DDGE: Disentangled Dirichlet Geodesic Evaluation for Robust Few-Shot Learning

Yifan Mei[1], **Kun Zhou**[2,3], **Zelei Wu** [1], **Jieyu Zhao**[1], **Xulun Ye**[1,†]
[1] School of Artificial Intelligence, Ningbo University, Ningbo, China
[2] School of Architecture & Urban Planning, Shenzhen University, Shenzhen, China
[3]Intelligent Computing Center, Shenzhen University
{2411100090, zhao_jieyu, yexulun}@nbu.edu.cn, zhoukun@szu.edu.cn

Abstract

Recent studies have highlighted the crucial role of distance metrics in noisy few-shot learning. However, existing approaches heavily rely on Euclidean or cosine measurements, failing to recognize the "deceptive proximity" where heterogeneous samples from different manifold peaks appear spuriously close in the flat ambient space. In this paper, we present Disentangled Dirichlet Geodesic Evaluation (DDGE), a novel manifold-aware framework that unifies orthogonal disentanglement and conformal integrals to efficiently evaluate complex topological spaces. Specifically, we decouple features into orthogonal semantic subspaces and leverage a prior-guided Dirichlet Process to expand discrete samples into a continuous, noise-purified semantic terrain. Upon this landscape, we measure the intrinsic geodesic distance via a density-aware conformal line integral to capture the authentic class topology. Extensive experiments demonstrate the state-of-the-art performance of DDGE on multiple benchmark datasets. Code is available at: https://github.com/24sleeping90/DDGE

1 Introduction

Few-Shot Learning (FSL) aims to recognize novel categories from strictly limited samples [1–4]. In FSL, learning a generalizable metric space is crucial. Previous metric-based approaches, such as Prototypical Networks [3], Matching Networks [1], and Relation Networks [4], map support and query samples into a high-dimensional semantic space and conduct classification. These methods typically employ Euclidean distance or cosine similarity to measure absolute point-to-point distances [5]. This formulation implicitly relies on a flat Euclidean space assumption, treating feature representations as points in an isotropic space where geometric proximity equates to semantic similarity.

However, this linear assumption overlooks the intrinsic topological structure of high-dimensional representations. As foundational manifold learning theories suggest [6], high-dimensional sensory data typically reside on low-dimensional non-linear manifolds embedded within a noisy ambient space. Under real-world topological contamination, such as out-of-distribution (OOD) anomalies [7, 8], standard Euclidean metrics often struggle. Specifically, high-dimensional spaces contain complex landscapes with high-density semantic regions and low-density voids. Linear metrics can produce a "deceptive proximity," potentially misjudging heterogeneous samples on different semantic peaks as similar simply because a straight line crosses the empty ambient space [9]. Capturing true semantic relationships requires navigating intrinsic geodesic paths. Existing robust FSL approaches primarily rely on deterministic hard-truncation [10] or global reweighting [11]. Yet, constrained by their reliance on globally flat spaces, they may discard genuine hard positives and struggle to isolate camouflaged topological outliers.

[†]Corresponding author.

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

To address these limitations, we propose Disentangled Dirichlet Geodesic Evaluation (DDGE), a manifold-aware framework for robust FSL. We shift the metric evaluation from point-to-point linear measurements to path-aware manifold evaluation. Our approach evaluates the path integral resistance of two points on an estimated semantic manifold, rather than calculating their absolute displacement in an empty space. Specifically, to recover a reliable class manifold under extreme data scarcity, we design a metric pipeline that operates gradient-free during the meta-testing phase. First, during base training, we learn a set of orthogonal projections to decouple high-dimensional features into compact semantic subspaces. This promotes local noise isolation and improves the numerical stability of low-dimensional manifold estimation. Subsequently, during meta-testing, the sparse support and query samples are also projected into these subspaces. Here, we inject geometric priors into the isolated scattered points and utilize a density-aware Dirichlet Process [12] for soft-calibrated proliferation, forming a noise-purified Gaussian Mixture Model (GMM) density landscape. Finally, we introduce a Multi-View Conformal Line Integral to approximate the geodesic distance. This mechanism dynamically evaluates path resistance across all decoupled dimensions, imposing an exponential penalty on trajectories traversing low-density topological voids, which helps mitigate the impact of adversarial noise.

Our core contributions are summarized as follows:

- We propose DDGE, a manifold-aware framework for robust FSL. It addresses the limitations of globally flat Euclidean metric learning by introducing a density-aware path evaluation metric.
- We design a robust manifold construction pipeline. Orthogonal disentanglement provides structural noise isolation during representation learning, while a soft-calibrated Dirichlet Process expands sparse support samples into a continuous semantic landscape without gradient updates during inference.
- Extensive experiments demonstrate that DDGE achieves state-of-the-art performance on multiple benchmarks under varying OOD noise conditions. The multi-view conformal integral mechanism effectively suppresses complex topological contamination while maintaining highly competitive performance on standard clean datasets.

2 Related work

2.1 Metric-based few-shot learning

The dominant paradigm in Few-Shot Learning (FSL) revolves around metric learning, which identifies novel classes via learned distance functions [1, 2]. Foundational works [3, 1, 4] evaluate query-to-support similarities, while recent advances leverage foundation models [13, 14] alongside advanced adaptation paradigms (e.g., prompt tuning [15]) or complex alignment mechanisms [16, 17] to extract richer semantics. Despite promising results on well-curated datasets [5], these methods implicitly rely on linear metrics like Euclidean or cosine distance, assuming a globally flat geometric space. Consequently, they evaluate absolute distances across empty ambient dimensions, making them susceptible to deceptive proximity rather than capturing the intrinsic manifold topology.

2.2 Robust FSL and outlier mitigation

To address topological contamination, such as out-of-distribution (OOD) anomalies [7, 8], robust FSL methods typically employ attention-based reweighting [11], feature masking [9], collaborative feature disentanglement [18], or outlier rejection [19–21]. However, deterministic hard-truncation strategies [10] risk discarding genuine hard positives, particularly in the extreme low-data regime (e.g., 5-shot). Moreover, under adversarial feature contamination [22], holistic feature evaluations remain highly susceptible to camouflaged topological noise. In contrast, DDGE employs a leave-one-out soft calibration mechanism and orthogonal disentanglement, effectively isolating local noise without relying on aggressive truncation.

2.3 Distribution calibration and manifold approximation

To mitigate data scarcity, distribution calibration methods [23] transfer geometric priors from base classes. For instance, MetaDiff [24] hallucinates virtual features by borrowing Gaussian mixture

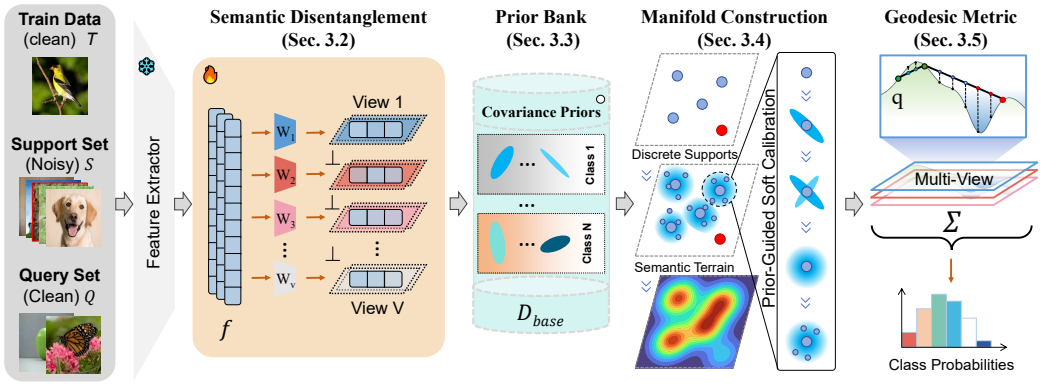


Figure 1: Overview of the DDGE framework. Operating on a frozen backbone (Φ), our pipeline comprises: (1) Orthogonal semantic disentanglement (Section 3.2); (2) Prior bank retrieval (Section 3.3); (3) DP-based manifold construction (Section 3.4); and (4) Geodesic metric evaluation (Section 3.5).

statistics [25]. However, directly transferring covariance matrices in ultra-high-dimensional spaces often encounters extreme rank deficiency. Standard Tikhonov regularization here inflates the null space into an isotropic noise sphere, distorting intrinsic semantic geometry. Unlike approaches that merely augment prototypes, DDGE applies orthogonal disentanglement [26] to bypass the rank paradox. This mathematically ensures a stable, noise-purified GMM landscape, providing a reliable geometric foundation for density-aware geodesic distance approximation [27].

3 Methodology

3.1 Problem formulation and overview

Problem statement. We consider the standard N-way K-shot few-shot learning (FSL) paradigm [1, 3]. Let $\mathcal{D}_{\text{base}}$ and $\mathcal{D}_{\text{novel}}$ denote the base and novel datasets with disjoint label spaces. In a specific task, the model is provided with a support set $\tilde{S} = \{(x_i, y_i)\}_{i=1}^{N\times K}$ and a query set Q. A pre-trained feature extractor $\Phi(\cdot)$ maps each image into a high-dimensional feature space $f \in \mathbb{R}^D$. To reflect real-world deployment, we assume the support set is subject to unknown topological contamination: $\tilde{S} = S_{\text{clean}} \cup S_{\text{noise}}$, where S_{noise} primarily comprises out-of-distribution (OOD) open-set anomalies [8]. Our objective is to learn a robust metric that accurately classifies Q by geometrically isolating S_{noise} without gradient updates during the meta-testing phase.

Overview. To achieve this, the proposed DDGE framework (Figure 1) bridges implicit foundation knowledge and explicitly computable metric spaces. Specifically, it sequentially decouples features into robust sub-spaces, inflates sparse supports into a continuous density landscape, and evaluates geodesic path resistance to reject topological outliers. Comprehensive mathematical formulations are provided in Appendix A.

3.2 Orthogonal semantic disentanglement

When dealing with global features $f \in \mathbb{R}^D$ extracted by pre-trained backbones, semantic concepts are often heavily entangled. Directly estimating geometric distributions in this high-dimensional space suffers from the curse of dimensionality and allows local topological contamination (e.g., a noisy background) to skew the global distance metric.

To address entanglement and establish a tractable density estimation foundation, we decompose the global feature f into V parallel, compact semantic views during base training [26]. We employ a set of independent projection matrices $\{W_v\}_{v=1}^{V}$, where $W_v \in \mathbb{R}^{d\times D}$ maps the global feature to a low-dimensional space ($d \ll D$). The projected feature for a sample x_i in view v is denoted as $x_i^{(v)} \in \mathbb{R}^d$. Restricting the subspace dimension to d mathematically ensures that these compact spaces are comprehensively spanned by the base training data, mitigating the impact of null-dimensions that cause rank deficiency in raw high-dimensional estimations.

To prevent these mappers from collapsing into redundant representations and to actively combat noise contamination, we impose an orthogonality constraint on the row spaces of the projection matrices:

$$\mathcal{L}_{\text{ortho}} = \sum_{i=1}^{V-1} \sum_{j=i+1}^{V} \|W_i W_j^\top\|_F^2 \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. By optimizing $\mathcal{L}_{\text{ortho}}$ alongside the standard metric-based classification loss [3] during base training, we structurally encourage the V semantic views to be mutually orthogonal.

This orthogonality structurally establishes a semantic isolation mechanism. If a corrupted support sample is contaminated by topological noise, the perturbation is highly likely confined to a limited subset of views, leaving the remaining orthogonal semantics intact. This isolation prevents local noise from counterfeiting the global topology, laying the structural groundwork for the multi-view rejection mechanism evaluated in Section 3.5.

3.3 Prior bank and soft calibration

Despite the mathematical tractability achieved via orthogonal projection, the discrete support samples in each view v still behave as dimensionless points. To explicitly parameterize this implicit geometry into a continuous density function, they require physical volume to formulate valid covariance matrices. To circumvent the difficulty of estimating empirical covariance from scarce samples (e.g., $K = 5$), we adapt a prior-guided distribution calibration [23] within each decoupled view. Specifically, within each sub-space, we construct a prior bank from $\mathcal{D}_{\text{base}}$, utilizing the L_2-normalized base class centroids as addressing keys $K_{\text{base}}^{(v)} \in \mathbb{R}^{N_{\text{base}} \times d}$ and their full-rank covariance matrices as geometric priors $\{\Sigma_b^{(v)}\}_{b=1}^{N_{\text{base}}}$. For a normalized support sample feature $\tilde{x}_i^{(v)} = x_i^{(v)} / \|x_i^{(v)}\|_2$, we compute semantic similarity via attention to retrieve a weighted covariance prior:

$$\Sigma_{\text{stable},i}^{(v)} = \sum_{b=1}^{N_{\text{base}}} \text{Softmax}\left(\tau \cdot (\tilde{x}_i^{(v)} (K_{\text{base}}^{(v)})^\top)\right) \Sigma_b^{(v)} + \epsilon I \tag{2}$$

where τ is the temperature scaling and ϵI represents the Tikhonov regularization. This interpolation inflates each discrete point into a valid d-dimensional Gaussian distribution $\mathcal{N}(x_i^{(v)}, \Sigma_{\text{stable},i}^{(v)})$.

Having established the geometric volume, we must evaluate the structural purity of each sample to mitigate outliers. Instead of relying on deterministic distance thresholds or hard-truncation [11, 10], which risk discarding genuine hard positives, we formulate Bayesian uncertainty through a leave-one-out (LOO) density estimation within the K support samples of each specific class. Reverting to the unnormalized projected space, for a given sample $x_i^{(v)}$ belonging to a specific class, the local density score $s_i^{(v)}$ is defined as the expected multivariate Gaussian likelihood of the raw feature $x_i^{(v)}$ relative to the distributions formed by the remaining intra-class samples:

$$s_i^{(v)} = \frac{1}{K-1} \sum_{j \neq i}^{K} \mathcal{N}\left(x_i^{(v)} \;\middle|\; x_j^{(v)}, \Sigma_{\text{stable},j}^{(v)}\right) \tag{3}$$

These raw density scores are translated into normalized generative quotas via a softmax function parameterized by a calibration temperature ϕ:

$$w_i^{(v)} = \frac{\exp(s_i^{(v)}/\phi)}{\sum_{j=1}^{K} \exp(s_j^{(v)}/\phi)} \tag{4}$$

Through this soft calibration, authentic samples near the manifold core are assigned higher distribution weights, while topological outliers are smoothly calibrated to near-zero confidence. This structurally mitigates outlier influence while preserving the intrinsic topology of the latent space.

3.4 Manifold construction via DP

Equipped with the validated geometric volume and generative quotas $w_i^{(v)}$, we explicitly fit the foundation model's underlying topology. We utilize the Dirichlet Process (DP) [12] as a robust

parametric fitting mechanism to translate implicit semantic representations into a mathematically computable, continuous density landscape, using $w_i^{(v)}$ as the routing probabilities.

To execute this proliferation while maintaining gradient-free efficiency, we employ the standard reparameterization trick [28]. For a selected support sample $x_i^{(v)}$ in the unnormalized space, we generate a set of virtual samples $x_{\text{virtual}}^{(v,m)} = x_i^{(v)} + L_i^{(v)} z^{(m)}$, where $L_i^{(v)}$ is the Cholesky decomposition of $\Sigma_{\text{stable},i}^{(v)}$, $z^{(m)} \sim \mathcal{N}(0, I)$, with a quota proportional to $N_{\text{virtual}} w_i^{(v)}$. These virtual samples bridge the sparse supports into a denser empirical distribution.

Defining a superset $S_{\text{total}}^{(v)}$ encompassing both original weighted samples and hallucinated virtual points, we formulate the noise-purified Gaussian Mixture Model (GMM) [25] landscape for view v:

$$P_{\text{GMM}}^{(v)}(x) = \frac{1}{|S_{\text{total}}^{(v)}|} \sum_{x_j \in S_{\text{total}}^{(v)}} \mathcal{N}\left(x \;\middle|\; x_j^{(v)}, \Sigma_{\text{stable},j}^{(v)}\right) \tag{5}$$

where each hallucinated virtual sample x_j inherits the stable covariance matrix Σ_{stable} of its parent support. This DP formulation structurally induces a concentration of density around genuine samples while limiting the propagation of corrupted outliers. Consequently, $P_{\text{GMM}}^{(v)}(x)$ accurately captures the true intrinsic topology of the novel class, providing a noise-purified continuous landscape for the subsequent geodesic evaluation.

3.5 Multi-view conformal integral

With the manifold explicitly parameterized by the differentiable GMM landscape $P_{\text{GMM}}^{(v)}(x)$, the underlying topology becomes geometrically calculable. Ideally, semantic similarity between a query sample q and a novel class should be measured by the exact geodesic path following the high-density regions of the data manifold. However, solving the Eikonal equation for this path [27] is computationally prohibitive for FSL fast adaptation.

To achieve an efficient approximation, we propose a density-aware conformal line integral that probes the Euclidean straight line, measuring path resistance over the local density landscape; the result serves as an upper-bound approximation to the geodesic distance, rather than the shortest geodesic path itself. For a specific decoupled view v, we define the parameterized linear path as $\gamma^{(v)}(t) = q^{(v)} + t(c^{(v)} - q^{(v)})$ for $t \in [0, 1]$, where $c^{(v)} = \sum_{i=1}^{K} w_i^{(v)} x_i^{(v)}$ is the density-weighted robust prototype of the novel class.

To penalize trajectories traversing low-density out-of-distribution regions [7], we formulate an exponential resistance function:

$$R^{(v)}(x) = 1 + \alpha \exp\left(-\beta \cdot P_{\text{GMM}}^{(v)}(x)\right) \tag{6}$$

where α defines the maximum resistance penalty for crossing empty space, and β controls the sensitivity to density variations. If the path traverses high-density semantic regions, the resistance decays to 1, degenerating to the standard Euclidean distance [3]. Conversely, traversing a low-density void scales the local resistance exponentially. Integrating this resistance along the parameterized path requires scaling by the physical arc length element. Thus, the conformal line integral distance in view v is efficiently approximated via a discrete Riemann summation with M uniformly spaced probes ($t_m = m/M$), scaled by the base Euclidean distance:

$$D_{\text{conformal}}^{(v)}(q, c) \approx \left\| c^{(v)} - q^{(v)} \right\|_2 \cdot \frac{1}{M} \sum_{m=1}^{M} R^{(v)}(\gamma^{(v)}(t_m)) \tag{7}$$

Ultimately, the global conformal distance is reconstructed by aggregating these independent integrals across all orthogonal sub-spaces: $D_{\text{total}}(q, c) = \sum_{v=1}^{V} D_{\text{conformal}}^{(v)}(q, c)$.

This multi-view integration ensures a robust evaluation. While a camouflaged noise sample might align with the clean distribution in one semantic dimension, the orthogonality constraint [26] prevents it from mimicking high-density paths across all sub-manifolds simultaneously. A deviation into a low-density void in any single view incurs a strict penalty. Consequently, DDGE effectively mitigates the limitations of linear metrics, ensuring that this upper-bound conformal distance reliably favors genuine intra-class samples. The complete inference pipeline is detailed in Appendix B.

Table 1: 5-way 5-shot classification accuracy (%) under varying OOD noise ratios on *mini*ImageNet and *tiered*ImageNet (grouped by backbones). Best results are in bold.

Method	*mini*ImageNet (5-shot)				*tiered*ImageNet (5-shot)			
	0%	20%	40%	60%	0%	20%	40%	60%
MAML (ICML'17)	$62.42_{\pm 0.18}$	$55.88_{\pm 0.19}$	$51.05_{\pm 0.19}$	$41.12_{\pm 0.18}$	$64.21_{\pm 0.19}$	$57.65_{\pm 0.19}$	$50.44_{\pm 0.20}$	$42.35_{\pm 0.21}$
ProtoNet (NeurIPS'17)	$70.45_{\pm 0.16}$	$64.52_{\pm 0.17}$	$56.12_{\pm 0.18}$	$46.45_{\pm 0.20}$	$71.62_{\pm 0.18}$	$68.22_{\pm 0.19}$	$60.45_{\pm 0.20}$	$49.33_{\pm 0.24}$
AM3 (NeurIPS'19)	$78.41_{\pm 0.27}$	$68.85_{\pm 0.27}$	$57.12_{\pm 0.29}$	$43.15_{\pm 0.36}$	$82.85_{\pm 0.22}$	$71.25_{\pm 0.21}$	$58.95_{\pm 0.30}$	$44.25_{\pm 0.31}$
RNNP (WACV'21)	$66.42_{\pm 0.17}$	$62.15_{\pm 0.18}$	$57.34_{\pm 0.19}$	$45.92_{\pm 0.20}$	$70.51_{\pm 0.18}$	$66.88_{\pm 0.19}$	$60.12_{\pm 0.20}$	$51.12_{\pm 0.21}$
SVAE-Proto (CVPR'21)	$83.42_{\pm 0.42}$	$71.05_{\pm 0.43}$	$56.25_{\pm 0.43}$	$48.15_{\pm 0.52}$	$85.15_{\pm 0.49}$	$72.85_{\pm 0.52}$	$62.12_{\pm 0.53}$	$52.85_{\pm 0.52}$
TraNFS (CVPR'22)	$68.35_{\pm 0.17}$	$65.55_{\pm 0.18}$	$59.88_{\pm 0.23}$	$48.95_{\pm 0.25}$	$71.42_{\pm 0.21}$	$68.65_{\pm 0.22}$	$63.25_{\pm 0.23}$	$52.55_{\pm 0.26}$
TSA (CVPR'22)	$76.55_{\pm 0.16}$	$68.55_{\pm 0.20}$	$59.25_{\pm 0.18}$	$46.35_{\pm 0.20}$	$78.95_{\pm 0.21}$	$69.95_{\pm 0.23}$	$60.45_{\pm 0.25}$	$47.35_{\pm 0.27}$
FGFL (ICCV'23)	$86.95_{\pm 0.61}$	$77.85_{\pm 0.71}$	$66.25_{\pm 0.75}$	$55.45_{\pm 0.81}$	$87.55_{\pm 0.59}$	$74.95_{\pm 0.63}$	$60.15_{\pm 0.70}$	$54.85_{\pm 0.75}$
DETA (ICCV'23)	$83.62_{\pm 0.34}$	$73.85_{\pm 0.29}$	$67.45_{\pm 0.35}$	$51.95_{\pm 0.39}$	$80.75_{\pm 0.30}$	$71.15_{\pm 0.36}$	$64.45_{\pm 0.31}$	$50.55_{\pm 0.33}$
MetaDiff (AAAI'24)	$81.55_{\pm 0.55}$	$75.95_{\pm 0.64}$	$68.25_{\pm 0.74}$	$52.15_{\pm 1.21}$	$85.65_{\pm 0.60}$	$77.95_{\pm 0.65}$	$72.85_{\pm 0.77}$	$53.65_{\pm 0.98}$
SemFew (CVPR'24)	$85.75_{\pm 0.55}$	$80.85_{\pm 0.68}$	$73.95_{\pm 0.76}$	$59.25_{\pm 0.98}$	$88.25_{\pm 0.51}$	$82.85_{\pm 0.62}$	$74.95_{\pm 0.72}$	$60.45_{\pm 0.82}$
FedFSL-CFRD (AAAI'25)	$71.42_{\pm 0.97}$	$61.95_{\pm 1.08}$	$57.35_{\pm 1.21}$	$45.45_{\pm 1.26}$	$78.42_{\pm 1.16}$	$65.85_{\pm 1.22}$	$52.12_{\pm 1.26}$	$51.65_{\pm 1.34}$
ECER-FSL (AAAI'25)	$92.65_{\pm 0.28}$	$85.25_{\pm 0.28}$	$75.45_{\pm 0.31}$	$60.95_{\pm 0.35}$	$91.65_{\pm 0.23}$	$85.95_{\pm 0.26}$	$74.45_{\pm 0.31}$	$57.85_{\pm 0.35}$
LDC (CVPR'25)	$91.65_{\pm 0.22}$	$84.15_{\pm 0.24}$	$76.25_{\pm 0.25}$	$57.45_{\pm 0.30}$	$90.65_{\pm 0.20}$	$82.85_{\pm 0.22}$	$74.15_{\pm 0.22}$	$57.25_{\pm 0.27}$
SemFew-Trans (CVPR'24)	$86.85_{\pm 0.49}$	$81.95_{\pm 0.54}$	$70.45_{\pm 0.59}$	$55.85_{\pm 0.67}$	$89.25_{\pm 0.54}$	$84.95_{\pm 0.59}$	$72.15_{\pm 0.64}$	$57.95_{\pm 0.68}$
DETA++ (TPAMI'25)	$90.65_{\pm 0.37}$	$79.85_{\pm 0.40}$	$73.45_{\pm 0.48}$	$57.65_{\pm 0.51}$	$85.75_{\pm 0.40}$	$76.85_{\pm 0.42}$	$70.45_{\pm 0.45}$	$56.45_{\pm 0.49}$
DDGE (ResNet-12)	$76.73_{\pm 0.64}$	$70.56_{\pm 0.67}$	$63.00_{\pm 0.80}$	$53.04_{\pm 0.85}$	$82.31_{\pm 0.67}$	$77.63_{\pm 0.79}$	$70.86_{\pm 0.90}$	$59.97_{\pm 0.92}$
DDGE (CLIP)	$91.57_{\pm 0.29}$	$\mathbf{90.57}_{\pm 0.45}$	$\mathbf{84.34}_{\pm 0.69}$	$\mathbf{71.94}_{\pm 0.88}$	$92.98_{\pm 0.35}$	$\mathbf{91.26}_{\pm 0.52}$	$\mathbf{86.84}_{\pm 0.74}$	$\mathbf{74.41}_{\pm 0.94}$
DDGE (Swin-T)	$\mathbf{93.16}_{\pm 0.23}$	$90.12_{\pm 0.33}$	$83.89_{\pm 0.54}$	$70.96_{\pm 0.80}$	$\mathbf{93.21}_{\pm 0.21}$	$91.06_{\pm 0.27}$	$85.02_{\pm 0.51}$	$72.26_{\pm 0.76}$

4 Experiments

4.1 Experimental setup

Datasets and baselines. We evaluate our framework on two standard FSL benchmarks, *mini*ImageNet [1] and *tiered*ImageNet [29], both of which are derived from the large-scale ImageNet dataset [30]. To validate generalizability, we extend evaluations to cross-domain (FC100 [31], CIFAR-FS [32]) and real-world noisy datasets (CIFAR-100N [33], Animal-10N [34], and Food-101N [35]). We compare against a comprehensive suite of state-of-the-art methods, spanning metric learning (AM3 [11]), transformer alignments (TSA [16]), distribution calibration (MetaDiff [24]), and robust frameworks (SemFew [19], ECER-FSL [20], and LDC [21]).

Noise protocol and implementation. Following robust FSL paradigms, we inject topological contamination into the 5-way 5-shot support sets by globally replacing a randomly sampled proportion $r \in \{0\%, 20\%, 40\%, 60\%\}$ of all support instances with Out-of-Distribution (OOD) open-set anomalies (or Systematic (SYS) label-swaps [8]), keeping the support-set size constant. Unless specified, we adopt frozen Swin-T [36] as the backbone. Computational analyses and structural parameter dynamics (e.g., V, N_{virtual}, M) are detailed in Appendices C and D.

4.2 Main results

Performance under Clean and Noisy Regimes. Table 1 summarizes the 5-way 5-shot classification performance on *mini*ImageNet and *tiered*ImageNet. The proposed DDGE framework demonstrates consistent improvements across the evaluated noise ratios. Under the noise-free (0%) setting, our method equipped with Swin-T achieves an accuracy of 93.16% on *mini*ImageNet and 93.21% on *tiered*ImageNet. This indicates that evaluating distances over a density-aware continuous landscape provides an effective metric space, even without explicit topological contamination.

DDGE's advantages become more pronounced under heavy noise. As the OOD noise ratio escalates to 60%, deterministic defenses experience notable performance drops. For instance, MetaDiff [24] and LDC [21] decrease to 52.15% and 57.45% on *mini*ImageNet, respectively. In contrast, our method maintains a 70.96% accuracy under the same 60% contamination. This performance gap supports the efficacy of the multi-view conformal integral, which structurally isolates noise and penalizes paths that traverse low-density regions, thereby preserving metric reliability.

Cross-Architecture Robustness. A prevalent challenge in robust FSL is the potential entanglement of performance gains with the specific inductive biases of the chosen backbone. To verify that our improvements originate from the manifold-aware metric rather than backbone-specific features, we evaluate DDGE across a spectrum of eight distinct architectures. This includes classical CNNs (e.g., ResNet-12 and ResNet-50 [37]), modern architectures like ConvNeXt [38], Vision Transformers

Table 2: Backbone generalization across *mini*ImageNet and *tiered*ImageNet. All results are evaluated using the DDGE metric under varying OOD noise ratios. Accuracy (%) is reported with 95% confidence intervals as subscripts.

Backbone	*mini*ImageNet (5-shot)				*tiered*ImageNet (5-shot)			
	0%	20%	40%	60%	0%	20%	40%	60%
ResNet-12 [37]	$76.73_{\pm 0.64}$	$70.56_{\pm 0.67}$	$63.00_{\pm 0.80}$	$53.04_{\pm 0.85}$	$82.31_{\pm 0.67}$	$77.63_{\pm 0.79}$	$70.86_{\pm 0.90}$	$59.97_{\pm 0.92}$
ResNet-50 [37]	$88.14_{\pm 0.33}$	$83.84_{\pm 0.42}$	$75.85_{\pm 0.62}$	$62.91_{\pm 0.82}$	$89.46_{\pm 0.28}$	$85.39_{\pm 0.40}$	$78.00_{\pm 0.66}$	$63.35_{\pm 0.86}$
ConvNeXt [38]	$91.86_{\pm 0.27}$	$90.16_{\pm 0.33}$	$83.46_{\pm 0.41}$	$70.67_{\pm 0.74}$	$92.48_{\pm 0.26}$	$90.85_{\pm 0.29}$	$84.74_{\pm 0.40}$	$71.08_{\pm 0.71}$
Swin-T [36]	$\mathbf{93.16}_{\pm 0.23}$	$90.12_{\pm 0.33}$	$83.89_{\pm 0.54}$	$70.96_{\pm 0.80}$	$93.21_{\pm 0.21}$	$91.06_{\pm 0.27}$	$85.02_{\pm 0.51}$	$72.26_{\pm 0.76}$
DeiT [39]	$88.64_{\pm 0.29}$	$84.99_{\pm 0.36}$	$76.42_{\pm 0.42}$	$63.09_{\pm 0.76}$	$89.93_{\pm 0.24}$	$86.34_{\pm 0.27}$	$79.48_{\pm 0.42}$	$65.23_{\pm 0.71}$
DINOv1 [40]	$89.28_{\pm 0.35}$	$84.74_{\pm 0.53}$	$77.93_{\pm 0.69}$	$63.36_{\pm 0.88}$	$91.10_{\pm 0.43}$	$86.98_{\pm 0.60}$	$80.25_{\pm 0.71}$	$67.51_{\pm 0.94}$
DINOv3 [14]	$93.04_{\pm 0.23}$	$89.71_{\pm 0.31}$	$83.44_{\pm 0.50}$	$71.78_{\pm 0.79}$	$\mathbf{93.38}_{\pm 0.33}$	$91.21_{\pm 0.39}$	$85.10_{\pm 0.62}$	$73.35_{\pm 0.85}$
CLIP [13]	$91.57_{\pm 0.29}$	$\mathbf{90.57}_{\pm 0.45}$	$\mathbf{84.34}_{\pm 0.69}$	$\mathbf{71.94}_{\pm 0.88}$	$92.98_{\pm 0.35}$	$\mathbf{91.26}_{\pm 0.52}$	$\mathbf{86.84}_{\pm 0.74}$	$\mathbf{74.41}_{\pm 0.94}$

(e.g., Swin-T [36], DeiT [39], and the DINO series [40, 14]), and vision-language models (CLIP [13]) (Table 2).

The empirical results indicate consistent robustness across all evaluated architectures. While lightweight models (e.g., ResNet-12) naturally exhibit a lower performance ceiling due to limited capacity, their relative degradation under extreme noise is consistently mitigated. Notably, under the challenging 60% noise regime, modern architectures (e.g., ConvNeXt, Swin-T, CLIP [13]) sustain an absolute accuracy above 70% on both datasets. Given that conventional Euclidean metrics typically experience substantial geometric collapse under such contamination, these results suggest that our density-aware geodesic evaluation provides generalizable robustness, largely decoupling metric performance from specific feature extraction biases.

4.3 Comprehensive ablation study

To analyze the contributions of our proposed components, we conduct an ablation study focusing on semantic disentanglement (S), the Dirichlet Process (D), and the geodesic metric (G). Table 3 details the quantitative performance across varying noise regimes.

Table 3: Targeted component ablation on *mini*ImageNet (5-shot, Swin-T). **S**: Semantic Disentanglement, **D**: Dirichlet Process, **G**: Geodesic Metric. Accuracy (%) is evaluated across varying OOD noise ratios.

Config	**S**	**D**	**G**	0% Noise	20% Noise	40% Noise	60% Noise
1 (Baseline)				$84.08_{\pm 0.35}$	$80.38_{\pm 0.45}$	$69.03_{\pm 0.63}$	$53.90_{\pm 0.85}$
2	✓			$89.56_{\pm 0.21}$	$85.57_{\pm 0.31}$	$76.82_{\pm 0.54}$	$62.79_{\pm 0.81}$
3		✓		$86.08_{\pm 0.31}$	$82.55_{\pm 0.41}$	$73.20_{\pm 0.66}$	$59.45_{\pm 0.85}$
4 (w/o G)	✓	✓		$91.35_{\pm 0.23}$	$87.35_{\pm 0.31}$	$78.51_{\pm 0.52}$	$65.23_{\pm 0.82}$
5 (w/o S)		✓	✓	$88.19_{\pm 0.28}$	$84.16_{\pm 0.46}$	$75.68_{\pm 0.66}$	$61.73_{\pm 0.93}$
6 (Full)	✓	✓	✓	$\mathbf{93.16}_{\pm 0.23}$	$\mathbf{90.12}_{\pm 0.33}$	$\mathbf{83.89}_{\pm 0.54}$	$\mathbf{70.96}_{\pm 0.80}$

Effectiveness of semantic disentanglement (S). Comparing Config 5 (without S) to the full model (Config 6) highlights the importance of subspace projection [26]. Without disentanglement, constructing the covariance matrix directly in the global ambient space leads to severe rank deficiency. Consequently, under 60% noise, accuracy drops significantly from 70.96% to 61.73%. This indicates that without decoupled isolation, adversarial noise can corrupt the global topology and bypass the subsequent geodesic evaluation.

Impact of the Dirichlet Process (D). The Dirichlet Process [12] provides the necessary geometric volume for path evaluation. Without this expansion (Config 2, S only), sparse support samples lack a continuous density field. Quantitatively, incorporating D (Config 4 vs. Config 2) establishes a stable density foundation, yielding consistent accuracy improvements across all noise regimes (e.g., from 76.82% to 78.51% under 40% noise) by mitigating the structural sparsity of the support set.

Role of the geodesic metric (G). While Config 4 successfully constructs a GMM terrain, it ultimately relies on the globally flat Euclidean distance for classification [3]. This linear path can incorrectly

indicate proximity by evaluating absolute displacement across low-density topological voids. Substituting it with the geodesic metric [27] (Config 6) allows the framework to dynamically evaluate path resistance, imposing a penalty on geometric anomalies. This effectively raises the accuracy from 78.51% to 83.89% under the 40% noise setting. The integration of S, D, and G ensures that topological voids are consistently exposed and penalized.

4.4 Robustness in real-world and systematic contamination

While synthetic out-of-distribution noise evaluates metric robustness under controlled settings, practical deployment frequently encounters natural topological distortions and domain shifts [16, 23]. We extend our evaluation to these real-world scenarios to validate the framework's broader applicability, benchmarking our DDGE against the state-of-the-art robust method, ECER-FSL [20].

Generalization to Cross-Domain and Natural Noise. Table 4 and Table 5 present DDGE's performance across cross-domain environments and datasets characterized by human-level annotation errors, respectively.

In cross-domain scenarios (Table 4) evaluated on datasets such as FC100 [31] and CIFAR-FS [32], the model maintains robust geometric transferability even under heavy artificial noise. Compared to the robust baseline ECER-FSL, DDGE exhibits a significant performance margin across all noise ratios. For instance, on CIFAR-FS under 40% noise, DDGE achieves 75.27%, outperforming ECER-FSL by a clear margin. This indicates that our prior-borrowing mechanism effectively transfers intrinsic geometric structures (covariance matrices) rather than overfitting to the visual patterns of the base domain. Furthermore, scaling from 5-shot to 10-shot consistently yields accuracy gains, indicating that denser support sets refine the Gaussian density landscape and enhance the precision of the conformal integral.

Simultaneously, Table 5 isolates the evaluation on real-world datasets (CIFAR-100N [33], Animal-10N [34], and Food-101N [35]) where class boundaries are inherently ambiguous due to organic label noise. Without injecting artificial contamination, deterministic baseline methods often struggle with this natural ambiguity. However, the soft calibration in our Dirichlet Process effectively accommodates it, achieving an impressive 5-shot accuracy of 84.27% on Food-101N and consistently outperforming the baseline across all unconstrained settings.

Table 4: Cross-domain FSL evaluated under artificial OOD noise. DDGE is compared with the robust baseline (ECER-FSL) to demonstrate transferability.

Dataset	Method	5-shot				10-shot			
		0%	20%	40%	60%	0%	20%	40%	60%
FC100	ECER-FSL	56.75	53.60	48.93	37.64	59.35	57.07	53.76	44.89
	Ours	**63.67**	**61.70**	**58.45**	**49.29**	**66.54**	**65.41**	**63.69**	**58.88**
CIFAR-FS	ECER-FSL	85.54	78.88	67.69	57.80	87.94	81.66	72.92	63.74
	Ours	**87.99**	**82.81**	**75.27**	**66.31**	**89.17**	**86.03**	**80.30**	**73.92**

Table 5: Real-world datasets with natural label noise (0% artificial noise).

Dataset	Method	Natural Noise	
		5-shot	10-shot
C-100N	ECER-FSL	62.25	64.14
	Ours	**63.58**	**65.41**
Ani-10N	ECER-FSL	76.87	79.20
	Ours	**82.58**	**84.37**
Food-101N	ECER-FSL	75.98	77.53
	Ours	**84.27**	**86.95**

Robust Defense Against Systematic (SYS) Noise. Beyond open-set anomalies, closed-set Systematic (SYS) noise [41] poses a fundamentally distinct challenge. By mislabeling genuine samples from valid classes, SYS noise forms deceptive, dense competing manifolds that exacerbate prototype distortion [3]. Table 6 contrasts the performance across various backbones under identical ratios of OOD and SYS noise. Although systematic noise inherently presents a more complex challenge, DDGE exhibits consistent robustness. For instance, with the Swin-T backbone at a 40% noise ratio, the model achieves a robust accuracy of 77.22% (compared to 83.89% under OOD noise). This defense is theoretically grounded in the decoupled semantic isolation (**S**) [26]. Even if a closed-set noise sample visually mimics a target class in one specific subspace, it mismatches the topology across orthogonal views. This deviation triggers the geodesic penalty (**G**), neutralizing camouflaged noise without relying on heuristic thresholds and preventing substantial geometric distortion.

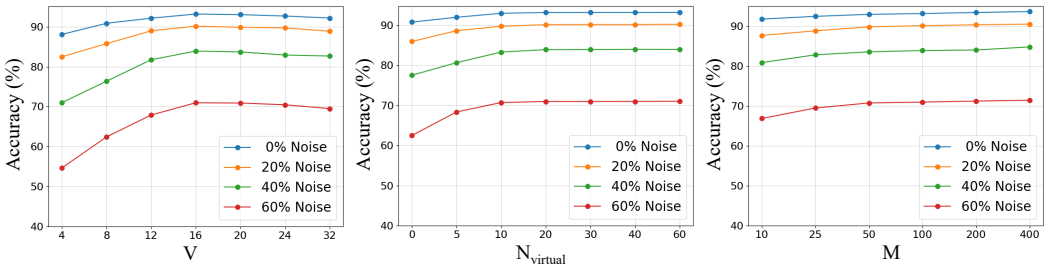
Table 6: Classification accuracy (%) of various backbones under different noise topologies on *mini*ImageNet (5-shot). The model exhibits robust resilience against both open-set (OOD) and closed-set (SYS) noise across varying contamination ratios.

Backbone	OOD Noise				SYS Noise			
	0%	20%	40%	60%	0%	20%	40%	60%
ResNet-12 [37]	$76.73_{\pm 0.64}$	$70.56_{\pm 0.67}$	$63.00_{\pm 0.80}$	$53.04_{\pm 0.85}$	$76.73_{\pm 0.64}$	$69.65_{\pm 0.73}$	$57.72_{\pm 0.89}$	$40.80_{\pm 0.99}$
ResNet-50 [37]	$88.14_{\pm 0.33}$	$83.84_{\pm 0.42}$	$75.85_{\pm 0.62}$	$62.91_{\pm 0.82}$	$88.14_{\pm 0.33}$	$80.86_{\pm 0.58}$	$70.37_{\pm 0.70}$	$55.49_{\pm 0.89}$
ConvNeXt [38]	$91.86_{\pm 0.27}$	$90.16_{\pm 0.33}$	$83.46_{\pm 0.41}$	$70.67_{\pm 0.74}$	$91.86_{\pm 0.27}$	$86.35_{\pm 0.58}$	$77.40_{\pm 0.75}$	$63.24_{\pm 0.91}$
Swin-T [36]	$\mathbf{93.16}_{\pm 0.23}$	$90.12_{\pm 0.33}$	$83.89_{\pm 0.54}$	$70.96_{\pm 0.80}$	$\mathbf{93.16}_{\pm 0.23}$	$87.00_{\pm 0.56}$	$77.22_{\pm 0.75}$	$65.44_{\pm 0.83}$
DeiT [39]	$88.64_{\pm 0.29}$	$84.99_{\pm 0.36}$	$76.42_{\pm 0.42}$	$63.09_{\pm 0.76}$	$88.64_{\pm 0.29}$	$82.46_{\pm 0.57}$	$73.27_{\pm 0.76}$	$57.34_{\pm 0.89}$
DINOv1 [40]	$89.28_{\pm 0.35}$	$84.74_{\pm 0.53}$	$77.93_{\pm 0.69}$	$63.36_{\pm 0.88}$	$89.28_{\pm 0.35}$	$84.86_{\pm 0.61}$	$75.05_{\pm 0.72}$	$59.19_{\pm 0.86}$
DINOv3 [14]	$93.04_{\pm 0.23}$	$89.71_{\pm 0.31}$	$83.44_{\pm 0.50}$	$71.78_{\pm 0.79}$	$93.04_{\pm 0.23}$	$88.11_{\pm 0.70}$	$80.22_{\pm 0.74}$	$66.75_{\pm 0.84}$
CLIP [13]	$91.57_{\pm 0.29}$	$\mathbf{90.57}_{\pm 0.45}$	$\mathbf{84.34}_{\pm 0.69}$	$\mathbf{71.94}_{\pm 0.88}$	$91.57_{\pm 0.29}$	$\mathbf{90.26}_{\pm 0.56}$	$\mathbf{85.75}_{\pm 0.66}$	$\mathbf{69.01}_{\pm 0.73}$

4.5 Boundary analysis and hyperparameter dynamics

Stress Testing at 80% Contamination. To explore the framework's physical boundaries, we push the topological contamination to an extreme 80% ratio (i.e., 1 genuine sample vs. 4 anomalies). Under this highly contaminated regime, the LOO density scoring risks inadvertently identifying the denser noise cluster as the genuine core. For unstructured OOD noise, the model's accuracy on *mini*ImageNet drops to 56.01%, yet it still outperforms the Euclidean baseline (40.37%) by a 15.64% margin, indicating that orthogonal disentanglement (**S**) still isolates scattered noise to some extent. Conversely, under 80% SYS noise, where label-swapped anomalies form a highly cohesive false manifold, the model experiences significant performance degradation (39.50%). This establishes a theoretical boundary: the geometric veto mechanism requires genuine samples to hold at least a minimal relative density advantage to initialize the correct manifold contours.

Structural Hyperparameter Dynamics. As introduced in Section 4.1, we analyze three core structural hyperparameters: semantic views (V), expansion size (N_{virtual}), and probing steps (M), which respectively govern the semantic disentanglement granularity, manifold smoothness, and geodesic approximation precision (Figure 2).



(a) Semantic views (V) (b) Expansion size (N_{virtual}) (c) Probing steps (M)

Figure 2: Dynamics of structural hyperparameters. **(a)** Semantic views V balance noise isolation against feature fragmentation. **(b)** Expansion size N_{virtual} rapidly injects geometric volume before safe saturation. **(c)** Probing steps M reveal a classic accuracy-latency trade-off.

Specifically, the view count V exhibits an inverted U-shape: a small V fails to isolate noise structurally, while an excessively large V over-fragments the semantic space. The expansion size N_{virtual} rapidly provides geometric volume and saturates safely, avoiding over-expansion that might amplify base-class estimation errors. Finally, the probing steps M exhibit diminishing returns regarding computational cost, with accuracy optimally plateauing around $M = 100$, where the discrete Riemann summation effectively converges to the continuous geodesic resistance [27]. This analysis supports our empirical configuration ($V = 16, N_{\text{virtual}} = 20, M = 100$) for fast and robust adaptation.

5 Conclusion

We presented DDGE, a robust few-shot learning framework that overcomes the limitations of globally flat linear metrics via a density-aware path evaluation paradigm. By integrating orthogonal semantic

disentanglement with a prior-guided Dirichlet Process, DDGE constructs a continuous, noise-purified Gaussian density landscape. Upon this geometry, a multi-view conformal integral dynamically evaluates path resistance, effectively neutralizing topological noise.

Limitations. While DDGE establishes a robust metric landscape, accurately recovering intrinsic class manifolds under extreme topological contamination—where anomalies heavily dominate the target distribution—remains a fundamental open challenge for current representation learning paradigms.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62606327, 62471266, 62006131, 62071260; the National Natural Science Foundation of Zhejiang Province under Grant LQ24F020001; and the Ningbo Major Research and Development Plan Project (No.2023Z224); and the Intelligent Computing Center of Shenzhen University.

References

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

[2] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[3] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[4] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[5] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[6] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[7] Duong Le, Khoi Duc Nguyen, Khoi Nguyen, Quoc-Huy Tran, Rang Nguyen, and Binh-Son Hua. Poodle: Improving few-shot learning via penalizing out-of-distribution samples. *Advances in Neural Information Processing Systems*, 34:23942–23955, 2021.

[8] Ji Zhang, Jingkuan Song, Lianli Gao, Nicu Sebe, and Heng Tao Shen. Reliable few-shot learning under dual noises. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[9] Xu Luo, Longhui Wei, Liangjian Wen, Jinrong Yang, Lingxi Xie, Zenglin Xu, and Qi Tian. Rectifying the shortcut learning of background for few-shot learning. *Advances in Neural Information Processing Systems*, 34:13073–13085, 2021.

[10] Ji Zhang, Lianli Gao, Xu Luo, Hengtao Shen, and Jingkuan Song. Deta: Denoised task adaptation for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11541–11551, 2023.

[11] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. Adaptive cross-modal few-shot learning. *Advances in neural information processing systems*, 32, 2019.

[12] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[14] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

[16] Wei-Hong Li, Xialei Liu, and Hakan Bilen. Cross-domain few-shot learning with task-specific adapters. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7161–7170, 2022.

[17] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Differentiable earth mover's distance for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5632–5648, 2022.

[18] Shanfeng Wang, Jianzhao Li, Zaitian Liu, Yourun Zhang, and Maoguo Gong. Fedfsl-cfrd: Personalized federated few-shot learning with collaborative feature representation disentanglement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21243–21251, 2025.

[19] Hai Zhang, Junzhe Xu, Shanlin Jiang, and Zhenan He. Simple semantic-aided few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28588–28597, 2024.

[20] Mushui Liu, Fangtai Wu, Bozheng Li, Ziqian Lu, Yunlong Yu, and Xi Li. Envisioning class entity reasoning by large language models for few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18906–18914, 2025.

[21] Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25411–25421, 2025.

[22] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in-and out-distribution improves explainability. In *European Conference on Computer Vision*, pages 228–245. Springer, 2020.

[23] Ji Zhang, Jingkuan Song, Lianli Gao, and Hengtao Shen. Free-lunch for cross-domain few-shot learning: Style-aware episodic training with robust contrastive learning. In *Proceedings of the 30th ACM international conference on multimedia*, pages 2586–2594, 2022.

[24] Baoquan Zhang, Chuyao Luo, Demin Yu, Xutao Li, Huiwei Lin, Yunming Ye, and Bowen Zhang. Metadiff: Meta-learning with conditional diffusion for few-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16687–16695, 2024.

[25] Stanislav Fort. Gaussian prototypical networks for few-shot learning on omniglot. *arXiv preprint arXiv:1708.02735*, 2017.

[26] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4136–4145, 2020.

[27] Ron Kimmel and James A Sethian. Computing geodesic paths on manifolds. *Proceedings of the national academy of Sciences*, 95(15):8431–8435, 1998.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[29] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[31] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31, 2018.

[32] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *arXiv preprint arXiv:1805.08136*, 2018.

[33] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

[34] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pages 5907–5915. PMLR, 2019.

[35] Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2018. doi: 10.1109/CVPR.2018.00571.

[36] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

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

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

[39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[40] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[41] Kevin J Liang, Samrudhdhi B Rangrej, Vladan Petrovic, and Tal Hassner. Few-shot learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9089–9098, 2022.

A Detailed Mathematical Formulations and Geometric Insights

In Section 3 of the main text, we introduced the Disentangled Dirichlet Geodesic Evaluation (DDGE) framework. Due to space constraints, specific mathematical implementations, numerical stabilization techniques, and approximation error bounds were abstracted. In this section, we provide the comprehensive mathematical derivations corresponding to the methodology, ensuring the theoretical reproducibility of our manifold-aware metric.

A.1 Physical Interpretations of Orthogonal Repulsion

In Section 3.2, we utilize a semantic disentanglement module to map the global features $f \in \mathbb{R}^D$ into V parallel sub-spaces of dimension d. The orthogonality constraint $\mathcal{L}_{\text{ortho}}$ serves as a structural regularizer that forces the latent space to adopt a block-diagonal covariance prior.

To formalize the optimization dynamics, the orthogonal loss is defined as:

$$\mathcal{L}_{\text{ortho}} = \sum_{i=1}^{V-1} \sum_{j=i+1}^{V} \|W_i W_j^\top\|_F^2 \tag{8}$$

By taking the partial derivative of $\mathcal{L}_{\text{ortho}}$ with respect to a specific projection matrix W_k, we can derive the gradient formulation:

$$\nabla_{W_k} \mathcal{L}_{\text{ortho}} = \frac{\partial}{\partial W_k} \sum_{j \neq k} \text{Tr}(W_k W_j^\top W_j W_k^\top) \tag{9}$$

$$\nabla_{W_k} \mathcal{L}_{\text{ortho}} = 4 \sum_{j \neq k} (W_k W_j^\top) W_j \tag{10}$$

In this derivation, the matrix product $(W_k W_j^\top)$ functions as a semantic coupling coefficient. During the base training phase, when two views encode redundant visual information, the gradient magnitude increases, dynamically pushing the row vectors of W_k toward the null space of W_j. This formulation minimizes the mutual information between sub-spaces, structurally isolating local topological noise into specific views without compromising the global metric space.

A.2 Numerical Robustness and Log-Density Estimation

In high-dimensional manifolds (e.g., $d = 128$), the direct evaluation of the multivariate Gaussian density is prone to significant numerical underflow. Because the Mahalanobis distance scales with dimensionality, linear likelihood values frequently fall below the machine epsilon, leading to computational instability during density estimation.

To execute the Leave-One-Out (LOO) density calibration robustly, we first ensure the strict positive-definiteness of the borrowed base-class priors by introducing a minimal Tikhonov regularization:

$$\Sigma_{\text{stable},i}^{(v)} = \sum_{b=1}^{N_{\text{base}}} A_{i,b}^{(v)} \Sigma_b^{(v)} + \epsilon I \tag{11}$$

where $\epsilon = 10^{-6}$.

Subsequently, we operate strictly within the logarithmic domain. To maintain mathematical clarity, we isolate the precision-weighted quadratic form $\mathcal{Q}_{ij}$ for any two intra-class samples $x_i^{(v)}$ and $x_j^{(v)}$:

$$\mathcal{Q}_{ij} = (x_i^{(v)} - x_j^{(v)})^\top (\Sigma_{\text{stable},j}^{(v)})^{-1} (x_i^{(v)} - x_j^{(v)}) \tag{12}$$

The exact log-likelihood $\log \mathcal{N}_{ij}$ is then computed without exponentiation:

$$\log \mathcal{N}_{ij} = -\frac{1}{2} \left[d \log(2\pi) + \log |\Sigma_{\text{stable},j}^{(v)}| + \mathcal{Q}_{ij} \right] \tag{13}$$

To compute the expected local density score $s_i^{(v)}$ and the generative quota $w_i^{(v)}$, we deploy the Log-Sum-Exp (LSE) stabilization technique. The normalized logarithmic quota is formulated as:

$$\log w_i^{(v)} = \frac{1}{\phi}\left(\log\frac{1}{K-1} + \text{LSE}_{j\neq i}[\log \mathcal{N}_{ij}]\right) - \log Z \tag{14}$$

$$\log Z = \text{LSE}_{k=1}^{K}\left[\frac{1}{\phi}\left(\log\frac{1}{K-1} + \text{LSE}_{j\neq k}[\log \mathcal{N}_{kj}]\right)\right] \tag{15}$$

where $\text{LSE}(\cdot) = \log\sum\exp(\cdot)$. This numerical stabilization prevents precision errors from causing a disproportionate concentration of generative quotas, ensuring that soft-calibration remains differentiable and stable even when samples are located in low-density ambient regions.

A.3 Dirichlet Proliferation and Quota Allocation

To actualize the manifold expansion via the Dirichlet Process, the continuous routing weights $w_i^{(v)} = \exp(\log w_i^{(v)})$ are translated into an integer allocation of virtual samples $q_i^{(v)}$ for each support point $x_i^{(v)}$.

This allocation is defined through an empirical discretization:

$$q_i^{(v)} = \max\left(1, \lfloor N_{\text{virtual}} \cdot w_i^{(v)} \rceil\right) \tag{16}$$

Let $S_{\text{total}}^{(v)}$ denote the expanded set comprising both the original K intra-class support samples and the hallucinated virtual samples generated via the standard Cholesky reparameterization $x_{\text{virtual}}^{(v,m)} = x_i^{(v)} + L_i^{(v)} z^{(m)}$. The cardinality of this superset is:

$$M_c^{(v)} = K + \sum_{i=1}^{K} q_i^{(v)} \tag{17}$$

The final noise-purified Gaussian Mixture Model (GMM) density field is computed as an unweighted arithmetic average over $S_{\text{total}}^{(v)}$, effectively augmenting the geometric volume around authentic semantic clusters:

$$P_{\text{GMM}}^{(v)}(x) = \frac{1}{M_c^{(v)}} \sum_{m=1}^{M_c^{(v)}} \mathcal{N}\left(x \;\middle|\; \mu_m^{(v)}, \Sigma_m^{(v)}\right) \tag{18}$$

A.4 Conformal Transformation and Geodesic Approximation

The calculation of the multi-view geodesic metric (Section 3.5) assumes the latent space acts as a Riemannian manifold. We define its intrinsic geometry using an isotropic conformal metric tensor $g_{ij}(x)$, scaled by the resistance function $R^{(v)}(x)$:

$$g_{ij}(x) = [R^{(v)}(x)]^2 \delta_{ij} = \left[1 + \alpha \exp(-\beta P_{\text{GMM}}^{(v)}(x))\right]^2 \delta_{ij} \tag{19}$$

Under this conformal metric, the generalized path length is:

$$L(\gamma) = \int_0^1 \sqrt{g_{ij}(\gamma(t))\dot{\gamma}^i(t)\dot{\gamma}^j(t)}\, dt = \int_0^1 R^{(v)}(\gamma(t))\|\dot{\gamma}(t)\|_2\, dt \tag{20}$$

Solving the exact geodesic trajectory via the non-linear Eikonal equation entails iterative optimization, which contradicts the fast-adaptation requirement of FSL. Thus, we constrain the probe trajectory $\gamma^{(v)}(t)$ to the straight Euclidean line for computational feasibility; consequently, the computed conformal line integral serves as an upper-bound approximation to the geodesic distance, rather than the shortest geodesic path itself. This approximation is practically informative because $R^{(v)}(x)$ grows exponentially in low-density topological voids, so any curved path attempting to bypass ridges incurs a substantial length penalty.

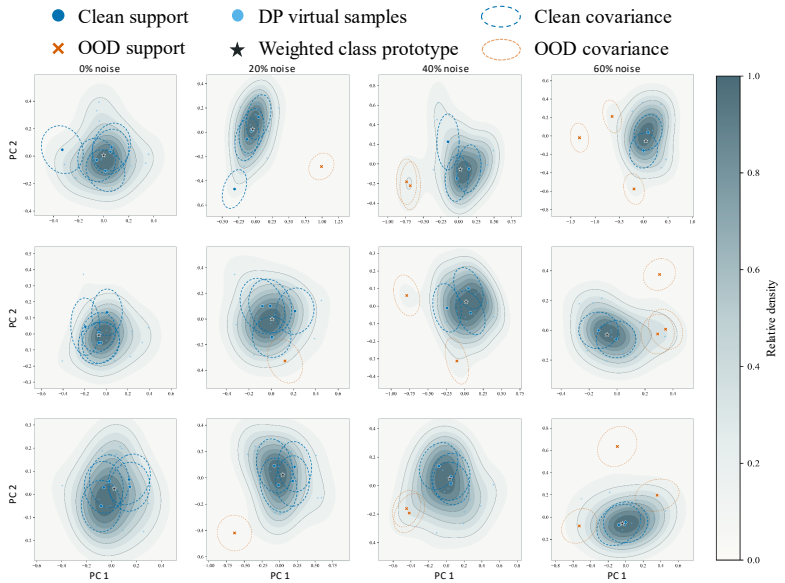


Figure 3: Density landscape visualization of the constructed manifold under OOD noise ratios (0%, 20%, 40%, 60%): clean samples form a concentrated density plateau, while OOD samples remain isolated in low-density valleys.

We evaluate this integral using a discrete Riemann summation over M uniform probing steps ($t_m = m/M$):

$$D^{(v)}_{\text{conformal}}(q, c) \approx \left\| c^{(v)} - q^{(v)} \right\|_2 \cdot \frac{1}{M} \sum_{m=1}^{M} R^{(v)} \left(q^{(v)} + t_m (c^{(v)} - q^{(v)}) \right) \tag{21}$$

Given that the GMM density field $P^{(v)}_{\text{GMM}}(x)$ is a summation of analytic exponential functions, it is infinitely differentiable (C^∞). The truncation error of this numerical integration is strictly bounded by $\mathcal{O}(1/M)$. By configuring $M = 100$, the discrete summation converges to the continuous conformal length with negligible numerical error, providing an effective upper-bound approximation of the geodesic distance.

A.5 Density Landscape Visualization

To verify that the estimated density landscape remains discriminative under noisy few-shot settings, Fig. 3 visualizes the constructed manifolds across three representative 5-way 5-shot evaluation tasks (top to bottom) under 0%, 20%, 40%, and 60% OOD contamination ratios (left to right). Each subplot depicts the 2D principal component (PC 1–PC 2) density contour map of a target novel class. Across all contamination regimes, Dirichlet Process (DP) virtual samples selectively proliferate around genuine clean supports, forming a concentrated Gaussian-mixture density plateau (the manifold core) centered around the weighted class prototype. Conversely, OOD outliers are assigned minimal generative quotas and remain isolated in low-density ambient regions. Even as increasing the noise ratio from 0% to 60% gradually weakens the clean manifold, OOD samples are never absorbed into its high-density core.

B Algorithm Summary and Pseudocode

Algorithm 1 summarizes the inference pipeline of the proposed Disentangled Dirichlet Geodesic Evaluation (DDGE) during the meta-testing phase.

The execution is structurally divided into two phases:

1. **Phase I: Manifold Construction.** The pipeline transforms sparse, discrete support features into a continuous probability density field $P^{(v)}_{\mathrm{GMM}}(x)$. This phase utilizes the prior-guided Dirichlet Process to explicitly parameterize the geometric volume of the novel class.
2. **Phase II: Density-Aware Path Integration.** The algorithm evaluates the intrinsic distance by measuring the cumulative resistance along the Euclidean straight-line probe connecting the query to the class prototype. Trajectories traversing low-density ambient regions incur an exponential penalty, thereby mitigating the impact of topological voids.

Algorithm 1: Disentangled Dirichlet Geodesic Evaluation (Meta-testing)

Input : Support set $\tilde{S} = \{(x_i, y_i)\}$, Query set $Q = \{q_j\}$;
Pre-trained backbone Φ, Disentanglement matrices $\{W_v\}_{v=1}^{V}$, Prior bank $\mathcal{B}^{(v)}$.
Output : Predicted labels $\{\hat{y}_q\}$ for all $q \in Q$.

`// Phase I: Manifold Construction (Class-wise Initialization)`
Extract global features: $f_i = \Phi(x_i), f_q = \Phi(q)$;
foreach *class* $c \in$ *Unique*$(\{y_i\})$ **do**
  foreach *view* $v \in \{1, \ldots, V\}$ **do**
    Project features: $x^{(v)}_{i,c} = W_v f_i, \forall i$ where $y_i = c$;
    Retrieve $\Sigma^{(v)}_{\mathrm{stable},i}$ from $\mathcal{B}^{(v)}$ via attention (Eq. 2);
    Calibrate generative quotas $w^{(v)}_i$ via LOO density estimation (Eq. 3-4);
    Construct GMM density field $P^{(v)}_{\mathrm{GMM},c}(x)$ via Dirichlet reparameterization (Eq. 5);
    Compute robust prototype: $c^{(v)} = \sum w^{(v)}_i x^{(v)}_{i,c}$;
  end
end

`// Phase II: Density-Aware Path Integration (Query Evaluation)`
foreach *query sample* $q \in Q$ **do**
  foreach *candidate class* $c \in \mathcal{C}$ **do**
    Initialize total distance $D_{\mathrm{total}}(q, c) \leftarrow 0$;
    foreach *view* $v \in \{1, \ldots, V\}$ **do**
      Define parameterized linear path: $\gamma^{(v)}(t) = q^{(v)} + t(c^{(v)} - q^{(v)})$, for $t \in [0, 1]$;
      Compute path resistance at M Riemann probing steps:
      $R^{(v)}(t_m) = 1 + \alpha \exp\left(-\beta \cdot P^{(v)}_{\mathrm{GMM},c}(\gamma^{(v)}(t_m))\right)$ (Eq. 6);
      Approximate conformal integral: $D^{(v)}_{\mathrm{conformal}} = \left\|c^{(v)} - q^{(v)}\right\|_2 \cdot \frac{1}{M} \sum_{m=1}^{M} R^{(v)}(t_m)$ (Eq. 7);
      Accumulate global conformal distance: $D_{\mathrm{total}}(q, c) \leftarrow D_{\mathrm{total}}(q, c) + D^{(v)}_{\mathrm{conformal}}$;
    end
  end
  Prediction: $\hat{y}_q = \arg\min_c D_{\mathrm{total}}(q, c)$;
end

C Computational Complexity and Mathematical Tractability

To provide a clear understanding of the computational overhead and theoretical basis of the proposed Disentangled Dirichlet Geodesic Evaluation (DDGE), we analyze its structural design and empirical latency compared to baseline approaches.

Computational Complexity. As DDGE operates in a fully gradient-free manner during meta-testing, it avoids the intensive backpropagation overhead required by optimization-based meta-learning methods. We focus our evaluation on the empirical inference latency per image, as summarized in Table 7.

Theoretically, constructing a continuous Gaussian density landscape from an ambient dimension D requires a covariance decomposition that scales with $\mathcal{O}(D^3)$. By projecting the global features

Table 7: Inference latency comparison. The empirical latency is measured in milliseconds per image (ms/img) on a single NVIDIA RTX 4090 GPU.

Method	**Probing Steps (M)**	**Infer. Latency (ms/img)**
Baseline (Euclidean)	-	1.54
Baseline (Global GMM)	-	1.59
DDGE (Ours)	100	3.33

into V orthogonal sub-spaces of dimension d (Phase I), DDGE bounds this matrix decomposition complexity to $\mathcal{O}(V \cdot d^3)$.

During Phase II, DDGE calculates the cumulative path resistance over M discrete probing steps. The theoretical complexity for this integration scales with $\mathcal{O}(V \cdot M \cdot |S_{\text{total}}| \cdot d)$. However, as shown in Table 7, while DDGE explicitly executes 100 discrete probing steps, the inference latency (3.33 ms/img) does not scale linearly by a factor of 100 compared to the Euclidean baseline (1.54 ms/img). The absolute computational overhead is marginal (approximately 1.79 ms/img). This practical efficiency is achieved because all density estimations are executed within the logarithmic domain using vectorizable Log-Sum-Exp operators. Since the resistance evaluations across the V views and M probing steps are structurally independent, the conformal integral is highly parallelized via tensor broadcasting on GPUs, effectively isolating the impact of topological noise with minimal latency sacrifice.

Mathematical Tractability. Beyond computational efficiency, the subspace projection mathematically ensures the validity of the borrowed geometric priors. In standard FSL benchmarks, the number of training samples per base class, denoted as N_{base}, is frequently smaller than the ambient dimension D (e.g., $N_{\text{base}} \approx 600 < 1024$ in *mini*ImageNet). Directly estimating a $D \times D$ covariance matrix from these samples leads to fundamental rank deficiency. Under such conditions, standard regularization techniques inject isotropic noise into the null space, which distorts the intrinsic geometric representation.

By projecting the global features into d-dimensional sub-spaces where the sample size strictly dominates the dimensionality ($N_{\text{base}} \gg d$), we ensure that the estimated base covariance matrices are full-rank and well-conditioned. This allows the soft-calibration mechanism to formulate valid geometric volumes in all projected directions, addressing the rank deficiency issue without inflating null spaces with isotropic noise.

D Functional Hyperparameter Sensitivity

Beyond the structural dimensions evaluated in the main text (e.g., semantic views V, expansion size N_{virtual}, and integration steps M), the DDGE framework utilizes hyperparameters to govern subspace disentanglement, regularize the density estimation in Phase I, and control the path resistance in Phase II. To assess the model's sensitivity and provide a comprehensive operational guide, we evaluate performance across a spectrum of parameter values on *mini*ImageNet under a 40% OOD noise ratio. This specific noise level is selected because the performance variations induced by noise-rejection parameters are more observable under moderate contamination compared to clean (0%) settings.

Path Resistance Parameters (α and β): In the conformal integral (Phase II), α dictates the maximum absolute penalty for traversing low-density regions, while β determines the exponential rate of resistance decay. As demonstrated in Table 8, a minimal margin scale ($\alpha = 1.0$) provides insufficient contrast between high-density ridges and topological voids, reducing the accuracy to 82.98%. On the other hand, an overly sensitive exponential decay (e.g., $\beta \geq 15.0$) abruptly truncates the manifold boundaries, which inadvertently penalizes genuine hard-positive queries that reside slightly away from the cluster centers. The framework exhibits a stable performance plateau within intermediate ranges ($\alpha \in [3.0, 7.0]$ and $\beta \in [3.0, 10.0]$), confirming that the multi-view geodesic evaluation relies on relative density contrast rather than exact scalar tuning.

Manifold Construction Temperatures (τ and ϕ): During Phase I, τ calibrates the attention distribution when retrieving base-class covariance priors, and ϕ acts as the temperature for the LOO

Table 8: Sensitivity analysis of hyperparameters on *mini*ImageNet (5-shot, 40% OOD Noise). Accuracy (%) is reported. The default values used in all main experiments are underlined, yielding the reported 83.89% accuracy.

Hyperparameter	**Evaluated Values & Accuracy (%)**				
Resistance Margin Scale (α)	1.0 82.98$_{\pm 0.54}$	3.0 **84.05**$_{\pm 0.49}$	<u>5.0</u> 83.89$_{\pm 0.54}$	7.0 83.57$_{\pm 0.55}$	9.0 82.65$_{\pm 0.56}$
Density Sensitivity (β)	1.0 81.46$_{\pm 0.54}$	3.0 82.44$_{\pm 0.60}$	<u>5.0</u> **83.89**$_{\pm 0.54}$	10.0 83.40$_{\pm 0.52}$	15.0 83.74$_{\pm 0.48}$
Prior Retrieval Temp. (τ)	1.0 82.19$_{\pm 0.56}$	5.0 82.56$_{\pm 0.53}$	<u>10.0</u> **83.89**$_{\pm 0.54}$	20.0 83.32$_{\pm 0.52}$	50.0 82.91$_{\pm 0.55}$
Calibration Smoothness (ϕ)	0.1 80.19$_{\pm 0.63}$	0.5 82.46$_{\pm 0.56}$	<u>1.0</u> 83.89$_{\pm 0.54}$	2.0 **84.04**$_{\pm 0.47}$	5.0 83.92$_{\pm 0.48}$
Orthogonal Loss Weight (λ)	0.01 83.86$_{\pm 0.51}$	<u>0.1</u> **83.89**$_{\pm 0.54}$	1.0 83.73$_{\pm 0.58}$	5.0 82.49$_{\pm 0.57}$	10.0 81.87$_{\pm 0.60}$
Subspace Dimension (d)	32 83.46$_{\pm 0.53}$	64 83.48$_{\pm 0.53}$	<u>128</u> **83.89**$_{\pm 0.54}$	256 83.71$_{\pm 0.53}$	512 79.90$_{\pm 0.71}$

density softmax, smoothing the allocation of generative quotas. Setting ϕ to a minimal low value (e.g., 0.1) results in a heavily skewed quota allocation. Mathematically, this approximates a Dirac delta function where only the single densest support sample undergoes Dirichlet proliferation, substantially reducing the geometric diversity and causing a performance drop to 80.19%. Conversely, utilizing overly large temperature values flattens the probability distributions, leading to over-smoothed geometric volumes that blur inter-class boundaries. Maintaining moderate values balances base-class prior fidelity with novel-class topological diversity.

Orthogonal Loss Weight (λ) and Semantic View Subspace Dimension (d): We further evaluate the orthogonal loss weight λ and the semantic view subspace dimension d on *mini*ImageNet (5-shot, 40% OOD noise, Swin-T). As shown in Table 8, performance remains stable over broad ranges ($\lambda \in [0.01, 1]$ and $d \in [32, 256]$), with the largest drops confined to the extreme values ($\lambda = 10$ and $d = 512$). This is consistent with the observation that over-emphasized orthogonality severs inherent semantic integrity (e.g., decoupling naturally co-occurring features), hurting classification. These empirical results validate that the robustness of DDGE stems from its intrinsic manifold-aware metric design, rather than relying on precise parameter tuning.

Parameter Selection Strategy: It is observable that specific configurations, such as $\alpha = 3.0$ or $\phi = 2.0$, yield marginally higher accuracies (84.05% and 84.04%, respectively) for this exact 40% noise setting. However, we consistently apply $\alpha = 5.0$ and $\phi = 1.0$ as the fixed default parameters across all noise levels and datasets reported in the main text. This standardized configuration is maintained to prevent dataset-specific tuning and ensure evaluation consistency.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: Abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

Justification: We have discussed our limitations in the conclusion section.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

   Justification: There is no theoretical result in this paper.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide the experimental setup in the main paper and detailed implementation settings in the appendix to support reproducibility.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: The current submission does not include an open-source code release.

   Guidelines:

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

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental setup is described in the main paper, and more detailed implementation settings and hyperparameter choices are provided in the appendix to support reproducibility.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We did report error bars or confidence intervals in the paper.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide information about the computer resources used in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our research complies with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: This work focuses on foundational metric learning and manifold evaluation. It does not have direct, foreseeable negative societal impacts.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risk.

Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Our paper gives proper acknowledgement of the original papers for the assets used in this work.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [N/A]

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

Guidelines:

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

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

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

Answer: [N/A]

Justification: Our paper does not involve human subjects and therefore does not require IRB approval.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: Our paper does not use LLMs as an important, original, or non-standard component.

Guidelines:

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