# Understanding CNN Robustness Under Label and Input Corruption

Jaeyeol Lim
Seoul National University
limlimlim00@snu.ac.kr

## Abstract

*Understanding the behavior of machine learning models under distributional shifts is essential for improving their reliability in real-world applications. In this study, we investigate the performance degradation of a convolutional neural network (CNN) trained on the CIFAR-10 dataset under various training-time perturbations. We evaluate the model's robustness against three distinct conditions: (1) randomly shuffled labels, (2) partial label noise affecting 20% of the training data, and (3) input perturbations including random cropping, blurring, color distortion, and grayscale conversion. Experimental results demonstrate that while the baseline model achieves a test accuracy of 89.36%, training with randomly shuffled labels leads to a complete collapse in generalization (9.58%). Partial label noise results in moderate degradation (76.87%), whereas input perturbations marginally improve performance (90.32%). These findings suggest that robustness to input-level corruption does not necessarily translate to robustness against label-level noise, underscoring the importance of addressing different types of training data imperfections separately. The code and experimental details are publicly available at https://github.com/limlimlim00/Mini-Project-CIFAR-10-image-classification.*

## 1. Introduction

Convolutional Neural Networks (CNNs) have demonstrated impressive performance on benchmark image classification datasets such as CIFAR-10. However, despite their success on clean and well-curated datasets, these models often exhibit degraded performance when exposed to real-world imperfections, including label noise, distribution shifts, and input corruptions. Understanding how such factors affect generalization is critical for building reliable machine learning systems in practical deployment scenarios.

While substantial progress has been made in optimizing models under standard training conditions, relatively fewer studies have systematically compared the effects of heterogeneous training-time data corruptions—such as label noise and input-level distortions—within a unified experimental framework. For example, a model may exhibit robustness to image perturbations while remaining highly vulnerable to even mild semantic inconsistencies in the labels. These asymmetries in robustness raise fundamental questions about what deep neural networks actually learn and to what extent they rely on data fidelity during training.

To investigate this issue, we conduct a controlled set of experiments using the CIFAR-10 dataset and a fixed CNN architecture, evaluating model robustness under four distinct training conditions:

- **Baseline**: Clean images with correct labels.
- **Random Label Shuffle**: Labels are randomly reassigned, eliminating any semantic alignment between inputs and outputs.
- **Label Noise (20%)**: A fixed 20% subset of training labels are randomly replaced with incorrect values.
- **Input Perturbation**: Input images are augmented using a composition of random resized cropping, horizontal flipping, color jittering, Gaussian blurring, and occasional grayscale conversion, while preserving the original labels.

To evaluate the impact of these corruption types, we analyze overall classification accuracy, training dynamics (loss curves), and confusion matrices under each condition. Our study is driven by the following research questions:

1. How does each type of data corruption influence model performance?
2. Are CNNs more susceptible to semantic (label-level) noise or superficial (input-level) distortions?
3. What patterns of failure and confusion emerge under different corruption scenarios?

Our findings reveal a sharp divergence in robustness across conditions: while the model achieves slightly improved performance under input perturbations, it fails to

achieve meaningful generalization under random label assignments, and exhibits moderate degradation in the presence of partial label noise. These results suggest a need for more nuanced training paradigms that can tolerate diverse imperfections in supervision without compromising generalization performance.

## 2. Related Works

Robustness in deep learning has been a longstanding research focus, especially in contexts involving noisy supervision or corrupted inputs. Prior work has highlighted the vulnerability of neural networks to both label noise and input perturbations.

**Label Noise.** Zhang et al. [1] demonstrated that deep neural networks can easily memorize random labels, raising questions about generalization. Arazo et al. [2] proposed methods to mitigate the impact of noisy labels using semi-supervised and sample reweighting techniques. Recent studies have further explored the challenges posed by real-world label noise, which can be more detrimental than simulated noise. For instance, Wu et al. [3] introduced NoisywikiHow, a benchmark designed to replicate real-world noise by explicitly constructing multiple sources of label noise to imitate human annotation errors. Additionally, Saeed et al. [4] proposed Few-Shot Human-in-the-Loop Refinement (FHLR), a method that enhances model generalization in the presence of noisy labels by incorporating minimal expert corrections.

**Input Perturbation.** Hendrycks and Dietterich [5] proposed a corruption benchmark (CIFAR-10-C) to evaluate models under common image corruptions such as blur, noise, and weather effects. Their findings emphasize the performance degradation under realistic distortions, motivating the need for data augmentation and adversarial training strategies. Further studies have investigated the robustness of models to various input perturbations. For example, a recent study introduced Perturbation-Rectified OOD detection (PRO), a post-hoc method that leverages the insight that prediction confidence for out-of-distribution inputs is more susceptible to reduction under perturbation than in-distribution inputs [6].

**Comparative Robustness Analysis.** Recent works suggest that generalization performance on test sets may not reflect true robustness to distributional shifts [7, 8]. However, few studies have explicitly compared the impact of label-level versus input-level perturbations under a controlled setting. Our study differs in that we adopt a single baseline CNN and evaluate its behavior across multiple controlled corruption settings, revealing asymmetries in robustness that are often overlooked.

## 3. Methods

This section outlines the experimental setup used to evaluate model robustness under different training-time corruption scenarios. We describe the dataset, model architecture, training configuration, and the specific corruption strategies applied during training.

### 3.1. Dataset

We use the CIFAR-10 dataset, which consists of 60,000 color images ($32\times32$ pixels) across 10 classes, with 50,000 training and 10,000 test images. The dataset is balanced across classes and widely adopted as a benchmark for image classification tasks.

### 3.2. Model Architecture

Our classifier is a convolutional neural network (CNN) inspired by the VGG16 architecture, which is characterized by its deep and uniform structure. The model comprises five convolutional blocks. Each block contains two convolutional layers with $3\times3$ kernels and padding of 1, followed by batch normalization and ReLU activation. Each block is followed by a $2 \times 2$ max pooling layer that reduces spatial resolution.

The number of channels increases progressively through the blocks: 64, 128, 256, 512, and 512, enabling the model to learn hierarchical representations. After the final convolutional layer, an average pooling layer is applied. The classifier consists of a flattening layer, a dropout layer ($p = 0.5$), and a fully connected layer mapping the 512-dimensional feature to 10 output classes.

The architecture is held fixed across all experimental conditions to ensure that performance differences are attributable to data corruption rather than model capacity.

### 3.3. Training Details

All models are trained from scratch using the Adam optimizer with a learning rate of $1 \times 10^{-3}$ and a batch size of 128. The loss function is cross-entropy loss. Training is performed for 100 epochs on a single NVIDIA A100 GPU provided by Google Colab. To ensure consistency across experimental conditions, no early stopping or learning rate scheduling is applied.

All images are normalized to have a mean and standard deviation of $(0.5, 0.5, 0.5)$. Two data augmentation strategies are used depending on the training condition:

- **Input Perturbation:** A complex augmentation pipeline is applied, including random resized cropping ($32 \times 32$, scale range: 0.5–1.0), horizontal flipping, color jittering (brightness, contrast, saturation set to

0.5), Gaussian blurring (kernel size 3, applied with 50% probability), and random grayscale conversion (applied with 20% probability). These augmentations simulate acquisition errors such as lens blur, framing inconsistencies, and lighting variations.

- **Other conditions:** Only horizontal flipping is used to maintain input variability while avoiding additional noise.

For all conditions, test-time preprocessing is kept fixed and consists of tensor conversion and normalization without any augmentation.

### 3.4. Corruption Settings

We evaluate model robustness under four training-time corruption scenarios:

- **Baseline:** Clean images and correct labels from the original CIFAR-10 dataset.
- **Random Label Shuffle:** Training labels are randomly reassigned independently of the input images, removing all semantic correlation.
- **Label Noise (20%):** 20% of the training labels are randomly replaced with incorrect labels drawn uniformly from the remaining 9 classes.
- **Input Perturbation:** The same augmentation pipeline described above is applied to simulate realistic input distortions, while labels remain unchanged.

Each setting is trained independently from scratch using the same model architecture and evaluated on the original clean test set. Evaluation metrics include overall classification accuracy, per-class accuracy, and confusion matrix analysis.

## 4. Experiments

This section presents a series of controlled experiments to evaluate the robustness of a VGG-based convolutional neural network (CNN) under various training-time corruption scenarios. By systematically isolating label-level and input-level perturbations, we examine how different forms of supervision noise and input distortion affect the model's generalization ability. All experimental conditions share an identical architecture and evaluation protocol to ensure a fair comparison.

### 4.1. Experimental Setup

All experiments are conducted on the CIFAR-10 dataset using the VGG-based CNN model described in Section 3. Each model is trained from scratch for 100 epochs using the Adam optimizer with a learning rate of $1 \times 10^{-3}$ and a batch size of 128. To ensure consistency across conditions, no early stopping or learning rate scheduling is applied.

Evaluation is performed on the clean CIFAR-10 test set based on three metrics: overall classification accuracy, per-class accuracy, and confusion matrix analysis. For consistency, all test-time inputs are identically preprocessed.

Training dynamics are visualized through loss and accuracy curves over epochs, while confusion matrices are row-normalized to aid interpretability.

### 4.2. Baseline

Under standard training conditions with clean images and correct labels, the model achieves strong performance and exhibits stable learning dynamics. As shown in Figure 1, the training loss decreases steadily while the test loss plateaus early, suggesting convergence without significant overfitting. The final training and test accuracies are 99.63% and 89.36%, respectively.

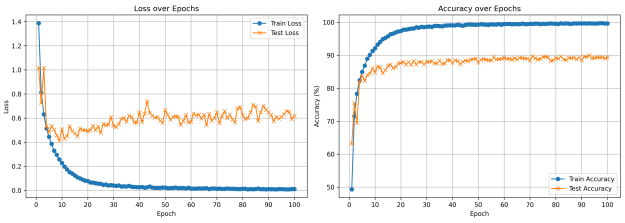


Figure 1. Training and test loss/accuracy curves for the Baseline condition.

The confusion matrix in Figure 2 confirms the model's ability to generalize well to all classes. Most predictions align closely with the true labels, as indicated by the strong diagonal pattern. Minor confusion occurs between semantically similar classes such as *cat* and *dog*, which is consistent with known challenges in CIFAR-10.

These results establish a strong reference point for evaluating the effects of label and input perturbations in subsequent sections.

### 4.3. Random Label Shuffle

Under the random label shuffle condition, each training label is reassigned independently of the corresponding input image. This eliminates any semantic relationship between inputs and outputs, making meaningful learning theoretically impossible.

As shown in Figure 3, both training and test loss remain nearly constant throughout training, and accuracy fluctuates around the expected chance level of 10%, indicating the model's inability to extract useful patterns. This behavior is consistent with prior findings that deep networks can memorize noise without generalizing [1].

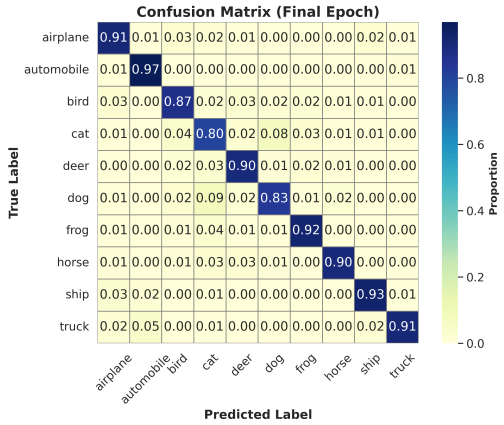


Figure 2. Confusion matrix for the Baseline model.

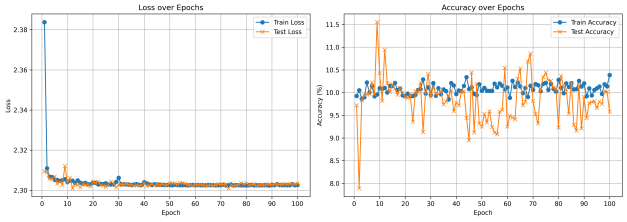


Figure 3. Training and test loss/accuracy curves for the Random Label Shuffle condition.

The confusion matrix in Figure 4 further confirms this failure. Predictions are nearly uniform across classes regardless of the input, and the diagonal structure observed in the baseline setting is absent. Most samples are classified as *automobile*, suggesting a collapse to a biased but non-informative prediction mode.

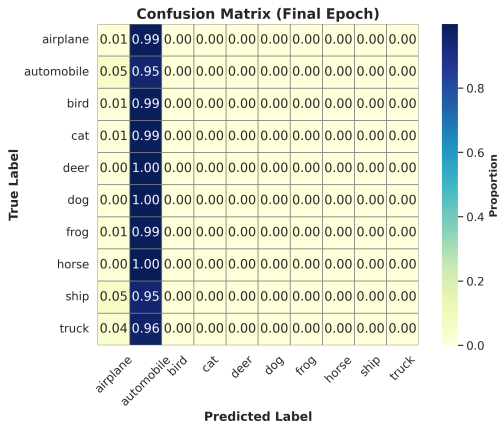


Figure 4. Confusion matrix for the Random Label Shuffle condition.

These results underscore the importance of preserving the semantic integrity of labels in supervised learning and highlight the model's inability to generalize when label supervision is entirely disrupted.

## 4.4. Label Noise (20%)

In this setting, 20% of the training labels are randomly replaced with incorrect values uniformly sampled from the remaining classes. Unlike the random label shuffle condition, most of the supervision remains valid, enabling partial learning while introducing semantic inconsistency.

As shown in Figure 5, the model is able to learn useful representations despite the noise. The training accuracy reaches 99.13%, indicating that the model still fits the noisy training data. However, test accuracy drops to 76.87%, reflecting a clear degradation in generalization due to corrupted labels. The test loss plateaus early and fluctuates throughout training, suggesting unstable validation behavior compared to the baseline.

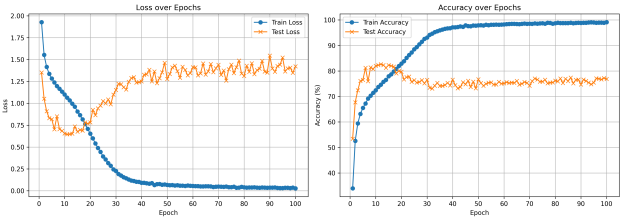


Figure 5. Training and test loss/accuracy curves for the Label Noise (20%) condition.

The confusion matrix in Figure 6 illustrates this performance drop more clearly. While the diagonal remains dominant, off-diagonal values are notably stronger than in the baseline, especially for visually similar classes such as *cat*, *dog*, and *deer*. This indicates that the model is more prone to inter-class confusion under noisy supervision.

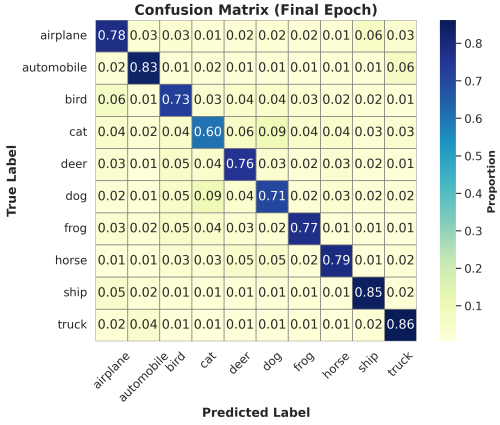


Figure 6. Confusion matrix for the Label Noise (20%) condition.

These results suggest that partial label noise does not completely prevent learning but introduces confusion in decision boundaries, especially for semantically overlapping categories.

### 4.5. Input Perturbation

In this condition, the input images are augmented using a combination of transformations including random resized cropping, horizontal flipping, color jittering, Gaussian blurring, and grayscale conversion. The goal is to simulate realistic acquisition noise such as lens blur, lighting variation, and framing inconsistency, while keeping the labels unchanged.

As shown in Figure 7, the model demonstrates stable training behavior. The training loss steadily decreases, and test accuracy improves beyond the baseline, reaching 90.32%. This suggests that the input perturbations act as an effective regularizer, enhancing generalization by encouraging the model to learn more robust feature representations.

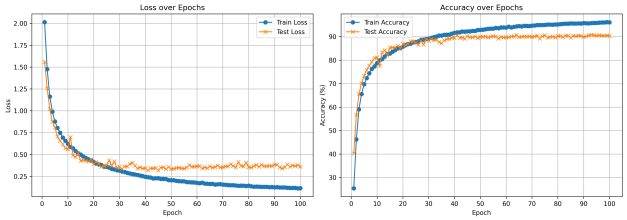


Figure 7. Training and test loss/accuracy curves for the Input Perturbation condition.

The confusion matrix in Figure 8 supports this finding. A strong diagonal structure is maintained, similar to the baseline, with slightly reduced off-diagonal noise even for visually similar classes. This indicates improved class separability and resilience to natural image variation.

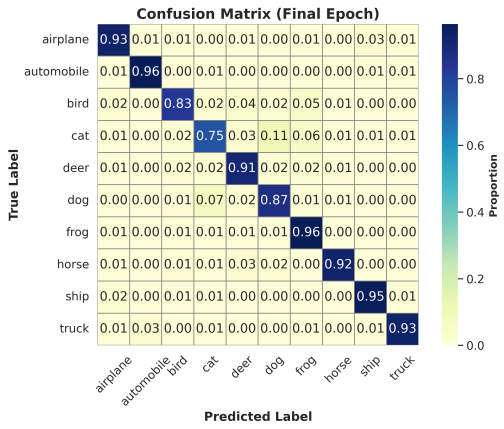


Figure 8. Confusion matrix for the Input Perturbation condition.

These results highlight that input-level distortions, when applied appropriately during training, can improve generalization performance and robustness without compromising class-level precision.

### 4.6. Summary and Comparative Insights

Table 1 summarize the overall performance across all training conditions. The results reveal striking differences in how label-level and input-level corruptions affect model learning and generalization.

| Condition | Train Acc. (%) | Test Acc. (%) |
|---|---|---|
| Baseline | 99.63 | 89.36 |
| Label Noise | 99.13 | 76.87 |
| Random Shuffle | 10.38 | 9.58 |
| Input Perturb. | 96.14 | 90.32 |

Table 1. Training and test accuracy under different corruption settings.

Random label shuffling results in a complete collapse of generalization, with both training and test accuracy near chance level. This confirms that the model cannot learn meaningful patterns when the supervisory signal is entirely disconnected from the input.

In contrast, when only a portion of labels are corrupted (Label Noise), the model still achieves high training accuracy but suffers a notable drop in test performance, especially on visually similar classes. This suggests that partial semantic inconsistency in supervision distorts decision boundaries without fully preventing learning.

Surprisingly, input perturbation slightly improves generalization performance compared to the baseline. This supports the notion that input-level distortions, when applied systematically, serve as an effective form of regularization by encouraging the model to focus on invariant features.

Overall, these findings demonstrate that models are far more sensitive to label noise than to input perturbations, and that not all forms of data corruption are equally harmful. Designing robust training pipelines therefore requires an understanding of which types of corruption the model can tolerate—and which it cannot.

## 5. Conclusion

This paper presents a systematic analysis of how training-time corruptions affect the generalization performance of a CNN classifier on CIFAR-10. Through controlled experiments involving label noise, random label shuffling, and input perturbations, we observe markedly different model behaviors depending on the type of corruption.

The results show that semantic label noise, even when limited to a fraction of the data, significantly impairs gen-

eralization by distorting decision boundaries. In contrast, input perturbations preserve or even enhance test performance, likely due to their regularization effects. When labels are randomly shuffled, eliminating all semantic alignment, the model completely fails to generalize.

These findings highlight the asymmetric effects of corruption sources in supervised learning, emphasizing that robustness to input noise does not imply robustness to label noise. Future work may explore strategies that explicitly address supervision quality, including noise-aware loss functions, semi-supervised refinement, or human-in-the-loop correction.

## References


[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Int. Conf. Learn. Represent.*, 2017. 2, 3

[2] Eric Arazo, Diego Ortego, Noel E. Albert, Kevin O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, 2019. 2

[3] Tianyu Wu, Xiaoyu Ding, Ming Tang, Hongyu Zhang, Bing Qin, and Ting Liu. Noisywikihow: A benchmark for learning with real-world noisy labels in natural language processing. *arXiv preprint arXiv:2305.10709*, 2023. 2

[4] Awais Saeed, Dimitrios Spathis, Jihye Oh, Euiyoung Choi, and Ali Etemad. Learning under label noise through few-shot human-in-the-loop refinement. *Scientific Reports*, 15:4276, 2025. 2

[5] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Int. Conf. Learn. Represent.*, 2019. 2

[6] Yan Jiang, Kimin Lee, Shiyu He, and Zhaoyang Zhang. Post-hoc ood detection via perturbation-rectified confidence. *arXiv preprint arXiv:2503.18784*, 2025. 2

[7] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. 2

[8] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv preprint arXiv:2007.00644*, 2020. 2