orthogonal Graph Neural Networks. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 3996–4004. AAAI Press.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS): 8527–8537.

Huh, J. E.; and Rebeschini, P. 2024. Generalization Bounds for Label Noise Stochastic Gradient Descent. In Dasgupta, S.; Mandt, S.; and Li, Y., eds., *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, 1360–1368. PMLR.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L. J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *35th International Conference on Machine Learning, ICML 2018*, 5(6): 3601–3620.

Karim, N.; Rizve, M. N.; Rahnavard, N.; Mian, A.; and Shah, M. 2022. UNICON: Combating Label Noise Through Uniform Selection and Contrastive Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 9666–9676. IEEE.

Kendall, M. G. 1938. A New Measure of Rank Correlation. *Biometrika*, 30(1/2): 81–93.

Kim, J.; Baratin, A.; Zhang, Y.; and Lacoste-Julien, S. 2023. CrossSplit: Mitigating Label Noise Memorization through Data Splitting. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 16377–16392. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, J.; Socher, R.; and Hoi, S. C. H. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Gool, L. V. 2017a. WebVision Database: Visual Learning and Understanding from Web Data. *CoRR*, abs/1708.02862.

Li, Y.; Yang, J.; Song, Y.; Cao, L.; Luo, J.; and Li, L. 2017b. Learning from Noisy Labels with Distillation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 1928–1936. IEEE Computer Society.

Lu, Y.; and He, W. 2022. SELC: Self-Ensemble Label Correction Improves Learning with Noisy Labels. In *International Joint Conference on Artificial Intelligence*.

Marriott, R. T.; Romdhani, S.; and Chen, L. 2021. A 3D GAN for Improved Large-Pose Facial Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 13445–13455. Computer Vision Foundation / IEEE.

Shi, J.; Gare, G. R.; Tian, J.; Chai, S.; Lin, Z.; Vasudevan, A. B.; Feng, D.; Ferroni, F.; and Kong, S. 2024. LCA-on-the-Line: Benchmarking Out of Distribution Generalization with Class Taxonomies. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Song, H.; Kim, M.; and Lee, J. 2019. SELFIE: Refurbishing Unclean Samples for Robust Deep Learning. volume 97 of *Proceedings of Machine Learning Research*, 5907–5915. PMLR.

Song, H.; Kim, M.; Park, D.; and Lee, J. 2020. Learning from Noisy Labels with Deep Neural Networks: A Survey. *CoRR*, abs/2007.08199.

Tan, C.; Xia, J.; Wu, L.; and Li, S. Z. 2021. Co-learning: Learning from Noisy Labels with Self-supervision. In Shen, H. T.; Zhuang, Y.; Smith, J. R.; Yang, Y.; Cesar, P.; Metze, F.; and Prabhakaran, B., eds., *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 1405–1413. ACM.

Tu, Y.; Zhang, B.; Li, Y.; Liu, L.; Li, J.; Wang, Y.; Wang, C.; and Zhao, C. 2023. Learning from Noisy Labels with Decoupled Meta Label Purifier. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 19934–19943. IEEE.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularization. In *CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 13723–13732. Computer Vision Foundation / IEEE.

Wei, H.; Zhuang, H.; Xie, R.; Feng, L.; Niu, G.; An, B.; and Li, Y. 2023. Mitigating Memorization of Noisy Labels by Clipping the Model Prediction. In *ICML 2023*.

Wei, J.; Zhu, Z.; Cheng, H.; Liu, T.; Niu, G.; and Liu, Y. 2022. Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. In *ICLR 2022*. OpenReview.net.

Yuan, C.; and Yang, L. 2022. Large margin projection-based multi-metric learning for classification. *Knowl. Based Syst.*, 243: 108481.

Zhang, Z.; Chen, W.; Fang, C.; Li, Z.; Chen, L.; Lin, L.; and Li, G. 2023. RankMatch: Fostering Confidence and Consistency in Learning with Noisy Labels. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 1644–1654. IEEE.