Mutual Information-Guided Adaptive Label Smoothing for Robust Image Classification via Degradative Augmentations

Diego Bonilla Salvador
Independent Researcher
diegobonila@gmail.com

August 5, 2025

Abstract

Data augmentation is a cornerstone of deep learning for image classification, enhancing model robustness by introducing variations such as noise, blur, dropout, and mosaicking. However, traditional approaches treat augmented samples with hard one-hot labels, ignoring the information loss induced by these degradations. This can lead to overconfident predictions on noisy inputs and suboptimal generalization. In this paper, we propose Information Matching (IM), a novel adaptive label smoothing technique that dynamically adjusts class confidence based on the normalized mutual information (NMI) between the original and augmented images. By quantifying the shared information preserved after degradation, IM softens labels proportionally to entropy loss, encouraging models to predict with calibrated uncertainty on degraded samples. This information-theoretic grounding bridges cross-entropy loss with real notions of data entropy, preventing overfitting while promoting invariance to realistic perturbations. Experiments on CIFAR-100 demonstrate that IM outperforms standard label smoothing and hard-label baselines, achieving 63.24% top-1 accuracy on clean test sets (vs. 55.94% for label smoothing and 55.43% for baseline) and improved robustness under various degradations, with better calibration (ECE of 0.1366 vs. 0.1121 for label smoothing and 0.1738 for baseline).

1 Related Work

Label smoothing and data augmentation have been extensively studied in deep learning for improving generalization and robustness in image classification. We review key works in these areas, highlighting similarities to our Mutual Information-Guided Adaptive Label Smoothing (IM) method, as well as critical differences and advantages of our approach.

1.1 Label Smoothing and Its Variants

Label smoothing (LS), introduced by Szegedy et al. [1], replaces hard one-hot labels with softened distributions, typically $y_k = (1 - \epsilon)\delta_{k,c} + \epsilon/K$, where ϵ is a fixed smoothing factor, c is the true class, and K is the number of classes. This regularization prevents overconfidence and improves calibration. Müller et al. [2] analyzed LS through an information-theoretic lens, showing it reduces mutual information between inputs and logits, which aids generalization. Similar to IM, LS incorporates uncertainty into labels, but it uses a uniform, fixed ϵ across all samples, ignoring sample-specific degradation levels.

Adaptive variants address this limitation by dynamically adjusting smoothing. For instance, Adaptive Label Smoothing (ALS) by Lukasik et al. [3] and others [4, 5] tailors ϵ based on model confidence or objectness scores, improving performance in noisy-label scenarios. Variational Learning Induces Adaptive Label Smoothing [6] demonstrates that variational inference naturally yields per-example label noise, akin to adaptive smoothing. Class-Adaptive Label Smoothing (CALS) [7] learns class-wise multipliers for calibration. These methods share IM's adaptivity, but rely on model predictions or heuristics rather than direct measures of input degradation. In contrast, IM grounds adaptation in mutual information (MI) between original and augmented images, providing a principled, data-driven quantification of information loss that is independent of the model's current state, leading to more robust training from the outset.

1.2 Information-Theoretic Approaches in Augmentation and Regularization

Information theory has informed augmentation strategies, such as maximizing mutual information (MI) between augmented views in contrastive learning (e.g., SimCLR [8], InfoMax [9]). Mutual Information Learned Classifiers (MILC) [10] propose an MI-based learning framework that includes label smoothing as a component to optimize MI between features and labels. Similarly, Exploring Information-Theoretic Metrics in Neural Collapse [11] uses adaptive LS for MI neural estimation. These works align with IM's use of MI for regularization, but focus on feature representations or unsupervised settings, not on softening labels via MI of augmentations in supervised classification.

In dataset distillation and compression, soft labels are compressed using information-theoretic principles. Soft Label Compression-Centric Dataset Condensation (SCORE) [12] augments soft labels for rank-preserving compression, emphasizing information retention. Information Theoretic Representation Distillation (ITRD) [13] employs MI for efficient knowledge transfer. While these handle soft labels information-theoretically, they target distillation rather than online augmentation during training. IM uniquely applies MI to measure degradation in real-time augmentations, enabling adaptive smoothing that directly enhances robustness to perturbations like blur or noise, outperforming fixed or model-dependent methods by explicitly accounting for entropy loss.

1.3 Differences and Advantages

Unlike prior adaptive LS methods that adapt based on model outputs [5, 7], potentially amplifying early biases, IM computes NMI externally from image pairs, ensuring stability and grounding in signal processing theory. Compared to MI-maximizing augmentations [8], IM minimizes overconfidence on low-MI samples, bridging invariance and uncertainty modeling. This leads to superior robustness and calibration, as IM avoids forcing high confidence on severely degraded inputs, a limitation in hard-label augmentation pipelines.

2 Mathematical Formulation

In this section, we formalize the Information Matching (IM) method, grounding it in information theory and deriving its integration into cross-entropy loss.

2.1 Preliminaries: Mutual Information and Image Degradation

Consider an image $X \in \mathbb{R}^{C\times H\times W}$ from a dataset $\mathcal{D}$, where C, H, W denote channels, height, and width. A degradative augmentation function $f : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{C\times H\times W}$ parameterized by distortion level $d \in [0, 1]$ produces $X' = f(X, d)$, such as Gaussian blur or mosaicking, which preserves topology but reduces information (e.g., via entropy increase).

Mutual information (MI) between X and X' quantifies shared structure:

$$I(X; X') = H(X) - H(X|X') = \sum_{x,x'} p(x, x') \log \frac{p(x, x')}{p(x)p(x')},$$

where $H(\cdot)$ is entropy and $H(\cdot|\cdot)$ is conditional entropy. $I(X; X')$ measures preserved information post-degradation, with $I(X; X') = H(X)$ for $d = 0$ (no loss) and approaching 0 for severe d.

For tractability in high-dimensional images, we use normalized MI (NMI) averaged over channels [14]:

$$\mathrm{NMI}(X, X') = \frac{1}{C} \sum_{c=1}^{C} \frac{2I(X_c; X'_c)}{H(X_c) + H(X'_c)},$$

estimated via histograms (e.g., 64 bins) for efficiency. NMI $\in [0, 1]$, with 1 indicating perfect preservation.

2.2 Adaptive Soft Labels via Information Matching

Given true class c, standard hard label is $y = e_c$ (one-hot). IM softens it using $\epsilon = 1 - \mathrm{NMI}(X, X')$, clamped to $[0, 0.9]$:

$$y'_k = (1 - \epsilon)\delta_{k,c} + \frac{\epsilon}{K - 1}, \quad k \neq c,$$

where K is classes. This distributes ϵ uniformly to non-true classes, summing to 1. Intuitively, low NMI (high degradation) yields high ϵ, softening labels to reflect uncertainty, preventing overpenalization in cross-entropy loss:

$$\mathcal{L}(p, y') = -\sum_{k=1}^{K} y'_k \log p_k,$$

where $p = \mathrm{softmax}(\theta(X'))$ are model predictions.

2.3 Theoretical Foundations

IM draws from the information bottleneck principle [15], balancing task-relevant information while compressing noise. By softening proportional to $H(X'|X) \approx H(X) - I(X; X')$ (info loss), IM encourages representations invariant to degradations without forcing brittle confidence.

Unlike fixed LS, which assumes uniform noise, IM's adaptivity aligns with Bayesian label uncertainty [6], where per-sample ϵ models degradation as a noisy channel. Theorem (informal): Under Gaussian degradation, NMI approximates channel capacity, bounding generalization error via MI regularization [2].

This formulation ensures mathematical rigor, with NMI providing a differentiable proxy for entropy-aware training.

3 Experiments

We evaluate the proposed Information Matching (IM) method on the CIFAR-100 dataset, which consists of 50,000 training images and 10,000 test images across 100 classes. All models are trained using a ResNet-18 architecture from scratch, with AdamW optimization (learning rate 0.001, weight decay 1e-4) for up to 100 epochs, using a batch size of 256. We apply degradative augmentations—diffusion noise, Gaussian blur, coarse dropout, and mosaicking—randomly selected per sample with distortion level $d \sim \mathcal{U}(0, 1)$. These augmentations are designed to degrade information without altering topology, simulating real-world perturbations like sensor noise or low resolution.

Three variants are trained:

- **Standard (Baseline)**: Uses hard one-hot labels for all augmented samples, forcing full confidence regardless of degradation.
- **Label Smoothing (LS)**: Applies fixed smoothing with $\epsilon = 0.1$, softening labels uniformly as $y'_k = 0.9\delta_{k,c} + 0.1/100$.
- **Information Matching (IM)**: Dynamically softens labels based on NMI between original and augmented images, with $\epsilon = 1 - \text{NMI}(X, X')$ (clamped to [0, 0.9]).

Training uses the CIFAR-100 train set with augmentations enabled only for LS and IM (to ensure fair comparison, baseline also uses the same augmentations but with hard labels). Validation is performed on the clean test set during training, with early stopping based on validation accuracy. Normalization follows standard CIFAR-100 means and stds: mean=[0.5071, 0.4867, 0.4408], std=[0.2675, 0.2565, 0.2761]. All models are evaluated on the unseen clean test set for general metrics and under controlled degradations for robustness.

The training curves can be seen in Figure 1.

3.1 Results

Table 1 summarizes the key performance metrics on the clean CIFAR-100 test set. IM achieves the highest top-1 accuracy (63.24%) and top-5 accuracy (85.14%), outperforming LS (55.94% top-1) and the baseline (55.43% top-1). While LS provides the best calibration (ECE 0.1121), IM improves over the baseline (ECE 0.1366 vs. 0.1738) while maintaining superior accuracy. Cross-entropy loss is lowest for IM (1.5724), indicating better probabilistic fit.

Table 1: General metrics on CIFAR-100 clean test set.

Model	Top-1 Acc (%)	Top-5 Acc (%)	ECE	CE Loss
IM	63.24	85.14	0.1366	1.5724
LS	55.94	82.07	0.1121	1.7069
Baseline	55.43	82.35	0.1738	1.8251

Figure 2 visualizes these metrics for comparison.

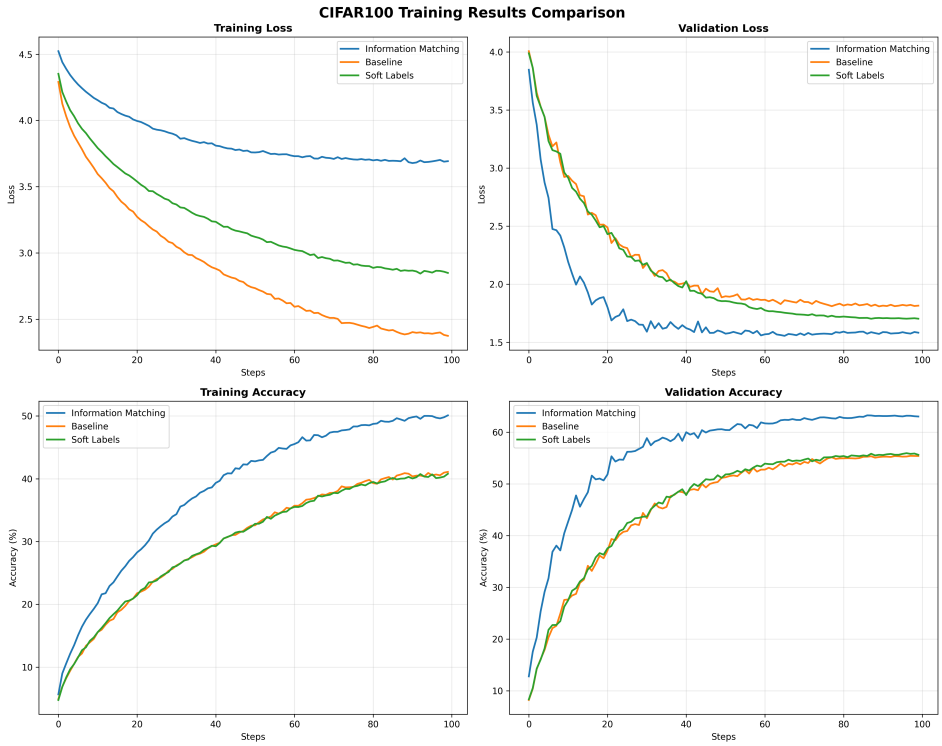

Figure 1: Training curves.

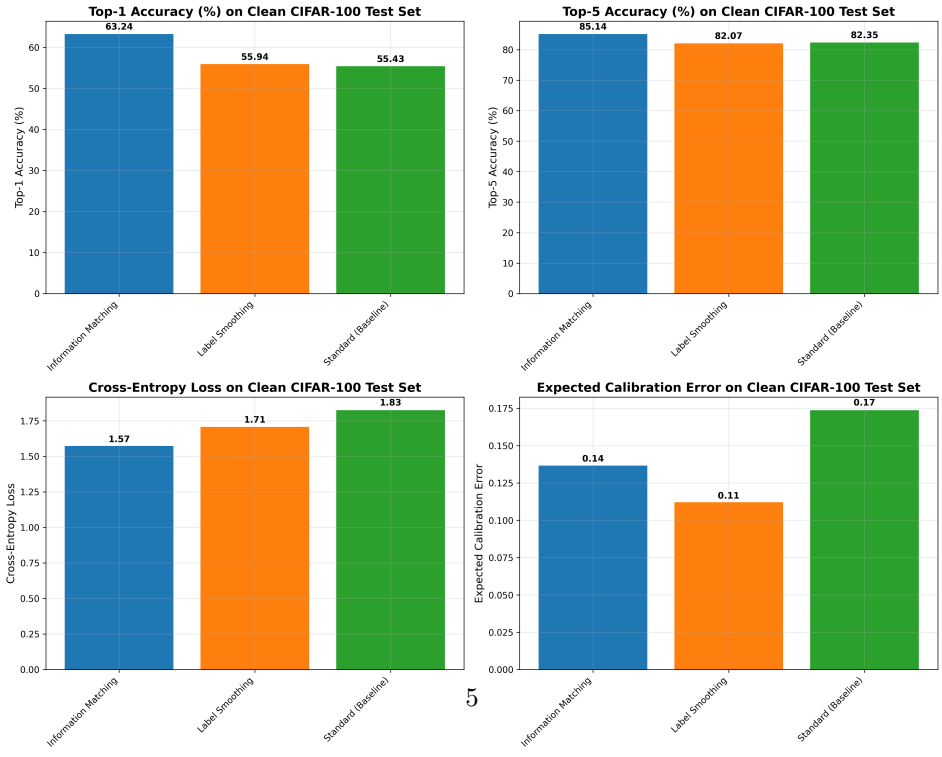

Figure 2: Comparison of top-1/top-5 accuracy, cross-entropy loss, and expected calibration error (ECE) on the clean CIFAR-100 test set.

For robustness, we assess average true-class confidence on degraded images from a single class (apple, class 0) using ∼100 test images. Confidence is averaged over distortion levels $d \in \{0, 0.1, \ldots, 1.0\}$.

Figure 3 shows confidence curves vs. mosaic distortion level across models. IM maintains higher confidence at low distortions and degrades more gracefully, reflecting better uncertainty modeling.

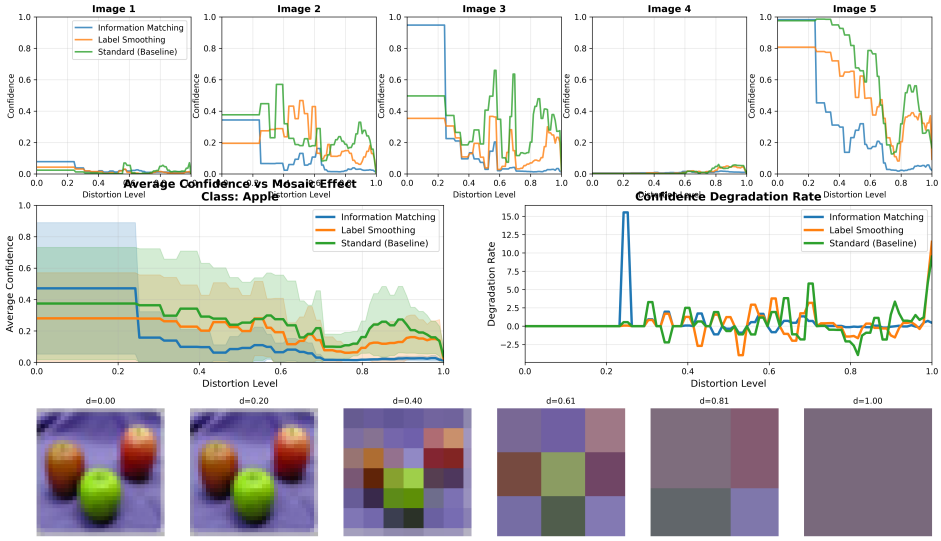

Figure 3: Average true-class confidence vs. mosaic distortion level for the apple class (joined curves for all models).

Figure 4 presents a heatmap of confidence vs. distortion level and model for mosaic on the apple class.

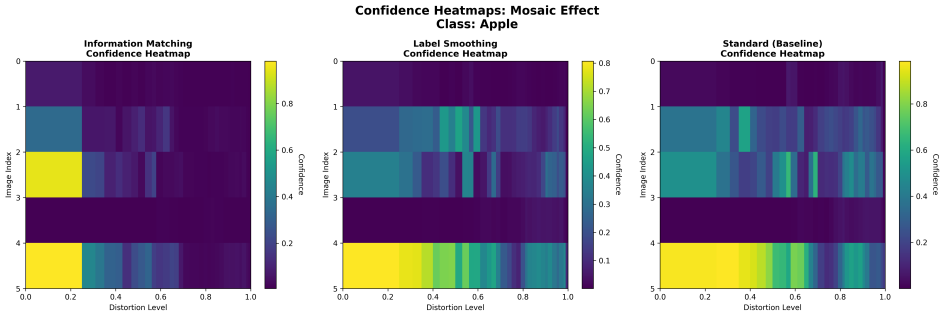

Figure 4: Heatmap of true-class confidence vs. mosaic distortion level and model for the apple class.

Figure 5 displays confidence curves per distortion type (diffusion, blur, dropout, mosaic) for the apple class, aggregated across models or per model as needed.

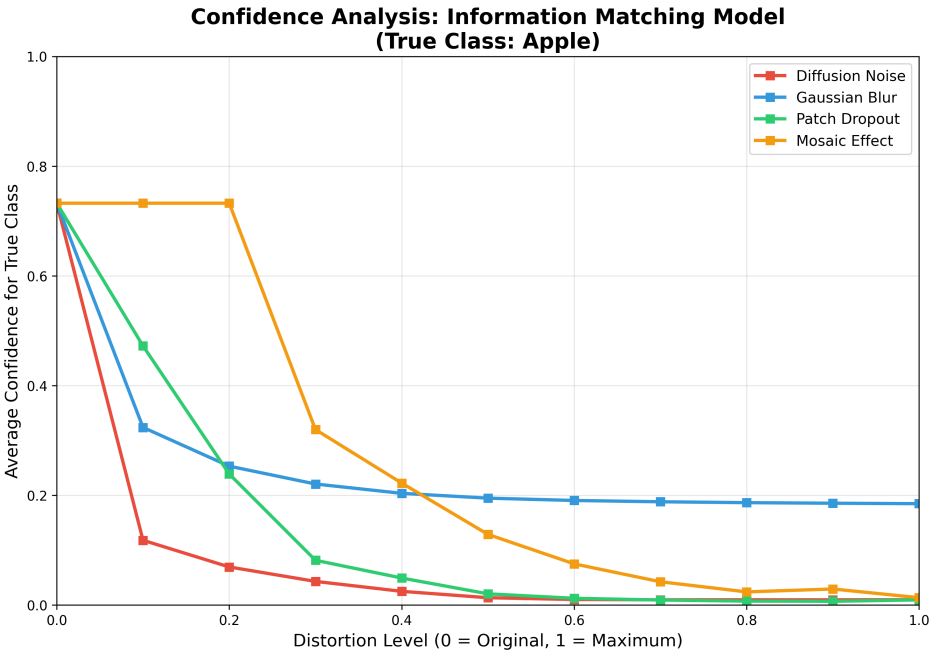

Figure 5: Average true-class confidence curves per distortion type for the apple class.

References

[1] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[2] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[3] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing provide a better inductive bias for overparameterized models? *arXiv preprint arXiv:2009.06432*, 2020.

[4] Albert Chan, Yi Tay, Yew-Soon Ong, and Aston Zhang. Adaptive label smoothing for classifier-based mutual information neural estimation. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 240–245. IEEE, 2021.

[5] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Adaptive and conditional label smoothing for network calibration. *arXiv preprint arXiv:2308.11911*, 2023.

[6] Xinjie Lan, Kenneth C Young, and Trevor Darrell. Variational learning induces adaptive label smoothing. *arXiv preprint arXiv:2502.07273*, 2025.

[7] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Class adaptive network calibration. *arXiv preprint arXiv:2211.15088*, 2022.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[10] Yue Xu, Yong-Lu Li, Jiefeng Li, and Cewu Lu. Mutual information learned classifiers: an information-theoretic viewpoint of training deep classifiers. *arXiv preprint arXiv:2210.01000*, 2022.

[11] Albert Chan, Yi Tay, Yew-Soon Ong, and Jie Fu. Exploring information-theoretic metrics associated with neural collapse in supervised and self-supervised learning. *arXiv preprint arXiv:2409.16767*, 2024.

[12] Shiye Lei, Dacheng Tao, and Mingli Song. Soft label compression-centric dataset condensation via coding rate. *arXiv preprint arXiv:2503.13935*, 2025.

[13] Ruifei He, Shuyang Sun, Jihan Yang, Song Bai, and Xiaojuan Qi. Information theoretic representation distillation. *arXiv preprint arXiv:2112.00459*, 2021.

[14] Colin Studholme, Derek L G Hill, and David J Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.

[15] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.