DARTS: Distance-Aware Robust Training for Selective Classification

A.Q.M. Sazzad Sayyed*, Nathaniel D. Bastian† and Francesco Restuccia*

*Northeastern University † United Stated Military Academy

Corresponding Author: sayyed.a@northeastern.edu

Abstract

Neural networks are vulnerable to overconfidence. Selective classification addresses this by abstaining models from uncertain predictions. Existing selective classification approaches do not consider the geometry of the feature space; as such, they require an auxiliary classification head or rely on computation-heavy methods. To address this, we propose DARTS, a new distance-to-boundary-aware training approach that significantly improves performance by explicitly shaping feature-space geometry. DARTS provides a closed-form expression for the distance of an input to its nearest decision boundary to encourage correctly classified samples to lie farther from the boundary, while constraining misclassified ones to remain closer. Being geometry-based, DARTS is architecture-agnostic and improves confidence separation and robustness without additional parameters or inference overhead. Experimental results across the CIFAR, Imagenet, CIFAR-C, and Imagenet-C benchmarks with ResNet, ConvNext, and ViT architectures show that DARTS improves Area Under Risk Coverage Curve (AURC) by upto 58% compared to state-of-the-art approaches, with consistent gains of over 20% on both CIFAR100 and Imagenet. We will release our code for reproducibility.

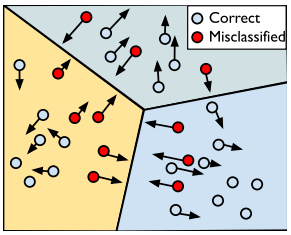
1. Introduction

Safety-critical applications require machine learning algorithms able to abstain from predictions when uncertain, rather than output a potentially harmful decision [39]. *Selective classification* allows a model to reject a subset of inputs, thereby trading coverage for higher reliability on the retained set [6, 18]. The effectiveness of selective classification depends on (1) a training objective that encourages geometrically well-separated and robust decision regions; and (2) a confidence score that reliably ranks samples by their likelihood of correctness.

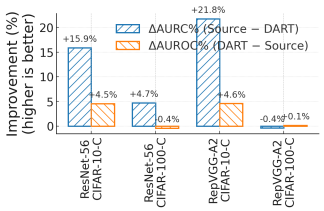
Prior Work. Chow [6] introduced the concept of classification with a reject option. Theoretical performance guarantees were then derived under various assumptions, such as realizability, Tsybakov noise conditions or margin-based separability [25]. Experimental methods then emerged in both classical and deep learning settings. Early approaches often relied on confidence thresholds derived from model outputs [17], while more recent work has explored learnable rejection mechanisms [36], uncertainty quantification via ensembles [28] or Bayesian methods [15], and game-theoretic formulations that jointly optimize prediction and rejection policies [4]. A prominent line of research focuses on *bounded-improvement* selective classifiers, which guarantee that the selective risk is never worse than that of a fixed baseline classifier while maximizing coverage [13, 46]. In the context of deep neural networks, SelectiveNet [18] trains a dedicated selection head alongside the predictor. Despite attempt to design proper confidence functions for selective classification and failure prediction [7], it has been shown by [14] that maximum softmax probability works best as a confidence metric.

Key Issues. Existing selective classification methods face two critical limitations. First, post-hoc approaches such as softmax entropy [17] or Monte Carlo dropout variance [15] do not explicitly account for the geometry of the decision boundary or the model's robustness in local neighborhoods. Consequently, confidence estimates can be poorly calibrated, especially near ambiguous or adversarially vulnerable regions [2, 41]. Ultimately, this leads to suboptimal risk-coverage trade-offs. Second, training procedures for selective classifiers often treat prediction and selection as decoupled objectives [16] or optimize them using surrogate losses [5, 8] that do not directly encourage well-separated decision boundaries with reliable confidence margins.

Core Innovation. We propose Distance-Aware Robust Training for Selective classification (DARTS), a training framework that explicitly shapes the *feature-space geometry* of a classifier through class-dependent margin regularization. Figure 1a overviews DARTS. In stark opposition with adversarial training methods that rely on input-space perturbations [26], DARTS operates in the *feature space*. By leveraging the closed-form distance in feature space between an input and decision boundaries of class pairs, DARTS enforces asymmetric geometric constraints that separate confidently correct predictions from uncertain or incorrect ones. Specifically, we treat correctly and incorrectly classified samples in a complementary fashion. We enforce a minimum margin from the nearest decision boundary (computed over top rival classes). Samples whose



(a) Conceptual intuition. DARTS encourages margin-aware feature geometry aligned with decision boundaries.



(b) Percent improvement. Percent improvement. ΔAURC% = (Source-DARTS)/Source; ΔAUROC% = (DARTS-Source)/Source.

Figure 1. (a) *Decision-aware training* shapes feature–boundary margins, reducing overlap between correct and incorrect confidence. (b) Across architectures (ResNet-56, RepVGG-A2) and datasets (CIFAR-10-C/100-C), DARTS yields consistent *percentage* improvements: lower selective risk (ΔAURC% ↑) and higher discriminability (ΔAUROC% ↑).

features lie too close to the decision boundary of the competing classes are penalized, thus improving robustness and calibration. For misclassified samples, rather than pushing them away from their incorrect prediction, we push them to approach and remain near the decision boundary (i.e., the hyperplane separating the incorrectly predicted and nearest class) within a maximum margin. This prevents overconfident errors and ensures that ambiguous samples remain near rejectable boundaries.

DARTS's approach follows a warm-up phase of standard cross-entropy training. A curriculum schedule gradually increases the penalty on misclassified samples, stabilizing optimization and aligning the learned geometry with confidence behavior. Importantly, DARTS's computation depends only on penultimate layer features, classifier weights, and classifier bias. As such, DARTS requires no auxiliary networks, additional parameters, or input-space perturbations, thus reducing the computational burden of training for selective classification. Moreover, being distance-based, DARTS is a general distance-based approach that is applicable to any neural network classifier, including convolutional- and transformer-based architectures.

Summary of Novel Contributions

• We introduce DARTS to explicitly shape the feature-space geometry of neural classifiers using closed-form distances to decision boundaries. The proposed *asymmetric margin regularization* enforces wide and stable margins for correctly classified samples while constraining misclassified ones near their top-2 decision boundary, thus improving reliability without auxiliary networks or perturbations. In addition, we propose a unified *geometric confidence measure* that is architecture-agnostic, invariant to logit scaling, and consistent with the training objective;

• We perform extensive experimental evaluation on CIFAR and Imagenet benchmarks as well as their corrupted versions with ResNet, ConvNext, and ViT architectures. We show that DARTS improves AURC by up to 58% compared to state-of-the-art approaches, with consistent gains of over 20% on both CIFAR-100 and ImageNet..

2. Related Works

Selective Classification: El-Yaniv and Wiener [13] and Wiener and El-Yaniv [46] introduce bounded-improvement selective classifiers that guarantee the selective risk never exceeds that of a fixed baseline, albeit under strong realizability assumptions. More practical variants include SelectiveNet [18], which trains an auxiliary selection head alongside the main predictor, and Deep Gamblers [34], which formulates abstention as a utility-maximization problem. However, these approaches either introduce additional network components or depend on heuristic confidence scores (e.g., softmax entropy), which often lack geometric interpretability and calibration.

Confidence and Uncertainty Quantification: Other approaches improve confidence reliability and uncertainty quantification. Classical approaches include Monte Carlo dropout [15], deep ensembles [28], and density-based methods [23, 45]. Subsequent works have proposed energy-based [31] and logit-space geometric measures such as ViM [44] and GEN [32] which focus on flagging the Out-of-Distribution (OOD) inputs. These methods are often computationally expensive (ensembles), require architectural changes (Bayesian networks), or conflate epistemic uncertainty with distributional shift. In contrast, our proposed confidence score is deterministic, parameter-free, and derived from the geometry of decision boundaries in feature space, which makes it more interpretable and efficient.

Margin-Based Score for Selective Classification Closely related to DARTS are margin-regularized representation learning methods such as Large Margin Softmax [30] and ArcFace [10], which enforce angular margins to improve discriminative power but assume correct labels and fail to handle misclassified samples. Confidence-Aware Learning [36] penalizes low-confidence correct predictions, but applies uniform penalties across samples. *DARTS* diverges from these approaches through an asymmetric margin objective: it enforces large minimum margins for correct predictions to form robust basins, while constraining misclassified samples near the top-M' decision boundary to prevent overconfident errors. This dual objective explicitly aligns geometric structure with the demands of selective classification.

Geometric Confidence Metrics. Decision-boundary distances have been used as confidence indicators in linear models and SVMs [38]. However, they are less common in deep networks due to their inherent non-linearity. [22] showed that ReLU networks yield piecewise-linear boundaries, enabling exact margin computation in special cases-though such methods are intractable for large-scale mod-

els. Our approach circumvents this issue by assuming a linear classifier head, a standard design in modern architectures (e.g., ResNets), which allows approximate yet effective boundary-distance computation using only the final weight matrix $\mathbf{W}$. This produces a confidence score that strongly correlates with correctness, leading to superior performance over softmax-based baselines with negligible computational cost.

Summary of Novelty

In stark opposition with prior work, DARTS requires no auxiliary networks, no input-space perturbations, and no post-hoc calibration. These features make it interpretable and easily integrable into existing deep learning pipelines. DARTS directly regularizes the decision geometry during training, thus unifying geometric margin shaping and confidence estimation within a single framework.

3. Problem Statement: Selective Classification

We adopt the standard formalism in literature[17, 18]. Accordingly, in *selective classification*, a predictor f is paired with a selection function $g : \mathcal{X} \to \{0, 1\}$ that decides whether to accept or abstain from the classification:

$$(f, g)(x) = \begin{cases} f(x), & g(x) = 1, \\ \text{reject}, & g(x) = 0. \end{cases} \tag{1}$$

Let $\ell(f(x), y)$ denote a per-sample loss (e.g., 0-1 or cross-entropy), and let the expectation be over $(x, y) \sim \mathcal{D}$. The *coverage* is defined as $\phi(g) = \mathbb{E}[g(x)]$, while the corresponding *selective risk* is defined as

$$R(f, g) = \frac{\mathbb{E}[\ell(f(x), y)\, g(x)]}{\phi(g)}, \tag{2}$$

which is the standard risk when $\phi(g) = 1$. The reliability/coverage trade-off is summarized by the *risk–coverage curve* (RCC) and its area under the curve (AURC) [18, 36].

Confidence Function. A *confidence function* $C : \mathcal{X} \to \mathbb{R}$ estimates the model's belief in correctness. A threshold-based selector defined as

$$g_\tau(x) = \mathbb{I}\big[C(x) \geq \tau\big] \tag{3}$$

accepts predictions above confidence level τ and abstains otherwise. Thus, an effective $C(x)$ must rank samples based on their true correctness probabilities, yielding an optimal risk–coverage trade-off.

Feature–Classifier Decomposition. Henceforth, we will consider models such as

$$f(x) = \nu(\psi(x)) = \mathbf{W}\, \mathbf{h}_\psi(x) + \mathbf{b}, \tag{4}$$

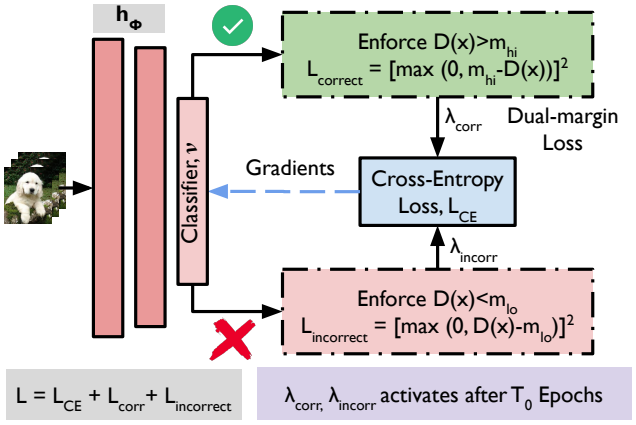


Figure 2. Training pipeline of the DARTS framework. Given an input sample x, the feature extractor h_ϕ and linear classifier ν produce logits and boundary distances $D(x)$. Correctly classified samples are pushed away from the decision boundary using a high margin m_{hi} via $\mathcal{L}_{\text{corr}}$, while misclassified samples are pulled closer with a low margin m_{lo} via $\mathcal{L}_{\text{wrong}}$. Both margin-based losses are combined with cross-entropy and activated after a warm-up phase, aligning geometric margins with selective confidence.

where $\psi : \mathcal{X} \to \mathbb{R}^d$ is a feature extractor (e.g., a CNN or ViT backbone), $\mathbf{h}_\psi(x) \in \mathbb{R}^d$ denotes the feature representation, and ν is a linear classifier parameterized by $(\mathbf{W}, \mathbf{b})$. This decomposition allows a geometric interpretation of model confidence; in other words, the decision boundaries between classes correspond to hyperplanes in the feature space, whose distances reflect the model's certainty. DARTS realizes a geometrically-grounded confidence measure which is based on distances from $\mathbf{h}_\psi(x)$ to these decision boundaries.

4. The DARTS Framework

We first introduce the estimate of geometric distance from the decision boundary in feature space in Section 4.1. Then we describe the overall loss function along with its components in Section 4.2. We conclude with a description of the inference procedure in Section 4.3.

4.1. Confidence via Decision Boundary Distance

Consider a model $f = \nu \circ \psi$, composed of a feature extractor ψ and a linear classifier ν:

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}\, \mathbf{h}_\psi(\mathbf{x}) + \mathbf{b}, \tag{5}$$

where $\mathbf{h}_\psi(x) \in \mathbb{R}^d$ denotes the penultimate-layer feature representation, and $\mathbf{W} = [\mathbf{w_1}, \ldots, \mathbf{w_K}]^\top \in \mathbb{R}^{K \times d}$ and $b \in \mathbb{R}^K$ are the classifier parameters for K classes. The decision boundary between two classes c_1 and c_2 is defined as the set of feature points h satisfying

$$\langle \mathbf{w_{c_1}}, \mathbf{h} \rangle + \mathbf{b_{c_1}} = \langle \mathbf{w_{c_2}}, \mathbf{h} \rangle + \mathbf{b_{c_2}}. \tag{6}$$

This boundary forms a hyperplane orthogonal to the vector $(\mathbf{w_{c_1}} - \mathbf{w_{c_2}})$. The signed Euclidean distance of a feature vector h to this hyperplane is therefore:

$$d(\mathbf{h}; c_1, c_2) = \frac{\langle \mathbf{w_{c_1}} - \mathbf{w_{c_2}}, \mathbf{h} \rangle + (\mathbf{b_{c_1}} - \mathbf{b_{c_2}})}{\|\mathbf{w_{c_1}} - \mathbf{w_{c_2}}\|_2}. \quad (7)$$

For input x with prediction $\hat{y} = \arg\max_c f_c(x)$, the *nearest decision boundary distance* is defined as the minimum absolute distance to any class other than the predicted:

$$D(\mathbf{x}) = \min_{c' \neq \hat{y}} \frac{|f_{\hat{y}}(x) - f_{c'}(x)|}{\|w_{\hat{y}} - w_{c'}\|_2}. \quad (8)$$

Intuitively, $D(\mathbf{x})$ measures how deeply $\mathbf{h}_\psi(x)$ lies within its predicted decision region. Unlike softmax-based confidence scores, $D(\mathbf{x})$ explicitly incorporates classifier geometry (via the pairwise weight directions) and the feature representation, yielding a geometrically interpretable normalized margin. In practice, we approximate Equation (8) by considering only the top-M' rival classes (highest logits excluding $\hat{y}$), which suffices to identify the closest boundary in high-dimensional settings. Computing $D(x)$ scales as $O(M'd)$ per sample, which is negligible compared to the cost of forward propagation. The proposed distance $D(x)$ serves as a unified geometric measure for both training regularization and for inference-time confidence estimation.

Geometric interpretation. Equation (8) reveals that each logit difference $f_{\hat{y}}(\mathbf{x}) - f_{c'}(\mathbf{x})$ acts as a proxy for the signed distance to a class-separating hyperplane, normalized by the separation between $w_{\hat{y}}$ and $w_{c'}$. Large $D(x)$ values indicate that the feature lies well within a class region, whereas small values correspond to boundary proximity and higher misclassification risk.

4.2. DARTS Training Procedure

We integrate the decision boundary distance into training via a dual-margin regularization scheme. DARTS enforces large margins for correctly classified samples and restricts overconfident errors for misclassified ones. The total loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{corr}} \cdot \mathcal{L}_{\text{corr}} + \lambda_{\text{wrong}}(t) \cdot \mathcal{L}_{\text{wrong}}, \quad (9)$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy and the two geometric regularizers are defined below.

(i) Correct-sample regularizer. For correct classifications, we encourage large distances to the decision boundary:

$$\mathcal{L}_{\text{corr}} = \frac{1}{|\mathcal{C}|} \sum_{x_b \in \mathcal{C}} \left[\max(0,\, m_{\text{hi}} - D(x_b)) \right]^2. \quad (10)$$

where $\mathcal{C}$ denotes the set of all correctly classified inputs and $D(\mathbf{x})$ represents the distance to the nearest decision boundary obtained using (8). $\mathcal{L}_{corr}$ improves margin consistency by expanding the classifier's confident regions.

(ii) Misclassified-sample regularizer. For misclassified samples ($\hat{y} \neq y$), we penalize excessively large distance to nearest decision boundary:

$$\mathcal{L}_{\text{wrong}} = \frac{1}{|\mathcal{W}|} \sum_{x_b \in \mathcal{W}} \left[\max(0,\, D(\mathbf{x}_b) - m_{\text{lo}}) \right]^2, \quad (11)$$

which decreases overconfidence by reducing the geometric margin in ambiguous regions.

Training schedule. As summarized in Algorithm 1 in supplementary SectionB, training begins with a warm-up phase (T_0 epochs) using pure cross-entropy, after which the margin-based losses are activated. Boundary constraints stabilize as the classifier geometry converges with the time-dependent weight $\lambda_{\text{wrong}}(t)$.

4.3. Inference via Nearest-Boundary Confidence

At inference, we reuse $D(x)$ as a confidence score for selective prediction. Given a test input x, we compute:

$$C(x) = \min_{c' \in \mathcal{R}} \frac{|f_{\hat{y}}(x) - f_{c'}(x)|}{\|w_{\hat{y}} - w_{c'}\|_2 + \varepsilon}, \quad (12)$$

where $\mathcal{R}$ denotes the top-M' rival classes. A prediction is accepted if $C(x) \geq \tau$ (fixed threshold) or if it belongs to the top-γ fraction of validation confidences (target coverage):

$$g(x) = \begin{cases} \hat{y}, & C(x) \geq \tau, \\ \text{reject}, & \text{otherwise}. \end{cases} \quad (13)$$

This rule naturally defines the risk-coverage trade-off curve, which we analyze in Section 7. Since $C(x)$ directly encodes the geometric proximity to decision boundaries, it provides a calibrated and interpretable confidence measure, as implemented in Algorithm 2 in supplementary SectionB.

Interpretation. DARTS aligns decision geometry with predictive confidence, i.e., correctly classified samples are repelled from class boundaries, while misclassified ones are constrained near their top-2 boundaries, reducing overconfidence. The resulting boundary-aware representation exhibits improved calibration and selective risk efficiency.

5. Theoretical Analysis

To guarantee convergence and error rate of DARTS, we analyze (i) optimization convergence of the boundary-aware objective and (ii) error-rate generalization via normalized margins. Detailed proofs are in the Supplement section.

Setup and notation. Let $f(x) = Wh_\phi(x) + b$, with $h_\phi : \mathcal{X} \to \mathbb{R}^d$ the feature extractor and (W, b) the linear head. For a labeled point (x, y), define the *normalized margin*

$$\gamma(x, y; W) = \min_{c \neq y} \frac{f_y(x) - f_c(x)}{\|w_y - w_c\|_2},$$

which coincides with the nearest decision-boundary distance in Section. 4.1. We assume (A1) $\|h_\phi(x)\|_2 \le B$ almost surely; (A2) $\|w_c\|_2 \le R$ and $\|w_y - w_c\|_2 \ge \rho > 0$; (A3) Lipschitz logits in h_ϕ and smooth losses. These are common assumptions and enforced in practice via weight decay/normalization and appropriate training.

Lemma 5.1 (SGD stability under Robbins-Monro)**.** *With step sizes $\{\eta_t\}$ s.t. $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$, stochastic gradients of*

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{corr}} \cdot \mathcal{L}_{\text{corr}} + \lambda_{\text{wrong}}(t) \cdot \mathcal{L}_{\text{wrong}}$$

have bounded variance. Partial iterates are bounded and $\mathcal{L}(W_t, b_t, \phi_t)$ converges.

Theorem 5.2 (Convergence to a margin-consistent equilibrium)**.** *Under (A1)-(A3) and Lemma 5.1, DARTS admits limit points at which*

$$\begin{aligned} \gamma_f(x, y) &\ge m_{\text{hi}} \quad \textit{for all correct } (x, y), \\ d_x &\le m_{\text{lo}} \quad \textit{for all misclassified } x. \end{aligned} \tag{14}$$

where $d_x = \frac{|z_{c_1} - z_{c_2}|}{\|w_{c_1} - w_{c_2}\|_2}$ is the top-2 normalized gap. Thus, training stabilizes a geometry with expanded safe regions *for correct points and* limited overconfidence *for errors.*

Observation. The asymmetric dual margins $(m_{\text{hi}}, m_{\text{lo}})$ and the curriculum weight $\lambda_{\text{wrong}}(t)$ directly implement the push-pull effect formalized in Theorem 5.2, aligning representation geometry with confidence (Section 4).

Theorem 5.3 (Margin-based generalization (normalized))**.** *Let $\widehat{\Pr}$ denote the empirical margin distribution on n samples and fix $\tau > 0$. With probability at least $1 - \delta$,*

$$\Pr(\text{err}(f)) \le \widehat{\Pr}\big(\gamma(x) \le \tau\big) + \mathcal{O}\left(\frac{RB\sqrt{K}}{\tau\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

If DARTS enforces $\gamma(x) \ge m_{\text{hi}}$ on a $(1 - \epsilon)$ fraction of points, setting $\tau = m_{\text{hi}}$ yields

$$\Pr(\text{err}(f)) < \epsilon + \frac{RB\sqrt{K}}{m_{\text{hi}}\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

Observation. Increasing m_{hi} tightens the bound while weight control (smaller R) improves constants. Both are explicitly encouraged by DARTS's correct-sample regularizer and standard regularization.

6. Experimental Setup

We evaluate DARTS across standard selective-classification benchmarks and architectures to assess robustness, calibration, and risk–coverage trade-offs. All experiments follow consistent training, evaluation, and metric protocols.

Datasets and Architectures. We use CIFAR-10 and CIFAR-100 [27] for small-scale evaluation and ImageNet-1k [9] for large-scale validation. To test robustness, we further report results on corrupted variants (CIFAR-10C and ImageNet-C). Following prior work [18, 23], models are trained on the standard training split and evaluated on the clean test set unless otherwise specified. We assess both convolutional and transformer-based backbones to examine the generality of DARTS 's feature-space margin regularization. For CIFAR datasets, we use ResNet-18 [21] and RepVGG-A2 [11]. For ImageNet, we include ResNet-50, ConvNeXt-Base [33], and ViT-Base/16 [12], representing classical CNNs, modern CNNs, and transformer.

Metrics and Baselines. We report standard accuracy, selective-classification, and misclassification-detection metrics: Area Under the Risk Coverage Curve (AURC), Excess-AURC (E-AURC), Normalized (N-AURC), Risk at 80% Coverage (Risk@80), False Positive Rate at 95% True Positive (FPR@95), and Area Under the Receiver Operating Characteristics (AUROC). All metrics are averaged over three random seeds. Extended results under corruptions and distribution shifts appear in the Supplementary. We compare DARTS against representative selective-classification and confidence-estimation methods: Softmax Entropy [23], SelectiveNet [18], DOCTOR [19], Deep Gamblers [34], and SURE [29], along with a vanilla training baseline. All are trained under identical architectures and hyperparameters for fairness. Full implementation details are provided in the Supplementary (D).

7. Results and Analysis

7.1. Comparison of DARTS with Baselines

CIFAR10 Becnhmark Table 1 reports selective classification results on CIFAR-10 benchmark. Across both RepVGG-a1 and ResNet56 architectures, DARTS consistently yields the best reliability across all metrics, indicating strong confidence calibration and margin control.

DARTS vs Vanilla training We refer to training with standard cross-entropy loss as *vanilla training*. Relative to the latter, DARTS substantially improves selective reliability even when using the same confidence function. For example, when confidence is measured with maximum softmax probability, DARTS reduces AURC from 11.1×10^{-3} to 6.1×10^{-3} on RepVGG-a2 and from 12.0×10^{-3} to 8.0×10^{-3} on ResNet-56 - a reduction of 39.2% on average. Similarly, the risk at 80% coverage drops from 9.2×10^{-3} to 5.9×10^{-3} (RepVGG-A2) and from 10.4×10^{-3} to 7.6×10^{-3} (ResNet-56) - a reduction of 31.4% on average - showing that DARTS training improves ranking ability even without changing the scoring rule. This demonstrates that DARTS 's feature-space margin regularization directly improves confidence reliability by shaping class decision regions during training.

Table 1. Selective classification performance on **CIFAR-10** ($\times 10^3$). Lower values indicate better reliability. All methods use identical backbones and training protocols. Mean $\pm$ standard deviation are reported over three runs.

Method	RepVGG-A2				ResNet56			
	Risk@80%	AURC	E-AURC	NAURC	Risk@80%	AURC	E-AURC	NAURC
SelectiveNet [18]	9.5±0.2	11.5±0.3	10.6±0.3	512.8±4.7	10.9±0.3	12.3±0.3	10.9±0.3	416.4±3.9
DOCTOR [19]	9.2±0.2	11.2±0.3	10.3±0.3	507.5±4.5	10.5±0.3	12.0±0.3	10.5±0.2	411.5±3.8
Deep Gamblers [34]	9.1±0.2	10.8±0.3	9.9±0.3	499.4±4.3	10.3±0.3	11.6±0.3	10.2±0.2	404.8±3.6
SURE [29]	8.9±0.2	10.6±0.3	9.7±0.2	492.2±4.2	10.2±0.2	11.3±0.3	9.9±0.2	398.6±3.5
Vanilla Training (Max Softmax)	9.2±0.2	11.1±0.3	10.2±0.2	505.7±4.4	10.4±0.2	12.0±0.3	10.5±0.2	409.7±3.7
Vanilla Training (Nearest-Boundary)	5.8±0.2	6.5±0.2	5.6±0.2	276.5±3.0	7.0±0.2	7.4±0.2	5.9±0.2	230.4±2.5
DARTS (Max Softmax)	5.9±0.1	6.1±0.1	5.1±0.1	238.8±2.5	7.6±0.2	8.0±0.2	6.2±0.1	**225.3**±2.3
DARTS (Nearest-Boundary)	**4.1**±0.1	**4.4**±0.1	**3.4**±0.1	**159.4**±2.0	**6.5**±0.2	**7.0**±0.1	**5.3**±0.1	189.6±2.1

Table 2. Selective classification performance on **CIFAR-100** ($\times 10^2$). Lower values indicate better reliability. All methods use identical backbones and training protocols. Mean $\pm$ standard deviation are reported over three runs.

Method	RepVGG-A2				ResNet56			
	Risk@80%	AURC	E-AURC	NAURC	Risk@80%	AURC	E-AURC	NAURC
SelectiveNet [18]	10.48±0.12	5.63±0.10	3.51±0.09	45.82±0.43	15.59±0.15	8.37±0.13	4.85±0.10	53.41±0.51
DOCTOR [19]	10.24±0.10	5.51±0.08	3.48±0.08	45.23±0.40	15.20±0.13	8.26±0.11	4.81±0.09	52.93±0.48
Deep Gamblers [34]	10.20±0.11	5.54±0.09	3.42±0.07	44.77±0.42	15.24±0.12	8.21±0.10	4.76±0.08	52.50±0.45
SURE [29]	10.13±0.10	5.49±0.08	3.38±0.07	44.12±0.38	15.18±0.11	8.15±0.09	4.72±0.08	51.83±0.41
Vanilla Training (Max Softmax)	**10.01**±0.10	5.47±0.08	3.44±0.07	44.69±0.39	**15.12**±0.11	8.19±0.09	4.74±0.08	52.17±0.40
Vanilla Training (Nearest-Boundary)	10.21±0.12	**5.05**±0.07	3.02±0.06	39.28±0.37	15.85±0.14	7.88±0.08	4.43±0.07	48.74±0.39
DARTS (Max Softmax)	11.19±0.11	5.55±0.08	3.16±0.06	38.85±0.35	15.31±0.12	**7.45**±0.07	**3.88**±0.06	**42.27**±0.34
DARTS (Nearest-Boundary)	11.46±0.10	5.30±0.07	**2.91**±0.06	**35.77**±0.33	15.63±0.11	7.50±0.07	4.12±0.06	44.90±0.35

Table 3. Average selective classification performance across architectures on ImageNet-1K ($\times 10^2$).

Method	Risk@80	AURC	E-AURC	NAURC
SelectiveNet [18]	9.54	6.63	4.46	63.25
DOCTOR [19]	9.38	6.44	4.40	61.77
Deep Gamblers [34]	9.31	6.34	4.33	61.10
SURE [29]	9.29	6.34	4.32	60.78
Vanilla Training (Max Softmax)	9.58	6.49	4.63	62.16
Vanilla Training (Nearest-Boundary)	9.24	4.91	3.05	40.90
DARTS (Max Softmax)	**8.95**	5.43	3.60	48.45
DARTS (Nearest-Boundary)	9.21	**4.84**	**3.01**	**40.57**

Comparison with existing selective methods. When compared to dedicated selective-classification models such as SelectiveNet, DOCTOR, Deep Gamblers, and SURE, DARTS achieves the lowest values on every metric while using no additional selection head or auxiliary uncertainty branch. For instance, DARTS coupled with distance to nearest boundary confidence metric attains AURC of 4.4×10^{-3} and E-AURC of 3.4×10^{-3} on RepVGG-A2-representing an improvement of over 60% relative to SelectiveNet (11.5×10^{-3} and 10.6×10^{-3} respectively). Even the weaker DARTS coupled with maximum softmax confidence variant outperforms all baselines on both networks, confirming that margin-aware training alone enhances the separability of correctly and incorrectly classified samples in confidence space. The gains are pronounced in normalized AURC (N-AURC), where DARTS reduces the score from 512.8 (SelectiveNet) to 159.4, a $1.49\times$ improvement.

Impact of distance to nearest-boundary confidence Even without DARTS training, replacing the softmax confidence with the distance to nearest boundary confidence metric yields a large improvement. Models yielded from vanilla training recipe, when coupled with this metric, achieves AURC of 6.5×10^{-3} as opposed to 11.1×10^{-3} for maximum softmax. cutting selective error by nearly half. This suggest that geometric boundary is a better proxy for reliability than raw softmax probability.

Results on CIFAR-100. Table 2 summarizes results on CIFAR-100, a more challenging 100-class dataset that tests the scalability of selective reliability. Even under this high-entropy regime, DARTS consistently outperforms or matches all competing methods across architectures. Compared to standard training, DARTS significantly improves selective classification across all metrics-for example, with ResNet-56, DARTS (Max Softmax) reduces AURC from 8.19×10^{-2} to 7.45×10^{-2} and N-AURC

Table 4. **DARTS (Nearest-Boundary) vs. Vanilla Training (Max Softmax)** on CIFAR-10-C and CIFAR-100-C. Values are averaged over all corruption types and severities. ↓ lower is better; ↑ higher is better.

Dataset	Architecture	AURC↓	E-AURC↓	N-AURC↓	FPR@95↓	AUROC↑	Relative Δ (%)
CIFAR-10-C	ResNet-56	**0.106** vs 0.142	**0.052** vs 0.075	**0.48** vs 0.74	**0.38** vs 0.57	**0.89** vs 0.85	AURC ↓25, FPR ↓33, AUROC ↑5
CIFAR-10-C	RepVGG-A2	**0.115** vs 0.155	**0.056** vs 0.081	**0.49** vs 0.79	**0.36** vs 0.60	**0.89** vs 0.84	AURC ↓26, FPR ↓40, AUROC ↑5
CIFAR-100-C	ResNet-56	**0.314** vs 0.398	**0.112** vs 0.146	**0.89** vs 1.23	**0.53** vs 0.66	**0.81** vs 0.77	AURC ↓21, FPR ↓20, AUROC ↑4
CIFAR-100-C	RepVGG-A2	**0.339** vs 0.426	**0.120** vs 0.158	**0.84** vs 1.27	**0.55** vs 0.70	**0.80** vs 0.76	AURC ↓20, FPR ↓22, AUROC ↑4

Table 5. Comparison between **DARTS** and **Vanilla-trained** models on ImageNet-Rendition [24] using ConvNext Base architecture. All values are lower is better except AUROC (higher-better).

Training	Scoring	Risk@80%	AURC↓	E-AURC↓	N-AURC↓	FPR@95↓ / AUROC↑
DARTS	Nearest-Boundary	0.610	**0.438**	**0.137**	**4.18**	**0.633** / 0.791
	MSP	**0.600**	0.442	0.142	4.32	0.664 / **0.812**
Vanilla	Nearest-Boundary	0.614	0.453	0.152	4.69	0.667 / 0.780
	MSP	0.605	0.529	0.228	7.03	0.925 / 0.746

from 52.17 to 42.27, indicating sharper risk–coverage behavior, while on RepVGG-A2, it lowers E-AURC from 3.44×10^{-2} to 3.16×10^{-2}, confirming that margin-aware training constrains confidence dispersion even in large-class settings. Relative to selective-classification baselines such as SelectiveNet, DOCTOR, Deep Gamblers, and SURE, DARTS achieves lower AURC and E-AURC values without auxiliary selection heads or ensemble mechanisms; notably, DARTS (Nearest-Boundary) attains an E-AURC of 2.91×10^{-2} on RepVGG-A2-about a 14% reduction compared to SURE - and the best N-AURC (35.77). Even in vanilla training, replacing softmax confidence with boundary-based confidence reduces AURC from 5.47×10^{-2} to 5.05×10^{-2}. When coupled with DARTS training, this effect compounds.

Results on ImageNet. As summarized in Table 3 (see Appendix Table 9 for detailed results), DARTS consistently improves selective reliability across architectures on ImageNet-1K. The distance to nearest-boundary variant achieves the lowest average AURC (4.84) and E-AURC (3.01), outperforming prior selective-classification methods (SelectiveNet, DOCTOR, Deep Gamblers, SURE) as well as the vanilla Nearest-Boundary baseline, demonstrating DARTS's superior uncertainty calibration over the full confidence spectrum. Similarly, coupling DARTS training with maximum softmax score attains the lowest Risk@80% coverage (8.95), improving by over 6% relative to vanilla softmax training while maintaining strong overall calibration. Notably, DARTS improves both probabilistic and geometry-based uncertainty estimators without altering model structure, with consistent reductions in AURC and NAURC across ConvNeXt-Base, ResNet-50, and ViT-Base/16. The improvements are most pronounced for convolutional architectures, yet also apply to transformers.

7.2. DARTS for Misclassification Detection

DARTS consistently enhances the ability to distinguish correctly classified samples from misclassified ones. Reductions in FPR@95 directly indicate that fewer incorrect predictions are assigned high confidence, while improvements in AUROC reflect better global separability between correct and incorrect instances across all confidence thresholds. On CIFAR-10, DARTS yields the most significant gains, achieving up to 47.7% lower FPR@95 and 5.4% higher AUROC compared to source training. Thus, boundary-aware regularization effectively suppresses over-confidence near decision margins and sharpens the geometric separation between feature subspaces of different classes.

Table 6. Misclassification detection results on CIFAR benchmarks (%). Δ denotes absolute improvement of DARTS over the corresponding source model (lower FPR@95 and higher AUROC indicate better detection).

Dataset	Scoring	RepVGG-A2		ResNet-56	
		FPR@95↓	AUROC↑	FPR@95↓	AUROC↑
CIFAR-10	Vanilla Training (MSP)	76.9	86.9	66.3	88.8
	DARTS (MSP)	29.2	92.3	27.5	92.0
	Improvement	**47.7**	**+5.4**	**38.8**	**+3.2**
	Vanilla Training (Nearest-Bound.)	49.8	90.9	26.4	92.1
	DARTS (Nearest-Bound.)	22.3	93.8	21.4	92.9
	Improvement	**27.5**	**+2.9**	**5.0**	**+0.8**
CIFAR-100	Vanilla Training (MSP)	48.2	86.5	52.9	84.9
	DARTS (MSP)	42.0	87.1	43.5	86.6
	Improvement	**6.2**	**+0.6**	**9.4**	**+1.7**
	Vanilla Training (Nearest-Bound.)	39.7	87.1	47.9	84.9
	DARTS (Nearest-Bound.)	38.9	87.6	43.1	85.4
	Improvement	**0.8**	**+0.5**	**4.8**	**+0.5**

On the more challenging CIFAR-100 benchmark, which exhibits higher inter-class overlap and lower margin separability, DARTS maintains consistent improvements of up to 9% reduction in FPR@95 and up to 1.7% increase in AUROC. Although absolute gains are smaller due to the increased class entropy, the trends remain uniform across both RepVGG-A2 and ResNet-56 backbones. This indicates that the proposed dual-margin training principle generalizes across architectures with distinct inductive biases (convolutional or re-parameterized) and reinforces the overall reliability of the predictive confidence under diverse data regimes. In summary, the improvements in both FPR@95 and AUROC demonstrate that DARTS not only calibrates model confidence but also aligns the decision geometry more closely with the underlying data structure.

7.3. Selective Classification under Distribution Shift

We compare DARTS, which uses distance to the nearest decision boundary for confidence, with conventional cross-entropy models using Max Softmax, the standard selective-

classification baseline [23]. As shown in Table 4, DARTS consistently improves reliability across architectures and datasets. On CIFAR-10-C, it reduces AURC by 25.6% and FPR@95 by up to 36.5%, while increasing AUROC by 3–5 points. It also maintains lower selective risk on CIFAR-100-C, demonstrating robustness under corruption. Unlike Max Softmax, which relies on output probabilities, DARTS derives confidence from feature-space geometric margins, yielding sharper risk–coverage behavior.

Table 5 extends this to ImageNet-Rendition with a ConvNeXt-Base backbone. DARTS improves calibration and error separation across scoring methods, achieving lower AURC, E-AURC, and N-AURC, as well as reduced FPR@95 and higher AUROC. The Nearest-Boundary score offers the most consistent reliability, while MSP benefits from DARTS 's improved boundary geometry. Overall, DARTS establishes a robust confidence structure through feature-space regularization, outperforming standard training without auxiliary calibration or post-hoc correction.

7.4. Efficiecny Comparison

As evident from 7, DARTS introduces negligible computational overhead. For ConvNext base architecture and Imagenet dataset, the per batch training latency increases only by 3 ms (3.5%) while inference latency is reduced drastically over 5× compared to maximum softmax probability.

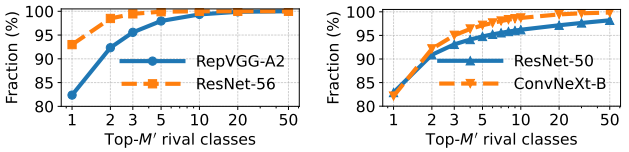
Table 7. Training and inference latency (ms) on ImageNet with ConvNeXt-Base (batch size = 32).

Method	Inference Latency (ms)	Training Latency (ms)
MSP (Vanilla Training)	70	82
DARTS (Nearest-Boundary)	13	85

7.5. Ablation Study

Distance to Boundary and top M' Rival Classes We identify the nearest rival class by searching among the top-M' classes ranked by logits. To validate this approximation, we measure the fraction of samples whose true nearest neighbor in geometric distance coincides with one of the top-M' logit classes. As shown in Fig. 3, the coverage rises sharply with M': on CIFAR-100, over 97% of samples are captured within the top-5 rivals, while on ImageNet-1K more than 95% fall within the same range and nearly all within the top-10. This rapid saturation demonstrates that decision-boundary geometry is largely governed by a small set of high-probability competitors, supporting our use of a top-M (with M'=10) truncation that provides substantial computational efficiency with negligible error.

Ablation on Loss Components Table 8 analyzes the effect of applying high-margin (m_{hi}) and low-margin (m_{lo}) constraints in DARTS training. Using only the high-margin



(a) CIFAR-100. (b) ImageNet-1K.

Figure 3. Fraction of samples where the nearest class (in geometric distance) appears within the top-M logits. On both CIFAR-100 and ImageNet-1K, the true nearest rival is captured by the top-5 logits in nearly all cases, justifying the top-M truncation used in our boundary approximation.

Table 8. Ablation study on the impact of high-margin (m_{hi}) and low-margin (m_{lo}) constraints. Results are shown on CIFAR-10 using RepVGG_a2. All values are lower is better.

Configuration	Scoring	Risk@95%	AURC↓	E-AURC↓	N-AURC↓
m_{hi} only	Nearest-Boundary	0.0071	0.0093	0.0084	0.4111
	Max Softmax	0.0088	0.0118	0.0109	0.5371
m_{lo} only	Nearest-Boundary	0.0063	0.0070	0.0062	0.3124
	Max Softmax	0.0089	0.0107	0.0098	0.4896
Both m_{hi} and m_{lo}	Nearest-Boundary	**0.0041**	**0.0044**	**0.0034**	**0.1594**
	Max Softmax	0.0059	0.0061	0.0051	0.2388

constraint enlarges the separation among correctly classified samples but provides limited robustness against misclassified points. Applying only the low-margin constraint helps suppress overconfident wrong predictions, leading to slightly better calibration and reduced false positives, although the overall improvement remains moderate. When both constraints are combined, DARTS achieves the best performance across all selective-classification and misclassification metrics, showing the lowest AURC, E-AURC, and N-AURC, together. These results confirm that jointly enforcing both margins shapes the decision boundaries more effectively by expanding correct-sample separation while constraining overconfidence on misclassified samples, yielding stronger overall reliability under selective prediction.

8. Conclusion

Our proposed DARTS takes a geometric point of view for improving the selective classification performance of Deep Neural Networks (DNNs) and separates the correct and incorrect inferences with asymmetric margin constraints. DARTS is lightweight with minimal overhead in computation and latency compared to the vanilla training with cross-entropy loss. Extensive experiments on six datasets and five different architectures show that DARTS not only improves selective classification performance on clean data but on distribution shifted inputs as well. Overall, DARTS establishes a new paradigm for reliable deep learning-where uncertainty is grounded in interpretable feature-space geometry rather than ad hoc output statistics.

Acknowledgements

This work has been supported by the National Science Foundation under grants CNS-2312875 and OAC-2530896; by the Air Force Office of Scientific Research under grant FA9550-23-1-0261; by the Office of Naval Research under grant N00014-23-1-2221; and by the Defense Advanced Research Projects Agency under Cooperative Agreement D25AC00374-00.

References

[1] Dimitri P. Bertsekas. Nonlinear programming. *Athena Scientific*, 1999. 10

[2] Luís Felipe Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks. In *Uncertainty in Artificial Intelligence (UAI) 2024*, 2024. 1

[3] Luís Felipe Prates Cattelan and Danilo Silva. How to fix a broken confidence estimator: Evaluating post-hoc methods for selective classification with deep neural networks, 2024. 14

[4] Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. A game theoretic approach to class-wise selective rationalization. *Advances in neural information processing systems*, 32, 2019. 1

[5] Mohammad Amin Charusaie et al. Sample efficient learning of predictors that complement humans. In *Proceedings of the 2022 International Conference on Machine Learning (ICML) Workshops*, 2022. 1

[6] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. 1

[7] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *Advances in neural information processing systems*, 32, 2019. 1

[8] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. *Journal of Machine Learning Research*, 17:1–42, 2016. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale Hierarchical Image Database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 2

[11] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 5

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929, 2020. 5

[13] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 327–334, 2010. 1, 2

[14] Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir H Abdi. Towards better selective classification. In *The Eleventh International Conference on Learning Representations*. 1

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1050–1059, 2016. 1, 2

[16] Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021*, page 1183–1193, 2021. 1

[17] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS) 30*, pages 4878–4887, 2017. 1, 3

[18] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, pages 2151–2159. PMLR, 2019. 1, 2, 3, 5, 6, 13, 15

[19] Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. Doctor: A simple method for detecting misclassification errors. *Advances in Neural Information Processing Systems*, 34:5669–5681, 2021. 5, 6, 13, 15

[20] 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. 14

[21] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 5

[22] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 41–50, 2019. 2

[23] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2017. 2, 5, 8, 12, 14

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 7, 14

[25] Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4):709–721, 2006. 1

[26] Tianjin Huang, Vlado Menkovski, Yulong Pei, and Mykola Pechenizkiy. Calibrated adversarial training. In *Asian Con-*

ference on Machine Learning, pages 626–641. PMLR, 2021. 1

[27] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. 5

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017. 1, 2

[29] Yuting Li, Yingyi Chen, Xuanlong Yu, Dexiong Chen, and Xi Shen. Sure: Survey recipes for building reliable and robust deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17500–17510, 2024. 5, 6, 13, 15

[30] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning (ICML) Workshops*, 2016. 2

[31] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 2, 14

[32] Xixi Liu, Yaroslava Lochman, and Zach Christopher. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 14

[33] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 5

[34] Zhen Liu, Yifei Wang, and Yisen Wang. Deep gamblers: A unified framework for selective classification and active learning. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 5, 6, 13, 15

[35] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018. 11

[36] Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 15634–15650, 2022. 1, 2, 3

[37] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015. 11

[38] JC PLATT. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999. 2

[39] Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. *Journal of Data-centric Machine Learning Research*, 2024. Reproducibility Certification. 1

[40] Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3): 400–407, 1951. 10

[41] Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023. 1

[42] Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, 1996. 11

[43] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019. 11

[44] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[45] Felix Wenzel, Kevin Roth, Bastiaan S Veeling, et al. Good, better, best: A comparison of uncertainty quantification methods in deep learning. *arXiv preprint arXiv:2006.07584*, 2020. 2

[46] Yair Wiener and Ran El-Yaniv. Agnostic selective classification. *Advances in Neural Information Processing Systems*, 24, 2011. 1, 2

A. Supplementary Material

A.1. Theoretical Analysis of DARTS

This appendix provides detailed proofs and auxiliary results supporting the theoretical analysis in Sec. 5. We retain all notation and assumptions introduced in the main text.

A.2. Convergence of Boundary-Aware Objective

Lemma A.1 (SGD boundedness and descent)**.** *Under Assumptions (A1)–(A3), suppose the stepsize sequence $\{\eta_t\}$ satisfies $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$. Then for stochastic gradient updates on*

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{corr}}\mathcal{L}_{\mathrm{corr}} + \lambda_{\mathrm{wrong}}(t)\mathcal{L}_{\mathrm{wrong}},$$

the iterates (W_t, b_t, ϕ_t) remain bounded and $\mathcal{L}(W_t, b_t, \phi_t)$ converges almost surely.

Proof. Each component loss is smooth and lower-bounded; their gradients have finite variance owing to (A1)(A3). Standard Robbins–Monro conditions [1, 40] imply almost sure convergence of stochastic approximation schemes to stationary points. Boundedness of (W_t, b_t) follows from coercivity induced by weight decay and the bounded features assumption. □

Proposition A.2 (Gradient dynamics of dual-margin terms)**.** *For a correctly classified sample (x, y) with $\gamma(x) < m_{\mathrm{hi}}$,*

$$\frac{\partial \mathcal{L}_{\mathrm{corr}}}{\partial \gamma(x)} = -2(m_{\mathrm{hi}} - \gamma(x)) < 0,$$

so the update increases $\gamma(x)$. For a misclassified sample with top-two gap $d_x > m_{\mathrm{lo}}$,

$$\frac{\partial \mathcal{L}_{\mathrm{wrong}}}{\partial d_x} = 2(d_x - m_{\mathrm{lo}}) > 0,$$

so the update decreases d_x.

Proof. Both follow from direct differentiation of the hinge-squared penalties in Eqs. (10–9). □

Theorem A.3 (Convergence to margin-consistent equilibrium)**.** *Combining Lemma A.1 and Proposition A.2, DARTS converges to*

a stationary configuration where $\gamma(x) \geq m_{\text{hi}}$ for correct samples and $d_x \leq m_{\text{lo}}$ for misclassified ones. Hence, the feature geometry reaches a stable equilibrium balancing margin expansion and overconfidence suppression.

Proof. Let $D(x) = \gamma(x, y; W, b)$. Then

$$\mathcal{L}_{\text{corr}} = \frac{1}{|\mathcal{C}|} \sum_{x_b} [\max(0, m_{\text{hi}} - D(x_b))]^2.$$

For active x_b with $D(x_b) < m_{\text{hi}}$,

$$\frac{\partial \mathcal{L}_{\text{corr}}}{\partial D(x_b)} = -2(m_{\text{hi}} - D(x_b)) < 0.$$

Since SGD updates in the **negative gradient** direction, $D(x_b)$ increases. Similarly, for $\mathcal{L}_{\text{wrong}}$, active terms with $d_x > m_{\text{lo}}$ yield $\frac{\partial \mathcal{L}_{\text{wrong}}}{\partial d_x} > 0$, so d_x decreases. □

Remark A.1 (Curriculum stabilization). The time-dependent weight $\lambda_{\text{wrong}}(t)$ guarantees smooth enforcement of the misclassified constraint, preventing oscillations in early training when class predictions are unstable.

A.3. Error-Rate Generalization via Normalized Margins

Lemma A.4 (Relationship between Normalized and Unnormalized margin)**.** *Let $z(x, y; W) := f_y(x) - \max_{c \neq y} f_c(x)$ be the unnormalized margin.*

$$z(x, y; W) = \max_{c \neq y} \left[\gamma_c(x, y; W) \|w_c - w_y\|\right].$$

where, $\gamma_c := \frac{(f_y - f_c)}{\|w_y - w_c\|}$. Thus, $\gamma(x, y; W) = \min_{c \neq y} \gamma_c$ satisfies $z(x, y; W) \geq \gamma(x, y; W)\rho$. Equality holds only if maximizing c for z is also the minimizer for γ.

Proof. The lemma follows directly from the definition. From the assumptions, $\|w_c - w_y\| \geq \rho$. Thus, $z \geq \gamma \|w_c - w_y\| \implies z \geq \gamma\rho$. □

Theorem A.5 (Generalization bound)**.** *Let $\mathcal{W}_R = \{W : \|w_c\|_2 \leq R\}$. For any $\tau > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$ over n i.i.d. samples,*

$$\Pr(\text{err}(f)) \leq \widehat{\Pr}(\gamma(x) \leq \tau) + \tilde{\mathcal{O}}\left(\frac{RB\sqrt{K}}{\tau\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Proof. The result adapts the standard margin-based generalization bound for linear classifiers [37] to normalized margins $\gamma(x)$ by noting that the effective classifier acts on bounded features of norm B and normalized direction vectors $\frac{w_y - w_c}{\|w_y - w_c\|_2}$.

(Error Decomposition) Let $S = \{(x_i, y_i)_{i=1}^{i=n}\}$ be the training set. Define the indicator function $\text{err}(f) := \mathbb{I}[y \neq \arg\max_c f_c(x)]$. So, a point is misclassified *iff* $z(x, y; W) \leq 0$ (hence, $\gamma \leq 0$). Thus,

$$\begin{aligned} \Pr(\text{err}(f) = 1) \leq \Pr\big(\gamma(x) \leq \tau\big) \\ + \Pr\big(\text{err}(f) = 1,\ \gamma(x) > \tau\big). \end{aligned} \tag{15}$$

The empirical counterpart of the first term is $\widehat{\Pr}(\gamma(x) \leq \tau)$. We bound the second via a surrogate.

(Surrogate Loss) Define *ramp loss* as convex surrogate to 0-1 loss:

$$\ell_\tau(z) := \min\big\{1,\ \max\{0,\ 1 - z/\tau\}\big\}.$$

As ramp upper-bounds the 0-1 loss, $\text{err}(x, y; f) \leq \ell_\tau(z(x, y; W))$ always. By A.4, if $\gamma > \tau$, then $z \geq \gamma\rho > \tau\rho$. So, for correctly classified samples, $z > 0 \implies z/\tau > \rho > 0$, so, $\ell_\tau(z) = 0$. If misclassified, $z \leq 0 \implies \ell_\tau(z) = 1$. Thus, $\Pr(\text{err} = 1, \gamma > \tau) \leq \mathbb{E}[\ell_\tau(z)]$.

□

(Function Class for the Ramp Loss) The ramp is $(1/\tau)$-Lipschitz. We define the unnormalized linear class

$$\begin{aligned} \mathcal{G} := \big\{(x, y) \mapsto (w_y - w_{c(x)})^\top h_\phi(x) \mid \\ W \in \mathcal{W}_R,\ c(x) = \arg\max_{c' \neq y} f_{c'}(x)\big\} \end{aligned} \tag{16}$$

The class of ramp losses is given by

$$\mathcal{F}_\tau := \{(x, y) \mapsto \ell_\tau(g(x, y)) \mid g \in \mathcal{G}\}.$$

The empirical Rademacher complexity is

$$\begin{aligned} \widehat{\text{Rad}}_n(\mathcal{F}_\tau) := \mathbb{E}_\sigma\Big[\frac{1}{n} \sup_{f \in \mathcal{F}_\tau} \sum_{i=1}^{n} \sigma_i f(x_i, y_i)\Big], \\ \sigma_i \sim \text{Rademacher} \end{aligned} \tag{17}$$

By Talagrand's contraction [42], $\widehat{\text{Rad}}_n(\mathcal{F}_\tau) \leq (1/\tau)\widehat{\text{Rad}}_n(\mathcal{G})$. For $\mathcal{G}$, we know that the class of linear functions with coefficients of norm $\leq 2R$ on features of norm $\leq B$ has log-covering number

$$\log \text{Cov}(\mathcal{G}, \epsilon, \ell_2(S)) \leq K \log\left(1 + \frac{2RB}{\epsilon}\right) + O(\log n),$$

Applying Dudley's entropy integral [43]:

$$\begin{aligned} \widehat{\text{Rad}}_n(\mathcal{G}) \leq & \frac{12}{\sqrt{n}} \int_0^{\sup_{g \in \mathcal{G}} \|g\|_{\ell_2(S)}} \sqrt{\log \mathcal{N}(\mathcal{G}, \varepsilon, \ell_2(S))}\, d\varepsilon \\ & \leq \tilde{\mathcal{O}}\left(RB\sqrt{\frac{K}{n}}\right). \end{aligned} \tag{18}$$

By scaling with $1/\tau$, we get:

$$\widehat{\text{Rad}}_n(\mathcal{F}_\tau) \leq \tilde{\mathcal{O}}\left(\frac{RB\sqrt{K}}{\tau\sqrt{n}}\right).$$

(Uniform Convergence) The ramp loss $\ell_\tau(z) \in [0, 1]$ is bounded, so standard uniform convergence bounds apply. By the bounded-difference inequality and symmetrization ([35]), with probability at least $1 - \delta$ over the draw of the sample S,

$$\mathbb{E}[\ell_\tau(z)] \leq \widehat{\mathbb{E}}_S[\ell_\tau(z)] + 2\,\widehat{\text{Rad}}_n(\mathcal{F}_\tau) + \sqrt{\frac{\log(2/\delta)}{2n}}. \tag{4.1}$$

As the ramp loss upper bounds the 0-1 error: $\mathbb{I}\{\text{err}(f) = 1\} \leq \ell_\tau(z)$ for all z, and moreover,

$$\widehat{\mathbb{E}}_S[\ell_\tau(z)] \leq \widehat{\Pr}_S(\gamma \leq \tau) + \widehat{\Pr}_S(\gamma > \tau, \text{err} = 1).$$

On the training set, if the classifier achieves zero error on points with margin $> \tau$ (as is typical in large-margin settings), the second term vanishes. In general, it is nonnegative but bounded by the empirical error, and is absorbed into the low-margin empirical mass $\widehat{\Pr}_S(\gamma \le \tau)$ for the purpose of upper bounds.

Final Assembly Recall from Step 1 that

$$\begin{aligned}\Pr(\mathrm{err}(f) = 1) \le \Pr\big(\gamma(x) \le \tau\big) \\ + \Pr\big(\mathrm{err}(f) = 1,\ \gamma(x) > \tau\big).\end{aligned} \tag{19}$$

Combining this with the uniform convergence bound and the Rademacher estimate,

$$\begin{aligned}\Pr(\mathrm{err} = 1) &\le \widehat{\Pr}(\gamma \le \tau) + \mathbb{E}[\ell_\tau(z)] \\ &\le 2\,\widehat{\Pr}(\gamma \le \tau) + \widetilde{\mathcal{O}}\left(\frac{RB\sqrt{K}}{\tau\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right).\end{aligned} \tag{20}$$

The factor of 2 is conventional in margin bounds and can be absorbed into the $\widetilde{\mathcal{O}}$ notation, yielding the claim.

Corollary A.6 (Error bound under enforced margins)**.** *If $\gamma(x) \ge m_{\mathrm{hi}}$ for at least $(1-\epsilon)$ fraction of the training data, then setting $\tau = m_{\mathrm{hi}}$ gives*

$$\Pr(\mathrm{err}(f)) \lesssim \epsilon + \frac{RB\sqrt{K}}{m_{\mathrm{hi}}\sqrt{n}} + \sqrt{\tfrac{\log(1/\delta)}{n}},$$

highlighting that larger enforced margins and smaller weight norms tighten generalization.

Proof. The corollary follows by setting $\tau = m_{\mathrm{hi}}$ and assuming $\widehat{\Pr}_S(\gamma \le m_{\mathrm{hi}}) \le \epsilon$ in Theorem A.5. □

Remark A.2. This bound formally justifies the high-margin constraint of DARTS 's correct-sample regularizer: expanding decision regions for confident samples directly lowers asymptotic risk.

A.4. Scale Invariance of DARTS

As an additional property of DARTS, we show that it is invariant to scaling of the classifier.

Proposition A.7 (Scale invariance under head rescaling)**.** *Fix any $\alpha > 0$ and define a rescaled classifier head by $\tilde{w}_c = \alpha w_c$ and $\tilde{b}_c = \alpha b_c$ for all c. Let $\tilde{f}_c(x) = \tilde{w}_c^\top z(x) + \tilde{b}_c$ and define $\tilde{S}_{DARTS}$ analogously from $\{\tilde{w}_c, \tilde{b}_c\}$. Then for every x,*

$$\tilde{S}_{DARTS}(x) \;=\; S_{DARTS}(x).$$

Consequently, the induced ranking of examples by confidence, and any coverage-based selective metrics (e.g., RC curves, AURC, AUROC of accept/reject) are invariant to positive common rescalings of the classifier head.

Proof. For any x and any $c \ne \hat{y}(x)$ we have $\tilde{f}_{\hat{y}}(x) - \tilde{f}_c(x) = \alpha\big(f_{\hat{y}}(x) - f_c(x)\big)$ and $\|\tilde{w}_{\hat{y}} - \tilde{w}_c\|_2 = \|\alpha(w_{\hat{y}} - w_c)\|_2 = \alpha\|w_{\hat{y}} - w_c\|_2$. Thus each candidate ratio is unchanged:

$$\frac{\tilde{f}_{\hat{y}}(x) - \tilde{f}_c(x)}{\|\tilde{w}_{\hat{y}} - \tilde{w}_c\|_2} = \frac{\alpha\big(f_{\hat{y}}(x) - f_c(x)\big)}{\alpha\|w_{\hat{y}} - w_c\|_2} = \frac{f_{\hat{y}}(x) - f_c(x)}{\|w_{\hat{y}} - w_c\|_2}.$$

Because $\alpha > 0$, $\arg\max_c f_c(x)$ (and therefore the set of top-M rivals by logit) is preserved, so the minimum over rivals is also unchanged. Hence $\tilde{S}_{\mathrm{DARTS}}(x) = S_{\mathrm{DARTS}}(x)$ for all x. □

Corollary A.8 (Ranking and RC-metric invariance)**.** *Under the conditions of Prop. A.7, the ordering of examples by S_{DARTS} is identical before and after rescaling. Any selective metric that depends only on this ordering at a given coverage (e.g., risk at fixed coverage, AURC/EAURC, AUROC of accept/reject) is invariant.*

Temperature scaling at test time. Consider post-hoc temperature scaling applied only to logits: $f_c^{(T)}(x) = f_c(x)/T$ with $T > 0$, while the denominator still uses $\|w_{\hat{y}} - w_c\|_2$ from the unscaled head. Then

$$S^{(T)}_{\mathrm{DARTS}}(x) = \min_{c \ne \hat{y}} \frac{f_{\hat{y}}(x) - f_c(x)}{T\,\|w_{\hat{y}} - w_c\|_2} = \frac{1}{T}\,S_{\mathrm{DARTS}}(x).$$

Thus the *values* scale by $1/T$, but the *ranking* (and any coverage-based metrics) remain unchanged. If true invariance of the *numerical score* under temperature scaling is desired, compute the denominator with the correspondingly rescaled head, i.e., replace $\|w_{\hat{y}} - w_c\|_2$ by $\|w_{\hat{y}} - w_c\|_2/T$ (equivalently, pretend the head is scaled by $1/T$), which restores $S^{(T)}_{\mathrm{DARTS}}(x) = S_{\mathrm{DARTS}}(x)$ exactly.

Remarks. (i) The assumption $\alpha > 0$ (and $T > 0$) is essential; negative scaling would flip logit order and invalidate $\hat{y}$. (ii) Because multiplying all logits by a positive constant preserves their order, the *top-M' rival set* used by DARTS is unaffected by either head rescaling or temperature scaling.

B. Algorithms

C. Evaluation Metrics.

We report standard accuracy, selective-classification, and misclassification-detection metrics:
AURC: Area under the risk–coverage curve; lower values indicate better selectivity.
E-AURC: Expected AURC, computed by subtracting the minimal achievable risk–coverage area to isolate the excess selective error.
N-AURC: Normalized AURC, scaling AURC into $[0, 1]$ for easier comparison across datasets and architectures.
Risk@80: Error among the top 80% most confident predictions, measuring ranking quality.
FPR@95: False-positive rate of misclassified samples when 95% of correct predictions are accepted.
AUROC: Separability between correct and incorrect predictions across confidence thresholds.
All metrics are averaged over three random seeds. Extended results under corruptions and distribution shifts appear in the Supplementary.

D. Brief Description of the Baselines

MSP [23] This seminal baseline estimates confidence as the maximum predicted softmax probability. Correctly classified or in-distribution samples typically yield higher maximum softmax scores, allowing MSP to serve as a simple yet strong detector of

Algorithm 1 DARTS Training (nearest decision boundary over top-M' rivals)

Require: Training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, backbone h_ψ, linear head $(\mathbf{W}, \mathbf{b})$, epochs T, warmup T_0, margins $m_{\text{hi}}, m_{\text{lo}}$, loss weights $\lambda_{\text{corr}}, \lambda_{\text{wrong}}(t)$, number of rivals M'
Ensure: Trained model $\mathbf{f}(\mathbf{x}) = \mathbf{W}\, h_\psi(\mathbf{x}) + \mathbf{b}$

Warmup: Train with cross-entropy for T_0 epochs
function BOUNDARYDISTANCE(z, a, c, W)
  return $\dfrac{|z[a] - z[c]|}{\|w_a - w_c\|_2}$
end function
for $t = T_0 + 1$ to T **do**
  for each minibatch $\{(x_b, y_b)\}_{b=1}^{B}$ **do**
    $z_b \leftarrow \mathbf{W}\, h_\psi(x_b) + \mathbf{b}$ ▷ logits
    $L_{\text{CE}} \leftarrow \frac{1}{B} \sum_b \text{CE}(z_b, y_b)$
    $\hat{y}_b \leftarrow \arg\max_c z_b[c]$ ▷ predicted class
    $R_b \leftarrow$ Top-M' indices of z_b excluding the anchor class ▷ anchor defined below

    Correct samples $\mathcal{C} = \{b : \hat{y}_b = y_b\}$ (anchor $= y_b$):
      $D_b \leftarrow \min_{c \in R_b}$ BOUNDARYDISTANCE(z_b, y_b, c, $\mathbf{W}$)
      $L_{\text{corr}} \leftarrow \frac{1}{|\mathcal{C}|} \sum_{b \in \mathcal{C}} \left[\max(0,\, m_{\text{hi}} - D_b)\right]^2$

    Misclassified samples $\mathcal{W} = \{b : \hat{y}_b \neq y_b\}$ (anchor $= \hat{y}_b$):
      $d_b \leftarrow \min_{c \in R_b}$ BOUNDARYDISTANCE(z_b, $\hat{y}_b$, c, $\mathbf{W}$)
      $L_{\text{wrong}} \leftarrow \frac{1}{|\mathcal{W}|} \sum_{b \in \mathcal{W}} \left[\max(0,\, d_b - m_{\text{lo}})\right]^2$

    Total loss: $L \leftarrow L_{\text{CE}} + \lambda_{\text{corr}}\, L_{\text{corr}} + \lambda_{\text{wrong}}(t)\, L_{\text{wrong}}$
    Update parameters of h_ψ and $(\mathbf{W}, \mathbf{b})$ via backprop
  end for
end for

misclassified and out-of-distribution examples. Despite its simplicity, MSP remains a common reference point for both confidence calibration and selective prediction.

SelectiveNet [18] SelectiveNet integrates prediction and abstention into a single deep network trained end-to-end. It employs a three-headed architecture consisting of a prediction head $f(x)$, a selection head $g(x)$, and an auxiliary head that regularizes shared representations. Training jointly minimizes selective risk under a coverage constraint enforced by an Interior Point Method (IPM) penalty, directly optimizing the risk–coverage curve rather than relying on post-hoc thresholding.

DOCTOR [19] DOCTOR formalizes misclassification detection as a binary hypothesis testing problem between trustworthy and untrustworthy predictions. It derives an optimal detector that balances false acceptance and rejection errors. The method operates in both Totally Black-Box (TBB) and Partially Black-Box (PBB) settings, using only softmax outputs with optional temperature scaling or small input perturbations. DOCTOR is lightweight, training-free, and specifically optimized for identifying model misclassifications.

Algorithm 2 Selective Inference with Nearest-Boundary Confidence

Require: Trained classifier $f(x) = \mathbf{W}\mathbf{h}_p si(\mathbf{x}) + \mathbf{b}$, test input $\mathbf{x}$, rejection threshold τ (or target coverage γ)
Ensure: Prediction $\hat{y}$ and decision: ACCEPT or REJECT

Compute logits: $z \leftarrow f(x) = Wh(x) + b$
Predicted class: $\hat{y} \leftarrow \arg\max_c z[c]$
Compute nearest decision-boundary distance:
  Let $z_{\hat{y}} = z[\hat{y}]$
  Mask $\hat{y}$: set $z[\hat{y}] \leftarrow -\infty$
  Find top-M' rival classes $\mathcal{R} = \{c'_1, \ldots, c'_{M'}\}$ with highest logits
  Compute distances:

$$d_{c'} \leftarrow \frac{|z_{\hat{y}} - z[c']|}{\|w_{\hat{y}} - w_{c'}\|_2 + \varepsilon}, \quad \forall c' \in \mathcal{R} \tag{21}$$

  Confidence score: $C(x) \leftarrow \min_{c' \in \mathcal{R}} d_{c'}$
if $C(x) \geq \tau$ **then**
  return ($\hat{y}$, ACCEPT)
else
  return (None, REJECT)
end if

Deep Gambler [34] Deep Gamblers introduces a game-theoretic confidence learning mechanism by allocating a fixed "betting budget" across all classes and an abstention option. The model learns to wager higher probabilities on confident predictions while reserving part of the budget for abstention in uncertain cases. This strategy encourages calibrated selective behavior and yields improved risk–coverage trade-offs without explicit threshold tuning.

SURE [29] SURE (Survey Recipes for Building Reliable and Robust Deep Networks) takes a holistic, synergistic approach to building robust and reliable networks. Rather than focusing on a single technique, SURE integrates diverse, complementary strategies across the model's training pipeline: regularization (e.g., RegMixup, Correctness Ranking Loss) to increase entropy for hard samples and improve feature separation; classifier design (e.g., Cosine Similarity Classifier) to encourage better feature alignment; and optimization (e.g., Sharpness-Aware Minimization, Stochastic Weight Averaging) to find flatter, more generalizable minima.

E. Training and Hyperparameter Details

CIFAR benchmarks For all CIFAR-10 and CIFAR-100 experiments, we adopt a standardized training setup across both vanilla baselines and our proposed DARTS models. All models are trained for 200 epochs with a batch size of 128 using SGD with momentum 0.9 and a weight decay of 5×10^{-4}. The initial learning rate is set to 0.1 and is decayed using a multi-step schedule with milestones at epochs 60, 120, and 160, with a decay factor of 0.2 at each step. Following standard practice, label smoothing with strength 0.1 is applied to stabilize training. Unless otherwise noted, experiments are conducted using models from the

`chenyaofo/pytorch-cifar-models` repository.

For all DARTS experiments, the optimization hyperparameters above remain unchanged; only the loss function is modified to incorporate decision-boundary–aware regularization. We use a high-margin threshold $m_{\text{hi}} \in \{0.1, 0.3, 0.5, 0.7, 1.0, 2.0\}$ for correctly classified samples and a low-margin threshold $m_{\text{lo}} \in \{0.05, 0.15, 0.3\}$ for incorrect ones. The corresponding margin penalties are weighted by λ_{corr} and $\lambda_{\text{wrong}}(t)$, where the latter is linearly increased from 2.0 to 3.0 over the training epochs. A warm-up period of $T = 50$ epochs is used before activating margin-based training to prevent early optimization instability. This ensures that any improvement in selective-classification performance arises solely from the proposed DARTS objective rather than differences in general training dynamics.

ImageNet Benchmark For all ImageNet experiments, we adopt a unified training–evaluation pipeline and adjust only model-specific optimization and Feature-RAT parameters. **ResNet-50** is trained for 30 epochs using SGD (initial learning rate 0.001, momentum 0.9, weight decay 1×10^{-4}) with label smoothing 0.1. Feature-RAT margins are set to $m_{\text{hi}} = 0.7$ and $m_{\text{lo}} = 0.05$, with loss weights $\lambda_{\text{corr}} = 1.0$ and λ_{wrong} linearly scheduled from 2.0 to 3.0. **ConvNeXt-Base** is trained for 10 epochs using SGD with a reduced initial learning rate (1×10^{-4}), momentum 0.9, and larger weight decay (0.02). Owing to stronger feature separation in ConvNeXt, we employ higher-margin constraints with $m_{\text{hi}} = 5$ and $m_{\text{lo}} = 1$, together with $\lambda_{\text{corr}} = 2.0$ and λ_{wrong} annealed from 2.0 to 4.0. **ViT-B/16** is trained for 50 epochs using SGD with a higher learning rate (0.005) and negligible weight decay (10^{-9}). Margins are tuned to $m_{\text{hi}} = 0.4$ and $m_{\text{lo}} = 0.05$, with $\lambda_{\text{corr}} = 1.0$ and λ_{wrong} ramped from 2.0 to 4.0. Across all architectures, we use a batch size of 256 and employ the same Top-M approximation with $M = 10$ rival classes when computing nearest-boundary penalties.

F. Detailed Results on ImageNet

The detailed results for the ImageNet benchmark are provided in Table 9.

G. Detailed Results on CIFAR-C

The detailed results on CIFAR-C benchmark are provided in Tables **??**-13. We compare the mean metrics for vanilla cross-entropy trained DNN with maximum softmax probability score and DARTS trained and distance to nearest boundary score.

H. Additional Comparison with Post-hoc Scoring Methods

We analyze the post-hoc misclassification detection and selective classification results reported in Tables 14-21. The DARTS in the tables here correspond only to the distance to nearest decision boundary score. We use MSP[23], maxlogit[24], pNorm [3], temperature scaling [20], generalized entropy [32], and [31]. The central observation is that DARTS consistently outperforms the standard post-hoc scores for both vanilla training and DARTS training.

Across datasets and architectures, AURC and E-AURC show substantial reductions after applying DARTS. This indicates that the risk–coverage behavior improves uniformly: the model becomes more reliable in ranking samples from easy to difficult. On CIFAR-10, these improvements are especially pronounced for RepVGG-A2, where AURC often decreases by 40–60%. Improvements remain visible on CIFAR-100 as well, although the gains are naturally smaller due to the higher intrinsic entropy and class granularity of the dataset. The consistent reductions in both AURC and E-AURC demonstrate that DARTS sharpens the confidence ordering and reduces local irregularities in the score distributions.

DARTS also improves AUROC for nearly all methods. For vanilla MaxSoftmax and MaxLogit, AUROC increases by 3–7 points on CIFAR-10 and by 1–3 points on CIFAR-100. This reflects improved separability between correct and incorrect predictions. These improvements hold across both architectures, reinforcing the observation that DARTS enhances the discriminative structure even without modifying the backbone.

FPR@95 remains a challenging metric, especially for CIFAR-100, where improvements are small and sometimes fluctuate. This is expected, as FPR@95 is highly constrictive with small error margin. Importantly, even when FPR@95 does not decrease, the other metrics—AURC, E-AURC, and AUROC—show consistent gains, indicating that DARTS improves the global geometry while FPR@95 mostly reflects dataset difficulty. On CIFAR-10, however, FPR@95 frequently improves by 5–15%, demonstrating that DARTS reduces high-confidence mistakes in lower-entropy settings.

Architecturally, RepVGG-A2 benefits more strongly from DARTS than ResNet-56. This supports the hypothesis that flatter, more linearly separable feature geometries respond better to margin-oriented supervision. Nonetheless, both architectures exhibit consistent improvements across most metrics. The fact that DARTS enhances performance without modifying backbone training confirms that the classifier head plays a central role in confidence calibration and selective reliability.

Overall, these results demonstrate that DARTS yields broad improvements in selective classification and misclassification detection. By enforcing cleaner geometric margins and more coherent nearest-class relationships, DARTS strengthens both metric-based scoring (pNorm, msp) and classical post-hoc confidence measures (MaxSoftmax, temperature scaling).

I. Margin Sensitivity Analysis

To assess the stability of the proposed DARTS objective under different geometric regularization strengths, we sweep the high-margin and low-margin thresholds over

$$m_{\text{hi}} \in \{0.5,\ 0.7,\ 1.0,\ 2.0\}, \qquad m_{\text{lo}} \in \{0.05,\ 0.15,\ 0.3\}.$$

Figure 4 reports the selective-classification metrics AURC and AUROC for the remaining settings. Two clear patterns emerge: (i) moderate high-margin thresholds ($m_{\text{hi}} \in [0.5, 0.7]$) combined with stronger penalties on hard samples (m_{lo}=0.3) yield the best overall calibration; and (ii) for larger high-margin thresholds ($m_{\text{hi}} \geq 1.0$), strong low-margin penalties become detrimental, and the best performance is achieved with m_{lo}=0.05.

In particular, the configuration

$$m_{\text{hi}} = 0.7, \qquad m_{\text{lo}} = 0.3$$

Table 9. Selective classification performance on **ImageNet-1K** (×10²). Lower values indicate better reliability. All methods use identical backbones and training protocols. Reported values are from a single run due to large-scale compute cost.

Method	ConvNeXt-Base				ResNet-50			
	Risk@80%	AURC	E-AURC	NAURC	Risk@80%	AURC	E-AURC	NAURC
SelectiveNet [18]	8.25±0.12	5.82±0.10	3.75±0.09	57.10±0.48	11.98±0.15	8.12±0.12	5.82±0.10	73.92±0.56
DOCTOR [19]	8.11±0.11	5.67±0.09	3.71±0.08	55.85±0.44	11.79±0.13	7.94±0.11	5.74±0.09	72.51±0.49
Deep Gamblers [34]	8.08±0.10	5.64±0.09	3.69±0.08	55.23±0.41	**11.66**±0.12	**7.70**±0.10	**5.58**±0.08	**71.62**±0.46
SURE [29]	8.02±0.10	5.61±0.08	3.65±0.07	54.62±0.38	11.73±0.11	7.75±0.09	5.63±0.08	71.84±0.41
Vanilla Training (Max Softmax)	7.76±0.09	5.58±0.08	4.14±0.07	60.77±0.40	11.51±0.10	7.83±0.09	5.71±0.08	73.10±0.39
Vanilla Training (Nearest-Boundary)	7.47±0.09	4.06±0.07	2.62±0.06	38.46±0.37	10.37±0.09	5.39±0.08	3.27±0.06	41.80±0.36
DARTS (Max Softmax)	**7.27**±0.08	4.28±0.07	2.84±0.06	41.75±0.35	10.26±0.09	6.09±0.08	4.00±0.07	51.43±0.34
DARTS (Nearest-Boundary)	7.46±0.08	**3.95**±0.06	**2.51**±0.06	**36.86**±0.33	10.47±0.09	5.35±0.07	3.26±0.06	41.90±0.32
Vision Transformer (ViT-Base/16)								
SelectiveNet [18]	8.40±0.12	5.94±0.09	3.82±0.08	58.73±0.49				
DOCTOR [19]	8.23±0.11	5.72±0.09	3.75±0.08	56.95±0.44				
Deep Gamblers [34]	8.19±0.11	5.69±0.08	3.73±0.07	56.44±0.42				
SURE [29]	8.12±0.10	5.65±0.08	3.69±0.07	55.87±0.40				
Vanilla Training (Max Softmax)	9.48±0.09	6.06±0.08	4.04±0.07	52.61±0.39				
Vanilla Training (Nearest-Boundary)	9.88±0.09	5.28±0.07	3.26±0.06	42.45±0.37				
DARTS (Max Softmax)	9.32±0.09	5.92±0.08	3.96±0.07	52.16±0.38				
DARTS (Nearest-Boundary)	9.71±0.09	**5.22**±0.07	**3.26**±0.06	**42.95**±0.36				

Table 10. Average corruption-level improvements on CIFAR-10-C using RepVGG-A2.

Corruption	ΔRisk80 ↑	ΔAURC ↑	ΔE-AURC ↑	ΔN-AURC ↑	ΔFPR95 ↑	ΔAUROC ↑
gaussian noise	0.0277	0.0490	0.0366	0.4936	0.2016	0.0042
shot noise	0.0155	0.0336	0.0296	0.3393	0.2840	0.0083
impulse noise	-0.0267	-0.0095	-0.0013	0.0820	0.1831	-0.0126
defocus blur	0.0005	0.0091	0.0101	0.3052	0.5173	0.0060
glass blur	-0.0243	0.0072	0.0166	0.1774	0.1782	-0.0330
motion blur	0.0011	0.0191	0.0189	0.3275	0.4234	0.0024
zoom blur	0.0016	0.0130	0.0134	0.3036	0.4146	0.0052
snow	-0.0095	0.0105	0.0126	0.2696	0.3973	-0.0052
frost	0.0098	0.0235	0.0222	0.3568	0.4488	-0.0131
fog	0.0044	0.0106	0.0106	0.3120	0.4868	0.0080
brightness	0.0056	0.0074	0.0075	0.3236	0.5802	0.0007
contrast	-0.0015	0.0043	0.0049	0.2011	0.3568	0.0015
elastic transform	-0.0045	0.0102	0.0120	0.2719	0.4038	0.0007
pixelate	0.0001	0.0200	0.0209	0.2974	0.3406	-0.0163
jpeg compression	0.0013	0.0172	0.0171	0.2573	0.3408	-0.0059

achieves the best balance across AURC, EAURC, Risk@95, and FPR@95, while

$$m_{\text{hi}} = 1.0, \qquad m_{\text{lo}} = 0.05$$

obtains the highest AUROC and lowest false positive rate. These trends confirm that DARTS is robust over a broad range of geometric regularization strengths, and that its improvements do not rely on fine-tuning a narrow set of hyperparameters.

Top-2 Logit Difference vs. Nearest-Boundary Distance

A natural baseline for boundary-aware confidence is the difference between the top-2 logits. While this quantity is often used as a proxy for margin, it fails to approximate the true decision-boundary distance used in our method.

Table 22 shows that replacing our nearest-boundary score with the top-2 logit gap during training leads to a substantial degradation across all selective-classification metrics: Risk@95 rises from 0.0041 to 0.0707, AURC increases by 8×, EAURC by 7×, and

Table 11. Average corruption-level improvements on CIFAR-100-C using RepVGG-A2.

Corruption	ΔRisk80 ↑	ΔAURC ↑	ΔE-AURC ↑	ΔN-AURC ↑	ΔFPR95 ↑	ΔAUROC ↑
gaussian noise	0.0100	0.0137	0.0097	0.2418	0.0183	0.0097
shot noise	0.0080	0.0176	0.0086	0.1477	0.0273	0.0062
impulse noise	-0.0035	0.0027	0.0050	0.1298	0.0149	0.0041
defocus blur	0.0040	0.0065	0.0062	0.1283	0.0510	0.0043
glass blur	0.0030	0.0139	0.0125	0.2426	0.0315	0.0062
motion blur	0.0020	0.0139	0.0132	0.2618	0.0363	0.0027
zoom blur	0.0025	0.0150	0.0146	0.2846	0.0385	0.0030
snow	0.0065	0.0166	0.0153	0.2325	0.0321	0.0074
frost	0.0072	0.0195	0.0169	0.2741	0.0364	0.0060
fog	0.0028	0.0104	0.0094	0.2144	0.0225	0.0044
brightness	0.0044	0.0108	0.0095	0.2458	0.0236	0.0035
contrast	0.0010	0.0030	0.0027	0.1321	0.0178	0.0004
elastic transform	0.0002	0.0055	0.0050	0.1740	0.0190	0.0026
pixelate	0.0028	0.0122	0.0109	0.1985	0.0221	0.0041
jpeg compression	0.0035	0.0138	0.0127	0.1897	0.0254	0.0051

Table 12. Average corruption-level improvements on CIFAR-10-C using ResNet-56.

Corruption	ΔRisk80 ↑	ΔAURC ↑	ΔE-AURC ↑	ΔN-AURC ↑	ΔFPR95 ↑	ΔAUROC ↑
gaussian noise	0.0606	0.0510	0.0337	0.6070	0.2920	0.0225
shot noise	0.0250	0.0458	0.0306	0.4540	0.3441	0.0201
impulse noise	-0.0121	0.0081	0.0084	0.2075	0.2624	0.0084
defocus blur	0.0011	0.0154	0.0133	0.3826	0.4847	0.0067
glass blur	-0.0183	0.0088	0.0110	0.2762	0.2298	-0.0146
motion blur	0.0020	0.0237	0.0202	0.4060	0.4520	0.0135
zoom blur	0.0044	0.0210	0.0172	0.3954	0.4221	0.0149
snow	-0.0071	0.0173	0.0145	0.3217	0.3892	0.0036
frost	0.0099	0.0261	0.0211	0.3891	0.4527	0.0044
fog	0.0067	0.0132	0.0113	0.3308	0.4138	0.0063
brightness	0.0075	0.0101	0.0088	0.3617	0.4228	0.0041
contrast	-0.0015	0.0054	0.0048	0.2421	0.3469	0.0012
elastic transform	-0.0033	0.0135	0.0118	0.2875	0.3726	0.0020
pixelate	0.0005	0.0225	0.0200	0.3087	0.2916	-0.0102
jpeg compression	0.0018	0.0187	0.0165	0.2684	0.3003	-0.0064

FPR@95 nearly doubles. AUROC drops from 0.9375 to 0.8691. This demonstrates that the top-2 gap is not conducive to selective classification. We believe the failure stems from the fact that the second-largest logit is not generally the closest class in weight-space, and logit differences are not normalized by classifier geometry.

Table 13. Average corruption-level improvements on CIFAR-100-C using ResNet-56.

Corruption	ΔRisk80 ↑	ΔAURC ↑	ΔE-AURC ↑	ΔN-AURC ↑	ΔFPR95 ↑	ΔAUROC ↑
gaussian noise	0.0114	0.0200	0.0137	0.4978	0.0106	0.0164
shot noise	0.0092	0.0219	0.0138	0.3521	0.0285	0.0109
impulse noise	-0.0041	0.0071	0.0062	0.1992	0.0217	0.0081
defocus blur	0.0020	0.0124	0.0105	0.3055	0.0477	0.0092
glass blur	0.0004	0.0141	0.0134	0.3418	0.0272	0.0119
motion blur	0.0014	0.0187	0.0170	0.3562	0.0300	0.0067
zoom blur	0.0023	0.0180	0.0157	0.3734	0.0313	0.0073
snow	0.0054	0.0222	0.0174	0.3127	0.0301	0.0135
frost	0.0062	0.0262	0.0210	0.3444	0.0317	0.0148
fog	0.0021	0.0095	0.0085	0.2812	0.0220	0.0065
brightness	0.0043	0.0087	0.0079	0.2748	0.0208	0.0042
contrast	0.0005	0.0041	0.0038	0.1957	0.0172	0.0023
elastic transform	-0.0001	0.0099	0.0091	0.2305	0.0204	0.0045
pixelate	0.0027	0.0161	0.0144	0.2536	0.0190	0.0087
jpeg compression	0.0031	0.0155	0.0138	0.2410	0.0202	0.0108

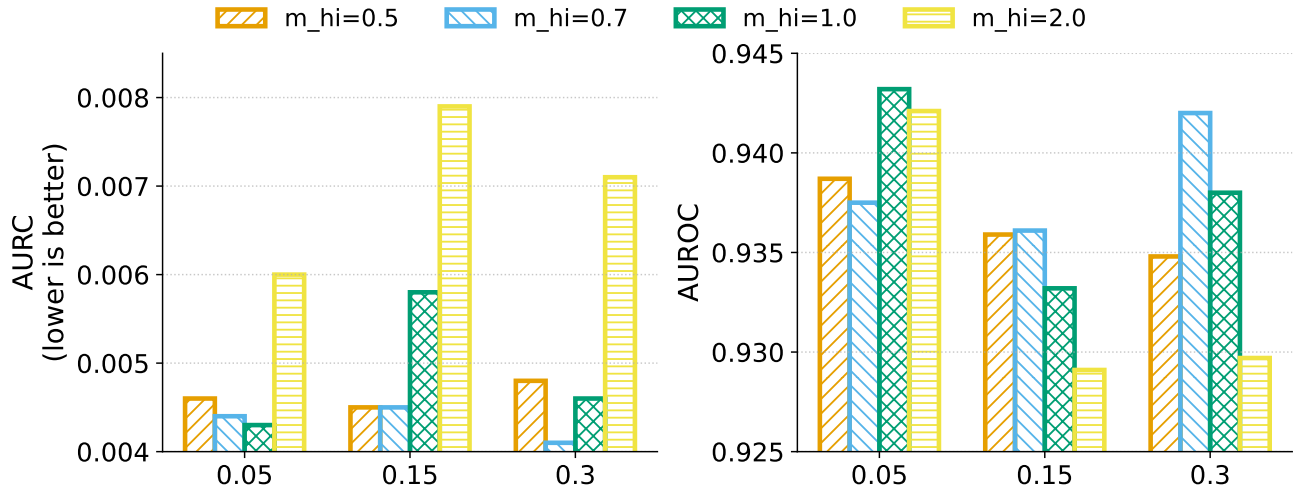


Figure 4. Sensitivity of AURC (left) and AUROC (right) to the margin thresholds $(m_{\text{hi}}, m_{\text{lo}})$. Moderate high-margin values ($m_{\text{hi}} \in [0.5, 0.7]$) with stronger low-margin penalties (m_{lo}=0.3) provide the best overall selective classification performance.

Table 14. Post-hoc misclassification detection and selective classification performance on **CIFAR-10** (RepVGG-A2, Baseline).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	**0.00651**	**0.00560**	**0.00575**	**0.90895**	**0.49755**
msp	0.01114	0.01024	0.00925	0.86897	0.76924
energy	0.01399	0.01308	0.01150	0.84230	0.80850
entropy	0.01171	0.01080	0.01000	0.86388	0.77529
maxlogit	0.01350	0.01259	0.01138	0.84828	0.80422
gen	0.00868	0.00777	0.00813	0.88720	0.71870
pnorm	0.01114	0.01024	0.00925	0.86897	0.76924
temp_scaling	0.01071	0.00980	0.00913	0.87247	0.77028

Table 15. Post-hoc misclassification detection and selective classification performance on **CIFAR-10** (RepVGG-A2, After DARTS).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	**0.00443**	**0.00341**	**0.00413**	**0.93748**	**0.22236**
msp	0.00613	0.00511	0.00587	0.92287	0.29229
energy	0.00937	0.00835	0.01000	0.87931	0.63798
entropy	0.00647	0.00544	0.00613	0.91731	0.33574
maxlogit	0.00786	0.00684	0.00787	0.90012	0.53402
gen	0.00588	0.00486	0.00587	0.92356	0.25838
pnorm	0.00473	0.00371	0.00538	0.93030	0.28570
temp_scaling	0.00599	0.00497	0.00550	0.92494	0.27261

Table 16. Post-hoc misclassification detection and selective classification performance on **CIFAR-10** (ResNet-56, Baseline).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	**0.00738**	**0.00589**	**0.00650**	**0.92114**	**0.26377**
msp	0.01196	0.01047	0.01038	0.88803	0.66350
energy	0.01761	0.01611	0.01675	0.83918	0.83677
entropy	0.01282	0.01133	0.01125	0.88085	0.70356
maxlogit	0.01609	0.01460	0.01438	0.85516	0.81594
gen	0.01015	0.00866	0.00900	0.89973	0.54953
pnorm	0.01196	0.01047	0.01038	0.88803	0.66350
temp_scaling	0.01139	0.00989	0.00975	0.89209	0.63696

Table 17. Post-hoc misclassification detection and selective classification performance on **CIFAR-10** (ResNet-56, After DARTS).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	**0.00703**	**0.00525**	**0.00700**	**0.92899**	**0.21382**
msp	0.00802	0.00625	0.00762	0.92049	0.27503
energy	0.01402	0.01224	0.01712	0.86670	0.57938
entropy	0.00880	0.00702	0.00862	0.91211	0.29979
maxlogit	0.01141	0.00963	0.01225	0.89318	0.44559
gen	0.00843	0.00665	0.00875	0.91387	0.27556
pnorm	0.00802	0.00625	0.00762	0.92049	0.27503
temp_scaling	0.00767	0.00589	0.00762	0.92330	0.26472

Table 18. Post-hoc misclassification detection and selective classification performance on **CIFAR-100** (RepVGG-A2, Baseline).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	**0.05052**	**0.03023**	0.10213	**0.87080**	**0.39739**
msp	0.05468	0.03440	0.10013	0.86455	0.48243
energy	0.06424	0.04396	0.11375	0.83261	0.58386
entropy	0.05753	0.03725	0.10625	0.85354	0.49783
maxlogit	0.06051	0.04022	0.10513	0.84819	0.55568
gen	0.05100	0.03071	**0.09913**	0.87013	0.41316
pnorm	0.05468	0.03440	0.10013	0.86455	0.48243
temp_scaling	0.05389	0.03360	**0.09913**	0.86680	0.46629

Table 19. Post-hoc misclassification detection and selective classification performance on **CIFAR-100** (RepVGG-A2, After DARTS).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	0.05297	0.02909	0.11462	0.87570	0.38918
msp	0.05548	0.03159	0.11175	0.87122	0.42034
energy	0.08864	0.06476	0.14587	0.77059	0.68604
entropy	0.06406	0.04017	0.12475	0.83930	0.48278
maxlogit	0.07478	0.05089	0.12613	0.81628	0.62323
gen	0.05786	0.03398	0.12013	0.85816	0.42414
pnorm	**0.05196**	**0.02808**	0.11300	**0.88075**	**0.38526**
temp_scaling	0.05367	0.02978	**0.11050**	0.87708	0.40388

Table 20. Post-hoc misclassification detection and selective classification performance on **CIFAR-100** (ResNet-56, Baseline).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	**0.07879**	**0.04426**	0.15850	0.84954	**0.47931**
msp	0.08194	0.04740	0.15150	0.84980	0.52911
energy	0.12993	0.09539	0.18263	0.74119	0.80414
entropy	0.09015	0.05561	0.16050	0.82693	0.59666
maxlogit	0.11260	0.07806	0.16363	0.78748	0.75995
gen	0.07935	0.04481	0.15475	0.85088	0.48838
pnorm	0.08194	0.04740	0.15150	0.84980	0.52911
temp_scaling	0.07972	0.04518	**0.15125**	**0.85432**	0.49866

Table 21. Post-hoc misclassification detection and selective classification performance on **CIFAR-100** (ResNet-56, After DARTS).

Method	AURC↓	E-AURC↓	Risk@80↓	AUROC↑	FPR@95↓
DARTS (Boundary Distance)	0.07692	0.04125	0.16200	0.85424	0.43095
msp	0.07447	0.03880	**0.15287**	0.86586	0.43511
energy	0.11554	0.07987	0.18312	0.76810	0.70407
entropy	0.08398	0.04830	0.16325	0.83667	0.49685
maxlogit	0.09841	0.06273	0.16412	0.81272	0.63844
gen	0.08313	0.04746	0.17250	0.83410	0.45739
pnorm	0.07356	0.03788	0.15500	0.86779	0.42451
temp_scaling	**0.07305**	**0.03737**	**0.15287**	**0.86965**	**0.42061**

Table 22. Substituting the top-2 logit difference for the nearest-boundary distance during training causes severe degradation across all selective-classification metrics.

Method	Risk@95	AURC	EAURC	NAURC	FPR@95	AUROC
Nearest-Boundary (DARTS)	0.0041	0.0044	0.0034	0.1594	0.2226	0.9375
Top-2 Logit Gap	0.0707	0.0358	0.0234	0.3654	0.4297	0.8691