Mask-guided Semantic Alignment: Robust Learning with Noisy Labels via Temporal Attention Stability

Yubo Nian[1] and Can Gao[1,2⋆]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2] Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen, China
2410105011@mails.szu.edu.cn, 2005gaocan@163.com

Abstract. Noisy labels pose a significant challenge in training deep neural networks, as corrupted annotations inevitably induce overfitting and significantly degrade generalization. However, existing methods typically rely on static metrics and overlook training dynamics, struggling to effectively separate clean samples from noisy ones under complex and high noise scenarios. In this paper, we propose a temporal attention-based framework for learning with noisy labels. Specifically, we introduce an attention temporal consistency module for reliable sample selection, which separates samples by quantifying the evolutionary stability of visual attention during training. Then, to address the scarcity of clean samples in high-noise scenarios, we design an adaptive attention masks module to synthesize images from clean and noisy samples, which employs differential masking strategies to generate high-quality training samples. Finally, we present a mask-guided feature alignment module, which introduces feature consistency regularization to facilitate the learning of robust representations for noisy data. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches, achieving accuracy gains of 1.6%/2.8% on CIFAR-10/100 under 90% symmetric noise, respectively. Code is available at https://github.com/iCAN-SZU/MSA

Keywords: Noisy labels · attention mask · sample selection · feature alignment

1 Introduction

The availability of large-scale datasets with high-quality annotations plays a critical role in enabling deep neural networks (DNNs) to achieve remarkable performance. However, collecting such accurately labeled data often presents dual challenges of high cost and time-intensive efforts. Despite cost-effectiveness advantages, crowdsourcing [14] inevitably introduces noisy labels due to annotator expertise variation. Pretrained models [10] and web-crawling [23] enhance

⋆ Corresponding author.

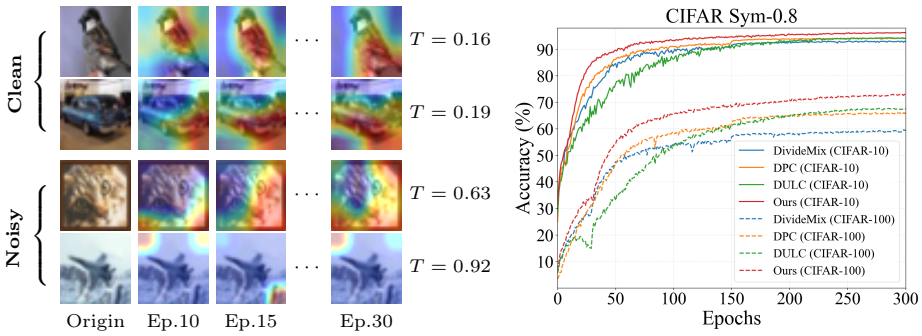

Fig. 1: Motivation and performance comparison. Left: Visualization of CAM evolution during training. Clean samples exhibit stable attention patterns with low temporal scores (T), whereas noisy samples show erratic attention shifts with high scores. Right: Comparison of test accuracy between our proposed method and other sample selection methods, demonstrating superior performance and stability.

efficiency but generate low-quality noisy samples from unreliable sources, which seriously affects the generalization of DNNs [1].

Learning with Noisy Labels (LNL) approaches have been proposed to mitigate the impact of corrupted annotations through regularization, sample selection, and label correction [36]. Regularization strategies apply explicit loss constraints [12,42,55] or implicit stochastic effects [21,41] to improve robustness but often suffer from generalization decline and spurious pattern overfitting under severe noise. Sample selection identifies reliable clean data to avoid noise interference [6,24,47] yet remains limited in distinguishing hard samples due to a heavy reliance on early-stage metrics. Label correction rectifies annotations through self-generated pseudo-labels [21,39,51] or auxiliary procedures [11,45,52]. However, incorrectly generated pseudo-labels can cause the model to overfit noisy patterns and further degrade performance.

DNNs exhibit a distinct memorization effect [1] where they tend to prioritize learning simple and generalized patterns inherent in clean data during the early training stages and then gradually overfit complex noise. This disparity implies that the historical training dynamics contain crucial information for distinguishing between correctly labeled and mislabeled samples, instead of relying solely on the final loss snapshot. To investigate these dynamics at a fine-grained level, we examine the micro-learning process through the lens of visual interpretability. As illustrated in Figure 1, we observe distinct disparities in the evolutionary trajectories of Class Activation Maps (CAMs) [53] between clean and noisy samples. Specifically, the trend in most cases is that the attention of correctly labeled samples converges rapidly and anchors firmly onto discriminative semantic objects. In contrast, mislabeled samples exhibit erratic visual behaviors where their activation regions fail to establish a stable focus and tend to drift aimlessly toward background noise or irrelevant peripheral areas as training progresses.

Motivated by this observation, we propose a temporal attention-based framework for learning with noisy labels. To overcome the limitations of conventional loss-based sample identification, we introduce an Attention Temporal Consistency (ATC) module for sample selection. This module accurately filters out hard noise by leveraging the temporal stability of visual attention, predicted class margin, and prediction loss. Building on this precise sample division, we design an Adaptive Attention Masks (AAM) module for image generation, which generates high-quality training samples by employing differential masking and background substitution, thereby enabling the network to localize authentic foreground objects and mitigate background interference. Finally, to obtain more robust feature representations, we introduce the Mask-guided Feature Alignment (MFA) module for robust representation learning, which aligns original views with their masked counterparts via consistency regularization, thereby compelling the network to no longer rely on localized features, ultimately facilitating robust representation learning for noisy data. To sum up, the main contributions of this study are as follows:

1) To accurately separate clean and noisy samples, an Attention Temporal Consistency (ATC) module is introduced to capture the evolutionary stability of visual attention, which enables the identification of hard noisy samples by integrating with static sample evaluation metrics.

2) To effectively exploit clean and noisy samples, an Adaptive Attention Masks (AAM) module is introduced to synthesize high-quality training samples by employing differential masking and background substitution, which enables the model to focus on authentic foreground objects and mitigate background interference.

3) To obtain more robust feature representations, a Mask-guided Feature Alignment (MFA) module is designed to align original views with information-restricted counterparts via consistency regularization , which compels the network to transcend local features and learn robust representations.

4) Extensive experiments are conducted on both synthetic and real-world datasets to verify the effectiveness of the proposed framework, with average accuracy gains of 1.6% on CIFAR and 1.4% on the Food-101N dataset.

2 Related Works

The existing LNL methods can be roughly divided into three categories: regularization, label correction, and sample selection.

Regularization. Deep neural networks tend to overfit corrupted data under cross-entropy loss [48], prompting the development of various robust regularization strategies [36]. Explicit regularization methods try to modify their objectives or add constraints, such as distinct parameter update strategies in REL [42], early regularization in ELR [26], geometric structure refinement in RRL [22], gradient rescaling via attention in NAL [27] and feature consistency in SSR+ [12]. Implicit methods utilize stochastic regularization effects to improve generalization. For instance, label smoothing [17, 28] prevents over-confidence, adversarial

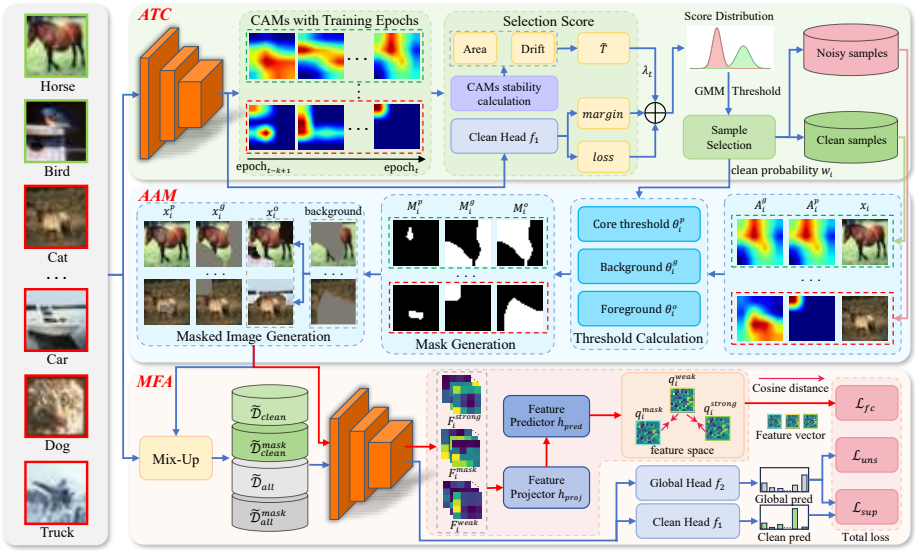

Fig. 2: The framework comprises three components: Attention Temporal Consistency (ATC), Adaptive Attention Masks (AAM), and Mask-Guided Feature Alignment (MFA). ATC partitions the dataset into clean and noisy subsets by tracking the temporal stability of CAM via a GMM. AAM generates differential masks based on sample selection results to erase background clutter and rectify feature representation. MFA aligns semantic representations between original and masked views using dual classification heads and a consistency loss.

erasing [41] masks high-activation regions, and mixup-based paradigms [7,21,49] enable the fusion of samples to alleviate the effects of noisy labels.

Label Correction. Label correction aims to identify true labels through various refinement strategies. This includes using self-predictions as pseudo-labels [21, 39, 46, 51], training meta-learning purifiers [29, 40, 52, 54], or leveraging multi-view consistency [24]. Other approaches involve shifted Gaussian models [9], cross-view semantic associations [11], and interpolation between noisy labels and model predictions [45] to generate soft targets that weaken the negative effects of label corruption.

Sample Selection. Sample selection identifies clean samples and isolates noise during training. Initial techniques primarily leveraged peer networks and small-loss criteria [15,47]. TheDivideMix [21] and its successors [2,7] employ Gaussian Mixture Models (GMM) combined with semi-supervised learning [3] to improve sample partition. Recent studies enhance sample identification by incorporating memory strength momentum [24], two-stage filtering processes [5], and feature-based discrimination [6,12,16] to achieve more robust results.

3 Proposed Method

Overview. LNL aims to deal with the problem of multi-class classification under label noise. Let $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$ denote the noisy training dataset containing N samples, where $x_i \in \mathcal{X}$ represents the input image and $\tilde{y}_i \in \{0,1\}^C$ is the provided noisy label, while the ground-truth label y_i is unavailable. To address this problem, we propose a novel model to learn with noisy labels, and its overall framework is illustrated in Figure 2. First, the Attention Temporal Consistency (ATC) module partitions samples into clean and noisy subsets by analyzing the evolutionary stability of their activation maps. Subsequently, the Adaptive Attention Masks (AAM) module applies differential masking strategies to suppress background noise for clean data and erase erroneously activated regions for noisy data to rectify their representations. Finally, the Mask-guided Feature Alignment (MFA) module enhances representation learning by aligning the semantic features across the original, augmented, and masked views.

3.1 Sample Selection with Attention Temporal Consistency

A critical challenge in learning with noisy labels is the memorization effect [1], where deep networks eventually overfit to mislabeled data, exhibiting high confidence similar to clean samples. Consequently, relying solely on static metrics (e.g., small-loss criteria) often fails to accurately distinguish clean samples from noisy ones, particularly in high-noise scenarios. However, we observe a distinct behavioral divergence in visual attention dynamics: clean samples maintain stable activation regions, whereas noisy samples exhibit erratic attention shifts due to incorrect semantic guidance. Capitalizing on this disparity, we propose an Attention Temporal Consistency (ATC) module for sample selection, which quantifies the temporal stability of CAMs to identify reliable clean samples.

Formally, for each input sample x_i, the backbone extractor $g(\cdot)$ generates a feature map F_i and further obtains the feature representation $\boldsymbol{v}_i$. Subsequently, the classifier head $f(\cdot)$ maps $\boldsymbol{v}_i$ to logits $\boldsymbol{z}_i = f(\boldsymbol{v}_i)$, yielding the class probabilities $\hat{p}_i = \mathrm{Softmax}(\boldsymbol{z}_i)$. Meanwhile, the F_i is utilized to generate two activation maps $A_i^g, A_i^p \in \mathbb{R}^{H \times W}$, conditioned on the provided noisy label $\widetilde{y}_i$ and the model's predicted class $\hat{y}_i$:

$$\begin{cases} A_i^g = \mathrm{CAM}(F_i, \tilde{y}_i) \\ A_i^p = \mathrm{CAM}(F_i, \hat{y}_i) \end{cases} \tag{1}$$

where $\hat{y}_i = \arg\max(\hat{p}_i)$ denotes the predicted class with the highest confidence, and $\mathrm{CAM}(\cdot,\cdot)$ means the function that extracts the activation map from the given feature map with the guidance of the class label.

To capture the evolutionary trajectory, the CAM $A_i^{g,(t)}$ for sample x_i is extracted at the t-th training epoch. Attention is focused on two geometric properties: the core area size and the centroid location. First, the core region $R_i^{(t)}$ is defined as the number of pixels within the activation map whose intensity

exceeds a specific threshold τ_r. The concentration of the model's attention is thereby quantified as follows:

$$R_i^{(t)} = \left| \left\{ (h,w) \mid A_i^{g,(t)}(h,w) > \tau_r \right\} \right|, 0 \le h \le H, 0 \le w \le W, \tag{2}$$

where $|\cdot|$ denotes the cardinality of the set.

Then, to quantify the spatial stability, the centroid $\mathbf{C}_i^{(t)}$ of the activation map $A_i^{g,(t)}$ is defined as the intensity-weighted mean of pixel coordinates, yielding a two-dimensional vector:

$$\mathbf{C}_i^{(t)} = \begin{bmatrix} c_H^{(t)} \\ c_W^{(t)} \end{bmatrix} = \frac{\sum_{h=1}^{H} \sum_{w=1}^{W} A_i^{g,(t)}(h,w) \cdot \begin{bmatrix} h \\ w \end{bmatrix}}{\sum_{h=1}^{H} \sum_{w=1}^{W} A_i^{g,(t)}(h,w)}. \tag{3}$$

This spatial descriptor enables the tracking of the structural shift in attention across epochs. The attention displacement between consecutive epochs can be defined as the Euclidean distance between centroids:

$$D_i^{(t)} = \left\| \mathbf{C}_i^{(t)} - \mathbf{C}_i^{(t-1)} \right\|_2. \tag{4}$$

To evaluate stability within a temporal window of k epochs, the moving averages of the core area and the centroid displacement are computed by:

$$\begin{cases} \overline{R}_i = \frac{1}{k} \sum_{j=t-k+1}^{t} R_i^{(j)} \\ \overline{D}_i = \frac{1}{k} \sum_{j=t-k+1}^{t} D_i^{(j)} \end{cases}. \tag{5}$$

Furthermore, let $\mathrm{Var}(R_i)$ and $\mathrm{Var}(D_i)$ denote the variances of the core area and displacement over the interval $[t-k+1, t]$. The temporal consistency score T_i is defined by combining the relative fluctuation of the core area with the spatial instability of the centroid:

$$T_i = \frac{\mathrm{Var}(R_i)}{\overline{R}_i + 1} + \mathrm{Var}(D_i) + \overline{D}_i. \tag{6}$$

This consistency score can be easily integrated into the existing sample selection framework. To accurately distinguish clean samples from noisy ones, the reliability of samples is evaluated from both static and dynamic perspectives. First, the static discriminative confidence of the model is quantified through two complementary metrics: the predictive margin m_i and the cross-entropy loss ℓ_i. The margin m_i measures the decision gap between the given label $\widetilde{y}_i$ and the most probable negative class, explicitly reflecting the classifier's certainty:

$$m_i = z_i^{\widetilde{y}_i} - \max_{j \neq \widetilde{y}_i} z_i^j. \tag{7}$$

Meanwhile, the cross-entropy term ℓ_i evaluates the overall fitting degree by quantifying the divergence between the predicted distribution and the label distribution. Clean samples typically exhibit larger margins and smaller losses. However, relying solely on these static indicators is insufficient, as deep networks can

memorize noisy labels, resulting in high confidence for mislabeled data. To address this, the temporal consistency score $\widetilde{T}_i$ is introduced as a dynamic penalty to filter out samples that are confidently predicted but lack semantic stability. The final sample selection score is derived by balancing the static confidence and the temporal consistency score:

$$\text{Score}_i = \underbrace{m_i - \ell_i}_{\text{Static Confidence}} - \underbrace{\lambda_t \cdot \widetilde{T}_i}_{\text{Dynamic Penalty}}, \tag{8}$$

where $\widetilde{T}_i = T_i \cdot \frac{10}{C}$ is the normalized consistency term scaled by the number of classes C, and λ_t is a dynamic weighting factor that gradually increases during training. By fusing this information, our metric can effectively identify hard noisy samples that resemble clean data under static predictions. Finally, following [21], a two-component Gaussian Mixture Model (GMM) is fitted to partition all samples into a labeled clean set $\mathcal{D}_{clean}$ and a noisy set $\mathcal{D}_{noise}$, by utilizing the posterior probability threshold τ_g. While the output probability w_i reflects the likelihood of x_i being marked as a clean sample.

3.2 Image Synthesis with Adaptive Attention Masks

In high-noise scenarios, the number of exploitable clean samples is limited. Standard data augmentation techniques [8, 30, 41] typically apply uniform strategies to all samples, neglecting the intrinsic semantic disparity between correctly labeled and mislabeled samples. Factually, for noisy samples, high activations typically correspond to misleading regions associated with the incorrect label. For clean samples, only masking high-activation regions may cause the model to focus on irrelevant backgrounds. Simultaneously masking activation regions and suppressing background interference is essential to fully capitalize on clean samples. Based on this fact, we propose the Adaptive Attention Masks (AAM) module for image synthesis, which generates high-quality samples by employing differential masking and background substitution, enabling the model to focus on real foreground objects and avoid the background interference.

Specifically, the AAM module is performed through three integrated stages: adaptive thresholding, differential mask generation, and semantic-preserving background mixing. To precisely calibrate the masking intensity, global reference thresholds are first computed based on the statistical mean of the activation maps. Let A_i^g and A_i^p denote the CAMs derived from the given label and the predicted label, respectively. The global thresholds are defined as:

$$\begin{cases} \theta_i^g = \text{mean}(A_i^g), \\ \theta_i^o = \text{mean}(A_i^p). \end{cases} \tag{9}$$

Furthermore, to dynamically adjust the erasure range, a prediction-guided threshold θ_i^p is introduced by incorporating the clean probability w_i of the sample. This threshold is adaptively defined with the sample's reliability:

$$\theta_i^p = \begin{cases} \max(A_i^p) \cdot \mu \cdot w_i \cdot \xi_i, & i \in \mathcal{D}_{clean} \\ \max(A_i^p) \cdot \mu \cdot (1 - w_i) \cdot \xi_i, & i \in \mathcal{D}_{noise} \end{cases} \tag{10}$$

where μ controls the base intensity of masking, and $\xi_i \sim U(\xi, 1)$ is a stochastic perturbation factor introduced to enhance robustness against overfitting.

Considering the difference in attention activation between clean and noisy samples, an adaptive binary mask M_i^g is generated. Specifically, for clean data, to suppress background clutter and compel the model to concentrate on the semantic object, regions with activation values below θ_i^g are masked. In contrast, for noisy data, to eradicate erroneous activations caused by label corruption, regions with activation values exceeding θ_i^g are masked. Formally, the mask M_i^g at coordinate (h, w) is determined by:

$$M_i^g(h,w) = \begin{cases} 1, & \text{if } i \in \mathcal{D}_{clean} \text{ and } A_i^g(h,w) < \theta_i^g, \\ 1, & \text{if } i \in \mathcal{D}_{noise} \text{ and } A_i^g(h,w) > \theta_i^g, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

The label-guided masked image x_i^g is then obtained by applying this mask: $x_i^g = x_i \odot (1 - M_i^g)$, where $\odot$ denotes element-wise multiplication.

To further prevent the model from overfitting to the most discriminative local regions (e.g., the head of an animal), an adversarial erasing strategy is employed based on the model's own predictions, which introduces a prediction-guided mask M_i^p to suppress peak activation regions:

$$M_i^p(h,w) = \begin{cases} 1, & \text{if } A_i^p(h,w) > \theta_i^p, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

and the prediction-guided masked image is $x_i^p = x_i \odot (1 - M_i^p)$.

Finally, to decouple the semantic object from its background context, a cross-sample background mixing strategy is used, and a foreground object mask M_i^o is defined using the separation threshold θ_i^o:

$$M_i^o(h,w) = \begin{cases} 1, & \text{if } A_i^p(h,w) > \theta_i^o, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Then, a new sample x_i^o is generated by preserving the foreground of the current sample x_i and filling the background region with a randomly permuted sample x_j:

$$x_i^o(h,w) = \begin{cases} x_i(h,w), & \text{if } M_i^o(h,w) = 1, \\ x_j(h,w), & \text{otherwise.} \end{cases} \tag{14}$$

This forces the model to recognize the semantic objects invariant to background changes.

3.3 Robust Representation Learning with Mask-guided Feature Alignment

Existing consistency regularization typically relies on random data augmentation, such as random cropping and color jittering, to learn robust representation [4]. However, random augmentations often fail to perturb key discriminative

regions, thereby resulting in overfitting to local irrelevantly noisy features. To address this, we introduce the Mask-guided Feature Alignment (MFA) module to learn robust feature representation, which aligns original views with masked counterparts via consistency regularization, thereby compelling the network to abandon its reliance on local information and instead explore complementary contextual semantics, and ultimately facilitating robust representation learning for noisy data.

Specifically, for a given training sample x_i, its weakly augmented view x_i^w is considered as the reliable *semantic anchor*, and the projection of this anchor is denoted as $\boldsymbol{q}_i^{weak} = h_{proj}(g(x_i^w))$, where h_{proj} means the projection head. To enforce consistency, the model is required to predict this anchor from two distorted views: the strongly augmented view x_i^s and the prediction-guided masked view x_i^p. The predicted embeddings are formulated as $\boldsymbol{q}_i^{mask} = h_{pred}(h_{proj}(g(x_i^p)))$ and $\boldsymbol{q}_i^{strong} = h_{pred}(h_{proj}(g(x_i^s)))$, where h_{pred} denotes the prediction head. To maximize the similarity between the anchor and the distorted predictions, the following consistency loss $\mathcal{L}_{fc}$ is formulated:

$$\mathcal{L}_{fc} = \underbrace{\mathcal{S}\left(\boldsymbol{q}_i^{mask}, \boldsymbol{q}_i^{weak}\right)}_{\text{Mask Alignment}} + \underbrace{\mathcal{S}\left(\boldsymbol{q}_i^{strong}, \boldsymbol{q}_i^{weak}\right)}_{\text{Augmentation Consistency}}, \tag{15}$$

where $\mathcal{S}(\cdot,\cdot)$ is the negative cosine similarity and is defined as $\mathcal{S}(u,v) = -\frac{u^\top v}{\|u\|_2\|v\|_2}$. The *Mask Alignment* term promotes the model to recover complete semantic information from masked inputs, preventing excessive focus on local noise. However, masking primarily addresses internal semantic completeness and does not account for variations in geometry and illumination. Therefore, the *Augmentation Consistency* term ensures the representation maintains invariance to severe global transformations. Integrating these objectives captures both the integral structure of the object and its invariance to environmental changes.

To fully utilize all available data, four sample sets are generated via MixUp augmentation [49]: the clean mixed set $\widetilde{\mathcal{D}}_{clean}$ and the global mixed set $\widetilde{\mathcal{D}}_{all}$, alongside their masked counterparts $\widetilde{\mathcal{D}}_{clean}^{mask}$ and $\widetilde{\mathcal{D}}_{all}^{mask}$. Subsequently, these augmented sets are used to optimize the feature extractor in a semi-supervised learning manner, followed by two distinct classification heads: a clean head f_1 and a global head f_2, both sharing the same backbone $g(\cdot)$. Since standard cross-entropy forces probability normalization, compelling the model to over-confidently predict heavily masked or noisy inputs. While Evidential Deep Learning (EDL) [32, 55] explicitly decouples confidence from evidence by treating logits as non-negative evidence counts. This allows the model to accumulate evidence only for reliable semantics while suppressing ambiguous inputs, naturally maintaining a separable logit distribution that ensures the reliability of our ATC selection module.

Specifically, class prediction probabilities are represented as a Dirichlet distribution $Dir(\boldsymbol{p}_i|\boldsymbol{\alpha}_i)$, where $\boldsymbol{p}_i$ is the class probability vector and $\boldsymbol{\alpha}_i$ are derived from the logits to represent the collected evidence. The evidential loss $\mathcal{L}_e$ is formulated as a combination of a negative log-likelihood term $\mathcal{L}_{nll}$ and a KL-

divergence term $\mathcal{L}_{kl}$:

$$\mathcal{L}_e(\mathcal{D}, f, \widetilde{\boldsymbol{y}}) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \left[\mathcal{L}_{nll}(\boldsymbol{\alpha}_i, \widetilde{\boldsymbol{y}}_i) + \beta \mathcal{L}_{kl}(\boldsymbol{\alpha}_i, \widetilde{\boldsymbol{y}}_i)\right], \tag{16}$$

where β is a balancing coefficient and α_i are derived from the network logits z_i such that the evidence is $e_i = \exp(z_i)$, which yields $\alpha_i = \frac{C}{10} e_i + 1$.

The total supervised loss $\mathcal{L}_{sup}$ aggregates the evidential losses from both heads across original and masked views:

$$\begin{aligned} \mathcal{L}_{sup} = \mathcal{L}_e(\widetilde{\mathcal{D}}_{clean}, f_1, \hat{\boldsymbol{y}}) + \mathcal{L}_e(\widetilde{\mathcal{D}}_{all}, f_2, \hat{\boldsymbol{y}}) \\ + \mathcal{L}_e(\widetilde{\mathcal{D}}^{mask}_{clean}, f_1, \hat{\boldsymbol{y}}) + \mathcal{L}_e(\widetilde{\mathcal{D}}^{mask}_{all}, f_2, \hat{\boldsymbol{y}}). \end{aligned} \tag{17}$$

Simultaneously, an unsupervised loss $\mathcal{L}_{uns}$ is further introduced to enforce prediction stability, with a Mean Squared Error (MSE) regularization term:

$$\mathcal{L}_{uns} = \eta \left[\mathcal{L}_e(\widetilde{\mathcal{D}}_{all}, f_2, \hat{\boldsymbol{y}}) + \mathcal{L}_e(\widetilde{\mathcal{D}}^{mask}_{all}, f_2, \hat{\boldsymbol{y}})\right] + \lambda_u \mathcal{L}_{mse}, \tag{18}$$

where $\hat{\boldsymbol{y}}$ means the pseudo-labels generated by the clean head f_1, η represents the noise ratio recognized by the model, λ_u controls the weight of the MSE loss term, which is defined as:

$$\mathcal{L}_{mse} = \frac{1}{|\widetilde{\mathcal{D}}_{all}|} \sum_{x_i \in \widetilde{\mathcal{D}}_{all}} \|\hat{\boldsymbol{y}}_i - \boldsymbol{p}(f_2(g(x_i)))\|_2^2 + \frac{1}{|\widetilde{\mathcal{D}}^{mask}_{all}|} \sum_{x_j \in \widetilde{\mathcal{D}}^{mask}_{all}} \left\|\hat{\boldsymbol{y}}_j - \boldsymbol{p}(f_2(g(x_j)))\right\|_2^2, \tag{19}$$

where $\boldsymbol{p}(\cdot)$ denotes the predicted class probability distribution after softmax normalization.

3.4 Training and Inference

During training, the feature extractor g, the classification heads f_1 and f_2, along with the projection head h_{proj} and prediction head h_{pred}, are jointly trained in an end-to-end manner. They are optimized by minimizing the total loss $\mathcal{L}_{total}$, which integrates the supervised, unsupervised, and feature alignment losses:

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \mathcal{L}_{uns} + \mathcal{L}_{fc}. \tag{20}$$

During inference, the final prediction for an unseen test image x is generated by using only the shared backbone and the clean head f_1, formulated as $\hat{y} = \arg\max f_1(g(x))$.

4 Experiments

4.1 Experiments Setting

Datasets. We evaluate our method on CIFAR-10/100 [18], Clothing1M [44], ANIMAL-10N [35] and Food-101N [20]. CIFAR-10/100 consists of 50k training

samples and 10k testing samples. Clothing-1M is a large-scale and real-world noisy dataset containing 1 million training samples crawled from online shopping websites, where their labels are nearly 40% mislabeled. ANIMAL-10N consists of human-annotated images from 10 animal categories, with nearly 8% mislabeled samples. The Food-101N contains about 31k food recipes images from 101 categories, and its estimated label noise rate is about 20%. To align with previous work [21], experiments on CIFAR-10/100 are conducted with symmetric noise rates of {20%,50%,80%,90%} and an asymmetric noise rate of 40%. Moreover, instance-dependent noise (IDN) at a 40% rate [43] is also considered.

Implementation Details. For CIFAR experiments, we use 18-layer PreAct ResNet [13] as the backbone. We train the networks using SGD optimizer from scratch for 300 epochs with momentum 0.9, batch size 128, initial learning rate 0.03, and weight decay 5e-4. The warmup phase lasts 10 epochs for CIFAR-10 and 30 epochs for CIFAR-100 with symmetric and asymmetric noise. The initial learning rate is 0.03 and is controlled by a cosine annealing scheduler. For Clothing1M and Food-101N, we follow the previous work [20, 21] and use ResNet-50 [13] with ImageNet [19] pretrained weights as the feature extractor for fair comparisons. The learning rate and the weight decay are set to 0.001 and 5e-4, respectively. For ANIMAL-10N, we train the backbone VGG-19 [34] for 150 epochs using SGD optimizer following [35]. More detailed implementations can be found in the *supplementary material*.

4.2 Experimental Results

We compare our method with multiple methods, including Co-teaching+ [47], DivideMix [21], AugDesc [30], LongReMix [7], DISC [24], SANM [41], HMW [50], CCLRL [11], DULC [45], Joint [39], PENCIL [46], ELR+ [26], RRL [22], CleanNet [20], Co-learning [38], PNP [37], SURE [25], CT [31], CA2C [33], SELFIE [35], PLC [51], SSR+ [12], and the naive cross-entropy (CE) method. For fair comparisons, we follow the standard LNL evaluation protocols [20,21,35] and adopt the same backbone as in previous works for all benchmarks.

Results on synthetic noisy datasets. Table 1 reports the test accuracy on CIFAR-10 and CIFAR-100. As shown, our approach consistently outperforms the best baseline by 0.5% - 3.0% across 20% to 90% symmetric noise levels. Notably, in the extreme 90% noise scenario, most methods like DivideMix and DISC struggle to maintain performance, while our method achieves a remarkable 95.1%, surpassing the second-best model DULC by 1.6%. On the more challenging CIFAR-100 dataset, our method reaches 55.6% accuracy under 90% noise, providing a substantial 2.8% improvement over DULC, which underscores our framework's capacity to extract reliable information from sparse clean data. Furthermore, under the IDN setting, which closely mimics real-world label noise, our method consistently outperforms all baselines on both datasets, verifying its practical effectiveness.

Results on real-world noisy datasets. Table 2 shows the results of different methods on Clothing-1M and Food-101N datasets. It can be seen that our method achieves 75.24% accuracy on Clothing-1M, surpassing the previous

Table 1: Performance comparison with other methods on CIFAR-10 and CIFAR-100 with synthetic noise. The best and second-best scores are **boldfaced** and underlined, respectively, and "-" indicates that results are not available in their papers.

Dataset	CIFAR-10						CIFAR-100					
Noise type	sym.				asym.	idn	sym.				asym.	idn
Method/Noise ratio	20%	50%	80%	90%	40%	40%	20%	50%	80%	90%	40%	40%
CE	84.5	71.6	48.4	40.9	82.5	67.6	53.1	33.3	9.6	7.3	33.2	43.2
Co-teaching+ [47]	84.6	80.2	55.2	32.0	83.8	73.8	54.5	46.5	18.4	10.3	41.2	24.5
DivideMix [21]	96.1	94.6	93.2	76.0	93.4	95.1	77.2	74.6	60.2	31.5	51.0	76.0
AugDesc [30]	96.3	95.4	93.8	91.9	94.6	-	79.5	77.2	66.4	41.2	63.5	-
LongReMix [7]	96.2	95.0	93.9	82.0	94.7	-	77.8	75.6	62.9	33.8	62.2	-
DISC [24]	96.3	95.4	92.9	57.6	93.4	95.9	78.6	76.3	59.3	24.2	76.5	78.4
SANM [41]	96.4	95.8	94.6	92.3	94.7	-	81.2	78.2	68.7	43.5	76.7	-
HMW [50]	93.5	95.2	93.7	90.7	93.7	-	76.6	75.8	63.4	43.4	72.1	-
CCLRL [11]	97.0	96.5	94.6	-	96.1	96.2	79.5	77.4	70.3	-	77.2	80.0
DULC [45]	96.6	96.0	95.0	93.5	95.2	-	79.4	76.4	67.7	52.8	75.8	-
Ours	**97.5**	**97.0**	**96.4**	**95.1**	**96.7**	**97.1**	**82.9**	**80.4**	**73.3**	**55.6**	**80.1**	**81.0**

Table 2: Performance comparison with other methods on Clothing-1M and Food-101N, where the marker "*" denotes results obtained by reproducing the method.

Clothing-1M		Food-101N	
Methods	Acc.	Methods	Acc.
CE	69.21	CE	81.7
Joint optimization [39]	72.16	CleanNet [20]	83.5
PENCIL [46]	73.49	Co-learning [38]	87.6
LongReMix [7]	74.38	PNP-hard [37]	87.3
DivideMix [21]	74.76	PNP-soft [37]	87.5
DISC [24]	74.79	DivideMix [21]	86.5
SANM* [41]	74.53	DISC [24]	87.3
ELR+ [26]	74.81	SURE [25]	88.0
RRL [22]	74.90	DISC+CT [31]	87.5
DULC [45]	75.09	CA2C [33]	86.8
Ours	**75.24**	Ours	**89.4**

SOTA methods and verifying that our approach is effective in large-scale real-world noise scenarios. Similarly, on Food-101N, our method obtains a superior accuracy of 89.4%. Moreover, as shown in Table 3, our method reaches 90.8% accuracy on Animal-10N, which brings a 1.1% improvement over the best baseline CCLRL, indicating its effectiveness in handling fine-grained real-world noisy datasets. More experimental results are provided in the *supplementary material*.

Table 3: Performance comparison with other methods on Animal-10N.

Method	CE	SELFIE [35]	PLC [51]	SSR+ [12]	DISC [24]	SANM [41]	CCLRL [11]	**Ours**
Accuracy	79.4	81.8	83.4	88.5	87.1	89.3	<u>89.7</u>	**90.8**

Table 4: Ablation study of different components on CIFAR-10/CIFAR-100.

Component			CIFAR-10 / CIFAR-100			
ATC	AAM	MFA	sym.20%	sym.50%	sym.80%	sym.90%
×	×	×	97.1 / 80.7	96.4 / 78.5	95.3 / 70.0	86.8 / 44.0
✓	×	×	97.1 / 81.7	96.4 / 79.5	95.6 / 70.5	94.6 / 49.6
✓	✓	×	97.4 / 82.1	96.7 / 80.0	96.1 / 72.3	94.8 / 53.3
✓	✓	✓	**97.5** / **82.9**	**97.0** / **80.4**	**96.4** / **73.3**	**95.1** / **55.6**

4.3 Ablation Study

To evaluate the contribution of each core component, we conduct ablation studies on CIFAR-10 and CIFAR-100 datasets under varying symmetric noise levels. As demonstrated in Table 4, the baseline model utilizing only the fundamental training framework without the three proposed modules exhibits a dramatic performance decline as noise intensity increases, particularly on the complex CIFAR-100 dataset, where the accuracy drops to 44.0% under 90% noise.

The integration of ATC serves as a pivotal foundation, yielding substantial gains in extreme noise scenarios. Specifically, it elevates the test accuracy from 86.8% to 94.6% on CIFAR-10 and from 44.0% to 49.6% on CIFAR-100 at the 90% noise level, validating the effectiveness of leveraging temporal activation stability for precise noise filtering. Building upon this, the introduction of AAM further refines representation learning by implementing differential masking and background substitution to generate high-quality training samples, while decoupling semantic objects from environmental context or misleading activations. This leads to consistent improvements; for instance, on CIFAR-100 with 80% noise, the accuracy increases from 70.5% to 72.3%. Finally, the inclusion of MFA enforces high-level semantic alignment between original and masked views, steering the model to focus on holistic features. This significantly enhances robust representation learning, ultimately achieving the best overall performance—notably reaching 55.6% accuracy on CIFAR-100 under 90% noise. These results underscore the synergistic effect of our framework: ATC ensures high-quality sample division, AAM generates high-quality training samples, and MFA consolidates representational robustness. collectively enabling stable learning even under severe label corruption. Refer to *supplementary material* for more ablation results.

4.4 Parameter Sensitivity

To investigate the impact of the mask ratio μ and the stochastic perturbation factor ξ, we conduct parameter sensitivity experiments on CIFAR-10 and CIFAR-

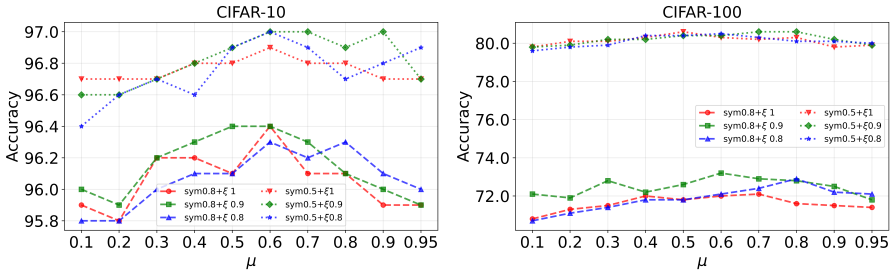

Fig. 3: Parameter sensitivity on CIFAR-10/100 with 50%/80% symmetric noise.

100 datasets under 50% and 80% symmetric noise. As shown in Figure 3 , the model accuracy exhibits a non-monotonic trend with respect to μ, where an optimal balance is achieved at moderate values. For instance, on CIFAR-100 with 80% noise and $\xi = 0.9$, the accuracy peaks at 73.3% when $\mu = 0.6$, whereas excessively low ratios fail to provide sufficient regularization against noisy labels and excessively high ratios result in a severe loss of semantic content that hinders feature discriminability. Meanwhile, the perturbation factor ξ introduces essential randomness to the threshold segmentation in the AAM module, with $\xi = 0.9$ consistently yielding superior or competitive performance across various settings. This randomness helps generalization while maintaining robustness under structural perturbations. Thus, these two parameters are set to $\mu = 0.6$ and $\xi = 0.9$ as they trade off noise suppression and feature discriminability.

5 Conclusion

In this paper, we accurately separate samples by quantifying visual attention stability, thereby transcending the limitations of traditional loss-based metrics. Building upon this, an adaptive differential masking strategy is introduced to synthesize high-quality training samples, compelling the model to localize authentic foreground objects while effectively mitigating background interference. Finally, by aligning original views with their masked counterparts, we successfully steer the model's focus away from unreliable localized regions toward robust holistic semantics, thereby achieving robust representation learning. Experimental results demonstrate that our method achieves state-of-the-art performance under different noise conditions. Despite promising results, it still has limitations: the model's performance may be affected by the initial quality of generated CAMs, and tracking temporal consistency introduces computational overhead during training. In the future, it would be worth designing a lightweight attention-tracking mechanism to reduce costs. Furthermore, exploring the applicability of temporal attention stability to open-set noise tasks deserves further investigation.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62476171), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011367), and the Guangdong Provincial Key Laboratory (No. 2023B1212060076).

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: ICLR. pp. 233–242 (2017)
2. Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., Liu, T.: Understanding and improving early stopping for learning with noisy labels. In: NeurIPS. pp. 24392–24403 (2021)
3. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A.: Mixmatch: A holistic approach to semi-supervised learning. In: NeurIPS (2019)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
5. Cordeiro, F.R., Belagiannis, V., Reid, I., Carneiro, G.: Propmix: Hard sample filtering and proportional mixup for learning with noisy labels. In: BMVC (2021)
6. Cordeiro, F.R., Carneiro, G.: Anne: Adaptive nearest neighbours and eigenvector-based sample selection for robust learning with noisy labels. PR **159**, 111132 (2025)
7. Cordeiro, F.R., Sachdeva, R., Belagiannis, V., Reid, I., Carneiro, G.: Longremix: Robust learning with high confidence samples in a noisy label environment. PR **133**, 109013 (2023)
8. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data (2018)
9. Englesson, E., Azizpour, H.: Robust classification via regression for learning with noisy labels. In: ICLR (2024)
10. Fabian Benitez-Quiroz, C., Srinivasan, R., Martinez, A.M.: Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: CVPR. pp. 5562–5570 (2016)
11. Fan, W., Li, K.: Combating semantic contamination in learning with label noise. In: AAAI. pp. 2870–2878 (2025)
12. Feng, C., Tzimiropoulos, G., Patras, I.: Ssr: An efficient and robust framework for learning with unknown label noise. In: BMVC (2022)
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV. pp. 630–645. Springer (2016)
14. Hossain, M., Kauranen, I.: Crowdsourcing: a comprehensive literature review. Strategic Outsourcing: An International Journal **8**(1), 2–22 (2015)
15. Huang, Z., Zhang, J., Shan, H.: Twin contrastive learning with noisy labels. In: CVPR. pp. 11661–11670 (2023)
16. Kim, T., Ko, J., Choi, J., Yun, S.Y., et al.: Fine samples for learning with noisy labels. In: NeurIPS. pp. 24137–24149 (2021)
17. Ko, J., Yi, B., Yun, S.Y.: A gift from label smoothing: robust training with adaptive label smoothing via auxiliary classifier under label noise. In: AAAI. pp. 8325–8333 (2023)

18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. (2009)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS (2012)
20. Lee, K.H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. In: CVPR. pp. 5447–5456 (2018)
21. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: ICLR (2020)
22. Li, J., Xiong, C., Hoi, S.C.: Learning from noisy data with robust representation learning. In: CVPR. pp. 9485–9494 (2021)
23. Li, W., Wang, L., Li, W., Agustsson, E., Van Gool, L.: Webvision database: Visual learning and understanding from web data. arXiv preprint arXiv:1708.02862 (2017)
24. Li, Y., Han, H., Shan, S., Chen, X.: Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In: CVPR. pp. 24070–24079 (2023)
25. Li, Y., Chen, Y., Yu, X., Chen, D., Shen, X.: Sure: Survey recipes for building reliable and robust deep networks. In: CVPR. pp. 17500–17510 (2024)
26. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. In: NeurIPS. pp. 20331–20342 (2020)
27. Lu, Y., Bo, Y., He, W.: Noise attention learning: Enhancing noise robustness by gradient scaling. In: NeurIPS. vol. 35, pp. 23164–23177 (2022)
28. Lukasik, M., Bhojanapalli, S., Menon, A., Kumar, S.: Does label smoothing mitigate label noise? In: ICLR. pp. 6448–6458 (2020)
29. Mu, C., Qu, Y., Yan, J., Yang, E., Deng, C.: Meta-learning dynamic center distance: Hard sample mining for learning with noisy labels. In: CVPR. pp. 415–425 (2025)
30. Nishi, K., Ding, Y., Rich, A., Hollerer, T.: Augmentation strategies for learning with noisy labels. In: CVPR. pp. 8022–8031 (2021)
31. Pan, W., Wei, W., Zhu, F., Deng, Y.: Enhanced sample selection with confidence tracking: Identifying correctly labeled yet hard-to-learn samples in noisy data. In: AAAI. pp. 19795–19803 (2025)
32. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. In: NeurIPS. vol. 31 (2018)
33. Sheng, M., Sun, Z., Zhou, T., Shu, X., Pan, J., Yao, Y.: Ca2c: A prior-knowledge-free approach for robust label noise learning via asymmetric co-learning and co-training. In: ICCV. pp. 901–911 (2025)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
35. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: ICML. pp. 5907–5915 (2019)
36. Song, H., Kim, M., Park, D., Shin, Y., Lee, J.G.: Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems **34**(11), 8135–8153 (2022)
37. Sun, Z., Shen, F., Huang, D., Wang, Q., Shu, X., Yao, Y., Tang, J.: Pnp: Robust learning from noisy labels by probabilistic noise prediction. In: CVPR. pp. 5311–5320 (2022)
38. Tan, C., Xia, J., Wu, L., Li, S.Z.: Co-learning: Learning from noisy labels with self-supervision. In: ACM MM. pp. 1405–1413 (2021)
39. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: CVPR. pp. 5552–5560 (2018)

40. Tu, Y., Zhang, B., Li, Y., Liu, L., Li, J., Wang, Y., Wang, C., Zhao, C.R.: Learning from noisy labels with decoupled meta label purifier. In: CVPR. pp. 19934–19943 (2023)
41. Tu, Y., Zhang, B., Li, Y., Liu, L., Li, J., Zhang, J., Wang, Y., Wang, C., Zhao, C.R.: Learning with noisy labels via self-supervised adversarial noisy masking. In: CVPR. pp. 16186–16195 (2023)
42. Xia, X., Liu, T., Han, B., Gong, C., Wang, N., Ge, Z., Chang, Y.: Robust early-learning: Hindering the memorization of noisy labels. In: ICLR (2020)
43. Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., Sugiyama, M.: Part-dependent label noise: Towards instance-dependent label noise. In: NeurIPS. pp. 7597–7610 (2020)
44. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: CVPR. pp. 2691–2699 (2015)
45. Xu, Y., Niu, X., Yang, J., Su, R., Zhang, J., Liu, S., Drew, S.: Revisiting interpolation for noisy label correction. In: AAAI. pp. 21833–21841 (2025)
46. Yi, K., Wu, J.: Probabilistic end-to-end noise correction for learning with noisy labels. In: CVPR. pp. 7017–7025 (2019)
47. Yu, X., Han, B., Yao, J., Niu, G., Tsang, I., Sugiyama, M.: How does disagreement help generalization against label corruption? In: ICML. pp. 7164–7173 (2019)
48. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: ICLR (2017)
49. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)
50. Zhang, S., Li, Y., Wang, Z., Li, J., Liu, C.: Learning with noisy labels using hyperspherical margin weighting. In: AAAI. pp. 16848–16856 (2024)
51. Zhang, Y., Zheng, S., Wu, P., Goswami, M., Chen, C.: Learning with feature-dependent label noise: A progressive approach. In: ICLR (2021)
52. Zheng, G., Awadallah, A.H., Dumais, S.: Meta label correction for noisy label learning. In: AAAI. pp. 11053–11061 (2021)
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)
54. Zhou, Y., Li, X., Liu, F., Wei, Q., Chen, X., Yu, L., Xie, C., Lungren, M.P., Xing, L.: L2b: Learning to bootstrap robust models for combating label noise. In: CVPR. pp. 23523–23533 (2024)
55. Zong, C.C., Wang, Y.W., Xie, M.K., Huang, S.J.: Dirichlet-based prediction calibration for learning with noisy labels. In: AAAI. pp. 17254–17262 (2024)