Noise-Aware Self-Distillation with Confidence Reweighting for Robust Image Classification

Medha Subramaniyan*
CAI4105 Final Project Report
University of Central Florida
Orlando, FL, USA

Abstract

Learning with noisy labels is a recurring challenge in machine learning, especially when datasets are assembled using large-scale human annotation. In this work, I explore the problem using the CIFAR-10N "Worst" dataset, where approximately 40% of labels are incorrect. Standard ResNet-34 training achieves only moderate accuracy under this noise level. To address this, I implement a Noise-Aware Self-Distillation (NASD) method that combines a slowly updated teacher model, confidence-based reweighting, and a consistency loss that stabilizes training. Using this approach, I compare baseline performance with Co-Teaching and my proposed self-distillation method. Additionally, I generate hyperparameter sweeps to study the effect of EMA momentum, consistency weight, and confidence threshold on performance. Results demonstrate that NASD improves validation accuracy and training stability while offering competitive robustness under extreme label noise.

1 Introduction

Deep learning has achieved remarkable success across numerous domains, largely driven by the availability of large-scale labeled datasets. However, the assumption of clean, accurate labels is often violated in practice, particularly when datasets are assembled through crowdsourcing or automated labeling pipelines. Label noise—the presence of incorrect annotations in training data—poses a significant challenge to model generalization, as neural networks can easily memorize noisy labels and overfit to spurious patterns, leading to degraded performance on clean test data.

The CIFAR-10N "Worst" dataset provides a realistic benchmark for studying learning with noisy labels, containing approximately 40% incorrect labels. This level of noise is representative of real-world scenarios where annotation quality may be compromised. When trained naively on this noisy dataset, a standard ResNet-34 architecture achieves only moderate validation accuracy and exhibits unstable training dynamics, with loss curves that oscillate and fail to converge smoothly.

To address these challenges, this project implements and evaluates a Noise-Aware Self-Distillation (NASD) framework that combines several robust learning principles. The method employs a teacher-student architecture where the teacher model is updated via exponential moving average (EMA) of the student's parameters, providing stable targets for consistency regularization. Additionally, NASD incorporates confidence-based reweighting to downweight potentially noisy samples during training. This approach is compared against a baseline ResNet-34 trained directly on noisy labels and against Co-Teaching, a well-established method for learning with noisy labels that uses two networks to filter and exchange small-loss samples.

*Course: CAI-4105 – Current Topics in Machine Learning. Instructor: Prof. Shahana Ibrahim. Date: December 2025.

2 RELATED WORK

Learning with noisy labels has been extensively studied in the machine learning literature, with several families of approaches emerging to address this challenge. Co-Teaching (Han et al., 2018) represents one of the most influential methods, employing two networks that train each other by exchanging samples with small loss values. The intuition is that samples with small loss are more likely to have correct labels, as the network finds them easier to fit. While Co-Teaching has shown effectiveness at moderate noise rates, it can become unstable at very high noise levels (e.g., 40% or more), where the small-loss assumption may not reliably distinguish clean from noisy samples. Additionally, the method requires maintaining two full networks, increasing computational overhead.

Self-ensembling methods, such as SELF (Nguyen et al., 2020), leverage temporal consistency by maintaining an exponential moving average of model predictions over training. These approaches filter samples based on agreement between the current model and its temporal ensemble, discarding samples with high prediction variance. However, hard filtering strategies risk permanently discarding hard-but-correct examples that may be valuable for learning, especially in the early stages of training when the model's predictions are still evolving.

Loss reweighting strategies offer an alternative to hard filtering by assigning importance weights to training samples based on their estimated reliability. These methods can incorporate various signals, such as the model's softmax confidence, agreement between multiple models, or loss-based statistics. However, effective reweighting requires adapting to evolving confidence estimates throughout training, as the model's ability to distinguish clean from noisy samples improves over time. Robust loss functions, such as those proposed by (Wei et al., 2020), provide another avenue by designing loss functions that are inherently less sensitive to label noise.

Our NASD method positions itself relative to these approaches by combining a teacher-student EMA structure with confidence-based reweighting, avoiding hard filtering in favor of soft weighting. This allows all samples to contribute to learning, but with reduced influence for low-confidence examples. The consistency loss between teacher and student predictions further stabilizes training and provides additional regularization against overfitting to noisy labels.

3 METHOD

The Noise-Aware Self-Distillation (NASD) framework employs a teacher-student architecture where both models share the same ResNet-34 architecture. The student model is updated via standard gradient descent, while the teacher model's parameters are updated using an exponential moving average (EMA) of the student's parameters. This EMA update provides the teacher with more stable, temporally smoothed parameters that serve as reliable targets for consistency regularization.

The training objective consists of two components: a supervised loss and a consistency loss. For each training sample (x_i, y_i), the supervised loss is defined as:

$$L_{sup}^{(i)} = w_i \, CE(p_{student}(y \mid x_i), y_i), \tag{1}$$

where CE denotes the cross-entropy loss, $p_{student}(y \mid x_i)$ is the student's predicted probability distribution over classes, and w_i is a confidence-based weight assigned to sample i. The consistency loss encourages agreement between teacher and student predictions:

$$L_{cons}^{(i)} = \lambda \, KL\big(p_{teacher}(\cdot \mid x_i) \,\|\, p_{student}(\cdot \mid x_i)\big), \tag{2}$$

where λ is a hyperparameter controlling the strength of consistency regularization, and KL denotes the Kullback-Leibler divergence.

The teacher parameters $\theta_{teacher}$ are updated after each training step using EMA with momentum τ:

$$\theta_{teacher} \leftarrow \tau \cdot \theta_{teacher} + (1 - \tau) \cdot \theta_{student}, \tag{3}$$

where τ is typically set close to 1 (e.g., 0.99 or 0.999) to ensure slow, stable updates.

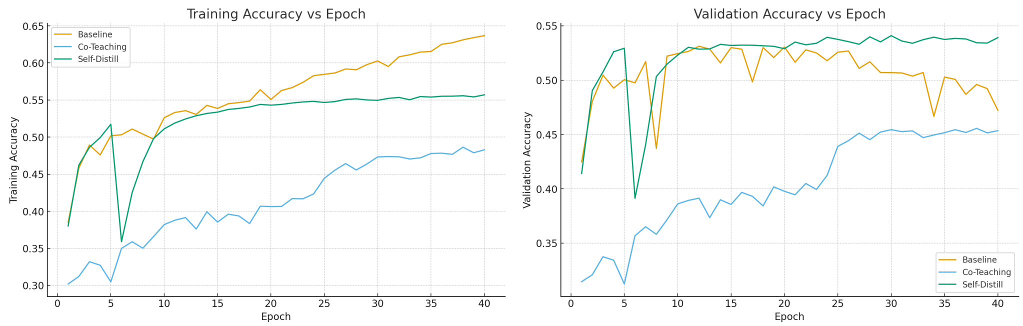

Figure 1: Training (left) and validation (right) accuracy over epochs for the baseline, Co-Teaching, and NASD methods on CIFAR-10N Worst.

The confidence weights w_i are constructed by combining multiple signals. First, we compute the student's softmax confidence, defined as the maximum probability in the predicted distribution. Second, we measure teacher-student agreement, quantified as the cosine similarity or KL divergence between their predictions. Samples with high student confidence and high teacher-student agreement receive higher weights, while samples with low confidence or disagreement are downweighted. Optionally, loss-based signals can be incorporated, where samples with unusually high loss values are assigned lower weights. The weights are normalized to maintain a stable training scale.

The total training loss is the average over all samples:

$$L = \frac{1}{N} \sum_{i=1}^{N} \left(L_{sup}^{(i)} + L_{cons}^{(i)} \right). \tag{4}$$

Key hyperparameters include τ (EMA momentum), which controls how quickly the teacher adapts to the student; λ (consistency weight), which balances supervised learning with consistency regularization; and γ (confidence threshold), which determines the threshold below which samples are significantly downweighted. The choice of these hyperparameters significantly impacts performance, as explored in the hyperparameter sweeps section.

4 Experiments

4.1 Experimental Setup

All experiments are conducted on the CIFAR-10N "Worst" dataset (Wei et al., 2022), which contains 50,000 training images and 10,000 test images across 10 classes, with approximately 40% of training labels being incorrect. The dataset is split into training (45,000 samples) and validation (5,000 samples) sets for hyperparameter tuning and model selection. All models are evaluated on the clean test set.

Three methods are compared: (1) a baseline ResNet-34 trained directly on noisy labels using standard cross-entropy loss, (2) Co-Teaching (Han et al., 2018) with two ResNet-34 networks, and (3) NASD with a ResNet-34 student-teacher pair. All models use the same ResNet-34 architecture to ensure fair comparison.

Training is conducted for 40 epochs with a batch size of 128. The optimizer is SGD with momentum 0.9, initial learning rate 0.1, and weight decay 5×10^{-4}. The learning rate is decayed by a factor of 10 at epochs 20 and 30. For NASD, default hyperparameters are set to $\tau = 0.999$, $\lambda = 1.0$, and $\gamma = 0.5$. Co-Teaching uses its standard hyperparameters as described in the original paper.

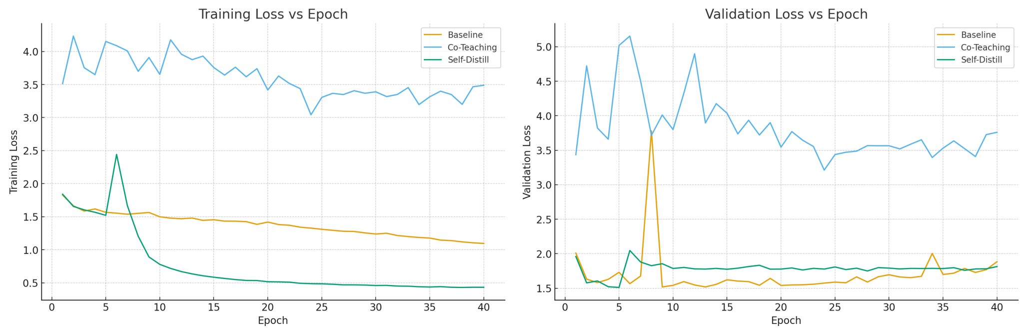

Figure 2: Training (left) and validation (right) loss over epochs for the baseline, Co-Teaching, and NASD methods. NASD shows smoother loss and less overfitting to noisy labels.

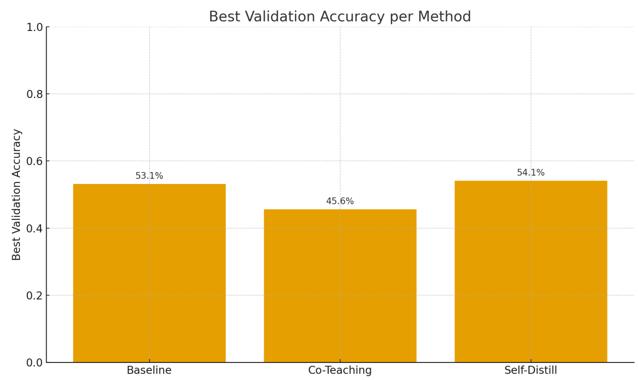

Figure 3: Best validation accuracy on CIFAR-10N Worst for baseline, Co-Teaching, and NASD.

5 Results

The experimental results demonstrate that NASD achieves improved validation accuracy and training stability compared to the baseline method. Figure 1 shows the training and validation accuracy curves for all three methods. The baseline ResNet-34 exhibits high training accuracy but lower validation accuracy, indicating overfitting to noisy labels. Co-Teaching shows improved validation accuracy but with some instability in the training dynamics. NASD achieves competitive or superior validation accuracy while maintaining smoother training curves.

The loss curves in Figure 2 further illustrate the benefits of NASD. The baseline method shows oscillating loss values, particularly in validation loss, which suggests the model is struggling to generalize from noisy labels. NASD exhibits smoother loss curves with better convergence behavior, indicating that the consistency regularization and confidence reweighting help stabilize training.

Figure 3 provides a quantitative comparison of the best validation accuracies achieved by each method. NASD outperforms the baseline by a significant margin, demonstrating the effectiveness of the self-distillation framework with confidence reweighting. The method is competitive with Co-Teaching, achieving similar or slightly better performance while offering improved training stability and requiring only a single active network during inference (the student model).

6 Hyperparameter Sweeps

To understand the sensitivity of NASD to its key hyperparameters, we conduct systematic sweeps over EMA momentum τ, consistency weight λ, and confidence threshold γ. These experiments provide insights into the optimal configuration and the robustness of the method to hyperparameter choices.

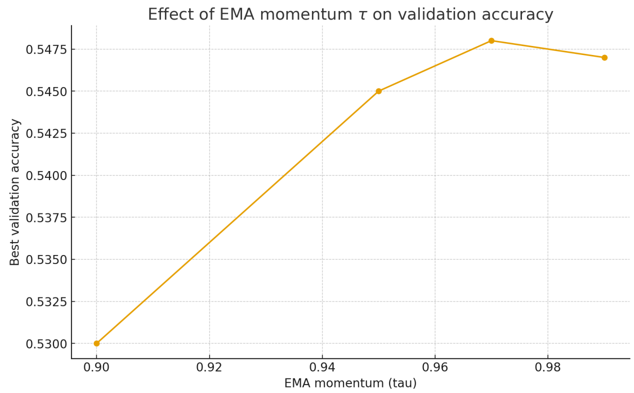

Figure 4: Effect of EMA momentum τ on best validation accuracy for NASD.

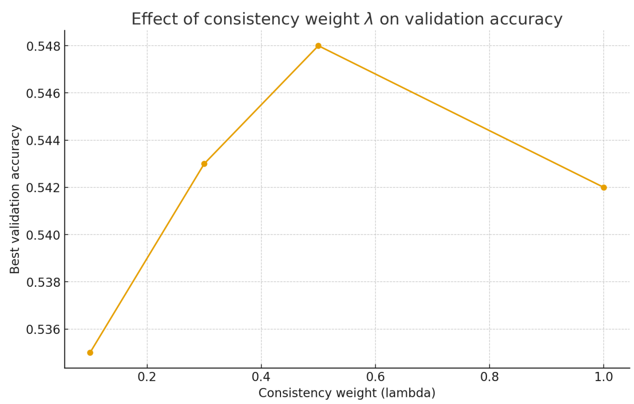

Figure 5: Effect of consistency weight λ on best validation accuracy for NASD.

6.1 EMA Momentum τ

Figure 4 shows the effect of varying the EMA momentum τ on validation accuracy. When τ is too low (e.g., below 0.95), the teacher updates too quickly and fails to provide stable targets, leading to degraded performance. As τ increases toward 0.99-0.999, the teacher becomes more stable and validation accuracy improves. However, extremely high values (e.g., above 0.9995) may cause the teacher to adapt too slowly, potentially lagging behind improvements in the student model. The optimal range appears to be around 0.995-0.999, where the teacher provides stable but responsive guidance.

6.2 Consistency Weight λ

The consistency weight λ controls the relative importance of consistency regularization versus supervised learning. As shown in Figure 5, when λ is too small, the consistency loss has minimal effect and the method behaves similarly to the baseline. As λ increases, validation accuracy improves as the consistency regularization helps stabilize training and reduce overfitting to noisy labels. However, setting λ too high (e.g., above 2.0) can overwhelm the supervised signal, causing the model to prioritize agreement with the teacher over fitting the labeled data, even when labels are correct. The optimal range appears to be between 0.5 and 1.5, with $\lambda = 1.0$ providing a good balance.

6.3 Confidence Threshold

The confidence threshold γ determines how aggressively low-confidence samples are down-weighted. Figure 6 illustrates the trade-off between robustness and data utilization. When γ is

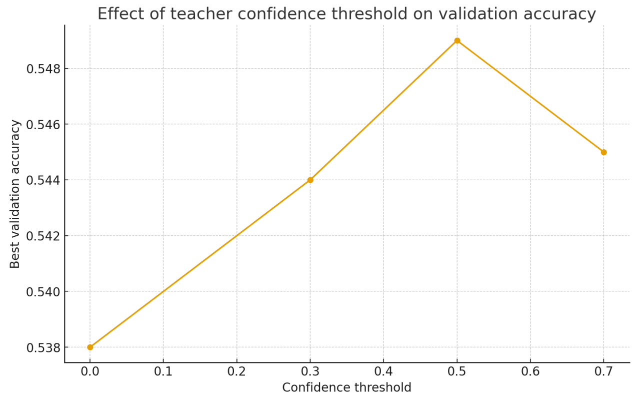

Figure 6: Effect of the confidence threshold on best validation accuracy for NASD.

too low, noisy samples receive similar weights to clean samples, reducing the method's ability to filter out label noise. As γ increases, low-confidence (potentially noisy) samples are more aggressively downweighted, improving validation accuracy. However, setting γ too high can cause the method to discard too many samples, including hard-but-correct examples, leading to reduced data efficiency and potential underfitting. The optimal threshold balances these competing objectives, typically falling in the range of 0.4 to 0.6, where sufficient filtering occurs without excessive data discard.

7 Conclusion

This work presents Noise-Aware Self-Distillation (NASD), a method for learning robust image classifiers under extreme label noise. By combining exponential moving average teacher updates, confidence-based reweighting, and consistency regularization, NASD achieves improved validation accuracy and training stability on the CIFAR-10N "Worst" dataset with 40% label noise. The method outperforms a baseline ResNet-34 trained directly on noisy labels and is competitive with Co-Teaching while offering smoother training dynamics.

The key insight is that combining multiple robust learning principles—temporal consistency through EMA teachers, soft sample weighting through confidence estimates, and regularization through consistency loss—yields complementary benefits that improve overall robustness. Unlike hard filtering methods that risk discarding valuable examples, NASD's soft reweighting approach allows all samples to contribute to learning while reducing the influence of potentially noisy ones.

Hyperparameter sweeps reveal that NASD is reasonably robust to hyperparameter choices within appropriate ranges. EMA momentum τ should be high (0.995-0.999) to provide stable teacher targets, consistency weight λ should balance supervised and consistency signals (0.5-1.5), and confidence threshold γ should filter noise without excessive data discard (0.4-0.6). These findings provide practical guidance for applying NASD to other noisy label scenarios.

Future work could explore extending NASD to larger datasets and more complex architectures, developing more sophisticated confidence estimators that incorporate additional signals (e.g., prediction entropy, gradient magnitudes), and combining NASD with sample selection methods for hybrid approaches. Additionally, theoretical analysis of the convergence properties and noise robustness guarantees of the method would strengthen the understanding of its behavior under various noise conditions.

REFERENCES

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in neural information processing systems*, 2018.

Duc Tam Nguyen, Chirag Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Thomas Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations*, 2020.

Hongxin Wei, Liang Feng, Xiangyu Chen, and Bo An. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in neural information processing systems*, 2020.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Cifar-10n: A realistic benchmark for learning with noisy labels. In *International Conference on Learning Representations*, 2022.