ENHANCING DNN RELIABILITY WITH REFINEMENT AND CALIBRATION

Ramya Hebbalaguppe* **Ajay Shastry** **Soumya Suvra Ghosal** **Chetan Arora**
SIT, Indian Institute of Technology Delhi, New Delhi, India
Project Webpage: `https://github.com/rhebbalaguppe/RefCal`

ABSTRACT

Although deep neural networks (DNNs) achieve high predictive accuracy, their confidence estimates are often unreliable, potentially compromising user trust in their decisions. This has driven research into `calibrated` models— where `calibration` measures the degree to which a model's predictive confidence matches the empirical probability of correctness. However, such a calibration metric can be changed by post-processing output to mimic the uncertainty of training time, without truly improving the model's understanding. Hence, statisticians recommend a model to be `refined` as well as calibrated. Intuitively, a model is said to be more refined if there is a large difference in its predictive confidence for correct and incorrect predictions (sometimes called `sharpness`). Calibration and refinement improve the statistical alignment between model uncertainty and real-world outcomes. We observe that common calibration methods often reduce model refinement. To address this, we propose: **(1)** a novel loss function that promotes refinement and is optimizable via supervised contrastive loss; **(2)** a unified training framework, `RefCal`, that jointly optimizes for calibration, refinement, and accuracy—enhancing DNN reliability. For eg., we report (accuracy↑, refinement↑, ECE↓) of (58.81, 95.67, 0.08) on CIFAR-100-LT dataset (10% class imbalance), surpassing (46.27, 93.7, 0.22) by the well known Correctness Ranking Loss Moon et al. (2020).

1 INTRODUCTION

Advances in datasets, model architectures, and computational resources have made it easier to achieve high accuracy with deep neural networks (DNNs). Consequently, research has shifted toward improving complementary performance metrics that enhance model trustworthiness, particularly reliability. This is crucial in safety-critical domains such as healthcare and autonomous vehicles, where models must not only be accurate but also properly quantify and communicate predictive uncertainty—exhibiting high confidence in correct predictions while reflecting uncertainty in cases of error. In this work, we examine reliability through both, calibration and refinement. Calibration measures how well predicted probabilities match true outcome frequencies, while refinement captures the sharpness of predictions. We formally define both dimensions and argue that both are essential for robust and trustworthy decision-making in DNNs.

Calibration. We consider a K-class classification problem, with $X \in \mathcal{X}$ as the input, $Y \in \mathcal{Y}$ as the target random variable, and $f(X)$ as the predicted confidence vector from a DNN based classification model, f. Let $\mathcal{P}$ as the set of distributions on $\mathcal{Y}$. Since, our focus is on classification, we use $\mathcal{P}_K$ to denote the K-dimensional simplex of corresponding categorical distributions. We use $\mathbb{P} \in \mathcal{P}_K$, and $\mathbb{P}_{Y|f(X)} \in \mathcal{P}_K$ to denote the distribution of Y, and conditional distribution given X, respectively.

Definition 1 (Calibration Gruber & Buettner (2022))**.** A model $f : \mathcal{X} \to \mathbb{P}$ is calibrated, if and only if, $f(X) = \mathbb{P}_{Y|f(X)}$.

An alternative formulation focuses on calibrating only the *top-label* prediction of a model $f : \mathcal{X} \to \mathcal{P}_K$. Let $C = \arg\max_k f_k(X)$, denote the predicted class with the highest confidence (top-label

*Contact: ramya.murthy@gmail.com

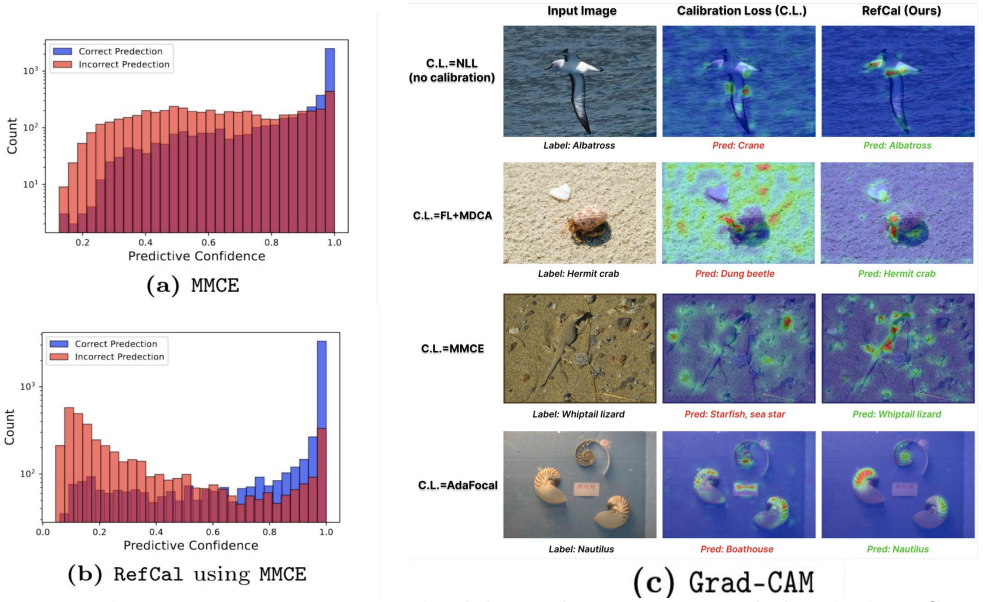


Figure 1: [Why `RefCal`?]: Our proposed training regime, `RefCal` optimizes both, **Ref**inement and **Cal**ibration. `(a)` shows when we calibrate a `ResNet-50` model using MMCE Kumar et al. (2018) calibration on `CIFAR100-LT`, it results in lower separation between the confidence values of the correct and incorrect predicted classes (red and blue). To this end, when we use the proposed proxy refinement loss along with calibration (For eg., using MMCE), it can be seen that the separability of correct and incorrect predictions is enhanced as shown in `(b)`, indicating better refinement. Quantitatively, (AUC (↑), ECE (↓)) for (refinement, calibration) improve from $(94.40, 0.25)$ for MMCE to $(95.03, 0.07)$ for MMCE + proposed refinement loss. `(c)` We provide Grad-CAM visualizations for `Resnet-18` trained on `ImageNet-LT`, using a particular calibration technique (each row, second column), and then by optimizing with the same calibration but adding our refinement loss (each, third column). As refinement loss forces a model to separate its confidence for correct, and incorrect predictions, it also leads the model to focus its attention on the salient object features, instead of the background.

softmax probability). In this setting, the calibration condition is expressed as:

$$f_C(X) = \mathbb{P}(Y = C \mid f_C(X)),$$

which states that the confidence assigned to the top-label C should match the true probability of C being the correct class.

Guo et al. (2019) showed that modern DNNs often produce overconfident but incorrect predictions, prompting various efforts to improve their calibration Platt et al. (1999); Kull et al. (2019); Kumar et al. (2019; 2018). Train time methods typically reduce over/under-confidence Kumar et al. (2018).

Pitfalls of Calibration Metrics

In a binary classification setting with validation confusion matrix $\{\{0.7, 0.3\}, \{0.2, 0.8\}\}$, one can post-process test predictions by assigning a fixed confidence vector $\{0.7, 0.3\}$ to all malignant predictions and $\{0.2, 0.8\}$ to all benign ones. This yields good calibration scores since predicted confidences match observed frequencies. However, this uniform assignment reduces confidence differentiation between true and false positives, thereby weakening the model's discriminative power while preserving accuracy. Post-hoc calibration can improve calibration metrics many times at the cost of predictive confidence differentiation. Calibration and refinement (sharpness) improve the statistical alignment between model uncertainty and real-world outcomes.

Refinement. Refinement was introduced by DeGroot and Fienberg in 1982, promotes sharper predictive distributions by encouraging confidence scores to be closer to 0 or 1 DeGroot & Fienberg

(1982). This behavior enhances the model's ability to distinguish between correct and incorrect predictions, thereby improving its discriminative power(See Fig. 1). The refinement trade-off was rediscovered recently, when Moon et al. (2020) proposed *correctness ranking loss* (CRL) to improve refinement by enforcing the following relationship for every pair of samples (x_i, y_i) and (x_j, y_j):

$$c_i \leq c_j \iff \mathbb{P}(\hat{y}_i = y_i \mid x_i) \leq \mathbb{P}(\hat{y}_j = y_j \mid x_j). \tag{1}$$

Here y denotes the target label, and $\hat{y}$ predicted label, for a sample x, and the predictive confidence, c. Recently, Yuan (2021); Yang et al. (2021) proposed a differentiable loss function that directly maximizes area under the ROC curve (AUC). They showed that for a two-class classification:

$$\texttt{AUC} = \mathbb{E}[\mathbb{I}(c_i > c_j)], \text{ s.t. } x_i \in \mathcal{S}_p, \text{and } x_j \in \mathcal{S}_n, \tag{2}$$

where $\mathcal{S}_p$ and $\mathcal{S}_n$ denote the set of correctly and incorrectly classified samples of a model on the test set. The result indicates that AUC can be used as metric to score a model's refinement performance. We show the improvement in refinement using our proposal with AUC as the metric.

Contributions. The main contributions of this work are:

- **A surrogate loss for refinement.** We propose a novel surrogate loss function that promotes refinement by encouraging sharper predictive distributions. We show, through mathematical analysis, that this loss can be minimized using the supervised contrastive loss Khosla et al. (2020), enabling the reuse of existing contrastive learning frameworks for refinement. On the STL10 dataset (in binary setting; see Tab. 3), our method achieves an AUC of 99.90, outperforming the state-of-the-art score of 98.86 Yuan (2021). **Scalable refinement for multi-class classification.** Existing refinement or AUC-based optimization approaches are typically limited to binary classification and require one-vs-all schemes for multi-class settings. In contrast, our formulation—based on supervised contrastive loss—extends naturally to multi-class classification.
- **RefCal: A two-stage framework for optimization of reliability objectives.** We introduce RefCal, a two-stage training procedure that first optimizes for refinement using supervised contrastive learning while the second stage fine-tunes the model using standard calibration and accuracy losses. On the CIFAR10-LT benchmark, RefCal achieves the best combined performance: {ECE: 1.05, AUC: 99.04, Accuracy: 88.11}, compared to the SOTA baseline Liu et al. (2022) with {ECE: 7.05, AUC: 98.59, Accuracy: 87.82}.
- **Revisiting the calibration–refinement trade-off for DNN classifiers.** Prior literature Murphy (1973); Singh et al. (2021) interprets Brier score decomposition to imply a trade-off between calibration and refinement. We clarify that this conflict arises only under the assumption of constant error. Our empirical results show that refinement can improve alongside calibration and accuracy (see Fig. 1).
- **Source code.** We will release code, models, and evaluation protocols upon acceptance.

2 Related work

Refinement. Refinement in literature is studied in two different styles. The first one aims at the separability of correct and incorrect predictions with a margin based on the predicted confidence and the second tries to optimize the ordinal ranking relationship between correctly classified an incorrectly classified samples. CRL Moon et al. (2020) is a popular technique in the first category which learns ordinal relationship by introducing margin-based penalties. In the second style for refinement, researchers have exploited its relationship with AUC (Eq. 2), and developed techniques to directly optimize AUC Yang (2022), or its proxy Yuan (2021). Margin maximization seems theoretically and practically superior, as it facilitates generalization error analysis and presents a clear geometric interpretation of the models being built. The resulting techniques have also been applied to large-scale medical image datasets Yang et al. (2023). However, the extensions to multi-class do not appear obvious, and calibration in conjunction with refinement has not been investigated systematically.

Calibration. Train time loss functions such as FL Mukhoti et al. (2020), MMCE Kumar et al. (2018) Patra et al. (2023)Hebbalaguppe et al. (2024)Rawat et al. (2021) Ghosal et al. (2025) Hebbalaguppe et al. (2025) etc., aim to alleviate the miscalibration in DNNs but do not address refinement. On the other hand, post-hoc calibration techniques Ding et al. (2020); Müller et al. (2019); Kull et al. (2019), tackle miscalibration by adding a parameterized calibration component to a DNN, which

can be fine-tuned using a separate validation set. Recent calibration metrics include Błasiok et al. (2023); Gruber & Buettner (2022); Nixon et al. (2019). Post-hoc methods typically have fewer learnable parameters compared to train-time techniques, resulting in limited calibration capabilities prompting us to employ train-time calibration in the proposed approach `RefCal`.

3 PROPOSED METHODOLOGY

We first design a surrogate loss for the refinement objective by proposing a new loss function optimized via supervised contrastive loss Khosla et al. (2020). Next, we perform calibration by training a classifier on frozen refined representations using standard calibration losses. This two stage procedure is termed `RefCal`, which achieves state-of-the-art calibration and refinement simultaneously.

3.1 PROPOSED REFINEMENT LOSS

Notation. Let $i \in \mathcal{I}$ denote indices of all the samples in our dataset. Let $z_i = f(x_i)$ be representation/embedding for a sample x_i, that we wish to learn through a DNN. We normalize z_i such that it lies on a unit hypersphere. The normalization is important to enable the use of the inner product to measure distances in the projection space. We also call z_i as the *anchor*, and use z_p/z_n to denote a sample with the same/different label as z_i. The set of all samples with the same label as z_i is called a *positive set* (denoted as $\mathcal{P}_i$). Similarly, the set of samples with a different label is called a *negative set* (denoted as $\mathcal{N}_i$). We define $\mathcal{A}_i = \mathcal{P}_i \cup \mathcal{N}_i \backslash z_i$. We propose a new surrogate loss to improve the refinement as follows:

Definition 2 (Proposed refinement loss)**.**

$$\mathcal{L}_{\text{ref}} = \sum_{i \in \mathcal{I}} \left(\min_{p \in \mathcal{P}_i} \frac{1}{2} \|z_i - z_p\|^2 - \min_{n \in \mathcal{N}_i} \frac{1}{2} \|z_i - z_n\|^2 \right). \tag{3}$$

Next, we show in the section below that the proposed refinement loss is a lower bound to the following supervised contrastive loss Khosla et al. (2020):

$$\mathcal{L}_{SC} = \sum_{i \in \mathcal{I}} -\frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \left(\frac{\exp(z_i \cdot z_p)/\tau}{\sum_{a \in \mathcal{A}_i} \exp(z_i \cdot z_a)/\tau} \right). \tag{4}$$

Here, $|\mathcal{P}_i|$ denotes the size of set $\mathcal{P}_i$. Now we state an important result required to establish the relationship between $\mathcal{L}_{\text{SC}}$ and $\mathcal{L}_{\text{ref}}$,

The cosine similarity between two d-dimensional vectors z_1 and z_2 with unit ℓ_2 norm, can be written as $z_1 \cdot z_2 \triangleq \frac{1}{2}(2 - \|z_1 - z_2\|^2)$.

Proof. Since we normalize the embedding vectors to lie on a unit hypersphere, we have $\|z_1\|^2 = \|z_2\|^2 = 1$ by definition. Rewriting $z_1 \cdot z_2 = \frac{1}{2}(2 - 2 + 2z_1 \cdot z_2)$ and using the expansion for $(z_1 - z_2) \cdot (z_1 - z_2)$, we have $z_1 \cdot z_2 = \frac{1}{2}(2 - \|z_1\|^2 - \|z_2\|^2 + 2z_1 \cdot z_2)$. This leads to $z_1 \cdot z_2 = \frac{1}{2}(2 - \|z_1 - z_2\|^2)$. □

The supervised contrastive loss, $\mathcal{L}_{\text{SC}}$, as given by Eqn. 4, is an upper bound to the refinement loss, $\mathcal{L}_{\text{ref}}$, as in Eqn. 3. That is: $\mathcal{L}_{\text{SC}} > \mathcal{L}_{\text{ref}}$.

Proof. Without loss of generality, but to keep the exposition simpler, we prove the result for temperature $\tau = 1$ in Eqn. 4. The result for $\tau > 0$ can be shown similarly. Setting $\tau = 1$ in Eqn. 4, and using Jensen's inequality: $-\mathbb{E}[\log(X)] \geq -\log(\mathbb{E}[X])$, on the inner summation:

$$\mathcal{L}_{\text{SC}} \geq -\sum_{i \in \mathcal{I}} \log \left(\frac{\sum_{p \in \mathcal{P}_i} \exp(z_i \cdot z_p)}{|\mathcal{P}_i| \sum_{a \in \mathcal{A}_i} \exp(z_i \cdot z_a)} \right), \tag{5}$$

$$= \sum_{i \in \mathcal{I}} \log \left(\frac{\sum_{a \in \mathcal{A}_i} \exp(z_i \cdot z_a)}{\sum_{p \in \mathcal{P}_i} \exp(z_i \cdot z_p)} \right) + \sum_{i \in \mathcal{I}} \log(|\mathcal{P}_i|). \tag{6}$$

Method	Venue	Ref.	Cal.	CIFAR100-LT					CIFAR100				
				Top-1 ↑	AUC ↑	ECE ↓	SCE ↓	ACE ↓	Top-1 ↑	AUC ↑	ECE ↓	SCE ↓	ACE ↓
NLL (CE)	-	✓	✗	47.57	94.00	21.70	0.61	0.52	73.89	98.34	5.95	0.22	**0.12**
RefCal (Ours)	Ours	✓	✓	**58.46**	**96.22**	**13.79**	**0.47**	**0.44**	**76.11**	**98.36**	**5.20**	**0.22**	0.15
LS Szegedy et al. (2015)	CVPR'15	✗	✓	47.84	92.11	9.21	0.44	0.55	74.66	98.30	11.02	0.31	0.40
RefCal (Ours)	Ours	✓	✓	**58.90**	**96.65**	**8.76**	**0.40**	**0.47**	**75.81**	**99.13**	**9.15**	**0.29**	**0.25**
CE + TS Guo et al. (2019)	ICML'17	✓	✓	45.35	94.40	28.25	0.71	0.57	73.89	99.10	8.00	0.28	0.23
RefCal (Ours)	Ours	✓	✓	**58.46**	**96.85**	**7.62**	**0.41**	**0.46**	**76.11**	**99.18**	**6.62**	**0.24**	**0.18**
MMCE Kumar et al. (2018)	ICML'18	✗	✓	49.11	94.40	25.90	0.65	0.50	72.68	98.10	8.67	**0.26**	**0.13**
RefCal (Ours)	Ours	✓	✓	**55.90**	**95.03**	**7.91**	**0.54**	**0.56**	**75.04**	**98.35**	**8.43**	0.30	0.25
MixUp Thulasidasan et al. (2019)	NeurIPS'19	✗	✓	52.90	95.82	**6.10**	**0.58**	**0.54**	**78.30**	**98.54**	**5.49**	**0.25**	**0.19**
RefCal (Ours)	Ours	✓	✓	**56.35**	**96.37**	18.84	0.69	0.72	74.81	98.22	27.02	0.60	0.56
CRL Moon et al. (2020)	ICML'20	✓	✗	46.27	93.70	22.03	0.63	0.54	73.89	98.29	5.94	0.22	0.12
RefCal (Ours)	Ours	✓	✓	**58.46**	**96.22**	**13.80**	**0.47**	**0.44**	**76.11**	**98.36**	**5.21**	**0.22**	**0.12**
FL + MDCA Hebbalaguppe et al. (2022b)	CVPR'22	✗	✓	46.17	94.16	11.32	0.52	0.48	73.46	**98.34**	5.71	**0.22**	0.14
RefCal (Ours)	Ours	✓	✓	**58.70**	**96.23**	**10.81**	**0.45**	**0.42**	**75.86**	98.29	**5.25**	**0.22**	**0.16**
AdaFocal Ghosh et al. (2022)	NeurIPS'22	✗	✓	47.68	95.49	29.00	0.72	0.55	68.49	98.70	16.04	0.22	0.14
RefCal (Ours)	Ours	✓	✓	**58.05**	**96.62**	**8.41**	**0.45**	**0.44**	**76.20**	**99.10**	**5.40**	**0.22**	**0.14**
MbLS Liu et al. (2022)	CVPR'22	✗	✓	48.10	93.00	**8.36**	0.45	0.47	75.92	98.46	**4.80**	**0.20**	**0.13**
RefCal (Ours)	Ours	✓	✓	**58.79**	**96.62**	8.60	**0.41**	**0.47**	**75.93**	**99.11**	8.46	0.28	0.23
LogitNorm Wei et al. (2022)	ICML'22	✓	✗	52.96	94.74	49.84	**0.13**	1.30	69.89	97.32	67.80	0.04	1.59
RefCal (Ours)	Ours	✓	✓	**55.89**	**96.30**	**25.97**	0.68	0.97	**72.10**	**97.55**	**19.63**	0.43	**0.19**

Table 1: Comparison of refinement and calibration performance on CIFAR100-LT and CIFAR100 using ResNet-50: RefCal (Ours) is our proposed method. We report `Top-1`, `AUC`, and calibration metrics: `ECE`, `SCE`, `ACE`. Bold entries denote the best scores. 'Ref' = refinement method; 'Cal' = calibration method. CIFAR100-LT is used with imbalance factor 0.1. All models selected based on best Top-1 accuracy overaged over 3 runs. Calibration metrics follow conventions used in prior work: 15 bins for ECE/SCE and adaptive binning for `ACE`. `NLL`: Negative log likelihood (cross-entropy loss). For all methods, we have used the code provided by the authors. All models used for inference were chosen based on best top 1% accuracy as per norm.

Splitting $\mathcal{A}_i$ into $\mathcal{P}_i$ and $\mathcal{N}_i$:

$$\mathcal{L}_{\text{SC}} \geq \sum_{i\in\mathcal{I}} \log\left(1 + \frac{\sum_{n\in\mathcal{N}_i}\exp(z_i\cdot z_n)}{\sum_{p\in\mathcal{P}_i}\exp(z_i\cdot z_p)}\right) + \sum_{i\in\mathcal{I}}\log(|\mathcal{P}_i|). \tag{7}$$

Since $\log(1+x) > \log(x)$ for $x > 0$

$$\mathcal{L}_{\text{SC}} > \sum_{i\in\mathcal{I}} \log\left(\frac{\sum_{n\in\mathcal{N}_i}\exp(z_i\cdot z_n)}{\sum_{p\in\mathcal{P}_i}\exp(z_i\cdot z_p)}\right) + \sum_{i\in\mathcal{I}}\log(|\mathcal{P}_i|). \tag{8}$$

$$\mathcal{L}_{\text{SC}} > \sum_{i\in\mathcal{I}} \left(\log\sum_{n\in\mathcal{N}_i}\exp(z_i\cdot z_n) - \log\sum_{p\in\mathcal{P}_i}\exp(z_i\cdot z_p)\right) + \sum_{i\in\mathcal{I}}\log(|\mathcal{P}_i|).$$

Using the inequality: $\max_i(x_i) \leq \log\sum_{i=1}^m \exp(x_i) \leq \max_i(x_i) + m$, we get:

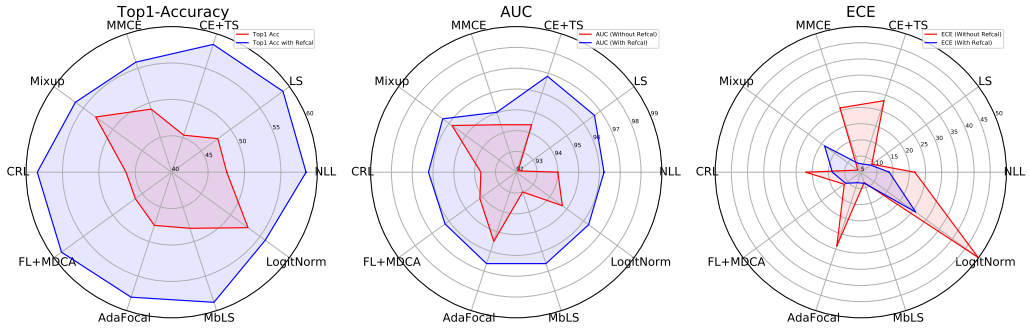
$$\mathcal{L}_{\text{SC}} > \sum_{i\in\mathcal{I}} \left(\max_{n\in\mathcal{N}_i}(z_i\cdot z_n) - \max_{p\in\mathcal{P}_i}(z_i\cdot z_p) - \log|\mathcal{P}_i|\right) + \sum_{i\in\mathcal{I}}\log|\mathcal{P}_i| \tag{9}$$

$$> \sum_{i\in\mathcal{I}} \left(\max_{n\in\mathcal{N}_i}\tfrac{1}{2}(2 - \|z_i - z_n\|^2) - \max_{p\in\mathcal{P}_i}\tfrac{1}{2}(2 - \|z_i - z_p\|^2)\right) \tag{10}$$

$$> \sum_{i\in\mathcal{I}} \left(\min_{p\in\mathcal{P}_i}\tfrac{1}{2}\|z_i - z_p\|^2 - \min_{n\in\mathcal{N}_i}\tfrac{1}{2}\|z_i - z_n\|^2\right). \tag{11}$$

where we use Lemma 3.1 to replace $z_i\cdot z_n$ and $z_i\cdot z_p$. Hence $\mathcal{L}_{\text{SC}} > \mathcal{L}_{\text{ref}}$.

□



(a) Top 1% Accuracy (b) AUC (c) ECE

Figure 2: Effect of RefCal Training: Spider plot comparing Top-1% Accuracy, AUC, and ECE with (Blue) and without (Red) RefCal training when ResNet-50 features extractor is used on CIFAR100-LT dataset. RefCal offers the best Top1% accuracy, AUC, and ECE in majority cases.

4 Optimizing for Proposed Refinement Loss

Geometric Interpretation of Refinement Loss: The refinement loss $\mathcal{L}_{\text{ref}}$ in Eqn. 3 promotes class-wise clustering by simultaneously: (a) Pulling the nearest positive neighbor closer (via the first term), thus tightening intra-class clusters; (b) Pushing the nearest negative neighbor farther (via the second term), widening inter-class margins. Because the margin between a sample and its nearest negative correlates with classifier performance, $\mathcal{L}_{\text{ref}}$ acts as a meaningful proxy for refinement. Furthermore, since $\mathcal{L}_{\text{SC}}$ upper-bounds $\mathcal{L}_{\text{ref}}$ (cf. Lemma 3.1), we minimize $\mathcal{L}_{\text{SC}}$ using the supervised contrastive loss proposed in Khosla et al. (2020) as a tractable surrogate.

On the Refinement Capabilities of Supervised Contrastive Loss: The central contribution of this paper lies in introducing a refinement loss whose effectiveness is validated by improved AUC scores as shown in the results. While supervised contrastive (SC) loss Khosla et al. (2020) is widely known for enhancing accuracy, we provide the first theoretical and empirical evidence that it also improves refinement—by showing that it upper bounds our proposed refinement loss. This connection enables the repurposing of SC optimization techniques for refinement tasks. Although we do not propose a novel calibration loss, our ability to leverage existing SC and calibration frameworks for refinement constitutes a key strength of this work.

4.1 RefCal Training Regime

We propose a training strategy to develop reliable DNN classifiers that simultaneously exhibit high refinement, high accuracy, and strong calibration (i.e., low calibration error). It consists of two stages: (a) **Refinement Stage:** We pretrain an encoder with supervised contrastive loss Khosla et al. (2020) to approximate our refinement loss, promoting discriminative and robust class-boundary representations. (b) **Calibration Stage:** Keeping the encoder frozen, we fine-tune a linear classifier head using a combination of standard classification loss and calibration-specific losses. We experiment with multiple calibration losses (see Tab. 1) to assess their impact. The resulting models demonstrate strong performance across all three axes: refinement, calibration, and accuracy. A detailed comparison using various loss combinations is reported in Tabs. 1,3.

5 Experimental Design

Datasets. We utilize well-established classification datasets: CIFAR10 Krizhevsky & Hinton (2009), CIFAR100, CIFAR100-LTLiu et al. (2019), and large-scale TinyImageNet Le & Yang (2015), ImageNet-1KDeng et al. (2009) and ImageNet-LT Liu et al. (2019). To evaluate the robustness of RefCal, we also present results on CIFAR100-C Hendrycks & Dietterich (2018) with varying degrees of corruption. For binary classification, we construct binary benchmarks from multi-class: CIFAR10 Krizhevsky & Hinton (2009),STL10 Coates et al. (2011).

Method	(a) ResNet18 on Imagenet-LT						(b) EfficientViT-M1 on CIFAR100					
	Top1 (%) ↑	AUC ↑	ECE (%) ↓	SCE (%) ↓	ACE(%) ↓	smECE(%) ↓	Top1 (%) ↑	AUC ↑	ECE (%) ↓	SCE (%) ↓	ACE(%) ↓	smECE(%) ↓
NLL (CE)	41.76	**98.30**	16.11	00.07	00.06	15.98	47.77	95.07	29.53	00.71	00.42	27.13
RefCal (Ours)	**42.18**	98.10	**10.82**	**00.05**	**00.04**	**10.38**	**56.78**	**96.00**	**09.67**	**00.28**	**00.19**	**09.17**
LS Szegedy et al. (2015)	41.73	97.70	**06.06**	00.05	00.06	**06.02**	48.97	92.79	**04.11**	00.30	00.25	**04.06**
RefCal (Ours)	**42.25**	**98.10**	08.46	**00.05**	**00.04**	08.19	**56.83**	**96.20**	08.32	**00.25**	**00.19**	07.90
CE+TS Guo et al. (2019)	41.76	**98.30**	09.13	00.06	00.05	08.98	47.77	95.40	15.25	00.43	00.35	14.99
RefCal (Ours)	**42.18**	98.20	**06.14**	**00.05**	**00.04**	**06.00**	**56.78**	**96.30**	**07.82**	**00.25**	**00.26**	**07.55**
MMCE Kumar et al. (2018)	41.89	98.10	15.91	00.07	00.06	15.78	48.70	95.11	28.82	00.70	00.42	26.61
RefCal (Ours)	**42.22**	**98.10**	**10.62**	**00.05**	**00.04**	**10.25**	**56.74**	**96.00**	**09.61**	**00.28**	**00.19**	**09.11**
CRL Moon et al. (2020)	41.61	98.10	16.08	00.07	00.06	15.95	49.07	95.11	28.53	00.69	00.41	26.45
RefCal (Ours)	**42.18**	**98.10**	**10.82**	**00.05**	**00.04**	**10.38**	**56.75**	**96.10**	**09.24**	**00.27**	**00.18**	**08.79**
FL+ MDCA Hebbalaguppe et al. (2022b)	40.68	**98.40**	08.04	00.06	00.06	08.03	47.25	95.25	17.76	00.51	00.37	17.69
RefCal (Ours)	**42.19**	98.20	**07.30**	**00.05**	**00.04**	**07.07**	**56.69**	**96.60**	**05.90**	**00.24**	**00.20**	**05.59**
AdaFocal Ghosh et al. (2022)	31.18	95.60	34.01	00.10	00.07	30.26	52.39	**96.63**	**14.35**	**00.43**	**00.29**	**14.36**
RefCal (Ours)	**41.83**	**97.80**	**19.25**	**00.06**	**00.04**	**18.83**	**53.10**	93.90	23.30	00.57	00.43	21.42
MbLS Liu et al. (2022)	41.14	97.70	**04.18**	00.06	00.05	**04.16**	49.90	93.85	08.78	00.33	00.25	08.78
RefCal (Ours)	**42.23**	**98.10**	08.63	**00.05**	**00.04**	08.30	**56.82**	**96.20**	**08.50**	**00.26**	**00.19**	**08.07**
LogitNorm Wei et al. (2022)	42.00	96.00	41.81	00.01	00.16	27.43	44.61	94.68	**16.38**	**00.47**	00.35	**16.36**
RefCal (Ours)	**42.04**	**96.30**	**41.90**	**00.01**	**00.16**	**25.77**	**55.93**	**95.30**	25.93	00.60	**00.34**	22.92

Table 2: (a) [Large Scale Experiments]:Comparison of reliability metrics of RefCal (Ours) vs. SOTA on ImageNet-LT using ResNet18. (b) [Experiments with Visual Transformers]: Comparison of reliability metrics of RefCal (Ours) vs. SOTA on Cifar-100 using Memory Efficient Vision Transformer architecture EfficientViT-M1 as feature extractor.

Baseline approaches. In our analysis, we incorporate several baseline methods that serve as a basis for comparison. These include models trained using Negative Log Likelihood NLL (also called Cross Entropy loss- CE), LS Szegedy et al. (2015), temperature scaling on top of CE, TS+CE, MixUpThulasidasan et al. (2019), Adafocal Ghosh et al. (2022), MMCE Kumar et al. (2018), CRL Moon et al. (2020), MDCA Hebbalaguppe et al. (2022b), MbLS Liu et al. (2022), LogitNorm Wei et al. (2022), and AUCM Yuan (2021).

Metrics. AUC: The area under the receiver operating characteristic curve measures separability between the classes and is a proxy for refinement. For measuring calibration, we use Expected Calibration Error (ECE), Static Calibration Error (SCE), and Adaptive Calibration Error (ACE), Smooth Calibration Error (smECE).

Implementation details. We implemented a multi-stage training methodology as outlined in Khosla et al. (2020), this is used to minimize our surrogate loss for refinement. The training process involves two phases. In the first stage, the emphasis is on feature extraction, which is achieved using a ResNet-50 He et al. (2016)/EfficientViT-M1Cai et al. (2022) network trained with the supervised contrastive loss for 1000 epochs. Subsequently, after the initial training phase, we freeze the parameters of the Stage 1 network, and introduce a classifier, constituting the second stage of our training procedure. In this second stage, a linear classifier was trained using a combination of classification and calibration losses for an additional 100 epochs.

Relevance and Applications:

Although we evaluate on moderate-scale datasets, the considered architectures (ResNet, Efficient-ViT, and MobileNet) are widely used in real-world vision systems ranging from medical imaging to edge deployment. Thus, our findings are directly relevant to practical deployment scenarios where model reliability and confidence calibration are critical. Future work will explore fine-tuning VLMs under the RefCal regime to enhance reliability.

6 RESULTS

6.1 MULTI-CLASS CLASSIFICATION

We conduct experiments on 8 datasets (ImageNet-1K, CIFAR100, ImageNet-LT, CIFAR100-LT, STL10, TinyImageNet, CIFAR100-C (Corrupted CIFAR100)) and CIFAR10). The results for the first 5 datasets are in the main text while the supplementary consists of TinyImageNet,

Method	Top-1 (%) ↑	AUROC ↑	ECE ↓	SCE ↓	ACE ↓	smECE ↓
NLL (CE)	96.71	99.37	1.21	1.64	1.47	1.20
RefCal (Ours)	**98.69**	**99.80**	**1.13**	**1.11**	**1.02**	**1.00**
LS Szegedy et al. (2015)	96.64	98.92	9.23	9.25	9.23	9.32
RefCal (Ours)	**98.58**	**99.90**	**3.02**	**3.21**	**2.97**	**3.03**
CE + TS Guo et al. (2019)	96.71	99.40	0.41	1.21	1.16	0.66
RefCal (Ours)	**98.69**	**99.90**	**0.26**	**0.51**	**0.46**	**0.47**
MMCE Kumar et al. (2018)	96.19	99.30	1.22	1.38	1.19	1.22
RefCal (Ours)	**98.74**	**99.88**	**1.13**	**1.16**	**1.00**	**0.98**
MixUp Thulasidasan et al. (2019)	96.58	99.36	6.01	6.30	6.26	6.04
RefCal (Ours)	**98.58**	**99.90**	**3.16**	**3.17**	**3.13**	**3.16**
CRL Moon et al. (2020)	96.88	99.47	1.46	1.45	1.40	1.50
RefCal (Ours)	**98.69**	**99.88**	**1.13**	**1.12**	**1.02**	**1.01**
AUCM Yuan (2021)	95.49	98.86	3.04	3.11	2.97	2.76
RefCal (Ours)	**98.56**	**99.90**	**0.56**	**0.81**	**0.82**	**0.64**
FL + MDCA Hebbalaguppe et al. (2022b)	95.49	98.23	**0.93**	1.97	1.98	**0.98**
RefCal (Ours)	**98.68**	**99.35**	1.30	**1.28**	**1.12**	**0.80**
AdaFocal Ghosh et al. (2022)	96.16	**99.11**	**1.10**	1.10	1.03	1.06
RefCal (Ours)	**98.69**	98.68	1.31	**0.67**	**0.64**	**0.67**
MbLS Liu et al. (2022)	96.98	99.42	1.09	1.18	0.85	1.15
RefCal (Ours)	**98.71**	**99.91**	**0.82**	**0.86**	**0.64**	**0.85**

Table 3: Comparison of reliability metrics for bi-

Method	CIFAR-10 FPR@TPR95 ↓	Det Error ↓	AUROC ↑	AUPR In ↑	AUPR Out ↑
NLL (CE)	30.50	08.80	95.90	96.90	94.70
+ RefCal (Ours)	**16.30**	**07.10**	**97.50**	**97.70**	**97.00**
LS Szegedy et al. (2015)	29.30	09.30	92.60	86.80	92.40
+ RefCal (Ours)	**17.00**	**08.20**	**97.10**	**97.60**	**96.30**
CE+TS Guo et al. (2019)	22.20	08.10	97.00	97.60	96.40
+ RefCal (Ours)	**14.10**	**06.80**	**97.90**	**98.20**	**97.60**
MMCE Kumar et al. (2018)	**12.10**	**06.90**	**98.30**	**98.40**	**98.20**
+ RefCal (Ours)	20.80	10.30	97.30	85.80	95.60
CRL Moon et al. (2020)	27.90	08.20	96.30	97.20	95.30
+ RefCal (Ours)	**16.30**	**07.10**	**97.50**	**97.70**	**97.00**
AUCM Yuan (2021)	38.30	12.10	93.40	93.60	92.40
+ RefCal (Ours)	**10.20**	**06.60**	**98.20**	**98.50**	**98.00**
FL+ MDCA Hebbalaguppe et al. (2022b)	20.50	08.00	97.10	**97.70**	**96.60**
+ RefCal (Ours)	**18.00**	**07.20**	**97.30**	97.40	96.50
MbLS Liu et al. (2022)	25.00	08.20	96.40	97.10	95.10
+ RefCal (Ours)	**16.90**	**07.80**	**97.20**	**97.70**	**96.60**

Table 4: [**Robustness to OOD data**]: Comparison of reliability metrics of RefCal (Ours) vs. SOTA. We use ResNet-50 backbone on in-distribution dataset CIFAR10 and test on OOD dataset SVHN.

CIFAR10 and CIFAR100-C results. Tab. 1 shows that training with RefCal achieves a significant improvement in refinement as measured by AUC and a reduction in calibration error (ECE, SCE, ACE) while also providing an accuracy increase. To understand the benefits of RefCal, observe the performance of our approach to non-calibrated/directly calibrated baseline models. Notice a jump in accuracy by over 10% in every case, similarly on AUC we do see a consistent increase. The reduction in calibration error (ECE and smECE) is also seen in majority of the cases.

Fig. 3 presents the error bars for 3 runs and comparative study highlighting AUC vs. calibration trade-offs associated with existing techniques and RefCal (Top-left is the most desirable location on the chart suggesting higher AUC and lower calibration error). Specifically, we found the mean and one standard scatter error for AUC and ECE plot shows the lower variances in case of RefCal variants emphasizing its reliability in comparison with SOTA. Note the variant RefCal +CE+TS (ours) offers the highest AUC-low calibration error with the next best being AUCMYuan (2021). We also compare Top 1% accuracy vs. calibration trade-off associated with contemporary techniques and RefCal. The mean and one standard scatter error bars for Top 1% accuracy and ECE reveal that RefCal variants not only offer lower variance but also reasonably good top 1% accuracy and calibration error trade-off.

Performance on various feature extractors and large scale datasets. RefCal continues to perform superior on ImageNet-LT/ImageNet and even using a different backbone architecture EfficientViT-M1Cai et al. (2022) as evidenced in the Tab. 2 in majority of the cases. Fig. 2 shows the spider plot illustrating that applying RefCal (blue) consistently improves Top1% Accuracy and AUC across different calibration methods compared to models without RefCal (red). Additionally, ECE is significantly reduced with RefCal, indicating better model reliability.

6.2 BINARY CLASSIFICATION

For binary classification, we employ multi-class classification datasets, transforming them into binary classification datasets. The categorization of classes within these datasets is done by grouping semantically relevant classes together. We construct binary benchmark dataset: STL10 Cao et al. (2019) following the settings mentioned in Yuan (2021). Tab. 3 reports results on binary classification. We outperform/are comparable to the SOTA on binary classification tasks in Top 1% accuracy, ECE, and AUC.

6.3 ROBUSTNESS OF **REFCAL**

Robustness to out-of-distribution (OOD) samples (semantic shift). A well-calibrated model should not only exhibit low confidence whenever it misclassifies but also in situations when it en-

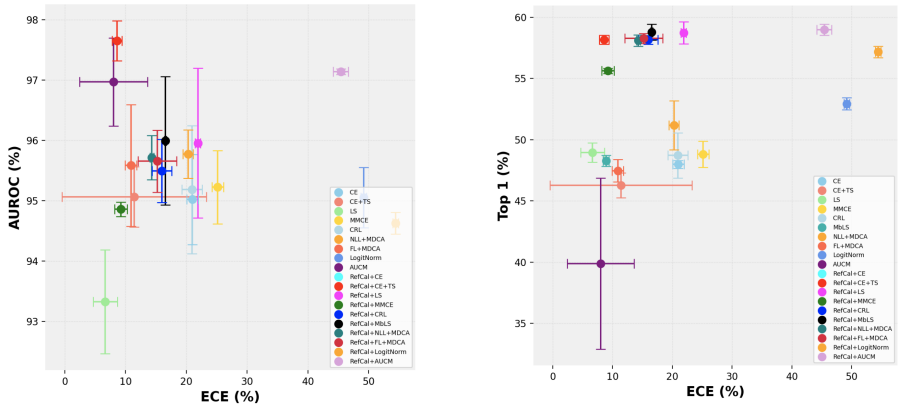


Figure 3: (Left) AUC vs. ECE trade-off and (Right) Top 1% accuracy vs. ECE trade-off for ResNet-50 on CIFAR100-LT (IF=10%). Higher AUC and Top 1% accuracy with lower ECE (top-left) are desirable. Lower variance in AUC and ECE highlights the reliability of RefCal variants.

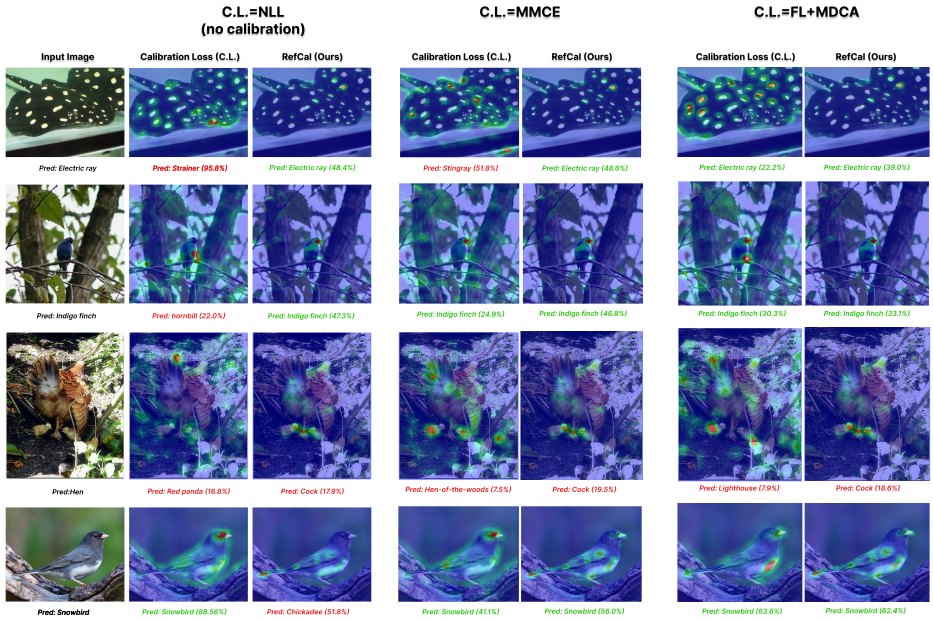


Figure 4: **[Grad-CAM]**: Our proposed training regime, RefCal allows for joint optimization of calibration and refinement. Grad-CAM visualizations for Resnet-18 trained on ImageNet-LT, using a particular calibration technique (column bearing title Calibration Loss (C.L.)), and then by jointly optimizing with the same calibration but adding our refinement loss (columns bearing title RefCal i.e., columns 3,5, and 7). Here, the Calibration Loss tested are: NLL (no calibration) in column 2, MMCEKumar et al. (2018) in column 4, and FL+MDCAHebbalaguppe et al. (2022a) in column 6. **Note:** Numbers in parentheses indicate predictive confidence for each method. Since, the refinement loss forces a model to separate its confidence for positive, and negative samples, it also leads the model to focus its attention on the salient object features, instead of the background.

Method	CIFAR-100C Top1 (%) ↑	AUROC ↑	ECE (%) ↓	SCE (%) ↓ $\times 10^{-2}$	ACE(%) ↓ $\times 10^{-2}$	smECE(%) ↓ $\times 10^{-2}$
NLL (CE)	24.81	84.08	24.73	00.80	00.79	23.84
+ RefCal (Ours)	**32.97**	**91.72**	**16.93**	**00.73**	**00.71**	**16.79**
LS Szegedy et al. (2015)	23.53	80.08	**03.36**	**00.35**	**00.49**	**03.32**
+ RefCal (Ours)	**30.01**	**91.75**	14.86	00.56	00.93	14.78
CE+TS Guo et al. (2019)	24.81	85.09	05.30	00.57	00.62	05.22
+ RefCal (Ours)	**32.97**	**92.54**	**02.17**	**00.59**	**00.69**	**02.18**
MMCE Kumar et al. (2018)	23.54	84.15	24.85	00.77	00.78	23.90
+ RefCal (Ours)	**32.22**	**93.21**	**02.00**	**00.65**	**00.76**	**02.00**
CRL Moon et al. (2020)	23.34	86.62	**07.21**	**00.62**	**00.64**	**07.11**
+ RefCal (Ours)	**32.96**	**91.72**	16.94	00.73	00.71	16.79
AUCM Yuan (2021)	19.02	83.71	**05.75**	00.68	**00.68**	**05.57**
+ RefCal (Ours)	**31.36**	**92.95**	29.47	**00.08**	01.34	27.36
FL+ MDCA Hebbalaguppe et al. (2022b)	22.83	85.03	11.90	**00.69**	00.72	11.89
+ RefCal (Ours)	**32.25**	**91.98**	**10.98**	**00.69**	**00.70**	**10.96**
AdaFocal Ghosh et al. (2022)	**34.76**	91.69	37.89	00.94	00.82	32.83
+ RefCal (Ours)	32.43	**92.61**	**07.78**	**00.68**	**00.70**	**07.78**
MbLS Liu et al. (2022)	23.47	0.80	19.12	00.64	**00.66**	18.97
+ RefCal (Ours)	**29.95**	**92.11**	**09.30**	**00.62**	00.85	**09.29**
LogitNorm Wei et al. (2022)	25.56	84.34	**24.03**	00.05	**01.01**	**23.18**
+ RefCal (Ours)	**33.93**	**90.67**	32.61	**00.03**	01.28	29.55

Table 5: [**Robustness to corruptions]: Comparison of reliability metrics of RefCal (Ours) vs. SOTA**]: We use ResNet-50 He et al. (2016) backbone on CIFAR100C Hendrycks & Dietterich (2018) The results are averaged with 5 degrees of severity of Gaussian noise. We observe that RefCal scores are consistently on-par/superior over train-time calibrators and refinement methods.

counters data that belongs to a new/different class different than the training classes. We investigate if RefCal can acquire representations that demonstrate increased resilience to OOD samples by assigning them low confidence. Tab. 4 summarizes the OOD detection performance of our model and highlights the capability to reject such unknown samples by increasing AUC between the in-distribution and OOD samples and decreasing both FPR@0.95TPR, and Detection Error.

Robustness to Natural corruptions. It is indeed encouraging to note that models trained on RefCal regime are not only better calibrated, refined, and accurate, but also seem to perform well when the test distribution is CIFAR100C (corrupted CIFAR 100 datasetHendrycks & Dietterich (2018) when the model is trained on CIFAR100 dataset. Tab. 5 (a) indicates higher reliability even in the case of natural corruptions.

Grad-CAM visualization. Figs. 1 and 4 compares Grad-CAM visualizations of RefCal and contemporary methods showing that refinement loss shifts attention from background to salient object features by separating confidence for correct and incorrect predictions.

Limitations. We observe that combining our loss $\mathcal{L}_{\text{ref}}$ with calibration methods such as AUCM Yuan (2021), MixUp Zhang et al. (2018), and Adafocal Ghosh et al. (2022) can be challenging on some datasets. Although $\mathcal{L}_{\text{ref}}$ provably improves refinement, joint optimization with multiple losses via SGD introduces complex interactions. We conjecture that some calibration methods produce representations more compatible with $\mathcal{L}_{\text{ref}}$ than others, which we defer to future work. Our approach uses SupCon Khosla et al. (2020) during training, incurring higher training cost than post-hoc calibration, but yielding substantial reliability gains with no additional inference overhead.

7 CONCLUSIONS

This paper demonstrates that existing calibration methods can improve calibration metrics while degrading prediction refinement. To address this, we propose a novel surrogate loss for refinement. We theoretically show that supervised contrastive loss upper-bounds our refinement loss, making it a suitable surrogate for minimization. Building on this, we introduce RefCal, a training framework that optimizes refinement, calibration, and accuracy. Experiments on standard classification benchmarks show that our approach outperforms or comparable to the SOTA methods in producing accurate, well-calibrated, and refined predictions, thereby enhancing the classification reliability.

REFERENCES

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *ACM STOC*, 2023.

Han Cai, Chuang Gan, and Song Han. Efficientvit: Enhanced linear attention for high-resolution low-computation visual recognition. *arXiv preprint arXiv:2205.14756*, 2022.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32, 2019.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Morris H DeGroot and Stephen E Fienberg. Assessing probability assessors: calibration and refinement. *Statistical decision theory and related topics III*, 1:291–314, 1982.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE CVPR*, pp. 248–255. IEEE, 2009.

Zhipeng Ding, Xu Han, Peirong Liu, and Marc Niethammer. Local temperature scaling for probability calibration. *CoRR*, abs/2008.05105, 2020.

Soumya Suvra Ghosal, Ramya Hebbalaguppe, and Dinesh Manocha. Better features, better calibration: A simple fix fornbsp;overconfident networks. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2025, Porto, Portugal, September 15–19, 2025, Proceedings, Part I*, pp. 231–247, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-05961-1. doi: 10.1007/978-3-032-05962-8_14. URL https://doi.org/10.1007/978-3-032-05962-8_14.

A. Ghosh, T. Schaaf, and M. Gormley. Adafocal: Calibration-aware adaptive focal loss. In *Advances in NeurIPS*, volume 35, pp. 1583–1595, 2022.

S. Gruber and F. Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in NeurIPS*, 35:8618–8632, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. *ICML*, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR*, pp. 770–778, 2016.

R. Hebbalaguppe, S. Ghosal, J. Prakash, H. Khadilkar, and C. Arora. A novel data augmentation technique for out-of-distribution sample detection using compounded corruptions. In *ECML*, pp. 529–545. Springer, 2022a.

R. Hebbalaguppe, M. Baranwal, K. Anand, and C. Arora. Calibration transfer via knowledge distillation. In *Proceedings of the ACCV*, pp. 513–530, 2024.

Ra. Hebbalaguppe, J. Prakash, N. Madan, and C. Arora. A stitch in time saves nine: A train-time regularizing loss for improved neural network calibration. In *Proceedings of the IEEE/CVF CVPR*, pp. 16081–16090, 2022b.

Ramya Hebbalaguppe, Tamoghno Kandar, Abhinav Nagpal, and Chetan Arora. Prompting without panic: Attribute-aware, zero-shot, test-time calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 289–305, 2025.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2018.

P. Khosla, P. Teterwak, and C. *et al.* Wang. Supervised contrastive learning. *Advances in NeurIPS*, 33:18661–18673, 2020.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

A. Kumar, S. Sarawagi, and U. Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th ICML*, Proceedings of Machine Learning Research, pp. 2805–2814, 2018.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. *arXiv preprint arXiv:1909.10155*, 2019.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

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

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE CVPR*, 2019.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *international conference on machine learning*, pp. 7034–7044. PMLR, 2020.

J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, and P. Dokania. Calibrating dnns using focal loss. *Advances in NeurIPS*, 33:15288–15299, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *arXiv preprint arXiv:1906.02629*, 2019.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

R. Patra, R. Hebbalaguppe, T. Dash, G. Shroff, and L. Vig. Calibrating deep neural networks using explicit regularisation and dynamic data pruning. In *Proceedings of the IEEE/CVF WACV*, pp. 1541–1549, January 2023.

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

Mrinal Rawat, Ramya Hebbalaguppe, and Lovekesh Vig. Pnpood : Out-of-distribution detection for text classification via plug andplay data augmentation. In *2021 International Conference on Machine learning (Workshop on Uncertainty & Robustness in Deep Learning)*, 2021.

A. Singh, A. Bay, B. Sengupta, and A. Mirabile. On the dark side of calibration for modern neural networks. *ICML-W*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, 2015.

S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *NeurIPS*, 2019.

H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022.

Tianbao Yang. Algorithmic foundation of deep x-risk optimization. *arXiv preprint arXiv:2206.00439*, 2022.

Zhiyong Yang, Qianqian Xu, Shilong Bao, Xiaochun Cao, and Qingming Huang. Learning with multiclass auc: Theory and algorithms. *IEEE Trans. PAMI*, 44(11):7747–7763, 2021.

Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, and Qingming Huang. Optimizing two-way partial auc with an end-to-end framework. *IEEE Trans. PAMI*, 45(8):10228–10246, 2023.

Z. et al.. Yuan. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *IEEE ICCV*, 2021.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.