LORENTZ EVIDENTIAL DEEP LEARNING FOR OUT-OF-DISTRIBUTION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidential Deep Learning (EDL) provides an efficient paradigm for open-set recognition by quantifying uncertainty in a single forward pass. However, existing Euclidean-based EDL methods struggle to directly adapt to hyperbolic representation learning, lacking standardized and interpretable rules for evidence construction and valid correlations between Dirichlet parameters and manifold geometric , which leads to inherent limitations in hyperbolic evidence modeling. To address these issues, we present the Lorentz Evidential Deep Learning (LEDL) framework to refine evidence modeling in hyperbolic space. Specifically, we first formulate the Lorentz-decoupling constraint to screen valid Lorentz embedding components from the perspective of semantic reliability and provide geometric rationale for reliable hyperbolic evidence construction. Second, we build a dedicated hyperbolic evidence inference pipeline based on the proposed constraint to achieve stable and standardized Dirichlet parameter generation. Finally, we develop Lorentz Prototype-guided Evidence Alignment to establish explicit mappings between evidence and manifold geometric confidence, enabling close alignment between evidential modeling and hyperbolic geometric properties. Extensive experiments demonstrate that the proposed LEDL achieves competitive open-set detection performance and robustness compared to existing baseline methods.

1 INTRODUCTION

Open-set recognition (Scheirer et al., 2013), the task of identifying unknown samples lying outside the training distribution, serves as a fundamental capability for numerous real world artificial intelligence systems. It finds broad applications in autonomous driving perception (Ho Yang et al., 2025; Asad et al., 2026; He et al., 2025; Shoeb et al., 2025), medical image processing (Cai et al., 2026; Vishal et al., 2026; Zhao et al., 2026), abnormal detection (Liu et al., 2025; Yuan et al., 2025; Xu & Shao, 2025), and is key to embodied intelligence (Ren et al., 2025; Gao et al., 2024). In such practical scenarios, models are required not only to produce predictions for known classes but also to reliably quantify uncertainty and reject unseen or anomalous inputs. Nevertheless, conventional closed-set classifiers tend to yield overconfident wrong predictions for out-of-distribution samples, which limits their deployment within dynamic real-world open environments. This practical demand has fostered substantial research on open-set recognition, covering distance-metric-based approaches, generative approaches, and vision-language strategies (Li et al., 2025). Among these alternatives, evidential deep learning (EDL) (Sensoy et al., 2018) offers a promising framework, since it enables uncertainty estimation within a single forward pass without costly generative sampling or heavy post-processing, bringing favorable computational efficiency.

Rooted in the Dempster-Shafer Theory (DST) (Shafer, 1976) and Subjective Logic (Jøsang, 2001; 2016), EDL generates Dirichlet parameters, treats class probabilities as random variables, and explicitly models uncertainty. Compared with Softmax-based classifiers that only produce pointwise probabilities, EDL exhibits inherent advantages in open-set recognition by explicitly capturing model uncertainty. However, existing EDL methods rely almost entirely on Euclidean representation spaces to construct Dirichlet parameters. In contrast, hyperbolic manifolds offer exponential representation capacity for hierarchical data structures, showing clear advantages in fine-grained and long-tailed recognition tasks. Despite progress in both Euclidean EDL and hyperbolic representation learning, their integration remains insufficiently explored. To bridge this gap, we study the synergy between hyperbolic manifold geometry and subjective evidential learning, developing an in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

terpretable uncertainty estimation framework tailored for hierarchical representations. Specifically, this integration faces three main challenges: the absence of principled Dirichlet parameters formulation in Lorentz space, geometric confounding introduced by the time-radial coordinate z_0 that distorts subjective-logic semantics, and the lack of a stable alignment mechanism between manifold-derived confidence and evidence strength.

To address these challenges, we propose the Lorentz Evidential Deep Learning (LEDL). We first analyze the semantic contamination caused by generating evidence directly from full Lorentz embeddings and formulate a design constraint for semantic validity, providing a basis for semantically reliable evidence construction in hyperbolic space. Building upon this constraint, we establish an evidence modeling paradigm tailored to hyperbolic representations to ensure semantic consistency throughout the evidence generation process. Furthermore, to bridge the gap between evidence strength and manifold-geometric confidence, we introduce Lorentz Prototype-guided Evidence Alignment (LPEA) that aligns Dirichlet parameters with manifold-derived confidence statistics. Extensive experiments on CIFAR10, CIFAR100, CUB-200, Flowers-102, and long-tailed variants CIFAR10-LT and CIFAR100-LT demonstrate the effectiveness of our approach for both classification and open-set recognition.

The main contributions of this work are as follows:

- We propose a Lorentz-decoupling constraint to eliminate geometric confounding induced by time-radial coordinates for establishing valid and reliable evidence generation guidelines in hyperbolic space.
- We build LEDL with a manifold-adaptive evidence mapping network to transform intrinsic directional semantics into Dirichlet parameters for reliable evidence generation.
- We introduce LPEA to bridge representation gaps between geometric confidence and Dirichlet distributions for stable alignment of evidential uncertainty.
- We conduct extensive evaluations across diverse benchmarks to demonstrate closed-set classification effectiveness for establishing reliable open-set rejection capability.

2 Related Work

2.1 Hyperbolic Learning

Hyperbolic learning addresses Euclidean bounds when modeling non-uniform, hierarchical, and long-tailed data. Standard representations include the Poincaré ball, Lorentz model, Klein model, and upper half-space model (Cannon et al., 1997). While the Poincaré ball started hyperbolic neural networks, it suffers from gradient instability near manifold boundaries (Ganea et al., 2018). In contrast, the Lorentz model offers cleaner inner products, providing robust high-dimensional numerical stability and easing gradient vanishing or explosion (Chen et al., 2022). Due to higher compute costs, Klein and upper half-space models see little use in visual tasks (Mettes et al., 2024). Current hyperbolic uncertainty methods generally fall into two categories. The first uses origin distance as an ad-hoc confidence proxy without extra heads. While straightforward, it lacks solid probabilistic grounding and yields inaccurate estimates on long-tailed or noisy inputs (Kasarla et al., 2026; Peng et al., 2026; Lu et al., 2026; Dengxiong & Kong, 2023). The second relies on Bayesian posterior sampling; though mathematically grounded, its heavy sampling burden hinders deployment on large classification benchmarks (Gong et al., 2026). Given its numerical stability and efficient pseudo-Minkowski operations, we build our method on the Lorentz manifold.

While hyperbolic methods excel at capturing non-Euclidean geometry and implicit hierarchies, most literature remains focused on deterministic representations. For visual classification, there is still a lack of evidential frameworks that output explicit uncertainty in a single forward pass, which constrains reliable decision-making under real-world shifts.

2.2 Evidential Deep Learning

Alongside progress in hyperbolic representation learning, EDL enables trustworthy uncertainty quantification grounded in DST and Subjective Logic. By outputting class-wise evidence to parameterize Dirichlet distributions, it computes uncertainty in a single forward pass without expensive

Bayesian sampling. Nevertheless, vanilla EDL inherently suffers from over-confidence, prompting various refinements in Euclidean embedding space. One line revises parameterizations and loss objectives: REDL (Chen et al., 2024) relaxes rigid hyperparameters and variance constraints to temper over-confident predictions, whereas $\mathcal{I}$-EDL (Deng et al., 2023) incorporates Fisher information to measure evidence quality for more dependable estimates. Another line introduces distributional statistics to handle domain shifts: DAEDL (Yoon & Kim, 2024) integrates feature-density statistics into evidential inference to grant distance awareness and boost out-of-distribution (OOD) detection. Beyond these, alternative formulations revisit subjective logic foundations directly: Fan et al. (Fan et al., 2023) decompose overall uncertainty into confusion and ignorance for open-set recognition, FEDL (Yoon & Kim, 2026) adopts evidence normalization to stabilize training dynamics, and HEDL introduces the hyper-opinion concept to refine multi-class evidence modeling (Li et al., 2024; Qu et al., 2024). Despite these multi-faceted refinements to vanilla EDL, all above variants remain strictly confined to Euclidean embedding space.

Consequently, standard EDL formulations and their variants are built on Euclidean assumptions and cannot cope with visual features carrying intrinsic non-Euclidean curvature. Forcing non-uniform image representations with implicit semantic hierarchies into flat Euclidean space distorts native geometric topology, introduces systematic fitting bias into Dirichlet distributions, and yields degraded uncertainty estimates. To bridge this research gap, we extend EDL directly into Lorentz hyperbolic space. By unifying non-Euclidean geometric representation with evidential uncertainty reasoning, our framework maintains topological integrity while delivering reliable uncertainty quantification for classification.

3 METHODS

This section details our LEDL for open-set recognition, as shown in Figure 1. We begin with a brief review of the Lorentz model and EDL. We then uncover the semantic confusion arising from directly applying naive Lorentz representations to evidence parameterization, and theoretically derive necessary geometric constraints to enforce valid evidence construction. To further enhance evidence calibration, we introduce LPEA, followed by the complete training and inference pipeline.

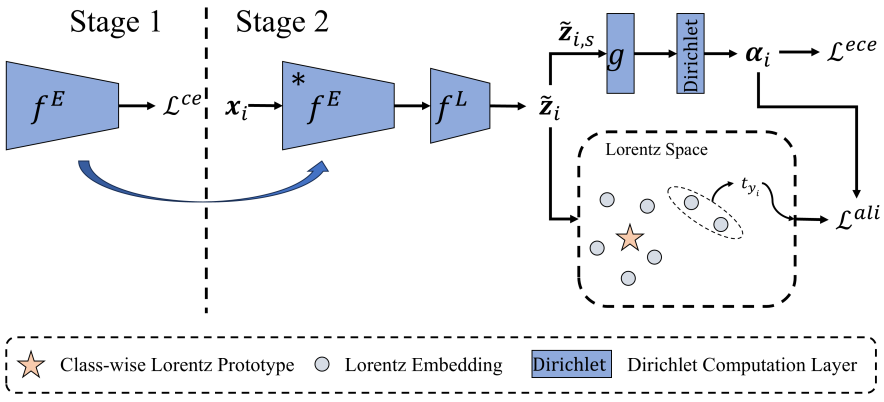


Figure 1: Overview of the LEDL pipeline. Stage-1 pretrains the Euclidean backbone with standard cross-entropy loss. In Stage-2, the backbone is fixed, and images are mapped into Lorentz space. Only the spatial component is used to generate Dirichlet parameters $\boldsymbol{\alpha}_i$. Class prototypes are maintained in Lorentz space to compute dynamic evidence targets for the loss $\mathcal{L}^{ali}$, which is optimized jointly with the Dirichlet cross-entropy loss $\mathcal{L}^{ece}$. The symbol $*$ denotes that the backbone is kept frozen during Stage-2 training.

3.1 PRELIMINARIES

Lorentz Model. The Lorentz model defines n-dimensional hyperbolic space as the upper-sheet hyperboloid embedded in an $(n+1)$-dimensional Minkowski space. For vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{n+1}$, the Lorentz inner product is $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_L = -u_0 v_0 + \sum_{i=1}^{n} u_i v_i$, where index 0 denotes the time coordinate and the remaining components denote space coordinates. The upper-sheet hyperboloid $\mathbb{L}_k^n$ with constant negative curvature $-1/k$ ($k > 0$) collects all vectors satisfying (Chami et al., 2019):

$$\mathbb{L}_k^n = \{\boldsymbol{z} \in \mathbb{R}^{n+1} \mid \langle \boldsymbol{z}, \boldsymbol{z} \rangle_L = -k,\ z_0 > 0\} \tag{1}$$

Any point on $\mathbb{L}_k^n$ decomposes as $\boldsymbol{z} = [z_0, \boldsymbol{z}_s]$. Under the hyperboloid constraint, the time component $z_0 = \sqrt{\|\boldsymbol{z}_s\|_2^2 + k}$ is fully determined by the space component norm, reflecting only radial geometric depth from the origin. In contrast, the spatial part $\boldsymbol{z}_s$ retains directional semantics for class discrimination. Distinguishing radial depth from directional semantics motivates our evidence parameterization.

Evidential Deep Learning. EDL, built upon DST and Subjective Logic, addresses the limitation of Softmax classifiers that predict point-wise probabilities without explicit uncertainty quantification. Instead of soft probabilities, an EDL model outputs non-negative raw evidence $\boldsymbol{e} = [e_1, \ldots, e_C]$, where e_c indicates evidence supporting class c. Raw evidence is converted to Dirichlet parameters via $\boldsymbol{\alpha} = \boldsymbol{e} + 1$, inducing a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha})$ over the $(C-1)$-dimensional probability simplex:

$$\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{c=1}^{C} p_c^{\alpha_c - 1} \tag{2}$$

where $B(\boldsymbol{\alpha})$ is the multivariate Beta function. Total Dirichlet strength is defined as $S = \sum_{c=1}^{C} \alpha_c$, and uncertainty is quantified by $u = C/S$. During inference, the expected class probability $\boldsymbol{p} = \boldsymbol{\alpha}/S$ serves as the classification prediction.

3.2 LORENTZ EVIDENTIAL DEEP LEARNING

Euclidean EDL maps features directly into evidence and Dirichlet parameters within flat spaces. Extending this paradigm to the Lorentz hyperboloid, however, presents distinct geometric conflicts. Two baseline adaptation strategies naturally arise: Strategy-1 directly passes Lorentz similarity scores through non-negative activations to yield evidence, whereas Strategy-2 projects hyperboloid embeddings onto the origin tangent space to execute standard Euclidean EDL pipelines. Appendix B.2 and Appendix B.3 detail their mathematical derivations and failure modes.

Under Strategy-1, Lorentz similarity calculations inherently bind the time-radial component z_0. By manifold construction, z_0 reflects radial depth on the hyperboloid without carrying directional class semantics. Although training can induce class-correlated variations in z_0 via spatial norm shifts, z_0 functions as a geometric confounder within the subjective logic framework. Incorporating z_0 into evidence mixes radial geometric offsets into class belief, disrupting semantic alignment. Conversely, Strategy-2 flattens manifold vectors through tangent projections, distorting learned pairwise distances and hierarchical topology. Thus, both naive options fail—either by corrupting semantic validity or by destroying non-Euclidean inductive biases. This dilemma motivates us to preserve Lorentz representations while isolating the position-dependent time coordinate, parameterizing subjective evidence exclusively through space components.

Given an input sample $\boldsymbol{x}_i$ in the batch $\mathcal{B}$ with size B, let $\boldsymbol{z}_i = [z_{i,0}, \boldsymbol{z}_{i,s}]$ be its Lorentz embedding. The Euclidean backbone f^E first extracts feature vector $\boldsymbol{z}_i^E \in \mathbb{R}^d$. We construct a tangent vector at the origin as $\boldsymbol{z}_i^T = \left(0, \boldsymbol{z}_i^E\right)$, mapping it onto the Lorentz manifold via the origin exponential map $\exp_o^k(\cdot)$:

$$\boldsymbol{z}_i = \exp_o^k\left(\boldsymbol{z}_i^T\right) = \left(\sqrt{k}\cosh\left(\frac{\|\boldsymbol{z}_i^E\|_2}{\sqrt{k}}\right), \sqrt{k}\sinh\left(\frac{\|\boldsymbol{z}_i^E\|_2}{\sqrt{k}}\right)\frac{\boldsymbol{z}_i^E}{\|\boldsymbol{z}_i^E\|_2}\right) \tag{3}$$

Lorentz transformation layer f^L (Shi et al., 2026) then refine this initial embedding $\boldsymbol{z}_i$ into the final vector $\tilde{\boldsymbol{z}}_i$, which strictly satisfies $\langle \tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{z}}_i \rangle_L = -k$. Importantly, the space component $\tilde{\boldsymbol{z}}_{i,s}$ directly captures flat coordinates within the origin tangent space. Because the origin tangent space preserves isotropic directional structures from the manifold, feeding $\tilde{\boldsymbol{z}}_{i,s}$ to subsequent Euclidean layers remains geometrically sound.

Assumption 1 (Semantic Alignment Premise)**.** *In classification-oriented evidential modeling, subjective evidence generated for parameterized Dirichlet distributions should encode only discriminative information relevant to class prediction, remaining unpolluted by geometry-driven radial offsets induced by manifold constraints.*

Principle 1 (Spatial Parameterization Principle for Lorentz Evidence)**.** *Let $\tilde{\boldsymbol{z}}_i = [\tilde{z}_{i,0}, \tilde{\boldsymbol{z}}_{i,s}] \in \mathbb{L}_k^{d+1}$ denote a Lorentz embedding on the hyperboloid. Under Assumption 1, deriving Dirichlet parameters $\boldsymbol{\alpha}_i$ from the full vector $[\tilde{z}_{i,0}, \tilde{\boldsymbol{z}}_{i,s}]$ causes evidential parameterization to absorb purely radial offsets from $\tilde{z}_{i,0}$. To enforce semantic alignment and semantic validity , $\boldsymbol{\alpha}_i$ must be parameterized strictly through the space component $\tilde{\boldsymbol{z}}_{i,s}$.*

Adhering to this criterion, we discard the time coordinate and feed only space features $\tilde{\boldsymbol{z}}_{i,s}$ into the evidence network g. Exponentiating these scalar outputs with 1 yields the Dirichlet concentration parameter vector $\boldsymbol{\alpha}_i$:

$$\boldsymbol{\alpha}_i = \exp\left(g(\tilde{\boldsymbol{z}}_{i,s})\right) + 1 \tag{4}$$

This guarantee ($\boldsymbol{\alpha}_i \geq 1$) ensures a mathematically valid Dirichlet distribution. Model optimization minimizes the expected negative log-likelihood under the Dirichlet distribution, yielding the closed-form formulation:

$$\begin{aligned} \mathcal{L}_i^{ece}(\boldsymbol{\alpha}_i) &= \int \left[\sum_{c=1}^{C} -y_{ic} \log p_{ic} \right] \mathrm{Dir}(\boldsymbol{p}_i | \boldsymbol{\alpha}_i) d\boldsymbol{p}_i \\ &= \sum_{c=1}^{C} y_{ic} \left(\psi(S_i) - \psi(\alpha_{ic}) \right) \end{aligned} \tag{5}$$

where y_{ic} represents one-hot labels, $\psi(\cdot)$ denotes the digamma function, and $S_i = \sum_c \alpha_{ic}$ reflects total Dirichlet strength for sample i. Averaging across a mini-batch yields the classification loss $\mathcal{L}^{ece}$:

$$\mathcal{L}^{ece} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_i^{ece}(\boldsymbol{\alpha}_i) \tag{6}$$

3.3 Lorentz Prototype-guided Evidence Alignment

While spatial-only evidence parameterization guarantees interpretable inference, this constraint alone cannot correct topological defects within learned Lorentz embeddings. Under manifold pathologies such as heavy intra-class dispersion or inter-class overlap, formally valid Dirichlet parameters still suffer from evidence miscalibration, degrading the reliability of uncertainty estimates. To address this limitation, we introduce LPEA to establish an explicit mapping between hyperbolic manifold deviation statistics and evidence strength. Our method extracts geometry-aware deviation metrics directly from hyperboloid embeddings to guide running target statistics, achieving precise alignment between evidence magnitudes and manifold representations.

Computing manifold deviation statistics from individual embeddings leaves the estimator sensitive to noise and outliers. We therefore maintain a set of class-wise Lorentz prototypes $\{\boldsymbol{\mu}_c\}_{c=1}^{C}$ as stable geometric anchors during training, freezing them during validation and testing. To avoid the heavy cost of exact hyperbolic Fréchet mean optimizations, we adopt momentum-based exponential moving averages ($m_p \in (0, 1)$) in the ambient Minkowski space followed by manifold re-projection. When class embeddings remain reasonably localized along radial directions, this ambient centroid scheme serves as an efficient surrogate. For each iteration, we compute mini-batch Euclidean centroids for present classes, map them to the hyperboloid as $\boldsymbol{\mu}^{\mathcal{B}}$, apply EMA updates, and project the updated vectors back onto the manifold. Missing classes in a batch retain their prior prototype states:

$$\boldsymbol{\mu}_c = \mathrm{project}\big(m_p \boldsymbol{\mu}_c^{old} + (1 - m_p)\, \boldsymbol{\mu}_c^{\mathcal{B}}\big) \tag{7}$$

where $\mathrm{project}(\cdot)$ maps ambient vectors back onto the Lorentz space.

To quantify sample manifold deviation relative to class anchors, we construct a prototype-aware deviation metric s_{ic} derived from the Lorentz inner product. Compared to full geodesic distance computations, this linearly scaled inner-product formulation serves as a first-order surrogate for local hyperbolic divergence, drastically reducing computational complexity while mapping deviation

signals into a bounded range:

$$s_{ic} = -\frac{\langle \tilde{\boldsymbol{z}}_i, \boldsymbol{\mu}_c \rangle_L + k}{2k} \tag{8}$$

When a sample closely aligns with its class prototype ($\langle \tilde{\boldsymbol{z}}_i, \boldsymbol{\mu}_c \rangle_L \to -k$), s_{ic} approaches 0, monotonically increasing as the embedding departs from the class anchor. This deviation metric ranks samples by manifold divergence and selects the top-s large-deviation samples within each class to establish target evidence baselines.

Static evidence targets fail to accommodate representation shifts during training. We instead track per-class dynamic target statistics $\{t_c\}_{c=1}^{C}$ that evolve alongside learned representations, bridging manifold deviation statistics with Dirichlet parameter scale. Under intra-class dispersion, these boundary-like samples help impose a lower bound for evidence strength and mitigate pathological low-evidence outputs. To filter noisy interference, we build a large-deviation subset $\mathcal{I}_c^{top}$ from samples ranking in the top-s fraction of prototype-relative deviation within class c (retaining at least one sample). Batch-level targets $t_c^{\mathcal{B}}$ are computed over this filtered subset when at least two samples are present:

$$t_c^{\mathcal{B}} = \begin{cases} \frac{1}{|\mathcal{I}_c^{top}|} \sum_{i \in \mathcal{I}_c^{top}} \alpha_{ic}, & |\mathcal{I}_c^{top}| \geq 2 \\ t_c^{old}, & \text{otherwise.} \end{cases} \tag{9}$$

Because the peripheral subset contains only a fraction of samples per class, batch-level targets are particularly sensitive to sampling fluctuation. We therefore smooth these scalar statistics across iterations using EMA temporal filtering with momentum $m_t \in (0, 1)$, securing a stable target trajectory t_c. Initialization starts from uninformative Dirichlet priors with a scale offset: $t_c^{(0)} = 1.0 + \lambda_{cls}\, C$:

$$t_c = m_t\, t_c^{old} + (1 - m_t)\, t_c^{\mathcal{B}} \tag{10}$$

Without explicit alignment constraints, evidence magnitudes can become mis-aligned against manifold-derived statistics under intra-class dispersion and manifold pathologies. We impose an evidence alignment loss $\mathcal{L}^{ali}$ to regularize ground-truth-class evidence toward dynamic manifold statistics, correcting misalignment between evidential outputs and geometric representations and meanwhile mitigating pathological low-evidence outputs.

$$\mathcal{L}^{ali} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{ReLU}\big(t_{y_i} - \alpha_i\big) \tag{11}$$

where y_i denotes the true class label for sample $\boldsymbol{x}_i$, and α_i is its corresponding ground-truth Dirichlet concentration parameter. This loss only activates when sample evidence falls below manifold-derived target statistics and imposes no penalty for evidence values exceeding the target.

3.4 Training and Inference Pipeline

The overall framework is trained in two distinct phases: we first pre-train the Euclidean backbone independently, and subsequently freeze its weights to optimize the Lorentz evidential components. The objective function for the second phase combines the evidential classification loss with the alignment term:

$$\mathcal{L}^{total} = \mathcal{L}^{ece} + w^{ali} \mathcal{L}^{ali} \tag{12}$$

During training, parameters of the Lorentz transform, evidence sub-network, and running EMA stats are updated. At inference time, all statistical states remain frozen, and Dirichlet parameters alongside uncertainty are calculated using only spatial parts of the Lorentz embeddings. We summarize the step-by-step execution flow in Algorithm B.6.

4 Experiments

4.1 Setup

Datasets. We use four datasets CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), Flowers102 (Nilsback & Zisserman, 2008) and CUB200 (Wah et al., 2011) for in-distribution

(ID) datasets, while other three datasets SVHN (Netzer et al., 2011), Describable Textures Dataset (DTD) (Cimpoi et al., 2014) and Places365 (Zhou et al., 2018) for OOD datasets. See Appendix C.1 for more details.

Evaluation Metrics. We adopt False-Positive-Rate at 95% True-Positive-Rate (FPR95), Area Under the Precision-Recall Curve (AUPR), Area Under the Receiver Operating Characteristic Curve (AUROC), and closed-set classification Accuracy (Acc) as our evaluation metrics; detailed definitions and mathematical formulations are provided in Appendix C.2.

Implementation Details. We largely follow the experimental protocol from HEDL (Qu et al., 2024). ResNet-18 is used for CIFAR10, CIFAR100 and long-tailed CIFAR-LT variants, while ResNet-34 serves as backbone for fine-grained Flowers102 and CUB200. Long-tail imbalance is controlled by parameter η, defined as the ratio between tail-class sample count and head-class sample count (Yoon & Kim, 2026),; $\eta = 0.1$ corresponds to moderate imbalance and $\eta = 0.01$ corresponds to severe imbalance. We adopt a two-stage training scheme: pre-train the backbone with softmax for 90 epochs then fine-tune our LEDL module for 10 epochs. The hyperboloid scale parameter is fixed to $k = 1.0$. All of our reproduced results are evaluated over five distinct random seeds, reporting mean and standard deviation. Full hyperparameter settings, hardware and preprocessing configurations are provided in Appendix C.3. Average OOD performance denotes metric values averaged over the three OOD datasets SVHN, DTD, and Places365.

4.2 STATE-OF-THE-ART

Table 1: Results of OOD detection performance for ours against other state-of-the-art methods with CIFAR10 and CIFAR100 as ID datasets. Except for Ours, the remaining compared methods are reported in the HEDL (Qu et al., 2024).

Method	CIFAR10				CIFAR100			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
MSP (Hendrycks & Gimpel, 2018)	56.47	79.90	89.54	95.06	82.91	69.46	77.47	77.25
ODIN (Liang et al., 2020)	56.51	58.22	78.86	95.06	83.13	66.62	76.77	77.25
openGAN (Kong & Ramanan, 2025)	99.02	38.31	44.12	95.06	86.21	59.96	74.12	77.25
GradNorm (Huang et al., 2021)	94.07	71.19	55.49	95.06	85.91	78.33	70.41	77.25
VIM (Wang et al., 2022)	27.57	87.98	94.46	95.06	74.18	73.74	81.46	77.25
KNN (Sun et al., 2022)	41.04	89.47	93.23	95.06	73.13	72.79	81.71	77.25
DICE (Sun & Li, 2022)	63.78	72.03	83.66	95.06	80.40	71.37	79.28	77.25
RankFeat (Song et al., 2022)	55.97	76.79	75.87	95.06	67.59	73.19	68.45	77.25
ASH (Djurisic et al., 2022)	82.04	85.32	76.93	95.06	56.74	84.47	81.69	77.25
SHE (Zhang et al., 2023)	74.57	88.87	83.61	95.06	65.89	80.68	76.97	77.25
GEN (Liu et al., 2023)	38.64	92.58	90.52	95.06	57.64	82.26	80.14	77.25
MCDropout (Gal & Ghahramani, 2016)	52.47	81.32	89.98	94.95	77.11	71.60	79.48	75.83
G-ODIN (Hsu et al., 2020)	23.84	89.39	94.67	94.70	69.44	76.09	84.21	74.46
CSI (Tack et al., 2020)	27.09	92.37	91.82	91.16	75.71	77.13	71.15	61.60
MOS (Huang & Li, 2021)	82.71	80.03	59.38	94.83	93.18	78.64	55.59	76.98
VOS (Du et al., 2022)	37.55	80.13	91.27	**95.82**	96.99	58.59	68.51	77.20
LogitNorm (Wei et al., 2022)	22.47	94.04	95.97	94.30	82.14	72.58	78.13	76.34
RED (Pandey & Yu, 2023)	74.87	40.41	46.84	95.80	71.40	74.58	82.05	**80.36**
Ours	**17.54**	**94.89**	**96.99**	95.32	**36.50**	**88.32**	**92.24**	80.10
±Std	0.90	0.25	0.16	0.09	2.04	0.69	0.43	0.13

Table 1 summarizes the performance comparison between our method and existing OOD baselines on CIFAR10 and CIFAR100. Existing heuristic, generative, and distance-based methods exhibit clear performance drops across benchmarks, while effective on simple CIFAR10, their OOD discrimination degrades when handling CIFAR100 with a larger class space. Our approach achieves strong OOD metrics on both CIFAR10 (FPR95 = 17.54%, AUROC = 96.99%) and CIFAR100 (FPR95 = 36.50%, AUROC = 92.24%). Meanwhile, these gains introduce no severe in-distribution degradation, maintaining ID accuracies of 95.32% and 80.10%, respectively. These empirical results indicate that combining Lorentz manifold representations with dynamic evidence regularization better balances ID classification fidelity and OOD sensitivity. Supplementary evaluations on fine-grained benchmarks (Flowers102 and CUB200) are provided in Appendix D (Table 6).

Table 2: OOD detection performance comparison between our method and other EDL-based state-of-the-art methods, using CIFAR10 and CIFAR100 as in-distribution datasets. 0.00* denotes an extremely small standard deviation (< 0.01) that rounds to 0.00 when formatted to two significant figures.

Method	CIFAR10				CIFAR100			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
EDL	32.15 ± 1.09	84.18 ± 0.40	91.81 ± 0.31	95.23 ± 0.04	79.09 ± 0.85	71.51 ± 0.41	75.63 ± 0.51	80.38 ± 0.13
$\mathcal{I}$-EDL	24.14 ± 0.70	94.92 ± 0.05	95.95 ± 0.07	95.29 ± 0.03	67.85 ± 0.62	83.69 ± 0.18	84.75 ± 0.18	80.03 ± 0.04
REDL	17.87 ± 0.21	95.30 ± 0.01	96.63 ± 0.01	95.12 ± 0.01	79.84 ± 0.31	77.20 ± 0.10	76.82 ± 0.15	78.86 ± 0.10
HEDL	36.54 ± 0.48	83.57 ± 0.15	91.06 ± 0.10	94.71 ± 0.03	83.10 ± 0.00*	63.31 ± 0.00*	67.79 ± 0.00*	77.16 ± 0.10
FEDL	**13.69 ± 0.24**	**95.73 ± 0.10**	**97.46 ± 0.06**	**95.36 ± 0.05**	47.79 ± 0.43	86.51 ± 0.13	89.37 ± 0.11	**80.49 ± 0.05**
Ours	17.54 ± 0.90	94.89 ± 0.25	96.99 ± 0.16	95.32 ± 0.09	**36.50 ± 2.04**	**88.32 ± 0.69**	**92.24 ± 0.43**	80.10 ± 0.13

Table 3: Results of OOD detection performance for ours against other EDL-based state-of-the-art methods with CIFAR100-LT ($\eta = 0.1$) and CIFAR100-LT ($\eta = 0.01$) as ID datasets. 0.00* denotes an extremely small standard deviation (< 0.01) that rounds to 0.00 when formatted to two significant figures.

Method	CIFAR100-LT ($\eta = 0.1$)				CIFAR100-LT ($\eta = 0.01$)			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
EDL	83.01 ± 1.08	68.80 ± 0.79	75.40 ± 0.75	**73.91 ± 0.13**	92.51 ± 0.62	58.90 ± 0.60	62.45 ± 0.71	58.13 ± 0.24
$\mathcal{I}$-EDL	86.60 ± 0.34	81.10 ± 0.13	81.06 ± 0.18	70.52 ± 0.07	86.02 ± 0.30	74.38 ± 0.30	77.38 ± 0.21	54.18 ± 0.33
REDL	94.93 ± 0.35	71.76 ± 0.28	69.70 ± 0.40	66.37 ± 0.21	90.23 ± 0.69	68.65 ± 0.87	71.49 ± 1.02	50.62 ± 0.43
HEDL	81.65 ± 0.00*	65.77 ± 0.00*	73.47 ± 0.00*	34.14 ± 0.26	96.68 ± 0.00*	48.27 ± 0.00*	51.52 ± 0.00*	29.58 ± 0.13
FEDL	55.34 ± 1.89	85.61 ± 0.28	88.94 ± 0.34	73.77 ± 0.20	73.59 ± 1.14	73.62 ± 0.51	79.71 ± 0.53	**58.70 ± 0.20**
Ours	**30.72 ± 2.85**	**87.51 ± 0.67**	**93.50 ± 0.57**	70.27 ± 0.33	**48.25 ± 3.33**	**79.08 ± 1.03**	**88.36 ± 0.90**	49.88 ± 0.62

We further compare against recent EDL variants in Table 2. Although the Euclidean baseline FEDL achieves leading metrics on CIFAR10, its FPR95 increases to 47.79% on the CIFAR100 benchmark. By exploiting the representation capacity of hyperbolic space along with dynamic evidence calibration, our method reduces the FPR95 on CIFAR100 to 36.50% and improves AUROC to 92.24%, with only a minor variation in ID accuracy (80.10% vs. 80.49%). We attribute this gain to the capacity of hyperbolic geometry in decoupling complex boundaries, which offers a useful inductive bias for evidential uncertainty estimation. Additional fine-grained evaluations are reported in Table 7 of Appendix D.

Table 3 evaluates long-tailed scenarios on CIFAR100-LT under imbalance factors $\eta = 0.1$ and $\eta = 0.01$. Most EDL baselines show clear OOD performance deterioration under class imbalance, as scarce tail samples fail to accumulate sufficient evidence strength. Assisted by the magnitude correction from our evidence-alignment loss, our framework obtains superior OOD detection metrics across both imbalance ratios (FPR95 = 30.72% for $\eta = 0.1$; FPR95 = 48.25% for $\eta = 0.01$). A drop in ID accuracy is observed under severe imbalance ($\eta = 0.01$), which we attribute to the regularization behavior induced by the evidence-alignment objective: by suppressing overconfident evidence extrapolation on feature-sparse tail inputs, the formulation encourages the network to reflect higher uncertainty rather than making overconfident misclassifications, aligning with the core objective of safe uncertainty estimation. Supplementary CIFAR10-LT results are reported in Table 8 of Appendix D.

4.3 Ablation Study

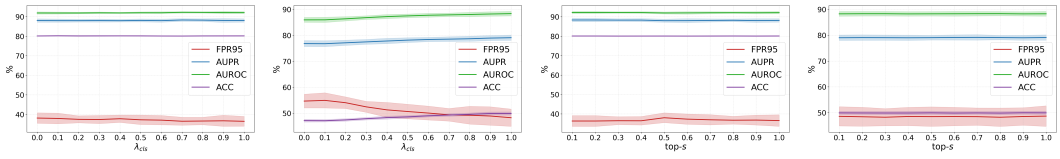
Evidence Construction from Lorentz Representations. We ablate alternative Dirichlet-evidence construction schemes from Lorentz representations in Table 4. Ours denotes our full pipeline using only spatial embedding components together with evidence-alignment loss. Strategy-1 leverages full embeddings including the time-radial component, which introduces semantic contamination from radial geometry. Strategy-2 projects features to tangent space for Euclidean EDL and discards

manifold inductive bias. Strategy-2 delivers marginally better OOD on balanced CIFAR100 (FPR95 = 38.50%), yet this advantage does not generalize to distribution-shift scenarios. On CIFAR100-LT ($\eta = 0.01$), Ours achieves FPR95 = 56.07% and substantially outperforms the other two strategies. This suggests full-embedding or tangent-projection strategies work adequately only for balanced data and lack robustness under distribution shift. Our spatial-component-only evidence construction preserves both evidence semantic validity and manifold priors for stronger cross-distribution robustness.

Table 4: Ablation on alternative Dirichlet parameters construction schemes from Lorentz representations, with CIFAR100 and CIFAR100-LT ($\eta = 0.01$) as ID datasets.

Method	CIFAR100				CIFAR100-LT ($\eta = 0.01$)			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
Ours	40.29 ± 1.14	**87.88 ± 0.52**	91.52 ± 0.18	80.13 ± 0.17	**56.07 ± 1.78**	**76.05 ± 1.04**	**84.92 ± 0.72**	**50.89 ± 0.46**
Strategy-1	51.72 ± 2.20	86.74 ± 0.49	89.45 ± 0.45	80.13 ± 0.06	88.60 ± 3.72	62.59 ± 0.90	67.36 ± 1.95	36.19 ± 0.19
Strategy-2	**38.50 ± 0.44**	87.86 ± 0.08	**91.70 ± 0.09**	**80.22 ± 0.12**	64.53 ± 2.27	69.92 ± 0.73	81.54 ± 0.89	37.79 ± 0.23

Hyperparameter Analysis for λ_{cls} and top-s. We evaluate two core hyperparameters: λ_{cls}, which controls the scale of per-class target-evidence priors ($\lambda_{cls} \in [0.0, 1.0]$, Figures 2a and 2b), and top-s, the ratio of large-deviation samples chosen for batch evidence estimation (top-$s \in [0.1, 1.0]$, Figures 2c and 2d). On CIFAR100, changing λ_{cls} brings marginal metric variation, with performance reaching its peak at $\lambda_{cls} = 0.7$ (36.50% FPR95 and 92.24% AUROC). In contrast, CIFAR100-LT ($\eta = 0.01$) shows clear sensitivity to λ_{cls}: increasing its value steadily improves performance, cutting FPR95 from 54.66% to 48.25% while raising in-distribution accuracy to 49.88%. This trend suggests that a stronger prior directly complements the evidence deficit inherent in tail classes under heavy imbalance. For top-s, performance across different deviation cutoff fractions remains smooth within a single standard deviation, indicating that our boundary sample selection fraction is non-critical to overall performance and tolerant of local noise. Following these empirical findings, we use $\lambda_{cls} = 0.7$ for balanced datasets, set $\lambda_{cls} = 1.0$ in long-tailed settings to balance prior and evidence terms, and fix top-$s = 0.3$ by default. Additional ablation studies can be found in the Appendix D.2.



(a) λ_{cls} (C100) (b) λ_{cls} (C100-LT, $\eta = 0.01$) (c) top-s (C100) (d) top-s (C100-LT, $\eta = 0.01$)

Figure 2: Ablation study of hyperparameters λ_{cls} and top-s. C100 denotes the dataset CIFAR100.

5 Conclusion

In this work, we present LEDL, a Lorentz evidential learning framework tailored for open-set recognition. To address the challenge of adapting EDL to hyperbolic networks, we clarify the geometric rationale for parameterizing subjective evidence solely from the spatial components of Lorentz embeddings, avoiding semantic contamination and evidence bias induced by radial coordinates. Leveraging this formulation, we establish a spatial-driven evidence paradigm and further introduce LPEA using temporal statistics to recalibrate class-wise evidence strength, enforcing strict consistency with the underlying manifold. Comprehensive evaluations demonstrate that LEDL enhances OOD detection efficiency while preserving in-distribution capability, showing solid robustness across complex long-tailed and distribution-shift scenarios.

Additional Notes. Discussions about limitations and future work, together with the AI-use disclosure statement, are provided in Appendix E and Appendix F, respectively. Ethics Statement and Reproducibility Statement can be found in Appendix G and Appendix H, respectively.

REFERENCES

Muhammad Asad, Ihsan Ullah, Ganesh Sistu, and Michael G. Madden. Contextual graph modeling for open-set object detection in multi-view autonomous driving. *IEEE Open Journal of Vehicular Technology*, 7:1323–1336, 2026. doi: 10.1109/OJVT.2026.3687800.

Xingyu Cai, Haiyan Yang, Jiayi Chen, Jun Shi, and Jun Wang. Adaptive multi-prototype open-set recognition network for medical image classification. *IEEE Transactions on Biomedical Engineering*, pp. 1–12, 2026. doi: 10.1109/TBME.2026.3721904.

James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31(59-115):2, 1997.

Ines Chami, Rex Ying, Christopher Re, and Jure Leskovec. Hyperbolic graph convolutional neural networks. Red Hook, NY, USA, 2019. Curran Associates Inc.

M. Chen, J. Gao, and C. Xu. R-edl: Relaxing nonessential settings of evidential deep learning. In *International Conference on Learning Representations*, volume 2024, pp. 35563–35588, 5 2024.

Weize Chen, Xu Han, Yankai Lin, Hexu Zhao, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Fully hyperbolic neural networks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5672–5686, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.389. URL https://aclanthology.org/2022.acl-long.389/.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Danruo Deng, Guangyong Chen, Yang Yu, Furui Liu, and Pheng-Ann Heng. Uncertainty estimation by fisher information-based evidential deep learning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Xiwen Dengxiong and Yu Kong. Ancestor search: Generalized open set recognition via hyperbolic side information learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3992–4001, 2023. doi: 10.1109/WACV56688.2023.00399.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. *arXiv preprint arXiv:2209.09858*, 2022.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022.

Lei Fan, Bo Liu, Haoxiang Li, Ying Wu, and Gang Hua. Flexible visual recognition by evidential modeling of confusion and ignorance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1338–1347, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. NIPS'18, pp. 5350–5360, Red Hook, NY, USA, 2018. Curran Associates Inc.

Chen Gao, Baining Zhao, Weichen Zhang, Jinzhu Mao, Jun Zhang, Zhiheng Zheng, Fanhang Man, Jianjie Fang, Zile Zhou, Jinqiang Cui, Xinlei Chen, and Yong Li. Embodiedcity: A benchmark platform for embodied agent in real-world city environment. 2024. URL https://arxiv.org/abs/2410.09604.

Yiwei Gong, Anna L. Smith, Dena Asta, and Catherine A. Calder. Hyperbolic latent space models for network embedding: Model specification and bayesian inference. 2026. URL https://arxiv.org/abs/2605.11340.

Zhuolin He, Xinrun Li, Jiacheng Tang, Shoumeng Qiu, Wenfu Wang, Xiangyang Xue, and Jian Pu. Toward camera open-set 3d object detection for autonomous driving scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 26(12):23190–23201, 2025. doi: 10.1109/TITS.2025.3618307.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 2018. URL https://arxiv.org/abs/1610.02136.

Jin Ho Yang, Woo Young Choi, and Chung Choo Chung. Physics-informed neural network-based open set classification of neighboring vehicle motion for decision-making in autonomous driving. *IEEE Access*, 13:168561–168579, 2025. doi: 10.1109/ACCESS.2025.3613724.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10948–10957, 2020. doi: 10.1109/CVPR42600.2020.01096.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8706–8715. IEEE, jun 2021.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Audun Jøsang. A logic for uncertain probabilities. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 9(3):279–311, June 2001. ISSN 0218-4885. doi: 10.1142/S0218488501000831. URL https://doi.org/10.1142/S0218488501000831.

Audun Jøsang. *Subjective Logic*, volume 3. Springer, 2016.

Tejaswi Kasarla, Max van Spengler, and Pascal Mettes. Balanced hyperbolic embeddings are natural out-of-distribution detectors. *International Journal of Computer Vision*, 134(5):202, 4 2026. ISSN 1573-1405. doi: 10.1007/s11263-026-02775-6. URL https://doi.org/10.1007/s11263-026-02775-6.

Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3233–3243, 2025. doi: 10.1109/TPAMI.2022.3184052.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Changbin Li, Kangshuo Li, Yuzhe Ou, Lance Kaplan, Audun Jøsang, Jin-Hee Cho, Dong Hyun Jeong, and Feng Chen. Hyper evidential deep learning to quantify composite classification uncertainty. In *International Conference on Learning Representations*, volume 2024, pp. 8826–8860, 2024.

Xiwen Li, Wei Quan, Jia Huang, and Peiyuan Hong. Recognizing unknowns: a survey on visual open-set recognition. *Applied Intelligence*, 55(16):1055, 2025. ISSN 1573-7497. doi: 10.1007/s10489-025-06956-7. URL https://doi.org/10.1007/s10489-025-06956-7.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. 2020. URL https://arxiv.org/abs/1706.02690.

Jing Liu, Yang Liu, Jieyu Lin, Jielin Li, Liang Cao, Peng Sun, Bo Hu, Liang Song, Azzedine Boukerche, and Victor C.M. Leung. Networking systems for video anomaly detection: A tutorial and survey. 57(10), May 2025. ISSN 0360-0300. doi: 10.1145/3729222. URL https://doi.org/10.1145/3729222.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23946–23955, 2023. doi: 10.1109/CVPR52729.2023.02293.

Yao Lu, Rohit Mohan, Florian Drews, Yakov Miron, and Abhinav Valada. Hyp2former: Hierarchy-aware hyperbolic embeddings for open-set panoptic segmentation. 2026. URL https://arxiv.org/abs/2605.02580.

Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132(9):3484–3508, 9 2024. ISSN 1573-1405. doi: 10.1007/s11263-024-02043-5. URL https://doi.org/10.1007/s11263-024-02043-5.

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, pp. 4. Granada, 2011.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47.

Deep Pandey and Qi Yu. Learn to accumulate evidence from all training samples: theory and practice. ICML'23. JMLR.org, 2023.

Kunyu Peng, Di Wen, M. Saquib Sarfraz, Yufan Chen, Junwei Zheng, David Schneider, Kailun Yang, Jiamin Wu, Alina Roitberg, and Rainer Stiefelhagen. Mitigating label noise using prompt-based hyperbolic meta-learning in open-set domain generalization. *International Journal of Computer Vision*, 134(3):99, 2 2026. ISSN 1573-1405. doi: 10.1007/s11263-025-02643-9. URL https://doi.org/10.1007/s11263-025-02643-9.

Jingen Qu, Yufei Chen, Xiaodong Yue, Wei Fu, and Qiguang Huang. Hyper-opinion evidential deep learning for out-of-distribution detection. NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Lei Ren, Jiabao Dong, Shuai Liu, Lin Zhang, and Lihui Wang. Embodied intelligence toward future smart manufacturing in the era of ai foundation model. *IEEE/ASME Transactions on Mechatronics*, 30(4):2632–2642, 2025. doi: 10.1109/TMECH.2024.3456250.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. NIPS'18, pp. 3183–3193, Red Hook, NY, USA, 2018. Curran Associates Inc.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

Xianglong Shi, Ziheng Chen, Yunhan Jiang, and Nicu Sebe. Intrinsic lorentz neural network. *arXiv preprint arXiv:2602.23981*, 2026.

Youssef Shoeb, Azarm Nowzad, and Hanno Gottschalk. Out-of-distribution segmentation in autonomous driving: Problems and state of the art. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4310–4320, 2025. doi: 10.1109/CVPRW67362.2025.00416.

Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:17885–17898, 2022.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 691–708, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_40. URL https://doi.org/10.1007/978-3-031-20053-3_40.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20827–20840. PMLR, 2022.

Jongheon Tack, Sungjun Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11839–11852, 2020.

Vishal, Arnav Aditya, Nitin Kumar, and Saurabh J. Shigwan. Dmdsc: A dynamic-margin deep simplex classifier for open-set recognition on medical image datasets. 2026. URL https://arxiv.org/abs/2605.00675.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. 2011.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23631–23644. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wei22d.html.

Fangdi Xu and Yunxue Shao. Multistage latent residual based anomaly detection. *Signal, Image and Video Processing*, 19(15):1303, 2025. ISSN 1863-1711. doi: 10.1007/s11760-025-04872-4. URL https://doi.org/10.1007/s11760-025-04872-4.

Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by density aware evidential deep learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Taeseong Yoon and Heeyoung Kim. Uncertainty estimation by flexible evidential deep learning. *Advances in Neural Information Processing Systems*, 38:118601–118641, 2026.

Yazhou Yuan, Ning Yu, Zhuolin Zheng, Yong Yang, Kai Ma, Zhixin Liu, Cailian Chen, and Jianmin Zhang. Enhancing network abnormal detection with nmf-secnn: Leveraging deep feature learning for high-precision traffic analysis. *IEEE Transactions on Network Science and Engineering*, 12 (3):2069–2080, 2025. doi: 10.1109/TNSE.2025.3544251.

Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, sep 2023.

Xinyu Zhao, Zhenquan Wu, Xuemei Zhu, Mengrui Zhang, Kaixuan Cui, Shaobin Chen, Sihui Zhang, Yarou Hu, Shirou Wu, Ruyin Tian, Yi Chen, Xiaorong Cheng, Na Duan, Ling Jiang, Na Li, Wenbin Wei, Wei Chi, Guishuang Ying, Baiying Lei, Jianhong Liang, and Guoming Zhang. Uncertainty-inspired open-set model for identifying infantile fundus abnormalities. *npj Digital Medicine*, 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-03118-y. URL https://doi.org/10.1038/s41746-026-03118-y.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

A MATHEMATICAL NOTATION

We summarise the key mathematical notations used throughout this paper in the table below for quick reference.

Table 5: Notation Definition

Notation	Definition
$\boldsymbol{x}_i$	the i-th sample
C	total number of classes
$\mathbb{L}$	Hyperboloid space
$\langle \cdot, \cdot \rangle_L$	Lorentz inner product
f^E	the Euclidean backbone
$\boldsymbol{z}_i^E$	Euclidean feature embedding of $\boldsymbol{x}_i$
f^L	Lorentz transformation layer
g	evidence network
$\tilde{\boldsymbol{z}}_i = [\tilde{z}_{i,0}, \tilde{\boldsymbol{z}}_{i,s}]$	Lorentz embedding of $\boldsymbol{x}_i$; $\tilde{z}_{i,0}$ is time component, and $\tilde{\boldsymbol{z}}_{i,s}$ is space component
$\boldsymbol{e}_i$	evidence vector of $\boldsymbol{x}_i$
$\boldsymbol{\alpha}_i$	Dirichlet parameters of $\boldsymbol{x}_i$
$\boldsymbol{p}_i$	class probality of $\boldsymbol{x}_i$
u_i	uncertainty of $\boldsymbol{x}_i$
$\boldsymbol{\mu}_c$	class-wise Lorentz prototype for class c
s_{ic}	prototype-aware similarity between $\boldsymbol{x}_i$ and $\boldsymbol{\mu}_c$
t_c	dynamic target evidence strength for class c
λ_{cls}	hyperparameter for dynamic target initialization offset
m_p	EMA momentum for Lorentz prototype update
m_t	EMA momentum for dynamic evidence target update
$\mathcal{L}^{ece}$	loss for classification
$\mathcal{L}^{ali}$	loss for Lorentz Prototype-guided Evidence Alignment
w^{ali}	loss weight for $\mathcal{L}^{ali}$

B PROOF

B.1 JUSTIFICATION OF PRINCIPLE 1

From the Lorentz hyperboloid constraint for sample i:

$$\langle \tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{z}}_i \rangle_L = -k \quad \Rightarrow \quad z_{i,0} = \sqrt{k + \|\tilde{\boldsymbol{z}}_{i,s}\|_2^2}. \tag{13}$$

Hence, time coordinate $z_{i,0}$ is completely determined by manifold curvature k and the Euclidean norm of space component $\tilde{\boldsymbol{z}}_{i,s}$, representing purely radial depth on the hyperboloid. Were Dirichlet parameters derived from the full embedding via $\boldsymbol{\alpha}_i = h(z_{i,0}, \tilde{\boldsymbol{z}}_{i,s})$, taking the partial derivative with respect to time coordinate yields:

$$\frac{\partial \boldsymbol{\alpha}_i}{\partial z_{i,0}} = \frac{\partial h}{\partial z_{i,0}} \neq \mathbf{0}, \tag{14}$$

demonstrating functional dependence on radial depth offsets. This violates Assumption 1, as class-agnostic geometric depth contaminates belief assignment. Enforcing the alignment constraint $\frac{\partial h}{\partial z_{i,0}} = \mathbf{0}$ forces $h(\cdot)$ to depend strictly on space components $\tilde{\boldsymbol{z}}_{i,s}$, satisfying Principle 1.

B.2 LIMITATIONS OF DISTANCE-BASED LORENTZ LOGITS FOR EDL

Note: This appendix analyzes the PLFC classifier from Shi et al. (2026). The original work uses $K < 0$ for sectional curvature; we adopt hyperboloid scale parameter $k > 0$ corresponding to sectional curvature $K_{\text{sec}} = -\frac{1}{k}$. All equations are adapted to our notation.

This subsection analyzes two cases: (1) standard sample-prototype negative Lorentz distances, and (2) intrinsic point-to-hyperplane PLFC logits. Even with advanced manifold-native classification heads, directly employing their outputs for Dirichlet parameterization exhibits fundamental flaws.

Case 1: Sample-Prototype Negative Distance Logit. The Lorentz geodesic distance between two hyperboloid points reads:

$$d_L(\boldsymbol{z}, \boldsymbol{\mu}_c) = \frac{1}{\sqrt{k}} \cosh^{-1}\left(\frac{\langle \boldsymbol{z}, \boldsymbol{\mu}_c \rangle_L}{k}\right), \quad d_L \geq 0. \tag{15}$$

where $\boldsymbol{\mu}_c$ is the prototype of class c. Conventional hyperbolic classification uses negative geodesic distance as classification logit: $s_c^{dist} = -d_L(\boldsymbol{z}, \boldsymbol{\mu}_c) \in (-\infty, 0]$. Applying naive mapping $\alpha_c = \exp(s_c^{dist}) + 1$ constrains evidence since $\exp(s_c^{\text{dist}}) \in (0, 1]$, yielding $\alpha_c \in (1, 2]$. Two critical issues emerge:

1. Concentration parameters remain bounded within a narrow range, crippling the model's ability to express high-confidence belief.
2. Distance evaluation explicitly couples time coordinate z_0, leaking radial geometric information into logits and triggering the semantic contamination warned in Principle 1.

Case 2: Intrinsic Point-to-Hyperplane PLFC Logit Shi et al. (2026). The point-to-hyperplane fully-connected layer (PLFC) provides a fully intrinsic Lorentz classifier whose output $logits_c$ represents signed geodesic distance from sample $\boldsymbol{z}$ to a learnable hyperboloid hyperplane:

$$\begin{aligned} logits_c(\boldsymbol{z}) &= \sqrt{k}\,\mathrm{sign}(\rho_c)\left|\sinh^{-1}\left(\frac{1}{\sqrt{k}}\frac{\rho_c}{\theta_c}\right)\right| \\ \rho_c &= \cosh\left(\frac{a_c}{\sqrt{k}}\right)\langle \hat{\boldsymbol{z}}, \boldsymbol{z}_s \rangle_{\mathbb{R}} - \sinh\left(\frac{a_c}{\sqrt{k}}\right)\|\hat{\boldsymbol{z}}\|_2\, z_0 \\ \theta_c &= \sqrt{\left\|\cosh\left(\frac{a_c}{\sqrt{k}}\right)\hat{\boldsymbol{z}}\right\|_2^2 - \left(\sinh\left(\frac{a_c}{\sqrt{k}}\right)\|\hat{\boldsymbol{z}}\|_2\right)^2} \end{aligned} \tag{16}$$

where $\hat{\boldsymbol{z}}$ defines the orientation of the learned Lorentz hyperplane, and scalar a represents its anchor's geodesic offset from the Lorentz origin. Inspecting ρ_c reveals that even for this fully intrinsic layer, the formulation explicitly relies on sample time coordinate z_0. Because z_0 encodes radial position decoupled from class semantics, it directly propagates into classification logits. Directly setting $\alpha_c = \exp(logits_c) + 1$ introduces clear drawbacks:

1. Time component z_0 leaks into logits and pollutes Dirichlet parameters after exponential mapping, violating Assumption 1.
2. PLFC logits prioritize classification margin over evidence magnitude, offering no built-in mechanism to guarantee well-calibrated evidential strength.

In summary, neither distance-based logits nor intrinsic point-to-hyperplane scores can be directly recycled for Dirichlet parameterization. This reinforces our primary design: routing spatial embedding components $\boldsymbol{z}_s$ into a dedicated evidence sub-network rather than reusing off-the-shelf classification outputs.

B.3 Limitations of Tangent-Space Projection for EDL

The second naive strategy projects Lorentz embeddings onto the origin tangent space to obtain Euclidean vectors before applying standard Euclidean EDL. This projection is mathematically well-posed and algebraicaly correct; nonethless, it distorts geometric structures and inductive biases learned on the hyperboloid, causing degraded empirical performance.

Let $\mathbb{L}_k^n$ denote the Lorentz hyperboloid with scale parameter $k > 0$. The manifold origin reads $\mathbf{0} = \left(\sqrt{k}, \mathbf{0}^\top\right)$. $\mathrm{Exp}_0(\cdot)$ and $\mathrm{Log}_0(\cdot)$ represent Riemannian exponential and logarithmic maps evaluated at origin $\mathbf{0}$.

For any hyperboloid point $\boldsymbol{z} = [z_0, \boldsymbol{z}_s]$, the logarithmic map projects it onto origin tangent space, yielding tangent vector $\boldsymbol{z}^{\tan} = \mathrm{Log}_0(\boldsymbol{z})$. The closed-form logarithmic map at the origin reads:

$$\mathrm{Log}_0(\boldsymbol{z}) = \frac{\cosh^{-1}\left(-\frac{1}{k}\langle \mathbf{0}, \boldsymbol{z} \rangle_L\right)}{\sqrt{k - \frac{1}{k}\langle \mathbf{0}, \boldsymbol{z} \rangle_L^2}}\left(\boldsymbol{z} - \frac{1}{k}\langle \mathbf{0}, \boldsymbol{z} \rangle_L \cdot \mathbf{0}\right) \tag{17}$$

where the Lorentz inner product expands as $\langle \mathbf{0}, \boldsymbol{z} \rangle_L = -0_0 z_0 + \langle \mathbf{0}_s, \boldsymbol{z}_s \rangle_{\mathbb{R}}$. Substituting the spatial origin $\mathbf{0}_s = \mathbf{0}$ simplifies this inner product:

$$\langle \mathbf{0}, \boldsymbol{z} \rangle_L = -\sqrt{k}\, z_0 \tag{18}$$

Substituting back into the logarithmic map, the tangent output $\boldsymbol{z}^{\tan}$ retains only space components (as the timelike tangent coordinate vanishes at the origin). The resulting Euclidean vector $\boldsymbol{z}_s^{\tan}$ is then processed by a Euclidean network to compute Dirichlet parameters. The root limitation stems from $\mathrm{Log}_0(\cdot)$ being a non-linear Riemannian mapping. Lorentz geodesic distances $d_L(\boldsymbol{z}_1, \boldsymbol{z}_2)$ transform into Euclidean distances $\|\mathrm{Log}_0(\boldsymbol{z}_1) - \mathrm{Log}_0(\boldsymbol{z}_2)\|_2$ post-projection, lacking linear proportionality.

Points far from the origin undergo severe distance stretching post-projection, distorting intrinsic hierarchical neighborhoods and weakening hyperboloid inductive biases. While a Euclidean network could theoretically approximate this warped space, doing so consumes excess capacity to offset geometric distortion—an overhead rarely compensated under finite data and compute, ultimately degrading empirical performance.

In short, while algebraicaly sound, tangent projection disrupts native hyperboloid inductive biases and introduces geometric distortion with fitting overhead, making it unsuitable as a primary solution.

B.4 Discussion on EMA-updated Prototypes: Approximation to Fréchet Mean

The prototype update scheme in the main text serves as a computationally scalable proxy for the exact Riemannian Fréchet mean on $\mathbb{L}_k^n$. Below we review the formal Fréchet mean, examine the sources of approximation error, and discuss why this proxy is valid in practice.

Given hyperboloid embeddings $\boldsymbol{z}_i$ for class c, the true Fréchet mean $\boldsymbol{\mu}_c^*$ minimizes the sum of squared geodesic distances:

$$\boldsymbol{\mu}_c^* = \arg\min_{\boldsymbol{\mu} \in \mathbb{L}_k^n} \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} d_{\mathbb{L}}^2(\boldsymbol{z}_i, \boldsymbol{\mu}). \tag{19}$$

Computing $\boldsymbol{\mu}_c^*$ exactly requires inner-loop Riemannian optimization at every training step, adding an $\mathcal{O}(|\mathcal{B}_c| \cdot T)$ computational overhead (where T is the number of inner steps) that quickly becomes intractable during online batch training.

To keep training fast and scalable, our surrogate procedure operates directly in the ambient Minkowski space:

1. Compute mini-batch Euclidean centroids over spatial-temporal ambient coordinates for class c, projecting the resulting mean onto the hyperboloid to form $\boldsymbol{\mu}_{\text{batch}}$;
2. Apply momentum aggregation in ambient coordinates before re-projecting: $\boldsymbol{\mu}_{new} \leftarrow \mathrm{project}\big(m_p\, \boldsymbol{\mu}^{old} + (1 - m_p)\, \boldsymbol{\mu}^{\mathcal{B}}\big)$.

This surrogate avoids costly Riemannian optimization loops entirely. The approximation error stems from two primary sources:

1. Ambient Minkowski centroids do not coincide with non-Euclidean Riemannian Fréchet means on $\mathbb{L}_k^n$;
2. Vector addition within EMA aggregation is performed linearly in ambient space rather than via hyperbolic gyrogroup addition.

This approximation holds well under the condition outlined in the main text: class-specific embeddings do not scatter too widely along radial directions. Within compact local neighborhoods on the manifold, the ambient centroid acts as a reliable first-order estimate of the true Riemannian mean, bypassing the iterative bottleneck without noticeably degrading representation quality.

B.5 Mathematical Justification of Lorentz Inner-Product Deviation Metric

This section details the geometric interpretation and validity of the prototype-aware deviation metric s_{ic}. On the Lorentz hyperboloid, the inner product between any sample embedding $\boldsymbol{z}_i$ and prototype

$\boldsymbol{\mu}_c$ obeys the Minkowski inequality $\langle \boldsymbol{z}_i, \boldsymbol{\mu}_c \rangle_L \leq -k$, with equality holding if and only if $\boldsymbol{z}_i = \boldsymbol{\mu}_c$. The exact geodesic distance is $d_{\mathbb{L}}(\boldsymbol{z}_i, \boldsymbol{\mu}_c) = \sqrt{k}\text{arccosh}\left(-\frac{\langle \boldsymbol{z}_i, \boldsymbol{\mu}_c \rangle_L}{k}\right)$. During training, as intra-class embeddings cluster within localized regions around their respective prototypes, a Taylor expansion of the inverse hyperbolic cosine around $\langle \boldsymbol{z}_i, \boldsymbol{\mu}_c \rangle_{\mathbb{L}} = -k$ demonstrates that the linear scaled form $-\frac{\langle \boldsymbol{z}_i, \boldsymbol{\mu}_c \rangle_L + k}{2k}$ acts as a valid first-order proxy for local manifold divergence. Driven by local compactness on the hyperbolic manifold, inner products between intra-class samples and their class prototypes reside predominantly within a compact neighborhood, ensuring well-behaved numerical bounds and stable gradient flow during evidential optimization.

B.6 Algorithm Details

Algorithm 1 details the step-by-step execution for both training and inference stages of the LEDL framework.

C Experimental Setting

C.1 Dataset Details

We adopt four in-distribution datasets (ID) and three out-of-distribution (OOD) datasets for our experiments. The ID datasets are CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), Flowers102 (Nilsback & Zisserman, 2008), and CUB200 (Wah et al., 2011). The OOD datasets are SVHN (Netzer et al., 2011), Describable Textures Dataset (DTD) (Cimpoi et al., 2014), and Places365 (Zhou et al., 2018).

CIFAR10 consists of 60000 color images with a spatial resolution of 32×32, which are partitioned into 10 mutually-exclusive semantic classes without class overlap. Under its standard partition, 50000 images are assigned to the training split for model parameter optimization, while the remaining 10000 images serve as the test split to evaluate both in-distribution classification and out-of-distribution detection performance.

CIFAR100 comprises 60000 32×32 resolution color images spanning 100 fine-grained classes, which are further grouped into 20 high-level super-classes. Following its official split, 50000 samples form the training set and 10000 samples form the test set. Compared with CIFAR10, it owns a larger number of target classes and raises higher classification difficulty, making it a widely-adopted benchmark for OOD detection.

Flowers102 is a fine-grained flower classification dataset containing 102 distinct flower species with 8189 images in total. Official pre-defined training, validation and test splits are provided. The sample quantity varies across different flower classes, which makes it suitable for examining the model's uncertainty estimation capability under fine-grained recognition scenarios.

CUB200 is a well-known fine-grained bird recognition dataset, covering 200 bird species with 11788 images in total. Official train-test partitions are released. It exhibits large intra-class variance and subtle inter-class discrepancy, posing substantial challenges for feature learning and uncertainty modelling.

SVHN is a real-world street-view house number dataset collected from outdoor street scenes. In our experiments, only its test subset is employed as out-of-distribution test samples. SVHN is never involved in model training to prevent extra class information from contaminating OOD evaluation.

DTD consists of diverse texture-focused images, which contain no explicit foreground objects and only preserve texture and material information. It acts as a classic benchmark in OOD detection research, and is frequently used to verify model uncertainty outputs when confronting completely unseen semantic patterns.

Places365 is a large-scale scene recognition dataset with massive real-world images covering 365 distinct scene classes. We use its test parts as OOD source. All input images are resized to guarantee consistent input resolution with in-distribution datasets.

Algorithm 1 Training and Inference of LEDL (Lorentz Evidential Deep Learning)

Require: Train Dataset $\mathcal{D}$, class number C, curvature scale $k > 0$; hyperparameters m_p, m_t, top-s ratio, weight w^{ali}, offset scale λ_{cls}; Euclidean backbone f^E, Lorentz transform f^L, evidence network g.

/* Phase 1: Backbone Pre-training */
Train f^E on $\mathcal{D}$ via standard cross-entropy loss; freeze all weights of f^E.
/* Phase 2: Initialization */
Initialize class prototypes $\{\boldsymbol{\mu}_c\}_{c=1}^{C}$ on hyperboloid;
Set initial target statistics: $t_c \leftarrow 1.0 + \lambda_{cls}\, C,\ \forall c \in \{1, \ldots, C\}$.
/* Phase 2: Lorentz Evidential Learning */
while not converged **do**
  Fetch mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B}$ from $\mathcal{D}$
  Forward Pass:
  $\boldsymbol{z}_i^E \leftarrow f^E(x_i)$
  $\boldsymbol{z}_i \leftarrow \exp_{\boldsymbol{o}}^k\left((0, \boldsymbol{z}_i^E)\right)$ ▷ Exponential map at origin $\boldsymbol{o} = (\sqrt{k}, \mathbf{0})$
  $\tilde{\boldsymbol{z}}_i \leftarrow f^L(\boldsymbol{z}_i)$
  $\tilde{\boldsymbol{z}}_{i,s} \leftarrow$ Slice spatial dimensions of $\tilde{\boldsymbol{z}}_i$ ▷ Discard temporal component
  $\boldsymbol{\alpha}_i \leftarrow \exp\left(g(\tilde{\boldsymbol{z}}_{i,s})\right) + 1$ ▷ Formulate Dirichlet parameters
  Evaluate classification loss $\mathcal{L}^{ece}$ using $\{\boldsymbol{\alpha}_i, y_i\}$
  Update prototypes $\{\boldsymbol{\mu}_c\}$ via ambient EMA aggregation
  for each class $c \in \{1, \ldots, C\}$ present in mini-batch **do**
    Compute s_{ic} for $y_i = c$; extract top-s subset $\mathcal{I}_c^{top}$
    if $|\mathcal{I}_c^{top}| \geq 2$ **then**
      Compute $t_c^{\mathcal{B}} \leftarrow \frac{1}{|\mathcal{I}_c^{top}|} \sum_{i \in \mathcal{I}_c^{top}} \alpha_{ic}$
      Update $t_c \leftarrow m_t\, t_c + (1 - m_t)\, t_c^{\mathcal{B}}$
    end if
  end for
  Evaluate alignment loss $\mathcal{L}^{ali}$
  Set total loss $\mathcal{L}^{total} \leftarrow \mathcal{L}^{ece} + w^{ali}\, \mathcal{L}^{ali}$
  Compute $\nabla \mathcal{L}^{total}$ to update parameters of f^L, g
end while

/* Inference Stage: Prototypes and t_c remain fixed */
function INFERENCE($\boldsymbol{x}$)
  $\boldsymbol{z}^E \leftarrow f^E(x)$
  $\boldsymbol{z} \leftarrow \exp_{\boldsymbol{o}}^k\left((0, \boldsymbol{z}^E)\right)$
  $\tilde{\boldsymbol{z}} \leftarrow f^L(\boldsymbol{z})$
  $\tilde{\boldsymbol{z}}_s \leftarrow$ Extract space components from $\tilde{\boldsymbol{z}}$
  $\boldsymbol{\alpha} \leftarrow \exp\left(g(\tilde{\boldsymbol{z}}_s)\right) + 1$
  Calculate total evidence $S = \sum_{c=1}^{C} \alpha_c$
  Derive class probabilities $\boldsymbol{p} = \boldsymbol{\alpha}/S$ and uncertainty $u = C/S$
  return predictive probabilities $\boldsymbol{p}$, uncertainty u
end function

C.2 Evaluation Metrics

This appendix defines all evaluation metrics adopted for open-set and out-of-distribution detection tasks, including three OOD detection metrics (FPR95, AUPR, AUROC) and the closed-set classification accuracy Acc.

Let S_{in} and S_{out} denote model-derived detection score sets for in-distribution and out-of-distribution samples, respectively. Ground-truth labels are set as $y = 1$ for in-distribution samples and $y = 0$ for out-of-distribution samples. Given a score threshold τ, we define true positives TP(τ), false

positives $\mathrm{FP}(\tau)$, true negatives $\mathrm{TN}(\tau)$, and false negatives $\mathrm{FN}(\tau)$.

$$\begin{aligned} \mathrm{TP}(\tau) &= \{s \in S_{\text{in}} \mid s \geq \tau\} \\ \mathrm{FN}(\tau) &= \{s \in S_{\text{in}} \mid s < \tau\} \\ \mathrm{FP}(\tau) &= \{s \in S_{\text{out}} \mid s \geq \tau\} \\ \mathrm{TN}(\tau) &= \{s \in S_{\text{out}} \mid s < \tau\} \end{aligned} \tag{20}$$

True-positive rate (TPR), false-positive rate (FPR), precision, and recall are defined as:

$$\begin{aligned} \mathrm{TPR}(\tau) &= \frac{|\mathrm{TP}(\tau)|}{|\mathrm{TP}(\tau)| + |\mathrm{FN}(\tau)|}, \\ \mathrm{FPR}(\tau) &= \frac{|\mathrm{FP}(\tau)|}{|\mathrm{FP}(\tau)| + |\mathrm{TN}(\tau)|}, \\ \mathrm{Precision}(\tau) &= \frac{|\mathrm{TP}(\tau)|}{|\mathrm{TP}(\tau)| + |\mathrm{FP}(\tau)|}, \\ \mathrm{Recall}(\tau) &= \frac{|\mathrm{TP}(\tau)|}{|\mathrm{TP}(\tau)| + |\mathrm{FN}(\tau)|}. \end{aligned} \tag{21}$$

FPR95 denotes the false-positive rate evaluated at threshold τ_{95} satisfying $\mathrm{TPR}(\tau_{95}) = 0.95$:

$$\mathrm{FPR95} = \mathrm{FPR}(\tau_{95}). \tag{22}$$

This metric quantifies out-of-distribution separation capability under the constraint of maintaining 95% true-positive recall for in-distribution data; lower values indicate stronger detection performance.

AUPR corresponds to the area under the precision-recall (PR) curve constructed by sweeping over feasible decision-threshold pairs of recall and precision. As noted above, in-distribution samples are designated as the positive class for computing AUPR. The metric takes values within $[0, 1]$, where higher values reflect improved out-of-distribution detection performance.

AUROC is the area under the receiver-operating-characteristic (ROC) curve, which characterizes the trade-off between true-positive and false-positive rates across decision thresholds. Statistically, AUROC equals the probability that a randomly drawn in-distribution sample receives a higher detection score than a randomly sampled out-of-distribution instance. It ranges over $[0, 1]$, with larger values indicating superior detection capability.

Acc is closed-set Accuracy, which measures multi-class prediction performance solely on in-distribution known-class samples. It is defined as the fraction of correctly predicted samples among all in-distribution test instances. It lies in $[0, 1]$, and higher values denote better closed-set classification performance.

C.3 Implementation Details

This appendix provides full implementation details, including backbone configurations, dataset splits, training configurations and dataset-wise hyperparameters.

For CIFAR10, CIFAR100 and their long-tailed variants CIFAR10-LT, CIFAR100-LT, ResNet-18 is adopted as the backbone. For fine-grained benchmarks Flowers102 and CUB200, ResNet-34 is used to acquire stronger feature representations. All experiments are implemented in PyTorch and run on a single NVIDIA GeForce RTX 3090 GPU. Official standard data splits are utilized for all datasets.

General training settings: the full training schedule contains 100 epochs. For the first 90 epochs, the backbone is pre-trained with softmax using the AdamW optimizer. In the subsequent 10 epochs, the backbone is frozen and our Lorentz evidential module is fine-tuned with RiemannianAdam. Both training stages adopt the same initial learning rate 0.0001 and batch size 128. All input images are resized to 224×224 at test time.

Hyperparameters are grouped by dataset families. For CIFAR10, CIFAR100, Flowers102 and CUB200: $\lambda_{cls} = 0.7$, $s = 0.3$, momentum terms $m_t = 0.95$, $m_p = 0.95$. For long-tailed datasets CIFAR10-LT ($\eta = 0.1, 0.01$) and CIFAR100-LT ($\eta = 0.1, 0.01$): $\lambda_{cls} = 1.0$, $s = 0.3$, $m_t = 0.95$, $m_p = 0.95$. The loss weight for loss $\mathcal{L}^{ali}$ is fixed to $w^{ali} = 0.1$ across all experiments.

D Additional Experimental Results

D.1 Additional state-of-the-art Experimental Results

Table 6: Results of OOD detection performance for ours against other state-of-the-art methods with Flowers102 and CUB200 as ID datasets. Except for Ours, the remaining compared methods are reported in the HEDL (Qu et al., 2024).

Method	Flowers102				CUB200			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
MSP (Hendrycks & Gimpel, 2018)	14.86	95.94	97.42	83.75	30.29	91.18	94.35	75.82
ODIN (Liang et al., 2020)	**4.36**	**97.63**	98.22	83.75	21.92	89.92	96.22	75.82
VIM (Wang et al., 2022)	6.34	96.70	97.94	83.75	6.71	97.27	98.26	75.82
GradNorm (Huang et al., 2021)	5.38	97.11	98.81	83.75	32.08	97.68	95.22	75.82
KNN (Sun et al., 2022)	18.45	88.83	95.30	83.75	14.35	88.63	97.40	75.82
DICE (Sun & Li, 2022)	4.64	97.62	98.95	83.75	25.82	88.83	96.00	75.82
RankFeat (Song et al., 2022)	96.57	76.62	60.98	83.75	74.68	83.38	71.09	75.82
ASH (Djurisic et al., 2022)	5.16	97.54	98.84	83.75	15.82	92.75	97.07	75.82
SHE (Zhang et al., 2023)	11.69	93.96	97.79	83.75	22.94	96.14	96.18	75.82
GEN (Liu et al., 2023)	5.25	97.55	**98.85**	83.75	15.88	92.74	97.06	75.82
MCDropout (Gal & Ghahramani, 2016)	14.77	96.22	97.41	83.98	42.46	87.08	91.76	75.83
G-ODIN (Hsu et al., 2020)	56.92	69.88	82.12	24.30	29.51	85.13	93.85	66.74
VOS (Du et al., 2022)	39.17	84.52	90.11	78.08	35.98	83.93	89.86	75.92
LogitNorm (Wei et al., 2022)	41.07	80.34	85.65	77.41	22.69	91.69	95.99	74.84
RED (Pandey & Yu, 2023)	95.87	80.10	76.45	84.63	36.01	94.58	94.89	**76.30**
Ours	12.66	95.99	97.81	**86.75**	**3.70**	**98.34**	**99.18**	74.18
± Std	10.23	1.96	1.58	1.18	0.82	0.28	0.13	0.10

Table 7: Results of OOD detection performance for ours against other EDL-based state-of-the-art methods with Flowers102 and CUB200 as ID datasets. 0.00* denotes an extremely small standard deviation (< 0.01) that rounds to 0.00 when formatted to two significant figures.

Method	Flowers102				CUB200			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
EDL	24.56 ± 0.27	87.38 ± 0.22	93.72 ± 0.08	89.14 ± 0.05	56.44 ± 0.52	62.09 ± 0.69	74.94 ± 0.77	$\mathbf{74.88 \pm 0.11}$
$\mathcal{I}$-EDL	97.97 ± 0.02	28.91 ± 0.01	25.84 ± 0.03	$\mathbf{89.34 \pm 0.02}$	78.89 ± 0.17	81.02 ± 0.05	83.89 ± 0.04	74.66 ± 0.05
REDL	98.31 ± 0.02	27.57 ± 0.01	21.37 ± 0.04	89.28 ± 0.02	79.27 ± 0.29	74.60 ± 0.06	80.96 ± 0.05	74.65 ± 0.03
HEDL	$\mathbf{5.98 \pm 0.00}$*	96.19 ± 0.00*	$\mathbf{98.60 \pm 0.00}$*	88.55 ± 0.05	62.08 ± 0.00*	58.54 ± 0.00*	79.68 ± 0.00*	74.07 ± 0.05
FEDL	8.84 ± 0.19	$\mathbf{96.91 \pm 0.02}$	97.49 ± 0.04	89.20 ± 0.07	18.33 ± 1.29	95.40 ± 0.16	96.12 ± 0.11	74.74 ± 0.07
Ours	12.66 ± 10.23	95.99 ± 1.96	97.81 ± 1.58	86.75 ± 1.18	$\mathbf{3.70 \pm 0.82}$	$\mathbf{98.34 \pm 0.28}$	$\mathbf{99.18 \pm 0.13}$	74.18 ± 0.10

Table 8: Results of OOD detection performance for ours against other EDL-based state-of-the-art methods with CIFAR10-LT ($\eta = 0.1$) and CIFAR10-LT ($\eta = 0.01$) as ID datasets. 0.00* denotes an extremely small standard deviation (< 0.01) that rounds to 0.00 when formatted to two significant figures.

Method	CIFAR10-LT ($\eta = 0.1$)				CIFAR10-LT ($\eta = 0.01$)			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
EDL	24.75 ± 1.28	88.74 ± 0.61	94.25 ± 0.33	92.95 ± 0.03	$\mathbf{33.41 \pm 0.75}$	84.11 ± 0.43	91.48 ± 0.21	87.29 ± 0.18
$\mathcal{I}$-EDL	38.97 ± 0.64	90.94 ± 0.16	93.79 ± 0.12	92.08 ± 0.08	56.42 ± 1.19	82.60 ± 0.33	89.12 ± 0.29	82.23 ± 0.19
REDL	35.04 ± 0.98	90.11 ± 0.14	93.76 ± 0.11	91.12 ± 0.06	62.17 ± 0.42	78.49 ± 0.13	86.38 ± 0.11	74.18 ± 0.29
HEDL	46.48 ± 0.72	81.73 ± 0.31	89.14 ± 0.22	58.65 ± 0.15	73.07 ± 0.00*	80.05 ± 0.00*	83.18 ± 0.00*	58.36 ± 0.21
FEDL	$\mathbf{19.24 \pm 0.34}$	$\mathbf{93.33 \pm 0.15}$	$\mathbf{96.50 \pm 0.08}$	$\mathbf{93.02 \pm 0.12}$	38.70 ± 2.16	$\mathbf{88.44 \pm 0.53}$	$\mathbf{92.81 \pm 0.42}$	87.38 ± 0.27
Ours	30.23 ± 2.68	90.30 ± 0.59	94.55 ± 0.48	93.00 ± 0.20	44.92 ± 2.39	80.87 ± 0.88	89.23 ± 0.72	$\mathbf{87.63 \pm 0.39}$

This subsection provides complete out-of-distribution detection results on fine-grained benchmarks Flowers102, CUB200, and long-tail CIFAR10-LT. Fine-grained datasets impose considerable chal-

lenges for open-set uncertainty estimation, since intra-dataset classes exhibit subtle visual distinctions and substantial feature overlap, which readily triggers evidence confusion among different classes. CIFAR10-LT is adopted to examine model stability under class-sample imbalance, covering both moderate imbalance ($\eta = 0.1$) and severe imbalance ($\eta = 0.01$). We report quantitative comparisons against both general-purpose open-set state-of-the-art baselines and competing EDL-family methods, and leverage concrete metrics to further verify the generalization capacity of our approach across diverse data distributions.

Table 6 shows OOD detection results against general open-set baselines on Flowers102 and CUB200. Our approach achieves top-tier performance (3.70% FPR95, 98.34% AUPR, and 99.18% AUROC) in CUB200. Meanwhile, certain post-processing techniques such as ODIN, DICE, and ASH record lower FPR95 values than our 12.66% in Flowers102. This outcome aligns closely with the characteristics of Flowers102, where the limited class count (102 classes) and uniform background textures allow post-processing heuristics to easily filter noise via logit clipping. However, these techniques leave the in-distribution accuracy constrained at 83.75%. In contrast, our framework explicitly optimizes evidence alignment and geometric projection over the frozen features, raising Flowers102 ID accuracy to 86.75%. These comparisons demonstrate that while post-processing easily fits OOD metrics on simple datasets with homogenous backgrounds, our approach establishes a better decision boundary while delivering competitive OOD sensitivity.

Table 7 evaluates EDL variants under fine-grained settings. A severe performance collapse occurs for $\mathcal{I}$-EDL and REDL on Flowers102 (AUROC falling to 25.84% and 21.37%), pointing to the vulnerability of standard evidence regularization under high class confusion. On Flowers102, HEDL and FEDL demonstrate strong performance (FPR95 of 5.98% and 8.84%, respectively), as the lower class cardinality allows Euclidean and hyperbolic baselines to establish effective decision boundaries. Nevertheless, our method maintains highly competitive overall capability (97.81% AUROC and 95.99% AUPR). More importantly, as the fine-grained scale increases on CUB200, both FEDL (18.33% FPR95) and HEDL (62.08% FPR95) suffer notable degradation. In contrast, our framework establishes superior OOD detection (3.70% FPR95 and 99.18% AUROC). This contrast highlights that while existing EDL variants perform well on smaller fine-grained subsets, our formulation scales far better to complex, high-cardinality manifolds.

Table 8 summarizes the performance of various EDL variants on CIFAR10-LT under imbalance ratios $\eta = 0.1$ and $\eta = 0.01$. On this dataset, FEDL yields lower FPR95 scores than our approach (19.24% vs. 30.23% for $\eta = 0.1$; 38.70% vs. 44.92% for $\eta = 0.01$). This outcome is expected given that CIFAR10-LT involves only 10 classes, where a low-cardinality label space allows Euclidean embeddings to map skewed evidence without severe spatial overlap. Even so, our framework retains strong in-distribution performance, securing the highest ID accuracy under severe imbalance ($\eta = 0.01$, 87.63% vs. FEDL's 87.38%). Importantly, FEDL's edge in low-dimensional setups fails to generalize to broader spaces; when extended to CIFAR100-LT (Table 3), its FPR95 sharply degrades to 73.59% at $\eta = 0.01$, whereas our method remains substantially more robust at 48.25%. These comparisons demonstrate the superior scalability of our dynamic evidence alignment when tackling complex, high-dimensional long-tailed distributions.

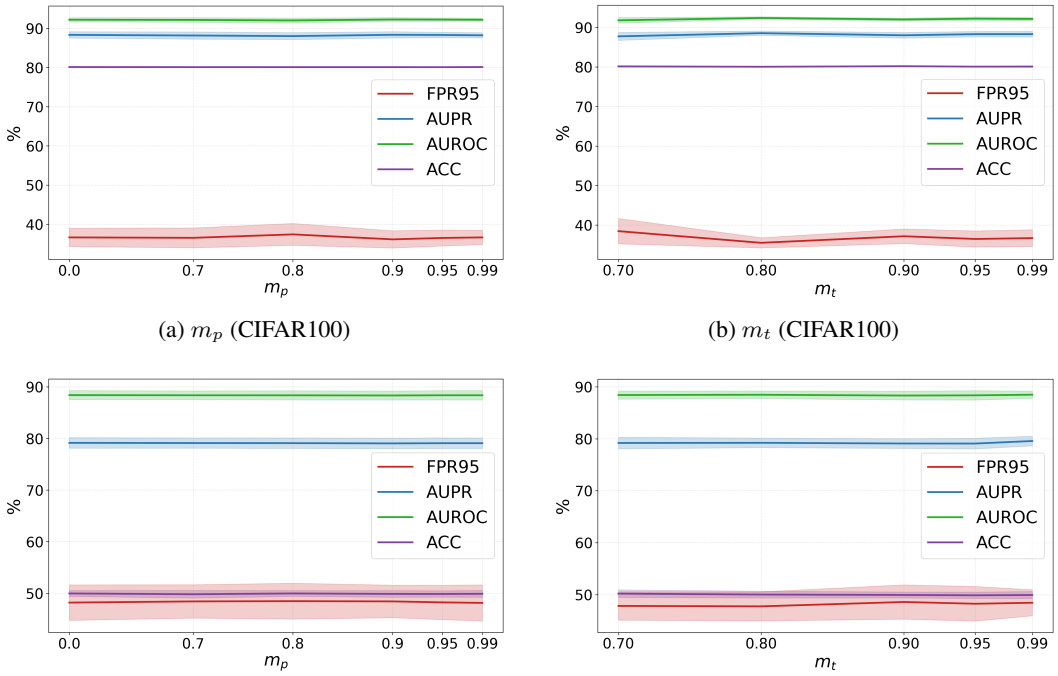
D.2 Additional Ablation Experimental Results

Table 9: Ablation study for loss $\mathcal{L}^{ali}$ using CIFAR100 and CIFAR100-LT ($\eta = 0.01$) as ID datasets.

Method	CIFAR100				CIFAR100-LT ($\eta = 0.01$)			
	Average OOD performance			ID data	Average OOD performance			ID data
	FPR95↓	AUPR↑	AUROC↑	Acc↑	FPR95↓	AUPR↑	AUROC↑	Acc↑
$\mathcal{L}^{ece}$	40.29 ± 1.14	87.88 ± 0.52	91.52 ± 0.18	80.13 ± 0.17	56.07 ± 1.78	76.05 ± 1.04	84.92 ± 0.72	50.89 ± 0.46
$\mathcal{L}^{ece}$ w/ $\mathcal{L}^{ali}$	36.50 ± 2.04	88.32 ± 0.69	92.24 ± 0.43	80.10 ± 0.13	48.25 ± 3.33	79.08 ± 1.03	88.36 ± 0.90	49.88 ± 0.62

Ablation of loss $\mathcal{L}^{ali}$. Table 9 compares model performance with and without the evidence-alignment term $\mathcal{L}^{ali}$. Following the setup in Sec. 3.1, Dirichlet parameters $\boldsymbol{\alpha}_i$ are parameterized exclusively from spatial embedding components $\tilde{\boldsymbol{z}}_{i,s}$. Though this choice yields semantic validity, discarding the temporal coordinate removes inherent magnitude constraints inherited from manifold embeddings. As shown in the table, enabling $\mathcal{L}^{ali}$ yields moderate FPR95 and AUROC

improvements on balanced CIFAR100, while in-distribution accuracy remains comparable. By contrast, larger OOD-detection gains are observed under the long-tailed CIFAR100-LT ($\eta = 0.01$) setting, where performance improvements exceed the range of seed-driven fluctuations within our evaluations. This discrepancy can be understood from the perspective of dataset-wise supervision conditions. Balanced datasets contain sufficient per-class samples such that gradient updates keep evidence magnitudes reasonable for most instances during training, rendering adverse effects from missing magnitude constraints modest. Under long-tailed conditions, sparse supervision for tail classes makes suppressed ground-truth Dirichlet concentrations more likely, aggravating uncertainty mixing between ID and OOD samples. Built upon dynamic targets t_c derived from large-deviation subsets $\mathcal{I}_c^{top}$, $\mathcal{L}^{ali}$ applies constraints on inadequately low evidence outputs to mitigate such undesirable behaviour. Taken together under the dataset configurations tested in this work, $\mathcal{L}^{ali}$ alleviates the lack of native magnitude constraints introduced by spatial-only evidence construction and brings beneficial effects for uncertainty estimation under imbalanced conditions.



(a) m_p (CIFAR100)

(b) m_t (CIFAR100)

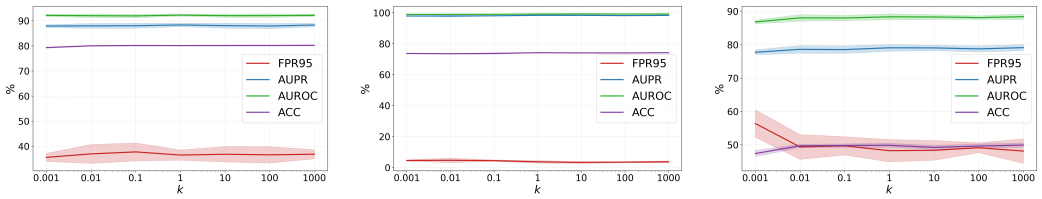
(c) m_p (CIFAR100-LT, $\eta = 0.01$)

(d) m_t (CIFAR100-LT, $\eta = 0.01$)

Figure 3: Ablation study of hyperparameters m_p and m_t.

We evaluate the sensitivity of two EMA momentum terms, m_p and m_t, as shown in Figure 3. This ablation verifies that our default setup ($m_p = 0.95, m_t = 0.95$) lies in a flat, robust regime rather than a fragile local peak. Fixing $m_p = 0.95$, we sweep $m_t \in \{0.7, 0.8, 0.9, 0.95, 0.99\}$. On CIFAR100, $m_t = 0.8$ yields an FPR95 of 35.53%, while our standard setting $m_t = 0.95$ reaches 36.50% FPR95 and 92.24% AUROC. On CIFAR100-LT, FPR95 stays between 47.75% and 48.59% across all values, with accuracy remaining within 49.88%–50.20%. These variations fall well within overlapping standard deviation bounds. Mechanically, m_t controls the temporal smoothing of target-evidence statistics. As long as a reasonable smoothing factor is used, accumulated statistics capture class-wise evidence behavior reliably. Small differences in update speed do not break the corrective property of the evidence alignment loss, confirming that adopting the conventional default $m_t = 0.95$ is safe and reproducible.

Next, fixing $m_t = 0.95$, we sweep $m_p \in \{0.0, 0.7, 0.8, 0.9, 0.95, 0.99\}$, where $m_p = 0.0$ disables EMA and uses raw batch centroids as prototypes. On CIFAR100, FPR95 spans a narrow interval of 36.14%–37.41%, with accuracy holding at roughly 80.10%. On CIFAR100-LT, FPR95 remains stable between 48.15% and 48.49%. Although individual mini-batches provide reasonable class representations under standard training conditions, EMA smoothing serves as a necessary stabilizing mechanism to prevent prototype oscillations under smaller batch sizes or imbalanced sampling.



(a) CIFAR100 (b) CUB200 (c) CIFAR100-LT ($\eta = 0.01$)

Figure 4: Ablation of scale parameter k. All subplots use log-scale for x-axis. Shaded bands denote standard deviation over five independent runs. Lower FPR95 is better; higher AUPR, AUROC and Acc are better.

These results indicate that the module is highly tolerant to the choice of momentum coefficient, maintaining consistent performance across a broad parameter range.

Taken together, momentum smoothing for target statistics and class prototypes shows broad tolerance to parameter variations. This insensitivity stems from the effectiveness of the loss constraint in the evidence space: as long as evidence alignment shapes the decision boundary properly, minor fluctuations in feature estimation do not disturb final predictions. This behavior indicates that the proposed approach operates reliably without dataset-specific hyperparameter tuning.

Ablation of hyperboloid scale parameter k. To investigate the effect of hyperboloid scale parameter k, we conduct ablation experiments over $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$. This parameter controls the bending of the Lorentz hyperboloid with sectional curvature $K_{sec} = -1/k$: large k approaches Euclidean geometry, while smaller k strengthens the hierarchical modelling capacity of hyperbolic representations. These ablation results are obtained with our full LEDL pipeline (spatial-component-only evidence construction plus evidence-alignment loss $\mathcal{L}^{ali}$); note that curvature-dependent model responses may differ if $\mathcal{L}^{ali}$ is disabled. Results are summarised in Figure 4. Each subplot reports FPR95, AUPR, AUROC and closed-set accuracy, with shaded bands showing standard deviations over five independent runs. From the empirical results, on balanced CIFAR100, metric fluctuations across tested k mostly stay within standard-deviation bounds and no clear monotonic trend emerges, suggesting that under our LEDL framework, balanced classification tasks exhibit low sensitivity to hyperbolic hierarchical inductive bias. For fine-grained CUB-200, OOD metrics degrade slightly for large-k near-Euclidean values; peak performance occurs roughly within $k \in [1, 10]$, and further increasing k only yields narrow-band metric oscillations. A noteworthy observation comes from heavily long-tailed CIFAR100-LT: at strongly-hyperbolic $k = 0.001$, FPR95 rises noticeably alongside reduced AUPR and AUROC. This outcome indicates that excessively strong hyperbolic bending induced by tiny k degrades performance, rather than Euclidean proximity. After leaving this extreme strongly-hyperbolic regime, performance does not monotonically improve with growing k but enters a stable plateau. This observation can be interpreted from two angles: smaller k improves the hierarchical representational potential of the hyperbolic manifold, yet very small k can also introduce undesired numerical effects. These competing tendencies together produce a broad operating parameter window. Taken together, the model's response to curvature varies with dataset characteristics; substantial performance degradation occurs only under the extreme strongly-hyperbolic small-k condition. Within reasonable hyperbolic parameter regimes, the joint influence of hierarchical representational potential and numerical side-effects keeps model metrics within a relatively stable range.

Uncertainty Distribution Visualization. Figures 5, 6, 7 show uncertainty densities for balanced CIFAR100 paired with SVHN, DTD and Places365, while Figures 8, 9, 10 correspond to long-tailed CIFAR100-LT with the same three out-of-distribution corpora. We plot uncertainty density curves for EDL, $\mathcal{I}$-EDL, REDL, HEDL, FEDL and our method to compare output behaviours across these variants. For cleaner density-plot visualization, we apply per-setup min-max normalization over raw uncertainty values u for each ID-OOD configuration: $u_{\text{norm}} = (u - \min u)/(\max u - \min u)$. We note that all reported quantitative OOD-detection metrics are computed using raw, unnormalized uncertainty scores. Taking CIFAR100 paired with SVHN as a representative example, the normalized density curves show that our method yields lower uncertainty for in-distribution samples and forms more compact distributions for out-of-distribution instances, yielding improved separability;

this general pattern is also observable for DTD and Places365. Under balanced CIFAR100, most baseline models show substantial overlap between ID and OOD uncertainty densities. Some baselines achieve partial separation yet still suffer from tail-region mixing. Upon moving to long-tailed CIFAR100-LT, separation performance drops for all compared approaches, and many in-distribution samples receive high uncertainty values which further enlarges distribution overlap. Across all six ID-OOD pairs, our method exhibits visually favourable distribution separation: in-distribution samples cluster at low uncertainty values while out-of-distribution samples shift toward higher scores. Even though separation degrades somewhat under long-tailed conditions, the overall overlap remains smaller relative to other compared approaches. These qualitative observations agree with FPR95, AUROC quantitative results from prior tables and provide supplementary distribution-level evidence for OOD-detection performance; mechanistic discussions can be found in the preceding ablation sections.

E Limitations and Future Work

While LEDL demonstrates solid uncertainty estimation across diverse open-set scenarios, several scope constraints and future directions remain. First, our evaluation focuses primarily on static 2D image classification tasks; extending LEDL to complex multi-modal settings, such as video action recognition and 3D point cloud understanding, will help test its spatiotemporal applicability. Second, the current evidential inference relies on prespecified prior hyperparameters for evidence scaling, where exploring data-driven adaptive uncertainty calibration mechanisms presents a promising path to enhance robustness under varying uncertainty regimes. Third, to better handle near-boundary out-of-distribution samples that lie extremely close to in-distribution margins, future work can incorporate dynamic metric learning and boundary refinement to explicitly optimize distance mappings near hyperbolic manifold boundaries, thereby improving discrimination against highly ambiguous instances.

F AI Tool Disclosure

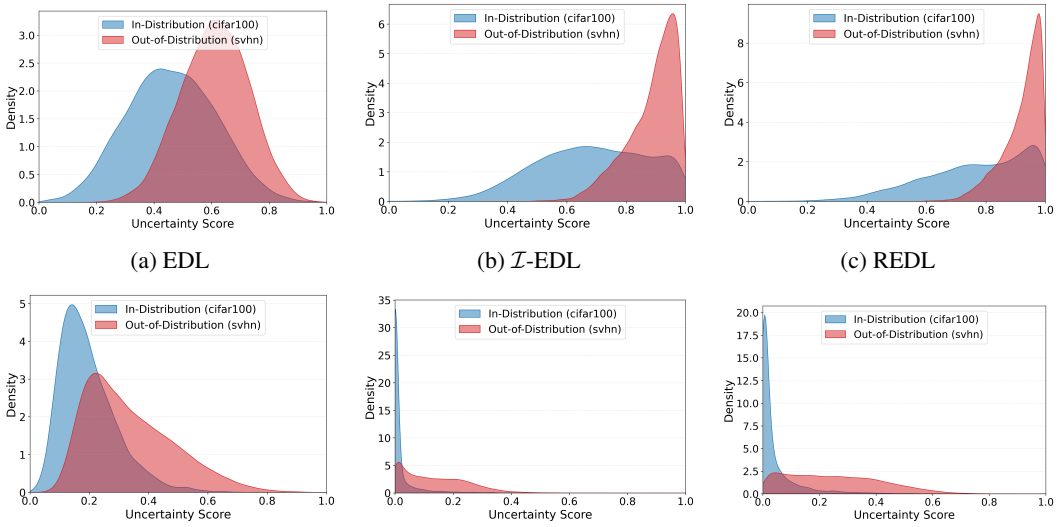
AI-assisted writing tools have been utilized to aid phrasing and polish the manuscript text. All scientific ideas, problem formulation, mathematical derivations, algorithm design, implementation, experimental results and analyses are completed entirely by the authors. No AI-generated scientific content is adopted in this work.

G Ethics Statement

This work involves no human participants, personally identifiable information or sensitive private data. All experiments are carried out on publicly-available benchmark datasets with their official public licenses; no proprietary or access-restricted datasets are adopted. Our approach targets out-of-distribution detection for image classification and does not create foreseeable safety risks. The authors confirm that this work adheres to the ICLR Code of Ethics and follows common standards for research integrity and transparency.

H Reproducibility Statement

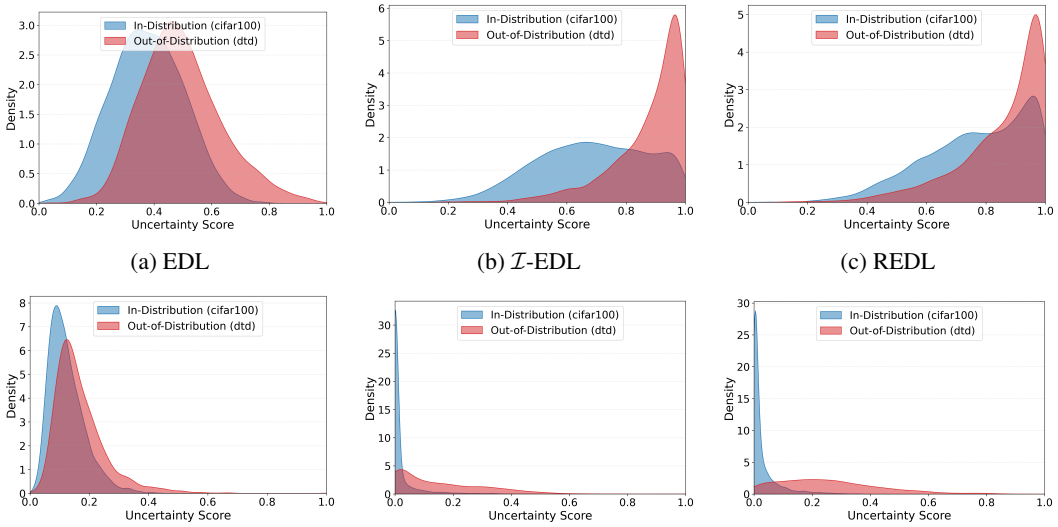
Our code is available at `https://github.com/anonymous/anonymous`. Full source code will be released upon paper acceptance.



(a) EDL (b) $\mathcal{I}$-EDL (c) REDL

(d) HEDL (e) FEDL (f) ours

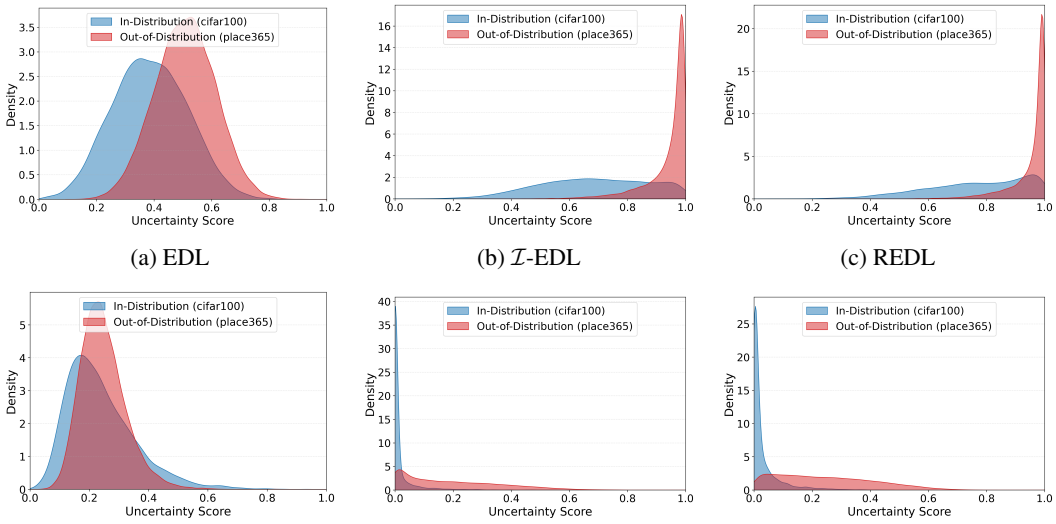
Figure 5: Uncertainty distribution on CIFAR100 as ID datasets and SVHN as OOD datasets.



(a) EDL (b) $\mathcal{I}$-EDL (c) REDL

(d) HEDL (e) FEDL (f) ours

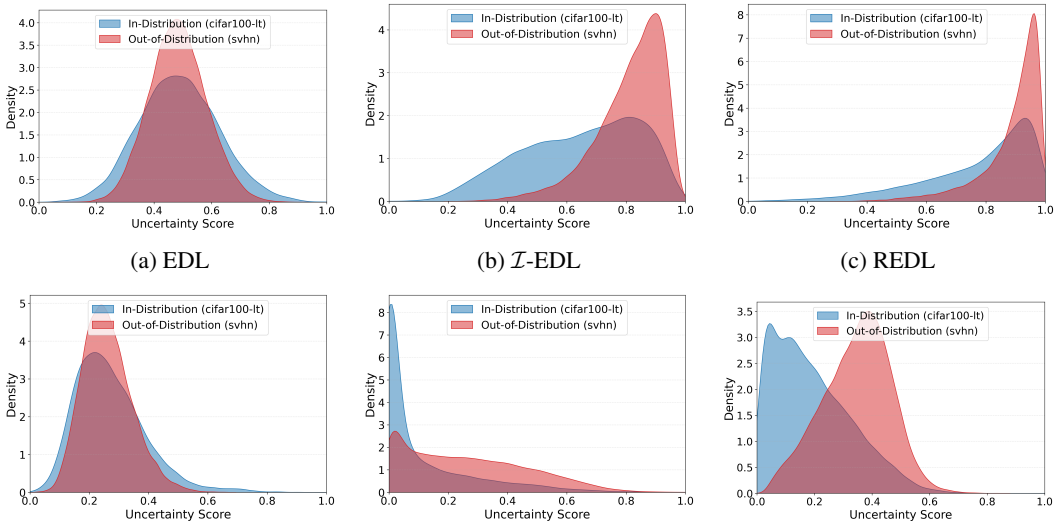
Figure 6: Uncertainty distribution on CIFAR100 as ID datasets and DTD as OOD datasets.



(a) EDL (b) $\mathcal{I}$-EDL (c) REDL

(d) HEDL (e) FEDL (f) ours

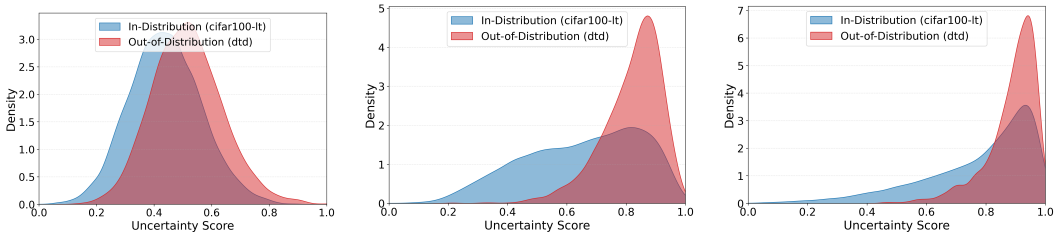
Figure 7: Uncertainty distribution on CIFAR100 as ID datasets and Place365 as OOD datasets.



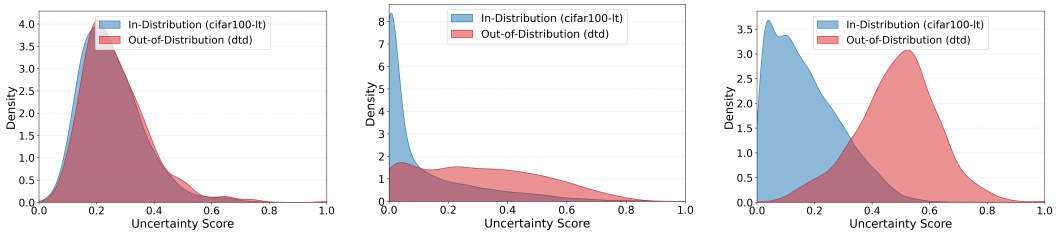
(a) EDL (b) $\mathcal{I}$-EDL (c) REDL

(d) HEDL (e) FEDL (f) ours

Figure 8: Uncertainty distribution on CIFAR100 as ID datasets and SVHN as OOD datasets.

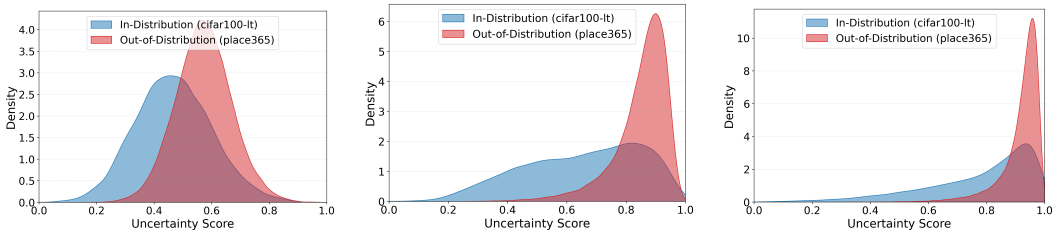


(a) EDL (b) $\mathcal{I}$-EDL (c) REDL

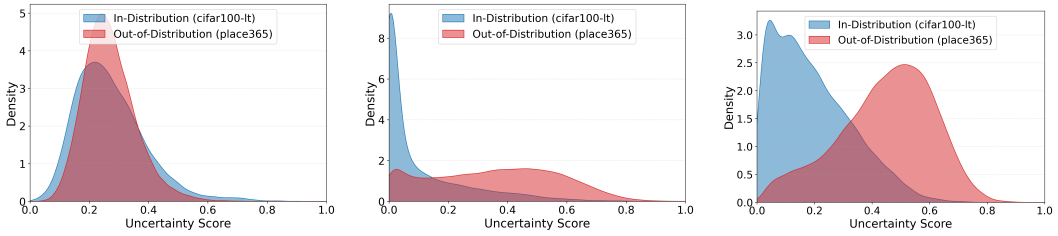


(d) HEDL (e) FEDL (f) ours

Figure 9: Uncertainty distribution on CIFAR100 as ID datasets and DTD as OOD datasets.



(a) EDL (b) $\mathcal{I}$-EDL (c) REDL



(d) HEDL (e) FEDL (f) ours

Figure 10: Uncertainty distribution on CIFAR100 as ID datasets and Place365 as OOD datasets.