Calibration Transfer via Knowledge Distillation

Ramya Hebbalaguppe[1,2§], Mayank Baranwal[2§], Kartik Anand[1], Chetan Arora[1]

[1]Indian Institute of Technology, Delhi [2]Tata Consultancy Services Research
** source code: `https://github.com/rhebbalaguppe/CalibrationXferViaKD`

Abstract. Modern deep neural networks often suffer from miscalibration, leading to overly confident errors that undermine their reliability. Although Knowledge Distillation (`KD`) is known to improve student classifier accuracy, its impact on model calibration remains unclear. It is generally assumed that well-calibrated teachers produce well-calibrated students. However, previous findings indicate that teachers calibrated with label smoothing (`LS`) result in less accurate students [45]. This paper explores the theoretical foundations of `KD`, revealing that prior results are artifacts of specific calibration methods rather than `KD` itself. Our study shows that calibrated teachers can effectively transfer calibration to their students, but not all training regimes are equally effective. Notably, teachers calibrated using dynamic label smoothing methods yield better-calibrated student classifiers through `KD`. We also show that transfer of calibration can be induced from lower capacity teachers to larger capacity students (aka `reverse-KD`). The proposed `KD` based `Calibration` framework, named `KD(C)`, leads to a state-of-the-art (`SOTA`) calibration results. More specifically, on `CIFAR100` using `WRN-40-1` feature extractor, we report an `ECE` of 0.98 compared to 7.61, 7.00, and 2.1 by the current `SOTA` calibration techniques, `AdaFocal` [9], `ACLS` [41], and `CPC` [5] respectively, and 11.16 by the baseline `NLL` loss (lower `ECE` is better).

Keywords: Confidence Calibration · Knowledge Distillation

1 Introduction

Calibration. Deep neural network (`DNN`) models have become increasingly prevalent in critical applications such as healthcare [21, 33], and autonomous driving [3]. In such applications, it is crucial for `DNN` predictions to not only be accurate but also trustworthy [7, 38]. Yet, it has been shown that the softmax probabilities (referred to as *predicted confidence* in this paper) produced by `DNNs` come with no formal probabilistic guarantees [10]. *Calibration* refers to the alignment between a `DNN` model's predicted confidence and the actual frequency of the event it represents. Calibration indicates model's ability to provide reliable uncertainty estimates, and many modern `DNNs` are shown to be miscalibrated.

Reasons for Miscalibration and Our Investigation. Mukhoti *et al.* [36] have shown that a `DNN` model overfitting on `NLL` loss is the main reason behind

** Equal contribution

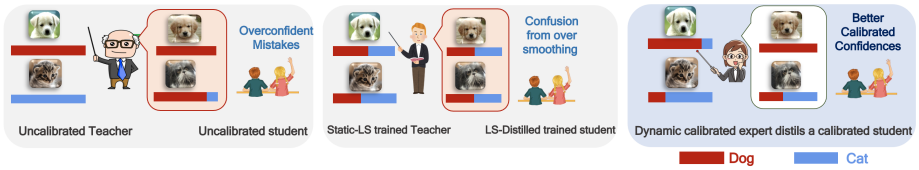

Fig. 1: Motivation. We examine a binary classification problem, using red and blue bars to show prediction confidence for dogs and cats. (**Left**): An uncalibrated teacher provides overconfident guidance, leading to student errors on challenging samples. (**Mid**): Static calibration methods like Label Smoothing (`LS`) [47]) excessively reduce confidence, potentially confusing students. (**Right**): Advanced calibration techniques (e.g., MDCA [13]) capture sample-level uncertainty, providing appropriate guidance and producing accurate, well-calibrated student models.

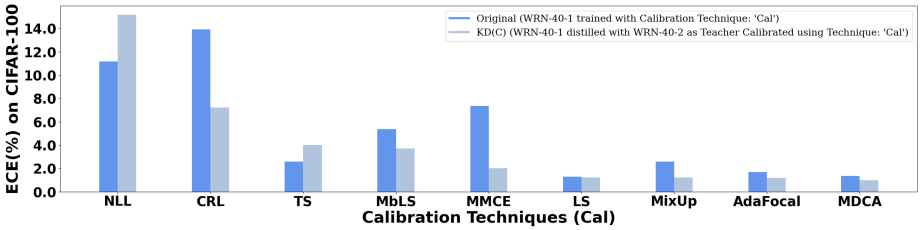

Fig. 2: Calibration via `KD`. We compare calibration performance (`ECE`, lower is better) of a student model (`WRN-40-1`) when directly calibrated (blue bars) versus trained via `KD` from a calibrated teacher model (`WRN-40-2`). `KD`-based calibration consistently outperforms direct calibration, except with `NLL` (no calibration), `TS`, and `LS` (sample-agnostic techniques). Our main contribution is a `KD`-based calibration framework compatible with various teacher models and sample-specific uncertainty calibration methods. Combined with `MDCA`, our approach achieves `SOTA` calibration performance.

highly overconfident predictions, leading to miscalibration. Further, [22] shows theoretically that `DNNs` with `ReLU` activation are susceptible to overfitting on `NLL` loss function. In this work, we explore if access to label uncertainties during training can prevent such overfitting and generate a calibrated classifier. Knowledge distillation (`KD`) has been used for transferring learned representations from a (typically large) teacher model to a (usually smaller) student model in the multitude of works. In this work, we investigate specifically if access to learned calibrated confidence through a teacher model also helps in the calibration of a student model.

Our Proposal: `KD` for Calibration. Unfortunately, existing explanations of the process of `KD` rarely go beyond simple qualitative statements attributing improved performance to learning from soft labels of the expert classifiers. Phuong and Lampert [42] provide first theoretical insight into the working mechanism of `KD`, albeit from an optimization viewpoint. Allen-Zhu and Li [1] elucidate the effectiveness of ensemble learning and `KD` in enhancing the test accuracy, but do

not focus on the transfer of calibration properties. In our work, we view the role of KD beyond its well-studied role of accuracy transfer and provide theoretical and empirical insights into the transfer of *calibration* to student classifiers. We show, for the first time, that calibrated teachers distill the best-calibrated students, and thus, propose a new recipe for producing an accurate and calibrated classifier using KD and a calibrated teacher model.

Departure from Current Belief: Does KD Conflict with Calibration? Interestingly, there is an influential prior work investigating the accuracy of a student model after KD from a teacher model trained using Label Smoothing (LS) [37]. However, LS was observed to impair KD, i.e., the accuracy of student classifiers degrade when teacher classifiers are calibrated with LS [45]. This discourages the use of KD for calibration. In our work, we show that this impairment is not the artifact of KD but of the LS itself, which when used to calibrate teacher classifiers and distill their representation to student classifiers at higher temperatures, ends up over-smoothing a student's predictions, thereby significantly degrading its accuracy [4]. We show that teachers trained via dynamic label-smoothing methods (e.g., [5, 9, 13]) consistently distil calibrated students across all temperatures. To this end, we highlight the role of KD in calibrating classifiers and argue strongly in favor of using knowledge sharing from calibrated experts to student classifiers as the most promising calibration technique.

Contributions. To achieve calibration of DNNs, we bring together two seemingly unrelated sub-fields: KD and confidence calibration. Our contributions include:

1. **Understanding calibration transfer via distillation**: We develop a theoretical framework to analyze KD and its ability to transfer the learning of a teacher to a student classifier and show, arguably for the first time, that calibrated teachers can distill calibrated students. We corroborate our proposed *calibration framework* through theoretical insights of calibration transfer for linear models backed by exhaustive experiments.
2. **Achieving best student DNN calibration**: Our experiments demonstrate that students trained via KD from teachers that are first calibrated using dynamic/adaptive label-smoothing, exhibit the best calibration compared to other train-time/post-hoc calibration techniques. (Sec. 4.3). Our framework is named KD(C) (**K**nowledge **D**istillation from a **C**alibrated teacher).
3. **Not all calibration techniques are compatible with KD**: It has been observed empirically that LS impairs KD [37]. This impairment is argued to be a high-temperature phenomenon [4]. In our experiments, we too observe a similar behavior when teacher classifiers are trained via static calibration methods, eg., LS. However, we show that when the teacher classifiers are calibrated using dynamic LS methods, the distillation produces calibrated student classifiers consistently across wide temperature regimes.
4. **Calibration distillation works both ways**: Similar to recent works in reverse-KD [18], where it has been shown that smaller teacher models can also distill accurate student models, we show that the same applies for calibration distillation as well, and smaller calibrated models can also yield better

calibrated larger models. The observation is consistent with our key insight that the availability of label ambiguities through soft-labels during training is extremely useful for calibration. This setting is relevant when large calibrated models or large datasets for training such models are not readily available. Reverse calibration significantly widens the applicability of our framework.

2 Related Work

2.1 Confidence Calibration

Train-time Calibration. Such techniques integrate model calibration during the training phase through suitable modification of loss function. E.g., label smoothing (LS) [47], originally proposed to improve the classifier accuracy by computing cross-entropy with a weighted sum of one-hot vector and the uniform distribution, was adopted by [37] for improving calibration. Most train-time methods for calibration inherently look to smooth confidence scores in a sample-agnostic manner [13, 20, 32, 35, 41].

Post-hoc Calibration. These techniques focus on post-training optimization using a separate hold-out set. [10], demonstrated that temperature scaling (TS), smoothing confidence scores by dividing the logits with a scalar $T > 1$, enhances its calibration. Other notable contributions in this category also include the studies by [2, 17, 24, 25, 43]. However, it was observed in [13] that train-time approaches offer superior performance over post-hoc methods.

Data Augmentation, and Bayesian Techniques for Calibration. Prominent examples of former include [49] and [12], whereas later methodologies are exemplified by [8, 28, 29, 39, 51]. In the context of our research, train-time and KD-based approaches are especially pertinent and discribed below.

2.2 Knowledge Distillation (KD) for Calibration

KD [15] was originally proposed to enhance the accuracy of student classifiers by transferring knowledge from high-capacity teacher classifiers. [40] propose explanation-based KD for enhanced accuracy. [46] identify several potential factors contributing to diminished fidelity in distillation, i.e., the student's capability to align with a teacher's predictions. These factors include student capacity, network architecture, data domain, optimization methods, among others. However, recent empirical evidence points to the regularization effects of KD over student classifiers akin to training classifiers separately via LS [48]. This indicates the potential calibration benefits of KD. It was shown in [53] that when the temperature parameter during KD is set to unity and the probability distribution of teacher classifiers are assumed to be uniform, KD via teacher classifier and LS of student classifier exhibit identical behaviors in terms of gradient propagation. The observations prompted us to explore the scenario where a teacher classifier itself is first calibrated via LS and then distills knowledge to a student classifier (i.e., distilling knowledge from an LS calibrated teacher) with the hope of

doubling the regularization benefits. However, it was observed in [37] that LS representations interfered with KD, thus nullifying any regularization benefits.

The above view was shown as incomplete [45] with an argument that such an impairment is only a high-temperature phenomenon. While this interplay between LS and KD provided some insights into the regularization benefits of KD, its role as a potential calibrator of student classifiers has not been studied so far.

In our work, we look beyond just the vanilla LS of teacher classifiers and provide direct theoretical and empirical evidence towards the benefits of working with calibrated teachers and how they distill student classifiers via KD with SOTA calibration performance. We also systematically analyze various calibration techniques for the teacher classifiers so that the resulting student classifiers exhibit significantly improved calibration performance over directly calibrating them via train-time or posthoc methods. We observe that *dynamic* LS methods, such as the MDCA [13], consistently exhibit better accuracy and calibration trade-off across wider temperature ranges.

3 Proposed Methodology

3.1 Theoretical Results

In this section, we analyze the mechanics of obtaining calibrated models via KD from a theoretical standpoint. We want to underscore that our primary objective is not the creation of a novel neural network or the formulation of a theoretical framework for KD representation learning. Rather, our aim is to gain insight into the calibration transfer behavior proposed in KD and then leverage the insights to develop a SOTA calibration framework. Hence, to keep the exposition simple, we focus on linear teacher and student networks in a binary classification problem. Such linear classifiers, which were initially explored in [42] to gain a general understanding of KD, have not been previously investigated for their potential to transfer learned representations, particularly calibration, to student networks. Furthermore, the authors in [42] utilized a simplified version of the KD loss function, which did not consider the significance of distillation weights and quadratic temperature scaling. These factors play a crucial role in showcasing the transfer of calibration to student models. It may be noted that our analysis can also be broadened to accommodate more than two classes, necessitating the substitution of the binary cross-entropy loss function with the multi-class cross-entropy loss and the replacement of the Sigmoid activation with Softmax activation. Nonetheless, for the sake of simplicity in mathematical treatment, we confine our focus to binary classification.

Definition 1 (Confidence Calibration). *Given a data distribution $\mathcal{D}$ of $(x, y) \in \mathcal{X} \times \{0, 1\}$ and let c be the predictive confidence, the predictor $f : X \to [0, 1]$ is said to be calibrated [6], if:* $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[y \mid f(x) = c\right] = c, \quad \forall c \in (0, 1)$

Notation. We represent an i^{th} training instance by $\mathbf{x}_i \in \mathbb{R}^d$, and the set of all training examples by $\mathbf{X} \in \mathbb{R}^{d\times N}$. We use $z_{i,s}$ and $z_{i,t}$ to represent logits of the

student and teacher networks for the i^{th} training instance, respectively. These logits can be converted into valid probability distributions $p_{i,s}$ and $p_{i,t}$, respectively, using Sigmoid activation function. In knowledge distillation, the output probabilities of the teacher network are softened using inverse temperature scaling of the logit, leading to prediction probabilities $\{\tilde{p}_{i,t}\}$. The true class labels are denoted by $\{y_i \in \{0,1\}\}$. Since, the teacher and student networks are assumed to be linear networks, an arbitrary deep network can equivalently be represented using a single layer network. We use $\mathbf{W}_s$ and $\mathbf{W}_t$ to represent weight matrices of the student and teacher networks, respectively. Finally, we use $T \in \mathbb{R}_+$ to depict temperature parameter for temperature scaling, while $\alpha \in [0,1)$ represents the relative importance of the student's binary cross-entropy loss. Below we list the key assumption before presenting our theoretical results.

Assumption 1 *The student and teacher networks are represented by linear networks.*

Remark 1. Assumption 1 ensures that both student and teacher networks can be compactly represented as single-layer linear networks. Though the assumption implicitly enforces the student network to be of the same capacity as that of the teacher network, as we show through our experiments, it helps us understand the mechanics of the KD, and design appropriate techniques which improve calibration even when the assumptions do not necessarily hold true. We wish to point out that the assumption in this work is the same as used in most contemporary works, including [42]. **Note:** We wish to emphasize that the incorporation of nonlinear activations makes the KD-loss function non-convex, and despite empirical visualization techniques discussed in visualizing the loss landscape of neural nets [31], there is currently no theoretical work, addressing KD for calibration in a general non-convex setting.

KD Problem Formulation. In KD, a student model minimizes the weighted combination of the binary cross-entropy ($\mathcal{L}_{\text{BCE}}$), and KD loss ($\mathcal{L}_{\text{KD}}$), given by:

$$\begin{aligned}
\mathcal{L}_{\text{BCE}} &= -\sum\nolimits_{i=1}^{N} \left[y_i \log p_{i,s} + (1-y_i) \log \left(1 - p_{i,s}\right) \right], \\
\mathcal{L}_{\text{KD}} &= -T^2 \sum\nolimits_{i=1}^{N} \left[\tilde{p}_{i,t} \log \tilde{p}_{i,s} + (1-\tilde{p}_{i,t}) \log \left(1 - \tilde{p}_{i,s}\right) \right], \\
\text{and} \quad \mathcal{L}_{\text{tot}} &= (1-\alpha)\mathcal{L}_{\text{BCE}} + \alpha \mathcal{L}_{\text{KD}}.
\end{aligned} \tag{1}$$

Here $p_{i,s} \coloneqq \sigma(\mathbf{W}_s^\top \mathbf{x}_i)$, $\sigma(\cdot)$ denotes the Sigmoid function,

$$\tilde{p}_{i,s} \coloneqq \sigma\left(\frac{\mathbf{W}_s^\top \mathbf{x}_i}{T}\right), \quad \text{and} \quad \tilde{p}_{i,t} \coloneqq \sigma\left(\frac{\mathbf{W}_t^\top \mathbf{x}_i}{T}\right).$$

Theorem 1. *Let $\mathbf{X} \in \mathbb{R}^{d \times N}$ be the data matrix, and $\mathbf{W}_s$ and $\mathbf{W}_t$ represent the parameters of the student and the teacher networks, respectively. Then, under*

Assumption 1 and using the gradient-descent algorithm, the parameters $\mathbf{W}_s$ *of the student network converge to:*

$$\mathbf{W}_s \approx \begin{cases} \alpha\mathbf{W}_t + 4(1-\alpha)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{Y}_{1/2}, & \text{if } N < d \\ \alpha\mathbf{W}_t + 4(1-\alpha)(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{Y}_{1/2}, & \text{else} \end{cases},$$

where $\mathbf{Y}_{1/2} \coloneqq \left[y_i - \frac{1}{2}\right]_{i=1}^{N}$ *is an* N*-dimensional vector.*

Please refer to supplementary material for the proof.

Remark 2. Theorem 1 shows that when α is close to unity, the weights of the student network are almost identical to those of the teacher network. Thus, properties of the teacher network transfer directly to the student. For $\alpha \neq 1$, the student also updates its weight from the labeled data.

Calibrated Teachers produce Calibrated Students. It is easy to see from Definition 1 that if a teacher network, with predicted probabilities $\{p_{i,t}\}$, is well calibrated, then the following holds:

$$\sum\nolimits_{i=1}^{N} p_{i,t} = \sum\nolimits_{i=1}^{N} y_i. \tag{2}$$

We now prove that calibrated teachers distill calibrated students. On the contrary, if the teacher classifier is not well-calibrated, it is impossible to distill well-calibrated student classifiers. The result extends our understanding of KD beyond accuracy transfer and formally characterizes the transfer of calibration from a teacher to student network.

Theorem 2. *Let Assumption 1 hold. Let* t_c *and* t_{uc} *be two teacher classifiers with output probabilities* $\{p_{i,t_c}\}$ *and* $\{p_{i,t_{uc}}\}$*, respectively. Also, let* s_c*,* s_{uc} *depict two student classifiers trained independently from the corresponding teacher classifiers* t_c *and* t_{uc} *through KD, with output probabilities* $\{p_{i,s_c}\}$ *and* $\{p_{i,s_{uc}}\}$*, respectively. If the teacher classifier* t_c *is well calibrated, then the student classifier* s_c *is also well calibrated. Conversely, if the teacher classifier* t_{uc} *is not well calibrated, the corresponding student classifier* s_{uc} *mimics a similar behavior, i.e.,*

$$\sum\nolimits_{i=1}^{N} p_{i,s_c} = \sum\nolimits_{i=1}^{N} y_i, \quad \textit{and} \quad \sum\nolimits_{i=1}^{N} p_{i,s_{uc}} \neq \sum\nolimits_{i=1}^{N} y_i.$$

Please refer to supplementary material for detailed proof.

3.2 Proposed Algorithm

Based on the theoretical results presented earlier, and the consequent discussion, we propose a two step procedure to train calibrated student models:

1. ***Train-time calibration of teachers*:** We draw inspiration from train-time calibration techniques that have shown superior performance than post-hoc calibration and have experimented with the following techniques: ([5, 13, 27, 35, 37]) to name a few. A simple gradient analysis [41] reveals that train-time calibration methods, such as MDCA ([13]) and ACLS ([41]), act as *dynamic/adaptive* label smoothing, which is arguably better than the traditional *static* label smoothing [37]

2. ***Knowledge distillation from calibrated teacher*:** Once trained for calibration and accuracy, teacher classifiers distil their behavior to student classifiers through KD loss (c.f. Eq. (1)). As a result, the student classifiers are both accurate and confidence calibrated (see Theorem 2).

The supplementary material contains a flow diagram illustrating the proposed KD(C) framework.

Remark 3. We do not strictly advocate a specific dynamic label smoothing calibration technique but rather a KD-style calibration where an expert model helps enhance the calibration performance of a student. However, based on our empirical observations, we recommend the usage of MDCA [13] and AdaFocal [9] as calibrators for teachers due to their consistent behavior across diverse tasks.

4 Experiments and Results

Evaluation Metrics:. We benchmark our framework KD(C) against other competing methods using **(a)** calibration error metrics (lower value is better), Expected calibration error (ECE) ([10]), Static Calibration Error (SCE) and Adaptive Calibration Error (ACE) [38], as well as **(b)** Top1 accuracy (higher value is better), indicative of generalization performance.

Datasets and Baselines. We use widely accepted diverse datasets, CIFAR10 [23], and CIFAR100 [23] for benchmarking. We give results for Tiny-ImageNet [30] in the supplementary. To test robustness of our approach, we report additional results on CIFAR100-C ([14]) in the supplementary. We could not experiment on ImageNet due to limited computational constraints in our lab. We include models trained through standard NLL, as well as LS [47], TS [10], MixUp [49], Adafocal [9], MMCE [26], CRL [35], CPC [5] MDCA [13], MbLS [32], PSKD [20], and ACLS [41]. Along with this, we include student distilled from an uncalibrated teacher obtained by training using NLL as one of the baselines, and refer to such student model as KD(UC).

Training details. The architectures used in the experiments include ResNet [11], MobileNetV2 [44], DenseNet [16], and WideResNet [54] architectures. The exact details of training and model hyperparameters, along with the details on compute resources are included in the supplementary material. Due to computational constraints, we used ImageNet-100 (IN100) as a proxy for ImageNet-1K performance. The PyTorch implementation for training ResNets on IN100 is available at tinyurl.com/mr3cr8rd.

4.1 Quantitative Results

Large calibrated teacher models distilling into smaller models. We now present compelling evidence supporting the superiority of our proposed KD(C) method over the SOTA train-time and post-hoc techniques for calibrating smaller student classifiers. To this end, we leverage distillation to create a smaller model

Table 1: Comparison of calibration performance on CIFAR100 dataset for student models calibrated using SOTA calibration techniques vs. proposed KD(C) framework employing a relatively larger calibrated teacher: WRN-40-2 (2.24M) on CIFAR100 dataset. We use WRN-40-1 [54] (0.56M) and MobileNetV2 [44] (2.25M) as the student models. For ECE/SCE computation, 15 bins were used by prior work. ACE uses an adaptive binning strategy. **Numbers in bold:** best performance; underlined: second best. Gray cells show the performance of the teacher model alone when trained with the respective calibrator. Similarly, Beige, and Blue cells show the performance of WRN-40-1 and MobileNetV2 models respectively, when trained directly with the respective calibrator. On the other hand, LightCyan cells show the performance when these models were trained with the proposed KD(C) framework but the teacher model was pre-trained using different calibration techniques. We observe that the proposed KD(C) framework consistently gives the best calibration performance. In terms of accuracy, PSKD manages to slightly improve KD(C), but this is consistent with the observation in the contemporary literature [13], where improvement in calibration performance often comes at a cost of a slight drop in accuracy. Note that the calibration performance (ECE) of PSKD is much worse than ours.

Architecture →	**WideResNet-40-2(T)**				**WideResNet-40-1(S)**				**MobileNetV2(S)**			
Method↓	Top1 (%) ↑	ECE (%) ↓	SCE (%) ↓	ACE (%) ↓	Top1 (%) ↑	ECE (%) ↓	SCE (%) ↓	ACE (%) ↓	Top1 (%) ↑	ECE (%) ↓	SCE (%) ↓	ACE (%) ↓
NLL	74.39	13.04	0.32	13.04	70.04	11.16	0.30	11.19	66.09	7.76	0.25	7.80
LS [47]	75.15	2.32	0.21	2.08	70.07	1.30	0.21	1.49	66.96	4.24	0.23	4.18
CE+TS [10]	74.39	2.68	0.20	2.60	70.04	2.57	**0.19**	2.50	66.09	2.33	**0.19**	2.37
MMCE [27]	72.82	5.62	0.22	5.58	69.69	7.34	0.25	7.37	62.90	3.21	0.21	3.15
MixUp [49]	76.23	4.76	0.22	4.63	72.04	2.57	0.21	2.52	67.53	8.69	0.28	9.73
CRL [35]	70.86	11.97	0.32	11.88	65.80	13.91	0.37	13.91	67.05	12.06	0.33	12.06
PSKD [20]	75.22	7.82	0.24	7.82	**72.56**	3.73	0.20	3.72	69.09	6.95	0.23	6.94
MDCA [13]	74.17	1.57	0.20	1.61	68.51	1.35	0.21	1.34	66.96	1.61	0.20	1.92
AdaFocal [9]	73.12	2.24	0.20	2.28	67.36	2.10	0.21	1.97	65.34	1.83	0.20	**1.53**
CPC [5]	74.92	10.70	0.27	10.65	69.99	7.61	0.23	7.55	67.30	4.17	0.22	4.07
MbLS [32]	74.79	7.81	0.23	7.78	69.97	5.37	0.22	5.37	67.32	2.27	0.20	2.33
ACLS [41]	74.90	6.03	0.22	6.01	69.92	7.00	0.23	6.99	66.26	3.27	0.21	3.35
KD distilled Student with NLL WRN-40-2 Teacher (KD(UC))					69.60	15.18	0.37	15.18	67.15	6.05	0.22	6.02
Ours (KD distilled Student with MixUp WRN-40-2 Teacher)					72.48	1.21	0.20	1.17	**71.97**	2.94	0.24	2.91
Ours (KD distilled Student with AdaFocal WRN-40-2 Teacher)					71.70	1.19	**0.19**	1.34	69.47	2.44	0.21	2.41
Ours (KD distilled Student with CPC WRN-40-2 Teacher)					70.00	9.02	0.26	9.01	67.78	3.64	0.20	3.66
Ours (KD distilled Student with MDCA WRN-40-2 Teacher)					71.07	**0.98**	0.20	**1.10**	68.67	**1.52**	0.20	1.64
Ours (KD distilled Student with MMCE WRN-40-2 Teacher)					72.08	2.02	**0.19**	1.95	68.49	1.83	**0.19**	1.68

(e.g., WRN-40-1/MobileNetV2) from a well-calibrated teacher model (e.g., WRN-40-2) and compare its performance with models directly subjected to train-time calibration techniques, as well as the progressive-KD (PSKD) method introduced by [20]. We also report the impact of distillation from an uncalibrated teacher model, denoted as KD(UC), as a baseline. The summarized results are detailed in Tab. 1. Notably, KD(C) demonstrates significantly lower calibration errors (ECE/SCE/ACE) while simultaneously achieving higher accuracy compared to models calibrated directly using the calibration techniques. Fig. 3 provides a visual representation of our findings, illustrating the mean and standard deviations of accuracy and calibration errors over three random runs. Notably, KD(C) variants exhibit (a) the best balance between accuracy and calibration while (b) displaying higher reliability, as evidenced by their lower variance. Importantly, our results confirm our theoretical findings discussed in Sec. 3.1, establishing that calibrated teachers are capable of effectively distilling calibrated students. This underscores the successful transfer of learned representations, encompass-

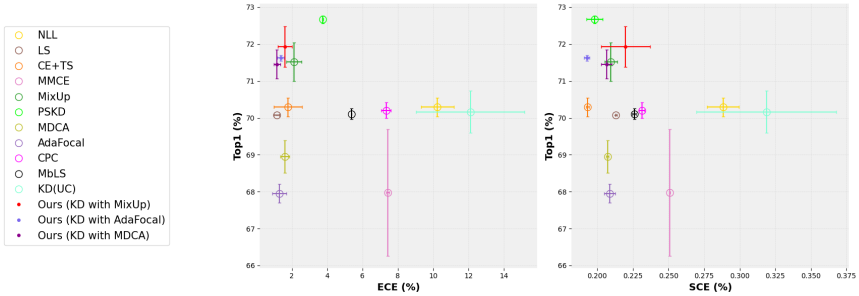

Fig. 3: Comparative study of accuracy vs. calibration trade-offs: (Models in top-left location are best) The mean and one standard scatter error bars for `Top1`, `ECE` and `SCE` of `WideResNet-40-1` trained on `CIFAR100` using `SOTA` calibration techniques. `WideResNet-40-2` was used as a Teacher. `KD(C)` variants (represented by small filled dots) achieve the best results in terms of `ECE`, and `SCE`, along with slight boosts in `Top1` (an inherent `KD`-property). Further, the lower variances emphasize the reliability of `KD(C)` variants. All plots were generated by training `WideResNet-40-1` models through every calibration technique on 3 runs.

ing both accuracy and calibration aspects, from a calibrated teacher model to a smaller student. In supplementary, we give additional results in showing our approach consistently yields improved calibration across various model architectures and reliability diagrams corresponding to Tab. 1.

Small calibrated teacher models distilling into large models. In settings where large trained models are not available, it is desirable to be able to distill the knowledge from smaller models to larger models (referred to as `reverse-KD`). [19] have shown that smaller models can also be valid teachers for large students, however, it was observed that the gains in accuracy were not significant as compared to distilling from a large teacher comparatively. Our results for the configuration are summarized in Tab. 2**(a)**, where smaller model `MobileNetV2` is used as teacher network to calibrate `ResNet-50`. We report a similar behavior in terms of accuracy, where there is marginal to no improvement in the accuracy. However, even with smaller teacher, we notice a significant improvement in calibration performance of bigger student model using proposed `KD(C)` framework. This shows the potential impact of the proposed technique. We discuss our thoughts on the reasons why calibration works differently than accuracy in Sec. 4.3.

Self-distillation. A significant question that arises pertains to the generalizability of insights gleaned from the prior set of experiments. Particularly whether these insights can be extended to produce accurately calibrated classifiers with identical architecture and capacity. Our research demonstrates that this process, referred to as "self-distillation", results in classifiers that exhibit superior calibration compared to their teachers. However, similar to `reverse-KD`, the increase in accuracy is only marginal, likely due to the absence of distillation from a teacher with greater capacity. Similarly, calibration improvement also follows

Table 2: Results on CIFAR-10 dataset for **(a)** reverse-KD from a small model (MobilNetV2 with 2.25M weights) to a big model (ResNet-50 with 23.53M weights); and **(b)** Self-Distillation on MobileNetV2. Even when the accuracy stays more or less the same, we observe significant improvement in the calibration performance using the proposed KD(C) framework. Numbers in **bold**: best, underlined: second best.

Calibration Techniques	(a) Small to Big				(b) Self-Distillation			
	ResNet-50 (T=MobileNetV2)				MobileNetV2 (T=MobileNetV2)			
	Top1 (%) ↑	ECE (%) ↓	SCE (%) ↓	ACE (%) ↓	Top1 (%) ↑	ECE (%) ↓	SCE (%) ↓	ACE (%) ↓
NLL	88.55	6.65	1.39	6.65	89.87	3.30	0.75	3.28
LS [47]	87.73	6.17	1.36	7.69	89.60	7.10	1.78	6.75
CE+TS [10]	88.55	1.91	0.65	2.18	89.90	0.96	**0.40**	0.77
MMCE [26]	87.73	2.74	0.59	2.58	89.38	1.20	0.51	0.94
MixUp [49]	88.49	5.95	1.63	5.93	89.57	9.42	2.07	9.41
CRL [35]	84.04	10.30	2.14	10.29	**90.31**	2.92	0.72	2.81
PSKD [20]	88.01	2.13	0.84	1.74	89.21	3.27	0.93	3.25
MDCA [13]	87.16	1.18	0.75	1.27	88.74	0.99	0.46	0.80
AdaFocal [9]	85.07	1.08	0.63	1.17	88.98	0.79	0.44	0.86
CPC [5]	88.30	7.13	1.47	7.12	89.26	3.47	0.79	3.44
MbLS [32]	88.19	6.94	1.43	6.90	89.86	2.83	0.69	2.78
ACLS [41]	87.97	5.83	1.22	5.86	89.53	1.80	0.56	2.66
KD with NLL	**89.01**	4.24	0.91	4.22	89.88	0.99	0.43	0.82
Ours(KD with TS)	88.91	1.07	**0.44**	1.26	90.23	0.51	0.41	0.59
Ours(KD with MMCE)	88.59	1.02	0.52	0.90	89.97	0.85	0.54	0.84
Ours(KD with MDCA)	88.12	**0.66**	0.45	0.77	88.79	**0.48**	0.48	**0.54**
Ours(KD with AdaFocal)	88.71	0.68	0.49	**0.70**	89.56	0.63	0.41	0.65

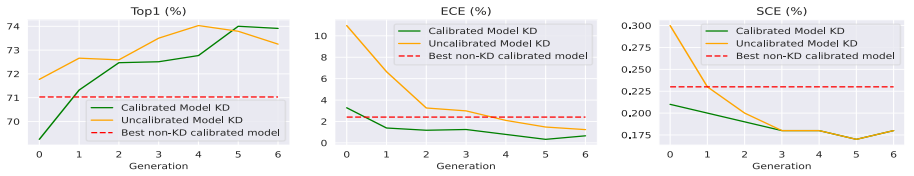

Fig. 4: Iterative self distillation CIFAR100 using ResNet56. We use KD with dynamic label smoothing technique (MDCA) for calibration in this experiment.

a similar trend, with the student model showing significant calibration performance improvement using proposed framework, compared to training directly with a particular calibration technique. Our findings on the CIFAR-10 dataset are succinctly presented in Tab. 2**(b)**. It is worth noting that, unlike the PSKD approach proposed by [20], which progressively distills knowledge from the previous epoch's model, KD(C) employs self-distillation just once with a fixed teacher throughout the training process, following a methodology akin to [55].

Iterative self-distillation. We are intrigued by the experiments in [20, 34, 52], and perform a similar investigation if KD(C) can iteratively distill more accurate and calibrated models. Here both teacher and student have identical architectures, and a student in t^{th} iteration (called *generation* hereon) becomes teacher for $(t+1)^{th}$ generation. We refer to this as iterative self-distillation. Fig. 4 shows the self-distillation process for six generations. As expected the gap between KD(UC) and KD(C) gradually diminishes with each generation. Recall that the only difference between the two is initialization: generation zero teacher is uncal-

ibrated in KD(UC) but calibrated in KD(C). Our observation aligns with findings from [20, 55].

Results on NLP datasets: Table 3 demonstrates KD(C)'s effectiveness on NLP tasks, showing results for DistilBERT (66.97M) on 20-Newsgroups sequence classification, with BERT (109.50M) as teacher. KD(C) variants outperform direct calibrators in this transformer distillation task.

Other results included in the supplementary material. We report **(a)** calibration performance under dataset drift; **(b)** ablation study on the effect of hyper-parameters like T (temperature) and α (distillation weight) in the supplementary along with experiments involving other DNN architectures.

Calibration	Top1 (%)	ECE (%)	AECE (%)
NLL	91.75	2.70	2.85
LS	<u>92.02</u>	5.87	5.60
MMCE	89.18	2.37	2.83
PSKD	**92.12**	1.84	2.33
MDCA	91.22	2.93	2.67
AdaFocal	90.93	5.60	5.33
CPC	91.62	3.23	2.98
Ours (KD with TS)	91.06	**1.41**	<u>1.86</u>
Ours (KD with MDCA)	90.19	1.90	2.16
Ours (KD with LS)	90.88	5.33	4.69
Ours (KD with CPC)	90.64	<u>1.69</u>	**1.62**

Table 3: Results for DistilBERT (66.97M) for Seq. Classification on 20-Newsgroups dataset. For KD+UC and KD+C variants BERT (109.50M) was used as Teacher. Observe that KD(C) variants achieve competitive calibration of DistilBERT model.

4.2 Visualization

Unlike traditional LS, dynamic/adaptive regularization [13, 41] offers sample-specific label-smoothing. However, not many methods are able to capture inter-class semantics during the train-time, which become easier to guide in the proposed KD(C) framework through the interplay of teacher-based knowledge transfer and learning directly from data. Inspired from [4], we give the additional rationale for the superior performance of the proposed KD(C) framework using "penultimate layer visualizations".

Penultimate layer visualization. [37] visualized penultimate activations by projecting them onto a hyperplane defined by weights of three selected classes.

Systematic diffusion. The concept of "systematic diffusion", introduced by [4], was developed to address discrepancies observed in prior studies, particularly the contradictions between [45] and the insights presented in LS literature [37]. This concept aims to elucidate the compatibility of LS with KD. The findings from [4] indicate that when KD is conducted at elevated temperatures from a teacher model trained with LS, it results in a systematic shift in the relationships between classes. Specifically, for semantically similar classes, the inter-cluster distance decreases, while for the remaining classes, it increases relatively.

Our observations. In Fig. 5, we provide visual evidence of the limitations associated with LS-trained teachers compared to MDCA teachers [13] used in conjunction with proposed KD(C) framework. These penultimate layer visualizations

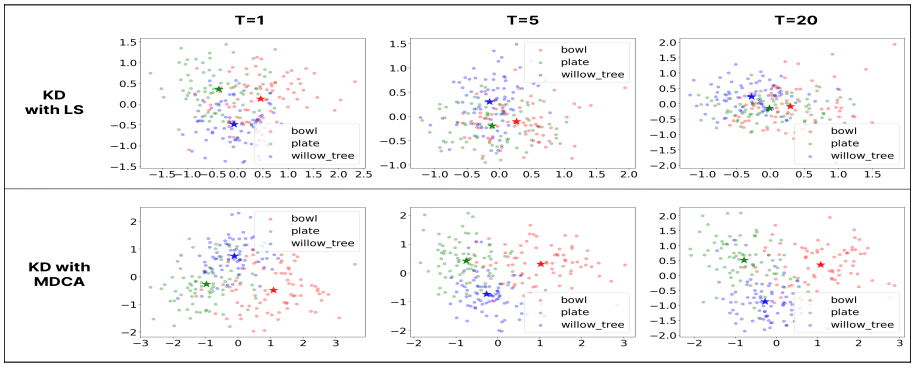

Fig. 5: Visualization of penultimate layer's activations. We train ResNet8 on CIFAR100 using ResNet56 as teacher. First row shows results for teacher calibrated using LS (static label smoothing), and the second row shows results for MDCA (dynamic label smoothing). We follow the same setup and procedure used in [37,45], with two semantically similar classes (bowl, plate) and one semantically dissimilar class (willow_tree). A '*' in the plot represents cluster's centroid. A well-calibrated teacher, with dynamic label smoothing based MDCA [13], can effectively capture the inter-class relationships. Observe that the classes: bowl and plate are visually similar and hence the penultimate visualizations of these classes should be closer than the dissimilar class: willow_tree. As temperature T is increased the similar classes diffuse into one for the case of KD with LS while KD with MDCA offers better separation, retaining the semantic similarity while well separated from the dissimilar class.

reveal that semantically similar classes experience systematic diffusion when using LS, whereas this phenomenon is not observed with KD when dynamic smoothing regularization based MDCA is used. We notice a trend where distilled student models are most calibrated when the distillation temperature (T) is set to 1. We hypothesize that increasing T leads to the destruction of discriminating features, as outlined by [4], due to systematic diffusion among highly similar classes as seen in the penultimate representations. These discriminating features are crucial for achieving calibration by resolving confusion among similar classes.

4.3 Discussion

We use this section to give our perspective on experimental observations:

Why post-hoc calibration works? Mukhoti et al. [36] identified overfitting to NLL loss as a major cause of overconfident DNN predictions. This overconfidence arises from one-hot encoding's inability to capture subtle inter-class similarities, such as between similar cat and dog images. Lacking sample-specific uncertainty, one-hot encoding leads to overconfidence even in overfitted models. Post-hoc calibration methods like TS address this by adjusting output probabilities through inverse temperature scaling.

Why Train-time calibration works better than post-hoc calibration? Train-time calibration techniques refine target probability vectors for models. For instance, LS assigns a uniform probability vector as the target, while MDCA uses batch-wise label frequency alongside the standard one-hot label vector.

Why KD based calibration would work best? Train-time calibration techniques lack sample-specific, adaptive uncertainty during model training. KD addresses this gap. When a student model is trained via KD from a calibrated teacher, it learns to align its probability vector with the teacher's, incorporating crucial sample-specific, adaptive uncertainty. Our findings show that the effectiveness of calibration techniques depends on the degree of sample-wise, adaptive uncertainty provided to the student during training, along with accuracy.

Why LS interferes with KD? Penultimate visualizations explain why LS can interfere with KD and give lower accuracy. LS tends to smooth the output probability vector, thereby reducing class structure information in the resultant vector [4]. Since KD relies upon the exact class structure in a teacher's probability vector (or logits), excessive LS ends up killing the information needed by KD.

Why Self-distillation is not as effective? Kim *et al.* [20] propose a self-distillation technique (called PSKD), which we showed in our experiments is not as effective as ours. Our theoretical results also indicate why this should be the case. Until the model is fully calibrated, by nudging the student towards the previous epoch's probability vector as a teacher, one ends up giving a rich, but uncalibrated probability vector to a student model. In contrast, in our proposed design, one trains a teacher model fully, and then only transfers this calibrated class representation to a student to match. This explains the reason why despite PSKD also using distillation as us, is still less effective for calibration.

5 Conclusions

We present a novel calibration technique using KD, validated across diverse scenarios including large-to-small, small-to-large, and self-distillation settings with various architectures and datasets. Our work clarifies misconceptions about using calibrated teachers with KD, demonstrating that most modern calibration techniques, particularly MDCA and ACLS, can be effectively combined with KD to produce SOTA calibrated student models. This approach is supported by a robust theoretical foundation for transferring calibration and accuracy between teacher and student DNNs. Our findings are particularly relevant as KD becomes increasingly important for developing lightweight, trustworthy models for edge computing, neural architecture search, and model compression.

Limitations and Future Work. (1) We leave a comprehensive and exhaustive study for explaining the compatibility of calibration techniques, such as CPC [5], TS [10], ACLS [41] etc., with KD for future work. **(2)** We acknowledge the need to broaden our theoretical results and encompass one hidden-layer DNNs with ReLU activation in our future work, since for such networks in a non-KD setting, the cross-entropy loss function is found to be locally strongly convex [50].

6 Acknowledgement

We thank Jatin Prakash and Neelabh Madan for initial experiments, and Lovekesh Vig, Ashwin Srinivasan and Gautam Shroff for insightful technical discussions.

References

1. Allen-Zhu, Z., Li, Y.: Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=Uuf2q9TfXGA
2. Bohdal, O., Yang, Y., Hospedales, T.: Meta-calibration: Meta-learning of model calibration using differentiable expected calibration error. In: ICML Uncertainity in Deep Learning Workshop (2021)
3. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
4. Chandrasegaran, K., Tran, N.T., Zhao, Y., Cheung, N.M.: Revisiting label smoothing and knowledge distillation compatibility: What was missing? In: International Conference on Machine Learning. pp. 2890–2916. PMLR (2022)
5. Cheng, J., Vasconcelos, N.: Calibrating deep neural networks by pairwise constraints. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13699–13708 (2022). https://doi.org/10.1109/CVPR52688.2022.01334
6. Dawid, A.P.: The well-calibrated bayesian. Journal of the American Statistical Association (1982)
7. Dusenberry, M.W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K., Dai, A.M.: Analyzing the role of model uncertainty for electronic health records. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 204–213 (2020)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
9. Ghosh, A., Schaaf, T., Gormley, M.: Adafocal: Calibration-aware adaptive focal loss. In: Advances in Neural Information Processing Systems. vol. 35, pp. 1583–1595 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/0a692a24dbc744fca340b9ba33bc6522-Paper-Conference.pdf
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. CoRR **abs/1706.04599** (2017), http://arxiv.org/abs/1706.04599
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hebbalaguppe, R., Ghosal, S.S., Prakash, J., Khadilkar, H., Arora, C.: A novel data augmentation technique for out-of-distribution sample detection using compounded corruptions. European Conference of Machine Learning (2022)
13. Hebbalaguppe, R., Prakash, J., Madan, N., Arora, C.: A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16081–16090 (June 2022)

14. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2018)
15. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015), http://arxiv.org/abs/1503.02531
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
17. Islam, M., Seenivasan, L., Ren, H., Glocker, B.: Class-distribution-aware calibration for long-tailed visual recognition. ICML Uncertainity in Deep Learning Workshop (2021)
18. Jiang, X., Deng, X.: Knowledge reverse distillation based confidence calibration for deep neural networks. Neural Processing Letters **55**(1), 345–360 (2023)
19. Jiang, X., Deng, X.: Knowledge reverse distillation based confidence calibration for deep neural networks. Neural Processing Letters **55**(1), 345–360 (2023)
20. Kim, K., Ji, B., Yoon, D., Hwang, S.: Self-knowledge distillation with progressive refinement of targets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6567–6576 (October 2021)
21. Kononenko, I.: Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine **23**(1), 89–109 (2001)
22. Kristiadi, A., Hein, M., Hennig, P.: Being bayesian, even just a bit, fixes overconfidence in relu networks. In: International conference on machine learning. pp. 5436–5446. PMLR (2020)
23. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tiny ImageNet Dataset (2009)
24. Kull, M., Perello-Nieto, M., Kängsepp, M., Song, H., Flach, P., et al.: Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. arXiv preprint arXiv:1910.12656 (2019)
25. Kull, M., Silva Filho, T., Flach, P.: Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In: Artificial Intelligence and Statistics. pp. 623–631. PMLR (2017)
26. Kumar, A., Liang, P., Ma, T.: Verified uncertainty calibration. arXiv preprint arXiv:1909.10155 (2019)
27. Kumar, A., Sarawagi, S., Jain, U.: Trainable calibration measures for neural networks from kernel mean embeddings. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2805–2814. PMLR (10–15 Jul 2018), http://proceedings.mlr.press/v80/kumar18a.html
28. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles (2017)
29. Laves, M.H., Ihler, S., Kortmann, K.P., Ortmaier, T.: Well-calibrated model uncertainty with temperature scaling for dropout variational inference. arXiv preprint arXiv:1909.13550 (2019)
30. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
31. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. Advances in neural information processing systems **31** (2018)
32. Liu, B., Ben Ayed, I., Galdran, A., Dolz, J.: The devil is in the margin: Margin-based label smoothing for network calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 80–88 (2022)

33. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics **19**(6), 1236–1246 (2018)
34. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: Proceedings of the AAAI conference on artificial intelligence. pp. 5191–5198 (2020)
35. Moon, J., Kim, J., Shin, Y., Hwang, S.: Confidence-aware learning for deep neural networks. In: international conference on machine learning. pp. 7034–7044. PMLR (2020)
36. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, F., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems **33**, 15288–15299 (2020)
37. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Advances in neural information processing systems **32** (2019)
38. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR Workshops (2019)
39. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift (2019)
40. Parchami-Araghi, A., Böhle, M., Rao, S., Schiele, B.: Good teachers explain: Explanation-enhanced knowledge distillation (2024)
41. Park, H., Noh, J., Oh, Y., Baek, D., Ham, B.: Acls: Adaptive and conditional label smoothing for network calibration. In: Proceedings of the IEEE/CVF ICCV (2023)
42. Phuong, M., Lampert, C.: Towards understanding knowledge distillation. In: International Conference on Machine Learning. pp. 5142–5151. PMLR (2019)
43. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
44. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
45. Shen, Z., Liu, Z., Xu, D., Chen, Z., Cheng, K.T., Savvides, M.: Is label smoothing truly incompatible with knowledge distillation: An empirical study. arXiv preprint arXiv:2104.00676 (2021)
46. Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does knowledge distillation really work? Advances in Neural Information Processing Systems **34**, 6906–6919 (2021)
47. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
48. Tang, J., Shivanna, R., Zhao, Z., Lin, D., Singh, A., Chi, E.H., Jain, S.: Understanding and improving knowledge distillation. arXiv preprint arXiv:2002.03532 (2020)
49. Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In: Advances in neural information processing systems (2019)
50. Wang, Y., Lacotte, J., Pilanci, M.: The hidden convex optimization landscape of regularized two-layer relu networks: an exact characterization of optimal solutions. In: International Conference on Learning Representations (2021)

51. Wenzel, F., Snoek, J., Tran, D., Jenatton, R.: Hyperparameter ensembles for robustness and uncertainty quantification. NeurIPS (2020), https://arxiv.org/abs/2006.13570
52. Yalburgi, S., Dash, T., Hebbalaguppe, R., Hegde, S., Srinivasan, A.: An empirical study of iterative knowledge distillation for neural network compression. In: ESANN. pp. 217–222 (2020)
53. Yuan, L., Tay, F.E.H., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: CVPR (2021)
54. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
55. Zhang, Z., Sabuncu, M.: Self-distillation as instance-specific label smoothing. Advances in Neural Information Processing Systems **33**, 2184–2195 (2020)