AUGLOSS: A Learning Methodology for Real-World Dataset Corruption

Kyle Otstot, John Kevin Cava, Tyler Sypherd, Lalitha Sankar
Arizona State University, Tempe, Arizona, 85281
{kotstot, jcava, tsypherd, lsankar}@asu.edu

Abstract—Deep Learning (DL) models achieve great successes in many domains. However, DL models increasingly face safety and robustness concerns, including noisy labeling in the training stage and feature distribution shifts in the testing stage. Previous works made significant progress in addressing these problems, but the focus has largely been on developing solutions for only one problem at a time. For example, recent work has argued for the use of tunable robust loss functions to mitigate label noise, and data augmentation (*e.g.,* AUGMIX) to combat distribution shifts. As a step towards addressing both problems simultaneously, we introduce AUGLOSS, a simple but effective *methodology* that achieves robustness against both train-time noisy labeling and test-time feature distribution shifts by unifying data augmentation and robust loss functions. We conduct comprehensive experiments in varied settings of real-world dataset corruption to showcase the gains achieved by AUGLOSS compared to previous state-of-the-art methods. Lastly, we hope this work will open new directions for designing more robust and reliable DL models under real-world corruptions. The GitHub link to the paper and code repository is: https://github.com/SankarLab/AugLoss.

1 INTRODUCTION

DEEP learning (DL) models achieve great successes in many domains. With such great successes, DL models have been deployed in many applications, even safety-critical applications (*e.g.,* autonomous driving and healthcare). Modern DL models rely heavily on the ability of training data to estimate and represent the data encountered during deployment. However, such a design introduces problems in both the training and testing stages of the machine learning (ML) pipeline. In the training stage, data curating procedures are often imperfect, leading to errors in the labeling process. For instance, recent results highlight that the amount of label noise in publicly-available image datasets is $\approx$ 8% to 38.5% [1]. This, in turn, affects the robustness and reliability of the DL models.

In the testing stage, the trained and deployed model often encounters new scenarios, thereby introducing a mismatch between the train and test distributions. State-of-the-art (SoTA) DL models exhibit overconfident predictions when the train and test sets are independent and identically distributed [2], which leaves them particularly vulnerable to test-time feature distribution shifts [3]. For instance, [4] shows that adding small common corruptions (*e.g.,* snow, blue, pixelation) to the test images are enough to subvert existing classifiers; specifically, the test error of a ResNet-50 model rises from 24% on ImageNet to 76% on ImageNet-C.

The above two problems threaten the trustworthiness and adoption of ML algorithms in safety and security-critical domains; therefore, it is critical to develop techniques that robustify ML models under both noisy labeling in the training stage and feature distribution shift in the testing stage. There has been steady progress made in addressing these problems, but focus has largely been on developing solutions for only one problem at a time. For instance, a range of methods including meta-learning [5], [6], loss correction [7], and parameterized loss functions [8], [9], [10], [11] have been proposed to robustify models against noisy labeling. However, it is unclear that such methods may continue to be robust in the face of a test-time distribution shift. Similarly, many methods have been shown to improve a model's generalization ability against feature distribution shift, including data augmentation [12], [13], adversarial learning [14], and pre-training [15], [16]. However, such methods have only been considered when the dataset is exclusively corrupted by distribution shifts; their ability to handle noisy training labels are unknown and unexplored.

In this work, we highlight the necessity of considering *both* problems simultaneously while designing robust DL models for real-world systems. In doing so, we present a robust methodology to tackle corruptions in both training (*e.g.,* noisy labeling) and testing (*e.g.,* feature distribution shift) stages. To this end, we select two representative methods, namely, robust (parameterized) loss functions and data augmentation, that have been often used to address training- and testing-stage corruptions, respectively. Unifying data augmentation techniques with robust loss functions, we propose AUGLOSS as a simple, yet effective *methodology* to enhance robustness against both train-time noisy labeling and test-time distribution shifts.

We consider three realizations of AUGLOSS by pairing each of three well-studied robust loss functions – focal loss [17], normalized cross entropy + reverse cross entropy (NCE+RCE) [9], and α-loss [11] – with AUGMIX data augmentation [18]. For comparison with oft-used methods, our experiments also include the following: (i) cross entropy (CE) loss without AUGMIX, (ii) AUGMIX without a robust loss (*i.e.,* with CE), (iii) each of the three robust losses

without AUGMIX. We evaluate all of these options for augmentation-loss choices by conducting comprehensive experiments with the CIFAR-10/100 datasets. For each realization, we produce WideResNet [19] models by first training on varied settings of synthetic and human-annotated label noise [20], then testing on corrupted images of CIFAR-10/100 [4]. We summarize our key contributions below.

- To the best of our knowledge, our proposed *methodology*, AUGLOSS, is the first to combine data augmentation and robust loss functions, thus offering an effective solution to train-time noisy labeling and test-time distribution shift. More broadly, AUGLOSS provides a blueprint to study the efficacy of different augmentation and loss combinations.
- Our comprehensive experiments in varied settings of train-label and test-feature corruptions showcase the gains achieved by AUGLOSS compared to all other augmentation-loss combinations listed earlier. Our key contribution is in highlighting the efficacy of different combinations for different corruption scenarios. In particular, we show that AUGLOSS (AUGMIX + robust loss) methods consistently outperform the other combinations across every tested setting of label noise.
- Lastly, we observe that although AUGLOSS outperforms SoTA methods, our results also suggest that no single robust loss function within the AUGLOSS framework is a universal "best fit" across all tested settings of label noise. In particular, our results suggest that the focal loss performs well on CIFAR-100 corrupted by asymmetric noise, the NCE+RCE loss generally performs well under settings of high label noise (0.3,0.4), and the α-loss performs well on both CIFAR-10 and CIFAR-100 datasets when corrupted by symmetric label noise.

Our results, detailed in the sequel, highlight that our proposed AUGLOSS methodology can enhance the performance of DL models under both train-time noisy labeling and test-time feature distribution shifts. We believe this is an important step toward real-world robust DL systems. The rest of the paper is organized as follows. In Section 2, we review the literature of robust loss functions and data augmentation techniques. In Section 3, we present the background, problem setup, and formulation of both robust loss and data augmentation. Section 4 outlines the AUGLOSS framework, and Section 5 details our experimental setup, results, and discussion. Finally, we conclude in Section 6.

2 RELATED WORK

2.1 Loss Functions for Noisy Labeling

Using loss functions to mitigate noisy labels during training is a strong tradition in machine learning [7], [21], [22]. A common approach is the enhancement of the cross entropy loss with hyperparameters (and possibly additional regularizing terms) that enable the practitioner to tailor the loss to the desired application. In this vein, an important hyperparameterized loss is the generalized cross entropy (GCE), which was motivated by the Box-Cox transformation in statistics [10]. GCE was experimentally shown to be robust to noisy labels in neural-networks. A generalization of GCE, called α-loss, was motivated by Arimoto entropies arising in information theory [11]. The α-loss was experimentally shown to be robust to noisy labels during training across several algorithms (neural-networks, logistic regression, and boosting), along with theoretical support of these robustness characteristics [23]. Another important hyperparameterized loss is the focal loss [17]. While the focal loss was initially used for dense object detection with great success, it has recently received deeper theoretical scrutiny [23], [24], been observed to improve neural-network calibration [25], and also applied to the problem of noisy labels [9]. Lastly, an important hyperparameterized loss function for noisy labels is NCE+RCE, which stems from the framework of active-passive losses [9]. Unlike α-loss and focal loss, NCE+RCE employs two hyperparameters and is a linear combination of losses (see Section 3.2.1 for more details). Important for our purposes, NCE+RCE has been successfully employed in computer vision applications, particularly in the very high noise regime. Overall, α-loss, focal loss, and NCE+RCE have all been shown to be robust to label noise in the training data, and hence comprise a strong representative subset of the robust loss function literature (for more examples, see [26], [27], [28], [29]). However, to the best of our knowledge, each of these loss functions have not been previously considered in the joint setting of training *and* test domain shift, which we argue is the real-world scenario addressed by our proposed AUGLOSS methodology.

2.2 Data Augmentation for Domain Adaptation

The issue of domain shifts is an active field. In particular, DL models learning on train data may not generalize well to the test data if the train and test distributions are misaligned. For example, adding small corruptions (*e.g.,* snow, blue, pixelation) to test images nearly triples the generalization error of ResNet-50 models training on ImageNet [4]. Consequently, the field of domain adaption has received much attention, including the proposal of data augmentation techniques that help robustify models under domain shift. Specifically, Cutout, the occlusion of square input regions during training, was shown to improve generalization error on clean images [30]. Building on this, CutMix replaces the occluded regions with portions of different images, and yields better test performance on out-of-distribution examples [31]. Similarly, Mixup synthesizes information from two images by training on convex combinations of two feature-target pairs [32]. Taking a different approach, AutoAugment searches for improved data augmentation policies – successive operations (*e.g.,* translation, rotation) that preserve image semantics – achieving state-of-the-art validation accuracy on clean datasets [33]. Combining the previous two approaches, AUGMIX enhances the training set with mixtures of three (or less) augmented images, each generated by its own chain of stochastically-sampled operations [18]. Out of the proposed data augmentation methods, AUGMIX achieves the best performance for out-of-distribution testing; however, descendants of AUGMIX, namely AugMax [14] and NoisyMix [34], combine data augmentation with adversarial learning and stability training, respectively, to achieve even better results under domain shift. In this work, we consider the unification of data augmentation with robust loss functions – one that, to our best knowledge, is unexplored.

3 PRELIMINARIES

3.1 Image Classification

Consider a feature-label pair of random variables $(X, Y) \sim q_{X,Y}$ sampled over feature space $\mathcal{X}$ and label space $\mathcal{Y}$. In this, we acknowledge the existence of a ground-truth posterior distribution $q_{Y|X}$ where $q_{Y|X}(y|x)$ is the probability that feature $x \in \mathcal{X}$ is representative of label $y \in \mathcal{Y}$, and a ground-truth marginal (or prior) distribution q_X over the the feature space, where $q_X(x)$ is the probability that we observe $x \in \mathcal{X}$. The underlying distribution $q_{X,Y}$ is unknown, but we have access to some dataset serving as a finite representation of $q_{X,Y}$. We then look to develop a method that learns a best estimate $\hat{q}_{Y|X}$ of the posterior $q_{Y|X}$ given the observed dataset. An effective method in this context should work to minimize the difference between $\hat{q}_{Y|X}$ and $q_{Y|X}$.

Oftentimes, the inner workings of the data-generating process are inaccessible, and consequently we are only provided the outcome (*i.e.*, the observed dataset), which has encouraged the development of practical methods biased to several convenient assumptions about the dataset. For example, it is common to assume that all dataset samples are i.i.d: independent, and *identically distributed* according to the evaluation samples. This way, a model fitting on a subset of the data can generalize to the unforeseen examples reflective of the ground truth. However, the data-generating process is susceptible to particular obstacles that prevent the satisfaction of i.i.d. and consequently hinder the efficacy of learning methods depending on these assumptions. Unaddressed limitations of data collection can mislead the model's perspective of $q_{Y|X}$ and q_X – the two factors composing $q_{X,Y}$ – and ultimately welcome data drawn from some perturbation of the ground-truth. We note these limitations below.

3.1.1 Mislabeling of the Sampled Features

Supervised learning requires each observed feature in the dataset to be properly annotated, but the accuracy is often mitigated by factors including the subversion of experienced domain experts [35], the unreliable nature of crowdsourcing platforms [36], and the threat of adversarial label-flip attacks [37]. As a result, large-scale datasets are prone to noisy labeling, where a fraction of the true labels are flipped to false classes. In this case, the train labels are drawn from some perturbation $\tilde{q}_{Y|X}$ of the true posterior $q_{Y|X}$.

In response to the growing concern of noisy labeling, recent work has been done on synthetic label noise generation within clean datasets [38]. Standard methods consider both symmetric (random labels flipped to other uniform-random classes) and asymmetric (random labels flipped to other visually-similar classes) approaches to noise generation. Since label noise is frequently associated with human error, real-world examples of asymmetric generation have recently received much interest: CIFAR-10N and CIFAR-100N (*a.k.a.* CIFAR-*N), for instance, contain human-annotated noisy labeling for the CIFAR-10 and CIFAR-100 images, collected by Amazon Mechanical Turk [20].

DL models that exhibit robustness under label corruption (noise) should give accurate classifications even after fitting on a dataset with noisy labeling. Therefore, the standard approach to evaluating a model's robustness includes: (1) training on noisy labels, generated symmetrically or asymmetrically; and (2) testing on clean labels.

Fig. 1: The 15 common corruptions on a HORSE-labeled image, found in the CIFAR-10-C dataset.

3.1.2 Imperfect Sampling of the Feature Space

The data collection process requires each image in the dataset to be sampled randomly from the feature space. However, the execution of random sampling in practice is often challenged by several factors. For example, the feature-generating distribution may evolve over time and experience a distributional shift during deployment [2]. Additionally, image datasets tend to reflect built-in biases in which the curated features are drawn from a small subset of the underlying distribution [7]. In either case, a distribution mismatch exists between the training and testing stages, and as a result, the train features are sampled from a perturbation $\tilde{q}_X$ of the true prior q_X.

Recent work has been done to address scenarios of unforeseen data shifts in the prior distribution, including the generation of real-world feature noise on clean datasets. CIFAR-10-C and CIFAR-100-C (*a.k.a.* CIFAR-*C), for instance, use fifteen algorithmically-generated corruptions – including real-world noise, blur, weather, and digital categories – on the CIFAR-10 and CIFAR-100 datasets to generate a new image set reflective of data shifts potentially encountered in practice [4]. Figure 1 shows the fifteen corruptions on a CIFAR-10 image.

The classification ability of robust, real-world DL models should generalize to corrupted features after fitting on a clean dataset. Hence, the standard approach to evaluating a model's robustness under data shift should include: (1) training on clean features; and (2) testing on noisy features, generated by common corruptions.

3.2 Dataset Corruption & Proposed Remedies

In this paper, we consider a dataset to be corrupted if its train labels and features are drawn from a misaligned posterior $\tilde{q}_{Y|X}$ and misaligned prior $\tilde{q}_X$, respectively. With the increasing prevalence of such large-scale datasets, we argue that errors in the data collection process are inevitable,

so effective learning methods require a particular degree of robustness under realistic settings of dataset corruption. Our methodology is primarily motivated by previous work done on model robustness under label noise or feature noise, but not both. Specifically, we consider two proposed techniques that robustify models under one of the two noise types. The techniques are presented below.

3.2.1 Robust Loss Functions

The basic loss function $\ell : \mathcal{P}(Y|X) \times \mathcal{Y} \to \mathbb{R}^+$ maps an estimated posterior distribution and true class label to a performance measure with the intent of learning a posterior distribution $\hat{p}$ that minimizes the expectation of $\ell\left(\hat{p}(x), y\right)$ across all (x, y) pairs in the train set. For any reasonable label corruption $D_{\text{train}} \mapsto \tilde{D}_{\text{train}}$ to the train set, we generally expect a loss function ℓ to be *robust* when the following property holds:

$$\operatorname*{argmin}_{\hat{p}} \mathbb{E}_{\tilde{D}_{\text{train}}} \ell\left(\hat{p}(x), y\right) \approx \operatorname*{argmin}_{\hat{p}} \mathbb{E}_{D_{\text{train}}} \ell\left(\hat{p}(x), y\right). \quad (1)$$

However, in this paper, we relax the intuitive representation of "robust loss" by asserting the following: loss functions exhibit robustness if they achieve better performance, in settings of train-time noisy labeling, than the standard loss function in image classification, cross entropy (CE)

$$\ell_{\text{CE}}(\hat{p}, y) = -\log\left(\hat{p}(y)\right), \quad (2)$$

which has been shown to be non-robust under label noise [9]. Moreover, recent progress has been made in the formulation of robust *tunable loss function families*, including focal loss [39], active-passive loss [9], and α-loss [11]. Specifically, a loss function family ℓ_θ parameterized by θ in the space Θ is considered robust when there exists a tuning $\theta^* \in \Theta$ such that ℓ_{θ^*} is robust.

The **focal loss** family parameterized by $\gamma \in [0, 5]$,

$$\ell_{\text{FL}}(\hat{p}, y; \gamma) = -\left(1 - \hat{p}(y)\right)^\gamma \log\left(\hat{p}(y)\right), \quad (3)$$

is experimentally shown to perform better than CE loss under label noise when γ is tuned above 0 [39] ($\gamma = 0$ is equivalent to CE). The **NCE+RCE loss** family parameterized by $(\beta_1, \beta_2) \in \mathbb{R}^2_+$,

$$\ell_{\text{NCE+RCE}}(\hat{p}, y; \beta_1, \beta_2) = \beta_1 \cdot \ell_{\text{NCE}}(\hat{p}, y) + \beta_2 \cdot \ell_{\text{RCE}}(\hat{p}, y), \quad (4)$$

is an example of active-passive loss – a proposed group of loss function families that linearly combine an *active* loss function and *passive* loss function. In this case, the active loss is normalized cross entropy (NCE)

$$\ell_{\text{NCE}}(\hat{p}, y) = \frac{\ell_{\text{CE}}(\hat{p}, y)}{\sum_{y' \in \mathcal{Y}} \ell_{\text{CE}}(\hat{p}, y')}, \quad (5)$$

and the passive loss is reverse cross entropy (RCE)

$$\ell_{\text{RCE}}(\hat{p}, y; \delta) = \delta \sum_{y' \neq y} \hat{p}(y'), \ \delta > 0. \quad (6)$$

In the NCE+RCE paper, δ is fixed to 4. Lastly, the **α-loss** family parameterized by $\alpha \in (0, \infty]$,

$$\ell_\alpha(\hat{p}, y; \alpha) = \frac{\alpha}{\alpha - 1}\left(1 - \hat{p}(y)^{1 - 1/\alpha}\right), \quad (7)$$

encapsulates the exponential ($\alpha = 1/2$), cross entropy ($\alpha = 1$), and 0-1 ($\alpha \to \infty$) losses. This family is shown to perform better than CE loss under label noise when $\alpha > 1$ [11].

3.2.2 Data Augmentation

Motivated by previous work on data augmentation [40], [41], we generalize the augmenter $A : \mathcal{X} \to \mathcal{X}^n$ to return an n-tuple

$$A(x) = \left(x_{\text{orig}}, x_{\text{aug1}}, x_{\text{aug2}}, \ldots, x_{\text{aug}(n-1)}\right) \quad (8)$$

from a given feature $x \in \mathcal{X}$, where $x_{\text{orig}} := x$ and each $x_{\text{aug}(i)}$ is a unique transformation of x. The model then learns n distinct distributions

$$\hat{P} := \left(\hat{p}_{\text{orig}}, \hat{p}_{\text{aug1}}, \hat{p}_{\text{aug2}}, \ldots, \hat{p}_{\text{aug}(n-1)}\right), \quad (9)$$

where $\hat{p}_{\text{orig}}$ fits on the train set D_{train}, and each $\hat{p}_{\text{aug}(i)}$ fits on the dataset $\{(x_{\text{aug}(i)}, y) : (x, y) \in D_{\text{train}}\}$.

The use of augmentation warrants a loss supplement $\ell_2 : \mathcal{P}(Y|X)^n \to \mathbb{R}^+$ that synthesizes the information of each learned distribution in $\hat{P}$ into a measure of similarity. Assuming true class preservation within $A(x)$, an effective form of regularization will ensure that $\forall_{i \in [n-1]}\ \hat{p}_{\text{aug}(i)} \approx \hat{p}_{\text{orig}}$, which aims to improve the model's robustness under varied settings of feature corruption. Training with data augmentation makes use of the general loss function $\mathcal{L}$

$$\mathcal{L}(\hat{P}, y; \lambda) = \ell_1(\hat{p}_{\text{orig}}, y) + \lambda \cdot \ell_2(\hat{P}), \quad (10)$$

where ℓ_1 is a basic loss function (*e.g.,* CE loss) and ℓ_2 is an augmentation regularizer scaled by some constant $\lambda \in \mathbb{R}^+$.

4 AugLoss Framework

Previously, we made note of the following:

1) Robustness under *label* corruption should be evaluated by training on noisy labels, then testing on clean labels.
2) Robustness under *feature* corruption should be evaluated by training models on clean features, then testing them on noisy features.

Motivated by robustness under real-world dataset corruption, we choose to unify the above two statements and propose that robust DL models should perform well on the *novel task* of training on noisy labels and clean features, followed by testing on clean labels and noisy features. In this section, we introduce AugLoss: a learning methodology that combines the known techniques of data augmentation and robust loss functions, formulating an effective solution to our novel task. The two stages of this methodology are illustrated in Figure 2 and outlined below.

Stage 1: Classification

The classification stage begins with a random sampling $(x, y) \in D_{\text{train}}$ of the train set; in practice, samples are done in batches, but for clarity of exposition we notate a single example. In the first step, data augmentation is performed with an augmenter-of-choice A. The augmenter runs the original image x through a number of distinct transformations, returning $x_{\text{orig}} := x$ along with the newly-transformed images. Appropriate transformations for the task-at-hand should be nontrivial: $x_{\text{aug}(i)} \neq x_{\text{orig}}$; uniquely generated: $x_{\text{aug}(i)} \neq x_{\text{aug}(j)}, \forall_{i \neq j}$; visually realistic: $q_{X = x_{\text{aug}(i)}} \approx q_{X = x_{\text{orig}}}$; preservative of class representations: $q_{Y|X = x_{\text{aug}(i)}} \approx q_{Y|X = x_{\text{orig}}}$; and computationally feasible.

Then, each image in $A(x)$ is fed through a network of sufficient capacity (relative to the dataset) to produce a tuple

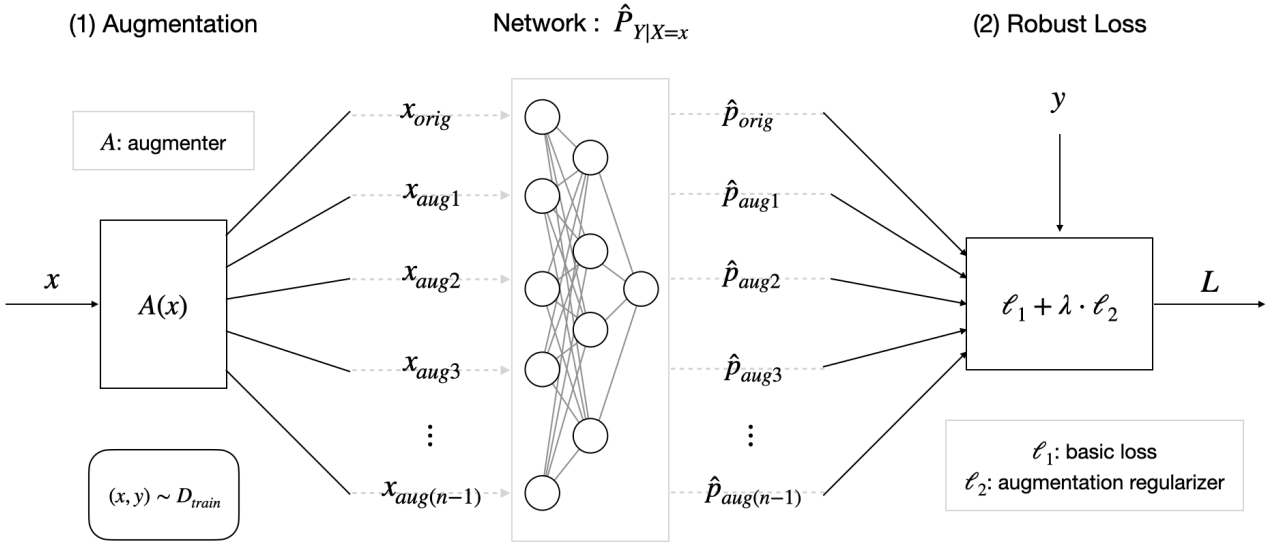


Fig. 2: AUGLOSS: the unification of data augmentation and robust loss functions.

of estimated posterior distributions. Each output $\hat{p}_{\text{aug}(i)}$ represents a distribution over the label space, conditioned on the augmented feature $x_{\text{aug}(i)}$.

Stage 2: Evaluation & Correction

The transformations undergone in the augmentation stage should work to preserve the true class representation of the original image, so an accurate classifier should embed each augmented image similarly to the original one. To ensure this, we train the model with the loss function $\mathcal{L}$ in Eq. 10, pairing a basic loss function $\ell_1(\hat{p}_{\text{orig}}, y)$ with an augmentation regularizer $\ell_2(\hat{P})$ that is minimized when the distributions in $\hat{P}$ are all the same. We set ℓ_1 to be an optimally-tuned instance of a robust loss function family.

As a result, the loss function takes the form $\mathcal{L} = \ell_1 + \lambda \cdot \ell_2$, where (1) the loss ℓ_1 works to robustify the model against label noise in the training stage, and (2) the regularizer ℓ_2 works to generalize the performance of the model to unforeseen data shifts in the testing stage. The regularization constant λ is fixed to a value that helps $\mathcal{L}$ balance between these two objectives. The optimization stage of the model searches for an estimated posterior distribution

$$\hat{p}^* := \underset{\hat{p}}{\operatorname{argmin}}\, \mathbb{E}_{D_{\text{train}}} \mathcal{L}\left((\hat{p}_{\text{orig}}, \hat{p}_{\text{aug1}}, ...), y; \lambda\right) \qquad (11)$$

by learning (on a noisy-labeled train set) the class representations invariant to common image corruptions.

5 EXPERIMENTS

In the context of our novel task, we showcase AUGLOSS against previously-accepted SoTA methods that have been shown to combat label or feature corruption. In doing so, we explore how several realizations of AUGLOSS perform on CIFAR-*-C after training on varied settings of synthetic noise in the traditional CIFAR datasets, or human-annotated noise in the CIFAR-*N datasets. In the next subsections, we detail the experimental protocol and provide brief justification for implementation choices.

5.1 Datasets & Metrics

Previous efforts have used synthetic approaches to noise generation – namely symmetric and asymmetric noise – to evaluate model robustness under label corruption. In the symmetric case, the probability of a label flip is distributed evenly among the false classes: if *A*, *B*, *C*, . . . , *J* are classes and label *A* were flipped to a different class, then *B*, *C*, . . . , *J* would be chosen with equal probability. Symmetric noise is a consequence of class-*independent* errors in the data-generating process – the same noise patterns are present no matter the difficulty in differentiating between classes. However, in crowdsourcing platforms (large-scale human-annotated labeling) inter-class visual similarity often influences the distribution of label noise in real-world datasets, giving rise to class-*dependent* asymmetric noise as an alternative synthetic approach.

Consistent with the standard, we train our models on the CIFAR-10 and CIFAR-100 datasets corrupted by varied levels of symmetric (Tables 1,2) and asymmetric (Tables 3,4) noise. Each level is defined by the noise rate (η), which is the probability that any given label is flipped. Our choice of noise rates $\eta \in \{0, 0.1, 0.2, 0.3, 0.4\}$ reflects the observation by [1] that approximately 8% to 38.5% of labels in publicly-available image datasets are noisy. For asymmetric noise, we adhere to the standard mappings [9] outlined below:

- **CIFAR-10:** some classes flip to other visually-similar ones – TRUCK $\mapsto$ AUTOMOBILE; BIRD $\mapsto$ AIRPLANE; DEER $\mapsto$ HORSE; CAT $\leftrightarrow$ DOG – and the rest simply flip to itself.
- **CIFAR-100:** we partition the label space of 100 classes into 20 superclasses[1], where each class belongs to a set containing four other visually-similar classes. Each class flips symmetrically to another member of its own superclass.

1. https://www.cs.toronto.edu/~kriz/cifar.html

Method		TABLE 1: **Symmetric** CIFAR-10					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	$\mathbf{27.57 \pm 0.61}$	35.79 ± 0.62	40.45 ± 1.37	45.19 ± 1.06	50.94 ± 0.34	43.09 ± 0.85
	Focal	33.81 ± 2.05	32.58 ± 0.95	37.98 ± 1.06	44.15 ± 0.22	50.47 ± 0.19	41.30 ± 0.61
	NCE+RCE	30.02 ± 0.30	30.42 ± 0.13	31.80 ± 1.20	34.20 ± 0.24	35.96 ± 1.43	33.10 ± 0.75
	α-loss	28.43 ± 0.47	$\mathbf{29.28 \pm 1.21}$	$\mathbf{30.23 \pm 0.81}$	$\mathbf{32.22 \pm 0.75}$	$\mathbf{35.16 \pm 0.51}$	$\mathbf{31.72 \pm 0.82}$
AUGMIX	CE	12.62 ± 0.40	15.14 ± 0.25	18.02 ± 0.11	21.23 ± 0.23	26.46 ± 0.59	20.21 ± 0.30
	Focal	13.81 ± 0.21	$\mathbf{12.59 \pm 0.05}$	13.91 ± 0.2	17.79 ± 0.46	23.15 ± 0.33	16.86 ± 0.26
	NCE+RCE	13.15 ± 0.16	13.72 ± 0.40	13.87 ± 0.09	14.19 ± 0.23	15.17 ± 0.12	14.24 ± 0.21
	α-loss	$\mathbf{12.55 \pm 0.09}$	12.82 ± 0.09	$\mathbf{13.23 \pm 0.13}$	$\mathbf{13.81 \pm 0.07}$	$\mathbf{15.10 \pm 0.27}$	$\mathbf{13.74 \pm 0.14}$

Method		TABLE 2: **Symmetric** CIFAR-100					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	53.84 ± 0.18	60.22 ± 0.35	64.93 ± 0.63	68.68 ± 0.23	72.74 ± 0.51	66.64 ± 0.43
	Focal	$\mathbf{53.50 \pm 0.35}$	57.85 ± 0.21	62.73 ± 0.46	67.18 ± 0.35	72.35 ± 0.49	65.03 ± 0.38
	NCE+RCE	54.66 ± 0.38	$\mathbf{55.34 \pm 0.54}$	56.91 ± 0.16	58.76 ± 0.60	61.92 ± 0.67	58.23 ± 0.49
	α-loss	54.43 ± 0.13	55.58 ± 0.24	$\mathbf{56.48 \pm 0.40}$	$\mathbf{58.11 \pm 0.61}$	$\mathbf{60.21 \pm 0.81}$	$\mathbf{57.60 \pm 0.52}$
AUGMIX	CE	37.80 ± 0.22	42.11 ± 0.19	44.86 ± 0.50	48.34 ± 0.18	51.65 ± 0.09	46.74 ± 0.24
	Focal	$\mathbf{36.11 \pm 0.2}$	38.55 ± 0.47	42.5 ± 0.26	46.97 ± 0.05	52.05 ± 0.09	45.02 ± 0.22
	NCE+RCE	41.53 ± 0.96	42.18 ± 0.10	42.84 ± 0.14	42.71 ± 0.46	$\mathbf{44.17 \pm 0.80}$	42.98 ± 0.38
	α-loss	37.66 ± 0.13	$\mathbf{38.46 \pm 0.26}$	$\mathbf{40.03 \pm 0.50}$	$\mathbf{41.90 \pm 0.51}$	44.54 ± 0.30	$\mathbf{41.23 \pm 0.39}$

Method		TABLE 3: **Asymmetric** CIFAR-10					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	$\mathbf{26.84 \pm 0.10}$	30.19 ± 0.31	32.77 ± 0.51	35.32 ± 0.11	38.42 ± 0.45	34.18 ± 0.35
	Focal	33.76 ± 2.00	32.73 ± 1.28	$\mathbf{29.33 \pm 1.01}$	$\mathbf{31.48 \pm 0.27}$	$\mathbf{33.99 \pm 0.86}$	$\mathbf{31.88 \pm 0.86}$
	NCE+RCE	30.02 ± 0.30	29.68 ± 0.70	30.57 ± 0.51	32.01 ± 1.06	36.89 ± 0.20	32.29 ± 0.62
	α-loss	28.98 ± 0.85	$\mathbf{28.79 \pm 0.21}$	30.46 ± 0.51	32.98 ± 0.36	39.09 ± 0.78	32.83 ± 0.47
AUGMIX	CE	12.67 ± 0.38	13.42 ± 0.36	14.97 ± 0.37	16.63 ± 0.57	20.27 ± 0.70	16.32 ± 0.50
	Focal	13.81 ± 0.21	13.64 ± 0.15	14.17 ± 0.30	15.55 ± 0.34	20.28 ± 0.84	15.91 ± 0.41
	NCE+RCE	13.13 ± 0.15	13.45 ± 0.11	13.86 ± 0.17	$\mathbf{14.74 \pm 0.06}$	18.29 ± 0.52	15.08 ± 0.22
	α-loss	$\mathbf{12.55 \pm 0.09}$	$\mathbf{12.84 \pm 0.2}$	$\mathbf{13.44 \pm 0.08}$	14.88 ± 0.33	$\mathbf{17.95 \pm 0.36}$	$\mathbf{14.78 \pm 0.24}$

Method		TABLE 4: **Asymmetric** CIFAR-100					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	53.92 ± 0.06	59.20 ± 0.32	63.11 ± 0.09	66.58 ± 0.38	69.98 ± 0.22	64.72 ± 0.25
	Focal	$\mathbf{53.53 \pm 0.43}$	57.15 ± 0.40	60.28 ± 0.09	63.70 ± 0.67	67.64 ± 0.36	62.19 ± 0.38
	NCE+RCE	54.45 ± 0.05	$\mathbf{56.02 \pm 0.66}$	$\mathbf{57.19 \pm 0.26}$	$\mathbf{59.09 \pm 0.53}$	$\mathbf{61.34 \pm 0.26}$	$\mathbf{58.41 \pm 0.43}$
	α-loss	54.58 ± 0.12	56.07 ± 0.32	58.59 ± 0.25	61.41 ± 0.31	65.25 ± 0.26	60.33 ± 0.28
AUGMIX	CE	37.88 ± 0.29	41.14 ± 0.16	43.71 ± 0.12	46.37 ± 0.15	49.57 ± 0.44	45.20 ± 0.22
	Focal	$\mathbf{36.11 \pm 0.20}$	$\mathbf{37.64 \pm 0.18}$	$\mathbf{39.77 \pm 0.10}$	$\mathbf{42.37 \pm 0.16}$	$\mathbf{45.52 \pm 0.27}$	$\mathbf{41.32 \pm 0.18}$
	NCE+RCE	42.13 ± 0.30	42.81 ± 0.07	43.78 ± 0.15	44.77 ± 0.10	46.53 ± 0.45	44.47 ± 0.19
	α-loss	37.64 ± 0.12	39.21 ± 0.10	41.21 ± 0.07	43.27 ± 0.04	45.74 ± 0.24	42.36 ± 0.11

TABLES 1-4: **mCE (mean% $\pm$ std)** over three random trials for the CIFAR-10 and CIFAR-100 datasets corrupted by varied levels of *synthetic* (symmetric or asymmetric) label noise. Each combination of dataset, noise rate, augmentation, and loss function is considered, and the average mCE across all nonzero noise rates for each method is reported in the Noisy Avg. column. The best result for each augmentation + dataset combination is **boldfaced**.

Method		CIFAR-10N					CIFAR-100N
Augment	Loss	Aggregate	Random 1	Random 2	Random 3	Worst	Noisy Fine
NoAug	CE	32.24 ± 0.41	37.56 ± 0.18	37.66 ± 0.30	37.96 ± 0.13	49.25 ± 0.34	67.66 ± 0.15
	Focal	29.85 ± 0.42	34.84 ± 0.46	34.85 ± 0.52	35.20 ± 0.39	48.05 ± 0.96	66.13 ± 0.18
	NCE+RCE	30.18 ± 0.21	31.11 ± 0.73	31.49 ± 0.31	32.35 ± 1.80	$\mathbf{38.13 \pm 0.46}$	$\mathbf{62.82 \pm 0.34}$
	α-loss	$\mathbf{29.22 \pm 0.79}$	$\mathbf{30.71 \pm 1.18}$	$\mathbf{30.44 \pm 0.88}$	$\mathbf{31.34 \pm 0.36}$	39.93 ± 0.35	63.09 ± 0.38
AugMix	CE	15.40 ± 0.30	18.59 ± 0.15	18.76 ± 0.19	18.95 ± 0.17	29.73 ± 0.28	52.52 ± 0.32
	Focal	13.28 ± 0.16	$\mathbf{13.77 \pm 0.11}$	$\mathbf{13.60 \pm 0.30}$	$\mathbf{13.61 \pm 0.20}$	24.31 ± 0.18	49.47 ± 0.18
	NCE+RCE	13.72 ± 0.27	14.16 ± 0.03	13.85 ± 0.18	14.07 ± 0.09	$\mathbf{18.14 \pm 0.32}$	48.90 ± 0.05
	α-loss	$\mathbf{13.06 \pm 0.13}$	14.07 ± 0.28	14.04 ± 0.07	14.00 ± 0.06	21.25 ± 0.04	$\mathbf{48.78 \pm 0.29}$

TABLE 5: **mCE (mean% $\pm$ std)** over three random trials across CIFAR-*N for each combination of augmentation and loss function. *Aggregate* through *Worst* are corruptions of CIFAR-10 and *Noisy Fine* is the sole corruption of CIFAR-100. The best result for each augmentation + dataset combination is **boldfaced**.

Although synthetic approaches are commonplace in the evaluation of robust learning methods, we assert that human-annotated noise, when available, offers a more accurate representation of label corruption present in real-world systems. Specifically, we justify the choice of CIFAR-*N over synthetic noise with the following reasons [20]:

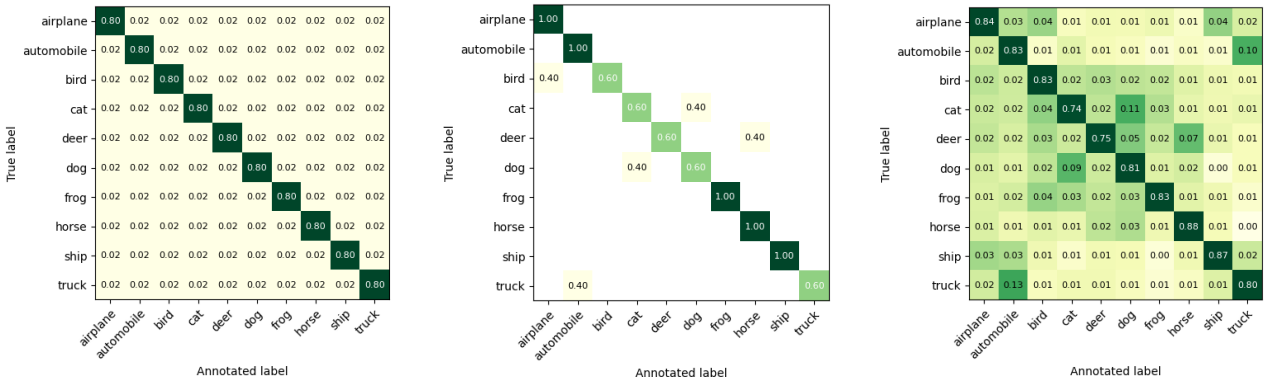
1) Datasets in CIFAR-*N provide more complex and diverse transition matrices compared to their synthetic counterparts, as shown in Figure 3.
2) Human-annotated noise is feature-dependent (*e.g.*, a cat-looking dog breed is more likely to be labeled a cat) while synthetic noise is strictly label-dependent.
3) CIFAR-*N takes into account the potential co-existence of two classes in a single image, often found in the CIFAR-100 dataset.

For these reasons, we additionally train our models on the CIFAR-*N datasets provided in Table 5. The generation process of CIFAR-*N collects data from *three* human annotators of CIFAR-10; therefore, each image in the dataset is annotated with *three* independent labels. Considering the three labels reported for each CIFAR-10 image, (1) the **Aggregate** dataset selects the most common label, (2) the **Random 1/2/3** datasets uniformly sample the labels, and (3) the **Worst** dataset uniformly samples one of the incorrect labels (if they exist). On the other hand, CIFAR-*N only collects data from *one* human annotator of CIFAR-100, and these results are directly reflected in the **Noisy Fine** dataset.

For similar reasons, we select CIFAR-*-C for evaluation. Since CIFAR-10-C (CIFAR-100-C) contains 15 corruptions of each image in CIFAR-10 (CIFAR-100), we consider the *mean corruption error* (mCE) – an average over the 15 individual corruption errors – as our primary performance metric. We consider the *clean error* – the test error on clean CIFAR datasets – as a baseline metric, and these results are reported in Tables A.1-A.5 in Appendix A.

5.2 Network Settings & Preprocessing

The following settings are fixed for all experiments: we use a WideResNet-40-2 model [19], and train for 100 epochs; the optimizer is SGD with a Nesterov momentum of 0.9 and weight decay of 5×10^{-4}; the learning rate scheduler is cosine annealing with an initial value of 0.1 and final value of 10^{-6}. Additionally, train batches of size 32 are preprocessed with random horizontal flips and batch normalization before being fed into the network.



(a) Symmetric (b) Asymmetric (c) CIFAR-10N Random 2

Fig. 3: Transition matrices for three methods of label noise generation, highlighting the greater diversity in human-annotated noisy labels (c) compared to its synthetically-generated counterparts (a, b).

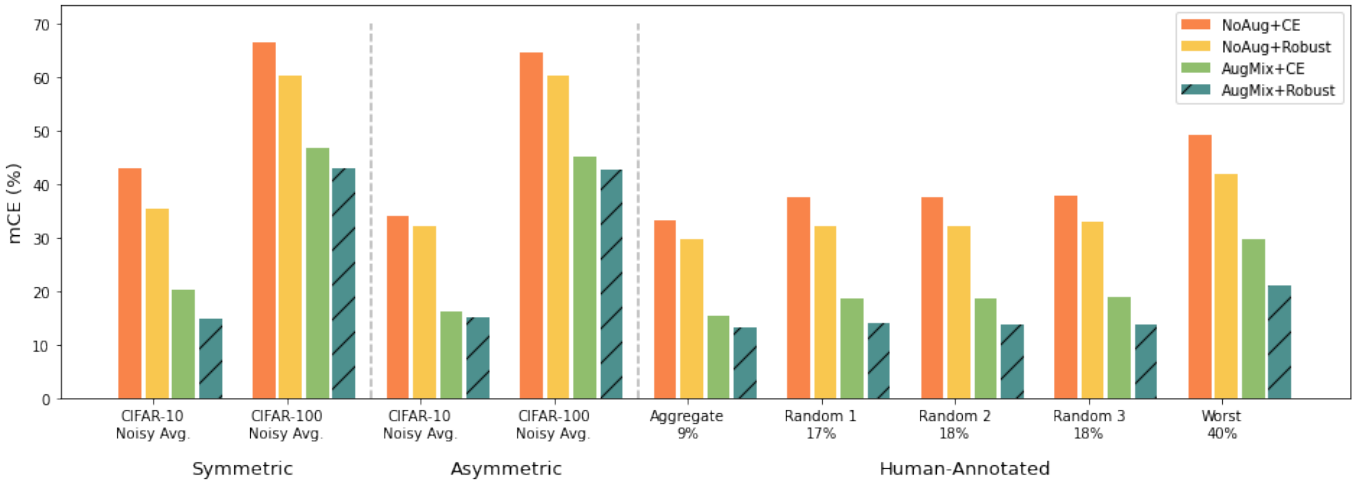


Fig. 4: The performances of each method type across symmetric, asymmetric, and human-annotated settings of label noise. The Noisy Avg. results for both CIFAR-10 and CIFAR-100 are included in the synthetic panels, while the CIFAR-10N results are included in the human-annotated panel. Hatched bars indicate the best performing method types for each setting. Note that our proposed methodology, AUGLOSS, *a.k.a.* Augmix+Robust, is the clear winner in all settings considered.

5.3 Augmentation + Loss Combinations

We compare the performances of several AUGLOSS-specific methods in the settings outlined above. These methods are created by pairing known data augmentation techniques with a variety of basic loss functions. Specifically, we consider two types of data augmentation: **NoAug** (no augmentation) and the SoTA example **AUGMIX**, proposed by [40]. The AUGMIX technique is defined by its two nontrivial transformations, each composed of stochastically-sampled operations (*e.g.*, autocontrast, rotate, solarize, *etc.*) organized in three chains of varied length, concluding with a mixture between each chain's output and the original image (see Figure B.1). Three generated images $x_{\text{orig}}, x_{\text{aug1}}, x_{\text{aug2}}$ are fed through the network to produce three posteriors $\hat{p}_{\text{orig}}, \hat{p}_{\text{aug1}}, \hat{p}_{\text{aug2}}$. The Jensen-Shannon divergence consistency loss (ℓ_{JS}) serves as the augmentation regularizer

$$\ell_{\text{JS}}(\hat{P}) = \frac{1}{3} \sum_{\hat{p}_i \in \hat{P}} \text{KL}\left(\hat{p}_i \| \hat{p}_{\text{mix}}\right), \tag{12}$$

where $\hat{p}_{\text{mix}}$ is the mean of the three posteriors in $\hat{P} := \left(\hat{p}_{\text{orig}}, \hat{p}_{\text{aug1}}, \hat{p}_{\text{aug2}}\right)$. Consistent with [40], λ is set to 12.

For basic loss functions, we consider the standard **cross entropy** loss along with three tunable robust loss function families: **focal** loss, **NCE+RCE** loss, and $\boldsymbol{\alpha}$-loss. Each family is optimally tuned on the CIFAR-10 and CIFAR-100 datasets under 20% symmetric noise, with the parameter search spaces motivated by [9], [11], [39] and reported in Table 6. We tune on symmetric noise because: (1) this approach is computationally inexpensive and requires no extra information on class relationships; (2) we seek to evaluate each loss' ability to generalize its robust behavior to *unforeseen* transition matrices and *varied* label noise rates.

5.4 Main Results & Discussion

For each pair of augmentation + loss combination and noisy-labeled train set, the mCE (mean% $\pm$ standard deviation) is reported over three random trials in Tables 1-5. Interpreting the results, two key observations are discussed below.

AUGLOSS best addresses dataset corruption.

As previously stated, there exist widely-accepted methods that effectively combat label noise in the training stage (*i.e.*, robust loss without augmentation) or feature noise in the testing stage (*i.e.*, AUGMIX+CE paired with Jensen-Shannon loss), compared to the oft-used methodology, namely, CE loss without augmentation. Our experimental setup encapsulates these baseline methods to showcase the nontrivial gains achieved by AUGLOSS methods when training on noisy CIFAR-10/100 and testing on CIFAR-*-C. Specifically, for analysis, we partition the augmentation + loss combinations into the following four method types:

- **NoAug+CE**: the baseline method of CE loss without AUGMIX. These results are primarily included to highlight the need for a solution to combat both train-time label noise and test-time feature distribution shifts.
- **NoAug+Robust**: the group of widely-accepted methods that train with a robust loss function, specifically to combat label noise in the training stage. In this case, we group together three specific instances – NoAug+*Focal*, NoAug+*NCE+RCE*, and NoAug+α*-loss* – by averaging over the three results in each setting of train data.
- **AUGMIX+CE**: the specific method proposed by [21], using AUGMIX data augmentation with CE loss to combat data distribution shifts in the testing stage.
- **AUGMIX+Robust**: the group of methods that follow the AUGLOSS framework. Each method trains with AUGMIX and a robust loss function (focal, NCE+RCE, or α-loss) in order to simultaneously harden the classifier against train-time label noise and test-time distribution shifts. As with *NoAug+Robust*, we average over the results given by the three robust loss functions.

In Figure 4, we compare the performances (with respect to mCE) of each method type – NoAug+CE, NoAug+Robust, AUGMIX+CE, and AUGMIX+Robust – under the different settings of label noise. Specifically, the two leftmost panels highlight the performance of each method type under synthetic (symmetric/asymmetric) label noise, while the right-

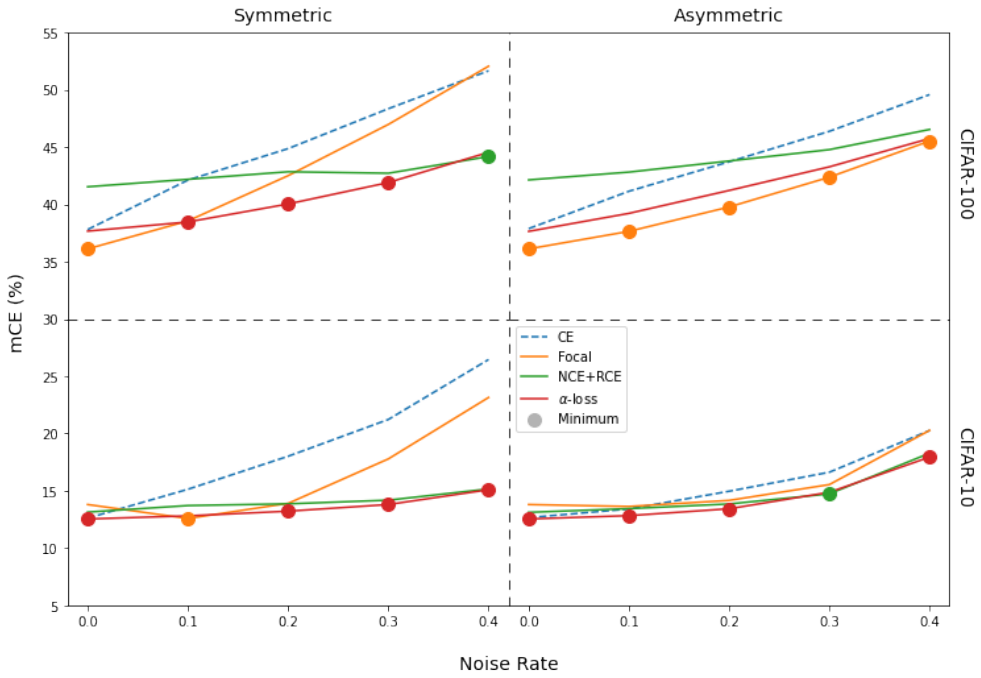


Fig. 5: The performances of each loss function across the synthetic (symmetric + asymmetric) settings of label noise. Dots are color-coded according to the best-performing loss functions at each setting and noise rate.

most panel considers the performance of each method type under human-annotated (CIFAR-10N) label noise. Taking both settings into account, the results show that the group of AUGLOSS methods (AUGMIX+Robust) consistently outperforms the other three types across all reported label noise settings. Although the well-known pairing of AUGMIX with CE loss already yields a drastic improvement in performance over the NoAug+CE baseline (*e.g.*, -18.9% mCE in the CIFAR-10N Random 2 setting), the novel pairing of AUGMIX with a *robust* loss showcases the potential for further improvement in real-world noisy settings (*e.g.*, -23.8% mCE in the CIFAR-10N Random 2 setting). Overall, these results underscore the pressing need for models to contain *both* data augmentation and robust loss functions (not just one) in order to simultaneously learn on noisy labels in the training stage, and generalize its classification ability to unforeseen feature distribution shifts in the testing stage.

No single robust loss function works best in all settings.

Within the AUGLOSS framework, we seek to evaluate and compare the performances of each robust loss function family – focal, NCE+RCE, and α-loss – paired with AUGMIX when training on noisy-labeled data and testing on CIFAR-*-C. In doing so, we observe that the best results in Tables 1-5 come from a *mixture* of the three losses, not just one. Figure 5, for example, illustrates the performance of each loss function in varied settings of synthetic label noise; here, we observe that α-loss generally performs the best in both symmetric settings (CIFAR-10 and CIFAR-100), as well as the asymmetric CIFAR-10 setting. Additionally, NCE+RCE loss tends to perform worse in low levels of label noise – especially with CIFAR-100 – but shows to be very competitive in high levels of label noise. In the three settings that α-loss achieves the best results, focal loss appears to approach CE loss – the non-robust baseline – as label noise increases; however, focal loss clearly outperforms the others across all noise rates in the asymmetric CIFAR-100 setting, which underscores the idea that no robust loss function is the universal "best fit" for all settings of *synthetic* label noise.

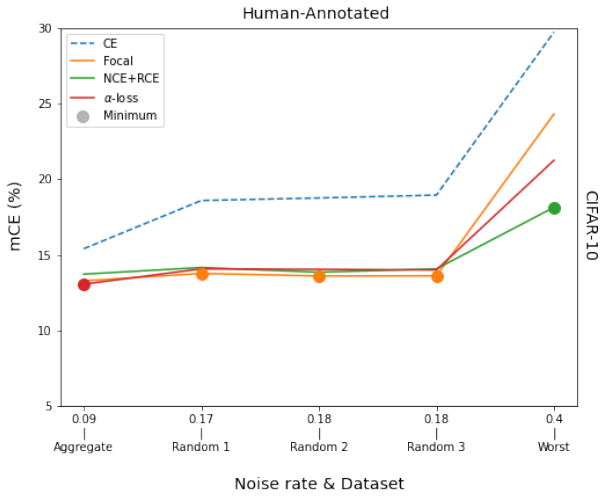


Fig. 6: The performances of each loss function across the five CIFAR-10N datasets. Dots are color-coded according to the best-performing loss functions for each dataset.

On the other hand, Figure 6 displays the performance of each loss function in varied settings of human-annotated label noise– namely CIFAR-10N. Specifically, this figure shows that in our experiment, α-loss performed the best with low-level noise (Aggregate), Focal loss performed the best with mid-level noise (Random 1-3), and NCE+RCE performed the best with higher levels of noise (Worst). Keeping in mind

that each family is optimized under 20% symmetric noise, these results suggest the following:

- α-loss may be the appropriate choice in settings of label noise over-estimation, where α is tuned against a noise rate (*i.e.*, 20%) higher than the true rate (*i.e.*, 9%).
- Focal loss shows to be effective when the true noise rate is predictable (*i.e.*, 20% $\approx$ 18%).
- NCE+RCE loss appears to be valuable in settings of label noise under-estimation – a useful property in safety-critical applications – where its parameters are tuned against a lower noise level but can generalize to higher levels (*i.e.*, 40%).

Since each loss clearly has its own merits, we conclude that no single choice appears to be the universal "best fit" within the AUGLOSS framework; rather, with more context on the task at hand, an appropriate loss function can be tailored to the specific needs of the problem.

6 CONCLUSION

In this work, we have proposed a novel *methodology*, AUGLOSS, wherein we have synthesized data augmentation techniques and robust loss functions in order to learn models that exhibit robustness to both train-time noisy labeling *and* test-time feature distribution shifts. In particular, using AUGMIX as the augmentation method, we have evaluated the performance of several tunable robust loss functions, including focal loss, NCE+RCE, and α-loss. Through our experimental procedure, we have demonstrated that on the whole, AUGLOSS yields much better mCE than either the robust loss functions or augmentation techniques alone. We have also observed that no tunable robust loss function stands out with different corruption settings handled better by different losses. Our novel methodology and the ensuing benchmarks significantly enhance existing results based on just loss functions or just augmentations. These benchmarks can further propel the field in identifying other augmentation techniques and robust loss functions to improve the robustness of models for out-of-distribution training and testing. In particular, a possible future effort is to leverage the recently introduced real-world WILDS dataset [42] to further evaluate the efficacy of AUGLOSS. Augmentations based on adversarial approaches [14] and NoisyMix [34] are other potential directions.

Kyle Otstot is a computer science M.S. student at Arizona State University. In 2022, he received B.S. degrees in computer science and mathematics from Arizona State University as a member of the Barrett Honors College. As an undergraduate at ASU, he served as a research assistant under Dr. Lalitha Sankar for one year, and his research interests currently include DL model robustness under real-world noisy labeling and feature domain shifts.

John Kevin Cava is a computer science Ph.D. student at Arizona State University. He received B.S. degrees in computer science, mathematics, and molecular biosciences and biotechnology from Arizona State University in 2018. His research interests include computer vision, natural language processing, and their use in molecular biology applications.

Tyler Sypherd is an electrical engineering Ph.D. student at Arizona State University. He received B.S.E. and M.S. degrees in electrical engineering also from Arizona State University in 2017 and 2020, respectively. His research interests include theoretical machine learning and information theory.

Lalitha Sankar received the B. Tech. degree from the Indian Institute of Technology, Bombay, the M.S. degree from the University of Maryland, and the Ph.D. degree from Rutgers University. She is currently an Associate Professor in the School of Electrical, Computer, and Energy Engineering at Arizona State University. Her research interests include applying information theory and data science to study reliable, responsible, and privacy-protected machine learning as well as cyber security and resilience in critical infrastructure networks. She received the National Science Foundation CAREER Award in 2014, the IEEE Globecom 2011 Best Paper Award for her work on privacy of side-information in multi-user data systems, and the Academic Excellence award from Rutgers in 2008.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Chaowei Xiao of NVIDIA for engaging with us in discussions and sharing invaluable feedback. This work is supported in part by NSF grants CIF-1901243, CIF-1815361, CIF-2007688, CIF-2134256, and SaTC-2031799 as well as a Google AI for Social Good grant.

ADDITIONAL EXPERIMENT DETAILS

Computing Resources

Our experiments were executed on a cluster of K80 GPUs. The installed CUDA version was v11.2. The settings with NoAug would generally take 3-4 hours to finish, while the settings with AUGMIX would generally take 8-10 hours.

Hyperparameter Tuning

The focal loss is parameterized by $\gamma \in \mathbb{R}^+$ and we consider the following search space [17]:

$$\gamma \in \{0.0, 0.5, 1.0, 2.0, 5.0\}.$$

Furthermore, the NCE+RCE loss employs two hyperparameters $(\beta_1, \beta_2) \in \mathbb{R}^2$ and motivated by [9], we consider the following search space:

$$(\beta_1, \beta_2) \in \{0.1, 1.0, 10, 99, 99.9\} \times \{0.1, 1.0, 10, 100\}.$$

Lastly, the α-loss is parameterized by $\alpha \in \mathbb{R}^+$ and we consider the following search space [11]:

$$\alpha \in \{1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 2.0, 3.0, 4.0\}$$

The best-performing parameters with respect to mCE for each family are shown below in Table 6.

	Focal: γ		NCE+RCE: (β_1, β_2)		α-loss: α	
	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
NoAug	5.0	5.0	(1.0,0.1)	(99.9,0.1)	3.0	2.0
AUGMIX	5.0	5.0	(1.0,0.1)	(99,1.0)	2.0	1.3

TABLE 6: Hyperparameters for each tunable loss function.

REFERENCES

[1] H. Song, M. Kim, and J.-G. Lee, "SELFIE: Refurbishing unclean samples for robust deep learning," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 5907–5915. [Online]. Available: http://proceedings.mlr.press/v97/song19b.html

[2] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," 2021.

[3] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. Mit Press, 2008.

[4] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019.

[5] G. Zheng, A. H. Awadallah, and S. Dumais, "Meta label correction for noisy label learning," in *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 2021.

[6] Z. Zhang, H. Zhang, S. O. Arik, H. Lee, and T. Pfister, "Distilling effective supervision from severe label noise," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9294–9303.

[7] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: a loss correction approach," 2017.

[8] D.-B. Wang, Y. Wen, L. Pan, and M.-L. Zhang, "Learning from noisy labels with complementary loss functions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, pp. 10 111–10 119, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17213

[9] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, "Normalized loss functions for deep learning with noisy labels," 2020.

[10] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," 2018.

[11] T. Sypherd, M. Diaz, J. K. Cava, G. Dasarathy, P. Kairouz, and L. Sankar, "A tunable loss function for robust classification: Calibration, landscape, and generalization," 2021.

[12] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[14] H. Wang, C. Xiao, J. Kossaifi, Z. Yu, A. Anandkumar, and Z. Wang, "Augmax: Adversarial composition of random augmentations for robust training," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[15] Z. Jiang, T. Chen, T. Chen, and Z. Wang, "Robust pre-training by adversarial contrastive learning." in *NeurIPS*, 2020.

[16] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang, "Adversarial robustness: From self-supervised pre-training to fine-tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 699–708.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.324

[18] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.

[19] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[20] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu, "Learning with noisy labels revisited: A study using real-world human annotations," 2021.

[21] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," *Advances in neural information processing systems*, vol. 26, 2013.

[22] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[23] T. Sypherd, R. Nock, and L. Sankar, "Being properly improper," 2021. [Online]. Available: https://arxiv.org/abs/2106.09920

[24] N. Charoenphakdee, J. Vongkulbhisal, N. Chairatanakul, and M. Sugiyama, "On focal loss for class-posterior probability estimation: A theoretical perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5202–5211.

[25] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania, "Calibrating deep neural networks using focal loss," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 288–15 299, 2020.

[26] Y. Wang, X. Ma, Z. Chen, Y. Luo, J. Yi, and J. Bailey, "Symmetric cross entropy for robust learning with noisy labels," 2019.

[27] X. Zhou, X. Liu, J. Jiang, X. Gao, and X. Ji, "Asymmetric loss functions for learning with noisy labels," 2021.

[28] Y. Liu and H. Guo, "Peer loss functions: Learning from noisy labels without knowing noise rates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6226–6236.

[29] Z. Leng, M. Tan, C. Liu, E. D. Cubuk, X. Shi, S. Cheng, and D. Anguelov, "Polyloss: A polynomial expansion perspective of classification loss functions," *arXiv preprint arXiv:2204.12511*, 2022.

[30] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *CoRR*, vol. abs/1708.04552, 2017. [Online]. Available: http://arxiv.org/abs/1708.04552

[31] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," *CoRR*, vol. abs/1905.04899, 2019. [Online]. Available: http://arxiv.org/abs/1905.04899

[32] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017.

[33] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *CoRR*, vol. abs/1805.09501, 2018. [Online]. Available: http://arxiv.org/abs/1805.09501

[34] N. B. Erichson, S. H. Lim, F. Utrera, W. Xu, Z. Cao, and M. W. Mahoney, "Noisymix: Boosting robustness by combining data augmentations, stability training, and noise injections," 2022.

[35] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.

[36] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 233–242.

[37] H. Xiao, H. Xiao, and C. Eckert, "Adversarial label flips attack on support vector machines," in *ECAI 2012*. IOS Press, 2012, pp. 870–875.

[38] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," 2017.

[39] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018.

[40] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," 2020.

[41] T. Huang, S. Halbe, C. Sankar, P. Amini, S. Kottur, A. Geramifard, M. Razaviyayn, and A. Beirami, "Dair: Data augmented invariant regularization," *arXiv preprint arXiv:2110.11205*, 2021.

[42] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5637–5664.

APPENDIX A
CLEAN ERROR RESULTS

Method		**Symmetric** CIFAR-10					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	$\mathbf{5.36 \pm 0.21}$	11.18 ± 0.23	15.84 ± 0.32	21.25 ± 0.30	28.21 ± 0.85	19.12 ± 0.42
	Focal	10.30 ± 1.79	8.56 ± 0.22	13.54 ± 0.21	19.25 ± 0.35	26.74 ± 0.49	17.02 ± 0.32
	NCE+RCE	7.66 ± 0.12	8.13 ± 0.26	9.50 ± 0.12	10.53 ± 0.13	12.10 ± 0.43	10.06 ± 0.23
	α-loss	6.15 ± 0.08	$\mathbf{7.22 \pm 0.13}$	$\mathbf{8.13 \pm 0.12}$	$\mathbf{9.45 \pm 0.25}$	$\mathbf{11.39 \pm 0.16}$	$\mathbf{9.05 \pm 0.16}$
AUGMIX	CE	$\mathbf{4.69 \pm 0.16}$	6.96 ± 0.18	9.70 ± 0.05	12.78 ± 0.29	18.32 ± 0.71	11.94 ± 0.31
	Focal	8.31 ± 0.03	6.46 ± 0.14	6.81 ± 0.08	9.88 ± 0.43	15.07 ± 0.50	9.55 ± 0.29
	NCE+RCE	6.25 ± 0.04	6.96 ± 0.18	7.33 ± 0.11	7.74 ± 0.17	8.83 ± 0.21	7.72 ± 0.17
	α-loss	5.15 ± 0.17	$\mathbf{5.70 \pm 0.21}$	$\mathbf{6.08 \pm 0.20}$	$\mathbf{7.07 \pm 0.18}$	$\mathbf{8.40 \pm 0.24}$	$\mathbf{6.81 \pm 0.21}$

TABLE A.1: **Clean error (mean% $\pm$ std)** over three random trials for the CIFAR-10 dataset corrupted by symmetric label noise. Each combination of dataset, noise rate, augmentation, and loss function is considered, and the average mCE across all nonzero noise rates for each method is reported in the Noisy Avg. column. The top result for each setting is **boldfaced**.

Method		**Symmetric** CIFAR-100					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	$\mathbf{24.67 \pm 0.39}$	32.01 ± 0.14	38.06 ± 0.18	44.08 ± 0.32	50.26 ± 0.71	41.1 ± 0.34
	Focal	25.8 ± 0.55	29.56 ± 0.39	35.16 ± 0.39	42.44 ± 0.24	49.7 ± 0.12	39.22 ± 0.28
	NCE+RCE	25.62 ± 0.25	$\mathbf{26.84 \pm 0.4}$	29.27 ± 0.35	31.63 ± 0.37	35.33 ± 0.08	30.77 ± 0.3
	α-loss	25.53 ± 0.36	27.08 ± 0.4	$\mathbf{28.75 \pm 0.28}$	$\mathbf{30.77 \pm 0.32}$	$\mathbf{33.61 \pm 0.41}$	$\mathbf{30.05 \pm 0.35}$
AUGMIX	CE	$\mathbf{23.04 \pm 0.23}$	27.74 ± 0.15	31.18 ± 0.61	35.2 ± 0.18	39.34 ± 0.17	33.36 ± 0.28
	Focal	25.18 ± 0.11	26.08 ± 0.36	29.25 ± 0.08	34.03 ± 0.22	39.47 ± 0.34	32.21 ± 0.25
	NCE+RCE	26.5 ± 0.67	26.98 ± 0.42	28.23 ± 0.2	29.77 ± 0.36	31.43 ± 0.23	29.1 ± 0.3
	α-loss	23.42 ± 0.26	$\mathbf{24.9 \pm 0.04}$	$\mathbf{26.92 \pm 0.27}$	$\mathbf{29.41 \pm 0.32}$	$\mathbf{32.43 \pm 0.52}$	$\mathbf{28.42 \pm 0.29}$

TABLE A.2: **Clean error (mean% $\pm$ std)** over three random trials for the CIFAR-100 dataset corrupted by symmetric label noise. Each combination of dataset, noise rate, augmentation, and loss function is considered, and the average mCE across all nonzero noise rates for each method is reported in the Noisy Avg. column. The top result for each setting is **boldfaced**.

Method		**Asymmetric** CIFAR-10					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	$\mathbf{5.21 \pm 0.26}$	8.4 ± 0.45	11.71 ± 0.25	15.71 ± 0.53	20.69 ± 0.17	14.13 ± 0.35
	Focal	9.83 ± 2.05	9.77 ± 0.06	$\mathbf{8.4 \pm 0.11}$	$\mathbf{10.56 \pm 0.18}$	$\mathbf{13.77 \pm 0.27}$	$\mathbf{10.62 \pm 0.16}$
	NCE+RCE	7.66 ± 0.12	8.22 ± 0.34	8.91 ± 0.26	11.16 ± 0.35	15.62 ± 0.33	10.98 ± 0.32
	α-loss	6.34 ± 0.13	$\mathbf{7.02 \pm 0.15}$	8.76 ± 0.12	12.29 ± 0.31	19.33 ± 0.48	11.85 ± 0.26
AUGMIX	CE	$\mathbf{4.72 \pm 0.01}$	5.92 ± 0.13	7.31 ± 0.3	9.3 ± 0.59	13.01 ± 0.09	8.88 ± 0.28
	Focal	8.31 ± 0.03	8.33 ± 0.08	8.5 ± 0.08	9.71 ± 0.28	14.21 ± 0.59	10.19 ± 0.26
	NCE+RCE	6.25 ± 0.04	6.72 ± 0.19	7.27 ± 0.08	8.47 ± 0.16	12.14 ± 0.47	8.65 ± 0.23
	α-loss	5.15 ± 0.17	$\mathbf{5.6 \pm 0.17}$	$\mathbf{6.44 \pm 0.05}$	$\mathbf{8.04 \pm 0.3}$	$\mathbf{11.54 \pm 0.42}$	$\mathbf{7.91 \pm 0.23}$

TABLE A.3: **Clean error (mean% $\pm$ std)** over three random trials for the CIFAR-10 dataset corrupted by asymmetric label noise. Each combination of dataset, noise rate, augmentation, and loss function is considered, and the average mCE across all nonzero noise rates for each method is reported in the Noisy Avg. column. The top result for each setting is **boldfaced**.

Method		**Asymmetric** CIFAR-100					
Augment	Loss	$\eta = 0$	0.1	0.2	0.3	0.4	Noisy Avg.
NoAug	CE	$\mathbf{24.4 \pm 0.09}$	30.95 ± 0.35	36.83 ± 0.19	41.88 ± 0.6	48.08 ± 0.67	39.44 ± 0.45
	Focal	25.72 ± 0.37	29.25 ± 0.29	32.92 ± 0.35	38.35 ± 0.23	44.69 ± 0.3	36.3 ± 0.29
	NCE+RCE	25.45 ± 0.14	$\mathbf{27.51 \pm 0.53}$	$\mathbf{29.24 \pm 0.16}$	$\mathbf{31.71 \pm 0.31}$	$\mathbf{35.12 \pm 0.37}$	$\mathbf{30.9 \pm 0.34}$
	α-loss	25.66 ± 0.24	28.11 ± 0.48	30.91 ± 0.35	34.53 ± 0.26	40.45 ± 0.26	33.5 ± 0.34
AUGMIX	CE	$\mathbf{23.28 \pm 0.45}$	26.76 ± 0.29	30.17 ± 0.32	33.02 ± 0.48	36.84 ± 0.51	31.7 ± 0.4
	Focal	25.18 ± 0.11	25.67 ± 0.33	$\mathbf{27.18 \pm 0.35}$	$\mathbf{29.54 \pm 0.18}$	$\mathbf{32.85 \pm 0.18}$	$\mathbf{28.81 \pm 0.26}$
	NCE+RCE	26.39 ± 0.43	27.31 ± 0.54	28.97 ± 0.1	30.82 ± 0.34	33.4 ± 0.54	30.12 ± 0.38
	α-loss	23.38 ± 0.26	$\mathbf{25.64 \pm 0.47}$	28.05 ± 0.11	30.35 ± 0.31	33.53 ± 0.64	29.39 ± 0.38

TABLE A.4: **Clean error (mean% $\pm$ std)** over three random trials for the CIFAR-100 dataset corrupted by asymmetric label noise. Each combination of dataset, noise rate, augmentation, and loss function is considered, and the average mCE across all nonzero noise rates for each method is reported in the Noisy Avg. column. The top result for each setting is **boldfaced**.

Method		CIFAR-10N					CIFAR-100N
Augment	Loss	Aggregate	Random 1	Random 2	Random 3	Worst	Noisy Fine
NoAug	CE	10.96 ± 0.29	16.09 ± 0.20	16.14 ± 0.36	15.98 ± 0.13	30.60 ± 0.27	47.21 ± 0.29
	Focal	9.37 ± 0.13	13.75 ± 0.25	13.92 ± 0.09	14.06 ± 0.08	29.50 ± 0.11	45.09 ± 0.47
	NCE+RCE	9.11 ± 0.20	9.86 ± 0.11	10.34 ± 0.26	10.21 ± 0.28	$\mathbf{17.13 \pm 0.39}$	$\mathbf{40.25 \pm 0.37}$
	α-loss	$\mathbf{8.23 \pm 0.26}$	$\mathbf{9.64 \pm 0.05}$	$\mathbf{9.94 \pm 0.25}$	$\mathbf{9.86 \pm 0.22}$	19.36 ± 0.38	40.94 ± 0.33
AUGMIX	CE	7.93 ± 0.28	10.93 ± 0.19	10.82 ± 0.21	11.24 ± 0.30	22.85 ± 0.17	41.01 ± 0.58
	Focal	8.00 ± 0.23	7.93 ± 0.15	7.62 ± 0.25	7.93 ± 0.04	17.85 ± 0.16	38.80 ± 0.27
	NCE+RCE	7.20 ± 0.13	7.94 ± 0.08	7.87 ± 0.11	7.70 ± 0.10	$\mathbf{12.81 \pm 0.30}$	38.10 ± 0.06
	α-loss	$\mathbf{6.29 \pm 0.05}$	$\mathbf{7.43 \pm 0.14}$	$\mathbf{7.53 \pm 0.12}$	$\mathbf{7.64 \pm 0.07}$	15.88 ± 0.39	$\mathbf{38.07 \pm 0.08}$

TABLE A.5: **Clean error (mean% $\pm$ std)** over three random trials across CIFAR-*N for each combination of augmentation and loss function. *Aggregate* through *Worst* are corruptions of CIFAR-10 and *Noisy Fine* is the sole corruption of CIFAR-100. The best result for each augmentation + dataset combination is **boldfaced**.

APPENDIX B
EXAMPLE OF AUGMIX

In Figure B.1, we illustrate one realization of AUGMIX, highlighting the preservation of image semantics.

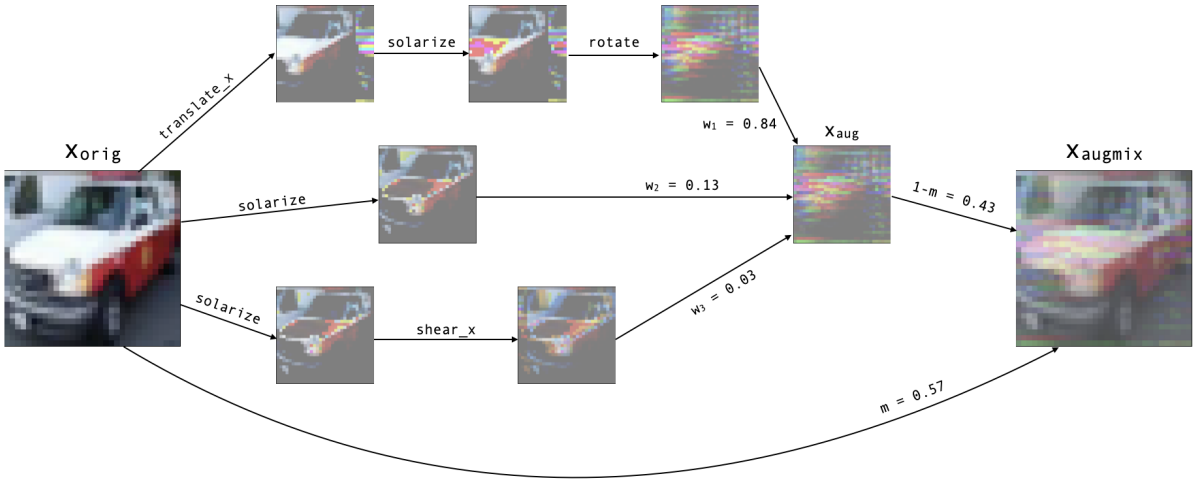


Fig. B.1: Realization of AUGMIX on a TRUCK-labeled image.