Improving Uncertainty Calibration of Deep Neural Networks via Truth Discovery and Geometric Optimization

Chunwei Ma[1] **Ziyun Huang**[2] **Jiayi Xian**[1] **Mingchen Gao**[*1] **Jinhui Xu**[*1]

[1]Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA
[2]Computer Science and Software Engineering, Penn State Erie, Erie, PA, USA

Abstract

Deep Neural Networks (DNNs), despite their tremendous success in recent years, could still cast doubts on their predictions due to the intrinsic uncertainty associated with their learning process. Ensemble techniques and post-hoc calibrations are two types of approaches that have individually shown promise in improving the uncertainty calibration of DNNs. However, the synergistic effect of the two types of methods has not been well explored. In this paper, we propose a truth discovery framework to integrate ensemble-based and post-hoc calibration methods. Using the geometric variance of the ensemble candidates as a good indicator for sample uncertainty, we design an accuracy-preserving truth estimator with provably no accuracy drop. Furthermore, we show that post-hoc calibration can also be enhanced by truth discovery-regularized optimization. On large-scale datasets including CIFAR and ImageNet, our method shows consistent improvement against state-of-the-art calibration approaches on both histogram-based and kernel density-based evaluation metrics. Our code is available at https://github.com/horsepurve/truly-uncertain.

1 INTRODUCTION

We live in an uncertain world. With the increasing use of deep learning in the real world, quantitative estimation of the predictions from deep neural networks (DNNs) must not be neglected, especially when it comes to medical imaging [Esteva et al., 2017] [Ma et al., 2019], disease diagnosis [De Fauw et al., 2018] [Ma et al., 2018], and autonomous driving [Kendall et al., 2017]. Uncertainty also plays an important role in differentially private data analysis [Bassily et al., 2013].

Modern deep neural networks, despite their extraordinary performance, are oft-criticized as being poorly calibrated and prone to be overconfident, thus leading to unsatisfied uncertainty estimation. The process of adapting deep learning's output to be consistent with the actual probability is called *uncertainty calibration* [Guo et al., 2017], and has drawn a growing attention in recent years.

For a better calibration of the uncertainty of DNNs, the efforts to date have been concentrated on developing more effective calibration and evaluation methods. Existing calibration methods roughly fall into two categories, depending on whether an additional hold-out calibration dataset is used. (1) Post-hoc calibration methods use a calibration dataset to learn a parameterized transformation that maps from classifiers' raw outputs to their expected probabilities. Quite a few techniques in this category can be used to learn the mapping, such as Temperature Scaling (TS) [Guo et al., 2017] [Kull et al., 2019], Ensemble Temperature Scaling (ETS) [Zhang et al., 2020], and cubic spline [Gupta et al., 2021], etc. However, the expressivity of the learnable mapping could still be limited in all of them. This is evidenced by the fact that in TS a single temperature parameter T is tuned, while ETS brings in three additional ensemble parameters. Thus, it is desirable to explore a more sophisticated form of the mapping function. (2) Another line of methods adapt the training process so that the predictions are better calibrated. Techniques in this category include mixup training [Thulasidasan et al., 2019], pre-training [Hendrycks et al., 2019a], label-smoothing [Müller et al., 2019], data augmentation [Ashukha et al., 2020], self-supervised learning [Hendrycks et al., 2019b], Bayesian approximation [Gal and Ghahramani, 2016] [Gal et al., 2017], and Deep Ensemble (DE) [Lakshminarayanan et al., 2017] with its variants (Snapshot Ensemble [Huang et al., 2017a], Fast Geometric Ensembling (FGE) [Garipov et al., 2018], SWA-Gaussian (SWAG) [Maddox et al., 2019]). Methods from these two categories thrive in recent years, and a natural idea is to combine them together. Recently, Ashukha et al. [2020] points out the

*Co-corresponding authors.

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

necessity of TS when DE is miscalibrated out-of-the-box. However, due to the intrinsic uncertainty and stochasticity of the learning process, it is possible that the ensemble members are not created equal, and simply averaging over all members may lead to a suboptimal ensemble result. Furthermore, when TS is appended after DE, the original ensemble members are all neglected. As a matter of fact, we still lack a method that can bridge the divide and make the most of the ensemble members.

A plethora of metrics have been developed during the past few years for the evaluation of calibration performance, such as Expected Calibration Error (ECE) [Naeini et al., 2015] and kernel density estimation-based ECE (ECE^{KDE}) [Zhang et al., 2020], Log-Likelihood (LL) [Guo et al., 2017] and calibrated LL [Ashukha et al., 2020], Kolmogorov-Smirnov (KS) test [Gupta et al., 2021], etc. Every such metrics has its strengths and weaknesses. For example, ECE could be easily biased by the binning scheme; LLs have also shown themselves to be closely correlated with accuracy ($\rho \sim 0.9$) [Ashukha et al., 2020]. Moreover, current post-hoc calibration methods usually restrict themselves to only one specific metric (e.g. LL), and it is believed that some metrics (e.g. ECE) can hardly be optimized directly [Ashukha et al., 2020]. Thus, there is a crucial need for an optimization framework that allows multiple metrics, including ECEs, to be considered at the same time.

To address these challenges, we propose in this paper a truth discovery-based framework and an accompanying geometric optimization strategy that is (a) more expressive, (b) metric-agnostic, and (c) beneficial to both ensemble-based and post-hoc calibration of DNNs.

Truth discovery, concerning about finding the most trustworthy information from a number of unreliable sources, is a well-established technique in data mining and theoretic computer science, with firm theoretic foundation [Ding and Xu, 2020, Huang et al., 2019, Li et al., 2020]. It finds applications in resolving disagreements from possibly conflicting information sources e.g., crowdsourcing aggregation. In this paper, we intend to answer the following question: *Can truth discovery be used to aggregate information from Deep Ensemble, and in turn to help uncertainty calibration?* This is conceivable because the perturbation within the opinions made by multiple classifiers may reflect the intrinsic uncertainty level of data: if an unanimity of opinion is reached by all classifiers, then the uncertainty level should be relatively low. More importantly, this unanimity has nothing to do with whether the opinions, i.e. predictions, are correct or not. Since the collections of classifiers may provide orthogonal information beyond a single classifier itself, we expect that this information could be unearthed via truth discovery.

Accordingly, in this paper, we propose truth discovery as an ideal tool for improving uncertainty calibration of deep learning, and make several contributions as follows:

1. We propose Truth Discovery Ensemble (TDE) that improves Deep Ensemble, and show that model uncertainty can be easily derived from truth discovery.
2. Considering that uncertainty calibration approaches may potentially cause a diminished accuracy, we further develop a provably accuracy-preserving Truth Discovery Ensemble (aTDE) via geometric optimization.
3. We propose an optimization approach that directly minimizes ECEs, works for both histogram-based and KDE-based metrics, and integrates multiple metrics via compositional training.
4. We further incorporate the discovered information (i.e. Entropy based Geometric Variance) into the post-hoc calibration pipeline (pTDE) and elevate the performance to a higher level.

To summarize, we show how truth discovery can benefit both ensemble-based and post-hoc uncertainty calibrations, and validate our proposed methods via experiments upon large-scale datasets, using both binning-based and binning-free metrics, along with comprehensive ablation studies.

2 PRELIMINARIES OF UNCERTAINTY CALIBRATION

For an arbitrary multi-class classifier (not necessarily neural network) $f_\theta : \mathcal{D} \subseteq \mathbb{R}^d \to \mathcal{Z} \subseteq \triangle^L$ that can make L predictions for L classes, its outputs (in any scale) can be transformed into a "probability vector" $\mathbf{z} \in \mathcal{Z}$ such that:

$$\sum_{l=1}^{L} z_l = 1, 0 \leq z_l \leq 1. \tag{1}$$

This can be done by the softmax function, which usually tails the last layer of a deep neural network. Here, $\triangle^L$ is the probability simplex in L dimensional space. Note that the classifier parameters can also be drawn from a distribution $\theta \sim q(\theta)$, e.g., ResNets with random initialization being the only difference.

Although $\mathbf{z}$ is in the probability simplex $\triangle^L$, its components may not necessarily have anything to do with, but sometimes are misinterpreted as, the probability of each class. Similarly, the maximum value of the L outputs, $\max_l z_l$, was used to represent the "confidence" that the classifier has on its prediction. To avoid possible misleading, $\max_l z_l$ is referred to as *winning score* v (i.e., $v = \max_l z_l$) [Thulasidasan et al., 2019] hereinafter.

For both ensemble-based and post-hoc calibration methods, the model is trained based on a set of N_t training samples $(\mathbf{x}^{(i)}, y^{(i)})_{i=1}^{N_t}, \mathbf{x}^{(i)} \in \mathcal{D}, y^{(i)} \in \{1, ..., L\}$. Let random variables X, Y represent input data and label, respectively. Then, another random variable $Z = f_\theta(X)$ stands for the probability vector. If z_l indeed represents the actual probability of class l (which usually not), then, the following should hold:

$$P(Y = l|Z = \mathbf{z}) = z_l. \tag{2}$$

At this time, we also call the classifier f_θ to be perfectly calibrated. It is well known that the probabilities $P(Y = l|Z = \mathbf{z})$ are hard to evaluate, since there is no ground-truth for the probability of an input $\mathbf{x}^{(i)}$ being misclassified as class $l \neq y^{(i)}$. In this paper, we focus on a variant of Eq. (2), which only measures the probability of the sample being correctly classified:

$$P(Y = y^{(i)}|Z = \mathbf{z}^{(i)}) = v^{(i)} \tag{3}$$

where $v^{(i)}$ is the winning score which is also the only value taken into consideration when evaluate top-1 accuracy (ACC).

Ensemble of Deep Neural Networks. Although it is computationally demanding, ensemble (Deep Ensemble [Lakshminarayanan et al., 2017], Snapshot Ensemble [Huang et al., 2017a], etc.) remains as a popular approach for uncertainty calibration. Formally, for a classifier f_θ with parameter distribution $q(\theta)$, the prediction of sample $\mathbf{x}^{(i)}$ is given by:

$$\mathbf{z}_{ens}^{(i)} = \int f_\theta(\mathbf{x}^{(i)})q(\theta)d\theta \tag{4}$$

which can be approximated by S independently trained classifiers as $\frac{1}{S}\sum_{s=1}^{S} f_{\theta^{(s)}}(\mathbf{x}^{(\mathbf{i})}), \theta^{(s)} \sim q(\theta)$. The S classifiers can be obtained either by independent random initialization (Deep Ensemble) or periodically convergence into local minimum via learning rate decay (Snapshot Ensemble). Since each $\mathbf{z}_\theta^{(i)} \in \triangle^L$, $\mathbf{z}_{ens}^{(i)}$ is also on the probability simplex.

Post-hoc Calibration of Deep Neural Networks. Until now, we have not yet introduced the concept of confidence, since all non-post-hoc approaches take the winning score as a representation of the confidence, based on which the calibration error is subsequently measured. In post-hoc calibration, on the other hand, a set of hold-out calibration samples $(\mathbf{x}^{(i)}, y^{(i)})_{i=1}^{N_c}$ is required to learn a mapping from $\mathbf{z}$ to another probability vector $\pi = \mathcal{T}(\mathbf{z})$ with a learnable function $\mathcal{T} : \triangle^L \to \triangle^L$, and $\max_l \pi_l$ is referred to as *confidence* w in this paper, i.e. $w = \max_l \pi_l$. Note that $\arg\max_l \pi_l$ is not necessarily equal to $\arg\max_l z_l$, and at this time the calibration may potentially decrease the accuracy, if the latter is the correct class. With this, our goal Eq. (3) now becomes $P(Y = y^{(i)}|Z = \mathbf{z}^{(i)}) = w^{(i)}$. Since we tackle both post-hoc and non-post-hoc calibrations in this paper, we distinguish *winning score* and *confidence* explicitly in that the former directly comes from the classifier f_θ while the latter is derived from a mapping deliberately learned for confidence modeling.

Calibration Error Evaluation. For the evaluation of a calibration algorithm, the calibration function $\mathcal{T}$ is applied on another evaluation dataset $(\mathbf{x}^{(i)}, y^{(i)})_{i=1}^{N_e}$ of size N_e which has no overlapping with neither the training nor the calibration datasets. In histogram-based evaluation metric, the N_e samples are split into B predefined bins. Formally, we define B pairs of endpoints $\{(\mu_b, \nu_b)\}_{b=1}^{B}, \nu_b = \mu_{b+1}$, and B point sets $\{P_b\}_{b=1}^{B} : P_b \subseteq \{w_i\}_{i=1}^{N_e}$ such that $\mu_b \leq w < \nu_b, \forall w \in P_b$. Then, the Expected Calibration Error (ECE) is defined as:

$$ECE(f_\theta) = \sum_{b=1}^{B} \frac{|P_b|}{K} |ACC(P_b) - conf(P_b)|, \tag{5}$$

which measures the empirical deviation of the sample accuracy in the b^{th} bin: $ACC(P_b) = \frac{1}{|P_b|}\sum_{j=1}^{|P_b|} \mathbf{1}(\arg\max_l z_l^{(j)} = y^{(j)})$ and the average confidence in it: $conf(P_b) = \frac{1}{|P_b|}\sum_{j=1}^{|P_b|} w^{(j)}$. The indicator function $\mathbf{1} : \mathcal{B} \to \{0, 1\}$ returns 1 if the Boolean expression is true and otherwise 0.

The ECE metric can be easily affected by the number of bins B and the positions of the endpoints. Without the use of binning, ECE^{KDE} estimates the calibration error by a kernel function $K : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with bandwidth $h > 0$. Based on Bayesian rule, ECE^{KDE} is given as:

$$ECE^{KDE}(f_\theta) = \int |w - \widetilde{ACC}(w)|\tilde{P}(w)dw \tag{6}$$

in which $\widetilde{ACC}(w)$ is the expected accuracy if the sample confidence is w. $\widetilde{P}(w)$ and $\widetilde{ACC}(w)$ are determined by kernel density estimation as:

$$\tilde{P}(w) = \frac{h^{-1}}{N_e}\sum_{i=1}^{N_e} K_h(w - w^{(i)}),$$

$$\widetilde{ACC}(w) = \frac{\sum_{i=1}^{N_e} \mathbf{1}(\arg\max_l z_l^{(i)} = y^{(i)})K_h(w - w^{(i)})}{\sum_{i=1}^{N_e} K_h(w - w^{(i)})}.$$

3 TRUTH DISCOVERY ENSEBMLE

Existing ensemble techniques in deep learning take the average of predictions made by multiple classifiers derived either from random initialization (Deep Ensemble [Lakshminarayanan et al., 2017]), periodically learning rate decay (Snapshot Ensemble [Huang et al., 2017a]), or from connected optima on the loss functions (Fast Geometric Ensembling [Garipov et al., 2018]), but scarcely utilize the sample level variance among the members of an ensemble. To make use of such information, we first introduce a vanilla truth discovery algorithm in Deep Ensemble context, and then extend it to one with accuracy-preserving guarantee.

3.1 TRUTH DISCOVERY WITHIN PROBABILITY SIMPLEX

To be consistent with truth discovery literature, we use *sources* to denote the S independently trained models. For every sample $(\mathbf{x}^{(i)}, y^{(i)})$ in the evaluation dataset, S independent predictions $\mathbf{z}^{(i,s)} = f_{\theta^{(s)}}(\mathbf{x}^{(i)})$ are made from all S sources (denoted by $\mathbf{z}_s$ hereinafter for brevity). Since the classifiers were trained with stochastic gradient descent (SGD), they may make wrong decisions on every $\mathbf{x}^{(i)}$. To model such a behavior, we assign a *reliability* value ω_s to each classifier $f_{\theta^{(s)}}$.

Definition 3.1 (Truth discovery [Li et al., 2014])**.** Given the set of points $\{\mathbf{z}_s\}_{s=1}^S \subseteq \triangle^L$ from S classifiers, truth discovery aims at finding the truth probability vector $\mathbf{z}^* \in \triangle^L$ and meanwhile the reliability ω_s for the s^{th} classifier, such that the following objective function is minimized:

$$\begin{aligned} \underset{\mathbf{z}^*, \{\omega_s\}_{s=1}^S}{\text{minimize}} \quad & \sum_{s=1}^{S} \omega_s ||\mathbf{z}^* - \mathbf{z}_s||^2 \\ \text{s.t.} \quad & \sum_{s=1}^{S} e^{-\omega_s} = 1. \end{aligned} \tag{7}$$

With this definition, interestingly, we can show a direct relationship between truth discovery and model uncertainty. We additionally define *uncertainty of source*, υ_s, as the opposite of source reliability, i.e. $\upsilon_s = e^{-\omega_s}$. Then, (7) can be written as:

$$\begin{aligned} \underset{\mathbf{z}^*, \{\upsilon_s\}_{s=1}^S}{\text{minimize}} \quad & \sum_{s=1}^{S} -\frac{||\mathbf{z}^* - \mathbf{z}_s||^2}{2} \ln \upsilon_s \\ \text{s.t.} \quad & \sum_{s=1}^{S} \upsilon_s = 1. \end{aligned} \tag{8}$$

This is essentially the *Cross Entropy* (CE) of source uncertainty υ_s and $||\mathbf{z}^* - \mathbf{z}_s||/2$, which is the similarity between the optimum probability vector to each source vector ($0 \leq \upsilon_s \leq 1, 0 \leq ||\mathbf{z}^* - \mathbf{z}_s||/2 \leq 1$). Thus, the minimization process is to ensure that the solution resolves the ambiguity of the system as much as possible. Hence, truth discovery can ideally benefit uncertainty calibration through finding the truth vector.

Algorithms for approximating the global optimum exist [Ding and Xu, 2020, Huang et al., 2019]. But here with the assumption Eq. (1) that all the possible truth vectors fall on the probability simplex $\triangle^L$, we adopt a simpler solution. Since both the truth vector $\mathbf{z}^*$ and source reliabilities are unknown, we can alternatively update the reliability/uncertainty and the truth vector. Specifically, if $\mathbf{z}^*$ is temporarily fixed, the optimum reliability values can be found through Lemma 3.1:

Lemma 3.1 ([Li et al., 2014])**.** If $\mathbf{z}^*$ is fixed, the following reliability value for each source ω_s minimizes the objective

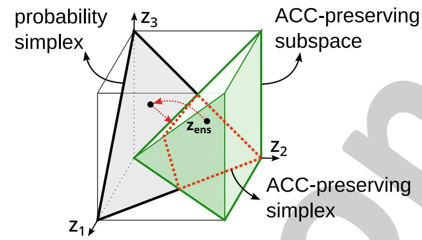


Figure 1: Geometric optimization of accuracy-preserving truth vector in $\mathbb{R}^3$. Here, 2 is the predicted class, and the accuracy-preserving simplex of class 2 is highlighted with red boundary.

function (7),

$$\omega_s = \ln\left(\frac{\sum_{t=1}^{S} ||\mathbf{z}^* - \mathbf{z}_t||^2}{||\mathbf{z}^* - \mathbf{z}_s||}\right). \tag{9}$$

After the reliabilities have been fixed, the new truth vector can be updated by simply taking the average of source vectors weighted by the found reliabilities, i.e., $\sum_{s=1}^{S} \omega_s \mathbf{z}_s$. It can be easily justified that the updated vector is still on the probability simplex. Initially, the ensemble vector $\mathbf{z}_{ens}$ can be an educated guess of $\mathbf{z}^*$. The iterative updating of the truth vector and the source reliability can be terminated if the position of the truth vector changed by less than ϵ within maximum I iterations. The process is summarized in Algorithm 1 (ignore line #5 at this time).

Algorithm 1: Optimization of the truth vector.

Data: $\{\mathbf{z}_s\}_{s=1}^S$
Result: $\mathbf{z}^*$
1 $\mathbf{z}^{*(0)} \leftarrow \mathbf{z}_{ens}$;
2 **for** $i \leftarrow 1, ..., I$ **do**
3 $\quad \omega_s^{(i)} \leftarrow \ln(\sum_{t=1}^{S} ||\mathbf{z}^{*(i-1)} - \mathbf{z}_t||^2 / ||\mathbf{z}^{*(i-1)} - \mathbf{z}_s||)$;
4 $\quad \mathbf{z}^{*(i)} \leftarrow \sum_{s=1}^{S} \omega_s^{(i)} \mathbf{z}_s$; ◁ update truth vector
5 $\quad \mathbf{z}^{*(i)} \leftarrow$ **Algorithm** $2(\mathbf{z}^{*(i)})$; ◁ preserve accuracy
6 $\quad$ **if** $||\mathbf{z}^{*(i)} - \mathbf{z}^{*(i-1)}||^2 < \epsilon$ **then**
7 $\quad\quad$ **return** $\mathbf{z}^{*(i)}$

3.2 ACCURACY-PRESERVING TRUTH DISCOVERY

Post-hoc calibration methods may potentially cause a decrease of the prediction accuracy, if the ranks of the scores for each class cannot be maintained. Hence, we usually anticipate the calibration algorithm to be *accuracy-preserving* [Zhang et al., 2020]. In the vanilla truth discovery algorithm, however, this requirement may not be satisfied, since $\mathbf{z}^*$ could change the predictions. Suppose that c is the predicted class derived from ensemble $\mathbf{z}_{ens}$, i.e., $c = \arg\max_l (z_{ens})_l$.

After the truth vector $\mathbf{z}^*$ is found, we want to at least maintain the accuracy of ensemble. Thus, Eq. (7) can be retrofitted to be an accuracy-preserving version.

Definition 3.2 (Accuracy-preserving truth discovery)**.** Given the set of points $\{\mathbf{z}_s\}_{s=1}^S \subseteq \triangle^L$ and the ensemble vector $\mathbf{z}_{ens}$, find the truth vector $\mathbf{z}^* \in \triangle^L$ and reliabilities $\{\omega_s\}_{s=1}^S$ such that the following objective function is minimized:

$$\begin{aligned} \underset{\mathbf{z}^*, \{\omega_s\}_{s=1}^S}{\text{minimize}} \quad & \sum_{s=1}^S \omega_s ||\mathbf{z}^* - \mathbf{z}_s||^2 \\ \text{s.t.} \quad & \sum_{s=1}^S e^{-\omega_s} = 1 \\ & \arg\max_l z_l^* = \arg\max_l (z_{ens})_l. \end{aligned} \tag{10}$$

It can be formulated as a geometric optimization problem. See Figure 1 for an illustration when $L = 3$. The constraint $\arg\max_l z_l^* = \arg\max_l (z_{ens})_l$ introduces a subspace $\Omega : z_l > z_m, \forall m \neq l$. The discovery of the truth vector must be performed within the accuracy-preserving simplex $\triangle_a = \triangle_L \cap \Omega$. Intuitively, when $\mathbf{z}^*$ falls outside of the accuracy-preserving simplex $\triangle_a$ in one iteration, we can find the projection of $\mathbf{z}^*$ onto $\triangle_a$ to pull it back to $\triangle_a$. This projection can be found by Algorithm 2 which is proved in Theorem 3.1. When the projection is done, Algorithm 1 continues until the desired truth vector is found.

Theorem 3.1. *Algorithm 2 preserves the accuracy of the prediction.*

Proof. The theorem can be proved using Lagrange Multipliers with Karush-Kuhn-Tucker (KKT) Conditions, see supplementary material Sec. A.1 for details. □

Algorithm 2: Projection from truth vector $\mathbf{z}$ onto accuracy-preserving simplex $\triangle_a$.

```
Data: z
Result: z̃
c ← argmax_l (z_ens)_l; z̃ ← z;
M ← ARGSORT({z_1, ..., z_L});
if M[1] = c then
    return z
else
    for l ← 1, ..., L do
        z̃ ← 1/(l+1) (z_c + z_M[1] + ... + z_M[l]);
        if z̃ > z_M[l+1] then
            z̃_c ← z̃; z̃_M[n] ← z̃, ∀n ≤ l;
            return z̃
```

4 TRUTH DISCOVERY-REGULARIZED POST-HOC CALIBRATION

Although ECE-like scores are difficult to be optimized directly, recent works have attempted to minimize ECE either by using maximum mean calibration error (MMCE), a kernelized version of the ECE [Kumar et al., 2018], or by rank preserving transforms [Bai et al., 2021]. To provide a better solution, we formulate the minimization of ECE (and also ECE^{KDE}) as an optimization problem in high dimensions, and show how it can be easily extended to incorporate the information gained from truth discovery.

Optimization of ECEs. For simplicity, we only consider the confidence for the ground-truth (namely, top-1) class, which is essentially the probability of a sample being correctly predicted. Then, our learnable mapping becomes $w = \mathcal{T}(\mathbf{z}) : \triangle^L \to \mathbb{R}$. One step further, if only the *winning score* is considered, then $w = \mathcal{T}(v) : \mathbb{R} \to \mathbb{R}$. Next, we find the specific form of $\mathcal{T}$. It has been shown that deep neural networks tends to be overconfident on most of the predictions [Guo et al., 2017]. Inspired by this, we impose an attenuation factor $\varphi(v)$ on every sample, which is a function of the winning score v so that the adjusted confidence becomes $w = v - \varphi(v)$. The simplest form of $\varphi(v)$ is to use a constant within a bin P_b. Thus, we define an attenuation weight ψ_b for the bin P_b. All the attenuation weights $\{\psi_b\}_{b=1}^B$ can be viewed as a point $\psi \in \mathbb{R}^B$. Now we have the definition of the mapping function:

$$w = v - \varphi(v) = v - \psi_\kappa \tag{11}$$
$$\text{s.t.} \quad \mu_\kappa \leq v < \nu_\kappa. \tag{12}$$

Notice that for consistency we call $\varphi(v)$ the attenuation factor, but it can also enhance the confidence w if $\varphi(v) < 0$ somewhere.

Now our goal is to find the location of ψ in $\mathbb{R}^B$ such that the expected calibration error is minimized:

$$\underset{\{\psi_b\}_{b=1}^B}{\text{minimize}} \quad ECE(\{P_b\}_{b=1}^B), \tag{13}$$

where ECE is shown in Eq. (5). Since all the computations in ECE (and ECE^{KDE}) are differentiable, the minimization of ECEs can be done by gradient descent methods. Here, a mini-batch Stochastic Gradient Descent (SGD) approach is used. In each epoch, a subset of calibration data is sampled, based on which the attenuation weights are updated. The optimization process is encapsulated in Algorithm 3, in which ECE and ECE^{KDE} can be used interchangeably. The algorithm can be easily implemented by virtue of automatic differentiation libraries e.g. PyTorch [Paszke et al., 2019].

Compositional Approach for the Optimization of ECEs. Even an accelerated method for KDE computation is used [O'Brien et al., 2016], KDE-based metric is still much more

Table 1: Comparison of TDE/aTDE with Deep Ensemble (DE) before post-hoc calibration.

		CIFAR100						CIFAR10					
		50 sources			100 sources			50 sources			100 sources		
Model	Method	$ECE^{KDE}\downarrow$	$ECE\downarrow$	$ACC\uparrow$	$ECE^{KDE}\downarrow$	$ECE\downarrow$	$ACC\uparrow$	$ECE^{KDE}\downarrow$	$ECE\downarrow$	$ACC\uparrow$	$ECE^{KDE}\downarrow$	$ECE\downarrow$	$ACC\uparrow$
PreResNet110	DE	2.88	2.50	82.83	3.06	2.58	**83.02**	1.31	**0.48**	**96.38**	1.15	**0.43**	96.41
	TDE	1.60	1.98	**82.89**	1.81	**1.94**	83.01	**1.05**	0.75	**96.38**	**1.01**	0.59	**96.42**
	aTDE	**1.55**	**1.92**	82.83	**1.78**	1.95	**83.02**	1.07	0.75	**96.38**	1.02	0.60	96.41
PreResNet164	DE	2.39	2.09	**83.52**	2.46	2.00	**83.55**	1.32	**0.34**	**96.68**	1.31	**0.35**	96.67
	TDE	**1.31**	**1.72**	83.49	1.44	**1.61**	83.54	1.14	0.65	96.66	**1.08**	0.52	**96.68**
	aTDE	1.33	1.76	**83.52**	**1.42**	1.62	**83.55**	**1.13**	0.63	**96.68**	**1.08**	0.53	96.67
WideResNet	DE	6.45	5.85	**84.38**	6.40	5.78	84.28	**1.06**	**0.35**	**97.17**	1.20	0.41	**97.20**
	TDE	**5.48**	**5.03**	84.37	5.58	5.07	**84.29**	1.14	0.49	97.15	**1.10**	**0.40**	97.17
	aTDE	5.49	5.05	**84.38**	**5.57**	**5.06**	84.28	1.13	0.47	**97.17**	1.12	0.41	**97.20**
		ImageNet						ImageNet (Snapshot Ensemble)					
		25 sources			50 sources			25 sources			50 sources		
ResNet50	DE/SE	3.11	3.12	**79.25**	3.24	3.17	**79.37**	1.85	2.11	**78.46**	1.73	**2.03**	**78.52**
	TDE	**2.16**	**2.42**	79.22	**2.41**	**2.58**	79.35	**1.69**	**2.08**	78.45	**1.66**	2.10	78.50
	aTDE	2.19	2.45	**79.25**	2.43	2.60	**79.37**	1.70	2.10	**78.46**	1.69	2.13	**78.52**

Algorithm 3: Optimization of ECEs.

Data: $\{(v^{(i)}, y^{(i)})\}_{i=1}^{N_c}$
Result: $\{\psi_b\}_{b=1}^{B}$
for *epoch* $\leftarrow 1, ..., \#epoch$ **do**
  sample $\{(v^{(j)}, y^{(j)})\}_{j=1}^{n_c} \sim \{(v^{(i)}, y^{(i)})\}_{i=1}^{N_c}$;
  $w = v - \varphi(v)$; ◁ apply attenuation factor
  loss $\leftarrow$ ECE or ECE^{KDE} ; ◁ forward propagation
  update $\{\psi_b\}_{b=1}^{B}$; ◁ backward propagation
return $\{\psi_b\}_{b=1}^{B}$

time-consuming compared to histogram-based metrics. A natural question that arises here is whether the minimization of ECE^{KDE} can be speeded up by the minimization of ECE. To answer this, we first find the attenuation weights by using ECE as the loss function, and then use the obtained $\{\psi_b\}_{b=1}^{B}$ as an initial guess for the minimization of ECE^{KDE}. This approach enables the compositional optimization of histogram-based and KDE-based calibration errors.

ECE Optimization Regularized by Truth Discovery. Given a discovered truth vector $\mathbf{z}^*$, let V denote the total squared distance to $\mathbf{z}^*$ (i.e., $V = \sum_{s=1}^{S} ||\mathbf{z}^* - \mathbf{z}_s||^2$) and q_s denote the contribution of each $\mathbf{z}_s$ to V (i.e., $q_s = ||\mathbf{z}^* - \mathbf{z}_s||^2/V$). Then, the entropy induced by $\{q_s\}_{s=1}^{S}$ is:

$$H = -\sum_{s=1}^{S} q_s \log q_s = \frac{1}{V}\sum_{s=1}^{S} ||\mathbf{z}^* - \mathbf{z}_s||^2 \log \frac{V}{||\mathbf{z}^* - \mathbf{z}_s||^2}. \tag{14}$$

Based on these, we can define the *Entropy based Geometric Variance* (HV):

Definition 4.1 (Entropy based Geometric Variance [Ding and Xu, 2020])**.** Given the point set $\{\mathbf{z}_s\}_{s=1}^{S}$ and a point $\mathbf{z}^*$, the Entropy based Geometric Variance (HV) is $H \times V$ where H and V are defined as shown above.

With this definition, it is easy to see that the objective function of truth discovery (7) is exactly the entropy based geometric variance (HV) and the optimization problem (7) is equivalent to finding a point $\mathbf{z}^*$ to minimize HV.

If the truth vector $\mathbf{z}^*$ has been determined, then HV is an indicator of the ambiguity of the system, and can be borrowed as an external information for our ECE optimization. Despite being in the same bin and overconfident, the sample confidences should not be attenuated at the same scale. Instead, the sample with higher HV (i.e. higher variance and uncertainty) is to be attenuated by a larger magnitude. Consequently, the mapping function can be reshaped as:

$$w^{(i)} = v^{(i)} - \varphi(v^{(i)}) = v^{(i)} - \alpha_1\psi_\kappa - \alpha_2 HV^{(i)}, \tag{15}$$

where α_1, α_2 are hyperparameters, and $HV^{(i)}$ is essentially the value of the objective function of truth discovery as computed in Section 3.2 for each sample $(x^{(i)}, y^{(i)})$ in a total of $N_c + N_e$ calibration and evaluation samples. The learning of the mapping function $\mathcal{T}$ from the calibration data is a supervised learning problem. Hence, $\mathcal{T}$ is expected to be overfitted to calibration data. By incorporating orthogonal information (i.e. HV) acquired from the truth discovery of multiple ensemble classifiers, we can learn a mapping $\mathcal{T}$ that generalizes better on the evaluation datasets.

5 EXPERIMENTS

The main goals of our experiments are to: (1) compare Truth Discovery Ensemble (TDE), especially the accuracy-preserving version (aTDE), with ensemble-based calibration

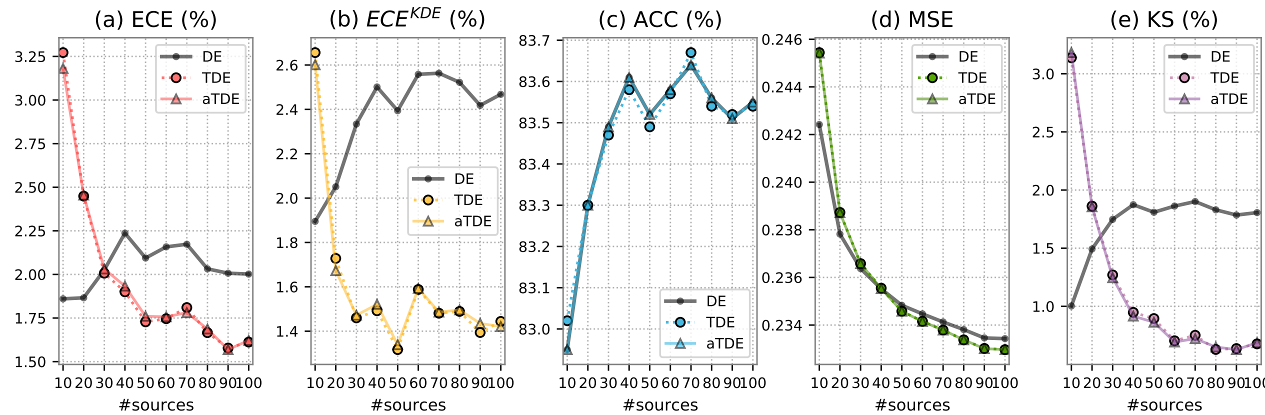


Figure 2: Results of DE/TDE/aTDE on PreResNet164 trained upon CIFAR100 in an unsupervised manner (i.e. no hold-out calibration set). The number of sources is increased from 10 to 100. Our method works favorably with various metrics for the evaluation of calibration.

Table 2: Comparison of post-hoc calibration methods. The **best** results are highlighted in bold. We also underline the best results excluding our method. ECEs are reported in terms of mean±standard deviation obtained from 5 random replications.

		ECE↓					ECE^{KDE}↓				
Dataset	Model	DE	TS	ETS	IRM	pTDE	DE	TS	ETS	IRM	pTDE
CIFAR100	ResNet18	2.86±0.42	2.70±0.36	2.31±0.30	2.52±0.35	**1.64±0.37**	2.41±0.35	2.37±0.37	2.25±0.33	2.09±0.36	**1.63±0.37**
CIFAR100	DenseNet121	1.69±0.15	1.68±0.14	1.45±0.15	1.70±0.19	**1.29±0.19**	1.67±0.07	1.66±0.08	1.65±0.12	1.66±0.15	**1.49±0.12**
CIFAR100	ResNeXt29	1.92±0.26	1.93±0.27	1.54±0.22	1.88±0.36	**1.20±0.34**	1.89±0.12	1.86±0.07	1.87±0.11	1.74±0.24	**1.63±0.12**
CIFAR10	ResNet18	0.94±0.14	0.65±0.13	0.65±0.13	0.67±0.24	0.75±0.10	1.50±0.11	1.48±0.06	1.48±0.06	1.35±0.08	**1.21±0.18**
CIFAR10	DenseNet121	0.88±0.11	0.65±0.13	0.65±0.13	0.55±0.06	**0.50±0.14**	1.78±0.17	1.76±0.16	1.76±0.16	1.56±0.15	**1.45±0.20**
CIFAR10	ResNeXt29	0.46±0.09	**0.33±0.10**	0.34±0.10	0.45±0.13	0.39±0.15	1.38±0.06	1.43±0.09	1.43±0.09	1.42±0.10	**1.36±0.06**
	#sources	DE/SE	TS	ETS	IRM	pTDE	DE/SE	TS	ETS	IRM	pTDE
ImageNet	10 (DE)	3.06±0.12	1.59±0.09	0.96±0.07	1.93±0.16	**0.88±0.11**	3.13±0.11	1.16±0.14	1.06±0.11	1.79±0.11	**0.93±0.08**
ImageNet	30 (DE)	3.19±0.06	1.47±0.15	0.87±0.10	1.89±0.12	**0.81±0.16**	3.26±0.05	1.10±0.12	0.97±0.09	1.75±0.07	**0.95±0.11**
ImageNet	10 (SE)	2.32±0.09	1.56±0.05	1.04±0.10	1.74±0.04	**0.71±0.17**	2.18±0.09	1.14±0.06	1.07±0.08	1.51±0.07	**0.95±0.07**
ImageNet	30 (SE)	2.13±0.07	1.57±0.07	0.95±0.15	1.70±0.07	**0.76±0.22**	1.89±0.09	1.16±0.08	0.95±0.17	1.43±0.11	**0.87±0.22**

Table 3: Ablation Study of the proposed truth discovery-regularized post-hoc calibration (pTDE). For the four variants of pTDE, the blue/red color denotes if compositional training (Comp.)/truth-discovery regularization (Truth. Reg.) is utilized.

		ECE↓				ECE^{KDE}↓			
Dataset	Model	opt^{hist}	opt^{KDE}	$pTDE^{hist}$	$pTDE^{KDE}$	opt^{hist}	opt^{KDE}	$pTDE^{hist}$	$pTDE^{KDE}$
Comp.	**Truth. Reg.**	✗ ✗	✓ ✗	✗ ✓	✓ ✓	✗ ✗	✓ ✗	✗ ✓	✓ ✓
CIFAR100	ResNet18	1.74±0.22	1.68±0.38	**1.59±0.37**	1.64±0.37	1.76±0.33	1.73±0.31	1.71±0.35	**1.63±0.37**
CIFAR100	DenseNet121	**1.29±0.17**	1.31±0.15	1.29±0.20	1.29±0.19	1.56±0.09	1.53±0.11	1.52±0.09	**1.49±0.12**
CIFAR100	ResNeXt29	1.27±0.33	1.24±0.35	**1.19±0.35**	1.20±0.34	1.68±0.12	1.67±0.12	1.64±0.11	**1.63±0.12**
CIFAR10	ResNet18	0.65±0.20	0.72±0.15	**0.65±0.17**	0.75±0.10	1.33±0.12	1.26±0.18	1.31±0.12	**1.21±0.18**
CIFAR10	DenseNet121	0.45±0.06	0.46±0.16	**0.44±0.04**	0.50±0.14	1.54±0.16	1.46±0.17	1.55±0.18	**1.45±0.20**
CIFAR10	ResNeXt29	0.41±0.10	0.41±0.13	0.40±0.11	**0.39±0.15**	1.37±0.07	**1.36±0.07**	1.37±0.06	1.36±0.06
	#sources	opt^{hist}	opt^{KDE}	$pTDE^{hist}$	$pTDE^{KDE}$	opt^{hist}	opt^{KDE}	$pTDE^{hist}$	$pTDE^{KDE}$
ImageNet	10 (DE)	**0.76±0.13**	0.83±0.13	0.81±0.09	0.88±0.11	0.99±0.08	**0.92±0.07**	1.01±0.08	0.93±0.08
ImageNet	30 (DE)	0.88±0.13	0.85±0.11	0.87±0.16	**0.81±0.16**	1.09±0.12	1.02±0.09	1.03±0.10	**0.95±0.11**
ImageNet	10 (SE)	0.80±0.19	0.73±0.19	0.75±0.19	**0.71±0.17**	1.04±0.12	0.96±0.09	1.01±0.11	**0.95±0.07**
ImageNet	30 (SE)	0.81±0.16	**0.76±0.20**	0.78±0.17	0.76±0.22	0.98±0.19	0.90±0.22	0.95±0.22	**0.87±0.22**

methods; (2) for post-hoc calibration scheme, compare truth discovery-regularized calibration (pTDE) with state-of-the-art methods on a wide range of network architectures and datasets; investigate if different components of our methods collaboratively contribute to the overall elevation of performance by ablation studies.

5.1 IMPROVED DEEP ENSEMBLE BY TRUTH DISCOVERY

Experimental Setup. For a fair comparison, we downloaded the trained models* of Ashukha et al. [2020] including PreResNet110/164 [He et al., 2016] and WideResNet28x10 [Zagoruyko and Komodakis, 2016] trained on CIFAR10/100 [Krizhevsky, 2009] (10/100 classes), and ResNet50 trained on ImageNet [Deng et al., 2009] (10000 classes). All 3 network architectures on CIFAR10/100 were trained 100 times (i.e. $S = 100$) following the Deep Ensemble (DE) workflow, while ResNet50 was trained on ImageNet resulting in 50 models either by Deep Ensemble or by Snapshot Ensemble (i.e. $S = 50$). For every sample in the standard testing dataset, S ensemble members were generated from the S models. While looking for the truth vector, for both the vanilla Truth Discovery Ensemble (TDE) and its accuracy-preserving counterpart (aTDE), we set $\epsilon = e^{-8}$ in all experiments, and observed a convergence within typically 5 iterations.

Results. The comparison of truth discovery ensemble methods with Deep Ensemble on CIFARs ($S = 50$ or 100), and with Deep Ensemble/Snapshot Ensemble on ImageNet ($S = 25$ or 50) is shown in Table 1. Clearly, TDE ameliorates either ECE or ECE^{KDE} by a large margin in most of the experimental settings, especially on datasets with higher complexity (i.e., CIFAR100 and ImageNet), but fails at maintain accuracy. The accuracy-preserving version aTDE, on the other hand, successfully preserves the accuracy, with nearly the same capability of lowering the ECEs, which validates the correctness of our accuracy-preserving Algorithm 2. It can also be concluded from Table 1 that higher ACC and lower ECEs are hard to be reached simultaneously, but the metrics contributed to by the two variants TDE/aTDE are usually very similar. The KS metric recently proposed by [Gupta et al., 2021] which measures the maximal distance between the accumulated output probability to the actual probability, is a binning-free calibration evaluator that different from ECEs. The KS error is also measured for all the experiments. By leveraging truth discovery, our TDE method lowers the KS to as low as 0.6% (100 sources) as shown in Figure 2(e), even without any hold-out calibration sample.

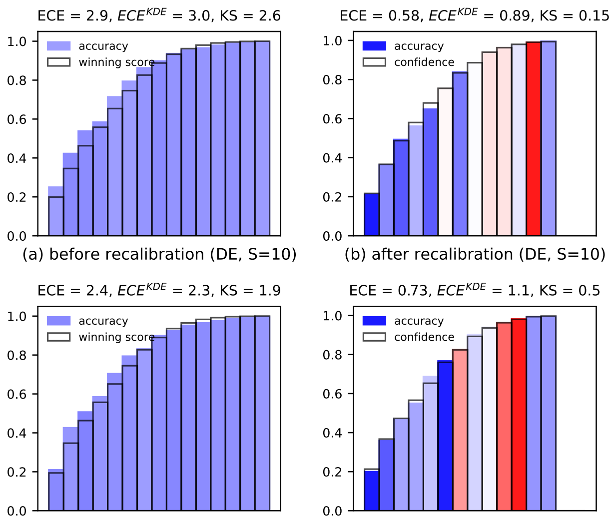
Further, we investigate the stability of TDE/aTDE with different number of sources by changing S by an interval of 10 for CIFARs and 5 for ImageNet, as illustrated in Figure 2 and Tables A1, A2, A3, and A4. Interestingly, Deep Ensemble tends to be overconfident with larger number of sources, i.e., ensemble members, while TDE/aTDE works favorably with even larger amount of available sources, and this is when a high accuracy is usually reached (Figure 2c), suggesting TDE and aTDE's superior ability in utilizing information from multiple sources than Deep Ensemble.

5.2 IMPROVED POST-HOC CALIBRATION BY TRUTH DISCOVERY-REGULARIZED OPTIMIZATION

Experimental Setup. In this section, we evaluate the performance of our truth discovery-regularized post-hoc calibration methods (pTDE), to which the information elicited from ensemble-based methods is incorporated. To this end, we train ResNet18, DenseNet121 [Huang et al., 2017b] and ResNeXt29 [Xie et al., 2017] on CIFAR10/100 datasets with Snapshot Ensemble scheme and obtain 200 ensemble members as the sources (i.e., $S = 200$). See Section A.2 for details. The pre-trained ImageNet models are also used with source numbers set at 10 and 30 for both DE and SE. We first train the vanilla optimization method using histogram-based ECE as the loss function (opt^{hist}) as described in Section 4 with batch size at 1000 for CIFARs and 10000 for ImageNet for 70 epochs. Then, we apply compositional training by switching to KDE-based loss function (opt^{KDE}) for 5 additional epochs. To leverage the information gained from truth discovery, the entropy based geometric variance (HV) values are computed for all $(N_c + N_e)$ samples. By taking HV into the training process, opt^{hist} is promoted to truth discovery-regularized post-hoc calibration pTDE^{hist}, which is subsequently optimized for 5 additional epochs using ECE^{KDE} as the loss function to be pTDE^{KDE}. Finally, pTDE^{KDE} (or pTDE for short) is compared with several state-of-the-art post-hoc calibration methods, namely, Temperature Scaling (TS) [Guo et al., 2017], Ensemble Temperature Scaling (ETS) [Zhang et al., 2020], and multi-class isotonic regression (IRM) [Zhang et al., 2020].

Results. The attenuation/enhancement factor we apply in Eq. (11) enables an iterative rearrangement of samples across bins, until a small discrepancy between confidence and actual accuracy is achieved within every bin. Figure 3 shows how this recalibration (vanilla opt^{hist}) affects confidence distribution. From Table 2, we can see that our method (pTDE) significantly outperforms all competing methods (except on CIFAR10 with ECE). It is worth noting that when pTDE is excluded, there is no sweeping method under every circumstances, but pTDE overall shows better consistency especially with the ECE^{KDE} metric which is not susceptible to the binning strategy, e.g., the number and positions of the bins. To inspect the individual contributions

*downloaded from https://github.com/bayesgroup/pytorch-ensembles.



(a) before recalibration (DE, S=10) (b) after recalibration (DE, S=10)

(c) before recalibration (SE, S=10) (d) after recalibration (SE, S=10)

Figure 3: Winning score/confidence vs. actual accuracy on the unseen ImageNet evaluation dataset for Deep Ensemble (a, b) and Snapshot Ensemble (c, d), both with 10 sources, before (a, c) and after (b, d) the post-hoc calibration. The initial bins in (a, c) are selected such that all samples are evenly distributed over the bins, as suggested by [Zhang et al., 2020]. Then we fix the positions of bins in (b, d) for a better illustration. The color temperature of each bin reflects the number of samples that fall into that bin. The warmer the color is, the more samples the bin contains, and vice versa.

of compositional training and truth discovery regularization to the total performance, we conduct ablation study for the two components as shown in Table 3. Generally, the truth discover-regularized methods (pTDE^{hist} and pTDE^{KDE}) perform better. Depending on which kind of calibration error is used as loss function, it is purposefully minimized. This indicates that pTDE^{hist} performs better with ECE metric, while pTDE^{KDE} favors ECE^{KDE} metric. Notably, although ECE^{KDE} is targeted by pTDE^{KDE}, ECE can still be decreased on ImageNet, for example, when 30 sources from Deep Ensemble is used, showing the effectiveness of our proposed compositional training and truth discovery regularization. To get more insight into why truth discovery improves calibration performance, we show the relationship of the computed Entropy based geometric variance (HV) with the winning score in Figures A1 A2 A3 showing that truth discovery indeed provides information that orthogonal to winning score itself, and thus to prevent overfitting to the calibration dataset.

Finally, all the calibration results are further measured upon KS, shown in Table 4, and surprisingly, although pTDE is not specifically designed for the optimization of KS, its fully-fledged version pTDE^{KDE} is competitive, or even better than Spline, showing that the truth discovery-based regularizer is

Table 4: KS error (in %) on ImageNet evaluation dataset by various post-hoc calibration methods including four variants of our proposed method. The best results are shown in bold.

	Deep Ensemble		Snapshot Ensemble	
Method	$S = 10$	$S = 30$	$S = 10$	$S = 30$
DE/SE	2.71±0.10	2.86±0.03	1.83±0.08	1.52±0.09
TS	0.88±0.13	1.03±0.09	1.02±0.12	1.12±0.15
ETS	0.59±0.13	0.42±0.10	0.50±0.11	0.49±0.15
IRM	0.97±0.13	0.93±0.10	0.75±0.09	0.66±0.14
Spline	0.38±0.09	0.34±0.07	0.27±0.08	**0.30**±**0.11**
opt^{hist}	0.43±0.20	0.39±0.16	0.41±0.06	0.36±0.17
opt^{KDE}	0.37±0.15	0.33±0.11	0.31±0.07	0.33±0.14
pTDE^{hist}	0.42±0.18	0.36±0.19	0.26±0.06	0.34±0.18
pTDE^{KDE}	**0.37**±**0.14**	**0.27**±**0.12**	**0.25**±**0.04**	**0.31**±**0.13**

also beneficial to the minimization of the KS metric.

6 CONCLUSION

In this work, we first present Truth Discovery Ensemble (TDE) that neither requires hold-out calibration data nor alters any training process, but significantly surpasses the original ensemble result, and in the meanwhile preserves the accuracy (aTDE). For post-hoc calibration, the superiority of our final methods (pTDE) is attributed not only to truth discovery, but also to the compositional training strategy. In conclusion, truth discovery is well positioned to assist both ensemble-based and post-hoc calibration. We hope that the proposed calibrators augmented by truth discovery can enlarge the arsenal of uncertainty calibration methods for deep learning. Our source code is available at https://github.com/horsepurve/truly-uncertain.

Acknowledgements

This research was supported in part by NSF through grants CCF-1716400 and IIS-1910492.

References

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BJxI5gHKDr`.

Yu Bai, Tengyu Ma, Huan Wang, and Caiming Xiong. Improved uncertainty post-calibration via rank preserving transforms, 2021. URL `https://openreview.net/forum?id=jsM6yvqiT0W`.

Raef Bassily, Adam Groce, Jonathan Katz, and Adam D. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 439–448, 2013.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

Hu Ding and Jinhui Xu. Learning the truth vector in high dimensions. *Journal of Computer and System Sciences*, pages 78–94, 2020.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete Dropout. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8803–8812, 2018.

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.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=eQe8DEWNN2W`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 770–778, 2016.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019a.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pages 15637–15648, 2019b.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *5th International Conference on Learning Representations, ICLR*, 2017a.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2261–2269, 2017b.

Ziyun Huang, Hu Ding, and Jinhui Xu. A faster algorithm for truth discovery via range cover. *Algorithmica*, 81(10): 4118–4133, 2019.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical report*, 2009.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, pages 12295–12305, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, Proceedings of Machine Learning Research, pages 2810–2819, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6405–6416, 2017.

Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *International Conference on Management of Data, SIGMOD*, pages 1187–1198, 2014.

Shi Li, Jinhui Xu, and Minwei Ye. Approximating global optimum for probabilistic truth discovery. *Algorithmica*, pages 3091–3116, 2020.

Chunwei Ma, Yan Ren, Jiarui Yang, Zhe Ren, Huanming Yang, and Siqi Liu. Improved peptide retention time prediction in liquid chromatography through deep learning. *Analytical Chemistry*, 90(18):10881–10888, 2018.

Chunwei Ma, Zhanghexuan Ji, and Mingchen Gao. Neural style transfer improves 3D cardiovascular MR image segmentation on inconsistent data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 128–136. Springer, 2019.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, pages 13132–13143, 2019.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4696–4705, 2019.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2901–2907, 2015.

Travis A. O'Brien, Karthik Kashinath, Nicholas R. Cavanaugh, William D. Collins, and John P. O'Brien. A fast and objective multidimensional kernel density estimation method: fastkde. *Computational Statistics & Data Analysis*, 101:148–160, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 13888–13899, 2019.

Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5987–5995, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC*, 2016.

Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. Mix-n-match : Ensemble and compositional methods for uncertainty calibration in deep learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, volume 119, pages 11117–11128, 2020.