Content Diversity-guided Ambiguity Mitigation for Open-Set Noisy Label Learning

Zhihao Zhou [1], **Rui Li** [1*], **Xueying Li** [1]

[1]School of Data Science, Qingdao University of Science and Technology, China
4023111009@mails.qust.edu.cn, rui.li@qust.edu.cn, ponneylxy@163.com

Abstract

Open-set noisy label learning faces a critical challenge in maintaining robust DNN performance when training data contain both in-distribution noisy (IDN) and out-of-distribution (OOD) samples. These noisy samples induce overconfident but erroneous predictions due to their ambiguous positions relative to category boundaries. Current methods address this by filtering noisy samples based on visual features alone, they fail to resolve the semantic ambiguity near decision boundaries, where limited visual cues lead to unreliable sample purification. To this end, we propose Content Diversity-guided Ambiguity Mitigation (**CDgAM**), a novel framework that leverages diverse contents to mitigate visual ambiguity in open-set noisy label learning. CDgAM leverages textual descriptions of intra-class commonality and inter-class disparity to dynamically refine semantic boundaries, reducing bias in prototype learning. To further suppress early-stage uncertainty in visual representations, we design a region-sensitive distillation regularization that transfers boundary-aware knowledge from a multimodal large language model to the target DNN. Extensive experiments conducted on various datasets with different noise levels demonstrate the effectiveness of our CDgAM, outperforming state-of-the-art methods for open-set noisy label learning.

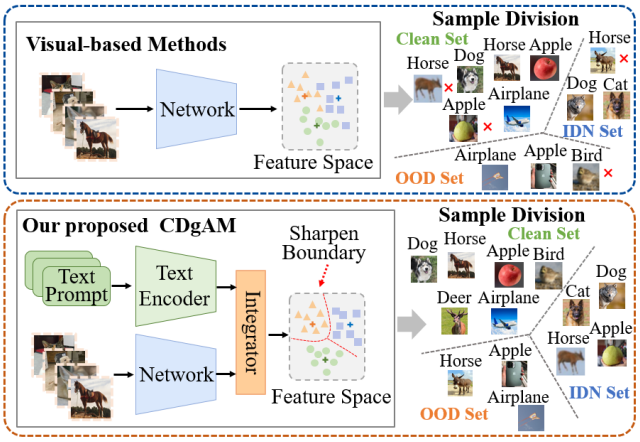


Figure 1: Visual-based methods vs. CDgAM. Visual-based methods are susceptible to assigning visually analogous or inferior-quality samples to erroneous sets owing to intrinsic noise and quality discrepancies. Our CDgAM can mitigate the ambiguity of category boundaries through multimodal semantic integration, enabling more precise sample division.

Introduction

Deep learning has achieved remarkable success by leveraging large-scale datasets. However, the data collection process often introduces incorrect annotations, resulting in noisy labels. These misleading labels can cause models to learn spurious features that misrepresent target categories, ultimately degrading both accuracy and generalization performance. A further challenge arises from the presence of out-of-distribution (OOD) samples—data points that fall outside the predefined classes. Due to their strong fitting capacity, deep neural networks (DNNs) may inadvertently learn erroneous patterns from OOD samples, further compromising model performance. To address these issues, open-set noisy label learning has emerged as a promising approach, which requires models to maintain robustness under the coexistence of in-distribution noisy (IDN) samples and OOD samples. This necessitates a dual-task framework, filtering out OOD samples while simultaneously selecting and correcting labels of IDN samples during training.

Most existing methods to noisy label learning primarily focus on IDN samples, with approaches such as (Tanaka et al. 2018; Zheltonozhskii et al. 2022; Li, Socher, and Hoi 2020) employing various ways to enhance DNN robustness and mitigate the effects of label noise. However, real-world data frequently contain OOD noise—samples that not only carry incorrect labels but also deviate from the training data distribution. While existing algorithms may perform well on purely in-distribution data, they often fail to generalize effectively when confronted with OOD samples. Accordingly, some recent works (Wang et al. 2018; Wu et al. 2021; Sun et al. 2020) consider OOD samples in the data and filter these samples through strategies such as density estimation, predicted confidence, small loss criterion, and so on. However, these methods are inevitably influenced by the visual features of the samples, resulting in low-quality and ambiguous category boundaries. As demonstrated in Figure 1,

*Corresponding author.

visual-based methods for clean sample selection are susceptible to assigning visually analogous or inferior-quality samples to erroneous sets owing to intrinsic noise and sample quality discrepancies. On the one hand, only a portion of labels are correct in the open-set noisy label learning task, manifesting in both limited intra-class diversity and poor inter-class discriminability. Consequently, the visual features learned only from images are inadequate to serve as comprehensive class descriptors. On the other hand, noisy labels, particularly those of OOD samples, often lead the model to assign high confidence to incorrect predictions, manifesting overconfident behavior. This bias continuously accumulates and intensifies as training progresses, further degrading the quality of the learned representations.

In this paper, we propose a Content Diversity-guided Ambiguity Mitigation (**CDgAM**) framework to address semantic deficiencies in open-set noisy label learning by leveraging diverse textual information. As demonstrated in Figure 1, our CDgAM mitigates the inherent ambiguity of category boundaries through multimodal semantic integration, enabling more precise sample division. Concretely, we design a Content Diversity-guided Semantic Synchronization (**CDgSS**) module, motivated by a key observation in open-set noisy label learning: Noisy samples are more prone to exacerbate the misguidance of categories with vague boundaries, making them hard to distinguish. Sole reliance on visual feature modeling is insufficient to ensure clear separability under mixed clean, IDN, and OOD samples. Specifically, CDgSS generates diverse contents of intra-class commonality and inter-class disparity to model holistic category relationships, thereby enhancing global category discriminability. And then multimodal synchronization is performed to establish precise semantic correspondences between visual samples and global category descriptions. These synergistic features are subsequently utilized to learn robust class prototypes, while a consistency-based strategy is adopted to disentangle clean, IDN, and OOD samples. Furthermore, we also introduce a Region-Sensitive Distillation Regularization (**RSDR**) term, which enables the early-stage DNNs to gain insight from the deeper semantics of the multimodal large language model. This facilitates bidirectional information exchange during training, thus further enhancing prototype learning and DNNs' prediction. Overall, our main contributions are summarized as follows:

- We propose a novel CDgAM framework for open-set noisy label learning, which leverages diverse contents to compensate for semantic deficiencies, category boundary ambiguity and mitigating prototype deviation to achieve precise sample purification.
- To mitigate early-stage representation uncertainty, we design a region-sensitive distillation regularization term that transfers deeper semantic knowledge from a multimodal large language model to the target DNN, thereby reducing error propagation during training.
- We evaluate our method on various datasets with different noise levels, demonstrating its effectiveness. In addition, multiple ablation experiments are conducted to analyze each component of our proposed CDgAM.

Related Work

Learning with Open-Set Noisy Data

The open-set noisy learning task presents a significant hurdle for DNN training as it introduces the complication of OOD errors that fall outside the defined class boundaries. Previous methods (Li, Xiong, and Hoi 2020) (Sun et al. 2020) (Wei et al. 2020) identify OOD samples based on simple criteria such as density estimation, predicted confidence, and standard loss criterion. These sample screening methods, largely inherited from in-distribution noise identification, are generally less effective at detecting OOD samples.

Recently, some efforts have been made to improve the performance of open-set noisy label learning. For example, PNP (Sun et al. 2022) introduces a low-cost projection head to compute the probabilities of a sample belonging to different noise types, which is then used as a basis for subsequent weight adjustment. MFT (Li, Yi, and Huang 2024) adopts the prototype-based paradigm and leverages the information from two parallel networks to collaboratively filter potential OOD samples. Despite the notable advancements introduced by these methods, they overlook the ambiguous of category boundaries caused by limited visual semantics. Besides, the early-stage prediction inaccuracies of DNNs in these methods may introduce cumulative errors, ultimately compromising their long-term effectiveness. In contrast, we leverage diverse texts to mitigate visual ambiguity in open-set noisy label learning and introduce a region-sensitive distillation regularization to compensate for initial visual uncertainty.

Large Model

Large models, especially large language models (LLMs, such as ChatGPT (Yang et al. 2023), Bert (Devlin 2018), Gemini (Team et al. 2024)) and multimodal large language models (MLLMs, such as CLIP (Radford et al. 2021), BLIP (Li et al. 2022)), have seen rapid development in recent years. They are trained on vast amounts of data, which enables them to acquire a breadth of knowledge that traditional smaller models cannot achieve. Due to large-scale pretraining and learning from massive amounts of data, large models reduce their reliance on labeled data, allowing for easy knowledge transfer and adaptation to various downstream tasks with only a small amount of fine-tuning data. For example, GLaMM (Rasheed et al. 2024) seamlessly integrates generative large language models with object segmentation masks, leveraging the large model to facilitate fine-grained image segmentation. GeoChat (Kuckreja et al. 2024) leverages the strong transfer learning capabilities of large models, fine-tuning them on remote sensing data to extend LLM applications across categories of high-resolution remote sensing images. Motivated by this, we exploit the vast knowledge contained within LLMs to generate diverse content that captures intra-class commonality and inter-class disparity, modeling holistic concept relationships. Moreover, we introduce the region-sensitive distillation regularization term, enabling knowledge transfer between early-stage DNNs and the deeper semantics of MLLMs, which further enhances the robustness of DNNs to noisy labels.

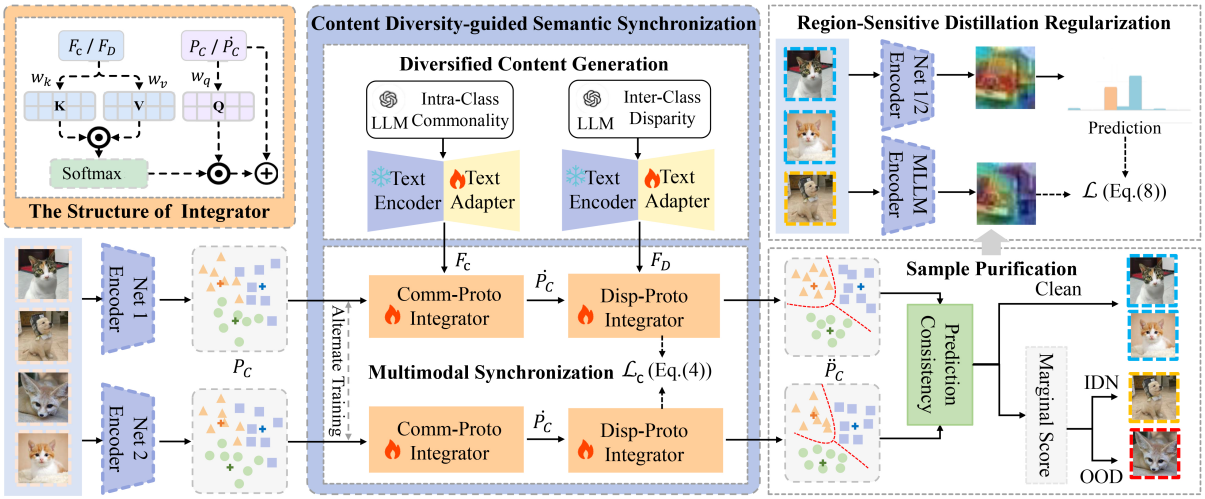


Figure 2: Overview of the proposed CDgAM. Content Diversity-guided Semantic Synchronization leverages an LLM to generate both intra-class commonality and inter-class disparity descriptions, which are aligned with visual features to enhance prototype learning. Then a consistency-based strategy separates clean, IDN, and OOD samples. Finally, Region-Sensitive Distillation Regularization is introduced to transfer MLLM semantics for guiding model training.

The Proposed Method

Preliminary

Problem Formulation. We consider a classification problem with a training set $D_{train} = \{x_i, y_i\}_{i=1}^{N}$, where x_i represents an image sample, and $y_i \in \mathbb{C} = \{1, \ldots, C\}$ is the label. In the open-set noisy label learning task, the true label y_{gt} of a sample in D_{train} is unknown, and only the potentially noisy label y_i is available. Based on the reliability of these labels, the training set can be divided into three subsets: **Clean samples**, where the true class of the sample belongs to the set $\mathbb{C}$ and the given label is correct, i.e., $y_i = y_{gt} \in \mathbb{C}$. **In-distribution noisy (IDN) samples**, where the true class of the sample belongs to the set $\mathbb{C}$ but the given label is incorrect, i.e., $y_{gt} \in \mathbb{C}$ and $y_i \neq y_{gt}$. **Out-of-distribution (OOD) samples**, where the true class does not belong to the set $\mathbb{C}$ and the given label is also incorrect, i.e., $y_{gt} \notin \mathbb{C}, y_i \in \mathbb{C}$ and $y_i \neq y_{gt}$. While IDN samples can benefit training if relabeled, OOD samples are harmful and should be removed. Our objective is to purify D_{train} by filtering out OOD samples, retaining only clean and IDN samples for robust model training.

Overview. As shown in Figure 2, given the training set D_{train} with IDN and OOD samples, we train two networks with identical structures in parallel following the strategy in (Zheltonozhskii et al. 2022; Li, Socher, and Hoi 2020) to preliminarily recognize the clean samples, which are used to construct vanilla prototypes to serve as compact class representations. To mitigate the ambiguous of category boundaries, we design a Content Diversity-guided Semantic Synchronization (CDgSS) module that leverages LLM-generated descriptions of intra-class commonality and inter-class disparity to extract discriminative text features. Through multimodal synchronization, CDgSS establishes precise semantic correspondences between visual samples and category descriptions, yielding more robust class prototypes. This enables effective separation of clean, IDN and OOD samples via a consistency-based strategy. Furthermore, our Region-Sensitive Distillation Regularization (RSDR) transfers deeper semantic knowledge from MLLMs to guide early-stage DNN training.

Content Diversity-guided Semantic Synchronization

Sample selection methods based on class prototypes are commonly used in learning with noisy labels (Li, Yi, and Huang 2024; Zhang et al. 2024). However, when datasets contain IDN and OOD samples, directly computed prototypes often capture irrelevant features due to inter-class feature ambiguity and intra-class feature dispersion, ultimately degrading model performance. To address this, we enhance prototype quality by leveraging LLM-generated diverse descriptions to explicitly sharpen category boundaries, which suppresses feature ambiguity and filters out noise-induced perturbations, ensuring prototypes concentrate on discriminative characteristics crucial for robust sample selection.

Diversified Content Generation. By leveraging the abundant prior knowledge embedded in LLMs, we can generate detailed category descriptions and emphasize the characteristics that can represent the commonness of each class. To generate discriminative yet generalizable features within each class, effective prompt templates for LLMs are essential. In this work, we design the following prompt templates of ***intra-class commonality descriptions*** for each class:

- *Describe what a [CLASS] looks like in pictures from multiple perspectives in [L] sentences.*

We replace *[CLASS]* with the class name and vary *[L]* to produce a different number of descriptions for each class. By feeding the specified text into GPT-4o without any images,

we can generate a variety of descriptions for each class. The generated description set can be represented as: $T = \{T_c^l\}$, where $c \in \{1, \ldots, C\}$ is the class index, and $l \in \{1, \ldots, L\}$ is the number of descriptions.

While different categories typically have distinct characteristics, noisy samples can blur their decision boundaries. Learning discriminative inter-class discrepancy descriptors helps suppress this ambiguity, though pairwise methods (Esfandiarpoor and Bach 2023) become impractical for large category spaces (e.g., CIFAR-100/WebFG-496). In this work, we develop scalable class-wise disparity descriptions that efficiently capture each class's unique characteristics relative to all others, avoiding exhaustive comparisons while handling both large-scale classification and label noise. Specifically, we leverage the extensive knowledge base of GPT-4o to generate ***inter-class disparity descriptions*** $T^* = \{T_c^*\}$, where $c \in \{1, \ldots, C\}$, guided by the following prompts:

- *Please describe the [CLASS] in the image, highlighting its distinguishing features compared to other categories, and explaining how these features allow it to be differentiated from potentially unseen categories.*

The multimodal large language model CLIP (Radford et al. 2021) demonstrates exceptional performance in text processing by accurately capturing the semantic meaning of natural language, which makes it highly adaptable for a wide range of language-related tasks. In this paper, we extract the text features $E \in \mathbb{R}^{C \times L \times d_t}$ and $E^* \in \mathbb{R}^{C \times d_t}$ of the LLM-generated diversified content T and $T*$, using a frozen text encoder of CLIP in conjunction with a text adapter (Houlsby et al. 2019), where d_t is the dimension of the text features.

Multimodal Semantic Synchronization. In this paper, we follow the common criteria (Zhu et al. 2023) to calculate vanilla prototypes $\mathcal{P}_c$, where the weighted average of visual features is adopted as described in Eq. (1). To suppress erroneous feature representations introduced by IDN and OOD samples to class prototypes, we introduce multimodal semantic synchronization to incorporate diversified textual knowledge into vanilla prototypes to construct refined prototypes with sharper decision boundaries. Specifically, we design two integrators, i.e., the comm-proto integrator and disp-proto integrator, which measure the relationships between visual prototypes and diversified texts to realize multimodal synchronization.

$$\mathcal{P}_c = \frac{1}{|N_c|} \sum_{y_i = c} \mathcal{F}\left(x_i; \theta\right), \tag{1}$$

Here, $|N_c|$ is the number of samples in the c-th class, y_i is the noisy label of sample x_i, and $\mathcal{F}$ is a feature extractor with parameter θ.

The ***comm-proto integrator*** incorporates intra-class commonality text feature E into prototypes $\mathcal{P}_c$ to obtain the refined prototypes $\dot{\mathcal{P}}_c$ that can highlight the semantic commonness of categories, which are represented as:

$$\dot{\mathcal{P}}_c = \mathcal{P}_c + softmax(\mathbf{Q}_1 \mathbf{K}_1^\top)\mathbf{V}_1, c \in \{1, \ldots, C\}, \tag{2}$$

where $\mathbf{Q}_1 = \mathcal{P}_c W_{q1}$, $\mathbf{K}_1 = E_c W_{k1}$, and $\mathbf{V}_1 = E_c W_{v1}$. W_{q1}, W_{k1}, and W_{v1} are learnable matrixes implemented using 1×1 convolution kernels.

The ***disp-proto integrator*** incorporates the inter-class disparity text features into the refined prototypes $\dot{\mathcal{P}}_c$, which further obtains more discriminative class prototypes, as represented in Eq. (3).

$$\ddot{\mathcal{P}}_c = \dot{\mathcal{P}}_c + softmax(\mathbf{Q}_2 \mathbf{K}_2^\top)\mathbf{V}_2, c \in \{1, \ldots, C\}, \tag{3}$$

where $\mathbf{Q}_2 = \dot{\mathcal{P}}_c W_{q2}$, $\mathbf{K}_2 = E_c^* W_{k1}$ and $\mathbf{V}_2 = E_c^* W_{v2}$. W_{q2}, W_{k2}, and W_{v2} are learnable matrixs implemented using 1×1 convolution kernels.

The multimodal semantic synchronization achieves semantic enrichment by synchronizing text-derived intra-class commonality and inter-class disparity features into the visual prototypes. While the two integrators share an identical architecture (Figure 2), they maintain independent parameters. During training, we employ an alternating optimization strategy where one network updates its integrator parameters while the other network uses frozen parameters from the previous epoch's counterpart, thus ensuring stable training without parameter oscillation. As described in Eq. (4), we introduce a contrastive loss function to orchestrate the learning process by facilitating multimodal feature fusion in integrators and strengthening the discriminative capacity of DNNs, which can reduce feature ambiguity while achieving greater separability in the learned decision boundaries.

$$\mathcal{L}_c = -\frac{1}{N} \sum_{i=1}^{N} \log \left(\frac{\exp\left(q_1^i \cdot q_2^i / \tau\right)}{\sum_{j=1}^{N} \exp\left(q_1^i \cdot q_2^j / \tau\right)} \right) \tag{4}$$

where q_1^i denotes the prototype of the i-th category obtained by Net1, while q_2^i is the corresponding class prototype produced by Net2. The temperature parameter is denoted as τ.

Consistency-based Sample Purification

As mentioned above, we obtain the refined prototypes $\left\{\ddot{\mathcal{P}}_1^1, ..., \ddot{\mathcal{P}}_C^1\right\}$ and $\left\{\ddot{\mathcal{P}}_1^2, ..., \ddot{\mathcal{P}}_C^2\right\}$ of two parallel networks. Following this, we calculate the cosine similarity between each sample x_i and the refined prototypes, the class label with the highest similarity score is assigned to the sample, formulated as: $\hat{y}_i = \text{argmax}(sim(f(x_i, \theta), \ddot{\mathcal{P}}_c))$, where $f(\cdot, \theta)$ is the network encoder with paramater θ. The prediction results obtained from two network encoders are respectively denoted as $\hat{y}_i^1$ and $\hat{y}_i^2$.

Fundamentally, in-distribution samples should cluster around their class centroids in feature space, while OOD samples naturally reside in peripheral regions. Motivated by this, we proposed a consistency-based sample purification strategy that first identifies clean samples through triple-consistency verification (requiring agreement between the original label y_i and predicted labels $\hat{y}_i^1$, $\hat{y}_i^2$ from both networks' cosine classifiers). For the the remaining uncertain samples, we introduce a discriminative marginal score $Score = \hat{p}_{hig} - \hat{p}_{sec}$, where $\hat{p}_{hig}$ and $\hat{p}_{sec}$ denote the highest and second-highest predicted values based on cosine similarity, respectively. The marginal score exploits the

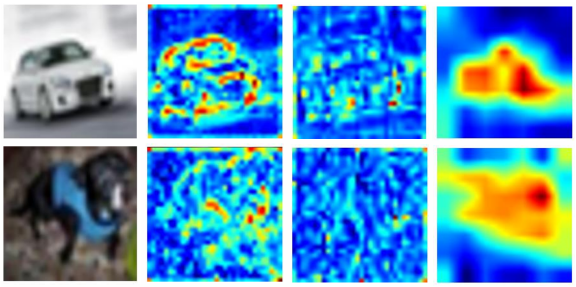


Figure 3: Visualization of the activation regions across different depth layers in CLIP.

observation that OOD samples exhibit near-uniform prediction probabilities due to their large feature-space distances from all class centroids. Subsequently, Gaussian Mixture Model (GMM) is used to model the margin score distribution, and a dynamic threshold γ is obtained, in which samples with scores higher than this threshold are merged into D_{idn}, while samples with scores lower than this threshold are regarded as OOD samples and discarded.

Region-Sensitive Distillation Regularization

While MLLMs like CLIP offer robust representations for noisy label learning, direct feature distillation often underperforms by overlooking local discriminative patterns crucial for classification. To address this challenge, we introduce a region-sensitive distillation regularization, which emphasizes knowledge transfer of critical local regions rather than global features. This strategy adaptively guides the network's attention to semantically important areas during its initial learning phase when representation quality is most vulnerable, thereby bridging the gap between the network's initially uncertain representations and the MLLM's mature feature space, leading to more reliable prototype discrimination and improved prediction accuracy.

Specifically, we treat our early-stage network as the student model and the CLIP as the teacher model. As visualized in Figure 3, the hierarchical attention mechanisms of CLIP exhibit distinct patterns: lower layers capture object boundaries while deeper layers focus on discriminative local features. We therefore select the final layer's attention regions as supervision signals due to their semantic richness. For each sample x_i, the output of the final convolutional layer from CLIP is denoted as M_i^{llm}, and the corresponding output from our network is denoted as M_i^f. Then the spatial alignment is achieved through bilinear interpolation (Zagoruyko and Komodakis 2016), producing normalized attention maps $\ddot{M}_i^{llm}$ and $\ddot{M}_i^f$ of consistent dimensions. Finally, the distillation learning is used as follows:

$$\mathcal{L}_{KD} = \frac{1}{|N_{id}|}\sum_{i=1}^{N_{id}} \frac{\ddot{M}_i^{llm}}{|\ddot{M}_i^{llm}|}\log\left(\frac{\ddot{M}_i^{llm}|\ddot{M}_i^f|}{\ddot{M}_i^f|\ddot{M}_i^{llm}|}\right) \tag{5}$$

$|N_{id}|$ is the number of in-distribution samples including clean and IDN samples $\{D_{clean}, D_{idn}\}$, and $|\cdot|$ is l_1 norm.

Overall Optimization. In this work, we remove OOD samples while retaining clean and IDN samples for training. The IDN samples are treated as unlabeled, and their new pseudo-labels denoted as $\tilde{y}_i$ are generated by averaging the prediction results from the dual networks. Concretely, we apply the mean squared error loss $\mathcal{L}$u for IDN samples and the standard cross-entropy loss $\mathcal{L}$x for clean samples.

$$\mathcal{L}_{\mathrm{u}} = \frac{1}{|N_{idn}|}\sum_{i=1}^{N_{idn}} (p_i - \tilde{y}_i)^2 \tag{6}$$

$$\mathcal{L}_{\mathrm{x}} = -\frac{1}{|N_{clean}|}\sum_{i=1}^{N_{clean}}\sum_{c=1}^{C} y_{i,c}\log p_{i,c} \tag{7}$$

where p_i is prediction of x_i, $|N_{clean}|$ and $|N_{idn}|$ are the numbers of D_{clean} and D_{idn}, respectively.

The overall training objective is represented as follows:

$$\mathcal{L} = \mathcal{L}_x + \lambda_u\mathcal{L}_u + \lambda\mathcal{L}_{KD} \tag{8}$$

where λ is the balancing weight of the designed distillation regularization and is set to 1 according to experience. λ_u is a balancing weight used to constrain IDN samples, which is set following (Li, Socher, and Hoi 2020).

Experiments

Experimental Settings

Datasets. We evaluate our experiment on multiple datasets. CIFAR-10 and CIFAR-100 (50K/10K train/test images) (Krizhevsky 2009) are widely used benchmarks for image classification, with CIFAR-10 containing 10 classes and CIFAR-100 spanning 100 classes. Following (Wu et al. 2021), we construct the OOD samples by injecting different amounts of samples from other datasets into CIFAR-10 and CIFAR-100 to assess the effectiveness of our method. We also conduct experiments on CIFAR-100N and CIFAR-80N, which are variations of the CIFAR-100 dataset. Adhering to the Jo-SRC (Yao et al. 2021) protocol, we create the closed-set CIFAR-100N and open-set CIFAR-80N datasets and introduce varying degrees of symmetric noise into both. Furthermore, we evaluate our framework on WebFG-496 (Sun et al. 2021), which is a fine-grained visual classification benchmark containing 496 categories with high-quality web images for robust model training and evaluation.

Implementation Details. Except for the WebFG-496 dataset, which uses ResNet50, all our experiments employ PreAct ResNet-18 as the base network and use SGD optimizer with momentum 0.9 and weight decay 5×10^{-4}. The CIFAR experiments use batch size set to 128 and epoch set to 360. For experiments conducted using CIFAR-10, we employ a 10-epoch warmup with an initial learning rate of 0.01, which is then decayed to 0.1 after 150 epochs. For experiments conducted using CIFAR-100 (including CIFAR-80N and CIFAR-100N), we use a 30-epoch warmup with the same initial learning rate and decay as CIFAR-10.

Performance Comparison with SOTA Methods

CIFAR-10 and CIFAR-100. We validate our method under combined IDN and OOD noise settings. For both CIFAR-10 and CIFAR-100, we first add 50% symmetric label

Methods	CIFAR-10				CIFAR-100	
	CIFAR-100		TinyImageNet		TinyImageNet	
	10k	20k	10k	20k	10k	20k
RoG (Lee et al. 2019)	63.01	62.56	61.69	63.15	52.65	50.40
ILON (Wang et al. 2018)	75.17	74.85	75.93	74.63	51.43	50.14
DividiMix (Li, Socher, and Hoi 2020)	92.73	92.26	94.08	93.83	70.38	69.89
ProtoMix (Li, Xiong, and Hoi 2021)	93.47	92.17	93.89	93.15	73.92	73.14
NGC (Wu et al. 2021)	93.69	92.31	93.73	93.54	74.57	73.49
LIOND (Zhao and Lee 2024)	-	93.50	-	94.04	-	75.35
Ours	**95.82**	**95.25**	**95.93**	**95.97**	**77.54**	**77.33**

Table 1: The performance comparison on CIFAR-10 and CIFAR-100 with 0.5 sym IDN samples and 10k/20k OOD samples. The accuracy (%) is reported and the best results are masked in **bold**.

Methods	CIFAR-100N		CIFAR-80N	
	sym-20%	sym-80%	sym-20%	sym-80%
JoCoR (Yao et al. 2021)	59.99	12.85	53.01	15.49
DISC (Li et al. 2023)	60.28	33.90	50.33	38.23
MPM (Zhang et al. 2025)	63.46	39.38	67.35	41.41
MoPro (Li, Xiong, and Hoi 2020)	65.60	30.29	54.22	28.32
Jo-SRC (Yao et al. 2021)	65.83	29.76	58.15	23.80
SED (Sheng et al. 2024b)	66.50	38.15	69.10	42.57
PNP (Sun et al. 2022)	67.00	34.36	64.25	31.32
Ours	**72.65**	**48.61**	**72.88**	**44.28**

Table 2: The performance comparison on CIFAR-100N and CIFAR-80N. The best results are masked in **bold**.

noise, then incorporate 10k/20k OOD samples from CIFAR-100/TinyImageNet. As presented in Table 1, we conduct a comprehensive comparison with various SOTA methods. Regarding LIONE's absence of performance metrics in the 10k setting, we note that LIONE neither reports results on 10k CIFAR-10/CIFAR-100 nor provides sufficient implementation details for full reproducibility. Ours consistently outperforms other methods across all experimental setups. This success stems from the content diversity-guided semantic synchronization, which effectively mitigates semantic constraints and class bias while improving sample purification via region-sensitive distillation regularization.

CIFAR-100N and CIFAR-80N. CIFAR-100N contains the same image data as CIFAR-100 but with labels manually annotated by non-expert annotators, introducing errors that better reflect real-world noise conditions compared to the synthetic noise in CIFAR-100. To incorporate OOD samples, CIFAR-80N is introduced, retaining 80 of the classes from CIFAR-100N and designating the remaining 20 classes as OOD samples. Our method outperforms these SOTA methods across various settings, as evidenced by the experimental results in Table 2. Notably, it achieves substantial performance gains over suboptimal methods, demonstrating remarkable robustness when handling challenging scenarios involving mixed IDN and OOD samples. These findings conclusively validate both the effectiveness and reliability of our proposed approach.

WebFG-496. We also test the performance on the real-world dataset WebFG-496, which contains 496 subcategories of three major categories: Web-car, Web-aircraft, and Web-bird. It is widely used to evaluate the robustness of deep learning models under noisy supervision. As demonstrated in Table 3, we compare our method with several SOTA approaches. Our method achieves performance improvements of 0.78%, 0.15%, and 0.8% over sub-optimal methods on Web-car, Web-aircraft, and Web-bird categories respectively. These results validate the effectiveness of our approach for real-world noisy label learning tasks.

Ablation Studies

Effects of Different Components. To further validate the effectiveness of different components, we conduct ablation studies on CIFAR-10 and CIFAR-100 with 0.5 symmetric IDN samples and 10k/20k OOD samples. Notably, we use dividemix as the baseline. Table 4 presents our experimental results, where models (A), (B), and (C) respectively represent the baseline, baseline w/CDgss, and baseline w/CDgss w/RSDR. It reveals that both of our proposed modules positively contribute to the model. The RSDR module effectively transfers prior knowledge from the pre-trained MLLM, while the CDgss enhances the ability to filter out OOD samples, thereby improving the model's generalization performance and accuracy.

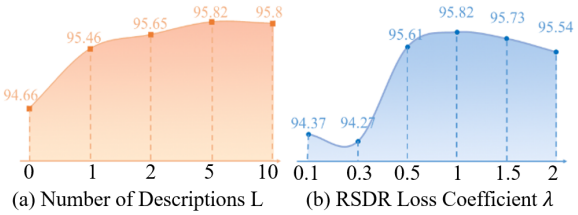
Effects of Hyperparameters. We analyze two key hyperparameters: L (number of LLM-generated class descriptions) and λ (RSDR loss coefficient). As shown in Figure 4(a), we observe that performance improves with increasing L compared to using no diversified descriptions, reaching

Methods	Web-car	Web-aircraft	Web-bird
Peer-learning (Sun et al. 2021)	82.48	78.64	75.37
DivideMix (Li, Socher, and Hoi 2020)	84.27	82.48	74.40
Jo-SRC (Yao et al. 2021)	88.13	82.73	81.22
NPN-Hard (Sheng et al. 2024a)	88.26	86.02	80.91
SED (Sheng et al. 2024b)	88.88	86.62	82.00
SimCore (Kim, Bae, and Yun 2023)	89.53	83.50	66.41
PNP (Sun et al. 2022)	90.11	85.54	81.93
Ours	**90.89**	**86.77**	**82.73**

Table 3: The performance comparison on WebFG-496. The accuracy (%) is reported and the best results are masked in **bold**.

Model	CIFAR-10				CIFAR-100	
	CIFAR-100		TinyImageNet		TinyImageNet	
	10k	20k	10k	20k	10k	20k
(A)	92.73	92.26	94.08	93.83	70.38	69.89
(B)	94.66	94.15	95.32	95.38	77.09	75.85
(C)	**95.82**	**95.25**	**95.93**	**95.97**	**77.54**	**77.33**

Table 4: Ablation studies of different components. The best results are shown in **bold**.



(a) Number of Descriptions L (b) RSDR Loss Coefficient λ

Figure 4: The line charts reflect the influence of the number of generated descriptions L and the loss coefficient λ.

its peak at $L = 5$. Beyond this point, further increases in L yield no additional gains, likely due to the inclusion of redundant semantic information in excessive descriptions. For the RSDR loss coefficient λ, Figure 4(b) reveals that performance initially improves with higher λ values but begins to decline after reaching its optimal point at $\lambda = 1$. Based on these findings, we ultimately set $L = 5$ and $\lambda = 1$.

Effects of Content Generation Methods. We compare four content generation methods: random irrelevant texts, class names only, and class-relevant texts from DeepSeek-V2 and GPT-4o. As shown in Table 5, DeepSeek-V2 and GPT-4o significantly outperform others by generating discriminative texts that capture intra-class commonality and inter-class disparity, while irrelevant texts degrade performance by introducing noise. This demonstrates that the high-quality and semantically coherent content generation is crucial for effective sample selection and class representation learning in our framework.

Decision Boundary Analysis. To evaluate the discriminative capability of our learned representations, we compare our method with state-of-the-art approaches under challenging noise conditions. As shown in Figure 5, while ProtoMix and NGC learn moderately separable features, they struggle to form well-defined decision boundaries, particularly for easily confusable classes. In contrast, our method explicitly enforces larger inter-class margins and tighter intra-class clustering, leading to more distinct class separation. This results in significantly sharper decision boundaries, effectively reducing ambiguity in classification.

Generation Method	Accuracy
Random text	93.86
Class name	95.28
Deepseek-V2	95.65
ChatGPT-4o	**95.82**

Table 5: Ablation studies of different content generation methods. The best results are shown in **bold**.

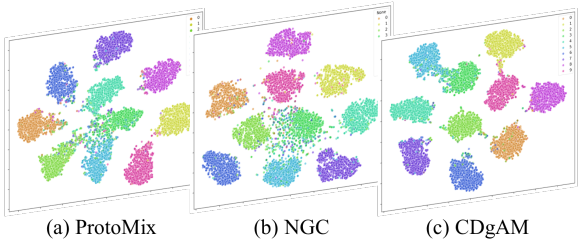

(a) ProtoMix (b) NGC (c) CDgAM

Figure 5: The T-SNE visualization of ProtoMix, NGC and ours on CIFAR-10 with 50% symmetric noise and 10k CIFAR-100 OOD noise.

Conclusion

In this paper, we propose CDgAM, a novel framework that enhances robustness against IDN and OOD samples by mitigating visual ambiguity through content diversity. Our method leverages textual descriptions of intra-class commonality and inter-class disparity to dynamically refine semantic boundaries, reducing prototype learning bias. Additionally, we introduce region-sensitive distillation to transfer knowledge from MLLMs' deep semantics to early-stage classifiers, improving prediction accuracy. Extensive experiments demonstrate our method's effectiveness in boosting model performance and generalization

Acknowledgments

This work is supported by the Qingdao Municipal Natural Science Foundation, China (No. 24-4-4-zrjj-93-jch), the Shandong Provincial Natural Science Foundation of China (No. ZR2024QF044, ZR2025MS998), and the Low-altitude Flight Intelligent Service Support Shandong Engineering Research Center Fund Project (No. KF2024SD007).

References

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Esfandiarpoor, R.; and Bach, S. H. 2023. Follow-Up Differential Descriptions: Language Models Resolve Ambiguities for Image Classification. *arXiv preprint arXiv:2311.07593*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Kim, S.; Bae, S.; and Yun, S.-Y. 2023. Coreset sampling from open-set for fine-grained self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7537–7547.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.

Lee, K.; Yun, S.; Lee, K.; Lee, H.; Li, B.; and Shin, J. 2019. Robust inference via generative classifiers for handling noisy labels. In *International conference on machine learning*, 3763–3772. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*.

Li, J.; Xiong, C.; and Hoi, S. C. 2020. Mopro: Webly supervised learning with momentum prototypes. *arXiv preprint arXiv:2009.07995*.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Learning from noisy data with robust representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9485–9494.

Li, X.; Yi, R.; and Huang, Y. 2024. Mutual Filter Teaching for Open-Set Semi-Supervised Learning. *IEEE Transactions on Multimedia*.

Li, Y.; Han, H.; Shan, S.; and Chen, X. 2023. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24070–24079.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Sheng, M.; Sun, Z.; Cai, Z.; Chen, T.; Zhou, Y.; and Yao, Y. 2024a. Adaptive integration of partial label learning and negative learning for enhanced noisy label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4820–4828.

Sheng, M.; Sun, Z.; Chen, T.; Pang, S.; Wang, Y.; and Yao, Y. 2024b. Foster adaptivity and balance in learning with noisy labels. In *European Conference on Computer Vision*, 217–235. Springer.

Sun, Z.; Hua, X.-S.; Yao, Y.; Wei, X.-S.; Hu, G.; and Zhang, J. 2020. Crssc: salvage reusable samples from noisy data for robust learning. In *Proceedings of the 28th ACM international conference on multimedia*, 92–101.

Sun, Z.; Shen, F.; Huang, D.; Wang, Q.; Shu, X.; Yao, Y.; and Tang, J. 2022. Pnp: Robust learning from noisy labels by probabilistic noise prediction. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5311–5320.

Sun, Z.; Yao, Y.; Wei, X.-S.; Zhang, Y.; Shen, F.; Wu, J.; Zhang, J.; and Shen, H. T. 2021. Webly supervised fine-grained recognition: Benchmark datasets and an approach. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10602–10611.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5552–5560.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wang, Y.; Liu, W.; Ma, X.; Bailey, J.; Zha, H.; Song, L.; and Xia, S.-T. 2018. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8688–8696.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13726–13735.

Wu, Z.-F.; Wei, T.; Jiang, J.; Mao, C.; Tang, M.; and Li, Y.-F. 2021. Ngc: A unified framework for learning with open-world noisy data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 62–71.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations

with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

Yao, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5192–5201.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

Zhang, Q.; Li, X.; Lu, J.; Qiu, L.; Pan, S.; Chen, X.; and Chen, J. 2024. ROG_PL: Robust Open-Set Graph Learning via Region-Based Prototype Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 9350–9358.

Zhang, Y.; Chen, Y.; Fang, C.; Wang, Q.; Wu, J.; and Xin, J. 2025. Learning from open-set noisy labels based on multi-prototype modeling. *Pattern Recognition*, 157: 110902.

Zhao, N.; and Lee, G. H. 2024. Robust Visual Recognition with Class-Imbalanced Open-World Noisy Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16989–16997.

Zheltonozhskii, E.; Baskin, C.; Mendelson, A.; Bronstein, A. M.; and Litany, O. 2022. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1657–1667.

Zhu, R.; Liu, H.; Wu, R.; Lin, M.; Lv, T.; Fan, C.; and Wang, H. 2023. Rethinking Noisy Label Learning in Real-world Annotation Scenarios from the Noise-type Perspective. *arXiv preprint arXiv:2307.16889*.