Appendix for The Devil is in the Margin: Margin-based Label Smoothing for Network Calibration

Bingyuan Liu[1]*, Ismail Ben Ayed[1], Adrian Galdran[2], Jose Dolz[1]

[1]ÉTS Montreal, Canada [2]Universitat Pompeu Fabra, Barcelona, Spain

A. Proof

Here we provide more details for the proof of Proposition 1 in the main text.

Proposition 1. *A linear penalty (or a Lagrangian) for constraint* $\mathbf{d}(\mathbf{l}) = \mathbf{0}$ *is bounded from above and below by* $\mathcal{D}_{KL}(\mathbf{u}||\mathbf{s})$*, up to additive constants:*

$$\mathcal{D}_{KL}(\mathbf{u}||\mathbf{s}) - \log(K) \overset{\text{c}}{\leq} \frac{1}{K}\sum_k (\max_j(l_j) - l_k) \overset{\text{c}}{\leq} \mathcal{D}_{KL}(\mathbf{u}||\mathbf{s})$$

where $\overset{\text{c}}{\leq}$ *stands for inequality up to an additive constant.*

Proof. Given the expression of the KL divergence:

$$\mathcal{D}_{\text{KL}}(\mathbf{u}||\mathbf{s}) = \frac{1}{K}\sum_k \log\left(\frac{1/K}{s_k}\right) \overset{\text{c}}{=} -\frac{1}{K}\sum_k \log(s_k)$$

where $\overset{\text{c}}{=}$ stands for equality up to an additive and/or non-negative multiplicative constants and $\mathbf{u}$ is the uniform distribution, and given the definition of softmax function:

$$s_k = \frac{e^{l_k}}{\sum_j^K e^{l_j}}$$

we have:

$$\begin{aligned}\mathcal{D}_{\text{KL}}(\mathbf{u}||\mathbf{s}) &\overset{\text{c}}{=} -\frac{1}{K}\sum_k \log\left(\frac{e^{l_k}}{\sum_j^K e^{l_j}}\right) \\ &= \frac{1}{K}\sum_k^K \left(\log\sum_j^K e^{l_j} - l_k\right)\end{aligned}$$

Then, considering the following well-known property of the LogSumExp function:

$$\max_j(l_j) \leq \log\sum_j^K e^{l_j} \leq \max_j(l_j) + \log(K)$$

We obtain :

$$\mathcal{D}_{\text{KL}}(\mathbf{u}||\mathbf{s}) - \log(K) \overset{\text{c}}{\leq} \frac{1}{K}\sum_k (\max_j(l_j) - l_k) \overset{\text{c}}{\leq} \mathcal{D}_{\text{KL}}(\mathbf{u}||\mathbf{s})$$

Furthermore, given the definition of the logit distances, i.e., $\mathbf{d}(\mathbf{l}) = (\max_j(l_j) - l_k)_{1\leq k\leq K} \in \mathbb{R}^K$, the penalty term, $\mathcal{D}_{\text{KL}}(\mathbf{u}||\mathbf{s})$, imposed by Label Smoothing (LS) is approximately optimizing a linear penalty (or a Lagrangian) for logit distance constraint:

$$\mathbf{d}(\mathbf{l}) = \mathbf{0}$$

which encourages equality of all logits.

□

B. Dataset Description and Implementation Details

In this section, we present the description of all the datasets used in our experiments, as well as the related implementation details.

CIFAR-10 [5] is an image classification dataset that includes a total of 60,000 images with size 32×32, divided equally into 10 classes. In our experiments, we use the standard train/validation/test split containing 45,000/5,000/10,000 images, respectively. During the experiments, we fixed the batch size to 128 and use SGD optimizer with a momentum of 0.9. The number of training epochs is set to 350, with a multi-step learning rate decay strategy, i.e., learning rate of 0.1 for the first 150 epochs, 0.01 for the next 100 epochs and 0.01 for the last 100 epochs. Data augmentation techniques like random crops and random horizontal flips are applied on the training set.

Tiny-ImageNet [1] is a subset of ImageNet containing 64×64 dimensional images, with 200 classes and 500 images per class in the training set, and 50 images per class in the validation set. Following the setting in [8], we use 50 samples per class (a total of 10,000 samples) from the training set as a validation set and the original validation set as a test set. The batch size is set to 64. We train for 100 epochs

*Corresponding author: bingyuan.liu@etsmtl.ca

Table 1. ECE for different methods with pre- and post-temperature scaling. Optimal T is indicated in brackets.

Dataset	Model	CE		LS		FL		FLSD		Ours	
		PreT	PosT	PreT	PosT	PreT	PosT	PreT	PosT	PreT	PosT
Tiny-ImageNet	R-50	3.73	1.86 (1.1)	3.17	1.79 (0.9)	2.96	1.74 (0.9)	2.91	1.74 (0.9)	**1.64**	**1.64** (1.0)
	R-101	4.97	2.01 (1.2)	2.20	2.20 (1)	2.55	2.22 (0.9)	4.91	1.64 (0.9)	**1.62**	**1.62** (1.0)
CIFAR-10	R-50	5.85	2.34 (3.9)	2.79	1.75 (0.9)	3.90	1.34 (0.7)	3.84	1.30 (0.7)	**1.16**	**1.16** (1.0)
	R-101	5.74	2.51 (3.9)	3.56	2.71 (0.9)	4.60	1.24 (1.4)	4.58	1.21 (1.9)	**1.38**	**1.13** (0.9)
CUB-200-2011	R-101	6.75	2.00 (1.2)	5.16	3.05 (0.9)	8.41	2.45 (0.8)	8.54	3.61 (3.8)	**2.78**	**1.72** (1.2)
20 News	GPCN	22.75	3.01 (3.1)	8.07	3.69 (1.2)	10.80	3.33 (1.4)	10.87	4.10 (1.4)	**5.40**	**2.09** (1.1)

with a learning rate of 0.1 for the first 40 epochs, of 0.01 for the next 20 epochs and of 0.001 for the last 40 epochs.

CUB-200-2011 [10] is the most popular fine-grained benchmarking dataset. As an extended version of the CUB-200 dataset, with roughly double the number of images per class and new part location annotations, it consists of $5,994$ training and $5,794$ test images, belonging to 200 bird species. We augment the images during training, i.e., we resize the images to 256×256 and then randomly crop patches of 224×224 from the scaled images or their horizontal flip as inputs. We initialize the model by pre-trained weights on ImageNet and then train on this dataset for 200 epochs. The batch size is set to 16 and SGD optimizer is used with a momentum of 0.9. The learning rate is initialized as 0.1 and decayed by a factor of 0.1 every 80 epochs. Note that, for margin m, we used the optimal m found on the validation set of Tiny-ImageNet (we did not use a validation set for CUB-200-2011).

PASCAL VOC 2012 [3] semantic segmentation benchmark contains 20 foreground object classes and one background class. The data is split into $1,464$ images for training, $1,449$ for validation and $1,456$ for testing. Note that the calibration performance on test set is unavailable, as the ground-truth on test set is not publicly released. Therefore, we only report the performances on validation set by using the best hyper-parameters found on the Tiny-ImageNet classification benchmark for all the methods, without any further tuning on the segmentation validation set. During training, we randomly crop the images to a 512 resolution, and apply other augmentations such as random horizontal flip, random brightness changes or contrast transformation. To train the segmentation model, we employ the popular public library[1], where the encoder is initialized with the weights pre-trained on ImageNet and the decoder is trained from scratch. The batch size is set to 8, and the momentum of the SGD optimizer to 0.9. The learning rate is initialized as 0.01, and decayed by a factor of 0.1 every 40 epochs. Finally, the network is trained for 100 epochs.

20 Newsgroups [6] is a popular text classification benchmark, containing 20,000 news articles, which are categorised evenly into 20 different groups based on their content. While some of the groups are significantly related (e.g. rec.motorcycles and rec.autos), other groups are completely unrelated (e.g. sci.space and misc.med). We use the standard train/validation/test split containing 15,098/900/3,999 documents, respectively. To train the Global Pooling Convolutional Network (GPCN) [7], we use Glove word embeddings [9]. Adam is used as optimizer with an initial learning rate of 0.001, and beta values equal to 0.9 and 0.999. The training is performed during 100 epochs, with a learning-rate decay by a factor of 0.1 after the first 50 epochs.

[1] https://github.com/qubvel/segmentation_models.pytorch

C. Ablation study on the balancing weight

We now investigate the impact of the balancing weight λ in our method, and compare it to the effect of α in Label Smoothing (LS), whose results are depicted in Figure 1. In particular, we show the evolution of calibration and classification metrics on Tiny-ImageNet validation and test sets. One may observe that, unlike LS, our method with margin is more robust with respect to the balancing weight in both subsets. Furthermore, the high similarity in the ECE curves of LS and Ours ($m = 0$) support our theoretical connections stating that LS approximates a particular case of the proposed loss when the margin is equal to 0.

D. Results with post temperature scaling

In Table 1, we compare with the method of applying post temperature scaling (PosT) [4] on the outputs of the CE-trained model. As this technique is orthogonal to the learning objectives, we also include the results when applying this post-processing to the proposed method. We can see that the PreT scores obtained by our method outperform the PosT results from CE across all the cases. Furthermore, our method with PosT also achieves the best performance across the datasets and backbones. It is worth noting that the proposed method has optimal temperature values very close to 1 (see Table 1), indicating that our models are already well calibrated. Note that the results of post-hoc scaling might be highly sensitive to the validation sets and data characteristics.

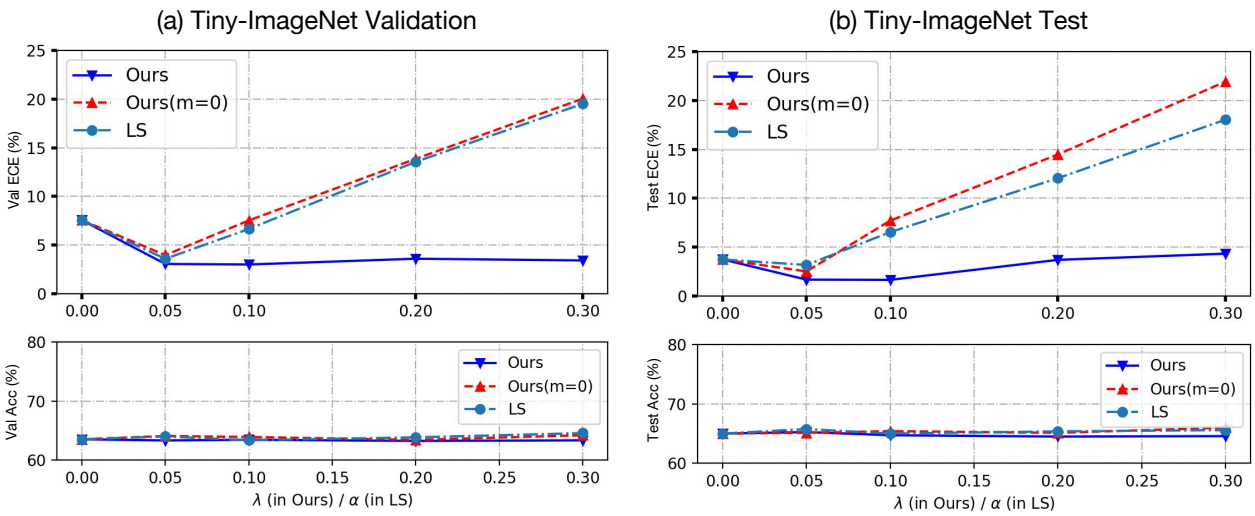

Figure 1. **Evaluating the effect of the balancing weight.** We present the variation of both ECE and Accuracy on the Tiny-ImageNet validation set (*left*) and on Tiny-ImageNet test set (*right*) using different balancing weight values, i.e., λ in our method and α in LS. The network used in this study is ResNet-50.

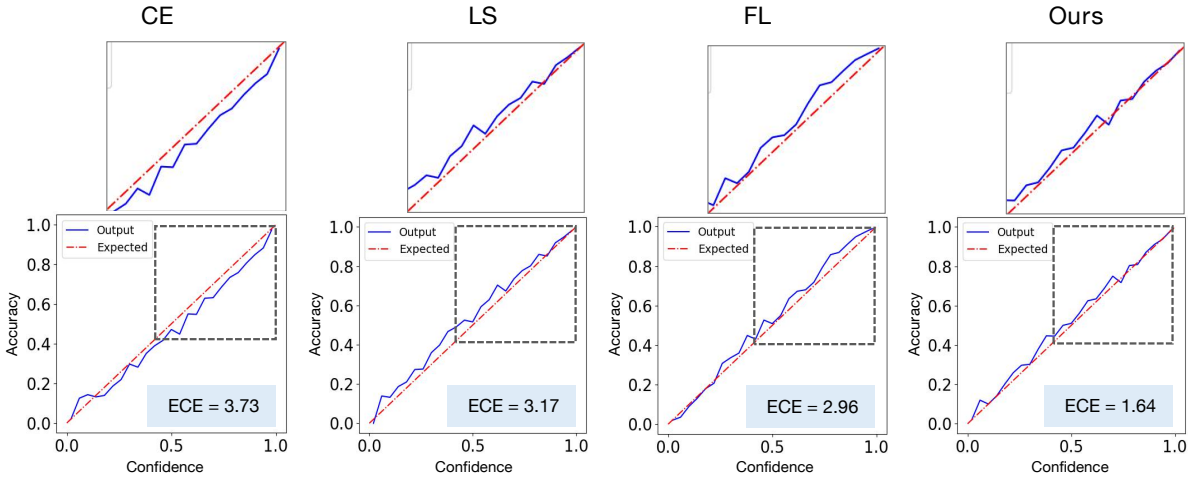

Figure 2. **Calibration visualizations of ResNet-50 on Tiny-ImageNet.** Reliability diagrams is computed with 25 bins. The zoom-in figures for part of the diagrams are also included, clearly showing the differences.

E. Results with Vision Transformers (ViT)

Table 2. Results with Vision Transformer (ViT) model.

Dataset	LS		FL		FLSD		Ours	
	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
CIFAR-10	**98.57**	1.39	98.49	1.20	98.55	1.13	**98.57**	**0.39**
Tiny-ImageNet	90.50	2.37	90.39	4.51	90.47	4.25	**90.65**	**1.26**

The recent study in [11] suggests that newer models, such as vision transformers (ViT) [2], are better calibrated than older models, such as convolutional neural networks. Inspired by these findings, we further evaluate the performance of the proposed method with ViT, whose results are presented in Table 2. In particular, we include the results obtained with a ViT on both CIFAR-10 and Tiny-ImageNet, demonstrating a similar trend, i.e., the proposed approach outperforms other calibration losses. This consolidates the message of this paper and further demonstrates the generalizability of the proposed loss.

F. Reliability diagram.

We further investigate the calibration behaviour of the proposed model with reliability diagrams, whose results for Tiny-ImageNet with ResNet50 are shown in Figure 2. What we expect from a perfectly calibrated model is that its reliability diagram matches the dashed red line, where the output likelihood predicts perfectly the model accuracy. We first observe that the model trained with the standard cross en-

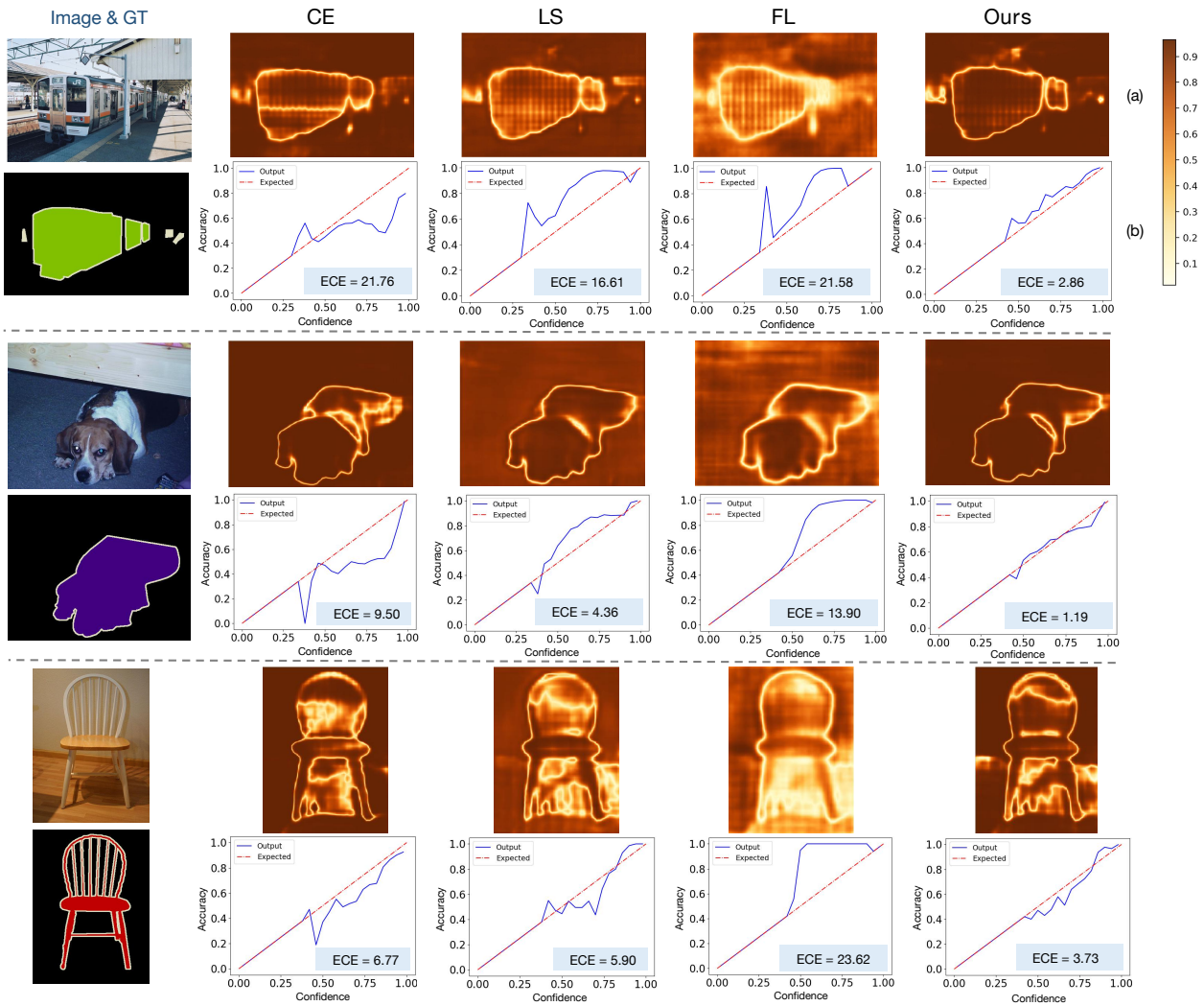

Figure 3. **Additional visual results on semantic segmentation.** We present additional examples from the qualitative segmentation results on the PASCAL VOC 2012 validation set, showing the superiority by our method, in terms of calibration performance. In the left, we give the original image with ground-truth (GT) mask, then we present the **confidence map (a)** and the **reliability diagram (b)** with the ECE (%) score for each method. The value of confidence map represent the predicted confidence, i.e., the element of the soft-max probability for the winner class. It is noted that deeper color denotes higher confidence in the map, as shown in the legend at the upper right corner.

tropy (*first plot*) is overconfident, as its accuracy is mostly below the confidence values. Both state-of-the-art methods (*second and third plots*) reverse this trend, and present reliability diagrams closer to the dashed line, which indicates that models trained with these losses are actually better calibrated. Even though both improve the calibration performance, an interesting observation is that the range accuracy vs confidence where they are better calibrated is indeed the opposite (LS provides better estimates for higher probabilities, whereas FL predictions are better calibrated in a low regime, close to 0). Last, we can observe that the reliability diagram slope provided by our method is much closer to a slope of 1, suggesting that the model is better calibrated. This observation is supported by the quantitative results reported in Section 5.1 of the main text.

G. Additional visual results on segmentation

In Figure 3, we present additional qualitative examples from the VOC segmentation model. As illustrated by the reliability diagrams (*b*) for different methods, our method achieves the best calibration performance. Regarding the confidence maps (*a*), the results from the proposed model are also consistent with the fact that uncertainty occurs mainly on the boundary while confidence is higher within and outside the segmentation regions. Note that all the trends are consistent with the examples shown in Figure 3 of the main text.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[3] Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 2

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 2

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1

[6] Ken Lang. Newsweeder: Learning to filter netnews. In *ICML*, 1995. 2

[7] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *ICML*, 2014. 2

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

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 2

[10] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2

[11] Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In *NeurIPS*, 2021. 3