Imbalanced Image Classification with Complemented Cross Entropy

Yechan Kim[1] Moongu Jeon[1,2]

[1]School of Electrical Engineering and Computer Science, [2]Korea Culture Technology Institute

Gwangju Institute of Science and Technology (GIST)

Gwangju, South Korea

yechankim@gm.gist.ac.kr, mgjeon@gist.ac.kr

***Abstract*—Recently, deep learning models have achieved great success in computer vision applications, relying on large-scale balanced datasets. However, imbalanced class distributions which are far from such datasets still limit the wide applicability of these models due to degradation in performance. To solve this problem, we focus on the study of the cross entropy that mostly ignores output scores on incorrect classes. In this work, we discover that leveraging predicted probabilities on incorrect classes helps improve the performance of imbalanced image classification. This paper proposes a simple but effective loss named Complement Entropy Loss (CCE) based on our finding. Our loss is complemented by entropy of incorrect classes for neutralizing their predicted probabilities without additional training procedures, *e.g.*, bi-objective optimization. Along with it, this loss facilitates the deep learning model to learn the key information especially from samples on minority classes for accurate and robust classification results. It also prevents several disadvantages from minority classes during updating the network parameters. Extensive experiments on imbalanced datasets demonstrate the effectiveness of our method compared to other state-of-the-arts.**

***Index Terms*—deep learning, class imbalance, classification, loss function, training objective, complement entropy**

I. Introduction

In recent years, computer vision algorithms led by deep neural networks (DNNs) have achieved remarkable success in many tasks such as image classification [1–4], object detection [5, 6], and text recognition [7, 8]. Such widespread adoption is attributable to the existence of large-scale datasets with a vast number of annotations. However, various emerging datasets typically exhibit extremely imbalanced class distributions, which largely limits the capability of the DNN model in terms of generalization. Although such imbalanced distributions in the existing real-world data is obviously a crucial challenge, not much research has been conducted yet.

To solve this issue, one common strategy is to resample the dataset, *e.g.*, oversampling on minority classes [9–11], undersampling on majority classes [12–14], and a hybrid of both [15, 16]. Another approach is to employ cost sensitive learning, *e.g.*, reweighting sample-wise loss proportionally to the class-wise inverse frequency, and penalizing hard-classified samples (typically, minority classes) by assigning relatively higher loss [17, 18]. However, these approaches typically neglect the fact that samples on minority classes may have noise or false annotation. This means that training criterion which

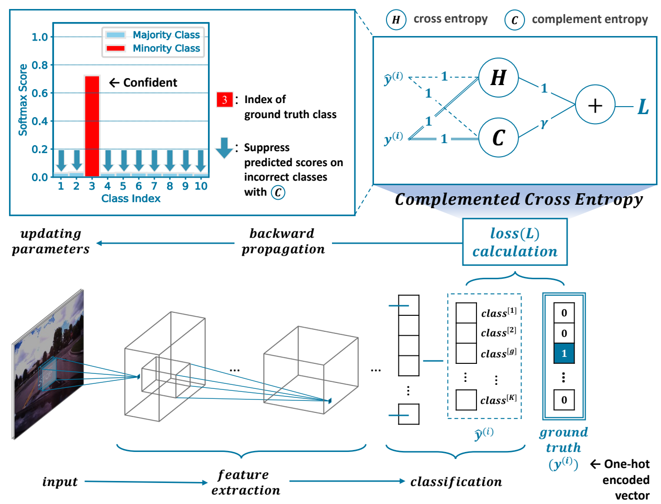


Fig. 1: An overview of training a classification model with the proposed loss function, named *Complemented Cross Entropy* (CCE). CCE attempts to suppress predicted probabilities on incorrect classes. It encourages the model to extract the key information from samples on minority classes and to ensure robust classification results.

largely focuses on the minority classes rather than majority classes might cause poor generalization in performance [19].

In order to address this problem, we revisit the cross entropy as primary objective function and observe many degradation problems in imbalanced datasets. To define cross entropy, let $\mathbf{y}^{(i)}$ be the label of one-hot encoded $(i)^{th}$ vector, $\hat{\mathbf{y}}^{(i)}$ be the class-wise estimated probability vector for given input sample $(\mathbf{x}^{(i)})$. Cross entropy, $H(\mathbf{y}, \hat{\mathbf{y}})$ can be defined as eq. (1):

$$\begin{aligned} H(\mathbf{y}, \hat{\mathbf{y}}) &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\mathbf{y}^{(i)[j]} \times \log \hat{\mathbf{y}}^{(i)[j]} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\mathbf{y}^{(i)T} \cdot \log \hat{\mathbf{y}}^{(i)} \\ &= -\frac{1}{N}\sum_{i=1}^{N}\log \hat{\mathbf{y}}^{(i)[g]} \end{aligned} \tag{1}$$

where g denotes the index of the ground truth class and $\hat{\mathbf{y}}^{(i)[g]}$

represents the predicted probability of the class g on given input $(i)^{th}$ sample (N: the number of samples in a mini-batch, K: the number of classes). While the use of cross entropy as a loss function is fairly widespread among deep learning classification tasks, one limitation exists in this entropy: All softmax probabilities on incorrect classes ($\hat{\mathbf{y}}^{(i)[j \neq g]}$) are treated as zero, since $\mathbf{y}^{(i)[j \neq g]}$ is always zero. It means that $\hat{\mathbf{y}}^{(i)[j \neq g]}$ in the cross entropy is totally ignored so that inaccurate estimation probabilities may produce a cumulative error. In order to overcome this problem of imbalanced classification, complement objective training (COT) was proposed by Chen *et al.*, where the core idea is leveraging information not only from correct but also from incorrect classes [20].

In this work, we introduce a novel loss function named **complemented cross entropy (CCE)** to tackle such performance degradation problem on imbalanced dataset. Our method does not require additional augmentation of samples or upscaling loss scales for the minority classes. Instead, the proposed method utilizes information on incorrect classes to train a robust classification model for imbalanced class distribution. Therefore, we argue that this strategy provides better learning chances for samples on minority classes because it encourages the correct class (including minority one) to overwhelm its softmax score across all the other "incorrect" classes.

Our key contributions are as follows: (i) We present a novel training loss function for imbalanced classification. It reduces the risk of overfitting or losing discriminative information on majority classes; (ii) We experimentally demonstrate the effectiveness of the proposed method for classification on imbalanced datasets.

In the further sections, we first introduce a concept of complement entropy and a training algorithm for imbalanced image classification in section II. We then present the experimental results of the proposed method in section III. Finally, conclusion and discussion are described in section IV.

II. Proposed Approach

In this section, we first provide a brief concept of complement objective training. We then propose our complemented cross entropy loss for imbalanced image classification.

A. Complement Entropy

Complement entropy, $C(\mathbf{y}, \hat{\mathbf{y}})$ calculates the mean of sample-wise entropy on incorrect classes in one single batch. The formulation can be defined as eq. (2):

$$C(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1, j \neq g}^{K} \frac{\hat{\mathbf{y}}^{(i)[j]}}{1 - \hat{\mathbf{y}}^{(i)[g]}} \times \log \frac{\hat{\mathbf{y}}^{(i)[j]}}{1 - \hat{\mathbf{y}}^{(i)[g]}} \quad (2)$$

where $\frac{1}{1-\hat{\mathbf{y}}^{(i)[g]}}$ is the normalizing factor. The inverse of $(1 - \hat{\mathbf{y}}^{(i)[g]})$ normalizes $\hat{\mathbf{y}}^{(i)[j]}$ to make $C(\mathbf{y}, \hat{\mathbf{y}})$ imply the information underlying probability distribution on just incorrect classes. The purpose of this entropy is to encourage the predicted probability of the ground truth class ($\hat{\mathbf{y}}^{(i)[g]}$) to be larger among the other incorrect classes. One way to achieve this goal is by flattening softmax scores on the incorrect classes. This means that the more we neutralize the distribution of predicted probabilities for the incorrect classes, the more confident the prediction for the correct class becomes. To this end, the optimizer of each model maximizes complement entropy, since the entropy becomes maximized when the probability distribution is uniform.

Our work is motivated by the concept of complement entropy. Since the predicted probability on the "correct and minority" class is less vulnerable to the probabilities on the other "incorrect and majority" classes with adopting this concept, it enables the model to find better hidden patterns in samples for the minority classes.

B. Complement Objective Training (COT)

Algorithm 1 describes the original form of COT. Before going further, we first define balanced complement entropy. This entropy, $C'(\mathbf{y}, \hat{\mathbf{y}})$ is designed to match the scale between cross entropy and complement entropy. The formulation can be defined as eq. (3):

$$C'(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{K-1} \times C(\mathbf{y}, \hat{\mathbf{y}}) \quad (3)$$

where $\frac{1}{K-1}$ is the balancing factor. At each iteration in training, cross entropy, $H(\mathbf{y}, \hat{\mathbf{y}})$ is first used to update the model parameters; (balanced) complement entropy, $C'(\mathbf{y}, \hat{\mathbf{y}})$ is then needed to update the parameters again. Extensive experiments have already been conducted by Chen *et al.* and they demonstrate the effectiveness of complementing cross entropy with complement entropy for stable training [20]. Despite its efficacy, it has one crucial limitation: it induces a training time approximately two times longer due to the required additional back-propagation in this training mechanism. On the other hand, we propose a single training loss function that efficiently performs like the training objective in the COT. It also makes the model optimizer to back-propagate only once rather than twice at each iteration.

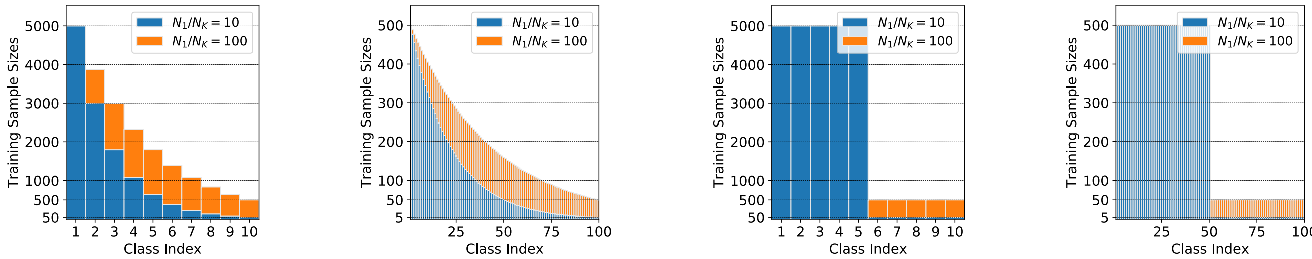
Algorithm 1: Training with a bi-objective concept: cross entropy and complement entropy (COT)

1 **for** $t \leftarrow 1$ **to** n_{train_steps} **do**
2 $\quad \mathbf{x}, \mathbf{y} \leftarrow$ mini_batch(t);
3 $\quad \hat{\mathbf{y}} \leftarrow$ model($\mathbf{x}, \mathbf{y}$);
4 $\quad$ cross_entropy $\leftarrow H(\mathbf{y}, \hat{\mathbf{y}})$;
5 $\quad$ complement_entropy $\leftarrow C'(\mathbf{y}, \hat{\mathbf{y}})$;
6 $\quad$ primary_optimizer.step(∂ cross_entropy);
7 $\quad$ secondary_optimizer.step(∂ complement_entropy);

C. Complemented Cross Entropy (CCE)

In contrast to COT, we replace training process by combination of cross entropy and complement entropy with one single entropy (see line 6). Algorithm 2 depicts a training procedure with our loss function. For modulating the complement entropy, we add γ to the complement entropy as eq. (4):

$$\tilde{C}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\gamma}{K-1} \times C(\mathbf{y}, \hat{\mathbf{y}}) \quad (4)$$



(a) CIFAR-LT-10 (b) CIFAR-LT-100 (c) CIFAR-Step-10 (d) CIFAR-Step-100

Fig. 2: Class-wise sample distributions on imbalanced variants of CIFAR-10/100. CIFAR-LT-10/100 exhibit long-tailed distribution (a, b); while CIFAR-Step-10/100 show step distribution in (c, d).

where the modulating factor, γ should be tuned to decide the amount that complements cross entropy, *e.g.*, $\gamma = 5$.

Algorithm 2: Training with CCE loss (Proposed)

for $t \leftarrow 1$ **to** n_{train_steps} **do**
  $\mathbf{x}, \mathbf{y} \leftarrow$ mini_batch(t);
  $\hat{\mathbf{y}} \leftarrow$ model($\mathbf{x}, \mathbf{y}$);
  cross_entropy $\leftarrow H(\mathbf{y}, \hat{\mathbf{y}})$;
  complement_entropy $\leftarrow \tilde{C}(\mathbf{y}, \hat{\mathbf{y}})$;
  final_loss $\leftarrow$ cross_entropy + complement_entropy;
  optimizer.step(∂ final_loss);

III. Experiments

We conduct various experiments on image classification. In this section, we briefly overview the datasets and implementation details. We then present the experimental results.

A. Datasets

Imbalanced CIFAR-10 and 100. CIFAR-10 and 100 [21] contain RGB images of real-world things (32×32 pixels): 50,000 samples for training and 10,000 samples for testing. The number of classes in both datasets is 10 and 100 respectively. The original version of CIFAR-10 and 100 is completely class-balanced. In order to conduct the experiments, we construct imbalanced variants of CIFAR-10 and 100 by removing samples randomly. More precisely, we first define *imbalance ratio* as $\frac{N_1}{N_K}$, where (i) N_i denotes training sample sizes of class index, i; (ii) K is the maximum number of index; (iii) each class index is sorted in descending order by samples sizes. We then consider two types of imbalance in exactly a same way as [22]: (a) *long-tailed distribution*: we make the training sets to follow an exponential decay distribution in sample sizes per each class.; (b) *step distribution*: we make all majority classes to own the same size of samples, so does minority classes (see fig. 2). The testing sets still have a class-balanced distribution. We only apply zero padding, random cropping, and horizontal flipping to the training sets, not the testing sets. We normalize both training sets and testing sets with mean and variance of the training ones.

TABLE I: The number of samples per class on Road Marking Dataset: It has "long-tailed' distribution originally.

Class	Sample Sizes	Class	Sample Sizes
leftturn*	705	forward & left	6
35*	112	clear	6
rightturn*	101	keep	6
rail*	90	stripe	3
forward*	80	school	3
40*	69	hump	3
xing*	64	30	2
ped*	54	slow	2
stop*	49	speed	2
bike	41	car	1
25	15	diamond	1
forward & right	13	lane	1
yield	7	pool	1
X-crossing	6		**(Total: 1,443)**

Road Marking Dataset. It is consisted of colored 1,443 samples on road markings such as "35," "40," "forward," and "stop" [23]. The number of classes in this dataset is 27. All existing examples were taken on clear and sunny days. Table I describes a class-wise sample distribution which is originally long-tailed. For our image classification experiments, we crop out the backgrounds in all images with given annotations. Also, we only use examples on classes marked '*' on table I for fair comparison of our method with the other state-of-the-arts results. We then split the data into training set and testing set at a ratio of 8:2. We make the testing set to still have a class-balanced distribution. We only apply zero padding, random cropping, and horizontal flipping to the training set, not the testing set. We also normalize both training set and testing set with mean and variance of the training one.

B. Experimental Setup

Baselines. We compare the proposed method (CCE) with the following techniques: (i) empirical risk minimization (ERM): we train models with only standard softmax cross entropy; (ii) complement objective training (COT): there are two optimizers for training: one for softmax cross entropy and the other one for softmax complement entropy; (iii) focal loss (FL): it uses sigmoid cross entropy with a modulating factor to concentrate on hard samples [17].

Training Details. We train each convolutional neural networks (CNNs) with mini-batch size 128, stochastic gradient descent (SGD) optimizer, where momentum of weight is 0.9 and weight decay is 5e-4. We set the maximum number of training epochs to 200 for CIFAR and to 100 for Road Marking Dataset. The CNNs used for our experiments are as follows: ResNet, MobileNet, SqueezeNet, ResNeXt, and EfficientNet [4, 24–27]. The learning rate is initially set to 1e-1, and dropped by a factor of 0.5 at 60, 120, and 160 epochs in the same manner of [28]. We also take the "linear learning rate warm-up" strategy [29] in the early 5 epochs. Hyperparameter, γ should be tuned for CCE: we set γ to 5 over all experiments. All the models are implemented by PyTorch framework [30]. The experiments are conducted on Nvidia GTX Titan X (GPU) and AMD Ryzen 7 3700X (CPU).

C. Experimental Results on CIFAR

This subsection presents the classification experimental results of the proposed loss on variants of CIFAR. To begin with, we first conduct experiments of the proposed loss, CCE on the original CIFAR in order to demonstrate the effectiveness of ours not only in class-imbalanced datasets but also in balanced sets. We then perform experiments of ours on imbalanced variants of CIFAR. As table II, III, and IV show, our loss outperforms the other methods in all experiments. We also observe that performance is improved by replacing the cross entropy to COT, which is our motivation in this work. Note that the proposed loss, CCE provides further enhancements in performance by exploiting entropy on predicted probabilities of incorrect classes per each mini-batch, demanding for less time-complexity than the COT.

TABLE II: Classification test accuracy (%) on balanced CIFAR-10 and 100.

	CIFAR-10			CIFAR-100		
	ERM	COT	**CCE**	ERM	COT	**CCE**
ResNet-18	93.86	93.92	**94.19**	76.68	75.85	**76.89**
ResNet-32	94.47	94.63	**94.80**	76.18	76.88	**77.19**
ResNet-50	94.35	94.50	**94.54**	77.14	76.48	**77.22**
ResNeXt-50	87.20	**87.76**	87.62	63.32	62.78	**64.10**
SqueezeNet	91.05	91.18	**91.23**	69.86	69.89	**69.92**
MobileNet	82.76	82.84	**82.91**	51.73	50.33	**52.85**

TABLE III: Classification test accuracy (%) on imbalanced variants of CIFAR-10 with ResNet-32.

	CIFAR-10			
	Long-tailed		Step	
Imbalance Ratio ($\frac{N_1}{N_K}$)	10	100	10	100
ERM	73.42	87.21	65.17	85.24
FL	73.54	88.16	65.31	85.52
COT	73.38	88.02	65.32	85.40
CCE	**74.64**	**88.37**	**65.69**	**86.73**

TABLE IV: Classification test accuracy (%) on imbalanced variants of CIFAR-100 with ResNet-32.

	CIFAR-100			
	Long-tailed		Step	
Imbalance Ratio ($\frac{N_1}{N_K}$)	10	100	10	100
ERM	43.49	62.35	40.77	60.17
FL	43.68	63.10	40.85	60.94
COT	43.54	62.59	40.74	60.42
CCE	**43.98**	**63.12**	**40.85**	**61.02**

D. Experimental Results on Road Marking

This subsection shows the classification experimental results of our method on Road Marking Dataset. We conduct experiments of our loss on various CNN models such as ResNet and EfficientNet. As table V and VI present, the proposed loss has also shown improvements in terms of accuracy on Road Marking Dataset. Especially in ResNet-101, our loss achieves significant performance improvement of 11.02% in terms of classification accuracy, compared to the ERM (see fig. 3).

TABLE V: Classification test accuracy (%) on Road Marking.

Model	ERM	COT	**CCE**
ResNet-50	98.78	98.78	**99.18**
ResNet-101	88.16	98.78	**99.18**
SqueezeNet	94.69	96.53	**96.53**
EfficientNet_b0	98.78	98.78	**99.80**
EfficientNet_b1	99.18	99.18	**99.39**
EfficientNet_b7	99.59	99.80	**99.80**

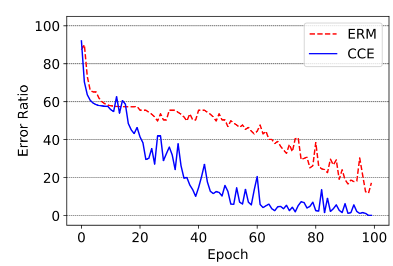


Fig. 3: Classification test error (%) of ResNet-101 on Road Marking over epochs.

TABLE VI: Comparison on results of our method with other state-of-the-arts methods on Road Marking.

Method	Test Accuracy (%)
Lee *et al.* (AlexNet + ERM) [6]	94.70
Lee *et al.* (GAN + Augmentation) [6]	98.80
Bailo *et al.* (PCANet + Logistic Regression) [31]	98.90
Balio *et al.* (PCANet + SVM) [31]	99.10
Ahmad *et al.* (LeNet$_{96}$ CP$_2$) [32]	99.05
Ours with Best	**99.80**

IV. Conclusion

In this paper, we presented a novel training loss function named Complemented Entropy Loss (CCE) for imbalanced image classification. We prove that exploiting predicted probabilities on incorrect classes helps the model to extract discriminative information especially from samples on minority classes and to prevent overfitting or performance degradation in class-imbalanced datasets. We believe that our work will open a new research direction to address various issues in imbalanced learning. In the future, we would like to extend our training method in other computer vision tasks such as object detection and visual question answering.

Acknowledgment

This work was supported by English II: Research Writing in Science and Engineering (GIST) in summer session, 2020. We sincerely appreciate professor. Paulina Martinez for her academic feedback on this paper.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[3] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[6] Younkwan Lee, Juhyun Lee, Yoojin Hong, YeongMin Ko, and Moongu Jeon. Unconstrained road marking recognition with generative adversarial networks. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1414–1419. IEEE, 2019.

[7] Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5571–5579, 2018.

[8] Younkwan Lee, Juhyun Lee, Hoyeon Ahn, and Moongu Jeon. Snider: Single noisy image denoising and rectification for improving license plate recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–10, 2019.

[9] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[10] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[11] Huaxiang Zhang and Mingfang Li. Rwo-sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion*, 20:99–116, 2014.

[12] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003.

[13] Show-Jane Yen and Yue-Shi Lee. Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009.

[14] Qi Fan, Zhe Wang, and Daqi Gao. One-sided dynamic undersampling no-propagation neural networks for imbalance problem. *Engineering Applications of Artificial Intelligence*, 53:62–73, 2016.

[15] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[16] Min Zeng, Beiji Zou, Faran Wei, Xiyao Liu, and Lei Wang. Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data. In *2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS)*, pages 225–228. IEEE, 2016.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[18] Serim Ryou, Seong-Gyun Jeong, and Pietro Perona. Anchor loss: Modulating loss scale based on prediction difficulty. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5992–6001, 2019.

[19] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343, 2018.

[20] Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Complement objective training. *arXiv preprint arXiv:1903.01182*, 2019.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[22] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019.

[23] Tao Wu and Ananth Ranganathan. A practical system for road marking detection and recognition. In *2012 IEEE Intelligent Vehicles Symposium*, pages 25–30. IEEE, 2012.

[24] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[25] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[27] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[28] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[29] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[31] Oleksandr Bailo, Seokju Lee, Francois Rameau, Jae Shin Yoon, and In So Kweon. Robust road marking detection and recognition using density-based grouping and machine learning techniques. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 760–768. IEEE, 2017.

[32] Touqeer Ahmad, David Ilstrup, Ebrahim Emami, and George Bebis. Symbolic road marking recognition using convolutional neural networks. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1428–1433. IEEE, 2017.