Supplementary Material
RankMixup: Ranking-based Mixup Training for Network Calibration

Jongyoun Noh Hyekang Park Junghyup Lee Bumsub Ham*

School of Electrical and Electronic Engineering, Yonsei University

1. Implementation details

Methods for comparison. We provide details of the hyperparameters used for reproducing the results of other methods as follows: (a) ECP [9]: We set the weight of entropy penalty to 0.1 following [9] (b) LS [8]: Following [5, 7], we report the results obtained with $\alpha = 0.05$. (c) FL [7]: We train the models with a fixed regularization parameter (γ) of 3. (d) Mixup [11]: We use a shape parameter α of 0.2, the best performing one in [11]. (e) FLSD [7]: Following the schedule in [7], we use the parameter γ of 5 for the samples whose output probability for the ground-truth class is within $[0, 0.2)$, otherwise we use the parameter γ of 3. (f) CRL [6]: We set the balancing parameter of the ranking loss as 1.0. (g) CPC [1]: We set the weights of binary discrimination and binary exclusion losses as 0.1 and 1.0, respectively. (h) MbLS [5]: We train the models with margins of 6 and 10 for the CIFAR10/100 datasets and Tiny-ImageNet dataset, respectively. (i) RegMixup [10]: We obtain the results with a shape parameter α of 10.0 as suggested in [10].

2. More results

Q and α. We perform experiments with various combinations of α and Q to further investigate the importance of diverse samples. We adopt the ResNet-50 [2] models trained with M-NDCG for the analyses, and show the calibration performances in terms of ECE and AECE on the validation sets of CIFAR10 [3] and Tiny-ImageNet [4] in Fig. 1 and 2, respectively. From the figures, we can observe three things: (1) Our models using α within the range of $[1, 3]$ achieve the best performances in most cases. As shown in Fig. 2 of the main paper, such distributions enable the model to sample more diverse and large mixup coefficients λ, suggesting that generating diverse samples with relatively strong interpolations is crucial for our framework. (2) Our models also perform better with larger αs than smaller ones, in contrast to vanilla mixup [11, 12]. However, in some cases, the performances worsen with very large α. (3) For most α values, both ECE and AECE improve as more augmented samples are used. This indicates that incorporating more diverse ranking relationships based on augmented samples is favorable in our framework. However, using more samples with very small or large α leads to degraded performance. Overall, these observations highlight the importance of using diverse samples and considering a range of α values for effective calibration in our framework.

References

[1] Jiacheng Cheng and Nuno Vasconcelos. Calibrating deep neural networks by pairwise constraints. In *CVPR*, 2022. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 2

[4] Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. *CS 231N*, 2015. 1, 2

[5] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *CVPR*, 2022. 1

[6] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *ICML*, 2020. 1

[7] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *NeurIPS*, 2020. 1

[8] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019. 1

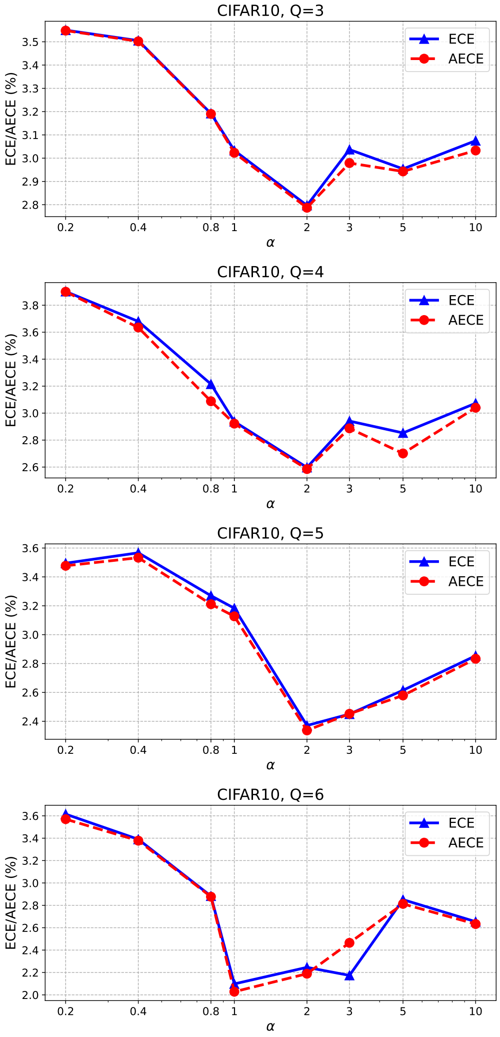
[9] Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *ICLR Workshop*, 2017. 1

[10] Francesco Pinto, Harry Yang, Ser-Nam Lim, Philip Torr, and Puneet K. Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *NeurIPS*, 2022. 1

[11] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019. 1

[12] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1

*Corresponding author.



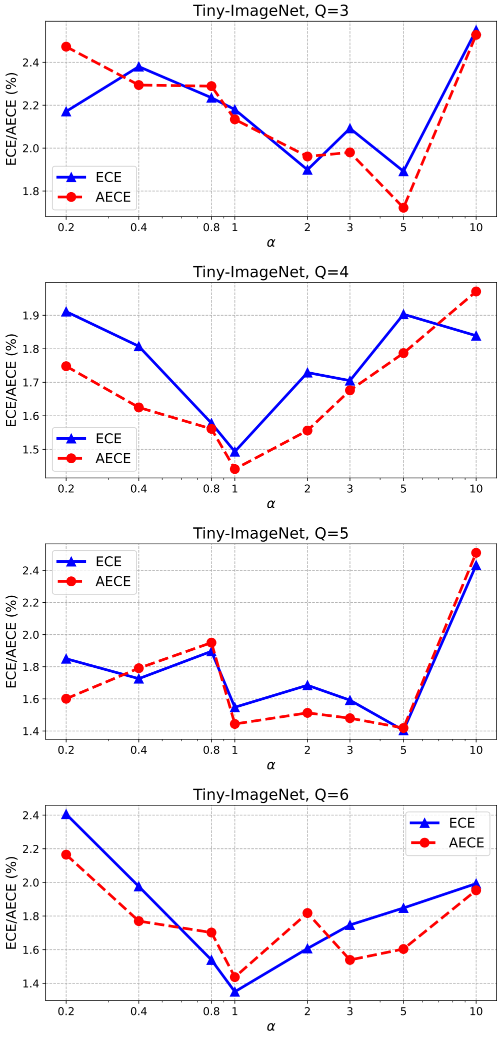


Figure 1: Quantitative results with various combinations of the parameter of *Beta* distribution (α) and the number of aumgented samples (Q). We plot the variation of both ECE (%) and AECE (%) on the validation split of CIFAR10 [3]. Best viewed in color.

Figure 2: Quantitative results with various combinations of the parameter of *Beta* distribution (α) and the number of aumgented samples (Q). We plot the variation of both ECE (%) and AECE (%) on the validation split of Tiny-ImageNet [4]. Best viewed in color.