Two Sides of Mis-Calibration: Identifying Over and Under-Confidence Prediction for Network Calibration

Abstract

Proper confidence calibration of deep neural networks is essential in safety-critical tasks for reliable predictions. Miscalibration can lead to model over-confidence and/or under-confidence; i.e., the predicted confidence can be greater or less than model accuracy. Recent studies have highlighted the over-confidence issue by introducing calibration techniques and demonstrated success on various tasks. However, miscalibration through under-confidence has not yet to receive much attention. In this paper, we address the necessity of paying attention to the under-confidence issue. We first introduce a novel metric, a miscalibration score, to identify the overall and class-wise calibration status, including being over or under-confident. Our proposed metric reveals the pitfalls of existing calibration techniques, where they often overly calibrate the model and worsen under-confident predictions. Then we utilize the class-wise miscalibration score as a proxy to design a calibration technique that can tackle both over and under-confidence. We report extensive experiments that show that our proposed methods substantially outperform existing calibration techniques. We also validate the automatic failure detection of the proposed calibration technique using our class-wise miscalibration score with a risk-coverage curve. Results show that our methods significantly improve failure detection as well as trustworthiness of the model.

1 INTRODUCTION

Effective confidence calibration for deep learning models improves the reliability of model predictions, which is imperative for practical deployment in safety-critical tasks including decision-making in medical diagnosis [Raghu et al., 2019, Dusenberry et al., 2020], autonomous driving systems [O'Brien et al., 2020, Zhang et al., 2018], assistance systems for socially vulnerable groups [Ashmore et al., 2021], and nuclear power plant monitoring [Linda et al., 2009]. A well calibrated model implies that the confidence scores for its predictions reflect its accuracy over a dataset, so that the model does not exhibit high confidence with low accuracy and vice versa. Due to the capacity of the model and complexity of the dataset, the model inherits a certain degree of miscalibration during training. Intuitively, the average predicted confidence can be higher or lower than accuracy, namely over-confidence and under-confidence. It is essential to address both for reducing calibration errors. Prior studies focus on over-confidence alone as many deep neural networks have been shown to be over-confident in fields such as computer vision, natural language processing and object detection [Müller et al., 2019, Guo et al., 2017, Lin et al., 2017]. Under-confidence, the other aspect of miscalibration, has not been addressed much in prior work. Under-confidence can be as problematic as over-confidence for applications such as automatic failure detection [Hendrycks and Gimpel, 2017, Corbière et al., 2019, Band et al., 2021]. Failure detection is a primary assessment for safety-critical tasks, whereby the system can decide to keep its predictions, reject them, or hand them to human experts for verification. One popular way of setting the threshold is by model confidence [Band et al., 2021, Corbière et al., 2019], where data points that fall below the confidence threshold are considered wrong predictions. If the model is under-confident, more samples will be predicted with lower confidence, including more correct predictions, unnecessarily increasing the burden on human experts to verify, and thus limiting the practical applications of the decision-making system.

As interest in tackling miscalibration has increased, several calibration techniques have been introduced and proven to be effective, such as label smoothing [Müller et al., 2019], temperature scaling [Guo et al., 2017], focal loss [Lin et al., 2017], and Deep Ensembles [Lakshminarayanan et al., 2017]. To categorize their methodologies, these techniques

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

aim to soften or squeeze either ground truth, logits or predicted probabilities. Despite their success, these techniques mainly target over-confidence. For example, in temperature scaling, Guo et al. [2017] pointed out that the temperature value T should be greater than 1 to soften the softmax predictions. This statement might not be suitable when the model is under-confident, as the way to calibrate the under-confident model should be to scale up the confidence instead of softening them. Furthermore, a technique designed for over-confidence may worsen under-confident predictions, which could diminish the interpretability of the model.

There are several metrics introduced to quantify calibration errors of the predicted confidence, such as Expected Calibration Error (ECE) [Naeini et al., 2015] and its variant to evaluate class-wise ECE [Obadinma et al., 2021], Static Calibration Error (SCE) [Nixon et al., 2019], Thresholded Adaptive Calibration Error (TACE) [Nixon et al., 2019] and Brier Score (BS) [Ashukha et al., 2020]. However, current evaluation metrics are unable to reflect the under or over-confidence status of the model, either in an overall or class-wise way, as they quantify the absolute calibration error. Identifying a model being over or under-confident in general, as well as class-wise miscalibration status, could be a guidance to fix the calibration error in a better way.

In this work, we quantify models as being under and over-confident, both overall and class-wise by designing novel metrics: Mis-Calibration Score (MCS) and class-wise Mis-Calibration Score (cwMCS). We use the class-wise mis-calibration score as the guidance to design calibration techniques that can tackle both over and under-confidence issues. In addition, we investigate pitfalls of the SOTA calibration techniques and find that most of them overly calibrate the model, leading to more severe under-confidence issues. Our contributions and findings are summarized as below:

1. We design metrics to identify over and under-confidence, both overall and class-wise.

2. We propose novel calibration techniques by utilizing mis-calibration score that can tackle both under and over-confidence issues.

3. With extensive experiments and observations we find that: (i) Unexpectedly, many of the recent SOTA models tend to be under-confident (e.g, Vision Transformers (ViT) [Dosovitskiy et al., 2020] and its variants); (ii) Many of existing calibration techniques can overly calibrate the model, leading to worse under-confidence; (iii) Our proposed calibration methods can benefit automatic failure detection by producing well-calibrated confidence estimates.

2 RELATED WORK

2.1 CALIBRATION TECHNIQUES

To alleviate the mis-calibration problem for deep neural networks, several techniques have been proposed and then widely applied as in-training or post-processing approaches. Label Smoothing (LS) [Szegedy et al., 2016] reduces over-confidence by computing the cross-entropy loss with uniformly squeezed labels instead of one-hot labels. Extensive work of LS such as Margin-based Label Smoothing (MBLS) [Liu et al., 2022] further provides a unifying constrained-optimization perspective of calibration losses. Focal Loss (FL) [Lin et al., 2017] adds a focusing factor to the standard cross-entropy loss to deal with an imbalanced dataset. The recent work sample-dependent focal loss (FLSD) [Mukhoti et al., 2020] investigate the effect of the loss on the training data and achieves impressive performance in calibration. Temperature Scaling (TS) [Guo et al., 2017] is a simple and straightforward method by applying a single-parameter variant of Platt Scaling [Guo et al., 2017] to logits produced from classifiers.

2.2 CALIBRATION METRICS

To measure the performance of calibration methods, the Expected Calibration Error (ECE) [Naeini et al., 2015] was proposed and widely applied in various tasks, such as image classification [Guo et al., 2017, Nixon et al., 2019] and sentiment analysis [Müller et al., 2019, Obadinma et al., 2021]. ECE splits the data into bins, calculates for each bin the average confidence and average accuracy, and averages over all bins. Despite the success of ECE, studies have shown certain limitations of it. First, ECE is sensitive to the number of bins. Theoretically, the larger number of bins indicates a smaller bias in the approximation, but the variance will increase with the growing number of bins due to fewer data points to calculate in some bins [Nixon et al., 2019]. In addition, ECE is unable to reveal the class-wise ECE status as it runs on the data as a whole. More precisely, if under and over-confident data points fall into the same bin, the calibration error for this bin can be severely underestimated [Nixon et al., 2019]. To better assess the degree of model calibration, Obadinma et al. [2021] proposed Contraharmonic Expected Calibration Error (CECE), Macro Subset Expected Calibration Error (MSECE) and Weighted Subset Expected Calibration Error (WSECE), respectively, to take the class-wise calibration error into consideration, especially for imbalanced datasets. However, the metric to reveal the miscalibration status as over or under-confident is still lacking in current calibration measurements.

2.3 CONFIDENCE CALIBRATION FOR FAILURE DETECTION

One main practical goal for confidence calibration is failure detection, which allows the model to detect misclassified samples by filtering out samples with the low-confidence score. More specifically, failure detection aims at distinguishing correct and incorrect predictions based on their confidence or uncertainty ranking. With a fixed or learned pre-defined threshold, the erroneous samples will be rejected by the model, then passed to human experts or other systems for a second opinion. Failure detection has been widely used in medical and manufacturing fields, such as AI for breast cancer screening [Leibig et al., 2022] and decision-making models for low-power Internet of Things (IoT) devices [Cho et al., 2020]. Hence calibrated confidence is essential to verify the trustworthiness of the model. Despite the significance, how calibration could benefit misclassification detection still needs more attention. Zhu et al. [2022] claimed that most calibration techniques are harmful for failure detection as they worsen the confidence separation between correct and incorrect samples. In this work, we argue that with proper techniques, calibration can benefit failure detection in certain tasks.

3 METHODOLOGY

3.1 PRELIMINARIES

Consider input samples $X = \{x_1, x_2, \ldots, x_N\}$ with N as the sample size and K as the numbers of classes.

Temperature Scaling (TS) As a simple extension of Platt scaling [Guo et al., 2017], Temperature Scaling (TS) [Guo et al., 2017] is a post-hoc approach to rescale logits with a single scalar T. A logit vector z is generated by a neural network with a given input x. T is applied on the logit z then it passes to the softmax function (denoted as σ), the new predicted confidence $\hat{p}^{TS}$ is:

$$\hat{p}^{TS} = \max \sigma_(z/T) \tag{1}$$

Expected Calibration Error (ECE) The Expected Calibration Error (ECE) [Naeini et al., 2015] is one of our base evaluation metrics for calibration performance. This metric represents the absolute difference between the predicted model confidence and accuracy. The predictions are sorted and grouped into M equal-spaced bins. Let B_m denote the set of samples with predicted value falling into the mth bin, where the interval is $\left[\frac{i-1}{M}, \frac{i}{M}\right]$. Given $\mathbb{I}$ as the indicator function, the average accuracy acc_m of B_m is: $acc_m = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{I}(\hat{\mathbf{y}}_i = \mathbf{y}_i)$. Likewise, the average confidence $conf_m$ of B_m is calculated by averaging the confidence of all samples in the bin, which is written as $conf_m = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$. ECE then can be estimated by taking a weighted average of the absolute difference between the accuracy and confidence in each bin, which can be formulated as:

$$ECE = \sum_{m}^{M} \frac{|B_m|}{N} |conf_m - acc_m| \tag{2}$$

where N is the number of samples.

Note that ECE fails to show whether the the calibration error is caused by under or over-confident predictions.

Weighted Subset Expected Calibration Error (wsECE) To present the class-wise ECE, the data points are divided into bins in a class-wise way. First, divide the dataset X, Y into K subsets where each sample x_n has the target $y = k$. Then pass each of these subsets to the ECE formula and get the result of individual ECEs with the size of K. Hence the class-wise ECE (cwECE) for each class is the set:

$$cwECE = \{ECE_1,\ ECE_2,\ \ldots,\ ECE_K\} \tag{3}$$

To represent the class-wise ECE comprehensively, Obadinma et al. [2021] proposed the method called weighted subset Expected Calibration Error (wsECE). As in imbalanced datasets, calibration error is unlikely to be the same for all classes. Error on minority classes can be much worse than the overall calibration performance. Hence the weight of each subset is significant to measure the overall calibration. Each subset ECEs is weighted by the percentage of data points of the class, and wsECE calculates the weighted average of all ECEs. Suppose the sample size in each class is $\{n_1, n_2, \ldots, n_K\}$, the calculation of wsECE can be written as:

$$wsECE = \frac{n_1}{N} ECE_1 + \frac{n_2}{N} ECE_2 + \ldots \frac{n_K}{N} ECE_K \tag{4}$$

wsECE gives an insight into the calibration performance of each class, including marginal classes of most under and over-confident ones. However, wsECE also does not distinguish between under and over-confidence.

3.2 PROPOSED METHODS

In this section, we design a novel metric, Mis-Calibration Score (MCS), to measure the calibration error, taking into account whether the model is under or over-confident. We also extend MCS to weighted subset MCS by considering the sample size in each class for imbalanced datasets. Then we integrate MCS with Temperature Scaling (TS) to address limitations with traditional TS techniques.

Mis-Calibration Score (MCS) To reveal the under and over confidence levels in each class, we first take the difference $conf_m - acc_m$ instead of the absolute value shown in equation (2). In this case, positive values indicate over-confidence, and negative values indicate under-confidence. With this slight modification, the Mis-Calibration Score (MCS) is formulated as:

$$MCS = \sum_{m}^{M} \frac{|B_m|}{N} \left(conf_m - acc_m\right) \tag{5}$$

To obtain the class-wise MCS (cwMCS), similar to the calculated method of cwECE (Eq. 3), the MCS for each subset is $cwMCS = \{MCS_1, MCS_2, \ldots, MCS_K\}$, which clearly shows the level of over or under confidence for each class.

The MCS can be positive (+) for over-confidence and negative (-) for under-confidence. Given, a dataset with K classes in total, the number of under-confident classes is k^-, the number of over-confident classes is k^+, the sample size for each under-confident class is $\{n_{1^-}, n_{2^-}, \ldots, n_{k^-}\}$, and the sample size for each over-confident class is $\{n_{1^+}, n_{2^+}, \ldots, n_{k^+}\}$. The corresponding MCS in each under and over-confidence class are: $\{MCS_{1^-}, MCS_{2^-}, \ldots, MCS_{k^-}\}$ and $\{MCS_{1^+}, MCS_{2^+}, \ldots, MCS_{k^+}\}$.

Due to the learning dynamics of the model and the complexity of the dataset, even for the balanced dataset, the number of under and over-confident classes is not likely to be equal. Therefore, inspired by the wsECE [Obadinma et al., 2021], which considers the class size, we introduce a weighted subset MCS (wsMCS) to address severe imbalanced miscalibration status (under or over-confidence) in predictions. First, we group all under-confidence and over-confidence classes and calculate their MCS with the sample size in each class, namely $wsMCS^-$ and $wsMCS^+$. The miscalibration error from the group with minority classes can be worse than what is indicated by the overall calibration, which can dominate the entire calibration performance. Then, to avoid this issue, we take the under and over-confidence class number into consideration to obtain overall wsMCS, and the final formula can be written as:

$$wsMCS^+ = \frac{n_{1^+}}{N} \mathrm{MCS}_{1^+} + \frac{n_{2^+}}{N} \mathrm{MCS}_{2^+} + \ldots \frac{n_{k^+}}{N} \mathrm{MCS}_{k^+} \tag{6}$$

$$wsMCS^- = \frac{n_{1^-}}{N} \mathrm{MCS}_{1^-} + \frac{n_{2^-}}{N} \mathrm{MCS}_{2^-} + \ldots \frac{n_{k^-}}{N} \mathrm{MCS}_{k^-} \tag{7}$$

$$wsMCS = \frac{k^+}{K} wsMCS^+ + \frac{k^-}{K} wsMCS^- \tag{8}$$

cwMCS Temperature Scaling (cwMCS TS) A scalar value of temperature (T) is unable to fulfill the calibration requirement for both over and under-confident classes, especially for marginal cases. We hereby integrate the class-wise Mis-Calibration Score (cwMCS) with T to formulate a vector (T^{cwMCS}) with the size of K:

$$T^{\mathrm{cwMCS}} = T \cdot (1 + \gamma \cdot \mathrm{cwMCS}) \tag{9}$$

where $cwMCS$ is max-normalized, γ is the weighting parameter to control the effect of class-wise miscalibration score into temperature T, which is tuned in the range of > -1 to $< +1$ with the increment of 0.001. The new confidence prediction $\hat{p}^{T^{\mathrm{cwMCS}}}$ with T^{cwMCS} is now:

$$\hat{p}^{T^{\mathrm{cwMCS}}} = \max\ \sigma\left(z / T^{\mathrm{cwMCS}}\right) \tag{10}$$

where σ is the softmax function. As MCS can be negative or positive, this allows for scaling up or scaling down, which is capable to fix under-confidence as well as over-confidence.

4 EXPERIMENTAL SETUP

4.1 DATASETS AND BASELINES

We validate the proposed calibration method with four benchmark image classification datasets: ImageNet 2012 (IN) [Russakovsky et al., 2015], Tiny-ImageNet (Tiny-IN) [Deng et al., 2009], and CIFAR10 (C10) and CIFAR100 (C100) [Krizhevsky et al., 2009]. For baselines, we use state-of-the-art (SOTA) Vision Transformer (ViT) [Dosovitskiy et al., 2020]r and its variants such as Swin-Transformer (SwinT) [Liu et al., 2021], Data-efficient Image Transformers (DeiT) [Touvron et al., 2021a], Class-Attention in Image Transformers (CaiT) [Touvron et al., 2021b], Bidirectional Encoder representation from Image Transformers (BEiT) [Bao et al., 2021], Co-scale Conv-attentional image Transformers (CoaT) [Xu et al., 2021], Cross-Attention Multi-Scale Vision Transformer (CrossViT) [Chen et al., 2021], ConvMixer [Trockman and Kolter, 2022] with the ImageNet pretrained weights from TIMM [1] library. To report comprehensive results on various models architectures, we also use the Convolutional neural networks (CNNs) in our experiments, namely EfficientNet [Tan and Le, 2019], DenseNet121 [Huang et al., 2017], ResNet34 [He et al., 2016], and VGG16 [Simonyan and Zisserman, 2014]. All models are with pretrained weights of ImageNet dataset. For recent SOTA calibration techniques MBLS [Liu et al., 2022] and FLSD [Mukhoti et al., 2020], we utilize the pre-trained model and official implementation from the repository [2].

We also verify our proposed methods on three text classification dataset: Stanford Sentiment Treebank v2 (SST-

[1] https://github.com/rwightman/pytorch-image-models

[2] https://github.com/by-liu/MbLS

Table 1: Results of baselines, TS and our proposed class-wise miscalibration score TS (cwMCS TS). The image datasets IN, Tiny-IN, C100 and C10 refer to ImageNet, Tiny-ImageNet, CIFAR100 and CIFAR10, and text dataset SST2, EMO, SENT mean Stanford Sentiment Treebank v2, TWEETEVAL Emotion and Sentiment respectively. ECE and wsECE are shown as percentages. wsECE is weighted subset ECE, and wsMCS is our weighted subset miscalibration score. Best calibration results for each row are shown in bold. The best result for wsMCS is calculated using the absolute value.

Dataset	Model	Acc (%)	Baseline			TS			cwMCS TS		
			ECE(%)	wsECE(%)	wsMCS	ECE(%)	wsECE(%)	wsMCS	ECE(%)	wsECE(%)	wsMCS
IN	ViT	83.30	2.20	13.40	0.393	0.90	14.00	-0.353	**0.70**	13.80	**-0.219**
	SwinT	84.40	8.70	18.50	-3.354	2.10	14.10	-0.384	**1.40**	**13.80**	**-0.135**
	DeiT	81.30	6.40	18.60	-2.477	3.30	15.30	1.620	**2.90**	**15.30**	**0.180**
	CaiT	83.00	3.60	15.80	-0.815	3.10	14.90	-0.032	**2.90**	**14.90**	**-0.002**
	BeiT	85.90	5.70	16.10	-2.292	4.20	13.40	-0.503	**2.00**	**13.20**	**-0.335**
	CoaT	80.70	7.60	19.50	-2.916	3.00	15.80	0.940	**2.20**	**15.70**	**0.126**
	CrossViT	81.90	5.20	17.90	-2.060	2.80	15.50	0.950	**2.50**	**15.50**	**0.071**
	ConvMixer	79.30	17.40	25.10	-6.720	8.90	16.20	-0.780	**0.80**	**16.10**	**-0.254**
	ConvNext	83.50	3.60	15.00	-0.294	3.40	14.60	0.174	**1.30**	**4.20**	**0.078**
	ResNet34	71.20	4.00	**18.30**	1.116	**1.60**	18.40	**-0.004**	**1.60**	18.40	**0.105**
	DenseNet121	71.80	3.00	**18.20**	0.182	1.40	18.40	**-0.017**	**1.30**	18.40	-0.097
	VGG16	69.90	3.30	18.30	0.908	**1.80**	18.50	**-0.046**	**1.80**	18.50	-0.049
	EfficientNet	75.60	14.50	24.80	-5.508	14.40	25.60	-4.409	**1.40**	**20.70**	**0.018**
Tiny-IN	ResNet34	59.70	9.30	18.30	0.330	8.30	16.90	-0.140	**4.30**	**14.90**	**-0.034**
	DenseNet121	61.80	13.00	19.10	0.507	9.20	16.40	-0.330	**2.10**	**14.40**	**-0.032**
	VGG16	61.60	15.60	20.10	0.617	82.90	16.60	-0.032	**2.80**	**10.20**	**-0.024**
C100	ResNet34	70.80	13.40	16.10	0.134	5.60	11.60	0.062	**1.60**	**1.20**	**-0.010**
	DenseNet121	66.20	14.80	17.70	0.147	7.20	12.30	-0.060	**2.10**	**4.60**	**-0.010**
	VGG16	64.30	9.90	14.00	0.097	5.40	11.10	-0.076	**1.30**	**1.10**	**-0.006**
C10	ResNet34	90.20	4.10	4.60	0.041	1.10	2.60	0.002	**0.50**	**1.50**	**0.001**
	DenseNet121	81.70	11.30	11.30	0.114	1.70	4.30	0.012	**0.80**	**2.20**	**0.001**
	VGG16	87.60	7.10	7.40	0.071	3.70	3.50	0.037	**1.70**	**1.60**	**0.012**
SST-2	RoBERTa	94.00	4.29	4.60	0.081	3.21	2.80	0.045	**1.21**	**1.70**	**0.005**
	BERT	92.40	6.09	6.20	0.107	3.10	4.60	0.064	**1.09**	**2.60**	**0.014**
	DistilBERT	91.10	8.00	8.00	0.147	4.17	6.90	0.069	**3.56**	**5.00**	**0.054**
EMO	RoBERTa	83.40	3.66	7.20	0.194	2.70	6.20	**0.036**	**1.84**	**5.80**	0.074
	BERT	81.00	4.52	8.40	-0.096	1.66	7.50	0.070	**1.29**	**7.20**	**0.058**
	DistilBERT	80.20	15.70	19.50	0.765	6.52	9.80	0.303	**4.31**	**8.20**	**0.260**
SENT	RoBERTa	72.40	2.16	**7.90**	0.048	1.96	8.00	**-0.018**	**1.54**	8.00	0.039
	xlm_RoBERTa	68.20	4.64	9.60	0.052	3.35	10.80	-0.084	**2.10**	**9.00**	**-0.023**

2) [Socher et al., 2013], TWEETEVAL-Emotion and TWEETEVAL-Sentiment [Barbieri et al., 2020]. For baselines, we use BERT [Devlin et al., 2018], RoBERTa [Liu et al., 2019] and DistilBERT [Sanh et al., 2019] with Hugging Face implementations [3].

As our experiment is in a post-processing approach, we primarily utilize each dataset's test set. For the ImageNet dataset, we equally divide its original test set of 50,000 images into validation and test sets for a fair comparison. For Tiny-ImageNet, CIFAR10/100 and all three dataset, an 80/10/10 for training/validation/test split is applied. The dataset setting is the same as other experiments for the investigation of existing calibration techniques (shown in table 3).

4.2 IMPLEMENTATION DETAILS

For the fair comparison, we utilized existing pretrained weights for most of our investigation and experimentation. For example, all the observations with ImageNet and existing calibration techniques are from pretrained weights except Tiny-ImageNet and CIFAR10/100 dataset are applied for model training and testing with the networks of ResNet34, DenseNet121 and VGG16. The models are trained with stochastic gradient descent (SGD) optimizer with momentum of 0.9 and weight decay of 0.0005. The initial learning rate for ResNet34 and DenseNet121 is 0.1, and 0.001 for VGG16. The batch size for all training is 1024 and for validation and test is 2048. The GPU of the Nvidia Tesla P40 is used for all experiments. For the evaluation metric of ECE, wsECE and our proposed method, we set the number of bins as $M = 15$.

5 RESULTS

In this work, we first conduct the experiment to identify the under-confident using our proposed metric of wsMCS. Then we observe the pitfalls of the existing calibration techniques and the limitation of the conventional calibration metric of ECE. Finally, we show that cwMCS based TS can effectively tackle both under and over-confidence. We also validate the efficacy of our cwMCS TS on a real-world application of automatic failure detection. All the experiment and results are shown in the Tables of 1 3, and 2, and Figures of 1 and 2.

[3] https://huggingface.co

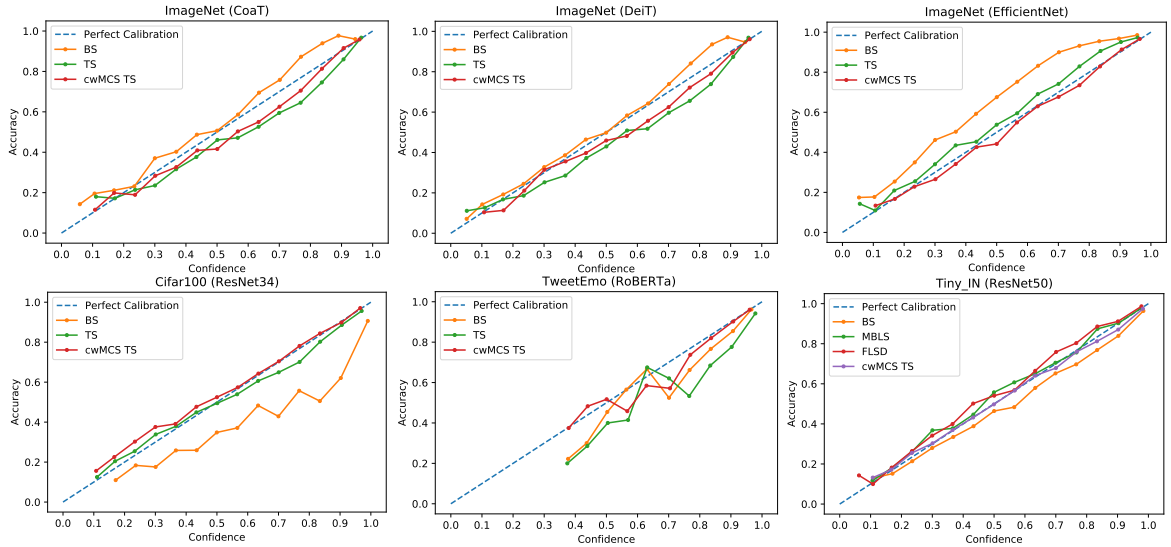

Figure 1: Reliability Diagram (RD) for ImageNet, Tiny-ImageNet, CIFAR100 and TweetEval-Emotion with calibration techniques. BS, TS, MBLS, FLSD, cwMCS TS mean baseline, temperature scaling, margin-based label smoothing, sample-dependent focal loss and class-wise Miscalibration Score TS respectively. First row from left to right: RD for ImageNet with CoaT, DeiT and EfficientNet. Second row from left to right: RD for CIFAR100 with ResNet34, TweetEval-Emotion with RoBERTa, and the comparison of SOTA calibration techniques on Tiny-ImageNet with RestNet50.

Table 2: Results of average under-confidence mis-calibration score (UC MCS) and average over-confidence miscalibration score (OC MCS) for baseline, TS and our proposed method cwMCS TS.

		Baseline		TS		cwMCS TS	
Dataset	Model	UC MCS (%)	OC MCS (%)	UC MCS (%)	OC MCS (%)	UC MCS (%)	OC MCS (%)
IN	ViT	-4.2 (48.3)	7.4 (51.7)	-5.2 (61.3)	7.2 (38.7)	**-3.7 (58.3)**	**0.5 (41.7)**
	SwinT	-11.6 (85.8)	9.2 (14.2)	-5.3 (64)	9.2 (36)	**-0.5 (65.7)**	**8.7 (34.3)**
	DeiT	-10.4 (79.6)	9.0 (20.4)	-5.2 (54.7)	9.6 (45.3)	**-3.2 (55.3)**	**5.6 (44.7)**
	ConvMixer	-19.0 (90.4)	**8.1 (9.6)**	-8.9 (61.5)	8.6 (38.5)	**-5.7 (59.7)**	8.4 (40.3)
Tiny-IN	ResNet34	**-3.1 (11.5)**	10.6 (88.5)	-5.8 (58)	6.3 (42)	-5.4 (52)	**4.3 (48)**
	DenseNet121	-1.0 (2.5)	13.4 (97.5)	-5.5 (58.5)	7.6 (41.5)	**-5.2 (56.5)**	**6.6 (43.5)**
	VGG16	**-0.8 (1.5)**	15.9 (98.5)	-5.2 (57.5)	5.1 (42.5)	-4.9 (52.5)	**3.1 (47.5)**
C100	Res34	**-0.1 (5)**	13.4 (95)	-5.2 (61)	6.2 (39)	-5.0 (57)	**4.2 (43)**
	DenseNet121	**-1.3 (1)**	15 (99)	-5.7 (55)	6.3 (46)	-4.3 (58)	**5.6 (42)**
	VGG16	**-1.3 (3)**	10.3 (97)	-4.4 (56)	8.7 (44)	-4.4 (56)	**3.7 (44)**
C10	ResNet34	0.0 (0)	4.1 (100)	**-1.4 (30)**	2.1 (70)	-1.5 (30)	**1.5 (70)**
	DenseNet121	0.0 (0)	11.3 (100)	-1.6 (40)	5.4 (60)	**-0.2 (30)**	**5.2 (70)**
	VGG16	0.0 (0)	7.1 (100)	-0.9 (40)	2.8 (60)	**-0.7 (50)**	**0.2 (50)**

5.1 IDENTIFYING UNDER-CONFIDENCE

To identify the under-confidence prediction, we adopt several Transformers, CNN BERT-based architectures and measure the accuracy, ECE, wsECE and proposed wsMCS. Table 1 presents the performance with image and text classification. Surprisingly, wsMCS for most of the Transformer and BERT-based architectures are negative for baselines which reveals the under-confidence prediction. It is also observed that the Transformer-based architecture (e.g., ConvMixer) with lower accuracy yields the higher ECE or wsECE but it is unable to reveal the specific level of the miscalibration. On the other hand, our wsMCS produces a high, prediction for the lowest-performed architecture. In terms of CNN-based architecture for image classification, the results suggest from wsMCS that most of these predictions are over-confident. In Table 2, almost more than half of the classes become under-confident after applying TS, indicating that TS can overly calibrate models. Our proposed method cwMCS TS significantly improves the mean over-confidence score and contributes to better calibration for under-confident classes. Additional observations of under-confidence can be found in appendix A.

5.2 PITFALLS OF EXISTING CALIBRATION TECHNIQUES

We first investigate several widely applied calibration techniques of LS, FL, TS, and recent works of MBLS and FLSD to check their calibration performance. Table 3 shows

Table 3: The comparison of SOTA calibration techniques and ours. Datasets Tiny-IN, C10 refer to Tiny-ImageNet and CIFAR10. All models are with pretrained weights of ResNet50 from Liu et al. [2022]. UC and OC are average under and over-confidence miscalibration score. wsMCS is overall miscalibration score and Acc is accuracy. All results except wsMCS are shown in percentage for clarity.

Dataset	Method	ECE(%)	wsMCS	UC(%)	OC(%)	Acc(%)
Tiny-IN	CE	3.77	6.356	-4.00	6.60	64.80
	TS	1.40	1.340	-4.50	5.60	64.80
	LS	3.07	-5.094	-60.00	4.80	65.50
	MBLS	1.86	-3.288	-5.50	5.20	64.70
	FL	3.10	-5.365	-6.30	4.30	63.20
	FLSD	2.84	-5.466	-3.90	5.00	63.90
	Ours	1.05	0.640	-1.90	5.00	64.90
C10	CE	6.51	0.650	0.00	65.00	92.40
	TS	5.34	0.532	0.00	32.00	92.40
	LS	2.79	0.332	-8.40	4.10	94.90
	MBLS	1.49	0.290	-0.40	2.70	94.70
	FL	3.78	-0.339	-4.20	0.30	94.30
	FLSD	3.61	-0.334	-33.40	0.00	94.20
	Ours	1.38	0.224	-2.50	22.00	92.40

the ECE, wsMCS, average under-confidence and over-confidence MCS and accuracy for each calibration method. The cross-entropy (CE) baseline is always over-confident as the wsMCS is positive for all experiments. However, after applying calibration techniques, wsMCS of FL and FLSD turns to negative values in results both dataset, which indicates the model is overly calibrated. On the other hand, the traditional calibration metric, ECE, cannot reveal the issue as shown in Table 3.

We further investigate TS with the most recent architectures of ViT and its variants with IN, Tiny-IN, C10/100 datasets in Table 1. Similar to other calibration techniques, TS also tends to over-calibrate the Transformer-based models. We also observe that more than half the classes become under-confident on average after the calibration (as shown in Table 2). These results suggest that models are overly calibrated as existing calibration techniques are designed to focus on the over-confidence issue. This is also observed in the reliability diagram in Figure 1. For example, after applying TS in ImageNet dataset with EfficientNet, FLSD and MBLS in Tiny-ImageNet dataset with ResNet50, the model turns into under-confidence as the curve of accuracy is clearly higher than the confidence. Therefore, techniques that can tackle both under and over-confident issues are required for better calibration.

5.3 EFFECTIVENESS OF PROPOSED CALIBRATION METHOD

Table 1 shows the calibration results TS and the proposed $\mathrm{cwMCS}\,TS$ validated with four image and three text dataset and various model architectures. In most experiments, our proposed methods outperform baselines and the TS method, with significant improvement of ECE and mis-calibration score. Notably, the accuracy remains the same for all methods in each experiment.

5.4 APPLICATION TO FAILURE DETECTION

To verify the effectiveness of our proposed method cwMCS TS on failure detection (FD), we utilize the predictive uncertainty to distinguish correct from incorrect samples by following [Corbière et al., 2019]. More specifically, we calculate the entropy of each sample's confidence. A smaller entropy means a bigger distance between class predictions within a predicted distribution, which suggests a higher probability for correct predictions. In addition, entropy is a symmetric measure regardless of the position in a distribution, where [0.8, 0.2] and [0.2, 0.8] produce the same entropy. It may cause an issue for class probabilities as each probability points to a target. From our observation, our approach can successfully discriminate the incorrect from correct predictions regardless of similar distributions.

Figure 2 shows the risk-coverage curves for Tiny-ImageNet with ResNet50 with the application of SOTA calibration techniques (Left), and ImageNet dataset accessed with vanilla (Right) and calibrated (Middle) architectures. The different percentages of 'Referred to Experts' means the samples removed from the test set, and the accuracy is calculated with the remaining data. We use the predictive uncertainty to rank all the samples in the test set, where samples with high uncertainty will be referred. A higher accuracy rates is better, as it suggests errors are concentrated in the low-confidence samples; i.e the method is better at discriminating correct and incorrect samples.

In Figure 2 (left), with the increment of percentage referred to experts, only our proposed method cwMCS and TS beat the baseline, where our proposed method outperforms TS especially between 0.15 and 0.4 proportion. LS is with the highest accuracy with the entire test set, but it does not continue until 5% samples removed. FLSD reaches the same performance as TS with the proportion of 0.2, then later falls behind LS and MBLS. The results shows that our proposed method performs better in error detection than SOTA calibration methods.

Figure 2 (Right) shows the FD performance of uncalibrated models, where SwinT is with the highest accuracy before any data removed. However, SwinT is surpassed at the point of proportion 0.15, then ViT outperforms all other models with the increment of samples referred. After calibrated with our proposed method (shown in Figure 2 (Middle)), SwinT maintains the highest accuracy until proportion 0.3, indicating the better trustworthiness with our method.

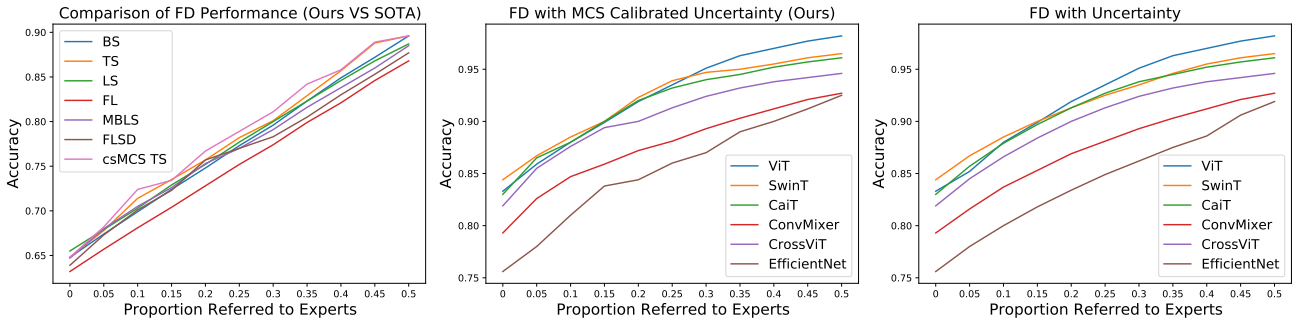

Figure 2: Risk-coverage Curves for failure detection (FD) performance. "Proportion Referred to Experts (x-axis)" means the percentage of data discarded, and "Accuracy (y-axis)" is the accuracy of remaining samples of the test set. Left: the comparison of FD on SOTA calibration techniques and ours on ResNet50 with Tiny-ImageNet. Middle: FD for ImageNet with calibrated models (our method); Right: FD for ImageNet with uncalibrated models (baseline).

6 DISCUSSION

In this work, we introduced a novel Mis-Calibration Score (MCS) to quantify over or under-confident predictions of models, both overall and class-wise. One limitation of the overall MCS is that it mostly indicates the trend of over or under-confidence, but not the magnitude of the miscalibration error. The reason is that we average the actual values of the difference between prediction confidence and accuracy instead of the absolute values. This results in lower scores as negative and positive (under and over-confidence) cancel each other out. However, the model still needs to be calibrated. To fix this issue, we calculate MCS for over and under-confident classes separately, namely average under-confidence MCS and average over-confidence MCS respectively, as shown in Table 2. The corresponding percentage of classes being under or over-confident is also presented in parentheses, from where we can tell how many classes contribute to each calibration status. However, this percentage distribution does not decide whether a model is well-calibrated or not. In other words, neither more under-confidence classes nor more over-confidence classes could guarantee less calibration error. The better calibration happens when their absolute value tends towards zero.

After quantifying the over and under-confidence for overall and class-wise ways (cwMCS), we utilize the cwMCS to design the MCS-aware calibration technique. Our proposed method effectively improve both ECE and WSECE (shown in Table 1), as well as tackle under and over-confidence issue separately (shown in Table 2). The reliability diagram in Figure 1 also validates it.

Even though some studies have claimed that the recent state-of-the-art deep neural networks tend to be over-confident, our findings in Table 1 and 2 are quite contrary. In the baseline of ViT and its variants, the overall MCS shown in Table 1 are mostly negative, and Table 2 suggests that the percentage of under-confident classes are much higher than over-confident ones. Furthermore, by comparing the recent calibration techniques and our proposed method, we find that many calibration methods overly calibrate the model. Table 3 and Figure 1 both show that the model turns under-confident after applying these techniques. For automatic failure detection result in Figure 2, the risk-coverage curve indicates that our method can filter incorrect samples more effectively within a smaller proportion. Hence solving calibration issue, especially the under-confidence, can improve the trustworthiness of the model. It also suggests that with proper calibration techniques, model confidence can help detect misclassification.

7 CONCLUSION

In conclusion, our work highlights that under-confidence is as crucial as over-confidence and hence should receive more attention from researchers. We introduced a novel metric Mis-Calibration Score (MCS), to identify under and over-confidence, overall and class-wise. Then we design class-wise MCS-aware calibration techniques to tackle both under and over-confidence issues. With extensive experiments, we validate that our proposed calibration technique can effectively address different miscalibration statuses. In addition, we observe that many of the current calibration techniques overly calibrate the model, inclined to a more severe under-confidence issue. Furthermore, our proposed method significantly improves error detection and proves that proper confidence calibration techniques can benefit failure detection for safety-critical tasks. For future work, we will apply our MCS-aware method to more calibration techniques, such as label smoothing and focal loss. As our experiments focus on image and text classification, we will also apply our proposed method to other tasks, such as text generation and image segmentation.

References

Rob Ashmore, Radu Calinescu, and Colin Paterson. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5): 1–39, 2021.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

Neil Band, Tim GJ Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*, 2020.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

Chiwoo Cho, Wooyeol Choi, and Taewoon Kim. Leveraging uncertainties in softmax decision-making models for low-power iot devices. *Sensors*, 20(16):4603, 2020.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems*, 32, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Michael W Dusenberry, Dustin Tran, Edward Choi, Jonas Kemp, Jeremy Nixon, Ghassen Jerfel, Katherine Heller, and Andrew M Dai. Analyzing the role of model uncertainty for electronic health records. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 204–213, 2020.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 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*, pages 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Christian Leibig, Moritz Brehmer, Stefan Bunk, Danalyn Byng, Katja Pinker, and Lale Umutlu. Combining the strengths of radiologists and ai for breast cancer screening: a retrospective analysis. *The Lancet Digital Health*, 4(7):e507–e519, 2022.

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

Ondrej Linda, Todd Vollmer, and Milos Manic. Neural network based intrusion detection system for critical infrastructures. In *2009 international joint conference on neural networks*, pages 1827–1834. IEEE, 2009.

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*, pages 80–88, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

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 Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, 2019.

Stephen Obadinma, Hongyu Guo, and Xiaodan Zhu. Class-wise calibration: A case study on covid-19 hate speech. In *Canadian Conference on AI*, 2021.

Molly O'Brien, William Goble, Greg Hager, and Julia Bukowski. Dependable neural networks for safety critical tasks. In *International Workshop on Engineering Dependable and Secure Machine Learning Systems*, pages 126–140. Springer, 2020.

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning*, pages 5281–5290. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021b.

Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9981–9990, 2021.

M Zhang, Y Zhang, L Zhang, C Liu, and S Khurshid. Deeproad: Gan-based metamorphic autonomous driving system. *Research Gate Publication*, 2018.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *European Conference on Computer Vision*, pages 518–536. Springer, 2022.