Towards Unbiased Calibration using Meta-Regularization

Anonymous authors
Paper under double-blind review

Abstract

Model miscalibration has been frequently identified in modern deep neural networks. Recent work aims to improve model calibration directly through a differentiable calibration proxy. However, the calibration produced is often biased due to the binning mechanism. In this work, we propose to learn better-calibrated models via meta-regularization which has two components: (1) gamma network (γ-Net), a meta learner to output sample-wise gamma value (continuous variable) for Focal loss for regularizing the backbone network; (2) smooth expected calibration error (SECE), a Gaussian-kernel based unbiased and differentiable surrogate to ECE that enables the smooth optimization γ-Net. We evaluate the effectiveness of proposed approach in regularizing neural network towards better and unbiased calibration on three computer vision datasets. We empirically demonstrate that: (a) learning sample-wise γ as continuous variables can effectively perform calibration; (b) SECE smoothly optimizes γ-Net towards unbiased and robustness to binning schemes; and (c) the combination of γ-Net and SECE achieves the best calibration performance across various calibration metrics and retains very competitive predictive performance as compared to multiple recently proposed methods.

1 Introduction

Deep Neural Networks (DNNs) have shown promising predictive performance in many domains such as computer vision Krizhevsky et al. (2012), speech recognition Graves et al. (2013) and natural language processing Vaswani et al. (2017). As a result, trained deep neural network models are frequently deployed and utilized in real-world systems. However, recent work Guo et al. (2017) pointed out that those highly accurate, negative log likelihood trained, deep neural networks are often poorly calibrated Niculescu-Mizil & Caruana (2005b). Their predicted class probabilities do not faithfully estimate the true probability of correctness and lead to (primarily) overconfident and under-confident predictions. Deploying such miscalibrated models into real-world systems poses a high risk, particularly when model outputs are directly utilized to serve customers' requests in applications like medical diagnosis Caruana et al. (2015) and autonomous driving Bojarski et al. (2016). Better calibrated model probabilities can be used as an important signal towards more reliable machine learning systems.

Recent trend tends to learn a well calibrated model direct optimising model towards improved calibration, one of representative work is from Kumar et al. (2018), they developed a differentiable equivalent of ECE, the Maximum Mean Calibration Error (MMCE). Mukhoti et al. (2020) found that a different training objective, focal loss, can effectively improve calibration and is a good alternative to regularized cross entropy. They also pointed out that the gamma parameter in focal loss plays a crucial role in making this approach effective, and proposed a sample-dependent schedule for gamma in focal loss (FLSD), which showed superior calibration performance as compared to baselines. Bohdal et al. (2021) proposed using a meta-learning based approach termed as meta-calibration. In their formulation, the backbone network learns to optimize the regularized cross-entropy loss, while a differentiable proxy for calibration error is used to tune the parameters of the weight regularizer.

Motivated by those advancements in this line, we introduce an approach to learn a sample-wise γ value via a meta network (named gamma network, i.e., γ-Net) for the focal loss as presented in Figure 1. Different to Focal loss and FLSD, where a global gamma parameter is used or scheduled during training, we learn a

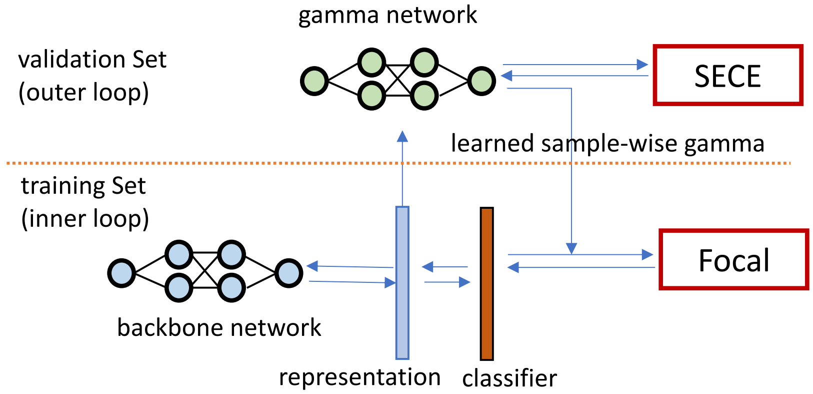


Figure 1: Our proposed approach for regularizing the base network towards better calibration with two new components: γ-Net and SECE . The inner loop optimizes the backbone network (e.g., resnet) which uses focal loss as an objective function. The γ-Net in the outer loop takes the extracted second last layer representation of backbone network as input, learns to output sample-wise γ for focal loss at a continuous space. The γ-Net is optimized by using the proposed SECE, a Gaussian-kernel based unbiased and differentiable calibration error.

gamma parameter for each sample. To achieve this, the γ-Net is optimised using smooth ECE (SECE), a differentiable surrogate to expected calibration error (ECE) Naeini et al. (2015). The output of γ-Net, i.e., the learned γ values, are then used to optimize the backbone network towards better calibration; pushing network weights and biases to accurately assess predictive *confidence* while retaining the original predictive *performance*.

Our contributions can be summarized as follows:

- We propose a γ-Net , as a meta-network to learn sample-wise γ for Focal loss as continuous variables, rather than using a single pre-defined γ value.
- We propose a kernel-based ECE estimator SECE , a variation of the kernel density estimate (KDE) approximation Zhang et al. (2020) to smoothly regularize the γ-Net towards better and unbiased calibration.
- We conduct extensive experiments and analysis to empirically show that γ-Net can effectively calibrate models, and SECE provides stable and smooth calibration. Their combination achieves competitive predictive performance and better scores across multiple calibration metrics as compared to baselines.

2 Related Work

Calibration of ML models using Platt scaling Platt et al. (1999) and Isotonic Regression Zadrozny & Elkan (2002) have showed significant improvements for SVMs and decision trees. With the advent of neural networks, Niculescu-Mizil & Caruana (2005a) showed those methods can produce well-calibrated probabilities even without any dedicated modifications.

Recently, Guo et al. (2017) and Mukhoti et al. (2020) have shown that modern NNs are noticeably more accurate but poorly calibrated due to negative log likelihood (NLL) overfitting. Minderer *et al*Minderer et al. (2021) revisited this problem and found that architectures with large sizes have a larger effect on calibration, but nevertheless, more accurate models tend to produce less calibrated predictions.

One of the ways to tackle calibration is to rely on post-hoc corrections using non-parametric approaches such as histogram binning Zadrozny & Elkan (2001), isotonic regression Zadrozny & Elkan (2002) and parametric methods such as Bayesian binning into quantiles (BBQ) and Platt scaling Platt et al. (1999). Beyond those four, temperature scaling (TS) is a single-parameter extension of Platt scaling Platt et al. (1999) and the most recent addition to the offering of post-hoc methods. Recent work has showed that a single parameter TS leads to good calibration performance with minimal added computational complexity Guo et al. (2017); Minderer et al. (2021). Extensions to temperature scaling include attention-based mechanisms to tackle noise in validation data Mozafari et al. (2018) and introduction of variational dropout based on TS to calibrate deep networks.

Calibration is often measured using the expected calibration error (ECE) - a difference between the model's confidence and accuracy. In-training methods minimize this measure directly by optimizing the ECE or its proxy. Kumar et al. (2018) developed a differentiable equivalent of ECE, the Maximum Mean Calibration Error (MMCE), which can be optimized directly to train calibrated models. Similarly, Bohdal et al. (2021) use a meta-learning approach Luketina et al. (2015) to train a model that minimizes the differentiable ECE (DECE) to find the optimal parameters of the L2 regularizers.

There are also methods that do not explicitly include the calibration objective but rather implicitly guide the training towards better calibration performance. Focal loss Lin et al. (2017) acts as a maximum entropy regularizer Mukhoti et al. (2020), label smoothing Müller et al. (2019) mitigates miscalibration by softening hard labels with an introduced smoothing modifier in the standard loss function (e.g., cross-entropy), while Mix-up training Thulasidasan et al. (2019) extends *mixup* Zhang et al. (2018) to generate synthetic samples during model training by combining two random elements from the dataset.

Most recently, Krishnan & Tickoo (2020) introduced a differentiable accuracy versus uncertainty calibration (AvUC) loss function that allows a model to learn to provide well-calibrated uncertainties. Liu et al. (2022) analyzed some current calibration methods such as label smoothing Müller et al. (2019), focal loss Lin et al. (2017), and explicit confidence penalty Pereyra et al. (2017) from a constrained-optimization perspective and pointed out those can be seen as approximations of a liner penalty. The authors proposed to add a marigin into learning objective. Cheng & Vasconcelos (2022) proposed calibration by pairwise constraints (CPC), which aims to strengthen the calibration supervision by casting multilclass searino into pairwise binary calibration. Thus authors proposed binary discrimination constraints (BDC) loss and binary exclusion constraint (BEC) loss as addition to standard cross-entropy loss to increase the supervision rate for calibration of the training process.

3 Preliminaries

3.1 Model Calibration

Calibration Guo et al. (2017) measures and verifies how the predicted probability estimates the true likelihood of correctness. Assume a model $\mathbf{m}$ trained with dataset $\{\mathbf{x}, y\}, \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}$. $\hat{\mathbf{p}}$ is the predicted softmax probability of all classes. If $\mathbf{m}$ makes 100 independent predictions, each with confidence $p = \arg\max(\hat{\mathbf{p}}) = 0.9$, a calibrated $\mathbf{m}$ is correct 90 times. Formally, $\text{accuracy}(\mathbf{m}(D)) = \text{confidence}(\mathbf{m}(D))$ if $\mathbf{m}$ is perfectly calibrated on dataset D.

Expected Calibration Error (ECE) Naeini et al. (2015). ECE computes the difference between model accuracy and confidence which, in the general form, can be written as $\mathrm{E}_{\hat{P}}[|\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) - p|]$ where the expectation is taken over all class probabilities (confidence) $\hat{P}$. ECE in this form is impossible to compute and it is approximated by partitioning data points into bins with similar confidence resulting in a new formulation of $\text{ECE} = \sum_{n=1}^{N} \frac{|b_n|}{N} |\text{acc}(b_n) - \text{conf}(b_n)|$, where N is the total number of samples and b_n represents a single bin.

Maximum Calibration Error (MCE) Naeini et al. (2015) is particularly important in high-risk applications where reliable confidence measures are absolutely necessary. It measures the worst-case deviation between accuracy and confidence, $\text{MCE} = max_{n \in \{1,\ldots,N\}} |\text{acc}(b_n) - \text{conf}(b_n)|$. For a perfectly calibrated model, the ideal ECE and MCE equal to 0.

Besides ECE and MCE, we also report *Classwise ECE* Nixon et al. (2019), *Adaptive ECE* Nixon et al. (2019) and *Reliability Diagrams* DeGroot & Fienberg (1983); Niculescu-Mizil & Caruana (2005a), we describe these metrics in Appendix A.1.

3.2 Focal Loss

For a classification task, the focal loss (FL) Lin et al. (2017) can be defined as $\mathcal{L}_{\gamma}^{f} = -(1 - p_{i,y_i})^{\gamma} \log p_{i,y_i}$ where γ is a hyper-parameter. It was originally proposed to handle imbalanced data distribution but Jishnu et al. Mukhoti et al. (2020) found the models trained with focal loss are better calibrated w.r.t. cross-entropy trained counterparts. This is because focal loss can be interpreted as a trade-off between minimizing Kullback–Leibler (KL) divergence and maximizing the entropy of predictions, depending on γ Mukhoti et al. (2020)[1]:

$$\mathcal{L}_f \geq \mathrm{KL}(q \parallel p) + \mathbb{H}(q) - \gamma \mathbb{H}(p) \tag{1}$$

where p is the prediction distribution, q is the one-hot encoded target class, and $\mathbb{H}(q)$ is constant. Models trained with this loss learn to strike a balance between narrow p (high confidence of predictions) due to the KL term but and broad p (avoiding overconfidence) due to the entropy regularization term. Original authors provided a principled approach to select the γ for focal loss based on the Lambert-W function Corless et al. (1996). Motivated by this, our work proposes to learn a more fine-grained sample-wise γ with a meta network.

3.3 Meta-Learning

Model calibration can be formulated as minimizing a multi-component loss objective where the base term is used to optimise predictive performance while a regularizer maintains model calibration Bohdal et al. (2021); Lin et al. (2017). Tuning the hyper-parameters of the regularizer can be an a difficult process when conventional methods are used Li et al. (2016); Snoek et al. (2012); Falkner et al. (2018). In this work, we adopt the meta learning approach Luketina et al. (2015). At each training iteration, the model takes a mini-batch from the training dataset D_{train} and the validation dataset D_{val} to optimize

$$\arg\min_{\theta,\phi} \; \mathcal{L}(\theta, \phi, D_{train}, D_{val}) = \mathcal{L}_{FL_\gamma}(\theta, D_{train}) + \mathcal{L}_{SECE}(\phi, D_{val}) \tag{2}$$

First, the base loss function $\mathcal{L}_{FL_\gamma}$ is used to optimize the parameter of the backbone model θ on D_{train}, note that the γ parameter for this step is predicted by γ-Net . Following that, the validation mini batch D_{val} is used to optimize the parameters of the meta-network (γ-Net) ϕ using a validation loss $\mathcal{L}_{SECE}$. The validation loss is a function of the backbone model outputs and does not depend on γ-Net directly. The dependence between validation loss and parameters of the meta-network is mediated via the parameters of the backbone model as discussed by Luketina et al. (2015).

4 Methods

This section introduces the two components of our approach: the γ-Net learns to parameterise the focal loss and the SECE provides differentiability to γ-Net towards calibration optimization.

4.1 γ-Net: Learning Sample-Wise Gamma for Focal Loss

Sample-dependent γ value for focal loss showed it effectiveness of calibrating deep neural networks Mukhoti et al. (2020). In this work, instead of scheduling γ value based on Lambert-W function Corless et al. (1996). We propose to learn a more fine-grained and local γ value, i.e., each sample has an individual γ, at a continuous space. Formally, γ-Net takes the representations from the second last layer of a backbone network (i.e. ResNet He et al. (2016). Let $\mathbf{x} \in \mathbb{R}^{b \times d}$ (b: batch size, d: hidden dimension) be the extracted representation, $\mathbf{A} \in \mathbb{R}^{d \times k}$ be a k-head self-attention matrix that followed by a liner layer with parameters $\mathbf{W} \in \mathbb{R}^{d \times 1}$. The γ-Net transforms the representation to sample-wise γ:

$$\mathbf{a} = \mathbf{x} \cdot \mathbf{A}, \;\; \in \mathbb{R}^{b \times k}, \;\; \mathbf{p} = \textsc{Softmax}(\mathbf{a}), \;\; \in \mathbb{R}^{b \times k} \tag{3}$$

$$\tilde{\mathbf{x}} = \mathbf{p} \cdot \mathbf{A}^{\top}, \;\; \in \mathbb{R}^{b \times d}, \;\; \gamma = \left| \tilde{\mathbf{x}} \cdot \mathbf{W} \right| / \tau, \;\; \in \mathbb{R}^{b \times 1} \tag{4}$$

[1]More theoretical findings can be found in the paper.

Here we use $|\cdot|$ to ensure the γ is positive valued and tune the temperature $\tau = 0.01$ as a hyperparameter. This is similar to the temperature setups as in meta-calibration Bohdal et al. (2021). Those operations result in a set of sample-wise γ to be used in focal loss $\mathcal{L}_r^f$. Note that, $f_\gamma(\gamma_i|x_i), x_i \in D$ is learned as a continous variable rather than discrete value as in original FL formulation Lin et al. (2017) and scheduled FL (FLSD) Mukhoti et al. (2020).

Practical Considerations To ensure $\gamma \geq 0$, we could apply operations based on min-max scaling or activation functions such as sigmoid, relu and softplus. However, our experimental evidences shows that the formulation presented in Equation 4 performs better across datasets.

4.2 SECE : Smooth Expected Calibration Error

Conventional calibration measurement via ECE discussed in Section 3.1, is computed over discrete bins by finding the accuracy and the confidence of examples in each interval. However, this approach is highly dependant on different settings of bin-edges and the number of bins, making it a biased estimator of the true value Minderer et al. (2021).

The issue of bin-based ECE can be traced back to the discrete decision of binning samples Minderer et al. (2021). The larger bin numbers, the lesser information retained. Small bins lead to inaccurate measurements of accuracy. For instance, in single example bins, the confidence is well defined but the accuracy in that interval of model confidences is more difficult to assess accurately from a single point. Using the binary accuracy of a single example leads us to the brier score Brier et al. (1950) which has been shown to not be a perfect measure of calibration.

To find a good representation of accuracy inside the single-example bin (representing a small confidence interval), we leverage the accuracy of other points in the vicinity of the single chosen example weighted by their distance in the confidence space. Formally, the soft estimation of a single-example bin accuracy:

$$\text{SACC } (b_i) = \sum_{j}^{M} \pi(x_i) K(z_i, z_j) \tag{5}$$

$$K\left(x_i, x_j'\right) = \exp\left(-\frac{\left\|x_i - x_j'\right\|^2}{2h^2}\right) \tag{6}$$

where b_i is the bin housing example x_i and $K(\cdot,\cdot)$ is a chosen distance measure, for example a Gaussian kernel and h is bandwidth, z_i and $\pi(x_i)$ are the confidence and accuracy of i^{th} example. As in meta-calibration Bohdal et al. (2021), we use the all-pairs approach Qin et al. (2010) to ensure differentiability. Having good measures of soft accuracy and confidence for each single-example bin, we can write the updated ECE metric as soft-ECE in the form

$$\text{SECE } = \frac{1}{M} \sum_{i}^{M} |\text{SACC } (i) - \text{conf}(i)|, \tag{7}$$

where i represents a single example and M is the number of examples. The new formulation is (1) differentiable as long as the kernel we use is differentiable and (2) enables smooth control over how accuracy over a confidence bin is computed via tuning the kernel parameters. Regarding (2), choosing a dirac-delta function recovers the original brier score while choosing a broader kernel enables a smooth approximation of accuracy over all confidence values.

Connection to KDE-based ECE Estimator Zhang et al. (2020) present a KDE-based ECE Estimator relying on kernel density estimation to approximate the desired metric. Canonically, ECE is computed as an integral over the confidence space as

$$\hat{\text{ECE}}^d = \int \|z - \hat{\pi}(z)\|_d^d \hat{p}(z) dz \tag{8}$$

where $z = \{z_1, z_2, ..., z_L\}$ denotes model confidence distribution over L classes, $\|\cdot\|_d^d$ denotes the d^{th} power of the ℓ_d norm, and $\hat{p}(z)$ represents the marginal density function of model's confidence on a given dataset.

The $\hat{p}(z)$ and $\hat{\pi}(z)$ are approximated using kernel density estimation. We argue that SECE is a special instance with $d = 1$ of $\hat{\text{ECE}}^d$ and estimate ECE with max probability $z_t, t = \arg\max\{z_1, z_2, ..., z_L\}$ for a single instance. And SECE is an upper bound of $\hat{\text{ECE}}^1$:

$$\hat{\text{ECE}} = \int |z - \hat{\pi}(z)|\, \hat{p}(z) dz \tag{9}$$

$$= \int |z_l - \hat{\pi}(z_l)|\, \hat{p}(z_l) dz_l \int |z_t - \hat{\pi}(z_t)|\, \hat{p}(z_t) dz_t \tag{10}$$

$$\leq \int |z_t - \hat{\pi}(z_t)|\, \hat{p}(z_t) dz_t = \text{SECE} \tag{11}$$

with $l = [1, L], l \neq t$ and density functions:

$$\hat{p}(z_t) = \frac{h^{-L}}{M} \sum_{i=1}^{M} K\left(z_t, z_i\right),\ \hat{\pi}_t(z_t) = \frac{\sum_{i=1}^{M} \pi(i) K\left(z_t, z_i\right)}{\sum_{i=1}^{M} K\left(z_t, z_i\right)} \tag{12}$$

where h is the kernel width and $\pi(i)$ represents the binary accuracy of point i. The $p(z_t)$ is a mixture of dirac deltas centered on confidence predicted for individual points in the dataset (used to approximate ECE). Replacing binary accuracy with accuracy computed using all-pairs approach Qin et al. (2010) results in SECE, which is differentiable and can be computed efficiently for smaller batches of data since the integral is replaced with a sum over all examples in a batch.

Discussion: Presented changes do not invalidate the analysis of the $\hat{\text{ECE}}$ as an unbiased estimator of ECE (Theorem 4.1 in Zhang et al. (2020)) as the all-pairs accuracy is an unbiased estimator of accuracy and the details of the $p(z)$ distribution are not used in the derivation beyond setting the bounds. As a result, the analysis described in Zhang et al. (2020) to show their KDE-based approximation to ECE is unbiased can be re-used to show the same property for SECE.

4.3 Optimising γ-Net with SECE

The newly introduced γ-Net and the SECE metric can be applied together to optimize selected ML models for both cross-entropy and calibration error using the meta learning scheme introduced in Section 3.3. As the probability calibration via maximising entropy is performed at a continuous space, we argue that SECE is an efficient learning objective to optimise γ-Net and provide calibration regularization to base network trained with focal loss.

Under the umbrella of meta-learning, two sets of parameters are learned simultaneously: θ, the parameters of base network f^c, as well as the ϕ, the parameters of γ-Net denoted as f^γ. The former, f^c, is used to classify each new example into appropriate classes and the latter, f^γ, is applied to find the optimal value of the focal loss parameter γ for each training example. Algorithm 1 in Supplementary Material describes the learning procedures.

5 Experiments

We conduct a series of experiments to accomplish following objectives:(1) examine both predictive and calibration performance of proposed method;(2) observe the calibration behaviors of proposed method during training;(3) empirically evaluate learned γ values for Focal loss and the robustness of different methods in terms of number of bins.

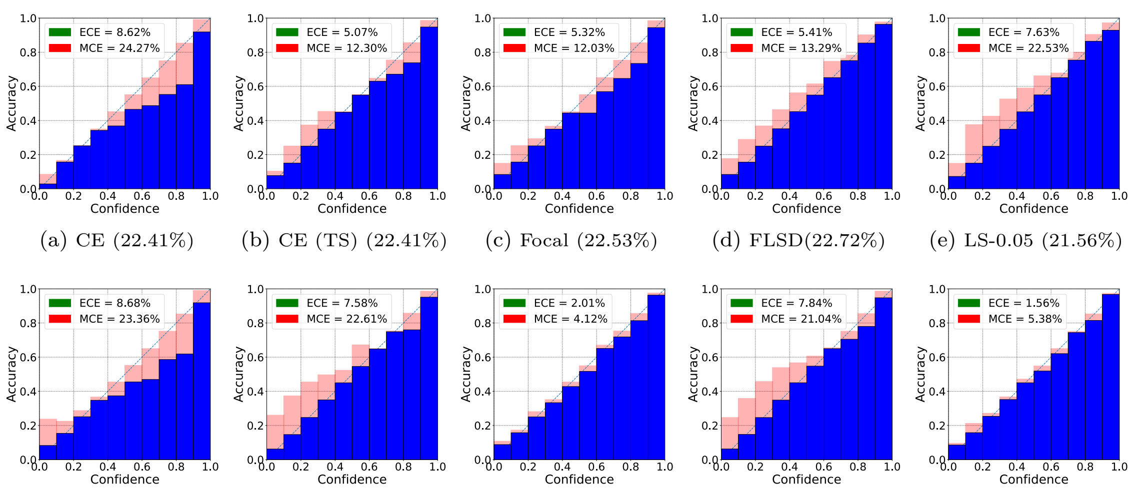
Implementation Details. We implemented our methods by adapting and extending the code from Bohdal et al. (2021) with Pytorch Paszke et al. (2019). For all experiments we use their default settings (using ResNet18 as base model, batch size 128, data augmented with random crop and horizontal flip) unless stated otherwise. We run each experiments for 5 repetitions with different random seeds and average results. We conducted our experiments on CIFAR10 and CIFAR100 (in Bohdal et al. (2021)) as well as Tiny-ImageNet Ya Le (2015); For meta-learning, we split the training set into 8:1:1 as training/val/meta-val with the original test sets are untouched. The experimental pipeline for all three datasets is identical. The models are trained with SGD (learning rate 0.1, momentum 0.9, weight decay 0.0005) for up to 350 epochs. The learning rate is decreased at 150 and 250 epochs by a factor of 10. The model selection is based on the best validation error.

The γ-Net is implemented with a multi-head attention layer with k heads and a fully-connected layer. The k is set to the number of categories. The hidden dim is set to 512, the temperature τ is fixed at 0.01. For SECE, we used the Gaussian kernel with bandwidth of 0.01 (selected with grid search) for both datasets. We initialize $\gamma = 1.0$. At inference stage, the meta network will not present except our ablation study on learned γ values in Section 5.2.

Baselines. We compare our method extensively with baselines including standard cross-entropy (CE), cross-entropy with post-hoc temperature scaling (TS) Platt et al. (1999), Focal loss with standard gamma value (Focal), $\gamma = 1$ Lin et al. (2017), Focal Loss with scheduled gamma (FLSD) Mukhoti et al. (2020), MMCE (Maximum Mean Calibration Error) Kumar et al. (2018) and Label Smoothing with smooth factor 0.05 (LS-0.05) or (LS-0.1) Müller et al. (2019) and Mix-Up ($\alpha = 1.0$) Thulasidasan et al. (2019). In meta-learning setting, we include CE-DECE (meta-calibrationBohdal et al. (2021) for learning unit-wise weight regularization), which is also used as meta-learning baseline, CE-SECE, FL$_\gamma$-DECE, focal loss with learnable sample-wise γ and FL$_\gamma$-DECE.

5.1 Predictive and Calibration Performance

Table 1 presents the performance comparison across approaches. Temperature Scaling (TS) can effectively reduce the errors on calibration metrics as compared to uncalibrated CE models (baseline). Label smoothing and Mixup achieve the best test errors on datasets but exhibit higher ECE and MCE scores. Focal and FLSD can improve calibration in general but we also observed high MCE score for FLSD. On the other hand, we also found that MMCE exhibits higher calibration errors as compared to baseline, this basically aligns with the findings from Bohdal et al. (2021).



(a) CE (22.41%) (b) CE (TS) (22.41%) (c) Focal (22.53%) (d) FLSD(22.72%) (e) LS-0.05 (21.56%)

(f) MMCE (22.30%) (g) CE-DECE(23.05%) (h) FL$_\gamma$-DECE(24.04%) (i) CE-SECE(23.58%) (j) FL$_\gamma$-SECE(23.78%)

Figure 2: The reliability diagram plots for models on CIFAR100 test set. The $(\cdot)$ presents test error. The diagonal dash line presents perfect calibration. The red bar presents the gap between the observed accuracy and the desired accuracy of the perfectly calibrated model (diagonal) - it is positive if the observed accuracy is lower and negative otherwise. The model from the 5^{th} run is used (same to Figure 4).

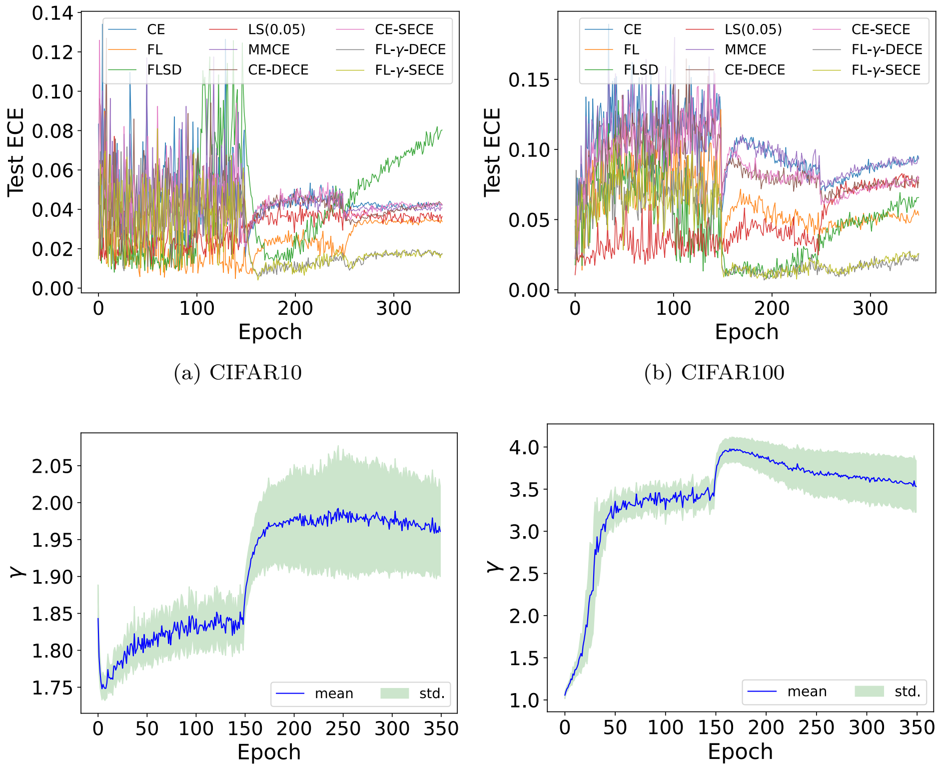
Figure 2 illustrates the reliability diagram plots of models on CIFAR100. The diagrams show that for CIFAR100, γ-Net based methods achieve better ECE compared to SECE and meta-loss for optimizing γ-Net ensures smooth and stable calibrations, and potentially reduces the calibration biases on bins. Similarly on CIFAR100, FL$_\gamma$-SECE shows improved ECE and MCE from CE-DECE.

Table 1: The predictive (test error) and calibration performance of different methods on CIFAR10 (Top), CIFAR100 (Middle) and Tiny-ImageNet (Bottom). The best scores are **bold**. The mean and standard deviation numbers are reported by averaging 5 runs with random seeds. As an alternative calibration method, our approach in general exhibits better calibration while retains competitive predictive performance as compared to conventional as well as meta-learning baselines.

Methods	Error	NLL	ECE	MCE	ACE	Classwise ECE
			CIFAR 10			
CE	4.812 ± 0.122	0.335 ± 0.01	4.056 ± 0.092	33.932 ± 5.433	4.022 ± 0.136	0.848 ± 0.023
CE (TS)	4.812 ± 0.122	0.211 ± 0.005	3.083 ± 0.140	26.695 ± 2.959	3.046 ± 0.157	0.656 ± 0.022
Focal	4.874 ± 0.100	0.207 ± 0.005	3.193 ± 0.104	28.034 ± 5.702	3.174 ± 0.098	0.690 ± 0.018
FLSD	4.916 ± 0.074	0.211 ± 0.005	6.904 ± 0.462	**19.246 ± 11.071**	6.805 ± 0.446	1.465 ± 0.088
LS (0.05)	4.744 ± 0.126	0.232 ± 0.003	2.900 ± 0.085	24.860 ± 8.599	3.985 ± 0.154	0.727 ± 0.009
LS(0.1)	4.918 ± 0.085	0.266 ± 0.004	7.566 ± 0.41	16.033 ± 3.783	7.611 ± 0.161	1.637 ± 0.056
Mixup(α=1.0)	**4.126 ± 0.068**	0.273 ± 0.033	12.863 ± 3.2	20.739 ± 4.205	12.833 ± 3.161	2.678 ± 0.615
MMCE	4.808 ± 0.082	0.333 ± 0.012	4.027 ± 0.082	41.647 ± 10.275	4.013 ± 0.091	0.845 ± 0.014
CE-DECE	5.194 ± 0.161	0.301 ± 0.038	4.106 ± 0.402	41.346 ± 13.325	4.088 ± 0.395	0.868 ± 0.074
CE-SECE	5.222 ± 0.168	0.289 ± 0.027	4.062 ± 0.241	50.81 ± 21.705	4.049 ± 0.251	0.852 ± 0.040
FL$_\gamma$-DECE	5.434 ± 0.095	**0.193 ± 0.009**	2.257 ± 0.787	56.633 ± 23.856	2.396 ± 0.669	**0.557 ± 0.165**
FL$_\gamma$-SECE	5.428 ± 0.144	**0.193 ± 0.010**	**2.138 ± 0.819**	22.725 ± 5.756	**2.357 ± 0.541**	**0.556 ± 0.165**
			CIFAR 100			
CE	22.570 ± 0.438	0.997 ± 0.014	8.380 ± 0.336	23.250 ± 2.436	8.347 ± 0.344	0.233 ± 0.006
CE (TS)	22.570 ± 0.438	0.959 ± 0.008	5.388 ± 0.393	13.454 ± 2.377	5.360 ± 0.315	0.208 ± 0.003
Focal	22.498 ± 0.214	0.900 ± 0.007	5.044 ± 0.203	12.454 ± 0.893	5.015 ± 0.207	0.203 ± 0.004
FLSD	22.656 ± 0.113	0.876 ± 0.005	5.956 ± 0.804	14.716 ± 1.387	5.958 ± 0.802	0.241 ± 0.008
LS (0.05)	21.810 ± 0.172	1.070 ± 0.011	8.108 ± 0.346	20.268 ± 1.536	8.106 ± 0.346	0.272 ± 0.006
LS(0.1)	22.244 ± 0.155	1.052 ± 0.011	4.754 ± 0.709	17.228 ± 0.923	4.777 ± 0.647	0.239 ± 0.004
Mixup(α=1.0)	**21.210 ± 0.227**	0.917 ± 0.017	9.716 ± 0.754	16.01 ± 1.335	9.722 ± 0.74	0.315 ± 0.011
MMCE	22.490 ± 0.143	1.021 ± 0.007	8.713 ± 0.245	23.565 ± 1.141	8.670 ± 0.305	0.238 ± 0.004
CE-DECE	23.406 ± 0.323	1.148 ± 0.006	7.309 ± 0.245	22.565 ± 1.446	7.253 ± 0.315	0.241 ± 0.002
CE-SECE	23.448 ± 0.302	1.153 ± 0.015	7.668 ± 0.330	24.261 ± 1.614	7.609 ± 0.295	0.244 ± 0.002
FL$_\gamma$-DECE	23.712 ± 0.204	0.888 ± 0.009	**1.879 ± 0.440**	8.271 ± 2.651	**1.838 ± 0.371**	0.195 ± 0.005
FL$_\gamma$-SECE	23.686 ± 0.377	**0.877 ± 0.004**	1.940 ± 0.365	**7.480 ± 1.867**	**1.939 ± 0.379**	**0.192 ± 0.006**
			Tiny-ImageNet			
CE	40.110 ± 0.110	1.838 ± 0.171	8.059 ± 1.296	15.73 ± 1.905	8.006 ± 1.282	0.154 ± 0.001
Focal	39.415 ± 0.625	1.896 ± 0.009	7.600 ± 0.309	13.771 ± 0.897	7.469 ± 0.301	0.152 ± 0.002
FLSD	39.705 ± 0.075	1.904 ± 0.025	14.501 ± 1.078	21.528 ± 2.116	14.501 ± 1.078	0.202 ± 0.006
LS (0.1)	**39.395** ± 0.305	2.185 ± 0.001	16.777 ± 0.476	29.088 ± 1.835	16.901 ± 0.460	0.199 ± 0.001
Mixup(α=1.0)	39.890 ± 0.271	1.932 ± 0.054	12.133 ± 2.069	31.440 ± 0.968	12.028 ± 2.079	0.193 ± 0.009
MMCE	40.310 ± 0.100	1.826 ± 0.177	8.206 ± 1.219	16.802 ± 2.339	8.165 ± 1.269	**0.149** ± 0.001
CE-DECE	41.350 ± 0.000	2.228 ± 0.033	10.694 ± 0.503	20.888 ± 0.430	10.553 ± 0.553	0.160 ± 0.000
CE-SECE	41.005 ± 0.145	2.213 ± 0.058	10.928 ± 1.125	21.362 ± 2.526	10.912 ± 1.069	0.157 ± 0.003
FL$_\gamma$-DECE	40.625 ± 0.095	**1.826** ± 0.007	5.944 ± 1.090	11.542 ± 1.990	6.077 ± 1.095	0.155 ± 0.007
FL$_\gamma$-SECE	40.850 ± 0.140	1.829 ± 0.005	**5.794** ± 0.756	**11.477** ± 1.563	**5.848** ± 0.751	0.156 ± 0.005

Among compared methods, our method FL$_\gamma$-SECE achieves considerably lower errors on calibration metrics. When comparing meta-learning based approaches: CE-DECE (meta-learning baseline), CE-SECE, FL$_\gamma$-DECE, FL$_\gamma$-SECE. We can see that our method (FL$_\gamma$-SECE) achieves comparable test error and better scores across calibration metrics. Particularly, it improves ECE by average 4.198% on three detests, in addition, by average 14.37% MCE and 3.917% ACE score respectively.

To observe the learning behavior of models, we plotted the changing curves of test error (included in Supplementary Material Section 3) and test ECE in Figure 3 (a-b). It is noted that though test error curves show similar behaviors, but the calibration behaviors are quite different across methods. The primary value of γ-Net with SECE is in the stable and smooth calibration behavior. After the 150^{th} epoch when the learning rate is decreased by a factor of 10, we can observe that the ECE score starts to increase for all approaches, particularly FLSD (green lines).



(a) CIFAR10 (b) CIFAR100

(c) CIFAR10 (d) CIFAR100

Figure 3: (a-b): ECE curves on the test dataset of CIFAR10 (a) and CIFAR100 (b).(c-d): The mean and standard deviation (std.) of γ on test dataset at each epoch. Low std. score indicates samples share similar gamma values, and high std. score indicates more samples have different γ values.

5.2 Ablation Study

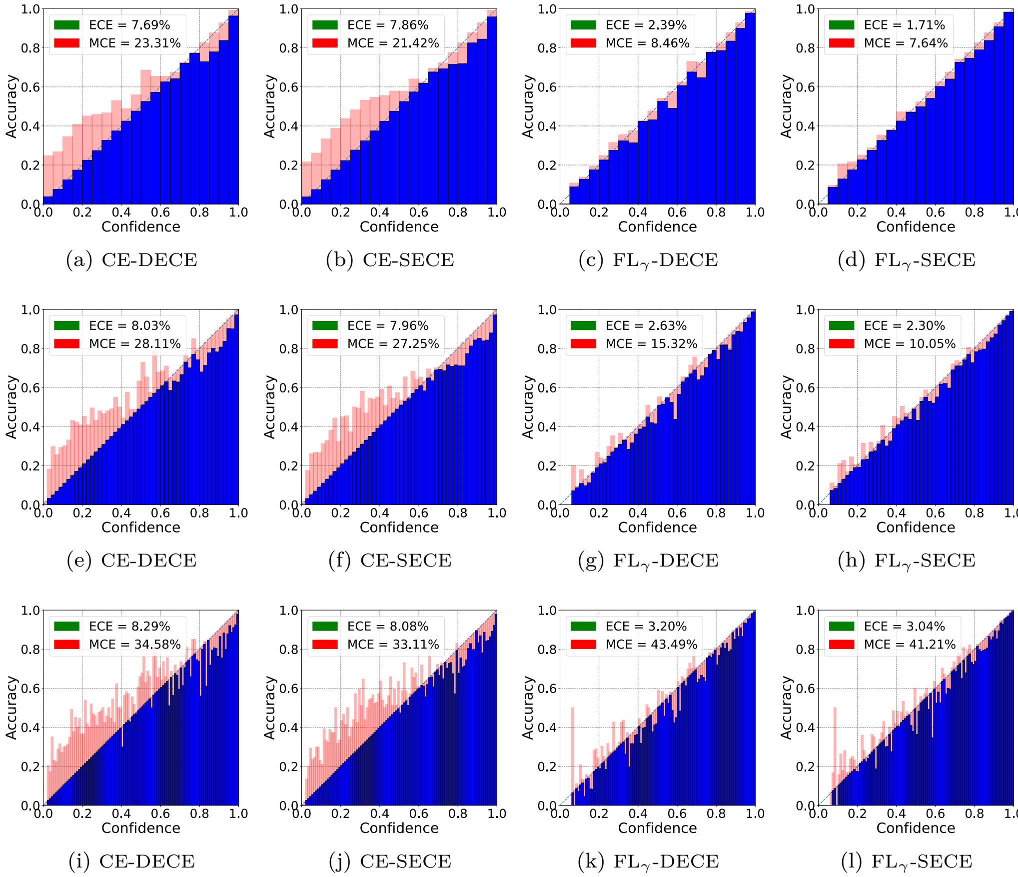
5.2.1 Learning γ as continous variables

Figure 3(c-d) presents the change of γ values on test dataset over epochs. At the earlier training stage, the samples γ_j for $x_j \in D_{test}$ have similar values (with low variance) because they are initialized with $\gamma_i = 1.0$ for $x_i \in D_{val}$. The γ parameter is observed to have higher standard deviation at the later training stage as γ-Net learns optimal value for each example in the dataset rather than relying on global values showcasing the flexibility of the network. It is also noted that γ is learned in a continuous space, different to the discrete values the pre-defined in Focal Loss Lin et al. (2017) and FLSD Mukhoti et al. (2020).

5.2.2 Calibration bias and robustness

In Figure 5, we examine the robustness of those methods with different binning schemes by varying the number of bins, which is one of causes of introducing calibration bias Minderer et al. (2021). It shows that γ-Net based approaches (FL$_\gamma$-DECE and FL$_\gamma$-SECE) maintain much lower ECE score when throughout all bin numbers from 10 to 1000 showing the trained based network is robustly calibrated. Furthermore FL$_\gamma$-SECE is also able to maintain lower MCE as compared to other methods.

Here we can clearly see the the role of SECE in stabilizing the calibration with γ-Net against the change of bin numbers. Table 2 shows the individual calibration gain from SECE as compared to using learnable γ only. Our method FL$_\gamma$-SECE. With SECE, we can see the improved both predictive and calibration performance. Importantly, from Figure 4 and 5, we observe that it reduces calibration bias that is introduced by different



(a) CE-DECE (b) CE-SECE (c) FL$_\gamma$-DECE (d) FL$_\gamma$-SECE

(e) CE-DECE (f) CE-SECE (g) FL$_\gamma$-DECE (h) FL$_\gamma$-SECE

(i) CE-DECE (j) CE-SECE (k) FL$_\gamma$-DECE (l) FL$_\gamma$-SECE

Figure 4: The reliability diagram plots on CIFAR100 with large bin numbers (top to bottom: 20, 50, 100).

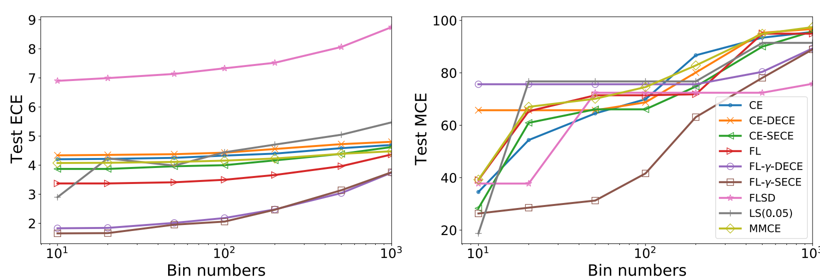


Figure 5: The ECE and MCE scores on CIFAR 10 test dataset with different bin numbers in $[10, 20, 50, 100, 200, 500, 1000]$. Our proposed approach shows better robustness on bin sizes. Similar plot for CIFAR100 is Figure 3, Supplementary Material.

binning mechanisms. This is mostly reflected on reduced ECE/MCE scores, i.e., the model exhibits high MCE on a particular bin if calibration bias exists.

6 Discussion

Novelties. Though we follow the general setup in Bohdal et al. (2021), our work makes several novel contributions: (1) Learnable sample-wise at as continuous variables for focal loss. To the best of our knowledge, this is the first work to introduce fine-grained γ for focal loss for the purpose of model calibration

Methods	Error	NLL	ECE	MCE	ACE	Classwise ECE
FL_γ	5.632 ± 0.118	0.197 ± 0.009	2.177 ± 0.619	46.172 ± 28.24	**2.319 ± 0.407**	**0.553 ± 0.12**
FL_γ-SECE	**5.428 ± 0.144**	**0.193 ± 0.010**	**2.138 ± 0.819**	**22.725 ± 5.756**	2.357 ± 0.541	0.556 ± 0.165
FL_γ	28.148 ± 8.127	1.051 ± 0.278	3.044 ± 1.542	10.082 ± 3.441	3.016 ± 1.511	0.226 ± 0.063
FL_γ-SECE	**23.686 ± 0.377**	**0.877 ± 0.004**	**1.940 ± 0.365**	**7.480 ± 1.867**	**1.939 ± 0.379**	**0.192 ± 0.006**

Table 2: SECE calibration gain on CIFAR10 (top) and CIFAR100 (bottom).

Methods	ECE(%)	Test Error(%)
CIFAR10		
Patra et al. (2023)	0.59	6.28
Hebbalaguppe et al. (2022)	0.93	7.18
Hebbalaguppe et al. (2022)	0.70	7.08
Liu et al. (2022) (without marign)	3.72	5.24
Liu et al. (2022)	1.16	4.75
Karandikar et al. (2021) (Focal+SB-ECE)	1.19	4.90
Karandikar et al. (2021) (Focal+SB-AvUC)	1.58	5.60
Meta-Regularization (Ours)	2.14	5.43
CIFAR100		
Patra et al. (2023)	1.74	26.57
Hebbalaguppe et al. (2022)	1.49	31.58
Hebbalaguppe et al. (2022)	0.72	29.80
Karandikar et al. (2021)(Focal+SB-ECE)	2.30	21.40
Karandikar et al. (2021) (Focal+SB-AvUC)	1.57	21.90
Meta-Regularization (Ours)	1.94	23.69

Table 3: Comparing meta-regularization (our) to recent (non-meta) regularization methods on CIFAR10 and CIFAR100. The meta-regularization gives very competitive predictive and calibration performance to conventional regularization methods (the numbers are obatined from the corresponding papers).

with learned from a meta-network. (2) The proposed SECE, which reduces model calibration bias against different binning mechanism. This is rarely discussed in the literature. Additionally, SECE is more also efficient than DECE in Bohdal et al. (2021) as it uses a simple summation, rather than predicting the bin assignment with networks. (3) We further showcase the importance and fesiablity of using gradient-based meta-learning Finn et al. (2017) in allevating model miscalibration. As shown in Table 3, the meta-regularization is able to achieve very competitive predictive and calibration performance to conventional regularization methods.

Limitation. Our proposed method uses γ-Net for learning γ, this increases parameter numbers by roughly 0.46% in training. However, as γ-Net is not present after training, there is no additional computational complexity in inference.

7 Conclusion

In this work, we presented a meta-learning based approach for learning well calibrated models and shown the benefits of two newly introduced components. Learning a sample-wise γ for Focal loss using γ-Net yields both strong predictive performance and unbiased and robust calibration. Optimising γ-Net with SECE plays an important role by ensuring stable calibration as compared to baselines. With extensive empirical results on three computer vision datasets, we showed that our method provides better calibration capability without changing the original networks.

References

Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Meta-calibration: Meta-learning of model calibration using differentiable expected calibration error. *arXiv preprint arXiv:2106.09613*, 2021.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Glenn W Brier et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Jiacheng Cheng and Nuno Vasconcelos. Calibrating deep neural networks by pairwise constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13709–13718, 2022.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the lambertw function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

Morris DeGroot and Stephen Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 1983.

Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. *35th International Conference on Machine Learning, ICML 2018*, 4:2323–2341, 7 2018. doi: 10.48550/arxiv.1807.01774. URL https://arxiv.org/abs/1807.01774v1.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. Ieee, 2013.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

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*, pp. 770–778, 2016.

Ramya Hebbalaguppe, Jatin Prakash, Neelabh Madan, and Chetan Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16081–16090, 2022.

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33:18237–18248, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18:1–52, 3 2016. ISSN 15337928. doi: 10.48550/arxiv.1603.06560. URL https://arxiv.org/abs/1603.06560v4.

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*, pp. 2980–2988, 2017.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 80–88, 2022.

Jelena Luketina, Mathias Berglund, Klaus Greff, and Tapani Raiko. Scalable gradient-based tuning of continuous regularization hyperparameters. *33rd International Conference on Machine Learning, ICML 2016*, 6:4333–4341, 11 2015. doi: 10.48550/arxiv.1511.06727. URL https://arxiv.org/abs/1511.06727v3.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. Attended temperature scaling: A practical approach for calibrating deep neural networks, 2018. URL https://arxiv.org/abs/1810.11586.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting Good Probabilities with Supervised Learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 2005a.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005b.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning, 2019. URL https://arxiv.org/abs/1904.01685.

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. *Advances in neural information processing systems*, 32, 2019.

Rishabh Patra, Ramya Hebbalaguppe, Tirtharaj Dash, Gautam Shroff, and Lovekesh Vig. Calibrating deep neural networks using explicit regularisation and dynamic data pruning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1541–1549, 2023.

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.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Tao Qin, Tie-Yan Liu, and Hang Li. A general approximation framework for direct optimization of information retrieval measures. *Inf. Retr.*, 13(4):375–397, aug 2010. ISSN 1386-4564. doi: 10.1007/s10791-009-9124-x. URL https://doi.org/10.1007/s10791-009-9124-x.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Adv. Neural Inf. Process. Syst. 25*, pp. 1–9, 2012. ISSN 10495258. doi: 2012arXiv1206.2944S. URL https://arxiv.org/pdf/1206.2944.pdf.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xuan Yang Ya Le. Tiny imagenet visual recognition challenge, 2015. URL http://vision.stanford.edu/teaching/cs231n/reports/2015/pdfs/yle_project.pdf.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 609–616. Citeseer, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699. ACM, 2002.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning, 2020. URL https://arxiv.org/abs/2003.07329.