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*

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

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 $\gamma(x) < m_{\text{hi}}$,*

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

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

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

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,*

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

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

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

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

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

Applying Dudley's entropy integral [43]:

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

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

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

(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,

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

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

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

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 $\tilde{S}_{\mathrm{DARTS}}$ analogously from $\{\tilde{w}_c, \tilde{b}_c\}$. Then for every x,*

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

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

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

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 $S_{\text{DARTS}}^{(T)}(x) = S_{\text{DARTS}}(x)$ 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

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 λ_{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, $\mathbf{W}$)
      $L_{\text{corr}} \leftarrow \frac{1}{|\mathcal{C}|} \sum_{b \in \mathcal{C}} \big[\max(0,\, m_{\text{hi}} - D_b)\big]^2$
    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, $\mathbf{W}$)
      $L_{\text{wrong}} \leftarrow \frac{1}{|\mathcal{W}|} \sum_{b \in \mathcal{W}} \big[\max(0,\, d_b - m_{\text{lo}})\big]^2$
    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

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

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 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.

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.

Table 9. Selective classification performance on **ImageNet-1K** ($\times 10^2$). 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

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

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

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

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

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

In particular, the configuration

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

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

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

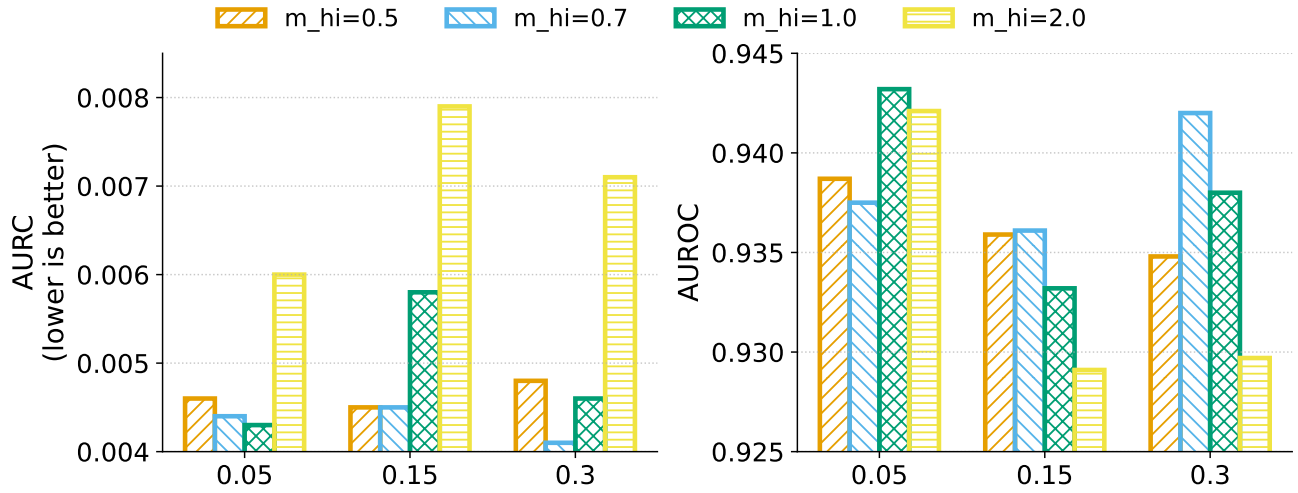

Figure 4. Sensitivity of AURC (left) and AUROC (right) to the margin thresholds $(m_{\text{hi}}, m_{\text{lo}})$. Moderate high-margin values ($m_{\text{hi}} \in [0.5, 0.7]$) with stronger low-margin penalties (m_{lo}=0.3) provide the best overall selective classification performance.

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

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.

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**

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 $8\times$, EAURC by $7\times$, and 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 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