Analyzing the Impact of Input and Label Modifications on CIFAR-10 Classification

June Young Yi
2019-13541, Computer Science Department
Seoul National University
julianyi1@snu.ac.kr

Abstract

The performance of deep learning models is intrinsically tied to the quality and nature of training data. Real-world datasets often suffer from imperfections such as noisy labels or distorted inputs, while augmentations intentionally modify data to improve robustness. This paper investigates the effects of systematic input and label modifications on a standard image classification task using the CIFAR-10 dataset and a ResNet-18 model. We evaluate four distinct scenarios: (1) a baseline with unchanged data, (2) completely randomized labels per sample, (3) 20% symmetric label noise, and (4) strong input image perturbations. Our findings reveal that while random labels lead to a catastrophic failure in learning (Cohen's $\kappa \approx -0.01$), the model exhibits resilience to 20% label noise, achieving 64.5% Top-1 accuracy (vs. 78.3% baseline). Strikingly, strong input perturbations applied during training act as a powerful regularizer, significantly improving generalization and yielding the highest test accuracy of 85.3% and a Cohen's Kappa of 0.837. Learning curve analysis further elucidates these effects, showing reduced overfitting with input perturbations. These results underscore the complex interplay between data quality, augmentation strategies, and model robustness, offering insights for developing more resilient machine learning systems.

1. Introduction

Deep neural networks have achieved state-of-the-art performance across a multitude of computer vision tasks, including image classification [5, 8]. A critical factor underpinning this success is the availability of large-scale, high-quality labeled datasets [2]. However, in practical applications, data acquisition and annotation processes are often imperfect, leading to various forms of data corruption. Inputs can be distorted due to sensor noise or varying environmental conditions. Labels can be erroneous due to human annotator mistakes, ambiguity in classes, or automated labeling inaccuracies [4, 10]. Conversely, intentional modifications like data augmentation are widely used to improve model generalization [14].

Understanding how these data imperfections and intentional modifications affect model training and generalization is crucial for building robust and reliable machine learning systems. While models can sometimes memorize even random labels [16], their ability to generalize to unseen data under such conditions is severely compromised. On the other hand, certain forms of input modifications can significantly enhance model robustness [1, 19].

This work aims to systematically study the impact of specific input and label modifications on the CIFAR-10 [7] image classification task. We employ a ResNet-18 [5] architecture and investigate four distinct experimental configurations:

1. **Baseline:** Training with original CIFAR-10 inputs and labels.

2. **Random Label Shuffle:** Training with original inputs but completely randomized labels, where each label is assigned independently and uniformly at random per sample.

3. **Label Noise (20%):** Training with original inputs, but for 20% of the samples, their labels are randomly flipped to a different incorrect class.

4. **Input Perturbation:** Training with original labels but heavily distorted input images, including random crops, blur, color jitter, and random erasing. Test images undergo a milder, fixed perturbation.

We evaluate performance using standard metrics such as Top-1/Top-5 accuracy, loss, Cohen's Kappa score, and per-class F1-scores. Our analysis focuses on quantifying performance changes, examining learning dynamics, and discussing implications for model robustness.

2. Related Work

Learning with Noisy Labels. The challenge of training models with noisy labels has been extensively studied. Approaches range from robust loss functions [3, 18], label correction mechanisms [13], sample selection strategies [4, 6], to meta-learning techniques [9]. Our "Label Noise (20%)" experiment investigates symmetric label noise.

Data Augmentation and Robustness. Data augmentation is a cornerstone technique for improving the generalization of deep learning models [14]. Common methods include geometric transformations, color space adjustments, and more advanced techniques like Mixup [17], CutMix [15], and AutoAugment [1]. Our "Input Perturbation" experiment employs a suite of strong augmentations, including Random Erasing [19].

Model Capacity and Memorization. Deep networks can fit random labels [16], highlighting their memorization capacity but questioning true generalization. Our "Random Label Shuffle" experiment probes this.

Covariate Shift. Input perturbations can be viewed as a form of covariate shift. Our "Input Perturbation" experiment, with different train/test distortions, touches upon this challenge [12].

3. Methods

3.1. Dataset

We use the CIFAR-10 dataset [7], comprising 60,000 32x32 color images in 10 classes. It's split into 50,000 training and 10,000 test images. We reserve 10% of training data (5,000 images) for validation (fixed seed 42), resulting in 45,000 training, 5,000 validation, and 10,000 test images.

3.2. Baseline Model and Training

A ResNet-18 architecture [5] is used, initialized from scratch. Training is for 200 epochs using SGD (momentum 0.9, weight decay 5e-4, batch size 128). Initial LR is 0.1, with a 5-epoch linear warmup, then Cosine Annealing. Main training seed is 42; hook-specific seed (Shuffle, Noise) is 0.

3.3. Experimental Configurations

1. Baseline. Original CIFAR-10 inputs/labels. Inputs normalized (CIFAR-10 mean/std).

2. Random Label Shuffle. Original inputs. Each sample's label (train, val, test) replaced by a random uniform choice from 10 classes.

3. Label Noise (20%). Original inputs. 20% of sample labels (train, val, test) flipped to a random incorrect class.

4. Input Perturbation. Original labels. **Training/Validation Inputs:** RandomResizedCrop (32x32, scale (0.6,1.0)), RandomHorizontalFlip (p=0.5), GaussianBlur (kernel 3, sigma (0.1,2.0)), ColorJitter (factors 0.4, hue 0.1), ToTensor, Normalize, RandomErasing [19] (p=0.25, scale (0.02,0.2)). Examples in Figure 2. **Test Inputs:** GaussianBlur (kernel 3, sigma 1.0), ToTensor, Normalize.

3.4. Evaluation Metrics

Metrics: Cross-Entropy Loss, Acc@1, Acc@5, Cohen's κ, and per-class F1-scores. For "Random Label Shuffle", Acc@1/Acc@5/Loss use shuffled test labels; κ and classification reports use original test labels. Others use their respective loader labels.

4. Experiments and Results

4.1. Implementation Details

Conducted using PyTorch [11]. Model, hyperparameters, data splits consistent (Sec. 3).

4.2. Overall Performance

Table 1 summarizes the key test set metrics.

Baseline. Achieves 78.26% Acc@1 and $\kappa = 0.7584$.

Random Label Shuffle. Fails to learn; Acc@1 on random labels is 10.11%. Against original labels, $\kappa = -0.0134$. Loss high (2.3219).

Label Noise (20%). Degrades to 64.52% Acc@1, $\kappa = 0.6058$. Model shows resilience.

Input Perturbation. Best performance: Acc@1 85.33% (+7.07% vs. baseline), $\kappa = 0.8370$. Test loss lowest (0.5327). Strong regularization.

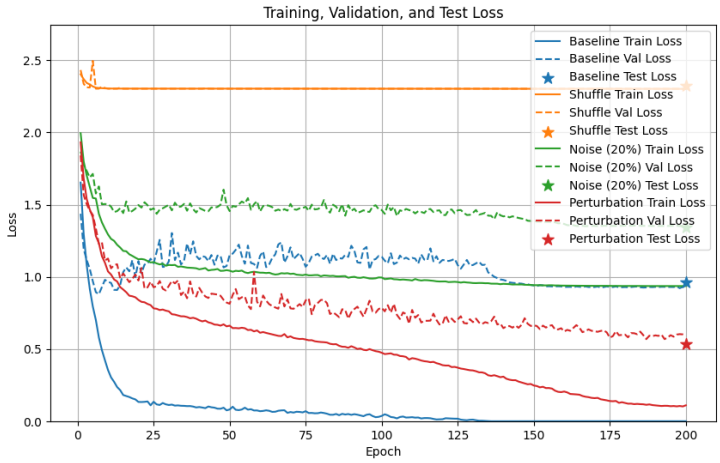
4.3. Training Dynamics

Learning curves for loss, Top-1 accuracy, and Top-5 accuracy are shown in Figure 1. Test points (stars) indicate final performance on the respective test sets.

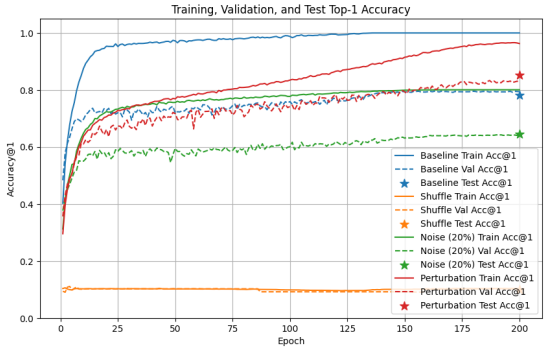
Baseline: Training Acc@1 reaches 100% (epoch 136), while validation Acc@1 plateaus around 79%; validation loss increases after epoch ∼140, indicating overfitting. Final test Acc@1 is 78.26%. **Random Label Shuffle:** Losses remain high (∼2.3); accuracies flat (∼10%). Final test Acc@1 on random labels is 10.11%. No meaningful learning. **Label Noise (20%):** Training Acc@1 reaches ∼80%; validation Acc@1 peaks at 64.28% (epoch 198).

Table 1. Test set performance. "Loss", "Acc@1", "Acc@5" are on loader labels. "Cohen's κ" for "Shuffle" is vs. original labels. Best Acc@1, Acc@5, κ bolded; lowest Loss (excl. Shuffle) bolded.

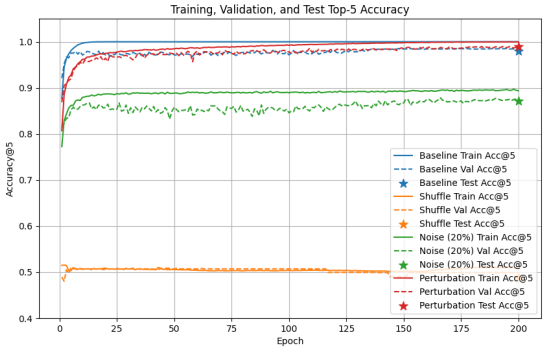
Experiment	Test Loss (loader labels)	Test Acc@1 (loader labels)	Test Acc@5 (loader labels)	Cohen's κ (report labels)
Baseline	0.9613	0.7826	0.9790	0.7584
Random Label Shuffle	2.3219	0.1011	0.4893	-0.0134
Label Noise (20%)	1.3409	0.6452	0.8707	0.6058
Input Perturbation	**0.5327**	**0.8533**	**0.9891**	**0.8370**



(a) Training, Validation, and Test Loss



(b) Training, Validation, and Test Top-1 Accuracy



(c) Training, Validation, and Test Top-5 Accuracy

Figure 1. Learning curves for all four experimental configurations, showing (a) Cross-Entropy Loss, (b) Top-1 Accuracy, and (c) Top-5 Accuracy for training (solid lines), validation (dashed lines), and final test points (stars) across 200 epochs.

The gap and later validation loss behavior suggest overfitting to noise. Test Acc@1 is 64.52%. **Input Perturbation:** Training Acc@1 reaches ∼96%. Validation Acc@1 closely tracks it, reaching ∼83%. The small train-val gap and low validation loss indicate improved generalization. Test Acc@1 is highest at 85.33%. Top-5 accuracy trends (Figure 1c) largely mirror Top-1, with higher absolute values. Input Perturbation also yields the best test Acc@5 (98.91%).

4.4. Per-Class Performance and Error Analysis

Table 2 presents per-class F1-scores. Figure 3 displays the confusion matrices.

The **Baseline** model shows typical difficulties with 'cat' and 'dog'. For **Random Label Shuffle**, against original labels, predictions are biased (Figure 3b), yielding near-zero F1 for most classes. With **Label Noise (20%)**, all class F1-scores drop. **Input Perturbation** improves F1-scores for all classes over baseline (Table 2), especially 'cat', 'dog',

Table 2. Per-class F1-scores on the test set. For "Random Label Shuffle", scores are against original CIFAR-10 labels.

Experiment	CIFAR-10 Classes (F1-Score)									
	Airplane	Automo-bile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Baseline	0.8126	0.8785	0.7200	0.6014	0.7541	0.6882	0.8279	0.8277	0.8796	0.8462
Random Label Shuffle (vs. original labels)	0.0000	0.1760	0.0135	0.0117	0.0000	0.0000	0.0284	0.0000	0.0000	0.0175
Label Noise (20%)	0.6546	0.7230	0.5939	0.5049	0.6481	0.5552	0.7001	0.6728	0.7008	0.6994
Input Perturbation	**0.8905**	**0.9188**	**0.8246**	**0.7000**	**0.8421**	**0.7538**	**0.8742**	**0.9039**	**0.9222**	**0.8909**

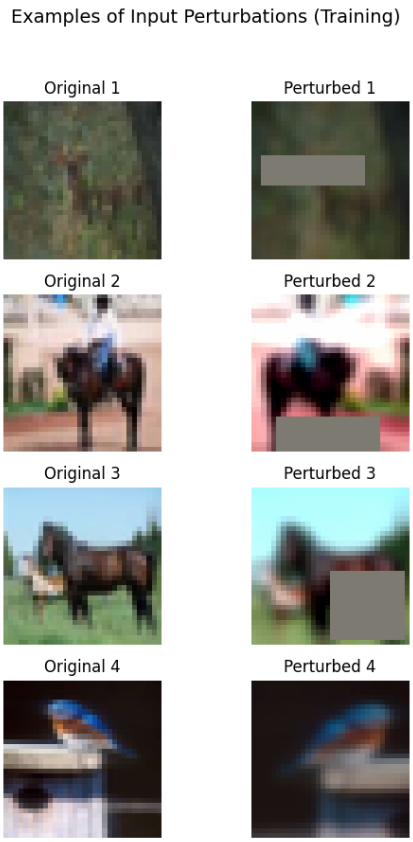


Figure 2. Examples of input images after applying the strong perturbation pipeline used in training. Each row shows an original image (left) and its perturbed version (right).

and 'bird', and its confusion matrix (Figure 3d) is visibly cleaner.

4.5. Qualitative Examples of Input Perturbations

Figure 2 illustrates the strong augmentations applied during training for the "Input Perturbation" experiment.

5. Conclusion

This study systematically investigated the impact of input and label modifications on a ResNet-18 model trained on CIFAR-10. **Random Label Shuffle** completely undermines learning, confirming the necessity of coherent label signals. Model performance against original labels was no better than random chance. **20% Symmetric Label Noise** significantly degraded performance. However, the model still learned effectively, showcasing some inherent robustness, though overfitting to noisy labels was evident. Most strikingly, **Strong Input Perturbations** acted as a powerful regularizer, leading to a substantial improvement in generalization and outperforming the baseline model. This configuration demonstrated reduced overfitting and better overall discrimination across classes.

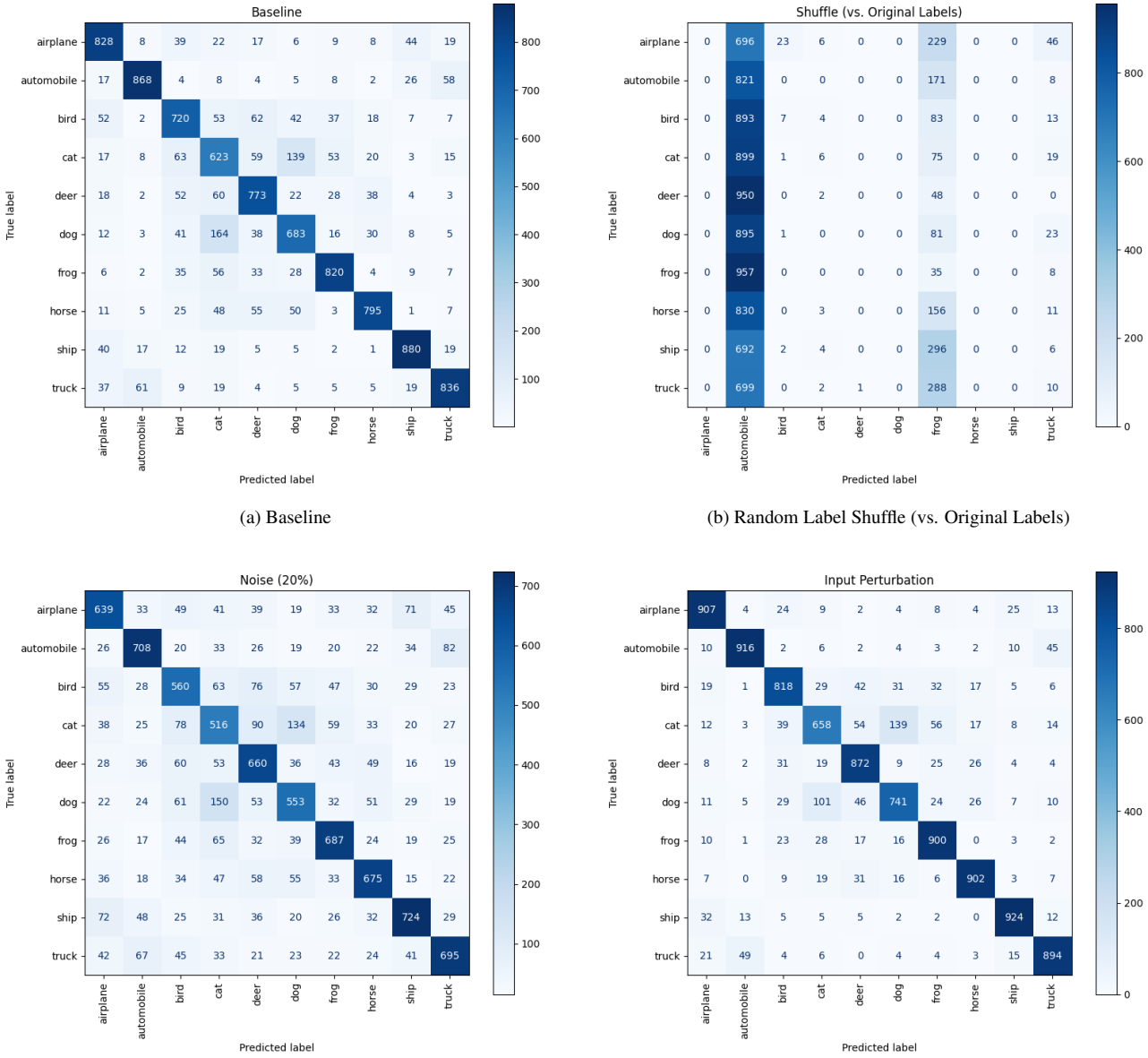
These results emphasize the critical role of data quality and appropriate data handling strategies. While models can tolerate moderate label noise or benefit greatly from well-designed augmentations, the integrity of the label signal remains paramount. The superior performance under strong input perturbation suggests that exposing models to diverse and challenging input conditions during training can be highly beneficial for learning robust, generalizable features.

Future work could explore the efficacy of explicit noise-robust training techniques under the 20% label noise condition, investigate a wider range of perturbation types and intensities, and extend this analysis to other datasets and model architectures.

References

[1] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2019. 1, 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, Miami, Florida, USA, June 20-25, 2009*, pages 248–255. IEEE Computer Society, 2009. 1



(a) Baseline

(b) Random Label Shuffle (vs. Original Labels)

(c) Label Noise (20%)

(d) Input Perturbation

Figure 3. Normalized confusion matrices for the test set. Diagonals represent per-class recall. For "Random Label Shuffle", matrix is vs. original true labels. For others, vs. their respective test loader labels.

[3] Aritra Ghosh, Himanshu Kumar, and P S Sastry. Robust loss functions under label noise for deep neural networks. In *AAAI*, 2017. 2

[4] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8535–8545, 2018. 1, 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1, 2

[6] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR, 2018. 2

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1, 2

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. 1

[9] Mengye Li, Bo An, Bradley Green, Sergey Levine, and Chelsea Finn. Learning to reweight examples for robust deep learning. In *ICML*, 2019. 2

[10] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, 2013. 1

[11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Z. Lin, N. Gimelshein, Luca Antiga, A. Desmaison, Andreas Kopf, Edward Z. Yang, F. DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, B. Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *NIPS*, 2019. 2

[12] Vishal M. Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Process. Mag.*, 32(3):53–69, 2015. 2

[13] Scott E. Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, D. Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR*, 2015. 2

[14] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 2019. 1, 2

[15] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, 2019. 2

[16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016. 1, 2

[17] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[18] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *ArXiv*, abs/1805.07836, 2018. 2

[19] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:13001–13008, 2020. 1, 2