MaxEnt Loss: Constrained Maximum Entropy for Calibration under Out-of-Distribution Shift

Dexter Neo, Stefan Winkler, Tsuhan Chen

School of Computing, National University of Singapore
e0534450@u.nus.edu, {winkler, tsuhan}@nus.edu.sg

Abstract

We present a new loss function that addresses the out-of-distribution (OOD) network calibration problem. While many objective functions have been proposed to effectively calibrate models in-distribution (ID), our findings show that they do not always fare well OOD. Based on the Principle of Maximum Entropy, we incorporate helpful statistical constraints observed during training, delivering better model calibration without sacrificing accuracy. We provide theoretical analysis and show empirically that our method works well in practice, achieving state-of-the-art calibration on both synthetic and real-world benchmarks. Our code is available at *https://github.com/dexterdley/MaxEnt-Loss*.

Introduction

Recent advances in machine learning have given rise to large neural networks with strong recognition performance in fields such as computer vision and natural language processing. Neural networks are increasingly used in areas where safety is a concern, such as self-driving cars (Bojarski et al. 2016), medical prognosis (Esteva et al. 2017; Bandi et al. 2019) and facial expression recognition (Vonikakis, Neo, and Winkler 2021; Neo, Chen, and Winkler 2023; Neo and Chen 2023b). Despite their popularity, deep neural networks have a tendency to be poorly calibrated.

Calibration refers to the model's correctness with regard to its predicted probabilities. In other words, models tend to overconfidently misclassify samples and erroneously recognize correct classes with low confidence. This often creates mistrust due to the mismatch between model correctness and confidence. For severe cases, uncalibrated models can cause harm, resulting in serious consequences such as overconfidently misidentifying cancerous cells (see Figure 1). This is especially common when uncalibrated models encounter samples that are out-of-distribution (OOD) from the training set. Ideally, a well-calibrated classifier should behave unconfidently and predict low probabilities whenever it misclassifies samples. Furthermore, we would like our models to not only remain accurate and well-calibrated in-distribution (ID) but to also provide further robustness against OOD shifts for safe deployment (Thulasidasan et al. 2019; Kumar, Liang, and Ma 2019).

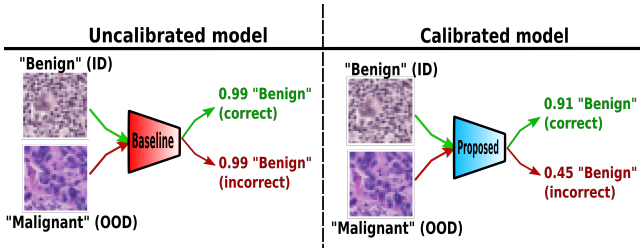

Figure 1: Uncalibrated models (left) make overconfident OOD misdiagnoses, resulting in dire consequences. Well-calibrated models (right) exhibit lower confidence, reflecting their uncertainty for OOD samples.

The current hypothesis as to why modern neural networks are miscalibrated is that these large models with millions of parameters have the capacity to learn and overfit to the given training data (Guo et al. 2017). This is especially true if model training is done using the cross-entropy (CE) loss, since the CE loss can only be fully minimized when probabilities $\hat{P}$ are equal to the one-hot ground truth y. This means that even though the accuracy is at 100%, CE loss can still be positive and minimized further by increasing the confidence of the predicted probabilities, resulting in overfitting and miscalibration. While many techniques have been proposed to address calibration, a category of methods include objective functions that pair CE loss with auxiliary terms. Specifically, objective functions such as Focal, Inverse Focal and Poly loss (Lin et al. 2017; Wang, Feng, and Zhang 2021; Leng et al. 2022) add an additional penalty term to control the confidence of predictions. Other objective functions such as AvUC and Soft-AvUC loss introduce a differentiable utility term based on accuracy and uncertainty (Krishnan and Tickoo 2020; Karandikar et al. 2021).

Contrary to these methods, we follow the work of Mukhoti et al. (2020) and further explore the Principle of Maximum Entropy (MaxEnt) (Jaynes 1957). We propose a novel regularization technique in the form of a general loss function for improving calibration based on constrained maximum entropy. Our method works by introducing additional constraints that complement loss functions typically used in supervised learning. We provide systematic comparisons of model accuracy and calibration for image classifi-

cation tasks. Our main contributions can be summarized as follows:

1. **MaxEnt Loss:** We explore the theoretical relations between the Principle of Maximum Entropy and Focal loss, based on which we propose a novel loss function with three different forms, showing how constraints can be introduced to further improve calibration.
2. **Automated Hyperparameter Tuning:** Based on the constraints, our framework provides an automated estimate of the optimal Lagrange multipliers, with no need for manual tuning.
3. **Evaluation on OOD shifts:** Our experiments show that MaxEnt loss remains robust in terms of accuracy and calibration for both synthetic and in-the-wild distribution shift benchmarks. We also analyze the ordering of the model's feature norms under increasingly shifted inputs.
4. **Ad-hoc calibration:** Our method is non-restrictive and works well in combination with popular ad-hoc calibration methods such as temperature scaling and label smoothing.

Related Work

Network Calibration: Existing methods for calibrating neural networks can be categorized roughly as follows: (1) Methods that approximate the true joint distribution or latent hidden vector z using generative models such as VAE (Kingma and Welling 2014), Cycle-GAN (Zhu et al. 2017). (2) Methods that directly regularize the output probabilities such as isotonic regression (Zadrozny and Elkan 2002), Bayesian binning (Naeini, Cooper, and Hauskrecht 2015), splines (Gupta et al. 2021), objective functions (Krishnan and Tickoo 2020; Karandikar et al. 2021; Wang, Feng, and Zhang 2021; Leng et al. 2022; Cheng and Vasconcelos 2022) and temperature scaling methods (Guo et al. 2017; Kull et al. 2019). A recent summary of model calibration can be found in (Minderer et al. 2021).

Calibration under OOD shift: For OOD problems, test inputs do not align with the training set (Ovadia et al. 2019). This phenomenon can be caused by either (1) completely OOD test inputs that belong to an OOD class not from the ground truth labels (Du et al. 2022), or (2) shifted OOD test inputs caused by perturbations and corruptions (Hendrycks and Gimpel 2017). Apart from the loss functions introduced previously, many other strategies have been proposed to tackle OOD calibration. This includes methods such as multi-domain temperature scaling (Yu et al. 2022), sampling Gaussians for domain drifts (Tomani et al. 2021) and transferable calibration (Wang et al. 2020). For this work, we mainly focus on loss functions for on-the-fly OOD calibration for both synthetic and in-the-wild data, rather than pre- or post-hoc techniques. For a recent review of OOD shifts, we refer to (Wiles et al. 2022).

Maximum Entropy: Pereyra et al. (2017) have shown that directly penalizing neural networks with the maximum entropy term helps prevent overconfidence, resulting in better generalization. The Principle of Maximum Entropy (Jaynes 1957) has a long standing in information theory where we maximize the model's entropy subject to constraints derived from the training set (Berger, Della Pietra, and Della Pietra 1996). Focal loss (Lin et al. 2017) was originally proposed for object detection, yet it can also be used for improving calibration. Mathematically, Focal loss is a general form of CE loss with an additional entropy term (Mukhoti et al. 2020) – reducing Focal loss simultaneously *minimizes* the KL divergence and *maximizes* the entropy, discouraging overconfidence. The MaxEnt method is also used in other tasks, such as Fine-Grained Visual Classification (Dubey et al. 2018), where classes may be visually similar, and reinforcement learning (Haarnoja et al. 2018; Neo and Chen 2023a), where high entropy policies tend to encourage stochasticity and improve exploration .

Preliminaries

Consider a classification task over a dataset D with N number of samples $(x_i, y_i)_{i=1}^N$, where X, Y denote input feature and label space, and $\mathcal{Y} = [1, 2, ..., K]$ is a fixed array containing all K class indices. Given an arbitrary input datum x_i, the task is modelled by a neural network with learnable parameters $\boldsymbol{\theta}$ and a penultimate layer containing K neurons, which output logits $g_i^\theta(x)$. The model learns to estimate the posterior distribution which are a set of valid probabilities such that $\sum_{k=1}^K P_i(y_k|x) = 1$, after the softmax function $\frac{\exp g_i^\theta(x)}{\sum_{k=1}^K \exp g_k^\theta(x)}$. The predicted top-1 class is then simply $\hat{y} := \arg\max g_i^\theta(x)$, with the corresponding confidence score $\hat{P} := \max P_i(y_k|x)$. In theory, a model is considered perfectly calibrated iff the model's probabilities match the true posterior distribution, satisfying the definition $\mathbb{P}(\hat{y} = y|\hat{P} = P) = P \quad \forall \in P[0-1]$. Realistically, achieving this level of calibration is infeasible as the true posterior distribution remains unknown. In the following we list a few error metrics have been proposed to approximate calibration; for the definitions of additional calibration metrics and results please refer to the Appendix.

Expected Calibration Error (ECE): Calibration error is commonly estimated using ECE (Naeini, Cooper, and Hauskrecht 2015). It is computed by splitting the model's probabilities into B bins. Let n_b, acc and conf represent the number of samples, average accuracy and confidence for each partitioned bin. The weighted absolute differences between acc and conf for each bin is calculated using the following formula: $\text{ECE} = \sum_{b=1}^B \frac{n_b}{N} |\text{acc}(b) - \text{conf}(b)|$.

Classwise ECE (CECE): ECE only considers the $\max$ confidence of the predicted probabilities. In certain scenarios, one would also require the probabilities of all other classes to be well calibrated, therefore CECE is proposed as an simple extension of ECE, considering all K classes (Nixon et al. 2019): $\text{CECE} = \frac{1}{K} \sum_{b=1}^B \sum_{k=1}^K \frac{n_{b,k}}{N} |\text{acc}(b,k) - \text{conf}(b,k)|$.

Kolmogorov-Smirnov Error (KSE): The approximation of calibration errors often requires the histograms/binning of empirical distributions. This causes an over-reliance on binning, which is sensitive to the number of bins chosen. Inspired by the Kolmogorov-Smirnov test, (Gupta et al. 2021) propose a numerical approximation of the equality between two cumulative distributions without the need for binning. Using the authors' notation, the model's probabilities are ab-

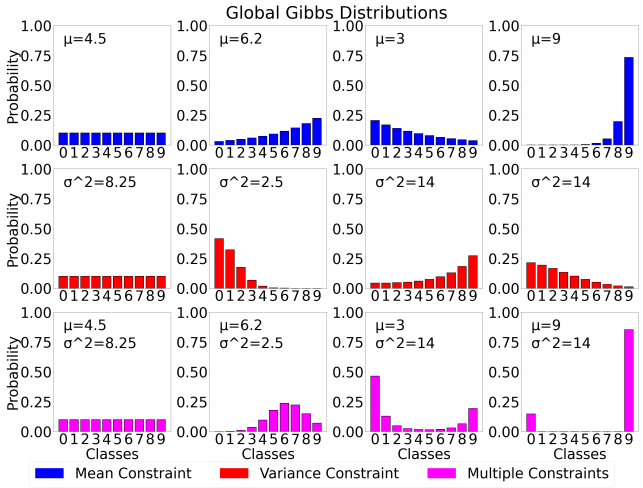

Figure 2: In the face of unknown OOD, we argue that predictions should not deviate too far from the observed global Gibbs distribution. For different values of constraints, higher μ values tend toward larger classes, while a higher σ^2 results in distributions being more spread out.

breviated as z_k with the integral form of the Kolmogorov-Smirnov error for top-1 classification defined as: KSE $= \int_0^1 |P(k|z_k) - z_k|P(z_k)dz_k$.

Methodology

In this section, we show the relationship between Maximum Entropy (Jaynes 1957), confidence penalty term (Pereyra et al. 2017) and Focal loss (Lin et al. 2017). We also show the effects of different constraints placed on the MaxEnt method and describe how they can be developed into a novel objective function for model calibration that we call MaxEnt Loss.

Principle of Maximum Entropy

The Principle of Maximum Entropy[1] is a probability distribution that maximizes the Shannon entropy subject to the given inequality constraints. Specifically, the general discrete form of the MaxEnt method is given by:

$$\max H(P_i(y_k|x)) = -\sum_{k=1}^{K} P_i(y_k|x)\log P_i(y_k|x)$$
$$\text{subject to} \sum_k P_i(y_k|x) f_n(\mathcal{Y}) \leq c_n \text{ for all constraints} f_n(\mathcal{Y}) \tag{1}$$

where $f_n(\mathcal{Y})$ is a function of the random variable vector $\mathcal{Y}$ and the respective Lagrange multipliers λ_n for each n number of constraints c_n. In theory, the MaxEnt method allows for any arbitrary function $f_n(\mathcal{Y})$ to be computed within the expectation, in practice we can simply assign a fixed vector $\mathcal{Y}$ containing the class labels as our function.

[1] We refer readers to (Jaynes 1957) along with examples included in the Appendix for an extended background of the Principle of Maximum Entropy.

Following Equation (1), we build a simple MaxEnt model using a ten-class example and illustrate the effects of different values of c_n in Figure 2. We show the global Gibbs distribution as a function of the given random variable and subjected constraints. The constraints are scalars computed from the expectation of the random variable such as $\mathbb{E}[\mathcal{Y}]$ and $\mathbb{E}[\mathcal{Y}^2]$. For example, if the class label frequency/prior distribution $P(\mathcal{Y})$ of the training set is uniform, then the expected average μ would be $\sum_{k=1}^{K=10} \mathcal{Y}P(\mathcal{Y}) = 4.5$, with the corresponding expected variance σ^2 given as $\sum_{k=1}^{K=10} (\mathcal{Y}-\mu)^2 P(\mathcal{Y}) = 8.25$ (see column 1 of Figure 2). In row 1, increasing $\mu > 4.5$ results in a shift towards larger classes and increasing $\sigma^2 > 8.25$ will give a distribution with a higher variance. For row 3, when both constraints are combined, the distributions shift to jointly accommodate both the mean and variance.

Relation to Focal Loss

Next, we show the relation between the Principle of Maximum Entropy and Focal loss, which reduces the emphasis on easily classified samples (Lin et al. 2017). Consider the multi-class form of Focal loss where $\mathcal{L}_F = -\sum_k (1 - P_i(y_k|x_i))^\gamma \log P_i(y_k|x_i)$, with the hyperparameter $\gamma \geq 0$. By setting $\gamma = 1$, $\mathcal{L}_F$ can be expanded and re-written as the CE loss with a confidence penalty term (Shannon's entropy):

$$\mathcal{L}_F = -\sum_k \underbrace{\log P_i(y_k|x)}_{\text{CE Loss}} - \underbrace{H(P_i(y_k|x))}_{\text{Shannon term}} \tag{2}$$

Supposedly, even if other values of $\gamma > 1$ are chosen, Equation (2) still holds true such that the Shannon term is paired with a polynomial. This additional entropy term is useful for preventing peaked distributions and delivers better generalization (Mukhoti et al. 2020).

Connecting Equation (1) and Equation (2), we argue that maximizing the model's entropy subject to constraints computed from the prior knowledge observed would be a possible approach for OOD scenarios. Since it is impossible to know beforehand the type or intensity of OOD shift, utilizing any additional information during training can be useful OOD. For example, if the class label frequencies of the training set are uniform, we should expect the expectations of the classifier's predictions to be closer to that of a uniform distribution, especially if the inputs are progressively shifted from the training samples.

MaxEnt Loss for End-to-End Training

Given the above, we now demonstrate how constraints can be added to the Focal loss. We propose three forms of MaxEnt Loss for end-to-end model calibration:

Definition 1 (Mean Constraint)**.** Consider the expected average μ of the target distribution to be constrained, we can add the following mean constraint terms to Equation (2) and aim to minimize the following objective function:

Algorithm 1: Constrained MaxEnt Loss Optimization

Data: Given training set $D = (x_i, y_i)_{i=1}^N$

1: Initialize neural network parameters $\boldsymbol{\theta}$ and learning rate schedule α
2: Compute the global and local expectations for the mean and variance constraints μ, σ^2
3: $\hookrightarrow$ $\mathbb{E}[\mathcal{Y}] = \mu$ and $\mathbb{E}[\mathcal{Y}^2] = \sigma^2$
4: Solve numerically for $\lambda_n \leftarrow$ `NewtonRaphson()` // Use a root-finder to obtain λ_n
5: **for** $e \in epochs$ **do**
6: **for** $i \in B$ **do** // Sample mini-batch of size B
7: Calculate MaxEnt Loss: $\mathcal{L}_{ME} = \mathcal{L}_F + \sum_{n=1}^{M} \lambda_n \left(\sum_{k=1}^{K} \mathcal{Y} P_i(y_k|x_i) - c_n \right)$
8: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \Delta \mathcal{L}_{ME}$ // Update parameters $\boldsymbol{\theta}$ by gradient descent
9: **return** $\boldsymbol{\theta}$
10:
11: **Function** `NewtonRaphson()`:
12: δ = 1e-15 // A small tolerance or stopping condition
13: **while** $g(\lambda) > \delta$ **do**
14: $\lambda_{n+1} = \lambda_n - \frac{g(\lambda)}{g'(\lambda)}$ // Update Lagrange Multipliers λ_n
15: **return** λ_n

$$\mathcal{L}_{ME}^{M} = -\sum_{k} \log P_i(y_k|x) - H(P_i(y_k|x)) + \lambda_\mu \Bigg[\underbrace{\sum_{k} \mathcal{Y} P_i(y_k|x) - \mu_G}_{\text{Global mean constraint}} + \underbrace{\sum_{k} \mathcal{Y} P_i(y_k|x) - \mu_{Lk}}_{\text{Local mean constraint}} \Bigg] \tag{3}$$

where μ_G is the global expected average $\mathbb{E}[\mathcal{Y}]$ computed from the prior distribution described earlier, and μ_{Lk} is the local expectation for the kth class, which can be computed from the target labels $\sum_k \mathcal{Y} y_k = \mu_{Lk}$ and λ_μ is the Lagrange multiplier for the mean constraint form.

Definition 2 (Variance Constraint)**.** Next, consider the case where the expected variance σ^2 of the target distribution is to be constrained, the variance constraints can be added to Equation (2):

$$\mathcal{L}_{ME}^{V} = -\sum_{k} \log P_i(y_k|x) - H(P_i(y_k|x)) + \lambda_{\sigma^2} \Bigg[\underbrace{\sum_{k} \mathcal{Y}^2 P_i(y_k|x) - \sigma_G^2}_{\text{Global variance constraint}} + \underbrace{\sum_{k} \mathcal{Y}^2 P_i(y_k|x) - \sigma_{Lk}^2}_{\text{Local variance constraint}} \Bigg] \tag{4}$$

where σ_G^2 is the global expected variance $\mathbb{E}[\mathcal{Y}^2]$ and $\sum_k \mathcal{Y}^2 y_k = \sigma_{Lk}^2$ is the expected local variance for the kth class with λ_{σ^2} as the corresponding Lagrange multiplier. For this form, we assume that there is no knowledge of the expected average/mean constraint.

Definition 3 (Mean and Variance Constraints)**.** Finally, when both the expected average and variance of the target distribution are to be considered, we can combine both constraints. The objective function for this form is given by:

$$\begin{aligned} \mathcal{L}_{ME}^{M+V} &= \mathcal{L}_F + \lambda_\mu \left[\sum_{k} \mathcal{Y} P_i - \mu_G + \sum_{k} \mathcal{Y} P_i - \mu_{Lk} \right] \\ &+ \lambda_{\sigma^2} \left[\sum_{k} (\mathcal{Y} - \mu)^2 P_i - \sigma_G^2 + \sum_{k} (\mathcal{Y} - \mu)^2 P_i - \sigma_{Lk}^2 \right] \end{aligned} \tag{5}$$

For this case, the variances are computed together with the expected average $\sum_k (\mathcal{Y} - \mu)^2 y_k = \sigma_{Lk}^2$, and the Lagrange multipliers λ_μ and λ_{σ^2} need to be solved simultaneously. For all three definitions, the respective Lagrange multipliers λ_n can be solved cheaply using traditional numerical root-finders that require only CPU. In our work, we select Newton Raphson's method as our root-finder which utilizes a helper function $g(\lambda)$ and its derivative $g'(\lambda)$ to solve for λ_n in $\mathcal{O}(n)$ time. This step is described in lines 4 and 11-15 of Algorithm 1, along with more details on the rest of our method.

We highlight that global expected mean and variance are computed from the prior distribution $P(y_k)$; the local mean and variances are computed for each class label. For example, in the case of one-hot labels the local averages are always $\mu_{Lk} = y_k$, however this is no longer true if the ground truth labels are not one-hot (e.g. label smoothing).

After the constraints are computed, each of the Lagrange multipliers λ_n are solved numerically before training is performed. The Appendix contains examples and proofs on this computation for single and multiple constraints. In our ablation study, we also include a discussion regarding the empirical effects of local constraints.

Experiments and Results

We perform experiments on six popular OOD shift image classification benchmarks and evaluate our method against recently proposed calibration objective functions. Specifically, we compare against the following baseline losses: CE, Focal, Inverse Focal (Wang, Feng, and Zhang 2021), AvUC

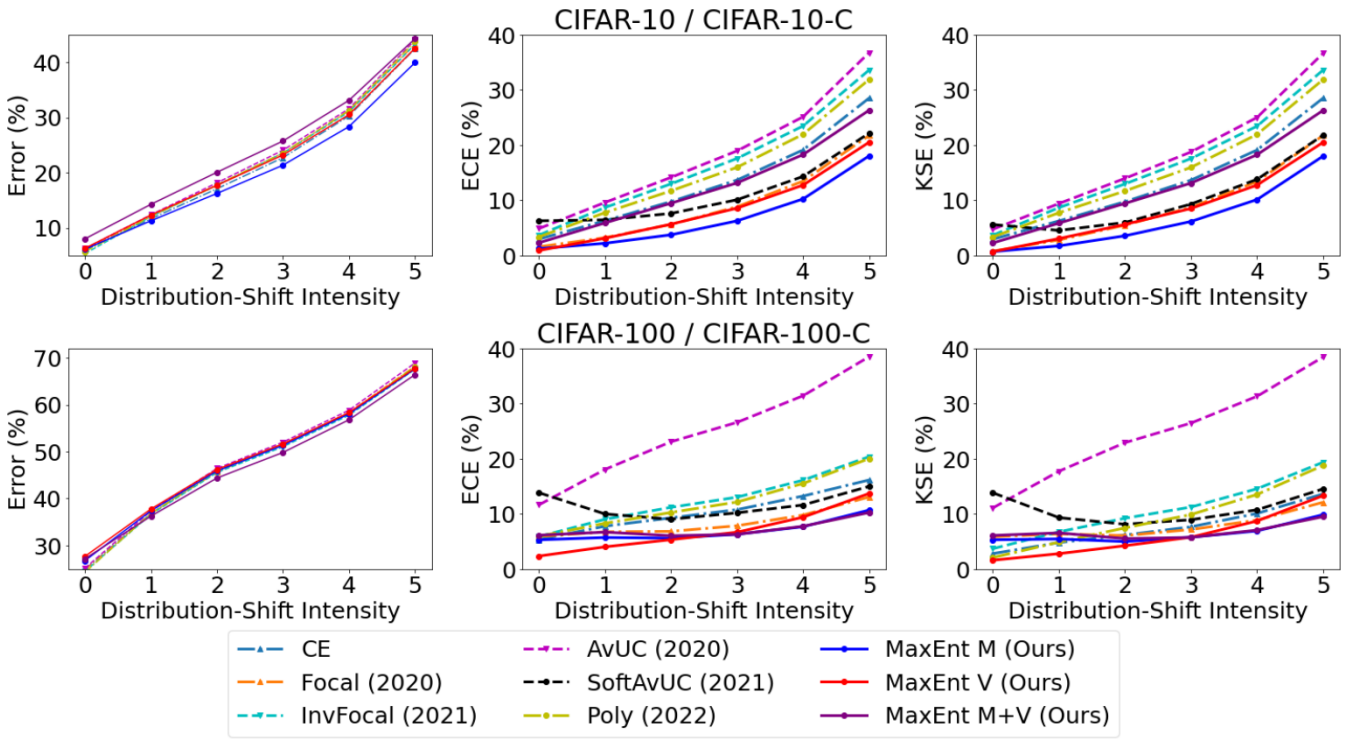

Figure 3: Test and calibration error curves highlighting the performance of different loss functions on CIFAR/CIFAR-C. As distribution shifts worsen from 0 to 5, all methods converge to similar test errors, while our method remains well calibrated.

(Krishnan and Tickoo 2020), Soft AvUC (Karandikar et al. 2021) and Poly loss (Leng et al. 2022).

For our analysis on synthetic OOD, we use ResNet-18 and ResNet-50 (He et al. 2016) with SGD optimizer for CIFAR and TinyImageNet, respectively. For in-the-wild OOD, we use ResNet-18 and DenseNet-121 (Huang, Liu, and Weinberger 2017) with Adam optimizer (Kingma and Ba 2015). Additional details about the OOD corruptions, experimental setup, and hyperparameters are provided in the Appendix. We show examples of each dataset in Figure 4 and describe the details of the following tasks.

Synthetic OOD: We make use of standard benchmarks for CIFAR10/CIFAR100/TinyImageNet and their corrupted forms CIFAR10-C/CIFAR100-C/TinyImageNet-C.

1. CIFAR10/CIFAR100 (Krizhevsky and Hinton 2009) contains RGB colored images (32x32) with ten or hundred classes. 45,000/5,000/10,000 images for training/-validation/testing.
2. TinyImagenet (Deng et al. 2009) is a subset of ImageNet with 200 classes, with images of size 64x64. 100,000 for training and 10,000 for validation/testing.
3. CIFAR10-C/CIFAR100-C/TinyImagenet-C (Hendrycks and Dietterich 2019) The corrupted form of CIFAR and TinyImagenet, comprising a total of 19 different transformations, with the initial 10,000 images of severity level one and the last 10,000 images of severity five.

Real-world OOD: We use the following in-the-wild computer vision datasets with their provided ID training sets and

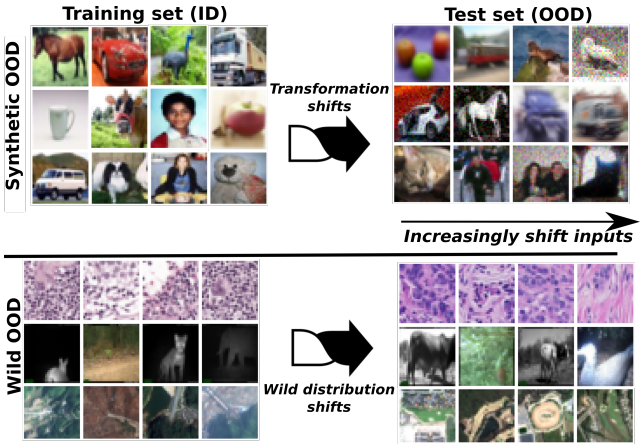

Figure 4: Samples from training and augmented validation/test sets for CIFAR10, CIFAR100 and TinyImageNet respectively (synthetic OOD, top 3 rows). Samples from Camelyon17-Wilds, iWildCam-Wilds and FMoW-Wilds are shown in the bottom 3 rows.

OOD sets for validation and testing. We denote real-world datasets with "Wilds" as per (Koh et al. 2021).

1. Camelyon17-Wilds (Bandi et al. 2019): Binary classification task on whether a (32x32) tissue slide contains any malignant/benign tumours.
2. iWildCam-Wilds (Beery, Cole, and Gjoka 2020): Static

Loss Fn.	(a) CIFAR10-C Accuracy	ECE	CECE	KSE	(b) CIFAR100-C Accuracy	ECE	CECE	KSE	(c) Tiny ImageNet-C Accuracy	ECE	CECE	KSE
CE	77.9±0.3	14.5±0.4	3.30±0.1	14.5±0.4	52.5±0.1	10.2±0.1	0.40±0.1	7.40±0.1	25.2±0.1	15.7±0.5	0.30±0.1	15.7±0.5
Focal	77.8±0.4	9.30±0.1	2.70±0.1	9.10±0.1	52.8±0.1	8.80±0.7	0.40±0.1	8.20±0.8	24.4±0.1	13.9±0.1	0.30±0.1	13.9±0.1
Inv Focal	77.9±0.1	16.4±0.2	3.60±0.1	16.4±0.2	52.8±0.1	13.1±0.1	0.50±0.1	11.8±0.1	26.4±0.1	16.3±0.4	0.30±0.1	16.3±0.4
AvUC	77.8±0.3	17.8±0.3	3.80±0.1	17.5±0.3	51.2±0.1	25.4±0.2	0.60±0.1	25.0±0.3	25.9±0.3	11.3±0.3	0.30±0.1	11.2±0.3
Soft AvUC	72.3±0.1	10.7±0.1	3.30±0.1	9.90±0.2	47.7±0.1	11.7±0.1	0.60±0.1	11.0±0.1	21.8±0.5	10.8±0.2	0.30±0.1	10.7±0.2
Poly	77.2±0.2	15.6±0.4	3.50±0.1	15.6±0.4	52.6±0.1	11.4±0.3	0.40±0.1	8.90±0.3	25.2±0.2	17.4±0.1	0.30±0.1	17.4±0.1
MaxEnt M	77.1±0.2	**8.70**±0.2	2.80±0.1	**8.50**±0.2	52.7±0.1	**6.30**±0.2	**0.40**±0.1	**5.90**±0.2	22.0±0.1	10.2±0.1	0.30±0.1	10.2±0.1
MaxEnt V	76.8±0.6	8.90±0.2	**2.70**±0.1	8.70±0.2	51.7±0.1	8.60±1.2	0.50±0.1	8.00±1.2	21.2±0.1	**9.40**±0.1	**0.30**±0.1	**9.40**±0.1
MaxEnt M+V	76.6±0.5	11.5±0.5	3.50±0.1	9.70±0.5	52.2±0.1	7.30±0.5	0.40±0.1	6.90±0.4	22.6±0.1	10.5±0.1	0.30±0.1	10.5±0.1
Loss Fn.	(d) Camelyon17-Wilds Accuracy	ECE	CECE	KSE	(e) iWildCam-Wilds Accuracy	ECE	CECE	KSE	(f) FmoW-Wilds Accuracy	ECE	CECE	KSE
CE	81.7±0.7	15.5±1.1	16.7±1.3	15.5±1.1	52.2±0.3	30.6±0.8	0.40±0.1	30.6±0.8	64.9±0.5	39.8±0.2	1.50±0.1	39.8±0.2
Focal	83.3±1.6	12.4±1.7	14.9±2.0	12.4±1.7	53.4±0.4	20.8±1.2	**0.30**±0.1	20.8±1.2	64.9±0.5	33.1±0.6	1.20±0.1	33.1±0.6
Inv Focal	84.3±2.7	14.2±2.7	15.0±2.8	14.2±2.7	55.9±0.6	29.9±0.8	0.40±0.1	29.9±0.8	64.8±0.4	15.7±2.0	0.70±0.1	15.7±2.0
AvUC	82.6±0.9	16.0±0.8	16.5±1.0	16.0±0.8	52.6±0.9	20.2±1.3	0.40±0.1	19.7±1.6	65.1±0.3	5.80±1.2	0.50±0.1	5.80±1.2
Soft AvUC	80.0±3.8	15.6±2.5	24.0±1.5	15.7±2.5	51.1±0.5	**18.7**±2.8	0.50±0.1	16.3±4.1	66.9±0.9	12.7±0.6	0.80±0.1	12.7±0.6
Poly	80.5±0.2	17.5±0.2	18.8±0.2	17.5±0.2	54.6±1.3	27.6±0.7	0.40±0.1	27.6±0.7	64.5±0.3	15.2±2.1	0.70±0.1	15.2±2.1
MaxEnt M	82.7±1.0	12.3±0.7	14.6±0.9	12.3±0.7	53.9±0.6	20.0±1.2	0.40±0.1	20.0±1.2	65.6±0.1	5.60±0.5	0.50±0.1	5.60±0.5
MaxEnt V	83.0±1.1	11.9±0.7	13.2±0.7	11.9±0.7	50.6±0.2	25.0±0.5	0.40±0.1	25.0±0.5	66.5±0.1	**4.60**±0.3	**0.50**±0.1	**4.70**±0.3
MaxEnt M+V	83.4±0.9	**8.30**±2.0	**12.2**±1.9	**8.30**±2.0	51.8±0.5	19.4±3.6	0.40±0.1	**12.2**±0.5	66.1±0.3	8.50±0.3	1.00±0.2	7.40±0.3

Table 1: Test scores (%) for synthetic (top) and real-world (bottom) OOD benchmarks computed across different approaches, with our method achieving state-of-the art OOD calibration. ± indicates the standard errors for 3 random seeds, with the best mean scores highlighted in bold.

camera traps deployed across different terrains with radical shifts in camera pose, background and lighting. The task is to identify the species in the photo out of 182 animal classes.

3. FMoW-Wilds (Christie et al. 2018): Satellite images across different functional buildings and terrain from over 200 countries. The task is to detect one out of 62 categories, including a "false detection" category.

Benchmarking Results

Synthetic OOD Evaluation We plot the test error, ECE and KSE in Figure 3 for each method on the different distribution shifts of CIFAR10/100-C. At level 0 of distribution shift (ID test set) we find that most methods are relatively well calibrated with $\leq 5\%$ ECE. However, they tend to become miscalibrated with higher test errors as the distributions increasingly shift away from the training set, with the poorest calibration coming from InvFocal and AvUC loss on both sets of CIFAR. On the other hand, Focal and Soft-AvUC loss perform relatively well OOD, with the best performance coming from our method. We note that the authors' original results for Soft-AvUC did not beat the Focal loss baseline, whereas in our experiments we choose the hyperparameters (T, κ) to the best of our ability, resulting in better performance than Focal loss OOD.

We compare the bin-strength and reliability diagrams (Niculescu-Mizil and Caruana 2005) for the different loss functions averaged across distribution shifts of CIFAR100-C in Figure 5. The bin-strengths plots show that the predictions of most methods remain concentrated in high-confidence bins, which suggests that these loss functions produce "peaky" distributions and over-confidence. In contrast, all three forms of our method mean (M), variance (V), mean plus variance (M+V) provide a better spread of predictions across bins, having significantly "softer" probabilities. From the reliability diagrams, the bars of most methods lie below the ideal diagonal, indicating that their predictions are over-confident and miscalibrated. On the other hand, our method produces bars that follow the diagonal more closely, demonstrating better calibration performance.

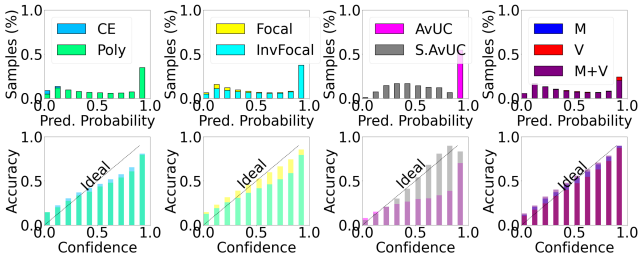

Figure 5: Bin-strength densities (top) and reliability diagrams (bottom) computed using $B = 10$ bins for different loss functions, evaluated on CIFAR100-C. MaxEnt Loss delivers a more uniform spread of probability densities and a reliability bar plot that better matches the ideal diagonal.

We further show the OOD results on the CIFAR10/100-C and TinyImageNet-C test sets, namely the accuracy, ECE, CECE (both computed with 15 bins) and KSE in Table 1. For the experiments shown in these tables, we use one-hot labels during training and report their performance without temperature scaling. We report the mean scores and the standard error (standard deviation divided by the square root of random seeds). When evaluated ID, all methods produce similar accuracies and are roughly within 95%, 75% and 50% for the original test sets of CIFAR and TinyImageNet respectively.

Real-world OOD Evaluation We further evaluate our method on real-world datasets, where distribution shifts occur naturally in-the-wild. Contrary to the synthetic datasets,

Dataset	MaxEnt Mean	MaxEnt Var	MaxEnt Mult	MaxEnt Mean w/ TS	MaxEnt Var w/ TS	MaxEnt Mult w/ TS	MaxEnt Mean w/ LS	MaxEnt Var w/ LS	MaxEnt Mult w/ LS
CIFAR10-C	8.70±0.2	8.90±0.2	11.5±0.5	↓6.90±0.1	↓**6.70**±0.2	↓9.30±0.5	9.60±0.7	9.80±0.9	11.6±0.1
CIFAR100-C	**6.30**±0.2	8.60±1.2	7.30±0.5	11.1±0.1	↓7.60±0.2	12.3±0.4	15.4±2.5	13.0±2.4	16.2±2.8
TinyImageNet-C	10.2±0.1	9.40±0.1	10.5±0.1	10.4±0.1	10.4±0.1	↓9.60±0.1	↓**9.00**±1.0	9.50±0.1	↓9.60±0.8
Camelyon17-Wilds	12.3±0.7	11.9±0.7	8.30±2.0	↓7.40±1.4	↓7.40±0.5	↓**4.60**±1.1	↓6.50±0.1	↓7.00±1.0	9.40±0.4
iWildCam-Wilds	20.0±1.2	25.0±0.5	19.4±3.6	↓8.20±0.8	↓**7.40**±1.0	↓11.8±2.0	↓7.40±1.1	↓9.20±0.4	↓18.5±0.1
FmoW-Wilds	5.60±0.5	4.60±0.3	8.50±0.3	↓**3.50**±0.4	↓4.40±0.3	↓4.00±0.9	6.20±0.9	4.60±0.3	↓7.60±2.3

Table 2: ECE (%) scores showcasing the effects of temperature scaling (TS) and label smoothing (LS) for the different OOD datasets. ↓ indicates improvements over the baseline and ± shows the standard errors with the best scores highlighted in bold.

the prior distribution for the Wilds datasets can be non-uniform, which might result in some form of bias for certain classes. Regardless, this does not negatively impact the performance of our method, as shown in Table 1. In general, we observe similar results as for synthetic OOD. Firstly, most loss functions produce relatively similar recognition accuracies across all datasets, regardless of synthetic or wild OOD. There are no significant drops in test set recognition when models are trained using our method. Secondly, models trained with MaxEnt loss are competitive and generally achieve state-of-the art performance alongside other baselines in terms of the various calibration metrics.

Importantly, MaxEnt loss is able to consistently deliver well calibrated models, even without the use of ad-hoc calibration techniques such as temperature scaling. We notice that the three forms of MaxEnt loss produce roughly similar performance, which is unsurprising since the constraints come from the same training set. However, as illustrated in Figure 2, these entropies are maximized subject to the given constraints, which can yield different results after optimization. This helps to explain why combining the mean and variance constraints does not necessarily lead to better calibration. In the Appendix, we report similar results when evaluating our method with different calibration metrics on both synthetic and wild OOD.

Pre- and Post-hoc Calibration

In this section, we discuss how MaxEnt loss complements commonly used ad-hoc techniques such as label smoothing and temperature scaling.

Pre-hoc Calibration: Label smoothing artificially softens the target distribution and seeks to encourage high entropy predictions (Müller, Kornblith, and Hinton 2019). Formally, the smoothed vector s_i is obtained after *uniformly* redistributing the probabilities of the correct class to other classes by a smoothing factor α: $s_i = (1-\alpha)y_k + \frac{\alpha}{K}$.

In contrast, our method performs smoothing *non-uniformly* with the help of constraints derived from the training distribution. In Appendix C.3, we can see that label smoothing only applies a linear shift in the bin-strength densities, whereas MaxEnt loss delivers different distributions with different global constraints. When label smoothing is applied together with our method, we are able to obtain models with even "flatter" bin-strength densities. We show the performance of our method with and without label smoothing in the right column of Table 2, using $\alpha = 0.01$. Label smoothing is only able to improve the ECE of our method in some cases. If the model is already well-calibrated, label smoothing can negatively affect calibration.

Post-hoc Calibration: For post-hoc calibration, we choose the standard temperature scaling technique, which linearly scales the classifier's output logits with a scalar $T > 0$. We follow the recommendations of Mukhoti et al. (2020) and perform grid-search over a typical range of temperature values [1.25, 1.50, 1.75, 2.00], picking the optimal temperature that minimizes the negative log-likelihood (NLL) (Hastie, Tibshirani, and Friedman 2001) of the validation set. We highlight that temperature scaling does not affect accuracy and is only helpful for model calibration under i.i.d. assumptions of the test set (Ovadia et al. 2019). In the case where the validation set is ID and not OOD shifted, temperature scaling may not be useful for improving calibration (Karandikar et al. 2021). As compared to label smoothing, our method exhibits the best calibration performance after temperature scaling, with significant improvements in ECE. This is particularly encouraging, because our method complements temperature scaling and does not restrict users with regards to post-hoc calibration.

Effects of Local Constraints

Figure 6 shows the empirical calibration error curves on CIFAR10-C, where we only consider the global constraints for each form of MaxEnt loss versus when the local constraints are included.

Firstly, without local constraints all three forms of our

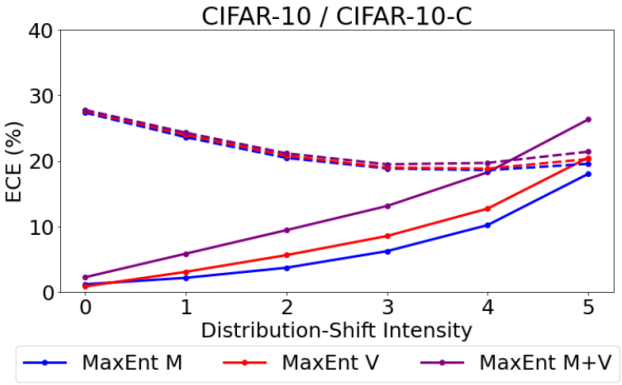

Figure 6: With only global constraints (dotted), predictions tend to be underconfident. Better calibration can be achieved by combining both global and local constraints (solid).

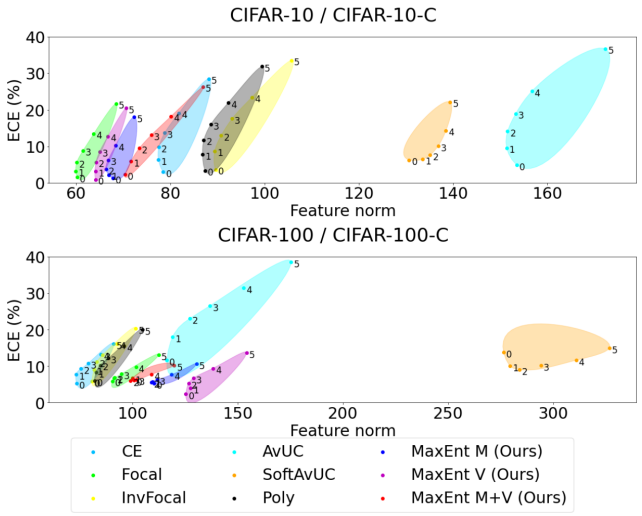

Figure 7: ECE vs. L2 norm of features from the penultimate layer of ResNet-18 trained using different methods, compared across CIFAR-C. Our method produces lower errors with smaller norms.

methods tend to be miscalibrated and underconfident for the lower intensity shifts. Secondly, all three global forms tend towards the same calibration performance across all shifts. With the inclusion of local constraints, all three forms of our method become better calibrated on lower intensity shifts, with the best calibration performance achieved by the mean constraint form.

Ordering of Feature Norms

We further analyze the performance of each loss function and show the ECE scatter plot as a function of the L2 norm of the learnt features in Figure 7. Specifically, each numbered point represents the average ECE per feature norm for the six intensity shifts extracted from the logits of the penultimate layer. We observe a correlation between ECE and feature norms, where lower feature norms tend to deliver smaller calibration errors. Apart from AvUC and SoftAvUC, most methods tend to support the findings from (Guo et al. 2017) that there is a strong relationship between miscalibration and overfitting. Our method generally produces small clusters with low calibration errors, which suggests that constraints provide robustness against distribution shifts and overfitting.

Limitations and Future Work

Choice of Constraints: In the current setup, we attribute the improvements in calibration to the given constraints. This assumes that the ratio/prior distribution $P(\mathcal{Y})$ observed during training is aligned with the test distribution. For test distributions that vary greatly from the training set, we believe that our method (along with many others) would not fare well. Since it is difficult to know the true distribution of the task (e.g. during model deployment), further approximations may be needed.

Unique Lagrange Multipliers: Our proposed framework approximates the Lagrange multipliers for each form of our method. Currently, we use a single Lagrange multiplier shared universally across both the global and local constraints. It may be possible to introduce unique Lagrange multipliers to control the trade-off between global and local constraints. However, this is non-trivial since it would require careful tuning and selection of hyperparameters, which is largely dependant on a separate validation set.

Conclusion

We presented MaxEnt loss, a novel loss function with multiple forms for calibrating deep neural networks across both synthetic and in-the-wild OOD computer vision datasets. We also showed the relationship between Focal loss and the Principle of Maximum Entropy. MaxEnt Loss achieves state-of-the art calibration with no significant increase in computation costs and requires only a few additional lines of code. Predictive uncertainty typically worsens under increasing dataset shift, whereas MaxEnt loss remains robust without any additional ad-hoc calibration. Furthermore, MaxEnt loss complements other ad-hoc calibration methods such as temperature scaling and label smoothing.

Acknowledgements

This research is supported in part by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). The authors would like to thank Tan Yingkiat and Bai Yunwei for their helpful suggestions and feedback.

References

Bandi, P.; Geessink, O.; Manson, Q.; Van Dijk, M.; Balkenhol, M.; Hermsen, M.; Bejnordi, B. E.; Lee, B.; Paeng, K.; Zhong, A.; et al. 2019. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2): 550–560.

Beery, S.; Cole, E.; and Gjoka, A. 2020. The iWildCam 2020 Competition Dataset. *arXiv*, abs/2004.10340.

Berger, A. L.; Della Pietra, S. A.; and Della Pietra, V. J. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1): 39–71.

Bojarski, M.; del Testa, D. W.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; Zhang, X.; Zhao, J.; and Zieba, K. 2016. End to End Learning for Self-Driving Cars. *arXiv*, abs/1604.07316.

Cheng, J.; and Vasconcelos, N. 2022. Calibrating Deep Neural Networks by Pairwise Constraints. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13709–13718.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional Map of the World. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6172–6180.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 248–255.

Du, X.; Wang, Z.; Cai, M.; and Li, S. 2022. Towards Unknown-aware Learning with Virtual Outlier Synthesis. In *Proc. International Conference on Learning Representations (ICLR)*.

Dubey, A.; Gupta, O.; Raskar, R.; and Naik, N. 2018. Maximum-Entropy Fine-Grained Classification. In *Advances in Neural Information Processing Systems*.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J. M.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542: 115–118.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *Proc. 34th International Conference on Machine Learning (ICML)*, volume 70, 1321–1330.

Gupta, K.; Rahimi, A.; Ajanthan, T.; Mensink, T.; Sminchisescu, C.; and Hartley, R. 2021. Calibration of Neural Networks using Splines. In *Proc. International Conference on Learning Representations (ICLR)*.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proc. International Conference on Machine Learning (ICML)*.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *Proc. International Conference on Learning Representations*.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proc. International Conference on Learning Representations (ICLR)*.

Huang, G.; Liu, Z.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.

Jaynes, E. T. 1957. Information Theory and Statistical Mechanics. *Physical Review*, 106: 620–630.

Karandikar, A.; Cain, N.; Tran, D.; Lakshminarayanan, B.; Shlens, J.; Mozer, M. C.; and Roelofs, R. 2021. Soft Calibration Objectives for Neural Networks. In *Advances in Neural Information Processing Systems*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *arXiv*, abs/1412.6980.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *arXiv*, abs/1312.6114.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; Lee, T.; David, E.; Stavness, I.; Guo, W.; Earnshaw, B.; Haque, I.; Beery, S. M.; Leskovec, J.; Kundaje, A.; Pierson, E.; Levine, S.; Finn, C.; and Liang, P. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proc. 38th International Conference on Machine Learning (ICML)*, volume 139, 5637–5664.

Krishnan, R.; and Tickoo, O. 2020. Improving model calibration with accuracy versus uncertainty optimization. In *Advances in Neural Information Processing Systems*.

Krizhevsky, A.; and Hinton, G. 2009. *Learning multiple layers of features from tiny images*. Master's thesis, Department of Computer Science, University of Toronto.

Kull, M.; Perello-Nieto, M.; Kängsepp, M.; de Menezes e Silva Filho, T.; Song, H.; and Flach, P. A. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *ArXiv*, abs/1910.12656.

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified Uncertainty Calibration. In *Advances in Neural Information Processing Systems*, volume 32.

Leng, Z.; Tan, M.; Liu, C.; Cubuk, E. D.; Shi, J.; Cheng, S.; and Anguelov, D. 2022. PolyLoss: A Polynomial Expansion Perspective of Classification Loss Functions. In *Proc. International Conference on Learning Representations (ICLR)*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. *arXiv*, abs/1708.02002.

Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F. A.; Zhai, X.; Houlsby, N.; Tran, D.; and Lucic, M. 2021. Revisiting the Calibration of Modern Neural Networks. In *Advances in Neural Information Processing Systems*.

Mukhoti, J.; Kulharia, V.; Sanyal, A.; Golodetz, S.; Torr, P. H.; and Dokania, P. K. 2020. Calibrating Deep Neural Networks using Focal Loss. In *Advances in Neural Information Processing Systems*.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When Does Label Smoothing Help? In *Advances in Neural Information Processing Systems*.

Naeini, M. P.; Cooper, G. F.; and Hauskrecht, M. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proc. 29th AAAI Conference on Artificial Intelligence*, 2901–2907.

Neo, D.; and Chen, T. 2023a. DSAC-C: Constrained Maximum Entropy for Robust Discrete Soft-Actor Critic. *ArXiv*, abs/2310.17173.

Neo, D.; and Chen, T. 2023b. FER-C: Benchmarking Out-of-Distribution Soft Calibration for Facial Expression Recognition. arXiv:2312.11542.

Neo, D.; Chen, T.; and Winkler, S. 2023. Large-Scale Facial Expression Recognition Using Dual-Domain Affect Fusion for Noisy Labels. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 5691–5699.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *Proc. International Conference on Machine Learning (ICML)*.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. *arXiv*, abs/1904.01685.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; and Snoek, J. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv*, abs/1906.02530.

Pereyra, G.; Tucker, G.; Chorowski, J.; Łukasz Kaiser; and Hinton, G. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. *arXiv*, abs/1701.06548.

Thulasidasan, S.; Chennupati, G.; Bilmes, J. A.; Bhattacharya, T.; and Michalak, S. E. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. In *Advances in Neural Information Processing Systems*.

Tomani, C.; Gruber, S.; Erdem, M. E.; Cremers, D.; and Buettner, F. 2021. Post-Hoc Uncertainty Calibration for Domain Drift Scenarios. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10124–10132.

Vonikakis, V.; Neo, D.; and Winkler, S. 2021. Morphset: Augmenting categorical emotion datasets with dimensional affect labels using face morphing. In *Proc. IEEE International Conference on Image Processing (ICIP)*.

Wang, D.-B.; Feng, L.; and Zhang, M.-L. 2021. Rethinking Calibration of Deep Neural Networks: Do Not Be Afraid of Overconfidence. In *Advances in Neural Information Processing Systems*.

Wang, X.; Long, M.; Wang, J.; and Jordan, M. 2020. Transferable Calibration with Lower Bias and Variance in Domain Adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 19212–19223.

Wiles, O.; Gowal, S.; Stimberg, F.; Rebuffi, S.-A.; Ktena, I.; Dvijotham, K. D.; and Cemgil, A. T. 2022. A Fine-Grained Analysis on Distribution Shift. In *Proc. International Conference on Learning Representations (ICLR)*.

Yu, Y.; Bates, S.; Ma, Y.; and Jordan, M. 2022. Robust Calibration with Multi-domain Temperature Scaling. In *Advances in Neural Information Processing Systems*.

Zadrozny, B.; and Elkan, C. 2002. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*.