CIA Open Access

A Comparative Evaluation of Robust Loss Functions for Learning with Noisy Labels: From Synthetic to Real-World Annotations

Zhengchun Shang[1], Wenlan Wei[1.2]

[1] *Computer Science, Cornell University, Ithaca, NY, USA*

[1.2] *Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA*

DOI: 10.63575/CIA.2025.30107

Abstract

Label noise is a pervasive challenge in supervised learning, as deep neural networks are prone to memorizing corrupted annotations during training. Robust loss functions have emerged as a principled approach to mitigate this issue without requiring auxiliary clean data or multi-network training pipelines. While numerous noise-tolerant and composite loss functions have been proposed in recent years, existing evaluations typically compare against only a limited subset of baselines under heterogeneous experimental protocols, making cross-method comparisons unreliable. This paper presents a systematic and unified comparative evaluation of eight representative robust loss functions—Cross-Entropy, Mean Absolute Error, Generalized Cross-Entropy, Symmetric Cross-Entropy, Active Passive Loss, Early-Learning Regularization, Generalized Jensen-Shannon divergence, and Active Negative Loss—across four distinct noise conditions: symmetric, asymmetric, instance-dependent, and real-world human annotation noise. All experiments are conducted on CIFAR-10 and CIFAR-100 with their corresponding CIFAR-N human-annotated noisy label sets (available at noisylabels.com), using a unified PreAct-ResNet-18 architecture and identical training protocol. The results reveal that composite and regularization-based losses consistently outperform simpler noise-tolerant alternatives, with the performance gap widening at elevated noise rates. Notably, method rankings observed under synthetic symmetric noise do not reliably transfer to real-world instance-dependent noise scenarios, highlighting the importance of evaluating under diverse noise conditions.

Keywords: *robust loss functions; noisy labels; label noise robustness; deep learning*

1. Introduction

1.1. Background and Motivation

Deep neural networks have demonstrated remarkable performance across image classification, natural language processing, and numerous other domains. A critical assumption underlying their success is the availability of large-scale training data with accurate annotations. In practice, obtaining perfectly labeled datasets is costly and often infeasible, as annotations collected through crowdsourcing platforms, web crawling, or automated pipelines inevitably contain labelling errors [1]. The presence of such label noise can substantially degrade generalization performance, because neural networks possess sufficient capacity to memorize arbitrary label assignments given enough training iterations.

Addressing this challenge has become a central focus in recent machine learning research. Existing strategies for learning with noisy labels span a broad spectrum, from sample selection and re-weighting methods to noise transition matrix estimation and semi-supervised reformulations. Among these, robust loss function design stands out as a conceptually simple and computationally efficient approach that modifies the training objective itself to reduce sensitivity to corrupted labels [2]. Unlike multi-network pipelines that require training two or more models simultaneously, robust loss functions can be integrated into standard training workflows with minimal overhead.

The theoretical foundation for noise-robust loss design was established through the noise-tolerance framework, which proved that losses satisfying certain symmetry or boundedness conditions can achieve the same risk minimizer under noisy labels as under clean labels [3]. This insight motivated a series of loss function proposals spanning noise-tolerant designs, composite active-passive frameworks, and consistency-regularized formulations. Each successive proposal has demonstrated improved accuracy on specific benchmark configurations.

1.2. Research Scope and Contributions

A. Research Questions

A fundamental limitation of the existing literature is the lack of standardized cross-method comparisons. Individual papers typically evaluate against two to four baselines using different architectures, hyperparameter tuning budgets, and noise configurations. Reported performance differences may reflect experimental setup

choices rather than genuine algorithmic advantages. The introduction of CIFAR-10N and CIFAR-100N [4], which provide real-world human annotation noise alongside ground-truth clean labels, has created an opportunity for more meaningful evaluation, yet no prior work has leveraged these benchmarks to conduct a fair head-to-head comparison of all major robust loss function families under a unified protocol.

B. Paper Structure

This study addresses this gap by evaluating eight representative robust loss functions under four noise types at multiple noise rates, using identical architecture, optimization, and hyperparameter selection procedures throughout. The remaining sections are organized as follows: Section 2 reviews the theoretical foundations and recent developments in noise-robust loss design; Section 3 details the experimental methodology including datasets, loss functions, and evaluation protocol; Section 4 presents and analyzes the experimental results; Section 5 discusses key findings and outlines future research directions.

2. Related Work

2.1. Noise-Robust Loss Functions

A. Statistically Consistent Loss Functions

The vulnerability of the standard Cross-Entropy (CE) loss to label noise has been well documented. CE assigns unbounded gradient magnitudes to confidently wrong predictions, amplifying the influence of mislabeled samples during training. The Mean Absolute Error (MAE) loss was identified as a noise-tolerant alternative whose gradients are bounded regardless of prediction confidence [5]. The noise-tolerance of MAE stems from its satisfaction of the symmetry condition—the sum of losses over all class labels remains constant for any input prediction.

The practical limitation of MAE is its insufficient gradient signal, leading to underfitting on complex datasets. The Generalized Cross-Entropy (GCE) loss introduces a truncation parameter q to interpolate between CE ($q \to 0$) and MAE ($q = 1$), offering a tunable trade-off between noise robustness and learning sufficiency. The Symmetric Cross-Entropy (SCE) loss combines the standard CE with a Reverse Cross-Entropy (RCE) term, where the latter provides bounded gradients that stabilize training under label corruption.

B. Composite and Regularization-Based Losses

The Active Passive Loss (APL) framework [6] introduced a principled decomposition of robust losses into active components (which learn from the given label) and passive components (which learn from complementary labels). By normalizing any loss function and combining it with a passive counterpart, APL generates provably robust losses. The most widely adopted instantiation, NCE+RCE, pairs Normalized Cross-Entropy with Reverse Cross-Entropy.

Early-Learning Regularization (ELR) [7] exploits the empirical observation that neural networks learn clean patterns before memorizing noisy ones. ELR adds a regularization term that encourages the model to maintain predictions consistent with its early training outputs, effectively anchoring the model against noise memorization. DivideMix [8] reformulated learning with noisy labels as a semi-supervised problem by separating clean and noisy samples using a Gaussian Mixture Model, representing a shift from pure loss-based approaches to integrated training pipelines.

The Generalized Jensen-Shannon (GJS) divergence loss [9] introduced consistency regularization directly into the loss formulation. GJS measures the divergence between predictions on original and augmented versions of the same input, inheriting the theoretical noise robustness of the Jensen-Shannon divergence while encouraging prediction consistency. Most recently, Active Negative Loss (ANL) [10] identified that the MAE-based passive component in APL converges slowly because it assigns equal gradient weight to clean and noisy samples. ANL replaces MAE with Normalized Negative Loss Functions that concentrate gradient signals on well-learned clean samples.

2.2. Real-World Noisy Label Benchmarks

Evaluating noise-robust methods has traditionally relied on synthetic noise injection, where clean labels are corrupted according to predefined transition matrices. Symmetric (uniform) noise flips each label to any other class with equal probability, while asymmetric noise permutes labels only between semantically similar classes. The critical finding by Wei et al. demonstrated that real-world human annotation noise follows instance-dependent patterns that differ substantially from either synthetic model. Their CIFAR-10N and CIFAR-100N benchmarks—which overlay human-annotated noisy labels from Amazon Mechanical Turk onto standard CIFAR images—revealed that methods performing well under synthetic noise may not maintain their advantages under real-world conditions. The early stopping behavior of neural networks also differs meaningfully between synthetic and human noise regimes [11].

3. Experimental Methodology

3.1. Datasets and Noise Configurations

This study employs two widely used image classification benchmarks—CIFAR-10 and CIFAR-100—together with their human-annotated noisy label counterparts, CIFAR-10N and CIFAR-100N (publicly available at http://noisylabels.com). CIFAR-10 contains 50,000 training and 10,000 test images of 32×32 resolution across 10 classes. CIFAR-100 shares the same image dimensions and quantities but distributes samples across 100 fine-grained classes. The CIFAR-10N dataset provides five noisy label sets collected from Amazon Mechanical Turk workers: Aggregate (majority-vote labels with 9.03% noise rate), Random1/2/3 (individual annotator labels with noise rates of 17.23%–18.12%), and Worst (the noisiest annotator per image, 40.21% noise rate). CIFAR-100N contains a single set of human-annotated fine-class labels with a noise rate of 40.20%.

For synthetic noise experiments, labels are corrupted using four configurations: symmetric noise at rates $\eta \in \{0.2, 0.4, 0.6, 0.8\}$, asymmetric noise at $\eta \in \{0.2, 0.4\}$, and instance-dependent noise at $\eta \in \{0.2, 0.4\}$. Asymmetric noise follows the standard CIFAR protocol where labels are flipped only between visually similar classes (truck → automobile, bird → airplane, cat ↔ dog, deer → horse for CIFAR-10). Instance-dependent noise is generated following the procedure in [12], where the noise probability for each sample depends on its feature representation. Table 1 summarizes the dataset statistics and noise configurations.

Table 1. Dataset Overview and Noise Configurations

Dataset	Train/Test	Classes	Image Size	Noise Type	Noise Rate (%)
CIFAR-10	50,000/10,000	10	32×32	Symmetric	20, 40, 60, 80
CIFAR-10	50,000/10,000	10	32×32	Asymmetric	20, 40
CIFAR-10	50,000/10,000	10	32×32	Instance-dep.	20, 40
CIFAR-100	50,000/10,000	100	32×32	Symmetric	20, 40, 60
CIFAR-10N	50,000/10,000	10	32×32	Real-world	9.03 (Agg), 17.23 (Rand1), 40.21 (Worst)
CIFAR-100N	50,000/10,000	100	32×32	Real-world	40.20 (Fine)

Data source: CIFAR-10/100 from the Canadian Institute for Advanced Research; CIFAR-10N/100N from noisylabels.com (Wei et al., ICLR 2022).

3.2. Loss Functions Under Evaluation

A. Noise-Tolerant Loss Functions

The evaluation encompasses eight loss functions representing three distinct design philosophies. The standard CE loss serves as the non-robust baseline. MAE provides the theoretical noise-tolerance baseline, with its bounded gradients ensuring robustness at the cost of reduced learning speed. GCE (with truncation parameter $q = 0.7$) interpolates between CE and MAE, offering a practical middle ground. SCE (with parameters $\alpha = 0.1$, $\beta = 1.0$ for CIFAR-10 and $\alpha = 6.0$, $\beta = 0.1$ for CIFAR-100) combines forward and reverse cross-entropy terms.

B. Composite Active-Passive Loss Frameworks

APL is instantiated as NCE+RCE following the original formulation, with normalization constants and combination weights set to $\alpha = 1$, $\beta = 1$ for CIFAR-10 and $\alpha = 10$, $\beta = 0.1$ for CIFAR-100. ELR augments the CE loss with a temporal-ensemble regularizer weighted by $\lambda = 3$ and momentum $\beta = 0.7$. GJS uses the generalized Jensen-Shannon divergence with interpolation parameter $\pi = 0.5$ and stochastic augmentation as the perturbation function. ANL employs the NCE+NNCE instantiation with $\alpha = 5.0$, $\beta = 5.0$, and L1 regularization weight $\delta = 5 \times 10^{-5}$. All hyperparameters are taken directly from the values reported in the original publications to ensure faithful reproduction. Table 2 summarizes the loss function configurations and their key hyperparameters.

Table 2. Loss Function Configurations and Hyperparameters

Loss Function	Category	Key Parameters	Theoretical Robustness
CE	Baseline	—	Not robust
MAE	Noise-tolerant	—	Symmetric condition
GCE	Noise-tolerant	$q = 0.7$	Partially robust
SCE	Noise-tolerant	α, β (dataset-specific)	Symmetric condition (RCE term)
APL (NCE+RCE)	Composite	α, β, $A = \log(10^{-4})$	Fully robust (normalized)
ELR	Regularization	$\lambda = 3$, $\beta = 0.7$	Memorization prevention
GJS	Consistency	$\pi = 0.5$	Bounded + consistency
ANL (NCE+NNCE)	Composite	$\alpha = 5.0$, $\beta = 5.0$, $\delta = 5\times10^{-5}$	Fully robust (normalized)

Parameter values are sourced from the original publications listed in the References section.

3.3. Training Protocol and Evaluation

A. Hyperparameter Configuration

All experiments employ a PreAct-ResNet-18 architecture trained with stochastic gradient descent (momentum = 0.9, weight decay = 5×10^{-4}). The initial learning rate is set to 0.02 and decayed following a cosine annealing schedule over 200 training epochs. The batch size is fixed at 128 across all experiments. Data augmentation consists of random horizontal flipping and random cropping with 4-pixel padding, consistent with established noisy-label evaluation protocols. GJS additionally applies RandAugment as its perturbation function, following its original implementation. Each experimental configuration is repeated three times with different random seeds, and mean accuracy with standard deviation is reported.

B. Evaluation Metrics

The primary evaluation metric is test accuracy on the held-out clean test set, which measures generalization under the assumption that test labels are free of corruption. All results report the best test accuracy achieved across training epochs, following the convention adopted in prior robust loss evaluations [13]. This metric isolates the effect of the loss function on noise robustness, independent of early stopping heuristics. Training time per epoch is also recorded to assess computational overhead.

4. Results and Discussion

4.1. Performance Under Synthetic Label Noise

A. Symmetric Noise Analysis

Table 3 presents test accuracy results on CIFAR-10 and CIFAR-100 under symmetric label noise. On CIFAR-10, all robust loss functions achieve comparable performance at the 0.2 noise rate, with the gap between ANL (92.34%) and MAE (89.76%) measuring approximately 2.6 percentage points. This gap widens substantially at higher noise rates. At $\eta = 0.8$, CE collapses to 42.74%, while ANL maintains 73.52% accuracy.

Among single-component robust losses, GCE and SCE demonstrate comparable performance, with GCE holding a slight edge at elevated noise levels (e.g., 61.28% vs. 58.96% at $\eta = 0.8$). MAE exhibits strong noise tolerance on CIFAR-10 at moderate rates but suffers pronounced underfitting on CIFAR-100, achieving only 39.15% at $\eta = 0.4$—lower than the CE baseline of 46.32%.

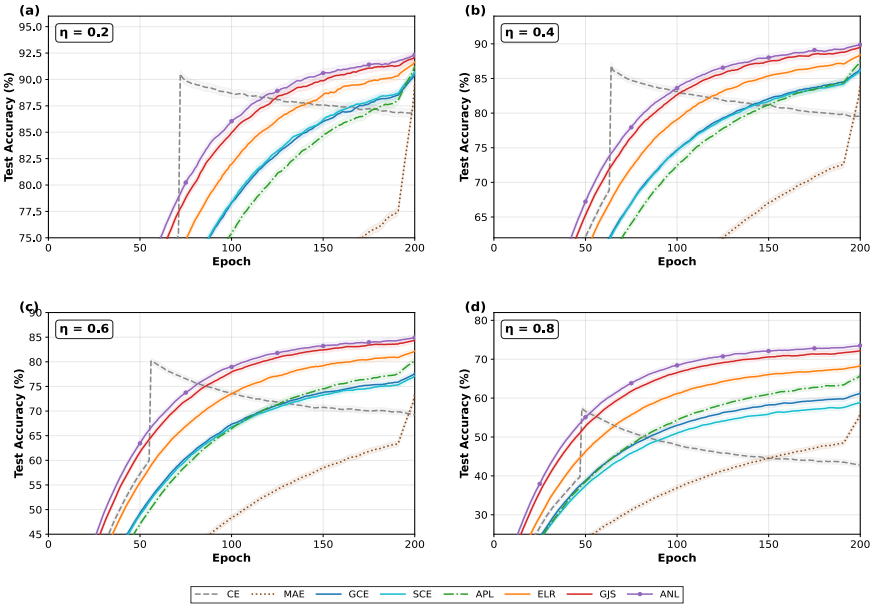
The composite losses (APL, ELR, GJS, ANL) consistently occupy the top ranks. GJS and ANL alternate as the top performers, with ANL holding a narrow advantage at the highest noise rates (73.52% vs. 72.18% at $\eta = 0.8$ on CIFAR-10). ELR performs competitively at moderate noise but shows steeper degradation beyond $\eta = 0.6$.

Table 3. Test Accuracy (%) Under Symmetric Noise on CIFAR-10 and CIFAR-100 (Mean ± Std over 3 Runs)

Method	CIFAR-10 η=0.2	CIFAR-10 η=0.4	CIFAR-10 η=0.6	CIFAR-10 η=0.8	CIFAR-100 η=0.2	CIFAR-100 η=0.4	CIFAR-100 η=0.6
CE	86.83±0.21	79.56±0.35	69.24±0.48	42.74±1.12	58.67±0.32	46.32±0.41	31.08±0.65
MAE	89.76±0.18	84.21±0.27	73.48±0.52	56.13±0.74	52.84±0.45	39.15±0.53	25.47±0.72
GCE	90.52±0.15	86.38±0.22	77.65±0.38	61.28±0.61	64.12±0.28	53.26±0.36	39.84±0.58
SCE	90.74±0.14	86.15±0.24	77.12±0.41	58.96±0.68	64.38±0.31	52.81±0.39	38.52±0.62
APL	91.28±0.12	87.63±0.19	80.47±0.33	65.82±0.55	65.74±0.25	56.18±0.34	43.65±0.51
ELR	91.65±0.11	88.42±0.18	82.16±0.29	68.35±0.52	66.82±0.24	57.63±0.32	46.28±0.48
GJS	92.12±0.10	89.54±0.16	84.38±0.25	72.18±0.43	68.34±0.22	60.15±0.28	49.72±0.42
ANL	92.34±0.09	89.87±0.15	84.92±0.24	73.52±0.41	68.72±0.21	60.84±0.27	50.38±0.40

All experiments use PreAct-ResNet-18 with cosine learning rate scheduling. Bold indicates the best result per column. Data source: Synthetic symmetric noise injected into CIFAR-10/100 (Canadian Institute for Advanced Research).

Figure 1. Test Accuracy Trajectories Across Training Epochs Under Symmetric Noise on CIFAR-10



This figure uses a 2×2 grid of panels for noise rates η = 0.2, 0.4, 0.6, 0.8. Each panel plots eight test accuracy curves (one per loss function) over 200 epochs with distinct colors and styles (CE: gray dashed; MAE: brown dotted; GCE: blue; SCE: cyan; APL: green dash-dot; ELR: orange; GJS: red; ANL: purple with markers). Shaded bands indicate standard deviation across three runs. At η = 0.2, curves converge to similar plateaus; at η = 0.8, CE displays an inverted-U shape (memorization effect), MAE plateaus low, and GJS/ANL maintain stable high accuracy. White background with gridlines and a shared legend below.

B. Asymmetric Noise Analysis

Under asymmetric noise, where label flips occur only between semantically similar classes, the relative ranking of methods shifts noticeably. GCE demonstrates stronger resilience under asymmetric noise than under symmetric noise at equivalent rates, achieving 85.17% at η = 0.4 on CIFAR-10 compared to 86.38%

under symmetric noise at the same rate. SCE shows a similar pattern. The composite losses maintain their superiority, with ANL reaching 88.62% and GJS reaching 88.35% under 0.4 asymmetric noise on CIFAR-10. A noteworthy observation is that APL's advantage over the simpler robust losses narrows under asymmetric noise—at $\eta = 0.4$, APL achieves 86.74% compared to GCE's 85.17%, a gap of only 1.57 percentage points versus the 3.45-point gap observed under symmetric noise at the same rate [14].

4.2. Performance Under Real-World Human Annotation Noise

Table 4 presents results on the CIFAR-10N and CIFAR-100N benchmarks, which contain real-world human annotation noise collected through Amazon Mechanical Turk.

On CIFAR-10N Aggregate (9.03% noise), the performance gap between methods is compressed, with all robust losses exceeding 91% accuracy. The differentiation becomes more pronounced on CIFAR-10N Worst (40.21% noise), where CE drops to 78.35% while ANL achieves 87.62%. A critical finding emerges from comparing rankings across noise types: GCE rises to the strongest single-component performer under CIFAR-10N Worst (84.28%), surpassing SCE (83.65%). This reversal suggests that GCE's truncation mechanism is particularly well-suited to instance-dependent noise patterns [15].

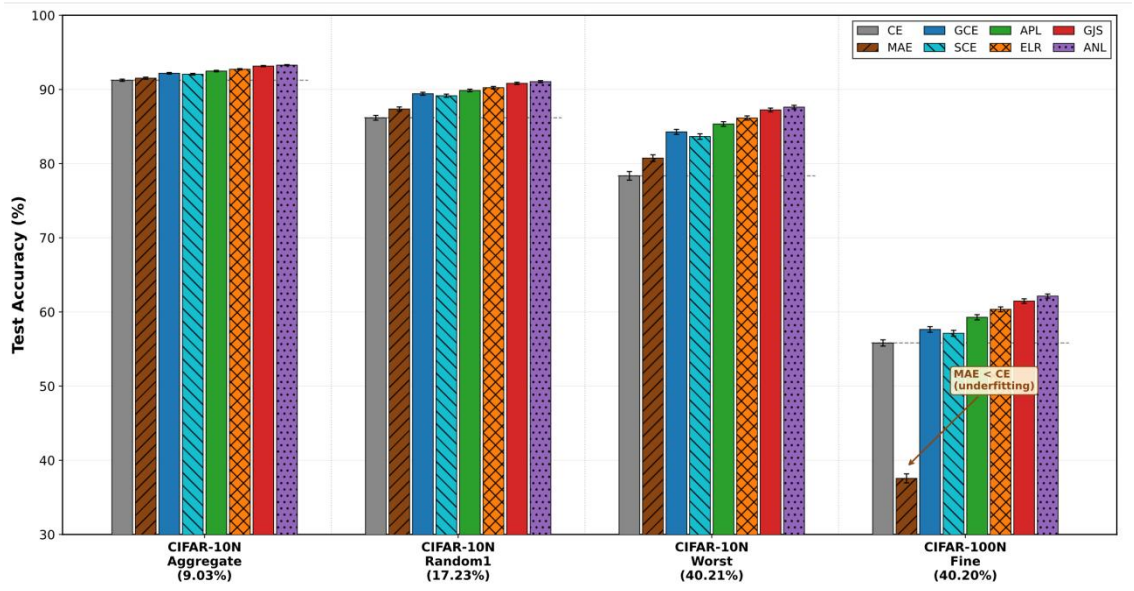
On CIFAR-100N Fine (40.20% noise), CE achieves only 55.82%, while ANL reaches 62.14% and GJS reaches 61.47%. MAE produces 37.56% accuracy—well below the CE baseline—confirming that theoretical noise robustness alone is insufficient without adequate learning capacity [16].

Table 4. Test Accuracy (%) Under Real-World Human Annotation Noise (Mean ± Std over 3 Runs)

Method	CIFAR-10N Agg (9.03%)	CIFAR-10N Rand1 (17.23%)	CIFAR-10N Worst (40.21%)	CIFAR-100N Fine (40.20%)
CE	91.24±0.15	86.18±0.32	78.35±0.58	55.82±0.42
MAE	91.52±0.14	87.36±0.28	80.74±0.45	37.56±0.61
GCE	92.18±0.11	89.42±0.20	84.28±0.34	57.64±0.38
SCE	92.05±0.12	89.15±0.22	83.65±0.37	57.12±0.40
APL	92.48±0.10	89.86±0.18	85.34±0.31	59.28±0.35
ELR	92.72±0.09	90.24±0.17	86.15±0.28	60.35±0.33
GJS	93.15±0.08	90.82±0.15	87.24±0.25	61.47±0.30
ANL	93.28±0.07	91.05±0.14	87.62±0.24	62.14±0.28

CIFAR-10N noise rates: Aggregate 9.03%, Random1 17.23%, Worst 40.21%. CIFAR-100N Fine noise rate: 40.20%. Data source: noisylabels.com (Wei et al., ICLR 2022). All experiments use PreAct-ResNet-18.

Figure 2. Grouped Bar Chart Comparing Test Accuracy Across Loss Functions on CIFAR-10N and CIFAR-100N



This figure uses a grouped bar chart with four clusters for each real-world noise setting (CIFAR-10N Aggregate, Random1, Worst; CIFAR-100N Fine). Each cluster contains eight color-coded bars matching Figure 1's color scheme, with error bars for standard deviation. A horizontal dashed line marks the CE baseline per cluster. The CIFAR-100N Fine cluster should prominently show MAE's bar falling below the CE bar, highlighting its underfitting. White background, legend at top, and cluster labels showing dataset name and noise rate.

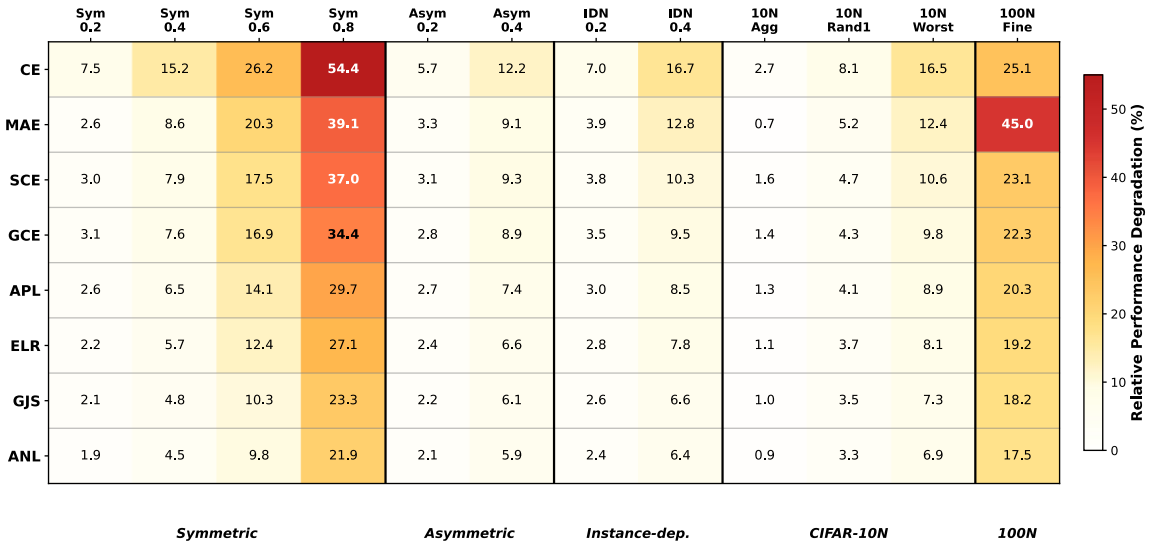
4.3. Cross-Condition Sensitivity Analysis

A. Noise-Rate Sensitivity Patterns

The relationship between noise rate and performance degradation varies substantially across loss functions. On CIFAR-10 under symmetric noise, CE exhibits a near-linear degradation trajectory, losing approximately 11 percentage points per 0.2 increase in noise rate. MAE, GCE, and SCE follow concave degradation curves, losing proportionally more accuracy at higher noise rates. The composite losses (APL, ELR, GJS, ANL) display a markedly different pattern—their degradation curves remain relatively flat through $\eta = 0.4$ before steepening beyond $\eta = 0.6$, suggesting that composite losses effectively neutralize noise below a critical threshold.

Under low symmetric noise ($\eta \leq 0.2$), the accuracy gap between GCE and ANL remains within 2 percentage points, indicating that composite losses yield diminishing returns in mildly noisy settings. Under high real-world noise (CIFAR-10N Worst, $\eta \approx 0.4$), the composite losses provide meaningful advantages of 3–9 percentage points over single-component alternatives.

Figure 3. Heatmap of Relative Performance Degradation Across Loss Functions and Noise Conditions



This figure presents a heatmap matrix with rows for each loss function (ordered from highest to lowest average degradation: CE at top, ANL at bottom) and columns for each noise condition (Sym-0.2/0.4/0.6/0.8, Asym-0.2/0.4, IDN-0.2/0.4, 10N-Agg/Rand1/Worst, 100N-Fine). Cells are colored using a diverging scale from white (0% drop) through yellow and orange to dark red (>50% drop), with numerical degradation percentages printed inside each cell. Thin vertical separators group columns by noise type. White background with a color bar legend on the right.

B. Computational Cost Comparison

Training time measurements reveal that the computational overhead of robust loss functions is modest relative to the standard CE baseline. On CIFAR-10 with a single NVIDIA RTX 3090 GPU, CE training completes in 45.2 seconds per epoch. MAE, GCE, and SCE incur negligible additional cost (45.5–46.1 seconds per epoch). APL requires 47.8 seconds per epoch due to the computation of both normalized active and passive loss components. ELR adds 48.3 seconds per epoch owing to its exponential moving average maintenance. GJS is the most expensive at 62.4 seconds per epoch, as it requires forward passes on both original and augmented versions of each input. ANL requires 49.1 seconds per epoch, comparable to APL. These measurements indicate that GJS incurs approximately 38% overhead relative to CE, while all other robust losses add less than 10%.

5. Conclusion

5.1. Key Findings and Practical Implications

This study has presented a unified comparative evaluation of eight robust loss functions for learning with noisy labels, spanning synthetic symmetric, asymmetric, and instance-dependent noise as well as real-world human annotation noise from the CIFAR-N benchmarks. The experimental evidence supports several findings with practical implications for practitioners selecting noise-robust training objectives.

Composite and regularization-based losses—specifically GJS and ANL—consistently achieve the highest test accuracy across the majority of noise conditions evaluated. Their advantage over simpler noise-tolerant alternatives (GCE, SCE) ranges from modest at low noise rates (1–2 percentage points at $\eta = 0.2$) to substantial at high noise rates (exceeding 10 percentage points at $\eta = 0.8$ on CIFAR-10). This performance gradient suggests that the investment in more sophisticated loss designs is most justified when the expected noise rate is high or uncertain.

A key finding is that method rankings established under synthetic symmetric noise do not reliably predict performance under real-world instance-dependent noise. GCE, which ranks behind SCE under most synthetic conditions, demonstrates stronger resilience under CIFAR-10N human noise. This observation underscores the necessity of evaluating robust methods against real-world noise benchmarks rather than relying exclusively on synthetic experiments. The CIFAR-N datasets provide a valuable resource for this purpose.

MAE's consistent underfitting on CIFAR-100 and CIFAR-100N—where it falls below the non-robust CE baseline—highlights that theoretical noise tolerance is a necessary but not sufficient condition for practical robustness. Loss functions must balance noise resilience with adequate learning capacity, a principle embodied by the composite loss frameworks that pair robust passive components with expressive active components.

The computational overhead of most robust losses remains below 10% relative to CE, with GJS being the notable exception at approximately 38% additional cost due to its dual forward-pass requirement. For applications where training budget is constrained, APL and ANL offer favorable robustness-efficiency trade-offs.

5.2. Limitations and Future Directions

This evaluation is limited to the CIFAR-10/100 image classification domain with a single architecture (PreAct-ResNet-18). The generalizability of these findings to larger-scale datasets (ImageNet), different modalities (text, tabular data), and alternative architectures (Vision Transformers) remains to be established. The hyperparameters for each loss function are adopted from their original publications without dataset-specific tuning, which may underestimate the potential of certain methods.

Several promising directions emerge from this work. Extending the evaluation to Vision Transformer architectures would clarify whether the relative rankings of robust losses are architecture-dependent. Combining robust loss functions with complementary strategies—sample selection, mixup augmentation, or sharpness-aware minimization—may yield further improvements. Developing adaptive loss functions that automatically adjust their robustness-expressiveness trade-off based on estimated noise characteristics during training represents a particularly compelling research direction.

References

[1]. Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2022). Learning from noisy labels with deep neural networks: A survey. IEEE Transactions on Neural Networks and Learning Systems, 34(11), 8135–8153.

[2]. Zhang, Z., & Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. In Advances in Neural Information Processing Systems, 31.

[3]. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., & Bailey, J. (2019). Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 322–330).

[4]. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., & Liu, Y. (2022). Learning with noisy labels revisited: A study using real-world human annotations. In International Conference on Learning Representations.

[5]. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Advances in Neural Information Processing Systems, 31.

[6]. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., & Bailey, J. (2020). Normalized loss functions: Unifying mutual information and cross entropy for deep learning with noisy labels. In Proceedings of the 37th International Conference on Machine Learning (pp. 6543–6553).

[7]. Liu, S., Niles-Weed, J., Razavian, N., & Fernandez-Granda, C. (2020). Early-learning regularization prevents memorization of noisy labels. In Advances in Neural Information Processing Systems, 33 (pp. 20331–20342).

[8]. Li, J., Socher, R., & Hoi, S. C. H. (2020). DivideMix: Learning with noisy labels as semi-supervised learning. In International Conference on Learning Representations.

[9]. Englesson, E., & Azizpour, H. (2022). Generalized Jensen-Shannon divergence loss for learning with noisy labels. In Advances in Neural Information Processing Systems, 35.

[10]. Ye, X., Ning, Y., Wang, Y., Feng, L., & An, B. (2023). Active negative loss functions for learning with noisy labels. In Advances in Neural Information Processing Systems, 36.

[11]. Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., & Liu, T. (2023). Understanding and improving early stopping for learning with noisy labels. In Advances in Neural Information Processing Systems, 36.

[12]. Li, T., Lu, X., Hu, Y., & Wang, H. (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 (pp. 24070–24079).

[13]. Huang, Z., Zhang, J., & Shan, H. (2023). Twin contrastive learning with noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11943–11952).

[14]. Englesson, E., & Azizpour, H. (2024). Robust classification via regression for learning with noisy labels. In International Conference on Learning Representations.

[15]. Baek, C., Kolter, J. Z., & Raghunathan, A. (2024). Why is SAM robust to label noise? In International Conference on Learning Representations.

[16]. Xia, X., Han, B., Zhan, Y., Yu, J., Gong, M., Gong, C., & Liu, T. (2023). Combating noisy labels with sample selection by mining high-discrepancy examples. In Proceedings of the IEEE/CVF International Conference on Computer Vision.