Generative Cross-Entropy: A Strictly Proper Loss for Data-Efficient Classification

Qipeng Zhan[*], **Zhuoping Zhou**[*], **Li Shen**
University of Pennsylvania
{qipengz@sas, zhuopinz@sas, li.shen@pennmedicine}.upenn.edu

Abstract

Cross-entropy (CE) is the default training loss for supervised classification, but its sample efficiency is limited when labels are scarce. Existing remedies primarily act on the data side, via augmentation, synthesis, or transfer from pretrained models; loss-side remedies are comparatively few, and none has displaced CE as the default. We revisit the training objective here. Drawing on the classical observation that generative classifiers reach their asymptotic error with fewer samples than discriminative ones, we propose *Generative Cross-Entropy* (GenCE), a drop-in replacement for CE that borrows the training objective of a generative classifier while the model itself remains a standard discriminative network, without altering the architecture or fitting a separate density model. GenCE follows from a Bayesian rewrite of the class-conditional likelihood and, in the mini-batch approximation, reduces to normalizing each sample's softmax score against the model's predictions on the batch, coupling the training signal across examples sharing a class. We extend the proper-scoring-rule framework to such non-local losses and prove that GenCE is strictly proper under a mild completeness condition: its population risk is uniquely minimized at the true posterior, and its mini-batch version is asymptotically strictly proper with an $O(1/B)$ bias in the batch size B. Across three image datasets, on two architectures and in both balanced small-data and class-imbalanced regimes, GenCE outperforms CE and other widely used losses, including losses designed for small data and baselines tuned per setting, while also producing better-calibrated probabilities and stronger out-of-distribution detection; the gains carry over to ImageNet-1K subsets with more classes than the batch size and to tabular and text data.

1 Introduction

The accuracy of deep neural networks for supervised classification depends strongly on the size of the labeled training set. In application domains such as medical imaging [49], rare-event detection [4], and specialized industrial inspection [56], gathering labels at the scale that modern architectures were designed for is either prohibitively expensive or simply impossible. As a consequence, learning algorithms that remain effective when only a few hundred to a few thousand labeled samples per class are available carry substantial practical value, and they have been studied under various headings: small-data learning [50, 8], low-shot classification [9, 38], and label-efficient training [30, 18].

Existing work in this area has converged on three broad strategies. *Data augmentation* [53, 52, 29] expands the effective training set through handcrafted or learned transformations. *Generative modeling* [41, 42] fits the data distribution and synthesizes additional samples for downstream training. *Transfer and meta-learning* [47, 44] sidesteps scarcity altogether by importing representations from related tasks or large pretrained models. What unifies these approaches is that they all act on the

[*]Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

data: they enlarge, replace, or borrow the training samples. The objective used to fit the classifier, almost invariably cross-entropy, is usually left untouched. Loss-side interventions for small data do exist [1, 20, 25], but they remain far fewer than data-side remedies, and none has displaced cross-entropy as the default [3].

This omission is somewhat curious, since classical theory gives a concrete reason to expect the loss itself to matter more when data are scarce. Ng and Jordan [36] showed that discriminative classifiers eventually attain a lower asymptotic error than their generative counterparts, but generative classifiers approach their asymptote with substantially fewer samples. Zheng et al. [55] later extended it to multiclass settings on top of deep features. Read together, these results suggest that the discriminative–generative trade-off is, at heart, a sample-efficiency trade-off, and that it tilts toward the generative side precisely when labels are scarce.

Building genuine generative classifiers in high-dimensional settings is, however, difficult. Modeling the class-conditional density $p(\boldsymbol{x}|y)$ or the joint density $p(\boldsymbol{x}, y)$ requires either strong distributional assumptions or complex generative architectures, and ensuring proper normalization $\int p(\boldsymbol{x}|y)\,\mathrm{d}\boldsymbol{x} = 1$ is non-trivial. The question we ask in this paper is therefore the following:

> *Can we retain the discriminative model architecture while incorporating the generative learning principle into the training objective?*

Our answer is yes. Applying Bayes' theorem to rewrite the class-conditional likelihood in terms of the discriminative posterior $p_\theta(y|\boldsymbol{x})$ and the marginal $p_\theta(y)$, and approximating the marginal by an empirical average within each mini-batch, we arrive at *Generative Cross-Entropy* (GenCE), a drop-in replacement for cross-entropy. The objective requires no auxiliary generator, no explicit density estimator, and no change to the network. The only structural difference relative to CE is that the per-sample score is normalized against the model's predictions for other examples in the batch, which couples the training signal across samples that share a class. The model itself remains discriminative: GenCE borrows the *training objective* of a generative classifier, not its density model, so [36, 55] motivate rather than guarantee its behavior.

This coupling places GenCE outside the scope of classical proper-scoring-rule theory, in which the per-sample loss is local, and properness can be analyzed pointwise. To handle this, we extend the framework to non-local scoring rules and prove (Theorem 1) that GenCE is strictly proper under a mild completeness condition on the data: in the population limit, its risk is uniquely minimized at the true posterior. GenCE thereby inherits the population-level guarantee that justifies CE in the first place, while attaining it through a generative rather than a pointwise mechanism. The mini-batch objective used in practice is asymptotically strictly proper, with $O(1/B)$ bias in the batch size B (Proposition 1).

Empirically, on CIFAR-10, CIFAR-100, and Mini-ImageNet, and across several network backbones, GenCE consistently outperforms cross-entropy and a range of standard alternatives (focal loss, GCE, MAE, Brier loss) in the small-data regime, and stays ahead of tuned, small-data, and mechanism-matched baselines, on ImageNet-1K subsets, and on tabular and text data. GenCE-trained models are also better calibrated, consistent with the strict-properness analysis, and more reliable on out-of-distribution detection.

Contributions. Our main contributions are summarized as follows:

- We propose GenCE, a loss that borrows the training objective of a generative classifier for discriminative classifier training without modifying the architecture or introducing a separate density model.
- We develop a non-local proper-scoring-rule framework for analyzing losses whose per-sample contributions are coupled, prove that GenCE is strictly proper under a mild completeness condition on the conditional distribution, and show that its mini-batch version is asymptotically strictly proper with $O(1/B)$ bias.
- We empirically show that GenCE improves accuracy, calibration, and OOD detection over cross-entropy and standard alternatives in the small-data regime across multiple datasets and architectures.

2 Related Work

Our work touches on three lines of research: methods that target the small-data regime, methods that import generative structure into classification, and methods that modify the classification loss itself. We summarize each below.

Learning with limited labels. The dominant response to label scarcity has been to act on the data. Classical augmentation expands the training set through geometric and photometric transformations such as cropping, flipping, color jitter, and noise injection, and it remains a standard component of any small-data pipeline. Mixup [53] reframed augmentation as interpolation in input and label space, training the network on convex combinations of pairs of examples and encouraging locally linear behavior between them. The idea has since branched into a family of related methods: Manifold Mixup [46] interpolates in feature space, CutMix [52] mixes localized image regions, and AdaMixUp [14] and AutoMix [29] learn the mixing policy itself. A second line of work replaces hand-designed augmentation with learned synthesis: GANs and VAEs are trained to model the data distribution and produce additional samples, and class-conditional adversarial generators have been used to manufacture examples near the decision boundary in extremely low-data settings. A third line, transfer and meta-learning, sidesteps the data problem altogether by importing representations from large pretrained models or related tasks. Although these three families differ in the mechanism, they share a common locus: all of them operate at the data level, enlarging, replacing, or borrowing samples while the loss used to fit the classifier remains cross-entropy.

Generative perspectives on classification. The discriminative–generative distinction predates deep learning. [36] showed that, for paired models such as logistic regression and naive Bayes, the discriminative model achieves a lower asymptotic error, while the generative model converges to its asymptote with far fewer samples. [55] extended this analysis to multiclass settings under linear evaluation, showing that the sample-efficiency gap between naive Bayes and logistic regression persists on top of deep features. A separate body of work pursues this idea constructively, building hybrid or fully generative deep classifiers, including energy-based models [12] that read off class probabilities from a learned joint density, normalizing-flow classifiers, and diffusion-based classifiers that compare class-conditional likelihoods. Earlier hybrid generative–discriminative formulations were investigated in [23, 39], and generative classifiers have been shown to provide robustness to adversarial attacks [26] and noisy labels [24]. These methods, however, commit to modeling $p(\boldsymbol{x}|y)$ explicitly, and inherit the architectural and optimization cost of doing so. Our approach borrows the motivation from this line of work but not its machinery: we keep a standard discriminative network and import the generative principle through the loss alone.

Modifications to the classification loss. A separate strand of work asks whether cross-entropy is the right objective even when data are plentiful. Focal loss [27] down-weights well-classified examples to focus learning on hard ones, and was later shown to improve calibration as a side effect [32, 45, 10]. Robust regression-style losses such as MAE and MSE on the softmax output have been studied as alternatives that are less sensitive to label noise [17]. Losses targeting small datasets include the cosine loss [1], t-vMF [20], and OLE [25], although carefully tuned CE remains hard to beat [3]. The confidence penalty [37], information maximization [11], and MILC [51] add entropy or class-marginal terms to CE with a tuned weight. Most of these losses are pointwise in the sense of Section 4; the marginal-based ones are non-local but are not analyzed as scoring rules. GenCE departs from this convention: each sample's loss depends on the model's predictions for the rest of the batch, and analyzing such non-local losses requires the framework we introduce.

3 Methodology

3.1 Problem Formulation

We consider a standard supervised K-class classification problem. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be a training set of N i.i.d. samples drawn from a joint distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the feature space and $\mathcal{Y} = \{1, 2, \ldots, K\}$ indexes the K classes. Without loss of generality we take $\mathcal{X} = \text{supp}(\mathcal{P}_\mathcal{X})$, since any region of $\mathbb{R}^d$ outside the support is $\mathcal{P}_\mathcal{X}$-negligible. A forecasting function (classifier) $f : \mathcal{X} \to \Delta^{K-1}$ maps each input feature to a distribution over $\mathcal{Y}$, where $\Delta^{K-1} = \{\boldsymbol{p} \in \mathbb{R}^K : p_k \geq 0, \sum_{k=1}^K p_k = 1\}$ denotes the probability simplex. In modern deep learning, f is instantiated as a parametrized neural network f_θ trained by minimizing a loss over

$\mathcal{D}$. Because $f_\theta(\boldsymbol{x})$ is intended to approximate the true posterior $p(y|\boldsymbol{x})$, we adopt the notation $p_\theta(y|\boldsymbol{x}) := [f_\theta(\boldsymbol{x})]_y$ for the predicted probability assigned to class y given $\boldsymbol{x}$, and write $p_\theta(\cdot|\boldsymbol{x})$ for the full predicted distribution. We further assume that, for each class k, the conditional feature distribution is continuous with density $p(\boldsymbol{x}|y=k)$, so that the true posterior $q(k|\boldsymbol{x}) = \mathcal{P}(y=k|\boldsymbol{x})$ is continuous on $\mathcal{X}$. The classifier $p_\theta(\cdot|\boldsymbol{x})$ is also continuous in $\boldsymbol{x}$, since it is the composition of continuous neural-network operations. These continuity properties enable the density-based analysis used in the derivations that follow and let us drop almost-sure qualifiers in the population-level statements of Section 4.

3.2 Cross-Entropy as Maximum Likelihood Estimation

Before introducing our objective, we briefly recall why cross-entropy is the standard choice. The justification is statistical: minimizing CE is equivalent to maximum likelihood estimation (MLE) under a categorical model. Given the predicted distributions $p_\theta(\cdot|\boldsymbol{x}_i)$, the likelihood of the observed labels under $\mathcal{D}$ is $\prod_{i=1}^N p_\theta(y_i|\boldsymbol{x}_i)$. Taking the negative logarithm and normalizing by N recovers the empirical cross-entropy

$$\mathcal{L}_{\text{CE}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(y_i|\boldsymbol{x}_i) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}_{[y_i=k]} \log p_\theta(k|\boldsymbol{x}_i), \tag{1}$$

where the second equality follows from the one-hot encoding of labels. Cross-entropy is, in this sense, MLE applied to the categorical likelihood.

3.3 Generative Cross-Entropy

While cross-entropy is effective for training classifiers, its purely discriminative nature has known limitations: it does not exploit the generative structure of the data. Prior work [36, 55] shows that generative classifiers, which model the class-conditional distribution $p(\boldsymbol{x}|y)$, often converge in fewer training samples and are less prone to overfitting than their discriminative counterparts, because they leverage additional information about the data distribution. Motivated by this observation, one might consider directly minimizing the conditional log-likelihood

$$\mathcal{L}_{\text{Gen}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i|y_i) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}_{[y_i=k]} \log p_\theta(\boldsymbol{x}_i|k). \tag{2}$$

Doing so, however, requires modeling the high-dimensional density and ensuring the normalization $\int p_\theta(\boldsymbol{x}|y)\,\mathrm{d}\boldsymbol{x} = 1$, which is impractical for the natural-image classifiers we care about. We therefore aim for a less ambitious goal: can we retain a discriminative model $p_\theta(y|\boldsymbol{x})$ while still optimizing the generative objective above? Bayes' theorem provides the bridge:

$$p_\theta(\boldsymbol{x}|y) \;=\; \frac{p_\theta(y|\boldsymbol{x})\,p(\boldsymbol{x})}{p_\theta(y)}, \tag{3}$$

where the marginal class probability $p_\theta(y)$ admits the integral representation

$$p_\theta(y) \;=\; \int p_\theta(y|\boldsymbol{x})\,p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \;=\; \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\big[p_\theta(y|\boldsymbol{x})\big]. \tag{4}$$

Given a training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, we approximate (4) by its empirical average over the training samples:

$$p_\theta(y) \;\approx\; \frac{1}{N}\sum_{j=1}^{N} p_\theta(y|\boldsymbol{x}_j). \tag{5}$$

Substituting (3) and (5) into the generative log-likelihood, the joint negative log-likelihood of the class-conditional model over the training set becomes

$$-\sum_{i=1}^{N} \log p_\theta(\boldsymbol{x}_i|y_i) \;\approx\; -\sum_{i=1}^{N} \log \frac{p_\theta(y_i|\boldsymbol{x}_i)\,p(\boldsymbol{x}_i)}{\frac{1}{N}\sum_{j=1}^{N} p_\theta(y_i|\boldsymbol{x}_j)}. \tag{6}$$

Since $p(\boldsymbol{x}_i)$ and the constant $1/N$ do not depend on θ, they can be omitted from the optimization. We thus define the training objective, termed *Generative Cross-Entropy* (GenCE), as

$$\mathcal{L}_{\text{GenCE}}(\theta) \;=\; -\frac{1}{N}\sum_{i=1}^{N}\log\frac{p_\theta(y_i|\boldsymbol{x}_i)}{\sum_{j=1}^{N}p_\theta(y_i|\boldsymbol{x}_j)}. \tag{7}$$

By construction, $\mathcal{L}_{\text{GenCE}}$ is equivalent to maximizing the class-conditional log-likelihood up to a constant independent of θ and the empirical approximation (5) of the marginal.

Practical considerations. In (7), the denominator $\sum_{j=1}^{N}p_\theta(y_i|\boldsymbol{x}_j)$ requires a sum over the entire training set, which is computationally prohibitive for large datasets. In practice, we approximate this sum within each mini-batch $\mathcal{B}$ of size B:

$$\mathcal{L}^{\mathcal{B}}_{\text{GenCE}}(\theta) \;=\; -\frac{1}{B}\sum_{i\in\mathcal{B}}\log\frac{p_\theta(y_i|\boldsymbol{x}_i)}{\sum_{j\in\mathcal{B}}p_\theta(y_i|\boldsymbol{x}_j)}. \tag{8}$$

The $j = i$ term is included, consistent with (6) and ensuring $\mathcal{L}^{\mathcal{B}}_{\text{GenCE}} \geq 0$. Proposition 1 quantifies the effect of this approximation, and Section 5.6 studies the batch size empirically.

In what sense "generative"? GenCE does not model $p(\boldsymbol{x}|y)$ or generate samples; the model is the same discriminative classifier trained by CE, and only the *training objective*, the class-conditional log-likelihood up to a θ-independent constant, is borrowed from generative classification. Equivalently, (8) is CE plus a batch-level class-marginal term,

$$\mathcal{L}^{\mathcal{B}}_{\text{GenCE}}(\theta) = -\frac{1}{B}\sum_{i\in\mathcal{B}}\log p_\theta(y_i|\boldsymbol{x}_i) + \frac{1}{B}\sum_{i\in\mathcal{B}}\log\sum_{j\in\mathcal{B}}p_\theta(y_i|\boldsymbol{x}_j), \tag{9}$$

whose unit weight is dictated by the likelihood rather than tuned, and is what Theorem 1 relies on. The sample efficiency of generative classifiers [36, 55] is thus a heuristic motivation rather than a result that transfers through the Bayes-rule rewrite; Appendix A.9 shows that generic confidence or marginal regularization does not explain the gains.

4 Theoretical Analysis

In this section, we establish the population-level foundation of GenCE. Our main result (Theorem 1) shows that, despite its non-local structure, GenCE is a strictly proper loss: its population risk is uniquely minimized at the true conditional distribution. As we discuss below, strict properness is precisely the statistical guarantee that aligns loss minimization with probabilistic calibration, and it is the property that justifies cross-entropy as a classification objective in the first place. Theorem 1 shows that GenCE inherits this guarantee.

4.1 From pointwise to non-local scoring rules

The classical justification for cross-entropy goes through the theory of proper scoring rules, which evaluate a forecast against an observed label on a per-sample basis.

Definition 1 (Scoring Rule)**.** *A scoring rule $\mathcal{S}$ is a function*

$$\mathcal{S} : \Delta^{K-1} \times \{1, \ldots, K\} \to \mathbb{R}. \tag{10}$$

Under this definition, the empirical risk decomposes as a mean of independent per-sample contributions, and properness can be analyzed sample-wise. GenCE, however, does not fit this framework. As shown in (7), the contribution of sample $(\boldsymbol{x}_i, y_i)$ to the GenCE objective involves the denominator $\sum_j p_\theta(y_i|\boldsymbol{x}_j)$, which couples the sample's score to every other training point sharing the same class. GenCE, therefore, cannot be written as an expectation of any pointwise scoring rule, and its statistical properties cannot be derived from the classical theory.

To analyze GenCE, we introduce a generalization in which the score is allowed to depend on the entire forecaster across the input space.

Definition 2 (Probability Kernel)**.** *A probability kernel is a mapping $\kappa : \mathcal{X} \to \Delta^{K-1}$, where*

$$\Delta^{K-1} = \{(a_1, \ldots, a_K) \in \mathbb{R}^K : a_k \geq 0, \sum_{k=1}^{K} a_k = 1\}. \tag{11}$$

For any $\boldsymbol{x} \in \mathcal{X}$, the output $\kappa(\boldsymbol{x})$ can be interpreted as a posterior distribution over the K classes conditioned on $\boldsymbol{x}$. With a slight abuse of notation, we write $q(\cdot|\cdot)$ in place of κ throughout the remainder of the paper.

Definition 3 (Non-local Scoring Rule)**.** *Let $\mathcal{K}$ denote the space of probability kernels $q(\cdot|\cdot) : \mathcal{X} \to \Delta^{K-1}$. A non-local scoring rule $\mathcal{S}$ is a functional*

$$\mathcal{S} : \mathcal{K} \times \mathcal{X} \times \{1, \ldots, K\} \to \mathbb{R}. \tag{12}$$

Every pointwise scoring rule is trivially a non-local scoring rule that ignores its first argument, so Definition 3 strictly generalizes Definition 1.

Definition 4 (Population Risk)**.** *Given the data distribution $\mathcal{P}_{\mathcal{X}}$, a probability kernel $q(\cdot|\cdot)$, and a classifier p_θ, the population risk of a non-local scoring rule $\mathcal{S}$ is defined as*

$$\mathcal{R}(\mathcal{S}; \mathcal{P}_{\mathcal{X}}, q, p_\theta) \;=\; \mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}_{\mathcal{X}}} \, \mathbb{E}_{y \sim q(\cdot|\boldsymbol{x})} \big[\mathcal{S}(p_\theta, \boldsymbol{x}, y)\big]. \tag{13}$$

4.2 Strict properness and calibration

We now ask the natural question: under its population risk, does GenCE prefer truthful forecasts?

Definition 5 (Proper and Strictly Proper)**.** *A non-local scoring rule $\mathcal{S} : \mathcal{K} \times \mathcal{X} \times \{1, \ldots, K\} \to \mathbb{R}$ is* proper *if, for every data distribution $\mathcal{P}_{\mathcal{X}}$, every probability kernel $q(\cdot|\cdot)$, and every classifier p_θ,*

$$\mathcal{R}(\mathcal{S}; \mathcal{P}_{\mathcal{X}}, q, q) \;\leq\; \mathcal{R}(\mathcal{S}; \mathcal{P}_{\mathcal{X}}, q, p_\theta), \tag{14}$$

and strictly proper *if equality holds only when $p_\theta = q$.*

Strict properness is the central population-level guarantee that one demands of a probabilistic classification loss. Taking q to be the true posterior $p(\cdot|\boldsymbol{x})$, strict properness states that the unique minimizer of the population risk is the true conditional distribution itself. In other words, a strictly proper loss aligns optimization with calibration: a forecaster that minimizes the population risk produces predictions $p_\theta(\cdot|\boldsymbol{x}) = q(\cdot|\boldsymbol{x})$ almost surely, which is the strongest form of probabilistic calibration. Conversely, any deviation from the true posterior incurs a strictly positive excess risk, so a strictly proper loss exerts a principled, distribution-aware pressure toward calibrated probabilities, even before any finite-sample considerations are taken into account. Cross-entropy is the canonical example of a strictly proper pointwise scoring rule. Our main result shows that GenCE inherits the same property under a mild condition on the data.

Definition 6 (Completeness of a Probability Kernel)**.** *A probability kernel $q(\cdot|\cdot) : \mathcal{X} \to \Delta^{K-1}$ is said to be* complete *if, for every $\boldsymbol{\alpha} \in \mathbb{R}^K$,*

$$\sum_{k=1}^{K} \alpha_k \, q(k|\boldsymbol{x}) = 0 \;\; \textit{for every } \boldsymbol{x} \in \mathcal{X} \iff \boldsymbol{\alpha} = \mathbf{0}. \tag{15}$$

The completeness condition is mild and typically satisfied in practice. Two sufficient conditions are:

1. *Anchor points.* [28] For each class $k \in [K]$ there is at least one point $\boldsymbol{x}_k \in \mathcal{X}$ with $q(k|\boldsymbol{x}_k) > 1/2$.
2. *Gaussian mixture model.* The class-conditional densities are pairwise distinct Gaussians with positive mixing weights.

Proofs of both sufficient conditions are given in Appendix A.2. Intuitively, rescaling $p_\theta(k|\cdot)$ by a class-wise constant leaves (7) unchanged, and completeness is what pins these constants to 1. On CIFAR-10/100 and Mini-ImageNet, every class has images with an unambiguous label, $q(k|\boldsymbol{x}_k) \approx 1$, so the anchor-point condition holds.

We are now ready to state our main theoretical result.

Theorem 1 (Strict Properness of GenCE)**.** *Suppose the probability kernel $q(\cdot|\cdot)$ is complete. Then $\mathcal{L}_{GenCE}$ is uniquely minimized at $p_\theta = q$; that is, $\mathcal{L}_{GenCE}$ is strictly proper.*

Proof. See Appendix A.1. □

Implications for calibration. Theorem 1 guarantees that, in the population limit, the GenCE-optimal classifier recovers the true posterior exactly: the same population guarantee enjoyed by cross-entropy, but reached through the class-conditional likelihood rather than pointwise likelihood matching. Because strict properness concerns the population minimizer, it bears most directly on calibration (Section 5.4); it does not by itself imply higher finite-sample accuracy, and we do not claim a sample-complexity advantage on theoretical grounds (see Section 6).

4.3 From the population risk to the mini-batch objective

Training minimizes (8), which replaces the class marginal $s_k(\theta) = \mathbb{E}_{\boldsymbol{x}}[p_\theta(k|\boldsymbol{x})]$ *inside the logarithm* by its batch average, so the batch loss is a biased estimate of the population risk. Proposition 1 (Appendix A.3) shows that the bias is $\beta(\theta)/B + O(B^{-2})$ for an explicit β, and that the minimizer of the expected batch objective, which SGD follows, deviates from the true posterior by $O(1/B)$. Mini-batch GenCE is thus *asymptotically strictly proper*, recovering Theorem 1 as $B \to \infty$; only the log-denominator contributes to the bias, since the CE part of (9) is unbiased.

5 Experimental Results

5.1 Experimental setup

Datasets. Our benchmarks are **CIFAR-10/100** [21], with 32×32 color images, and **Mini-ImageNet** [48], a 100-class subset of ImageNet [7] at 84×84. The small-data regime uses stratified subsets of size $N \in \{2000, 5000\}$. For class imbalance we adopt the class-imbalanced CIFAR-10/100 variants [6] with imbalance factors $\rho \in \{10, 100\}$. Out-of-distribution (OOD) detection takes CIFAR-10 as in-distribution data, paired with CIFAR-100 (Near-OOD, covariate shift) and SVHN [35] (Far-OOD, semantic shift).

Baselines. We compare GenCE against five widely used training losses: cross-entropy (CE); mean absolute error (MAE), the ℓ_1 distance between the softmax output and the one-hot label; Brier loss [17], its ℓ_2 counterpart; generalized cross-entropy (GCE) [54], a Box–Cox-style interpolation between MAE and CE robust to label noise; and focal loss [27], which down-weights easy examples and is known to improve calibration. GCE is included as the standard MAE–CE interpolation, not as a small-data method. In Tables 1 and 3, GCE and focal loss are tuned per cell by grid search (selected by accuracy); Tables 2 and 4 use the hyperparameters recommended in the original papers. Small-data and mechanism-matched baselines are covered in Section 5.2.

Training and evaluation. We use **ResNet-50** [15] and **VGG-19** [43], trained from scratch with SGD (momentum 0.9, weight decay 5×10^{-4}) for 200 epochs under a step schedule of $0.1 \to 0.01 \to 0.001$ at epochs 0, 100, 150. The default batch size is B=100, and the sensitivity study sweeps $B \in \{25, 50, 100, 200, 250, 300\}$. Augmentation comprises 4-pixel padded crops, horizontal flips, RandAugment [5], MixUp [53] and CutMix [52]. Each configuration is averaged over 5 random seeds; we report mean$_{\pm\text{std}}$ of accuracy, ECE [34] (M=15 equal-width bins), and AUROC for OOD. In wall-clock terms, the within-batch normalization is essentially free: GenCE is on average 0.3% slower per epoch than CE on a single A100.

Details on the datasets, baseline losses, evaluation metrics, and additional comparisons (small-data baselines, mechanism-matched regularizers, long-tail methods, ImageNet-1K, non-image data, supervised contrastive learning, and label smoothing) are provided in Appendices A.4–A.13.

5.2 Classification accuracy in the small-data regime

Table 1 reports test accuracy on twelve configurations; GenCE wins every cell, but the size of the win is what carries the message. Where CE already operates near a data-rich plateau (CIFAR-10, N=5000), the gap is around a point on either backbone. Where CE clearly struggles (CIFAR-100/VGG at N=2000), GenCE yields a gap of 9.5 points. The improvement is also asymmetric across backbones: on CIFAR-100 and Mini-ImageNet, VGG benefits more than ResNet at every sample size, exactly the regime where the network's own architectural prior contributes least. The within-batch coupling acts as a complement to the discriminative signal rather than a replacement, contributing little when CE already fits the data and increasing in importance as that signal weakens.

Stronger baselines and other domains. The gains do not stem from weak baselines (Appendices A.8–A.9). Table 1 already reports focal loss and GCE at their per-cell best γ and q; against t-vMF [20] at its best κ (which covers the cosine loss) it wins eleven and ties one. In the CIFAR-100/VGG/N=2000 cell, regularizers matching the CE-plus-marginal form of (9) [37, 11, 51] reach at most 15.64, versus 24.40 for GenCE. GenCE also beats CE on ImageNet-1K subsets with $K/B \approx 4$ (Appendix A.11) and all baselines on tabular and text data (Appendix A.13).

Table 1: Classification accuracy (%) on balanced small-data subsets across datasets and architectures. Each entry reports mean$_{\pm\text{std}}$ over 5 random seeds. The best result is highlighted in **bold**. †GCE and focal loss are reported at their per-cell best q and γ from a grid search (Appendix A.8), with the selected value in parentheses.

Dataset	N	Model	GenCE	CE	GCE ($q^\star$)†	Focal Loss ($\gamma^\star$)†	MAE	Brier Loss
CIFAR-10	5000	ResNet	$\mathbf{85.20}_{\pm 0.35}$	$84.05_{\pm 0.53}$	$82.87_{\pm 0.24}$ (0.3)	$83.96_{\pm 0.44}$ (1.0)	$83.66_{\pm 0.17}$	$83.82_{\pm 0.20}$
		VGG	$\mathbf{83.54}_{\pm 0.26}$	$83.00_{\pm 0.25}$	$82.71_{\pm 0.41}$ (0.1)	$82.38_{\pm 0.39}$ (2.0)	$82.70_{\pm 0.57}$	$82.71_{\pm 0.32}$
	2000	ResNet	$\mathbf{76.61}_{\pm 0.68}$	$72.51_{\pm 0.94}$	$72.57_{\pm 0.96}$ (0.3)	$74.13_{\pm 0.95}$ (2.0)	$74.17_{\pm 1.01}$	$72.94_{\pm 1.02}$
		VGG	$\mathbf{74.62}_{\pm 0.65}$	$70.38_{\pm 1.93}$	$69.20_{\pm 0.72}$ (0.7)	$69.74_{\pm 0.75}$ (2.0)	$71.14_{\pm 1.54}$	$71.06_{\pm 1.28}$
CIFAR-100	5000	ResNet	$\mathbf{51.74}_{\pm 0.29}$	$50.35_{\pm 0.16}$	$49.11_{\pm 0.34}$ (0.3)	$50.07_{\pm 0.35}$ (2.0)	$49.97_{\pm 0.21}$	$49.30_{\pm 0.44}$
		VGG	$\mathbf{41.30}_{\pm 0.81}$	$35.64_{\pm 2.20}$	$36.33_{\pm 1.26}$ (0.7)	$37.92_{\pm 0.32}$ (1.0)	$37.48_{\pm 1.71}$	$37.73_{\pm 0.76}$
	2000	ResNet	$\mathbf{32.06}_{\pm 0.47}$	$31.15_{\pm 1.11}$	$28.90_{\pm 1.52}$ (0.3)	$31.04_{\pm 1.21}$ (1.0)	$30.83_{\pm 0.31}$	$31.28_{\pm 0.55}$
		VGG	$\mathbf{24.40}_{\pm 1.12}$	$14.92_{\pm 0.58}$	$16.29_{\pm 0.96}$ (0.5)	$15.79_{\pm 0.28}$ (2.0)	$16.18_{\pm 0.59}$	$15.76_{\pm 0.68}$
Mini-ImageNet	5000	ResNet	$\mathbf{43.71}_{\pm 1.04}$	$41.94_{\pm 0.90}$	$41.77_{\pm 1.10}$ (0.1)	$42.30_{\pm 0.38}$ (2.0)	$41.94_{\pm 0.92}$	$42.35_{\pm 0.76}$
		VGG	$\mathbf{32.45}_{\pm 0.92}$	$31.50_{\pm 0.04}$	$32.29_{\pm 0.28}$ (0.3)	$31.46_{\pm 0.89}$ (1.0)	$30.57_{\pm 0.62}$	$30.98_{\pm 0.55}$
	2000	ResNet	$\mathbf{27.75}_{\pm 0.44}$	$24.91_{\pm 0.38}$	$24.32_{\pm 0.79}$ (0.1)	$24.82_{\pm 0.64}$ (0.5)	$25.03_{\pm 0.46}$	$24.76_{\pm 0.07}$
		VGG	$\mathbf{19.68}_{\pm 1.07}$	$16.76_{\pm 0.60}$	$18.65_{\pm 0.57}$ (0.3)	$17.42_{\pm 1.25}$ (1.0)	$17.06_{\pm 0.45}$	$17.09_{\pm 0.55}$

5.3 Class-imbalanced classification

Class-imbalanced labels represent data insufficiency: minority classes lack examples despite a sufficient training budget. This small-sample challenge, which motivates GenCE, reappears here, focused on the class axis rather than across the dataset. Table 2 extends the comparison to class-imbalanced CIFAR-10/100 on a ResNet-50 backbone. GenCE leads in all four cells, and the absolute gain over CE increases as ρ grows from 10 to 100. The mechanism behind this scaling is visible in the loss itself: the denominator $\sum_{j\in\mathcal{B}} p_\theta(y_i|\boldsymbol{x}_j)$ in (8) tracks the model's predicted class marginal, so a head class that is over-represented in the batch inflates its own denominator and dampens its gradient. In contrast, minority-class gradients pass through largely undamped. The result is an implicit rebalancing, no class weights, no resampling, that strengthens precisely as the imbalance worsens. Long-tail methods such as Balanced Softmax [40] and logit adjustment (LA) [31] revert to CE when labels are balanced, so we keep them out of the unified baseline set; because they shift the logits, they can also be composed with GenCE. Appendix A.10 reports both comparisons: adding GenCE on top of Balanced Softmax or LA improves accuracy in all four settings.

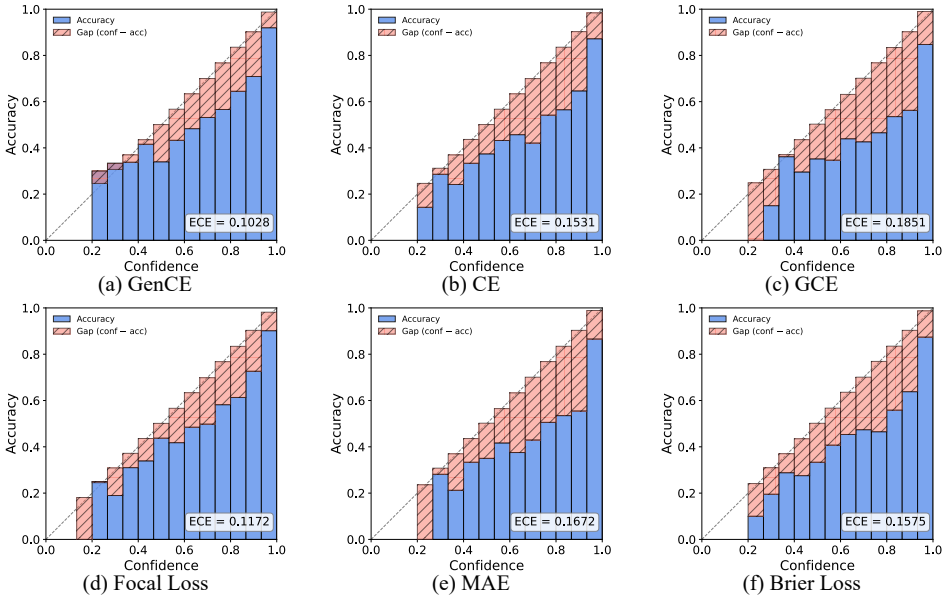
Table 2: Classification accuracy (%) on class-imbalanced CIFAR-10 and CIFAR-100 across different methods. All experiments use ResNet-50. The best result in each row is highlighted in **bold**.

Dataset	ρ	GenCE	CE	GCE	Focal Loss	MAE	Brier Loss
CIFAR-10	10	$\mathbf{87.00}_{\pm 0.23}$	$86.31_{\pm 0.23}$	$86.67_{\pm 0.19}$	$86.44_{\pm 0.32}$	$86.46_{\pm 0.34}$	$86.37_{\pm 0.30}$
	100	$\mathbf{70.84}_{\pm 0.55}$	$69.12_{\pm 0.52}$	$68.38_{\pm 0.11}$	$68.35_{\pm 0.56}$	$68.70_{\pm 0.20}$	$68.34_{\pm 0.41}$
CIFAR-100	10	$\mathbf{59.13}_{\pm 0.28}$	$57.69_{\pm 0.30}$	$56.65_{\pm 0.77}$	$57.64_{\pm 0.25}$	$58.10_{\pm 0.42}$	$57.96_{\pm 0.53}$
	100	$\mathbf{40.31}_{\pm 0.36}$	$38.42_{\pm 0.05}$	$37.36_{\pm 0.51}$	$37.94_{\pm 0.20}$	$37.98_{\pm 0.18}$	$38.42_{\pm 0.52}$

5.4 Calibration error

A classifier is perfectly calibrated when $\mathbb{P}(\hat{y} = y \mid \hat{p} = p) = p$ for every confidence level p. The expected calibration error [34] estimates the deviation by binning predictions into M equal-width confidence intervals and averaging the gap between bin accuracy and bin confidence,

$$\text{ECE} \;=\; \sum_{m=1}^{M} \tfrac{|B_m|}{N} \big|\text{conf}(B_m) - \text{acc}(B_m)\big|. \tag{16}$$



(a) GenCE (b) CE (c) GCE

(d) Focal Loss (e) MAE (f) Brier Loss

Figure 1: Reliability diagrams computed on CIFAR-10 with 15 bins, the dashed line ($y = x$) marks perfect calibration. Bars below (above) the diagonal indicate over- (under-) confidence; the gap between each bar and the diagonal contributes to ECE.

All ECE values are computed from the raw softmax output of the trained classifier; we do not apply temperature scaling [13] or post-hoc rescaling. While effective in data-rich settings, these corrections need a validation split large enough to fit rescaling parameters. The small-data setting of this paper does not support such a split, and using the training set for this purpose would undermine our comparison. GenCE attains the lowest ECE in five of six configurations of Table 3. The single exception is CIFAR-10 with N=5000, where focal loss, an objective engineered to suppress the high-confidence tail, is ahead. Still, it pays more than a point of accuracy for that calibration. The advantage over CE is most pronounced on Mini-ImageNet, where GenCE substantially reduces the calibration gap at both sample sizes. The shape of the loss makes this almost a corollary of how it optimizes: pointwise objectives reward inflating $p_\theta(y_i|\boldsymbol{x}_i)$ on each example independently, whereas GenCE divides that score by the model's average confidence on the same class across the batch, so a uniform inflation is a no-op at the objective level. Calibration improves not because of an extra regularizer but because indiscriminate confidence is no longer rewarded.

Table 3: Expected Calibration Error (ECE, %) across datasets and training-set sizes (ResNet). Lower is better. The best result in each row is in **bold**. GCE and focal loss use the per-cell hyperparameters selected for accuracy in Table 1.

Dataset	N	GenCE	CE	GCE	Focal Loss	MAE	Brier Loss
CIFAR-10	5000	$6.82_{\pm 0.55}$	$7.82_{\pm 0.67}$	$10.27_{\pm 0.27}$	$\mathbf{2.61_{\pm 0.25}}$	$9.17_{\pm 0.19}$	$7.65_{\pm 0.59}$
	2000	$\mathbf{10.64_{\pm 0.39}}$	$14.82_{\pm 0.74}$	$18.72_{\pm 0.45}$	$11.05_{\pm 0.60}$	$16.00_{\pm 0.55}$	$14.38_{\pm 1.28}$
CIFAR-100	5000	$\mathbf{10.25_{\pm 0.69}}$	$15.14_{\pm 0.44}$	$21.35_{\pm 0.18}$	$10.98_{\pm 0.66}$	$17.65_{\pm 0.43}$	$14.26_{\pm 0.77}$
	2000	$\mathbf{22.32_{\pm 0.80}}$	$28.34_{\pm 0.74}$	$37.40_{\pm 0.94}$	$22.62_{\pm 0.90}$	$31.18_{\pm 1.01}$	$27.58_{\pm 0.41}$
Mini-ImageNet	5000	$\mathbf{13.24_{\pm 0.92}}$	$23.00_{\pm 0.71}$	$26.50_{\pm 0.39}$	$18.23_{\pm 1.02}$	$25.04_{\pm 0.61}$	$22.07_{\pm 0.13}$
	2000	$\mathbf{23.82_{\pm 0.17}}$	$32.94_{\pm 0.93}$	$35.37_{\pm 0.41}$	$30.36_{\pm 0.45}$	$34.49_{\pm 0.87}$	$32.55_{\pm 0.59}$

Figure 1 shows the same effect graphically. The pointwise baselines (CE, GCE, Focal Loss, MAE, Brier Loss) all sit below the diagonal in the high-confidence bins—the bins that dominate ECE—while GenCE hugs the diagonal across the full confidence range.

5.5 Out-of-distribution detection

Following [16], we score each input by softmax entropy ($\text{AUC}_{\text{entropy}}$) and by negative maximum softmax probability ($\text{AUC}_{\text{confidence}}$); AUROCs are reported in Table 4. GenCE leads in every cell. The Near-OOD margin is modest (about a point), but the Far-OOD entropy gap is substantial, and

equally telling is the variance: GenCE's seed-to-seed standard deviation is 1.33, against CE's 3.43. The CE failure mode is concrete: on individual seeds, it produces highly confident labels for SVHN digits that bear no resemblance to any CIFAR-10 class, collapsing the entropy gap on those runs. Batch-level normalization quietly suppresses this: a confident answer must be relatively confident relative to the rest of the batch, which is hard to satisfy for uniformly unfamiliar inputs.

Table 4: AUROC (%) of models trained on CIFAR-10 as in-distribution data and evaluated on Near-OOD (CIFAR-100) and Far-OOD (SVHN) regimes. Each entry reports mean$_{\pm\text{std}}$ over 5 random seeds. Higher is better. The best value in each row is in **bold**.

Dataset	Metric	GenCE	CE	GCE	Focal Loss	MAE	Brier Loss
Near-OOD	$\text{AUC}_{\text{entropy}}$	$\mathbf{87.22}_{\pm 0.11}$	$86.24_{\pm 0.47}$	$86.60_{\pm 0.62}$	$86.36_{\pm 0.62}$	$86.18_{\pm 0.20}$	$86.79_{\pm 0.17}$
	$\text{AUC}_{\text{confidence}}$	$\mathbf{86.98}_{\pm 0.09}$	$86.06_{\pm 0.46}$	$86.40_{\pm 0.60}$	$86.14_{\pm 0.58}$	$86.03_{\pm 0.20}$	$86.56_{\pm 0.17}$
Far-OOD	$\text{AUC}_{\text{entropy}}$	$\mathbf{94.87}_{\pm 1.33}$	$88.40_{\pm 3.43}$	$90.41_{\pm 2.09}$	$87.57_{\pm 3.20}$	$91.61_{\pm 1.00}$	$93.84_{\pm 0.50}$
	$\text{AUC}_{\text{confidence}}$	$\mathbf{94.33}_{\pm 1.32}$	$88.16_{\pm 3.19}$	$90.00_{\pm 2.04}$	$87.31_{\pm 3.95}$	$91.34_{\pm 0.89}$	$93.33_{\pm 0.48}$

5.6 Sensitivity analysis

Effect of batch size. Because GenCE ties each sample's score to the rest of its mini-batch, a natural concern is whether the loss becomes more fragile under changes in B than its pointwise counterparts. Sweeping $B \in \{25, \ldots, 300\}$ on $N{=}2000$ CIFAR-10 (Figure 2), all losses peak at intermediate batch sizes: small batches give noisy updates (for GenCE also a noisier marginal estimate, Proposition 1), large ones too few updates per epoch. GenCE peaks at $B{=}100$ and CE at $B{=}200$, and both decline beyond. GenCE stays above CE at every B and is the best loss for all $B \leq 250$; only at $B{=}300$, where all losses have degraded, do GCE, MAE and Brier loss edge slightly ahead (Appendix A.12).

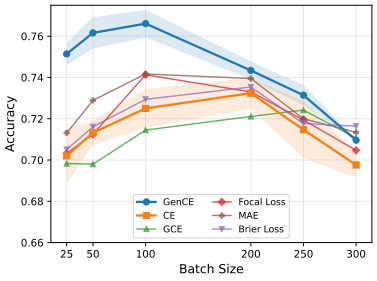


Figure 2: Accuracy vs. batch size B for all losses on $N{=}2000$ CIFAR-10.

6 Conclusion

Generative Cross-Entropy is a small, theoretically motivated change to the cross-entropy objective: each sample's softmax score is divided by the model's average response to the same class across the batch. The modification adds no parameters, no separate density model, and no architectural change, yet it brings the training objective of a generative classifier into a fully discriminative network. We extended the proper-scoring-rule framework to this non-local setting, showed that the population risk of GenCE is uniquely minimized at the true posterior under a mild completeness condition, and showed that the mini-batch objective approaches this guarantee at rate $O(1/B)$. Across CIFAR-10/100 and Mini-ImageNet, two backbones, and both balanced small-data and class-imbalanced regimes, GenCE-trained classifiers are more accurate, better calibrated, and more reliable on out-of-distribution inputs than CE, with the gap widening as data becomes scarcer or more imbalanced—evidence that the loss itself, more than the data, is a useful and under-examined axis for label-efficient learning.

Limitations and future directions. The "generative" interpretation is heuristic: the model remains discriminative, and our theory, which concerns the population minimizer and bears mainly on calibration, gives no finite-sample or sample-complexity result explaining the accuracy gains; this is the main open theoretical question. GenCE also relies on batch estimates of the class marginal: Proposition 1 bounds the bias at $O(1/B)$ for i.i.d. batches, but sensitivity to atypical or strongly imbalanced batch composition is not fully characterized, and class-stratified sampling is a natural remedy. Extending GenCE to speech and structured prediction, and combining it with augmentation or pretraining pipelines, is left to future work.

Acknowledgments

This work was supported in part by NIH grant R01 AG071470 and a PSOM AI2D Seeding Project.

References

[1] Björn Barz and Joachim Denzler. Deep learning on small datasets without pre-training using cosine loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1371–1380, 2020.

[2] Jock A. Blackard and Denis J. Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999.

[3] Lorenzo Brigato, Björn Barz, Luca Iocchi, and Joachim Denzler. Image classification with small datasets: Overview and benchmark. *IEEE Access*, 10:49233–49250, 2022.

[4] Ander Carreño, Iñaki Inza, and Jose A Lozano. Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. *Artificial Intelligence Review*, 53(5):3575–3594, 2020.

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

[7] 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.

[8] Bozheng Dou, Zailiang Zhu, Ekaterina Merkurjev, Lu Ke, Long Chen, Jian Jiang, Yueying Zhu, Jie Liu, Bengong Zhang, and Guo-Wei Wei. Machine learning methods for small data challenges in molecular science. *Chemical Reviews*, 123(13):8736–8780, 2023.

[9] Matthijs Douze, Arthur Szlam, Bharath Hariharan, and Hervé Jégou. Low-shot learning with large-scale diffusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3349–3358, 2018.

[10] Arindam Ghosh, Thomas Schaaf, and Matthew Gormley. Adafocal: Calibration-aware adaptive focal loss. *Advances in Neural Information Processing Systems*, 35:1583–1595, 2022.

[11] Ryan Gomes, Andreas Krause, and Pietro Perona. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems*, volume 23, 2010.

[12] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

[13] 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] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3714–3722, 2019.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[17] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

[18] Cheng Jin, Zhengrui Guo, Yi Lin, Luyang Luo, and Hao Chen. Label-efficient deep learning in medical image analysis: Challenges and future directions. *arXiv preprint arXiv:2303.12484*, 2023.

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[20] Takumi Kobayashi. t-vMF similarity for regularizing intra-class feature distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[22] Ken Lang. NewsWeeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[23] Julia A Lasserre, Christopher M Bishop, and Thomas P Minka. Principled hybrids of generative and discriminative models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 87–94. IEEE, 2006.

[24] Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, Bo Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *International conference on machine learning*, pages 3763–3772. PMLR, 2019.

[25] José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. OLÉ: Orthogonal low-rank embedding, a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8109–8118, 2018.

[26] Yingzhen Li, John Bradshaw, and Yash Sharma. Are generative classifiers more robust to adversarial attacks? In *International Conference on Machine Learning*, pages 3804–3814. PMLR, 2019.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[28] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

[29] Zicheng Liu, Siyuan Li, Di Wu, Zihan Liu, Zhiyuan Chen, Lirong Wu, and Stan Z Li. Automix: Unveiling the power of mixup for stronger classifiers. In *European Conference on Computer Vision*, pages 441–458. Springer, 2022.

[30] Zelun Luo, Yuliang Zou, Judy Hoffman, and Li F Fei-Fei. Label efficient learning of transferable representations acrosss domains and tasks. *Advances in neural information processing systems*, 30, 2017.

[31] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[32] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

[33] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[34] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[35] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, 2011.

[36] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

[37] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. ICLR Workshop.

[38] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018.

[39] Rajat Raina, Yirong Shen, Andrew Mccallum, and Andrew Ng. Classification with hybrid generative/discriminative models. *Advances in neural information processing systems*, 16, 2003.

[40] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, volume 33, pages 4175–4186, 2020.

[41] Lars Ruthotto and Eldad Haber. An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008, 2021.

[42] Ruslan Salakhutdinov. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2(1):361–385, 2015.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 403–412, 2019.

[45] Linwei Tao, Minjing Dong, and Chang Xu. Dual focal loss for calibration. In *International Conference on Machine Learning*, pages 33833–33849. PMLR, 2023.

[46] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019.

[47] Anna Vettoruzzo, Mohamed-Rafik Bouguelia, Joaquin Vanschoren, Thorsteinn Rögnvaldsson, and KC Santosh. Advances and challenges in meta-learning: A technical review. *IEEE transactions on pattern analysis and machine intelligence*, 46(7):4763–4779, 2024.

[48] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[49] Kai Wang. An overview of deep learning based small sample medical imaging classification. In *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, pages 278–281. IEEE, 2021.

[50] Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Small data machine learning in materials science. *npj Computational Materials*, 9(1):42, 2023.

[51] Jirong Yi, Qiaosheng Zhang, Zhen Chen, Qiao Liu, and Wei Shao. Mutual information learned classifiers: an information-theoretic viewpoint of training deep learning classification systems. *arXiv preprint arXiv:2209.10058*, 2022.

[52] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[53] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[54] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[55] Chenyu Zheng, Guoqiang Wu, Fan Bao, Yue Cao, Chongxuan Li, and Jun Zhu. Revisiting discriminative vs. generative classifiers: Theory and implications. In *International conference on machine learning*, pages 42420–42477. PMLR, 2023.

[56] Xiaoqing Zheng, Song Zheng, Yaguang Kong, and Jie Chen. Recent advances in surface defect inspection of industrial products using deep learning techniques. *The International Journal of Advanced Manufacturing Technology*, 113(1):35–58, 2021.

A Technical appendices and supplementary material

A.1 Proof of Theorem 1

The proof of Theorem 1 relies on an expectation form of the log-sum inequality, which we state and prove first; the strict-properness argument then follows.

Lemma 1 (Expectation form of log-sum inequality)**.** *For any two bounded measurable nonnegative functions $h, g : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and any distribution $\mathcal{P}_{\mathcal{X}}$,*

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\mathcal{X}}}\left[g(\boldsymbol{x})\log\frac{g(\boldsymbol{x})}{h(\boldsymbol{x})}\right] \;\geq\; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\mathcal{X}}}[g(\boldsymbol{x})]\log\frac{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\mathcal{X}}}[g(\boldsymbol{x})]}{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\mathcal{X}}}[h(\boldsymbol{x})]}, \tag{17}$$

with equality if and only if $g(\boldsymbol{x}) = c \cdot h(\boldsymbol{x})$ holds $\mathcal{P}_{\mathcal{X}}$-almost surely for some constant $c \geq 0$. Furthermore, if h and g are continuous, then equality holds if and only if $g(\boldsymbol{x}) = c \cdot h(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathcal{X}$.

Proof. Throughout, write $\mathbb{E}[\cdot]$ for $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_{\mathcal{X}}}[\cdot]$ and adopt the standard conventions $0\log 0 := 0$ and, for $a > 0$, $a\log(a/0) := +\infty$.

Reduction to the nondegenerate case. If $\mathbb{E}[h] = 0$, then $h = 0$ a.s. If furthermore $g = 0$ a.s., both sides of (17) equal 0 and equality holds with $c = 0$; otherwise $\mathbb{E}[g] > 0$ and both sides equal $+\infty$. If instead $\mathbb{E}[g] = 0$, then $g = 0$ a.s. and both sides equal 0. We may therefore assume $\mathbb{E}[h] > 0$ and $\mathbb{E}[g] > 0$. Moreover, if $\mathcal{P}_{\mathcal{X}}\{h = 0,\ g > 0\} > 0$, then $g\log(g/h) = +\infty$ on that set, so the left-hand side of (17) is $+\infty$ and the inequality is trivial. We may therefore further assume

$$g(\boldsymbol{x}) = 0 \quad \mathcal{P}_{\mathcal{X}}\text{-a.s. on } \{h = 0\}. \tag{18}$$

An auxiliary probability measure. Define a probability measure $\mathcal{Q}_{\mathcal{X}}$ on $\mathcal{X}$ by

$$\frac{\mathrm{d}\mathcal{Q}_{\mathcal{X}}}{\mathrm{d}\mathcal{P}_{\mathcal{X}}}(\boldsymbol{x}) \;=\; \frac{h(\boldsymbol{x})}{\mathbb{E}[h]}.$$

This is well defined because $h \geq 0$ and $\mathbb{E}[h] > 0$, and it satisfies $\mathcal{Q}_{\mathcal{X}}(\{h = 0\}) = 0$. Define $T : \mathcal{X} \to [0, \infty]$ by

$$T(\boldsymbol{x}) \;=\; \begin{cases} g(\boldsymbol{x})/h(\boldsymbol{x}), & h(\boldsymbol{x}) > 0, \\ 0, & h(\boldsymbol{x}) = 0; \end{cases}$$

the value on $\{h = 0\}$ is immaterial $\mathcal{Q}_{\mathcal{X}}$-a.s.

Jensen's inequality. The function $\varphi(t) = t\log t$, with $\varphi(0) := 0$, is continuous on $[0, \infty)$ and strictly convex on $(0, \infty)$. Jensen's inequality under $\mathcal{Q}_{\mathcal{X}}$ gives

$$\mathbb{E}_{\mathcal{Q}_{\mathcal{X}}}[\varphi(T)] \;\geq\; \varphi(\mathbb{E}_{\mathcal{Q}_{\mathcal{X}}}[T]). \tag{19}$$

Computing both sides under $\mathcal{P}_{\mathcal{X}}$ and using (18):

$$\mathbb{E}_{\mathcal{Q}_{\mathcal{X}}}[T] \;=\; \frac{1}{\mathbb{E}[h]}\,\mathbb{E}[g\,\mathbf{1}_{\{h>0\}}] \;=\; \frac{\mathbb{E}[g]}{\mathbb{E}[h]},$$

and, since $g\log(g/h) = 0$ on $\{h = 0\}$ by (18) and the convention $0\log 0 = 0$,

$$\mathbb{E}_{\mathcal{Q}_{\mathcal{X}}}[\varphi(T)] \;=\; \frac{1}{\mathbb{E}[h]}\,\mathbb{E}\Big[g\log\frac{g}{h}\Big].$$

Substituting into (19) yields

$$\frac{\mathbb{E}[g\log(g/h)]}{\mathbb{E}[h]} \;\geq\; \frac{\mathbb{E}[g]}{\mathbb{E}[h]}\log\frac{\mathbb{E}[g]}{\mathbb{E}[h]}.$$

Multiplying by $\mathbb{E}[h] > 0$ gives (17).

Equality case. Strict convexity of φ on $(0, \infty)$ implies that equality in (19) holds if and only if T is $\mathcal{Q}_{\mathcal{X}}$-a.s. constant; that is, there exists $c \geq 0$ such that $g(\boldsymbol{x})/h(\boldsymbol{x}) = c$ for $\mathcal{P}_{\mathcal{X}}$-a.e. $\boldsymbol{x} \in \{h > 0\}$. Combined with (18), this is equivalent to $g(\boldsymbol{x}) = c \cdot h(\boldsymbol{x})$ $\mathcal{P}_{\mathcal{X}}$-a.s. The converse is immediate. If, in addition, h and g are continuous, then the set $A = \{\boldsymbol{x} \in \mathcal{X} : g(\boldsymbol{x}) = c \cdot h(\boldsymbol{x})\}$ is closed (as the zero set of a continuous function). Since $\mathcal{P}_{\mathcal{X}}(A) = 1$ and $\mathcal{X} = \mathrm{supp}(\mathcal{P}_{\mathcal{X}})$ is the smallest closed set of full measure, $\mathcal{X} \subseteq A \subseteq \mathcal{X}$, and $g(\boldsymbol{x}) = c \cdot h(\boldsymbol{x})$ holds for every $\boldsymbol{x} \in \mathcal{X}$. □

We now turn to the proof of Theorem 1. Recall from Section 3.1 that the true posterior $q(\cdot|\cdot)$ and the classifier $p_\theta(\cdot|\cdot)$ are both continuous on $\mathcal{X}$; this lets us upgrade the equality case of Lemma 1 from "$\mathcal{P}_\mathcal{X}$-a.s." to "everywhere on $\mathcal{X}$" throughout the proof.

Proof of Theorem 1. The population risk of GenCE can be written as

$$\begin{aligned}\mathcal{R}_{\text{GenCE}}(\mathcal{P}_\mathcal{X}, q, p_\theta) &= -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\mathcal{X}}\,\mathbb{E}_{y\sim q(\cdot|\boldsymbol{x})}\left[\log\frac{p_\theta(y|\boldsymbol{x})}{\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{P}_\mathcal{X}}[p_\theta(y|\boldsymbol{x}')]}\right] \\ &= -\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\mathcal{X}}\sum_{k=1}^{K} q(k|\boldsymbol{x})\log\frac{p_\theta(k|\boldsymbol{x})}{\mathbb{E}_{\boldsymbol{x}'\sim\mathcal{P}_\mathcal{X}}[p_\theta(k|\boldsymbol{x}')]}.\end{aligned} \tag{20}$$

Denote $\pi_k := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\mathcal{X}}[q(k|\boldsymbol{x})]$ and $s_k := \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\mathcal{X}}[p_\theta(k|\boldsymbol{x})]$. The excess risk is

$$\begin{aligned}\Delta\mathcal{R} \;&:=\; \mathcal{R}_{\text{GenCE}}(\mathcal{P}_\mathcal{X}, q, p_\theta) - \mathcal{R}_{\text{GenCE}}(\mathcal{P}_\mathcal{X}, q, q) \\ &=\; \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\mathcal{X}}\sum_{k=1}^{K} q(k|\boldsymbol{x})\log\frac{s_k\,q(k|\boldsymbol{x})}{\pi_k\,p_\theta(k|\boldsymbol{x})} \\ &=\; \sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\mathcal{X}}\left[q(k|\boldsymbol{x})\log\frac{q(k|\boldsymbol{x})}{p_\theta(k|\boldsymbol{x})}\right] - \sum_{k=1}^{K}\pi_k\log\frac{\pi_k}{s_k},\end{aligned} \tag{21}$$

where we used $\mathbb{E}_{\boldsymbol{x}}[q(k|\boldsymbol{x})] = \pi_k$ in the last step. By Lemma 1 applied with $g(\boldsymbol{x}) = q(k|\boldsymbol{x})$ and $h(\boldsymbol{x}) = p_\theta(k|\boldsymbol{x})$ — both continuous on $\mathcal{X}$ by assumption — we have, for each $k \in [K]$,

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}_\mathcal{X}}\left[q(k|\boldsymbol{x})\log\frac{q(k|\boldsymbol{x})}{p_\theta(k|\boldsymbol{x})}\right] \;\geq\; \pi_k\log\frac{\pi_k}{s_k}, \tag{22}$$

with equality if and only if there exists a constant $c_k \geq 0$ such that $q(k|\boldsymbol{x}) = c_k\,p_\theta(k|\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathcal{X}$ (the continuous version of the equality case in Lemma 1). Summing over k shows $\Delta\mathcal{R} \geq 0$, with equality if and only if for every k there exists $c_k \geq 0$ with $q(k|\boldsymbol{x}) = c_k\,p_\theta(k|\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathcal{X}$.

It remains to show that, in the equality case, $c_k = 1$ for every k, so that $q = p_\theta$ on $\mathcal{X}$. First, $c_k > 0$ for every k: if $c_k = 0$ for some k, then $q(k|\boldsymbol{x}) = 0$ for every $\boldsymbol{x} \in \mathcal{X}$, which contradicts the completeness of q (taking $\boldsymbol{\alpha}$ to be the k-th canonical vector gives $\sum_j \alpha_j\,q(j|\boldsymbol{x}) = q(k|\boldsymbol{x}) = 0$ for every $\boldsymbol{x} \in \mathcal{X}$ with $\boldsymbol{\alpha} \neq \mathbf{0}$). Hence $p_\theta(k|\boldsymbol{x}) = q(k|\boldsymbol{x})/c_k$ for every $\boldsymbol{x} \in \mathcal{X}$. Combining $\sum_k p_\theta(k|\boldsymbol{x}) = 1$ with $\sum_k q(k|\boldsymbol{x}) = 1$,

$$\sum_{k=1}^{K}\frac{1-c_k}{c_k}\,q(k|\boldsymbol{x}) \;=\; \sum_{k=1}^{K}\left(\frac{1}{c_k}-1\right)q(k|\boldsymbol{x}) \;=\; \sum_{k=1}^{K}p_\theta(k|\boldsymbol{x}) - \sum_{k=1}^{K}q(k|\boldsymbol{x}) \;=\; 0 \quad \text{for every } \boldsymbol{x}\in\mathcal{X}. \tag{23}$$

Setting $\alpha_k := (1-c_k)/c_k$, completeness of q implies $\alpha_k = 0$ for all k, i.e. $c_k = 1$ for all k. Therefore $q(\cdot|\boldsymbol{x}) = p_\theta(\cdot|\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathcal{X}$, which establishes the strict properness of $\mathcal{L}_{\text{GenCE}}$. □

A.2 Sufficient conditions for completeness

We give two sufficient conditions under which the kernel $q(\cdot|\cdot)$ satisfies the completeness condition of Definition 6.

Example 1: anchor points

Assumption. For each class $k \in [K]$ there exists $\boldsymbol{x}_k \in \mathcal{X}$ such that $q(k|\boldsymbol{x}_k) > 1/2$.

Proof. Suppose $\sum_k \alpha_k\,q(k|\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X}$. Set $M := \max_k|\alpha_k|$ and pick $k^\star \in \arg\max_k|\alpha_k|$. We show $M = 0$.

Assume toward contradiction that $M > 0$.

Case 1: $\alpha_{k^\star} = M$. Evaluating the identity at $\boldsymbol{x}_{k^\star}$,

$$0 \;=\; \alpha_{k^\star}\,q(k^\star|\boldsymbol{x}_{k^\star}) + \sum_{k\neq k^\star}\alpha_k\,q(k|\boldsymbol{x}_{k^\star}).$$

Since $\alpha_k \geq -M$ for every k and $\sum_{k \neq k^\star} q(k|\boldsymbol{x}_{k^\star}) = 1 - q(k^\star|\boldsymbol{x}_{k^\star})$,

$$\sum_{k \neq k^\star} \alpha_k \, q(k|\boldsymbol{x}_{k^\star}) \;\geq\; -M\big(1 - q(k^\star|\boldsymbol{x}_{k^\star})\big).$$

Therefore

$$0 \;\geq\; M\, q(k^\star|\boldsymbol{x}_{k^\star}) - M\big(1 - q(k^\star|\boldsymbol{x}_{k^\star})\big) \;=\; M\big(2q(k^\star|\boldsymbol{x}_{k^\star}) - 1\big) \;>\; 0,$$

because $q(k^\star|\boldsymbol{x}_{k^\star}) > 1/2$, a contradiction.

Case 2: $\alpha_{k^\star} = -M$. By the symmetric bound $\alpha_k \leq M$,

$$0 \;\leq\; -M\, q(k^\star|\boldsymbol{x}_{k^\star}) + M\big(1 - q(k^\star|\boldsymbol{x}_{k^\star})\big) \;=\; -M\big(2q(k^\star|\boldsymbol{x}_{k^\star}) - 1\big) \;<\; 0,$$

again a contradiction. Hence $M = 0$ and $\boldsymbol{\alpha} = \mathbf{0}$. □

Example 2: Gaussian mixture model

Assumption. The class-conditional densities are $\mathcal{P}(\boldsymbol{x}|y = k) = \phi_k(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with $\boldsymbol{\Sigma}_k \succ 0$, mixing weights $\pi_k > 0$, and pairwise distinct components $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \neq (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ for $k \neq j$. Then by Bayes' rule,

$$q(k|\boldsymbol{x}) = \mathcal{P}(y = k|\boldsymbol{x}) = \frac{\pi_k \, \phi_k(\boldsymbol{x})}{\sum_j \pi_j \, \phi_j(\boldsymbol{x})}.$$

Proof. Suppose $\sum_k \alpha_k \, q(k|\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$. The denominator of $q(k|\boldsymbol{x})$ is strictly positive, so multiplying through gives

$$\sum_{k=1}^{K} \beta_k \, \phi_k(\boldsymbol{x}) = 0 \qquad \forall \boldsymbol{x} \in \mathbb{R}^d, \qquad \beta_k := \alpha_k \pi_k. \tag{24}$$

Since $\pi_k > 0$, it suffices to show that (24) implies $\boldsymbol{\beta} = \mathbf{0}$, i.e. that distinct Gaussian densities are linearly independent.

Step 1 (analytic continuation). The Fourier transform of ϕ_k is $\widehat{\phi}_k(\boldsymbol{\xi}) = \exp\big(i\boldsymbol{\mu}_k^\top \boldsymbol{\xi} - \tfrac{1}{2}\boldsymbol{\xi}^\top \boldsymbol{\Sigma}_k \boldsymbol{\xi}\big)$. Applying $\mathcal{F}$ to (24),

$$F(\boldsymbol{\xi}) := \sum_k \beta_k \exp\big(i\boldsymbol{\mu}_k^\top \boldsymbol{\xi} - \tfrac{1}{2}\boldsymbol{\xi}^\top \boldsymbol{\Sigma}_k \boldsymbol{\xi}\big) = 0 \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d.$$

Since F is entire on $\mathbb{C}^d$, the identity theorem gives $F \equiv 0$ on $\mathbb{C}^d$. Substituting $\boldsymbol{\xi} = -i\boldsymbol{\eta}$,

$$\sum_k \beta_k \exp\big(\boldsymbol{\mu}_k^\top \boldsymbol{\eta} + \tfrac{1}{2}\boldsymbol{\eta}^\top \boldsymbol{\Sigma}_k \boldsymbol{\eta}\big) = 0 \qquad \forall \boldsymbol{\eta} \in \mathbb{R}^d. \tag{25}$$

Step 2 (separating direction). Choose a unit vector $\mathbf{u} \in \mathbb{R}^d$ such that

- $\mathbf{u}^\top \boldsymbol{\Sigma}_k \mathbf{u} \neq \mathbf{u}^\top \boldsymbol{\Sigma}_j \mathbf{u}$ whenever $\boldsymbol{\Sigma}_k \neq \boldsymbol{\Sigma}_j$,
- $\mathbf{u}^\top \boldsymbol{\mu}_k \neq \mathbf{u}^\top \boldsymbol{\mu}_j$ whenever $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_j$ (so necessarily $\boldsymbol{\mu}_k \neq \boldsymbol{\mu}_j$).

The set of $\mathbf{u}$ violating either condition is a finite union of proper algebraic subvarieties of S^{d-1} (quadrics $\mathbf{u}^\top(\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_j)\mathbf{u} = 0$ and hyperplanes $\mathbf{u}^\top(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j) = 0$), hence has surface measure zero; such $\mathbf{u}$ exists.

Define $a_k := \mathbf{u}^\top \boldsymbol{\mu}_k$ and $b_k := \mathbf{u}^\top \boldsymbol{\Sigma}_k \mathbf{u} > 0$. By construction the pairs $\{(b_k, a_k)\}_{k=1}^K$ are pairwise distinct in lexicographic order. Setting $\boldsymbol{\eta} = t\mathbf{u}$ in (25),

$$h(t) := \sum_k \beta_k \exp\big(\tfrac{1}{2}b_k t^2 + a_k t\big) = 0 \qquad \forall t \in \mathbb{R}. \tag{26}$$

Step 3 (asymptotics). Let $k^\star$ be the unique index that maximizes (b_k, a_k) lexicographically. Dividing (26) by $\exp(\frac{1}{2}b_{k^\star}t^2 + a_{k^\star}t)$,

$$0 = \beta_{k^\star} + \sum_{k \neq k^\star} \beta_k \exp\big(\tfrac{1}{2}(b_k - b_{k^\star})t^2 + (a_k - a_{k^\star})t\big).$$

For each $k \neq k^\star$, either $b_k < b_{k^\star}$ or $b_k = b_{k^\star}$ with $a_k < a_{k^\star}$; in either case the exponent tends to $-\infty$ as $t \to +\infty$. Therefore $\beta_{k^\star} = 0$.

Step 4 (induction). Removing component $k^\star$ leaves $K-1$ pairwise distinct Gaussians satisfying the same identity. Iterating (formally, induction on K, with base case $K = 1$ trivial since a Gaussian is nowhere zero) yields $\beta_k = 0$ for every k, hence $\alpha_k = 0$ for every k. □

Remark 1. *Example 1 is a diagonal-dominance argument: the anchor point $\boldsymbol{x}_k$ isolates class k enough to dominate the contributions from other classes. Example 2 is the classical identifiability of finite Gaussian mixtures, recast as linear independence of the posterior kernel.*

A.3 Mini-batch GenCE: statement and proof

The practical loss (8) replaces $s_k(\theta) = \mathbb{E}_{\boldsymbol{x}}[p_\theta(k|\boldsymbol{x})]$ by $\hat{s}_B(k) = \frac{1}{B}\sum_{j\in\mathcal{B}} p_\theta(k|\boldsymbol{x}_j)$ inside the logarithm. Let $\hat{\mathcal{R}}_B(\theta)$ denote (8) minus the constant $\log B$ on a batch of B i.i.d. samples, $\pi_k = \mathbb{E}_{\boldsymbol{x}}[q(k|\boldsymbol{x})]$, and $\sigma_k^2(\theta) = \mathrm{Var}_{\boldsymbol{x}}[p_\theta(k|\boldsymbol{x})]$.

Proposition 1 (Mini-batch GenCE is asymptotically strictly proper)**.** *If $p_\theta(k|\boldsymbol{x}) \geq \eta > 0$ for all k and $\boldsymbol{x}$, then $\mathbb{E}[\hat{\mathcal{R}}_B(\theta)] = \mathcal{R}(\theta) + \frac{1}{B}\beta(\theta) + O(B^{-2})$ with $\beta(\theta) = \sum_k \Big[\frac{\mathrm{Cov}_{\boldsymbol{x}}(q(k|\boldsymbol{x}),\, p_\theta(k|\boldsymbol{x}))}{s_k(\theta)} - \frac{\pi_k\,\sigma_k^2(\theta)}{2\,s_k^2(\theta)}\Big]$. If, moreover, $p_{\theta^\star} = q$ for some $\theta^\star$ with q complete, and $\nabla^2\mathcal{R}(\theta^\star) =: H \succ 0$ (with the regularity conditions (R1)–(R2) below), the minimizer $\bar{\theta}_B$ of the expected batch objective, which SGD follows, satisfies $\bar{\theta}_B - \theta^\star = -\frac{1}{B}H^{-1}\nabla\beta(\theta^\star) + O(B^{-2})$.*

Let $\{(\boldsymbol{x}_j, y_j)\}_{j=1}^B$ be i.i.d. draws from $\mathcal{P}$, and abbreviate $p_k(\boldsymbol{x}) = p_\theta(k|\boldsymbol{x})$, $q_k(\boldsymbol{x}) = q(k|\boldsymbol{x})$, $s_k = \mathbb{E}_{\boldsymbol{x}}[p_k(\boldsymbol{x})]$, $\hat{s}_k = \frac{1}{B}\sum_{j=1}^B p_k(\boldsymbol{x}_j)$ and $D_k = \hat{s}_k - s_k$. Subtracting $\log B$ from (8),

$$\hat{\mathcal{R}}_B(\theta) = -\frac{1}{B}\sum_{i=1}^B \log p_{y_i}(\boldsymbol{x}_i) + \frac{1}{B}\sum_{i=1}^B \log \hat{s}_{y_i}, \qquad \mathcal{R}(\theta) = -\mathbb{E}\big[\log p_y(\boldsymbol{x})\big] + \sum_{k=1}^K \pi_k \log s_k,$$

where the expression for $\mathcal{R}$ follows from the first display in the proof of Theorem 1.

Part (i): bias of the risk. The first term of $\hat{\mathcal{R}}_B$ is an unbiased estimate of the first term of $\mathcal{R}$. By exchangeability,

$$\mathbb{E}[\hat{\mathcal{R}}_B] - \mathcal{R} = \sum_{k=1}^K \Big(\mathbb{E}\big[\mathbf{1}_{[y_1=k]}\log\hat{s}_k\big] - \pi_k \log s_k\Big).$$

Since $p_k \in [\eta, 1]$, both $\hat{s}_k$ and s_k lie in $[\eta, 1]$, and Taylor's theorem gives $\log\hat{s}_k = \log s_k + \frac{D_k}{s_k} - \frac{D_k^2}{2s_k^2} + \frac{D_k^3}{3s_k^3} + r_k$ with $|r_k| \leq D_k^4/(4\eta^4)$. Decompose $D_k = \frac{1}{B}(p_k(\boldsymbol{x}_1) - s_k) + \frac{B-1}{B}D_k'$, where D_k' is the centered average over $j \geq 2$; it is independent of $(\boldsymbol{x}_1, y_1)$, with $\mathbb{E}[D_k'] = 0$, $\mathbb{E}[D_k'^2] = \sigma_k^2/(B-1)$ and $\mathbb{E}[D_k'^3] = O(B^{-2})$. Using $\mathbb{E}[\mathbf{1}_{[y=k]}|\boldsymbol{x}] = q_k(\boldsymbol{x})$:

$$\mathbb{E}\big[\mathbf{1}_{[y_1=k]}D_k\big] = \tfrac{1}{B}\,\mathbb{E}\big[q_k(\boldsymbol{x})(p_k(\boldsymbol{x}) - s_k)\big] = \tfrac{1}{B}\,\mathrm{Cov}_{\boldsymbol{x}}\big(q_k(\boldsymbol{x}), p_k(\boldsymbol{x})\big),$$
$$\mathbb{E}\big[\mathbf{1}_{[y_1=k]}D_k^2\big] = \tfrac{(B-1)^2}{B^2}\,\pi_k\,\mathbb{E}[D_k'^2] + O(B^{-2}) = \tfrac{1}{B}\,\pi_k\sigma_k^2 + O(B^{-2}),$$

and, expanding the cube in the same way, $\mathbb{E}[\mathbf{1}_{[y_1=k]}D_k^3] = O(B^{-2})$; finally $\mathbb{E}[D_k^4] = O(B^{-2})$ for an average of bounded i.i.d. terms. Collecting terms,

$$\mathbb{E}[\hat{\mathcal{R}}_B(\theta)] - \mathcal{R}(\theta) = \frac{1}{B}\sum_{k=1}^K \left[\frac{\mathrm{Cov}_{\boldsymbol{x}}(q_k, p_k)}{s_k} - \frac{\pi_k\sigma_k^2}{2s_k^2}\right] + O(B^{-2}) = \frac{1}{B}\beta(\theta) + O(B^{-2}).$$

The first term in β stems from including $j = i$ in the denominator and is nonnegative whenever the model's class score is positively correlated with the true posterior; the second is the Jensen gap of the logarithm. At the true posterior, $p = q$ gives $s_k = \pi_k$ and $\beta = \sum_k \sigma_k^2/(2\pi_k) \geq 0$.

Part (ii): shift of the minimizer. We assume, in addition, that (R1) $\theta \mapsto p_\theta(k|\boldsymbol{x})$ is three times continuously differentiable on a neighborhood U of $\theta^\star$, with derivatives up to order three bounded uniformly in $\boldsymbol{x}$ and k; and (R2) $H = \nabla^2\mathcal{R}(\theta^\star) \succ 0$. Under (R1), differentiation and expectation commute, and $\nabla\mathbb{E}[\hat{\mathcal{R}}_B]$ and $\nabla^2\mathbb{E}[\hat{\mathcal{R}}_B]$ are expectations of smooth functions of batch averages of bounded i.i.d. quantities ($\hat{s}_k$, $\nabla\hat{s}_k$, $\nabla^2\hat{s}_k$), with the same single-observation dependence on y_1 as above. The expansion of Part (i), applied term by term, therefore yields $\mathbb{E}[\hat{\mathcal{R}}_B] = \mathcal{R} + \frac{1}{B}\beta + \rho_B$ on U with $\sup_U \|\nabla\rho_B\| = O(B^{-2})$ and $\sup_U \|\nabla^2(\beta/B + \rho_B)\| = O(B^{-1})$; we omit the routine bookkeeping.

By Theorem 1, $\theta^\star$ minimizes $\mathcal{R}$, so $\nabla\mathcal{R}(\theta^\star) = 0$. By (R2) and continuity, $\nabla^2\mathcal{R} \succeq \lambda I$ on a ball $U' \subseteq U$ around $\theta^\star$ for some $\lambda > 0$, so for B large enough $\mathbb{E}[\hat{\mathcal{R}}_B]$ is $\frac{\lambda}{2}$-strongly convex on U'. Its gradient at $\theta^\star$ is $\frac{1}{B}\nabla\beta(\theta^\star) + O(B^{-2}) = O(B^{-1})$, so it has a unique minimizer $\bar{\theta}_B \in U'$ with $\|\bar{\theta}_B - \theta^\star\| = O(B^{-1})$. Expanding the first-order condition $0 = \nabla\mathcal{R}(\bar{\theta}_B) + \frac{1}{B}\nabla\beta(\bar{\theta}_B) + \nabla\rho_B(\bar{\theta}_B)$ around $\theta^\star$,

$$0 = H(\bar{\theta}_B - \theta^\star) + O(\|\bar{\theta}_B - \theta^\star\|^2) + \tfrac{1}{B}\nabla\beta(\theta^\star) + O\big(B^{-1}\|\bar{\theta}_B - \theta^\star\|\big) + O(B^{-2}),$$

hence $\bar{\theta}_B - \theta^\star = -\frac{1}{B}H^{-1}\nabla\beta(\theta^\star) + O(B^{-2})$. Because fresh batches are drawn at every step and $\nabla\hat{\mathcal{R}}_B$ is an unbiased estimate of $\nabla\mathbb{E}[\hat{\mathcal{R}}_B]$, $\bar{\theta}_B$ is the point SGD with batch size B targets. As $B \to \infty$, $\bar{\theta}_B \to \theta^\star$, i.e., $p_{\bar{\theta}_B} \to q$, which recovers Theorem 1. □

Remark 2 (Fluctuation of a single-batch minimizer)**.** *The minimizer $\hat{\theta}_B$ of an individual batch objective $\hat{\mathcal{R}}_B$ additionally fluctuates around $\bar{\theta}_B$. Linearizing $\nabla\hat{\mathcal{R}}_B(\theta^\star)$ in the batch averages (delta method) gives $\nabla\hat{\mathcal{R}}_B(\theta^\star) = \frac{1}{B}\sum_j \psi(\boldsymbol{x}_j, y_j) + o_p(B^{-1/2})$ with mean-zero influence function*

$$\psi(\boldsymbol{x}, y) = -\nabla\log p_{\theta^\star}(y|\boldsymbol{x}) + \nabla\log s_y(\theta^\star) - \sum_{k=1}^{K} p_{\theta^\star}(k|\boldsymbol{x})\,\nabla\log s_k(\theta^\star).$$

Under standard M-estimation conditions, $\sqrt{B}\,(\hat{\theta}_B - \theta^\star) \xrightarrow{d} \mathcal{N}(0, H^{-1}\Sigma H^{-1})$ with $\Sigma = \mathrm{Cov}[\psi(\boldsymbol{x}, y)]$. This mean-zero $O_p(B^{-1/2})$ term accounts for the variability at small batch sizes; the systematic effect of the batch size is the $O(1/B)$ shift of Part (ii).

A.4 Datasets details

This subsection collects general background on the benchmarks used in Section 5; the precise splits and augmentation pipeline are specified in Section 5.1.

CIFAR-10 and CIFAR-100 [21]. Two widely used image-classification benchmarks consist of low-resolution natural images. CIFAR-10 covers ten coarse object categories (e.g. *airplane*, *automobile*, *bird*); CIFAR-100 partitions a similarly sized image collection into one hundred fine-grained classes grouped under twenty superclasses. Both datasets are class-balanced and have become a standard testbed for studying sample efficiency, calibration, and robustness because they are small enough to allow rapid iteration but rich enough to expose meaningful differences between methods.

Mini-ImageNet [48]. A 100-class subset of ImageNet [7], originally introduced as a few-shot benchmark and now used more broadly as a small-scale ImageNet substitute. Compared with CIFAR, Mini-ImageNet has higher input resolution and visually richer images, which makes it a useful intermediate setting between CIFAR-style benchmarks and full ImageNet.

Small-data subsets. The small-data regime ($N \in \{2000, 5000\}$) is constructed by stratified subsampling of the original training partition: each class contributes the same number of samples, drawn under seed-controlled randomness so that different seeds yield different subsamples.

Long-tailed CIFAR-10/100 [6]. A standard class-imbalanced variant of CIFAR-10/100 in which the per-class training count decays exponentially with the class index, controlled by an imbalance factor ρ. The exact schedule and values of ρ used here are stated in Section 5.1.

Out-of-distribution evaluation. The OOD study takes CIFAR-10 as the in-distribution (ID) data. For *Near-OOD*, CIFAR-100 serves as the negative set: it shares low-level image statistics with CIFAR-10 but introduces a semantic shift through new categories. For *Far-OOD*, we use SVHN [35], a digit-recognition dataset captured from natural street-view photographs, whose images differ from those in CIFAR-10 in both content and style.

A.5 Baseline losses

We compare GenCE against five widely used training losses with the formulations given in the corresponding original papers. In Tables 1 and 3, GCE and focal loss are tuned per cell (Appendix A.8); elsewhere, all losses use the hyperparameter values recommended in the original papers. Throughout, $\boldsymbol{p} = p_\theta(\cdot|\boldsymbol{x}) \in \Delta^{K-1}$ denotes the predicted distribution, y the integer label, and p_y the predicted probability of the true class.

- **Cross-entropy (CE):** $\mathcal{L}_{\text{CE}} = -\log p_y$.
- **Mean absolute error (MAE):** $\mathcal{L}_{\text{MAE}} = \sum_{k=1}^{K} |p_k - \mathbf{1}_{[k=y]}|$.
- **Brier loss [17]:** $\mathcal{L}_{\text{Brier}} = \sum_{k=1}^{K} (p_k - \mathbf{1}_{[k=y]})^2$.
- **Generalised cross-entropy (GCE):** $\mathcal{L}_{\text{GCE}} = (1 - p_y^q)/q$ for a hyperparameter $q \in (0, 1]$.
- **Focal loss [27]:** $\mathcal{L}_{\text{Focal}} = -(1 - p_y)^\gamma \log p_y$ for a focusing parameter $\gamma \geq 0$.

The additional comparisons of Appendices A.8–A.10 use the following losses, where $\boldsymbol{z} \in \mathbb{R}^K$ denotes the logits, $\mathbf{f}$ the penultimate feature, $\mathbf{w}_k$ the classifier weight of class k, $H(\cdot)$ the Shannon entropy, and $\bar{\boldsymbol{p}}_{\mathcal{B}} = \frac{1}{B} \sum_{j \in \mathcal{B}} p_\theta(\cdot|\boldsymbol{x}_j)$ the batch-averaged prediction.

- **t-vMF [20]:** CE on logits $z_k = s\, \phi_\kappa(\cos(\mathbf{w}_k, \mathbf{f}))$ with $\phi_\kappa(t) = \frac{1+t}{1+\kappa(1-t)} - 1$. The sharpness κ controls the heavy tail; κ=0 reduces ϕ_κ to the cosine similarity and corresponds to the cosine-similarity objective of [1], and κ=16 is the default of [20].
- **Confidence penalty [37]:** $\mathcal{L}_{\text{CP}} = -\log p_y - \beta\, H\big(p_\theta(\cdot|\boldsymbol{x})\big)$, i.e., entropy regularization.
- **Batch-marginal regularizer:** in the regularized-information-maximization family [11], $\mathcal{L}_{\text{BM}} = \mathcal{L}_{\text{CE}} - \beta\, H(\bar{\boldsymbol{p}}_{\mathcal{B}})$, which acts on the batch-averaged prediction.
- **MILC [51]:** trains on the mutual information $I(X;Y) = H(Y) - H(Y|X)$ rather than the conditional entropy alone; its objective is CE plus a batch class-marginal term weighted by λ.
- **Balanced Softmax [40]:** $-\log \frac{n_y e^{z_y}}{\sum_k n_k e^{z_k}}$, where n_k is the number of training samples of class k.
- **Logit adjustment (LA) [31]:** the logit-adjusted loss $-\log \frac{e^{z_y + \tau \log \pi_y}}{\sum_k e^{z_k + \tau \log \pi_k}}$ with empirical class prior π_k.
- **GenCE + Balanced Softmax / LA:** the long-tail logit shift is applied first, and GenCE (8) is evaluated on the softmax of the shifted logits; predictions at test time use the unshifted logits, as for the two base methods.

GenCE and all baselines share the same architecture, optimizer, learning-rate schedule, augmentation pipeline, and batch composition, so any performance differences are attributable solely to the loss.

A.6 Evaluation metrics

Top-1 accuracy. The fraction of test inputs whose top-scoring class matches the ground truth, evaluated under a deterministic forward pass with the model in evaluation mode.

Expected Calibration Error [34]. ECE quantifies the discrepancy between predicted confidence and empirical accuracy. Predictions are partitioned into M equal-width bins on the maximum-confidence axis, and the per-bin gap between mean confidence and empirical accuracy is averaged with bin-size weights:

$$\text{ECE} \;=\; \sum_{m=1}^{M} \frac{|B_m|}{N} \big|\text{conf}(B_m) - \text{acc}(B_m)\big|,$$

where N is the test-set size. ECE is a standard summary statistic for calibration, but is known to depend on bin count and binning scheme; we therefore fix the binning across all losses, datasets, and seeds for fair comparison. Reliability diagrams (Figure 1) provide a complementary view by plotting per-bin accuracy against confidence.

AUROC for out-of-distribution detection. We treat OOD detection as a binary classification problem: each test input is assigned a scalar score derived from the trained ID classifier, and AUROC is computed with the ID test set as the positive class and the OOD test set as the negative class. AUROC is threshold-free and equals the probability that a uniformly chosen ID input is scored higher than a uniformly chosen OOD input. No temperature scaling or post-hoc calibration is applied before scoring; both detectors operate on the raw softmax output of the trained classifier.

A.7 Extra Comparison

To position GenCE in a broader context, we briefly compare it with two well-known training paradigms—supervised contrastive learning and label smoothing—and report a head-to-head comparison in Table 5.

Supervised contrastive learning (SupCon) [19]. SupCon is a representation-learning objective designed to pull together examples of the same class on the unit hypersphere. Each sample i in a batch serves as an anchor; its ℓ_2-normalized embedding z_i is compared with every other in-batch embedding z_a via cosine similarity, and the loss for i takes the form

$$-\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i\cdot z_a/\tau)},$$

where $P(i)$ collects all other samples of the same class as i and $A(i)$ the rest of the batch. Training proceeds in two stages: the encoder is first optimized with this contrastive objective, then frozen while a separate linear head is fitted using cross-entropy. SupCon thus operates entirely in embedding space and explicitly distinguishes positives from negatives. GenCE differs in each of these dimensions. It operates directly on the classifier's softmax output rather than on ℓ_2-normalized embeddings; the comparison is between samples as candidates for a given class rather than between embedding pairs; and training is single-stage, no projection head, no augmentation-pair construction, and no separately fitted classifier. Whereas SupCon is most naturally analyzed in the mutual-information framework, GenCE admits a proper-scoring-rule analysis (Theorem 1) and yields a population-level uniqueness guarantee at the true posterior that contrastive losses do not provide.

Label smoothing (LS) [33]. Label smoothing replaces the one-hot target $\mathbf{e}_y$ with $(1-\varepsilon)\,\mathbf{e}_y+\varepsilon/K\,\mathbf{1}$, distributing a small mass uniformly across non-target classes; cross-entropy is then evaluated against this softened target. The mechanism is target-side rather than loss-side: the loss remains pointwise, only the regression target changes, and the practical effect is to discourage the network from saturating $p_\theta(y\mid \boldsymbol{x})$ at 1. This intervention sits at a different place in the pipeline than GenCE, which leaves the targets one-hot and instead reshapes the score itself by referencing the rest of the batch. Because the two modifications are largely orthogonal, they compose without conflict, and we evaluate the combination in Table 5 alongside their individual variants. We use the standard smoothing factor $\varepsilon = 0.1$ throughout.

Table 5: Classification accuracy (%) on balanced small-data subsets (N=2000, ResNet-50) across datasets. Each entry reports mean$_{\pm\text{std}}$ over 5 random seeds. Note that SupCon does not support soft targets and is therefore not evaluated under the Label Smoothing setting.

Dataset	GenCE	GenCE w/ Label Smoothing	CE	CE w/ Label Smoothing	SupCon
CIFAR-10	$\mathbf{76.61}_{\pm\mathbf{0.68}}$	$76.48_{\pm 0.50}$	$72.51_{\pm 0.94}$	$73.14_{\pm 0.15}$	$69.48_{\pm 0.43}$
CIFAR-100	$32.06_{\pm 0.47}$	$\mathbf{32.60}_{\pm\mathbf{0.50}}$	$31.15_{\pm 1.11}$	$30.80_{\pm 1.04}$	$23.44_{\pm 0.89}$
Mini-ImageNet	$27.75_{\pm 0.44}$	$\mathbf{27.98}_{\pm\mathbf{0.51}}$	$24.91_{\pm 0.38}$	$26.97_{\pm 0.71}$	$20.50_{\pm 0.43}$

Table 5 surfaces two clear patterns. SupCon lags noticeably behind all other methods on all three datasets, consistent with the well-documented difficulty of training a contrastive encoder when per-class sample density is low, precisely the regime that motivates this paper. The effect of label smoothing on CE, by contrast, is dataset-dependent: it helps on some benchmarks, is roughly neutral on others, and slightly hurts elsewhere, in line with prior reports that the gain from smoothed targets is sensitive to dataset and architecture. The same intervention, applied on top of GenCE, has only a marginal effect overall. Across all configurations, the best results are achieved by GenCE, with or without smoothing, and the gap to both CE+LS and SupCon is preserved.

A.8 Tuned baselines and small-data losses

Protocol. Following Section 5.1 with 5 seeds, we grid-search focal $\gamma \in \{0.5, 1, 2, 3, 5\}$ and GCE $q \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ separately in each of the twelve cells; Table 1 reports both losses at their per-cell best value, with the selected γ and q in parentheses, which favors the baselines more than any single global choice. We also train t-vMF [20] in all twelve cells, sweeping $\kappa \in \{0, 4, 16\}$; κ=16 is the default of [20], and κ=0 corresponds to the cosine-similarity objective [1], so the t-vMF column also covers the cosine loss. Table 6 compares GenCE with t-vMF at its per-cell best κ.

Table 6: Accuracy (%) of GenCE versus t-vMF tuned per cell (mean over 5 seeds); the selected κ is shown in parentheses. GenCE is reported as in Table 1. The best result in each row is in **bold**.

Dataset	N	**Model**	**GenCE**	**t-vMF ($\kappa^\star$)**
CIFAR-10	2000	ResNet	**76.61**	74.40 (4)
		VGG	**74.62**	73.42 (0)
	5000	ResNet	**85.20**	84.50 (0)
		VGG	**83.54**	83.31 (0)
CIFAR-100	2000	ResNet	**32.06**	30.50 (0)
		VGG	**24.40**	19.36 (4)
	5000	ResNet	**51.74**	49.21 (0)
		VGG	**41.30**	40.23 (0)
Mini-ImageNet	2000	ResNet	**27.75**	27.26 (0)
		VGG	**19.68**	18.71 (4)
	5000	ResNet	**43.71**	43.38 (16)
		VGG	**32.45**	**32.45** (4)

Tuning improves both focal loss and GCE relative to their default hyperparameters, but does not change the ranking: GenCE stays ahead of both in all twelve cells (Table 1). t-vMF is the strongest baseline overall and the closest competitor on Mini-ImageNet, yet GenCE outperforms it in eleven cells and ties in one (Mini-ImageNet, N=5000, VGG).

A.9 Mechanism-matched regularizers

By (9), GenCE can be read as CE plus a batch-level class-marginal term, which raises the question of whether its gains reduce to generic confidence or marginal regularization of a comparatively weak CE baseline. We test this in the cell with the largest gain, CIFAR-100/VGG-19, using the confidence penalty [37], a batch-marginal regularizer [11], and MILC [51]. MILC is the closest match, since $I(X;Y) = H(Y) - H(Y|X)$ makes its objective CE plus a class-marginal term. We search $\beta \in \{0.1, 0.5, 1.0\}$ for the first two and $\lambda \in \{0.1, 1, 10, 50\}$ for MILC, under the Section 5.1 protocol with 5 seeds, and report each method at its best coefficient in Table 7.

Table 7: Accuracy (%) of GenCE versus mechanism-matched regularizers on CIFAR-100 with VGG-19 (mean$_{\pm\text{std}}$ over 5 seeds). Each regularizer is shown at its best coefficient (in parentheses).

Loss	N=2000	N=5000
GenCE	$\mathbf{24.40_{\pm 1.12}}$	$\mathbf{41.30_{\pm 0.81}}$
CE	$14.92_{\pm 0.58}$	$35.64_{\pm 2.20}$
Confidence penalty	$15.00_{\pm 1.72}$ (β=1.0)	$38.30_{\pm 0.25}$ (β=0.5)
Batch-marginal regularizer	$15.64_{\pm 0.54}$ (β=0.1)	$37.86_{\pm 0.64}$ (β=0.1)
MILC	$13.91_{\pm 0.60}$ (λ=0.1)	$35.45_{\pm 1.16}$ (λ=0.1)

At N=2000 the strongest regularizer reaches 15.64, above CE but far below GenCE's 24.40, and no coefficient in the grid closes the gap. At N=5000 the confidence penalty and the batch-marginal regularizer recover part of the gap (38.30 and 37.86 versus 35.64), but GenCE remains about three points higher. Generic regularization therefore explains little of GenCE's advantage; the specific unit-weight coupling of (9) matters.

A.10 Composition with long-tail methods

Balanced Softmax [40] and LA [31] adjust the logits by the class prior, so they reduce to CE on balanced data and can be applied on top of GenCE (Appendix A.5). Table 8 compares each method with its GenCE composition on class-imbalanced CIFAR-10/100 (ResNet-50, Section 5.1 protocol, balanced test set, 5 seeds).

Table 8: Accuracy (%) of long-tail methods with and without GenCE on class-imbalanced CIFAR (ResNet-50, mean$_{\pm\text{std}}$ over 5 seeds).

Dataset	ρ	BalSoftmax	GenCE + BalSoftmax	LA	GenCE + LA
CIFAR-10	10	$87.83_{\pm 0.38}$	$\mathbf{88.11_{\pm 0.42}}$	$87.86_{\pm 0.53}$	$\mathbf{88.11_{\pm 0.31}}$
	100	$75.81_{\pm 0.97}$	$\mathbf{76.68_{\pm 0.54}}$	$75.93_{\pm 0.21}$	$\mathbf{77.22_{\pm 0.51}}$
CIFAR-100	10	$59.99_{\pm 0.44}$	$\mathbf{60.52_{\pm 0.13}}$	$59.97_{\pm 0.48}$	$\mathbf{60.64_{\pm 0.60}}$
	100	$42.37_{\pm 0.37}$	$\mathbf{42.87_{\pm 0.65}}$	$41.82_{\pm 0.69}$	$\mathbf{42.67_{\pm 0.48}}$

Adding GenCE improves both methods in all four settings, with the largest gains on CIFAR-10 at ρ=100 ($+0.87$ for Balanced Softmax and $+1.29$ for LA). The two interventions are complementary: the prior shift corrects the label marginal, while GenCE's normalization acts on the model's predicted class marginal within each batch.

A.11 Many classes per batch: ImageNet-1K

When K exceeds B, many classes appear once or not at all in a given batch, and the denominator for a target class is dominated by predictions on samples of other classes. We test this regime directly by training ResNet-50 on stratified subsets of ImageNet-1K [7] with $N \in \{10{,}000,\ 20{,}000\}$ training images and batch size B=256, so that $K/B \approx 4$. Table 9 reports the best top-1 accuracy reached by each loss.

Table 9: Top-1 accuracy (%) on ImageNet-1K subsets (ResNet-50, B=256, K=1000; mean$_{\pm\text{std}}$).

Training set size	GenCE	CE
N=10,000	$\mathbf{13.96_{\pm 0.14}}$	$13.37_{\pm 0.16}$
N=20,000	$\mathbf{20.16_{\pm 0.39}}$	$19.59_{\pm 0.02}$

The batch-level marginal estimate does not break down once K exceeds B: training is stable at $K/B \approx 4$, and GenCE outperforms CE at both sample sizes, the same pattern as on the smaller benchmarks.

A.12 Extended batch-size results

Table 10 lists the numbers behind Figure 2 on N=2000 CIFAR-10 (ResNet-50, 5 seeds), including $B \in \{250, 300\}$. All losses peak at intermediate batch sizes. GenCE peaks at B=100 and CE at B=200, and GenCE is ahead of CE at every batch size. GenCE is the most accurate loss for every $B \leq 250$; at B=300, where all losses have degraded, GCE, MAE and Brier loss are slightly ahead of it.

A.13 Non-image domains

To test whether the gains are specific to natural images, we evaluate all six losses with an MLP backbone (two hidden layers of width 256) on tabular and text data in the small-data regime: UCI Covertype [2] (54 features, 7 classes) and 20 Newsgroups [22] (TF-IDF with 2000 features, 20 classes), each with N=800 stratified training samples. We train with SGD for 200 epochs, with learning rate $0.1 \to 0.01 \to 0.001$ at epochs 100 and 150, no augmentation, and 5 seeds.

GenCE attains the best accuracy on both datasets, so its benefits are not restricted to vision and carry over to tabular and textual inputs.

Table 10: Accuracy (%) versus batch size on N=2000 CIFAR-10 (mean$_{\pm\text{std}}$ over 5 seeds for $B \in \{250, 300\}$ and for GenCE and CE; mean for the other losses at $B \leq 200$). The best loss at each batch size is in **bold**.

B	25	50	100	200	250	300
GenCE	$\mathbf{75.14_{\pm 0.55}}$	$\mathbf{76.16_{\pm 0.76}}$	$\mathbf{76.61_{\pm 0.67}}$	$\mathbf{74.34_{\pm 0.44}}$	$\mathbf{73.14_{\pm 0.49}}$	$70.96_{\pm 0.04}$
CE	$70.20_{\pm 1.35}$	$71.32_{\pm 0.59}$	$72.50_{\pm 0.95}$	$73.24_{\pm 0.74}$	$71.47_{\pm 1.38}$	$69.76_{\pm 0.62}$
GCE	69.83	69.80	71.45	72.11	$72.42_{\pm 0.33}$	$71.07_{\pm 0.71}$
Focal Loss	70.32	71.25	74.13	73.30	$71.94_{\pm 0.75}$	$70.47_{\pm 0.27}$
MAE	71.31	72.89	74.17	73.95	$72.00_{\pm 1.04}$	$71.34_{\pm 1.51}$
Brier Loss	70.51	71.61	72.94	73.54	$71.77_{\pm 1.26}$	$\mathbf{71.64_{\pm 0.98}}$

Table 11: Accuracy (%) on tabular (Covertype) and text (20 Newsgroups) data with an MLP, N=800 (mean$_{\pm\text{std}}$ over 5 seeds).

Dataset	**GenCE**	**CE**	**GCE**	**Focal Loss**	**MAE**	**Brier Loss**
Covertype	$\mathbf{49.56_{\pm 0.46}}$	$47.81_{\pm 1.00}$	$38.10_{\pm 6.37}$	$47.91_{\pm 1.07}$	$47.16_{\pm 0.79}$	$47.77_{\pm 1.05}$
20 Newsgroups	$\mathbf{41.84_{\pm 0.09}}$	$40.92_{\pm 0.64}$	$41.29_{\pm 0.39}$	$40.03_{\pm 0.54}$	$41.14_{\pm 0.49}$	$41.44_{\pm 0.13}$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 5 together reflect the three contributions enumerated at the end of Section 1: a new loss (Section 3.3), a strict-properness theorem under a stated completeness condition (Theorem 1), and an empirical study covering accuracy, calibration, and OOD detection (Sections 5.2–5.5).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated *Limitations and future work* paragraph in the Conclusion (Section 6).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The completeness condition is stated formally in Definition 6, the main theorem (Theorem 1) cites the assumption explicitly, and a complete proof, together with the supporting Lemma 1, is given in Appendix A.1. Two sufficient conditions for completeness, with proofs, are provided in Appendix A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5.1 specifies the datasets, architectures, baselines, and training protocol. Appendix A.4 provides further preprocessing and sampling details. The loss itself is fully specified in equations (7) and (8).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used (CIFAR-10/100, Mini-ImageNet, SVHN) are publicly available and cited in Section 5.1. We provide code in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 5.1 specifies the datasets, splits, baselines, architectures, optimizer family, augmentation, evaluation metrics, and number of seeds; per-loss hyperparameters and additional details are provided in Appendix A.4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Tables 1, 2, and 3 report $\text{mean}_{\pm\text{std}}$ across multiple random seeds, where the variability captures both the random seed and the random subsampling of the small-data training subsets.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The hardware used for our experiments is NVIDIA A100/A40 GPUs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work concerns a generic supervised classification objective and uses publicly available image-classification benchmarks; we have reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work proposes a generic training loss for supervised classification. Like any general-purpose classifier improvement, it could be deployed in beneficial applications (e.g., medical imaging in low-data settings) or harmful ones, but the paper does not introduce a new application or deployment, so we do not identify a direct societal impact specific to the contribution.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new dataset or pretrained model. The proposed loss is an algorithmic contribution and does not introduce assets with a high risk of misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and architectures are cited in Section 5.1 and used in accordance with their original licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper does not release new datasets or pretrained models. Code implementing the proposed loss will be released with the camera-ready version and documented at that time.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as part of the core methodology of this research. Any use of LLMs by the authors was limited to writing, editing, or formatting and did not affect the scientific content.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.