Loss Functions Robust to the Presence of Label Errors

Nicholas Pellegrino[1*]
David Szczecina[2*], Paul Fieguth[1]
[1]Vision and Image Processing Group, Systems Design Engineering, University of Waterloo
[2]Mechanical & Mechatronics Engineering, University of Waterloo
{npellegr,dszczeci,pfieguth}@uwaterloo.ca

Abstract

Methods for detecting label errors in training data require models that are robust to label errors (*i.e.*, not fit to erroneously labelled data points). However, acquiring such models often involves training on corrupted data, which presents a challenge. Adjustments to the loss function present an opportunity for improvement. Motivated by Focal Loss (which emphasizes difficult-to-classify samples), two novel, yet simple, loss functions are proposed that de-weight or ignore these difficult samples (*i.e.*, those likely to have label errors). Results on artificially corrupted data show promise, such that F1 scores for detecting errors are improved from the baselines of conventional categorical Cross Entropy and Focal Loss.

1 Introduction

Errors and noise are pervasive across datasets used to train and evaluate machine learning models, and may substantially impact the performance of such models. Errors, or noise, take two main forms:

1. Image Errors, and
2. Label Errors.

Image errors specifically refer to problems within the *images* of data samples (*e.g.*, out of focus, additive noise, object not present or occluded, *etc.*), whereas label errors specifically refer to problems with the *labels* of samples (*e.g.*, mislabelled). In literature, label errors are often referred to by the term label *noise* [1, 2, 3, 4]; however, here the term label errors is preferred, since the phenomenon involves *permutation* rather than the *addition* of a stochastic process, to which the concept of noise normally would refer. While both forms of errors are important, this work focuses on label errors.

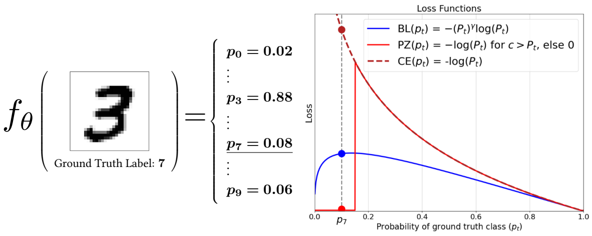

Figure 1: A model being trained operates on a corrupted data sample with a label error, and produces a low predicted probability for the ground truth class as labelled. Cross Entropy or Focal Loss would be sensitive to such samples, steering the model to fit to erroneous data. In contrast, the proposed loss functions, Blurry Loss (BL) and Piecewise-zero Loss (PZ), are insensitive and lead to robust training.

Fortunately, there exist methods for *detecting* label errors. These methods [5, 6, 7, 8] tend to operate on the premise that a well-trained model (*i.e.*, trained either on perfectly clean data, or in some manner robust to noise) will produce predicted class probabilities that conflict with the erroneous labels, thereby identifying samples likely to be mislabelled. However, obtaining such a model is not trivial. Indeed, perfect training data are difficult to come by (especially for every given domain), and training on a new model on a corrupted dataset can easily lead to fitting to the noise!

One dataset of particular interest is the BIOSCAN-5M insect dataset [9] (and its preceding BIOSCAN-1M dataset [10]) consisting of over five *million* hand-labelled images of insects. Labels consist primarily of both the *taxonomy* [11, 12] and a *genetic barcode* [13] for each sample. Due the the human process of taxonomic labelling, there exists a high likelihood of errors. Further, at finer-grained levels within the taxonomic hierarchy, the difference between categories of insects becomes increasingly subtle which further exacerbates the challenge of correct labelling. Yet another confounding issue is the lack of consensus among taxonomists, mean-

[*] Indicates equal contribution, joint first-authorship.

ing that what is considered correct by one may be incorrect by another!

Motivated by Focal Loss [14] which places emphasis on difficult-to-predict samples (*i.e.*, those where the predicted probability associated with the ground-truth label is low), two loss functions that largely *ignore* the difficult-to-predict samples are proposed. Samples with label errors would have low predicted probabilities for the as-labelled ground truth class, assuming the data has other correctly labelled samples upon which the model could learn and generalize. Therefore, by ignoring the difficult-to-predict samples, the loss functions would be robust to label errors in training data.

In this work, novel categorical loss functions are proposed to address the task of robustly training a model on corrupted data, such that existing frameworks for label detection may be employed to identify and remove (or even correct) incorrectly labelled samples within the dataset. Improvement is measured based on error detection F1 scores.

2 Background

Label errors can be found in many benchmark datasets [15], such as CIFAR [16] and ImageNet [17], and even MNIST [18] to a small extent. Many label error detection methods [5, 6, 7, 8] operate by training a model on the (possibly) corrupt data and rely on the model *not* fitting to the errors, such that predicted probabilities for erroneous data are low and may be detected through clever implementations of thresholding.

Note that experiments within this work specifically use the the framework of Confident Learning [8]; however, the proposed loss functions could be used, in principle, within many other frameworks. Furthermore, within the Confident Learning framework, detection using both the *'prune by noise rate'* and *'prune by class'* methods are used, whereby samples must simultaneously have both low probability of being their ground truth labelled class and high probability of being some other class.

The choice of loss function during model training can have a pronounced effect on the resulting model performance and generalizability. Focal Loss [14] extends conventional categorical Cross Entropy Loss [19] to improve the ability of models to train on data in which some classes are more difficult to learn than others. The most common use case is with imbalanced data, where some classes have far fewer samples and the model benefits from the additional weighting given to samples of these less common classes. For ground truth class t, the predicted probability, produced by the model, associated with this class is denoted p_t. With this notation, Cross Entropy Loss is defined (per sample) as

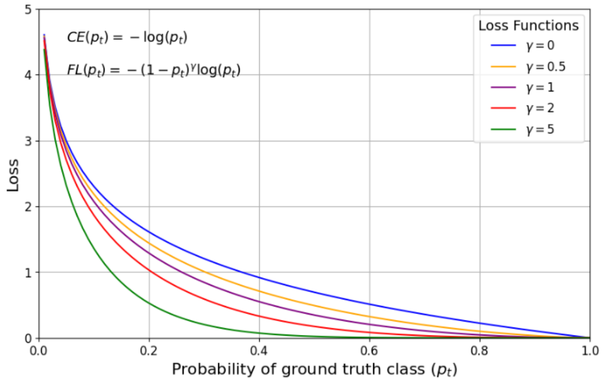

Figure 2: Cross Entropy (CE) and Focal Loss (FL) are compared. The effect of varying the weighting parameter γ is shown. Note that the $\gamma = 0$ case is equivalent to CE Loss. Observe that higher values of γ lead to more imbalanced weighting, focusing the training more on the difficult-to-classify samples.

$$\mathrm{CE}(p_t) = -\log(p_t), \tag{1}$$

and, with the addition of a weighting parameter, γ, Focal Loss is defined (per sample) as

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t). \tag{2}$$

Per batch or epoch, the individual sample losses are simply summed. The parameter γ determines the extent to which difficult-to-classify examples are weighted, based on p_t. For values of γ closer to 0, the weighting is more uniform, whereas for larger γ, closer to 5 for example, the weighting is more imbalanced, weighting samples with low p_t far greater than those with high p_t. Figure 2 illustrates the effects of varying γ. Conventionally, $\gamma = 2$ is used [14], based on empirical evidence indicating this value works well in general, though hyperparameter tuning is always an option when using Focal Loss.

A few loss functions designed to be robust to the presence of label errors have been proposed [2, 20, 21, 22], often incorporating additional loss terms, such as Mean Squared Error or Mean Absolute Error, to balance Cross Entropy. However, these tend to be theoretically complex (difficult to implement and interpret) and mainly benefit cases where data are very highly (*i.e.*, unrealistically) corrupted.

3 Method

Two major variations of categorical Cross Entropy Loss are proposed. The first is most closely related to Focal

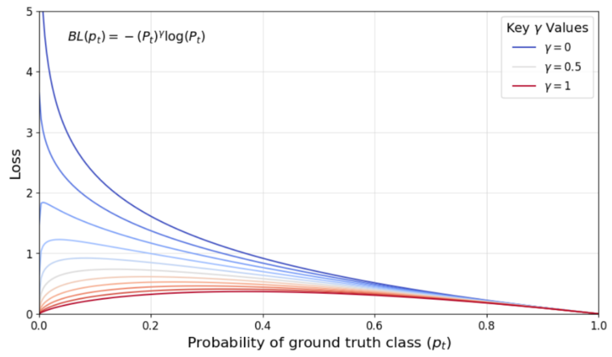

Figure 3: The Blurry Loss function is plotted here. Variations for the parameter γ are also shown, illustrating the change in imbalance of weighting.

Loss and takes a similar form. The second is a piecewise function that zeros the loss for samples with sufficiently low predicted probability, p_t, for the ground truth labelled category.

3.1 Blurry Loss

The first function is termed "Blurry" Loss in contrast to Focal Loss. Again, for ground truth labelled class t, predicted probability p_t, and a weighting parameter γ, the Blurry Loss is defined (per sample) as

$$\mathrm{BL}(p_t) = -(p_t)^\gamma \log(p_t). \tag{3}$$

Figure 3 illustrates this loss function and the effects of varying γ. For $\gamma = 0$, the Blurry Loss is equivalent to Cross Entropy Loss.

3.2 Piecewise-zero Loss

The second proposed loss function is designed in a piecewise manner such that samples for which the predicted probability, p_t, is beneath a cutoff are assigned a loss of zero, but otherwise Cross Entropy Loss is used. Most importantly, the *gradient* within the cutoff region is also zero, meaning that these samples with low p_t do not affect the training process (weights are not updated by zero-gradient samples). Again, the underlying assumption is that samples with label errors are most likely to have low p_t, given that the model has seen enough correctly labelled data to learn the correct classification. For ground truth labelled class t, predicted probability p_t, and a cutoff-position parameter c, the Piecewise-zero Loss is defined (per sample) as

$$\mathrm{PZ}(p_t) = \begin{cases} 0 & p_t \leq c, \\ \mathrm{CE}(p_t) = -\log(p_t) & p_t > c. \end{cases} \tag{4}$$

Figure 4 illustrates this loss function and the effects of varying c. For higher c, the loss function becomes more robust to label errors; however, this comes with the obvious downside that more of the data are simply ignored.

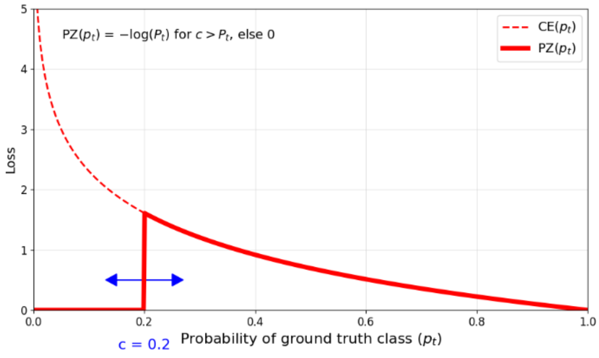

Figure 4: For $p_t < c$, the loss is zero. The cutoff parameter, c, may be adjusted to control the range of predicted probabilities that have zero loss assigned to them.

3.3 Loss Scheduling

During deep neural network (DNN) model training, the weights and biases are randomly initialized. Therefore, when training begins, the action of the model is entirely random and uncorrelated with the training labels. Directly using the proposed loss functions, particularly the Piecewise-zero Loss with a high cutoff, is likely to result in large amounts of correctly labelled data being de-weighted or outright ignored, to the detriment of the model.

To address this issue, beginning training with conventional loss functions, such as Cross Entropy, is suggested. A delay hyperparameter, d, is introduced to control at which epoch the transition to one of the proposed loss functions occurs. Under this loss scheduling scheme, the first $d - 1$ epochs use Cross Entropy, and starting at epoch d, one of the proposed loss functions is used.

4 Preliminary Experiments & Results

Experiments outlined in this section seek to demonstrate the efficacy of the proposed loss functions in robustly training models, as measured by their subsequent ability to identify label errors.

4.1 Experimental Setup

Datasets and Artificial Corruption To test the efficacy of the proposed loss functions, the labels of well known datasets are artificially corrupted. The method used is identical to that used in [23, 1], whereby labels are changed (*i.e.*, made incorrect) for a proportion of the data specified by a given corruption rate $\rho \in [0, 1]$. The datasets used are MNIST [18] and Fashion MNIST [24]. Both are similarly corrupted according to the same scheme. An underlying assumption is that these small-scale datasets have relatively low existing label error rates, such that nearly all errors consist of the artificially induced ones.

Baseline Loss Functions Categorical Cross Entropy and Focal Loss [14] will be used as baselines for the purpose of comparison.

Model A minimal convolutional neural network capable of scoring in excess of 99% accuracy (roughly the state-of-the-art) on MNIST will be used. The model consists only of two convolutional layers followed by two sets of alternating dropout and linear layers, and has a total of 1.2M trainable parameters.

Optimizer In all cases, the Adam [25] optimizer is used. This optimizer is adaptive (no hyperparameter tuning will be performed) and is a *de facto* gold standard in machine learning.

Hyperparameters The hyperparameter setting of γ, for Blurry Loss, and the cutoff, c, for the Piecewise-zero Loss are explored during these experiments. In particular, these parameters are swept across a range in order to study their effect on the ability of the trained model to detect label errors and find optimal values.

Metrics To assess the performance / efficacy of each loss function, the trained model will be used in the Confident Learning framework to detect label errors. These detections will be compared to a list of artificially corrupted samples. The F1 score [26, 27], a balanced metric of precision and recall, and a gold standard in detection, will be the primary metric of performance, while precision and recall will also be reported.

4.2 Experimental Results

To ensure fairness, models are trained only for the number of epochs at which point overfitting begins on the original *uncorrupted* datasets. A precursor experiment was performed for both uncorrupted datasets whereby the crossover point between training and testing loss is measured. In both cases, overfitting was found to occur starting at roughly 10 epochs. Therefore, the total number of epochs to be used in all experiments will be 10.

For *moderate* corruption rates of $\rho = 0.1$ and 0.2, as well as for a variety of loss function schedule delays, $d \in [0, 8]$, a suite of models are trained, each with a different variation of the loss function hyperparameters, γ and c. The primary results are shown as bar charts in Figure 5, in which the target (correct) and baseline-detected number of corrupted data samples are indicated. For each loss function, the number of total detections (blue bar) and correct detections (orange bar) are shown.

Table 1: F1 Score comparison across loss functions on corrupted datasets. Best results per row are bolded.

Dataset	F1 Score			
	CE	FL	BL	PZ
MNIST, $\rho = 0.1$	0.9668	0.9585	**0.9793** [γ = 0.3, d = 0]	0.9729 [c = 0.05, d = 6]
MNIST, $\rho = 0.2$	0.9708	0.9631	**0.9845** [γ = 0.4, d = 0]	0.9827 [c = 0.05, d = 4]
Fashion MNIST, $\rho = 0.1$	0.8222	**0.8332**	0.8237 [γ = 0.2, d = 0]	0.7954 [c = 0.05, d = 6]
Fashion MNIST, $\rho = 0.2$	0.8839	0.8814	**0.8892** [γ = 0.3, d = 0]	0.8745 [c = 0.05, d = 6]

To measure the effect loss scheduling (*i.e.*, delay prior to using proposed loss), loss functions with optimal parameters as found in Figure 5 are evaluated over a range of delays, $d \in [0, 8]$. Results are presented in Figure 6.

Lastly, for both datasets (MNIST and Fashion MNIST) and two corruption rates ($\rho = 0.1$ and 0.2), the performance (F1) for optimal loss function parameter settings are reported in Table 1.

5 Discussion

From Figure 5, observe that the performance of the proposed loss functions tends to exceed that of the baselines (CE and FL) for most parameter settings. There exist intermediate (*i.e.*, not extreme) parameter settings that yield optimal performance, which exceed that of the baseline in terms of both precision and recall (and F1). Notice that in these cases, such as $c = 0.05$ for Piecewise-zero Loss and $\gamma = 0.5$ for Blurry Loss, the total detections and correct detections more closely match the true corrupted data. Additionally, notice that performance is worse with Focal Loss than with Cross Entropy (or the proposed loss functions), as expected. This is likely a result of Focal Loss emphasizing difficult-to-classify samples, which in this case would often be those with corrupted labels, and thereby causing the model to fit to erroneous data.

The investigation of loss scheduling, shown in Figure 6, indicates that Blurry Loss tends to work best for no delay (*i.e.*, Blurry Loss may be used directly without CE for initial stages of training); however, the performance for the Piecewise-zero Loss is negatively impacted for either small or large amounts of delay. Having little or no delay likely results in some samples being forever ignored by the Piecewise-zero Loss and the model effectively 'seeing' less training data. On the other hand, with

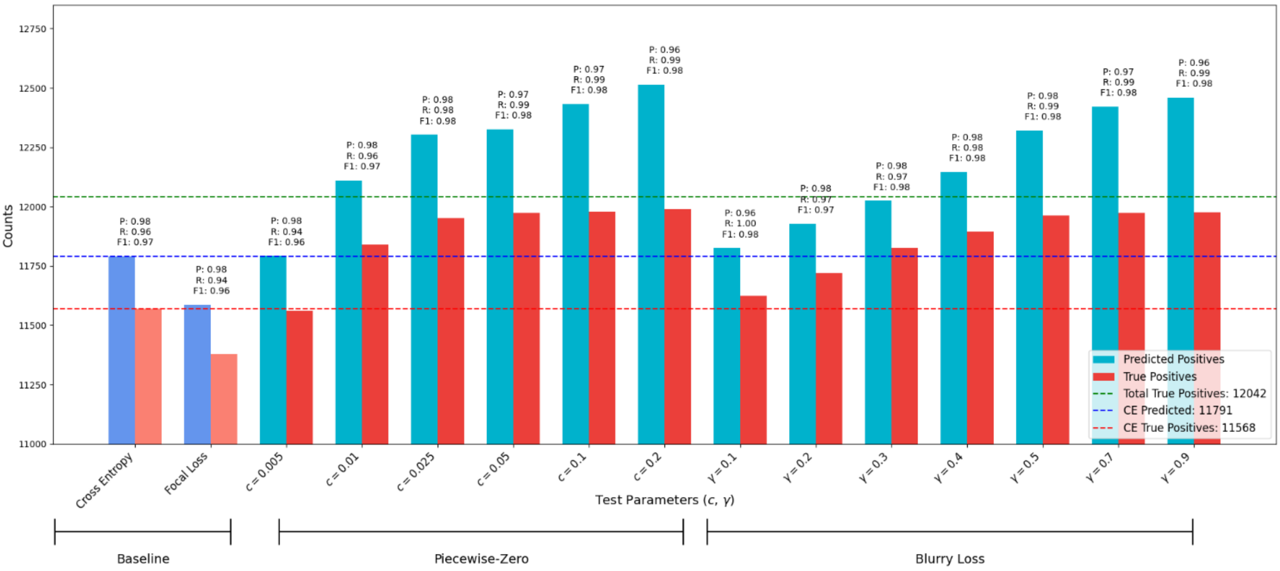

Figure 5: Total and correct detections (true positives) for baseline and experimental loss functions. Shown here are results for the MNIST dataset with corruption rate $\rho = 0.2$ and a delay of $d = 4$. Precision, recall, and F1 scores are presented above for each loss function parameter setting. Observe that F1 is improved compared to the baselines by the proposed loss functions for appropriate parameter settings, often trading-off precision and recall.

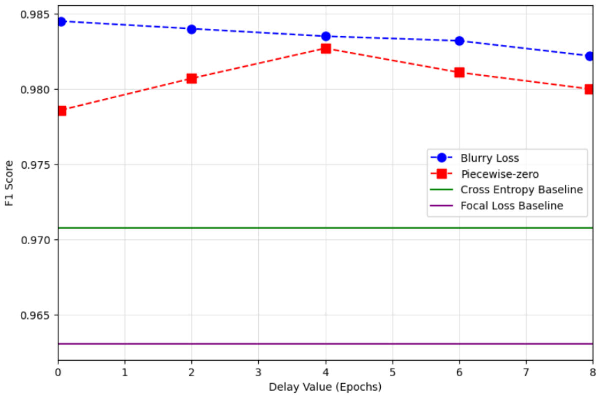

Figure 6: The effect of varying the amount of delay in loss the schedule before instituting alternate robust loss on performance (F1 score). Tests performed on MNIST dataset at a corruption rate of $\rho = 0.2$, and with parameters $\gamma = 0.4$ and $c = 0.05$.

too much delay, the model is mainly trained using Cross Entropy and the proposed loss function is used for very few epochs and has a reduced effect.

Based on the summary of optimal performance across datasets and corruption rates shown in Table 1, it appears that the impacts of the proposed loss functions are most pronounced at higher corruption rates (*i.e.*, difference in F1 scores between baselines and proposed are greater for $\rho = 0.2$ than 0.1). Further, the ability to detect label errors in Fashion MNIST was far reduced, and the proposed loss resulted in more modest improvements, if any, over the baselines. Indeed, for $\rho = 0.1$, Focal Loss results in the highest F1 score. The performance in this case may be a consequence of Fashion MNIST being a more difficult dataset than MNIST but at the same time still using the same minimal model (designed and suitable for MNIST). Nonetheless, at $\rho = 0.2$, the F1 score for Blurry Loss exceeds the others, even on Fashion MNIST. In all cases, the best performance of Blurry Loss slightly exceeds that of the Piecewise-zero Loss.

6 Conclusion

The proposed loss functions, Blurry Loss and Piecewise-zero Loss, significantly improved models' abilities to detect label errors when trained on artificially corrupted datasets. While current preliminary results show promise, a more detailed study involving comparisons with other robust loss functions will be required to adequately understand the impact of using such loss functions. In future work, it is hoped that these loss functions may be used on more realistic datasets, such as the BIOSCAN-5M dataset, where label errors are unknown and identification may lead to improvements in classifier performance and may even affect change in the current understanding of taxonomy and entomology.

Acknowledgments

This research was enabled in part by support provided by Calcul Québec (`calculquebec.ca`) and the Digital Research Alliance of Canada (`alliancecan.ca`).

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) via the NSERC-CGS D and NSERC Discovery programs (funding reference RGPIN-2020-04490).

Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), CRSNG-BESC D, et CRSNG Subvention à la Découverte (numéro de référence RGPIN-2020-04490).

Natural Sciences and Engineering Research Council of Canada
Conseil de recherches en sciences naturelles et en génie du Canada
Canada

References

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

[2] Z. Zhang and M. Sabuncu, “Generalized cross entropy loss for training deep neural networks with noisy labels,” *Advances in neural information processing systems*, vol. 31, 2018.

[3] G. Algan and I. Ulusoy, “Image classification with deep learning in the presence of noisy labels: A survey,” *Knowledge-Based Systems*, vol. 215, p. 106771, 2021.

[4] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, “Learning from noisy labels with deep neural networks: A survey,” *IEEE transactions on neural networks and learning systems*, vol. 34, no. 11, pp. 8135–8153, 2022.

[5] Z. Lipton, Y.-X. Wang, and A. Smola, “Detecting and correcting for label shift with black box predictors,” in *International conference on machine learning*. PMLR, 2018, pp. 3122–3130.

[6] J. Huang, L. Qu, R. Jia, and B. Zhao, “O2u-net: A simple noisy label detection approach for deep neural networks,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3326–3334.

[7] G. Pleiss, T. Zhang, E. Elenberg, and K. Q. Weinberger, “Identifying mislabeled data using the area under the margin ranking,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 044–17 056, 2020.

[8] C. Northcutt, L. Jiang, and I. Chuang, “Confident learning: Estimating uncertainty in dataset labels,” *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.

[9] Z. Gharaee, S. C. Lowe, Z. Gong, P. M. Arias, N. Pellegrino, A. T. Wang, J. B. Haurum, I. Zarubiieva, L. Kari, D. Steinke *et al.*, “Bioscan-5m: A multimodal dataset for insect biodiversity,” *arXiv preprint arXiv:2406.12723*, 2024.

[10] Z. Gharaee, Z. Gong, N. Pellegrino, I. Zarubiieva, J. B. Haurum, S. Lowe, J. McKeown, C. Ho, J. McLeod, Y.-Y. Wei *et al.*, “A step towards worldwide biodiversity assessment: The bioscan-1m insect dataset,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[11] G. C. Griffiths, “On the foundations of biological systematics,” *Acta biotheoretica*, vol. 23, no. 3-4, pp. 85–131, 1974.

[12] A. V. Brower and R. T. Schuh, *Biological systematics: principles and applications*. Cornell University Press, 2021.

[13] P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. DeWaard, “Biological identifications through dna barcodes,” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. 1512, pp. 313–321, 2003.

[14] T. Lin, “Focal loss for dense object detection,” *arXiv preprint arXiv:1708.02002*, 2017.

[15] C. G. Northcutt, A. Athalye, and J. Mueller, “Pervasive label errors in test sets destabilize machine learning benchmarks,” *Advances in Neural Information Processing Systems*, 2021.

[16] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[18] L. Deng, “The mnist database of handwritten digit images for machine learning research,” *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[19] I. J. Good, “Rational decisions,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 14, no. 1, pp. 107–114, 1952.

[20] A. Ghosh, H. Kumar, and P. S. 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.

[21] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, “Normalized loss functions for deep learning with noisy labels,” in *International conference on machine learning*. PMLR, 2020, pp. 6543–6553.

[22] X. Ye, X. Li, T. Liu, Y. Sun, W. Tong *et al.*, “Active negative loss functions for learning with noisy labels,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 6917–6940, 2023.

[23] N. Pellegrino, N. Zhao, and P. Fieguth, “The effects of label errors in training data on model performance and overfitting,” *Journal of Computational Vision and Imaging Systems*, vol. 9, no. 1, pp. 26–29, 2023.

[24] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[25] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[26] C. J. Van Rijsbergen, “Information retrieval. 2nd. newton, ma,” 1979.

[27] N. Chinchor and B. M. Sundheim, “Muc-5 evaluation metrics,” in *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993.